clacCÍRCULO
de
lingüística
aplicada a la
comunicación

63/2015

EVALUATING SHORT-TERM CHANGES IN L2 COMPLEXITY DEVELOPMENT

Bram Bulté and Alex Housen
Vrije Universiteit Brussel
Bram Bulte *at* vub ac be, alex housen *at* vub ac be

Abstract

This paper reports on a study on the nature and extent of the development of English L2 writing proficiency of 45 adult ESL learners over the time of an intensive short-term EAP program as evaluated by means of objective measures targeting different components of lexical and syntactic complexity. In addition, we compare the scores on these measures with more holistic and subjective ratings of learners' overall writing quality. Results reveal that some measures, but not necessarily the most popular linguistic complexity measures (e.g., subordination ratios and lexical richness measures), can indeed adequately and validly capture development in L2 writing in short-term ESL courses. Results further suggest that different subcomponents of syntactic and lexical complexity in L2 writing develop at different rates, which stressed the importance of calculating a sufficiently wide range of complexity measures in order to obtain a comprehensive picture of L2 development.

Bulté, Bram, and Alex Housen. 2015.
Evaluating short-term changes in L2 complexity development
Círculo de Lingüística Aplicada a la Comunicación 63, 42-76.
http://www.ucm.es/info/circulo/no63/bulte.pdf
http://revistas.ucm.es/index.php/CLAC
http://dx.doi.org/10.5209/rev_CLAC.2015.v63.50169

Key words: ESL, second language acquisition, complexity development, syntactic and lexical complexity.

Contents

1. Introduction

This paper reports on an investigation of the possibility of measuring short-term gains in L2 writing proficiency by instructed upper-intermediate learners of English, and of the validity of a range of quantitative metrics of L2 complexity against the benchmark of experienced judges' perceptions of L2 writing quality. Twenty-five years ago, Charles Alderson, one of the founding fathers of applied linguistics, in a wide-ranging review of language testing, argued that the development of progress-sensitive tests and measures was a major task for language assessors (Alderson 1990; see also Westeway et al 1990).

However, the complexity of this task was shown by the fact that many years later Alderson (2000) was still campaigning for ways to chart gains by learners on, for instance, English for academic purposes programmes, as the one studied in this paper.

Some commentators have doubted the possibility of much progress in speaking or writing skills over courses of two to four months, even with intensive study (Lennon 1995; Politzer & McGroarty 1985). While there are empirical findings supporting this pessimism (e.g., Rifkin 2005; Storch 2009), at least where typical standardised testing procedures are used, and in non-immersion situations, course providers, teachers and students still expect learners' productive skills to progress during such short intensive courses (Cumming 1995; Leaver & Shekhtman 2002; Tonkyn 2012; Wette 2010; White 1994). Therefore language assessors and instructors must take up Alderson's challenge of providing appropriate progress-sensitive measures for these contexts. Since productive L2 proficiency, and L2 progress, are typically measured by subjective ratings by skilled evaluators (e.g., teachers), it is also important to know which linguistic features of L2 performance correlate with, and may determine overall perceptions of progress by such judges. We investigate the possibility of measuring short-term gains in L2 writing proficiency in terms of features of linguistic *complexity*, and examine the adequacy of selected quantitative complexity measures as indicators of such proficiency and progress. In L2 research, as in L1 research, complexity has been proposed as a valid and basic descriptor of L2 performance, as an indicator of proficiency and as an index of language development and progress (Housen & Kuiken 2009).

The next section introduces the construct of L2 complexity and discusses how complexity can be, and has been, defined and operationalised in L2 research, and how it relates to other central notions in L2 research such as *proficiency*, *performance*, *progress*, and *development*. The 3rd section formulates the specific research questions we sought to answer and the design and methodology of the study. Section 4 presents the results of the analyses, and Section 5 interprets and discusses the results and relates them to the results of previous studies of complexity in L2 writing. Section 6 summarizes the main findings, discusses implications of the present study for L2 (writing) research and suggests directions for future complexity research on L2 writing and L2 writing development.

2. L2 Complexity

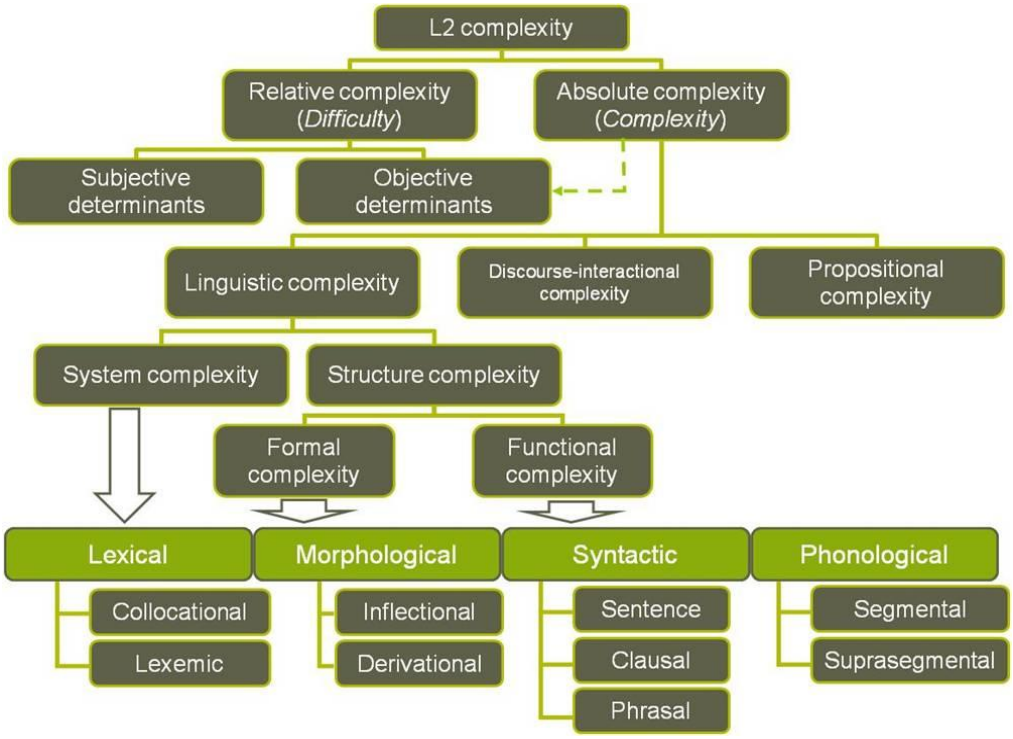
Complexity has become a fixture in contemporary science, particularly in biology, physics, chemistry, philosophy, psychology and sociology (see Mitchel 2009 for an overview). Complexity also figures prominently in the language sciences, particularly in functional approaches to language typology, evolution and contact (see Dahl 2004; McWorther 2001, 2011; Givon 2009; Sampson, Gil & Trudgill 2009). Complexity also has a long history in applied linguistics and second language acquisition (SLA) research. Already in the 1970s and 1980s the notions of *simplification* and *complexification* played an important role in the first models of SLA. Linguists such as John Schuman and Roger Andersen compared SLA to the processes of pidginisation and creolisation: SLA was seen as a process of gradual complexification, as the development from a lexically and structurally simple basilectal interlanguage variety to an increasingly complex acrolectal variety of the target language. Also the theoretical model that came out of the well-known ESF project invoked the notions of complexification and simplification to describe the development and structure of what they called the Basic Variety and learners' move beyond the Basic Variety.

However, despite the interest it has generated, there is no agreement in the L2 literature on the definition of complexity, and no consistency as to how it has been operationalised across (and sometimes even within) studies. This has led to terminological and conceptual confusion and has made it hard to interpret and compare over the results of individual studies (Bulté & Housen 2012; Norris & Ortega 2009; Pallotti 2009).

Norris & Ortega (2009) and others have argued that complexity is a highly complex construct, consisting of several sub-constructs, dimensions, levels and components, each of which can, in principle at least, be independently evaluated. Bulté & Housen (2012) attempt to capture this multidimensionality, by presenting a taxonomic model of different approaches to, and components of, language complexity as it has been applied and interpreted in L2 research (cf. Figure 1). A first and basic distinction is between absolute and relative complexity: "[i]n linguistics, complexity refers to both the […] internal structuring of linguistic units and to the psychological difficulty in using or learning them" (Crystal, 1997, p.76). Absolute complexity derives from *objective* inherent properties of linguistic units and/or systems thereof (as dictated by linguistic theories, hence 'objective') while relative complexity implies cost and difficulty of processing or learning, which could arise from both user/learner-related variables (and

hence 'subjective') (e.g. aptitude, age, motivation, stage of development) but also from more objective factors such as a language feature's input saliency and frequency as well as its objective inherent complexity properties. In what follows we use the term ‘difficulty’ to refer to relative (or: psychological, cognitive) complexity and reserve the term 'complexity' (and its derivatives such as 'L2 complexity') for absolute complexity as a manifestation of objective properties of linguistic units and (sub-)systems thereof (see also Skehan's (2003) distinction between *cognitive complexity* and *code complexity* respectively). Bulté & Housen (2012) further distinguish between three components of L2 complexity (in the narrow sense of the term): propositional complexity, discourse-interactional complexity and linguistic complexity. Of these three, linguistic complexity has received by far the most attention in L2 writing research, and this will also be the focus of the present study. Linguistic complexity can be investigated both at the level of the language system as a whole (or of its major subsystems and layers) as well as at the level of the individual linguistic features (forms, items, structures, patterns, rules) that make up such (sub-)systems. The complexity of these structures can in turn be studied in terms of their formal and functional properties (hence, structures may differ in terms of their *formal* and *functional complexity*). Finally, all these different types, components and subdimensions of complexity can be studied across various domains or layers of language such as the lexicon, syntax and morphology.

Figure 1: Taxonomic model of L2 complexity (Bulté & Housen 2012)



The taxonomy in Figure 1, though by no means exhaustive, aptly illustrates the multicomponential and multidimensional nature of complexity. This multidimensionality is still insufficiently reflected in empirical L2 research, including L2 writing research. Even though complexity figures prominently in recent L2 writing studies, few studies provide an explicit construct definition of complexity or a detailed characterisation of what type of complexity they investigate. Instead, complexity in L2 (writing) research is typically only operationalised as a behavioural or statistical construct (Bachman 2005), usually in terms of quantitative measures (Bulté & Housen 2012; Norris & Ortega 2009). General discussions of the validity and reliability of complexity measures in L2 (writing) research include Wolfe-Quintero et al (1998), Ellis & Barkhuizen (2005) and Ortega (2003) and Malvern et al (2004). There is a wealth of complexity measures available in the L1 and L2 acquisition literature – Bulté & Housen (2012) counted no fewer than forty different complexity measures in a sample of forty empirical L2 studies published between 2005 and 2008 (e.g. words/T-unit, clauses/sentence, number of subordinate clauses, dependent clauses/ total clauses, word types/word token, number of passive forms, number of relative clauses). With few exceptions, the current repertoire of measures in L2 research targets syntactic and lexical complexity - other dimensions and levels of linguistic complexity (e.g. morphological complexity) are rarely measured. Moreover, most L2 studies typically calculate only one or two or two complexity measures, usually from the same small set of 'popular' measures, viz. mean length of unit, subordination ratios, and lexical type/token-ratios. As a result, the multidimensional construct of complexity is reduced to one (or, at best, a few) of its many possible operationalisations and, as a result, complexity measurement practices in extant L2 research suffer from low content validity (Bulté & Housen 2012; Ortega 2012). Current L2 complexity measures reflect five related assumptions about linguistic complexity: 1) 'more is more complex' (e.g. more phonemes, inflectional forms or categories, grammatical derivations equate with more complexity), 2) 'longer linguistic units are more complex' (e.g. greater word, phrase, clause sentence, text, … length is assumed to index higher complexity), 3) 'more and/or more deeply embedded is more complex' (e.g. more recursion, more subordinated and more deeply embedded subordinated features is more complex), 4) 'more varied or diverse is more complex' (e.g. more different types of lexical or

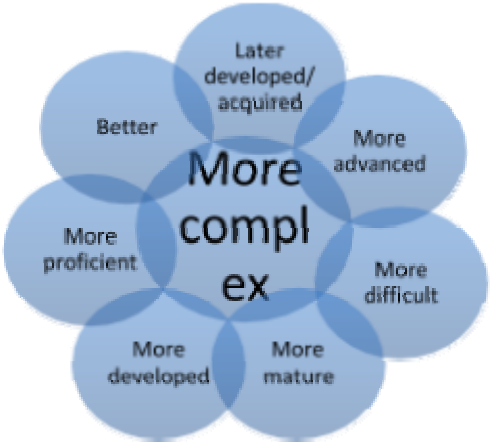
grammatical forms is more complex), 5) 'more marked, infrequent, sophisticated, semantically abstract, costly, cognitively difficult or later acquired features are more complex'. While all of these notions fall squarely within the remit of 'more is more complex', i.e. absolute-quantitative complexity (Ortega 2012, Szmrecsanyi & Kortmann 2012), many also come into the area of relative, cognitive complexity or *difficulty*. In this sense, most L2 complexity measures are hybrid measures, which detracts from their construct validity (Bulté & Housen 2012).

Ortega (2012) further points out that L2 researchers use complexity measures "with at least three main purposes in mind: (a) to gauge proficiency, (b) to describe performance, and (c) to benchmark development" (p. 128). Thus complexity has been investigated for a variety of purposes but rarely for its own sake in L2 (writing) research. Instead, complexity measures mainly serve as indicators, diagnostics or proxies for other, more general or higher-order constructs such as *L2 (writing) proficiency*, *L2 (writing) development* and *L2 (writing) quality* or *maturity* (Wolfe-Quintero et al 1998; Ortega 2003). L2 writing studies that have used complexity measures to these ends include studies of the effects on L2 writing performance and development of specific instructional treatments under (quasi-)experimental conditions (Kuiken & Vedder 2011; Hartshorn et al 2010), of specific second/foreign language program types such as immersion and EAP programmes (e.g. Byrnes, Maxim & Norris 2010; Vyatkina 2012) or of specific L2 learning contexts such as study-abroad vs. study-at home contexts (e.g. Storch 2009; Serrano et al 2012). More recently, DST-inspired approaches to SLA have analysed complexity in L2 writing by tracking the written productions of individual learners over longer periods of time in an attempt to reveal the internal developmental dynamics of L2 development (e.g. Verspoor et al 2008; Spoelman & Verspoor 2010; Verspoor, Schmid & Xu 2012). Collectively, these studies have shown that, with substantial exposure and/or intensive targeted writing instruction, L2 learners' scores on complexity measures increase over time or as general L2 proficiency develops (see also Ortega 2003), though many studies also failed to find statistically significant increases in complexity as assessed by such measures (Wolfe-Quintero et al 1998; Ortega 2003).

The fact that complexity has rarely been investigated for its own sake or as the central variable probably explains why the construct is still ill-defined in the L2 literature. As figure 2 shows, L2 complexity has been differentially characterized as 'difficult to

acquire or to produce', 'acquired late(r)', ‘developmentally advanced’, 'more proficient', 'more mature', 'of high(er) quality' or simply as 'better' (Bulté & Housen 2012).

Figure 2: L2 complexity and other constructs



Of particular relevance for this study is the use of complexity as a metric of *L2 development*, a practice that goes back to at least the mid-1970s (Larsen-Freeman & Strom 1977). This practice is based on the implicit yet widely held assumption that L2 complexity increases in the course of development, and that it increases linearly, so that more development leads to the use of more complex language and structures (i.e. a wider range of more sophisticated vocabulary, more sophisticated or more complex grammatical structures, etcetera).

However, linguistic complexity measures cannot be validated simply by showing that they increase in the course of L2 development. Developmental timing may give an indication of the *difficulty* (i.e. *cognitive* complexity) of an L2 feature or of a subsystem of L2 features, but as we have argued earlier, difficulty is conceptually distinct from *linguistic* or *structural* complexity (cf. Figure 1). Whether, or to what extent, linguistic complexity increases over time needs to be established empirically rather than be taken for granted. This is important to avoid the risk of circular reasoning which looms large in L2 complexity research. Many studies have interpreted L2 complexity in terms of one or several of the concepts listed in Figure 2, or have assumed that they are isomorphic: more complex language or more complex linguistic structures are taken to be more difficult (cognitively taxing) structures, more difficult structures are seen as structures that are developed or acquired late(r), and later acquired/developed structures are taken to be more advanced while the use of more difficult and more advanced language is in turn seen as a hallmark of 'better', more mature or more proficient language which in its

turn is taken to be more complex, etcetera. In order to avoid such circularity of reasoning, *complexity*, *development*, *proficiency* and the other constructs listed in Figure 2 need to be kept conceptually distinct from each other.

Exhaustively defining *complexity*, *development* and *proficiency* – one of the main goals of an entire subfield of linguistics and applied linguistics – is well beyond the scope of this paper. First, for present purposes, and following work in typological linguistics (e.g. Dahl 2004; Miestamo 2008), we propose to define *complexity* as much as possible as an absolute, objective and essentially quantitative property of language units, features and (sub)systems thereof in terms of (i) the number and the nature of discrete parts that the unit/feature/system consists of and (ii) the number and the nature of the interconnections between the parts. To put it simply, the more components a feature or system consists of, and the more and the more dense the relationships between its components, the more complex the feature or system is. Second, we define *L2 proficiency* somewhat loosely here as "a person's overall competence and ability to perform in L2" (Thomas 1994, p. 330). A learner's L2 proficiency is typically inferred from assessments of concrete instances of L2 use and production (e.g. essays). Finally, the notion of *L2 development* relates to the changes in the L2 proficiency of a learner over time and is again typically (though not exclusively) inferred from the observation of changes in concrete samples of L2 production collected at different times, such as essays or other writing samples in the case of writing production. *L2 development* is further often thought of in terms of 'growth' of the L2 system (knowledge) of a learner and in 'progress' towards a particular target or norm. Clearly, L2 proficiency and L2 development, like L2 complexity, are multidimensional and multicomponential constructs, and their different dimensions and components – of which complexity is but one – interact with each other over time.

3. Research questions

The main focus of this contribution is on the validity of complexity as a dimension of L2 writing development and L2 writing quality, and the validity of syntactic and lexical complexity measures as indicators thereof. Thus the general research questions that guide this study are as follows:

1. Which aspects of the linguistic (syntactic, lexical) complexity of the writing production of instructed intermediate/advanced learners of L2 English, as

measured by a battery of quantitative metrics, change (progress) during the relatively short term of a typical intensive EAP course?

2a. How do scores on quantitative linguistic complexity metrics correlate with subjective ratings of *writing quality* by experienced judges?

2b. Which linguistic complexity metric(s), or combinations thereof, best predict subjective ratings of *writing quality*?

4. Design, Materials and methods

4.1. Participants and data

The corpus[1] analysed in this study consists of 90 essays written by 45 L2 learners of English who were enrolled at Michigan State University (MSU) in a variety of study programmes, and who had to follow English for Academic Purposes (EAP) courses provided by the MSU English Language Center. These courses are taught over the course of one semester (12 weeks) for 3 hours per week. The participants in this study were enrolled in the third and fourth level courses, corresponding to the intermediate and lower-advanced levels of English proficiency. Little background information about the individual learners is available. The learners come from a variety of language backgrounds; from the content of the essays it can be inferred that most of the students come from Asian countries, such as Korea, Taiwan, and Thailand. The exact age of the learners is unknown though most of them were in their twenties at the time of data collection.

We analyse two essays per learner, one written at the beginning and one written at the end of the semester-long EAP programme. The essays comprise a wide range of topics, all related to the personal lives of the writers in order to allow them to be more engaged, as well as more at ease under the test conditions and thus better able to demonstrate their English writing abilities (Read, 2005, p. 198). Topics include descriptions of their families, of their high school in their home countries, holidays, or their experience when

[1] We thank Charlene Polio for providing us with the larger dataset the corpus that was used in this study was taken from.

they arrived in Michigan. The learners disposed of 30 minutes to complete each writing task, and they did not have access to outside sources.

4.2. Instruments and measures

The ninety essays were evaluated by means of both subjective ratings of writing quality as well as by a selection of quantitative measures gauging different aspects of L2 complexity.

Complexity measurement

We calculated a total of thirteen complexity measures, ten targeting different aspects of syntactic complexity, and three targeting lexical complexity.

Syntactic Complexity Measures

Table 1 lists the ten measures of syntactic complexity. These measures are based either on the *average length* (in words) of different linguistic units (sentences, T-units, finite clauses, noun phrases) or on a *ratio* of a specific subtype of a linguistic unit to a more general subtype or a higher-order unit. These ten syntactic complexity measures were chosen to gauge complexification at different layers of syntactic organisation, to wit the sentential, the clausal and the phrasal level. According to Norris & Ortega (2009), L2 learners complexify these different levels of syntactic organisation at different stages of development so that all three levels must be measured in order to cover the full trajectory of L2 development. Thus, three sets of measures targeting *sentential* syntactic complexity were selected, each capturing a different (though related) aspect of sentence complexity. The first set targets sentence complexity in terms of the mean length of sentential unit in words: mean length of sentence (MLS) and T-Unit (MLTU). The second set captures sentence composition in terms of clauses as defined by traditional grammars (e.g. Huddleston & Pullum 2002; Verspoor & Sauter 2000) and consists of four linguistically dependent measures: the simple sentence ratio (SSR), compound sentence ratio (CdSR), complex sentence ratio (CxSR) and the compound-complex sentences (CdCxSR). The third set gauges sentential syntactic complexification in terms

of proposition combining and clause integration strategies: the coordinate clause ratio (CCR), the subclause ratio (SCR).

Table 1: Syntactic complexity measures

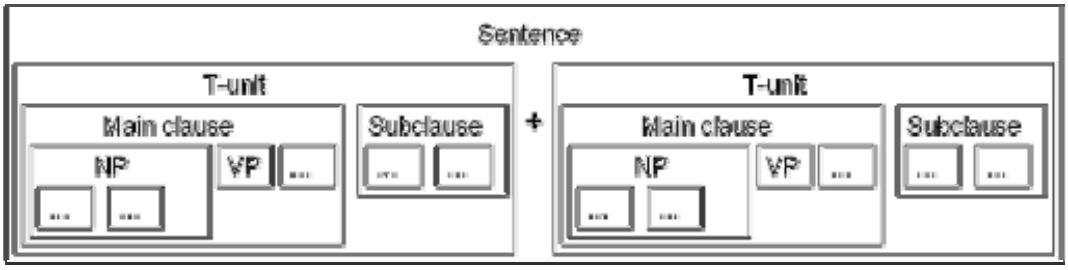
- **Syntactic sentential complexity:**
  - **Length of sentential unit**:
    - mean length of sentence (*MLS*)
    - mean length of T-unit (*MLTU*)
  - **Sentence composition**:
    - Simple sentence ratio (*SSR*)
    - Compound sentence ratio (*CdSR*)
    - Complex sentence ratio (*CxSR*)
    - Compound complex sentence ratio (*CdCxSR*)
  - **Proposition combining and clause linking**:
    - Coordinate clause ratio (coordinated clauses / sentence) (*CCR*)
    - Subclause ratio (subclauses / clause) (*SCR*)
- **Syntactic Clausal complexity** : mean length of finite clause (MLC_{fin}).
- **Syntactic Phrasal complexity** : mean length of noun phrase (*MLNP*).

Most of these eight sentential syntactic complexity measures are well tried metrics in L1 and L2 research (e.g. Ortega 2003, Ellis & Barkhuizen 2005, Spoelman & Verspoor 2010; Verspoor, Schmidt & Xu 2012; Kuiken, Vedder & Gilabert 2010). In contrast, the measurement of syntactic complexity at the clausal and phrasal level is only a fairly recent development in L1 and L2 complexity research, and the number of available measures is still limited. Therefore only one complexity measures for each of these two syntactic layers was calculated, mean length of finite clause (MLCfin) for clausal complexity and mean length of noun phrase (MLNP) for phrasal complexity.

For the syntactic complexity analysis, the ninety essays were segmented in clausal units in Microsoft Excel sheets and manually analysed and annotated for the following syntactic units and features: sentence type (simple, compound, complex, complex-compound); T-units; main, coordinated and subordinated clauses; finite verbs/clauses; noun phrases and words per noun phrase. Figure 3 shows how the different linguistic units targeted by these measures are hierarchically related.

All essays were analysed, annotated and counted by two researchers, and checked by a third, using a set of explicit guidelines and linguistic criteria outlined below. Inter-coder agreement initially varied from 85% (e.g. for the identification of NPs and of simple sentences vs. compound sentences with coordinated independent clauses) to nearly 100% (e.g. for the identification of subordinate clauses and finite verbs/clauses). All disagreements were discussed until agreement was reached. The unit counts were manually inserted in the spreadsheets where they served as input for the automatic calculation of the syntactic measures.

Figure 3: Linguistic units targeted by L2 complexity measures



Linguistic guidelines and criteria for analysis and coding: For the purpose of this study we adhered to a linguistic definition of a clause, as a unit consisting of a subject (explicit or implied) plus a predicate, i.e. a construction with a finite or non-finite predicator or verb as its nucleus. As a result, verb constructions such as *go look*, *tries finding, keeps shouting* and *starts to look* are all analysed as consisting of two clauses, a main clause plus a non-finite subordinate complement clause. Subclauses comprise adverbial clauses, complement clauses and relative clauses. A T-unit consists of one independent clause with all of its dependent (subordinate) clauses. In contrast to a T-unit, a sentence can also include two or more coordinated independent clauses. Sentences can become longer by adding more coordinated and/or subordinated clauses, when their constituent clause(s) contain more constituents and phrases, and when the phrases that make up these clauses contain more words. T-units, however, do not become longer when the writer adds coordinated clauses. In this sense, the mean length of sentence gives the same weight to coordinated and subordinated clauses, whereas the T-unit measure does not.

Noun phrases (NPs) can be made more complex through compounding, through multiple determiners (pre-, central and post-determiners) and through complementation and pre- and post-modification (by embedding other phrases or clauses). NPs can themselves be embedded in prepositional phrases (PPs), PPs can be embedded in NPs, and also different types of clauses (i.e. complement and relative clauses, each with their own inherent structure and possibly further embedded NPs) can be embedded in a NP. NPs with a pronoun as head were excluded from analysis as they do not allow for pre- or post-modification in English. NPs headed by a proper noun were excluded for the same reason. NPs embedded in PPs (e.g. *in [my country]*)were included in the analysis. Compound NPs (e.g. *boys and girls*) were counted as consisting of two (or more) separate heads (and therefore, in practice, also as two separate phrases), and the conjunction linking the nouns was counted only once. All words in clauses embedded in NPs were counted as words of the NP (e.g. *people [who come from outside this city]* contains seven words). NPs embedded in another NP (e.g. as part of a postmodifying or complementing PP as in *the repartition [of this population]*) were counted and analysed separately, but their constituent words were also included when calculating the length of the superordinate phrase (this is also true for NPs embedded in subordinate clauses that function as a dependent of a NP, such as relative clauses).

Lexical Complexity Measures

The three measures of lexical complexity target three related yet distinct aspects of lexical complexity (see Table 2): lexical diversity, richness and sophistication (Skehan 2009a,b). The diversity index *D* (Malvern et al 2004) served as an index of lexical diversity or variety. This index is a mathematical transformation of the standard type-token ratio (TTR) intended to reduce the intervening effects of text length and basically provides an indication of the degree of words repetition in a text. The fewer words are repeated, and thus the more different words that are used in a text, the higher the score for *D* (see however McCarthy & Jarvis 2007, 2010 for a critical discussion of this measure, and Jarvis 2013 for an in-depth discussion of the notion of lexical diversity). The second lexical complexity measure, the index of Guiraud *G* (Guiraud 1959), is

another transformation of the simple TTR. Most studies have used G as an index of lexical diversity, like D. However, we argue that G measures something more than sheer diversity. The mathematical transformation that the Guiraud index uses to control for text length effects in the calculation of the TTR (a square root in the denominator) has been shown to overcompensate for the decrease in scores with increasing text length, so that not only texts with fewer repetitions but also longer texts obtain higher scores for G (Bulté 2007; Bulté et al 2008). Since text length, or number of words produced (lexical *productivity*) has been considered a crude indicator of lexical complexity, and has been found to correlate well with proficiency level and subjective ratings of writing quality (Chuming 2005; Engber 1995; Daller & Phelan 2007; Larsen-Freeman 1978; Laufer & Nation 1995), we deemed G to be a useful complement to the Diversity index, especially for the analysis of timed writing samples that are not controlled for length as in the case of the MSU corpus. Finally, we calculated a measure of lexical sophistication, the advanced Guiraud index (AG; Daller et al 2003). AG indicates the extent to which a learner uses 'advanced' words, that is words that occur less frequently in language use. Whereas a link between the frequency of words (in the input) and their *difficulty* appears logical (though even this requires further empirical demonstration), the link between frequency and complexity as defined in this study is less obvious. The underlying logic is that the use of less frequent items would point to a larger vocabulary, consisting of more elements (see also Jarvis 2013). Also from an absolute, information-theoretic complexity point of view (Dahl 2004), frequent words can be argued to be less complex than non-frequent words because frequent items carry lower amounts of information (Juola 2008; Ehret & Szmrecsanyi, in press).

Table 2: Lexical complexity measures

Lexical Diversity: diversity/variation in the use of word types (lemmas): Diversity index (*D*)
Lexical Richness: variation in and number of word types used: Guiraud index (*G*)
Lexical Sophistication: variation in and number of 'basic' (frequent) vs. 'advanced' (less frequent) word types used: Advanced Guiraud (*AG*)

In contrast to the syntactic complexity analyses, the lexical complexity analyses were performed with the aid of automated tools after spelling mistakes in the essays had been corrected and proper nouns and interjections had been deleted. The lexical diversity index *D* was calculated with the VocD command in CLAN (MacWhinney 2000). RANGE (Heatley, Nation & Coxhead 2002) was employed to facilitate the calculation of the lexical sophistication measure *AG*. The distinction between 'basic' and 'advanced' words was made on the basis of word frequency lists compiled by Paul Nation (see Nation 2006). *Range* compares a written text against three ready-made word frequency lists derived from the statistical analysis of large language corpora, and it indicates how many words the text contains from each of the different word frequency levels (or 'bands'): (1) the list of the most frequent 1000 words, (2) the second 1000, and (3) the Academic Word List (Coxhead 2000). We considered word types to be advanced if they did not appear in the three previous lists.

Ratings

Two expert raters evaluated the essays using a rating scale, developed at MSU, targeting five different aspects of writing quality (see Appendix 1). For each of the five categories the students were given a score between 1 and 20. The mean score of the two raters' evaluations was used for our analyses. For the analysis of the relationship between the quantitative complexity measures and the subjective ratings we used three scores: (i) the mean total score of all five rating scales, and the scores of the individual scales (ii) Language Use and (iii) Vocabulary. These two specific scales are most closely related to the constructs of respectively syntactic and lexical complexity as targeted by our complexity measures. This is illustrated by the excerpts from the rubrics used for the rating scales in Table 3 (e.g. "frequent use of complex sentences", "range of vocabulary"). Other descriptors from these two rating scales are more closely related to other constructs such as accuracy (e.g. "no errors that interfere with comprehension", "idiomatic, native-like").

Table 3: Rating scales and descriptors

Mean total rating score	Language Use rating score	Vocabulary rating score
Combination of rating on 5 scales: Content Organisation Language use Vocabulary Mechanics	Descriptors: "no major error in word order and complex structures" "no errors that interfere with comprehension" "only occasional errors in morphology" "frequent use of complex sentences" "excellent sentence variety"	Descriptors: "sophisticated vocabulary" "choice of words" "range of vocabulary" "errors" "idiomatic, native-like" "academic register" "repetitive"

Statistical analyses

In order to investigate whether any changes had occurred over time in the L2 writings of the forty-five L2 learners (research question 1), we calculated mean scores and standard deviations for the two data collection points, and used paired samples t tests to check for the significance of the differences observed. Effect sizes (Cohen's *d*) were calculated to gauge the strength of the effect. Pearson correlations were used to assess the relationship between the quantitative measures and the subjective ratings (research question 2a), and a stepwise multiple linear regression with the mean overall writing quality rating scores as dependent variable and different complexity metric scores as independent variables was used to analyze which complexity measures best predict the subjective ratings of L2 writing quality (research questions 2b).

5. Results

5.1. Complexity Development

The first research question was whether and, if so, which aspects of the syntactic and lexical complexity of the L2 writing of the university-level learners of L2 English change over the course of their intensive EAP program. Table 4 shows the mean scores (and standard deviations) at the beginning (T1) and at the end (T2) of the semester for the thirteen measures of lexical and syntacic complexity as well as for the three subjective rating scales that were retained for this study, together with the p- and t-values of the paired samples t tests and the estimated effect sizes (Cohen's *d*). Statistically significant results are marked by one ($p \leq 0.05$) or two asterisks ($p \leq 0.01$).

Table 4: Developmental changes in complexity measures and holistic ratings over time

	Time 1	**Time 2**	**p (t value)**		***d***	
Syntactic complexity						
MLS	12.62 (2.98)	13.96 (3.03)	.005** (-2.956)		.441	
MLTU	10.78 (2.46)	11.75 (2.52)	.003** (-3.140)		.468	
SSR	41.3 % (15.1)	33.5 % (14.6)	.006** (2.887)		.430	
CdSR	6.9 % (6.0)	10.7 % (9.2)	.012* (-2.637)		.393	
CxSR	39.0 % (14.4)	41.4 % (14.6)	.330 (-0.984)		.147	
CdCxSR	13.8 % (11.1)	14.8 % (10.0)	.616 (-0.505)		.075	
CCR	0.18 (0.11)	0.24 (0.16)	.041* (-2.102)		.313	
SCR	0.40	0.41	.773		.043	

	(0.12)	(0.10)	(-0.290)			
MLCfin	7.25 (1.14)	7.86 (1.33)	.002** (-3.310)		.493	
MLNP	2.81 (0.65)	3.08 (0.70)	.009** (-2.726)		.406	
Lexical complexity						
G	8.33 (1.04)	8.31 (0.96)		.879 (0.153)		.023
D	71.95 (14.41)	69.75 (14.12)		.374 (0.899)		.134
AG	0.573 (0.28)	0.605 (0.40)		.534 (-0.627)		.094
Subjective ratings						
Overall Writing Quality	48.56 (10.56)	57.16 (8.24)		.000** (-6.896)		1.028
Language use	9.86 (2.16)	11.09 (1.78)		.000** (-4.817)		.718
Vocabulary	9.72 (1.76)	11.17 (1.43)		.000** (-7.367)		1.098

The scores on all syntactic complexity measures increase from T1 to T2 and for all but three sentential complexity measures (CsCR, CdCxR, SCR) the increase is statistically significant. With regard to sentential syntactic complexity, by the end of the four-month course, the learners wrote sentences that were on average around 1.4 words longer and T-units of around 1 word longer. When we look at sentential syntactic complexity as sentence composition in terms of clauses and the different mechanisms for combining propositions under clausal frames, we observe a significant decrease in the percentage of simple sentences (T1: 41.3%; T2: 33.5%) and a significant increase in compound sentences (T1: 6.9%; T2: 10.7%). Also the number of coordinated clauses per sentence increased significantly from T1 (0.18) to T2 (0.24). Suprisingly, the proportion of subordinated clauses and of sentences containing at least one subclause (i.e. complex

and compound complex sentences) did not change in a statistically significant way. At the level of the clause, a significant increase in finite clause length is observed (T1: 7.25 words/Clfin; T2: 7.86 words/Clfin). Also the length of NPs increases significantly (T1: 2.81 words/NP; T2: 3.08 words/NP), pointing to increased use of determiners and modifiers of the NP head. The highest effect sizes for the syntactic complexity measures were found for MLC_{fin} (d= 0.493), MLTU (d= 0.468) and MLS (d= 0.441).

In contrast to the syntactic complexity measures, only one out of the three lexical complexity measures, the lexical sophistication measure (AG), shows an increase from T1 to T2 but this increase is not statistically significant. The scores on the lexical diversity (D) and lexical richness (G) measures decreased slightly and non-significantly over time.

In comparison, the scores given by two raters on the three subjective rating scales Vocabulary, Language Use and Overall Writing Quality (the sum of five more specific rating scales, including Vocabulary and Language Use) all significantly increase from T1 to T2, suggesting a growth of perceived writing quality over time. The effect sizes show that the observed effect of this change over time is strong. The strongest effect size was found for the Vocabulary rating scale (d= 1.098), followed by the composite Overall Writing Quality scale (d= 1.028) and finally the scale for Language Use (d= 0.718).

It is further interesting to note that in this study the effect sizes for the subjective ratings of (different components of) writing quality are much higher than for the objective complexity measures when it comes to showing development over time. This might raise questions as to the progress-sensitivity of quantitative complexity measures. However, it should be pointed out that the quantitative measures calculated in this study target specific components and aspects of complexity, whereas the subjective ratings are more holistic in nature. In this sense, it would be worthwhile to look at the combined effect of complexity measures (see Byrnes et al 2010; Bulté 2013).

5.2. Objective complexity measures and subjective ratings of writing quality

Table 5 shows the correlations between the scores on the syntactic complexity measures and the subjective ratings for Overall Writing Quality and Language Use (the scale most closely related to syntactic complexity), and Table 6 does the same for the scores on the

lexical measures and the ratings for Overall Writing Quality and Vocabulary (the scale most similar to lexical complexity). Pearson correlation coefficients are provided, and significant correlations are again flagged with one ($p \leq 0.05$) or two asterisks ($p \leq 0.01$).

Table 5: Correlations between syntactic complexity measures and holistic ratings

	Overall WQ rating	**Language Use**
MLS	.413**	.423**
MLTU	.403**	.432**
SSR	-.431**	-.447**
CdSR	.110	.051
CxSR	.214*	.213*
CdCxSR	.179	.241*
CCR	.112	.089
SCR	.239*	.290**
MLCfin	.476**	.491**
MLNP	.358**	.373**

Table 6: Correlations between lexical complexity measures and holistic ratings

	Overall WQ rating	**Vocabulary**
G	.521**	.521**
D	.161	.178
AG	.068	.128

Significant modest-to-strong correlations are observed between the subjective writing quality ratings and slightly over half of the complexity metrics. Differences between the results for the overall writing quality scale and those for the two more specific scales are slight, which is not surprising given the strong correlations among the scores on the different rating scales themselves ($r=0.873$ between Language Use and Vocabulary). The strongest correlations are found for G ($r=0.521$), MLCfin ($r=0.476$) and SSR ($r=-0.431$). Non-significant and weak correlations characterized the relationships between writing quality ratings and clause coordination (CdS, CCR), lexical diversity (D), lexical sophistication (AG) and complex (compound) sentences (CxS, CdCxS).

Finally, we performed a stepwise multiple linear regression analysis to identify which (combination of) objective complexity metrics best predict the subjective ratings of Overall Writing Quality. For this purpose, we used Overall Writing Quality as dependent variable and entered the different quantitative measures as independents. The analysis yielded a significant model that explains 45% of the variance in perceived overall writing quality ($F(4, 89) = 17.672$; $p < 0.001$; $r = 0.67$; $R^2=0.45$) and includes the following four variables: the Guiraud index (G), mean length of noun phrase (MLNP), the proportion of simple sentences (MLC_{fin}) and the subclause ratio (SSR). Detailed statistics are reported in Table 7.

Table 7: Coefficients multiple linear regression model

	Unstandardised Coefficients		**Standardised Coefficients**	**t**	**Sig.**
	B	Std. Error	Beta		
(Constant)	20.609	10.233		2.014	.047
G	4.719	.898	.452	5.255	.000
MLNP	4.879	1.284	.323	3.801	.000
SSR	-27.026	7.495	-.398	-3.606	.001
SCR	-28.003	11.134	-.286	-2.515	.014

6. Discussion

6.1. Complexity development over time

A first objective of this study was to examine what changes, if any, occurred in the writing of English L2 learners during the relative short term of a typical intensive university ESL course - changes measurable in terms of lexical and syntactic complexity. In contrast to the observed changes in terms of syntactic complexity to be discussed shortly, the writings of the learners did not become more lexically diverse, rich or sophisticated in the course of the observation period. It appears then that lexical

and syntactic complexity do not develop in parallel in these data. This finding could be taken as support for Skehan's (2009a) and Foster & Tavakoli's (2009) proposal (based on empirical data and theoretical arguments in terms of processing mechanisms derived from Levelt 1989) that, at least for non-native users, lexical complexity constitutes a separate dimension of L2 performance and L2 proficiency, independent from grammatical complexity, rather than being different aspects of the same L2 performance-proficiency area. As these scholars themselves point out, this scenario for the development of lexical and grammatical complexity in SLA is still speculative and empirical research at other proficiency levels than the ubiquitously-studied low intermediates is vital in this regard. The written data in the present study, which include intermediate to advanced learners, while obviously not providing probative evidence, seem at least consistent with Skehan's et al account.

Syntactic complexity development is manifested in our analysis by a significant increase in the length of linguistic units at all levels of syntactic organisation (phrase, clause, sentence, T-unit). There is also an increase in clause coordination at the expense of the use of simple sentences but no significant increase in complex sentences or in subordination. Interestingly, the pattern of syntactic complexity development that emerges from our analyses does not correspond to the three-staged pattern for syntactic complexity development proposed by Norris & Ortega (2009). These authors argued that in the early stages of SLA, syntactic complexity is essentially established through clausal coordination. In a next, intermediate stage, subordination becomes the dominant means of syntactic complexification as the use of coordination trails off or diminishes. And at even more advanced stages of L2 development, further syntactic complexification would no longer be mainly established through subordination at the sentential level but increasingly through clausal and phrasal elaboration, that is, at the sub-sentential level. Our results point to a significant increase in both clausal coordination and phrasal elaboration, but not in subordination – which, to repeat, is one of the favorite complexity diagnostics in extant SLA research. This finding would suggest that changes in syntactic complexity do not follow the developmental pattern proposed by Norris & Ortega (2009) in any rigid or linear fashion.

Norris and Ortega (2009) and Ortega (2012) have also drawn on the distinction between dynamic styles (low formality, everyday contexts, typically oral) and synoptic styles

(high formality, specialized contexts, typically written) developed in Systemic Functional Linguistics (Halliday 1998) to account for L2 syntactic complexity development. In this framework, subordination, along with other clause combining mechanisms, is crucial for the development of dynamic styles but it is less relevant to the development of synoptic styles, which primarily rely on mechanisms such as grammatical metaphor through nominalizations and which are further characterized by higher lexical density, longer NPs through the use of multiple modifiers, as well as by a reduced number of combined clauses. The pattern which emerges from our analysis of the development of syntactic complexity features in the MSU data seems largely congruent with this account.

The developmental trends that emerge from our analyses of the dataset are probably most in line with dynamic systems accounts of L2 development (e.g. Verspoor et al 2008; Verspoor et al 2012; Larsen-Freeman 2006; Vyatkina 2012; Bulté 2013). These studies have indicated that complexity, accuracy and fluency, and selected sub-dimensions such as lexical and syntactic complexity and selected syntactic features such as subordination, coordination and phrasal elaboration, do not develop strictly successively nor linearly, especially when development is considered at short or medium-size time scales. Rather, their development is characterized by periods of growth and progress alternating with periods of stabilization or even temporary backsliding before progress picks up again (if at all). Since our study, as most previous empirical studies, only focuses on a rather short period within the entire developmental trajectory of our L2 learners, the patterns and trends we observed (or, indeed, failed to observe) may not be representative for the learners' overall long-term L2 development. Only longitudinal studies spanning sufficiently long observation periods (e.g. minimally three years) and with multiple and dense data collection points can adduce convincing evidence.

6.2. Link between quantitative measures and subjective ratings

Our second research questions targeted the link between the complexity measures and the subjective ratings. In this context, it has to be kept in mind that several of the descriptors used in the rating scales asked the researchers to focus on complexity or

complexity-related aspects of the essays[2]. The strongest correlations between complexity measures and perceived writing quality were found for lexical richness, clausal subordination, the mean lengths of (finite) clauses, sentences and of T-units and the proportion of simple sentences. Thus, particularly the use of many, and many different words in an essay (lexical richness) is seen as an indicator of higher writing quality, as is the use of longer units at the clausal and sentential level. Frequent use of simple sentences is perceived as a sign of lower writing quality. The use of compound and compound-complex sentences, clause coordination and, surprisingly, the use of more varied words (lexical diversity) and of less frequent words (lexical sophistication) contributes little or not significantly to the perception of a learner's overall writing quality. A combination of one lexical and three syntactic complexity measures (lexical richness, mean length of noun phrase, simple sentence ratio, subclause ratio) explains around 45% of the variance in the subjective ratings, and thus emerges as the prime aggregate complexity predictor of perceived Writing Quality in these data.

An interesting observation is that the measures that show significant development over time do not coincide with the measures that correlate significantly with the subjective ratings. For instance, the subjective ratings do not correlate with the measures of clausal coordination, even though these increased significantly over time. Conversely, the subordination ratio correlates well with the subjective ratings, although its scores did not increase from T1 to T2. Similarly, the highest correlation with the subjective ratings was observed for our measure of lexical richness (G) though this measure did not significantly increase over the course of the study.

In this respect it is interesting to note that whereas none of the measures of lexical complexity showed significant progress over the course of the study, the subjective ratings for 'Vocabulary' did increase significantly. This suggests either that the raters reacted to other aspects of the vocabulary of the essays which improved over time but which were not tapped by our lexical complexity measures (e.g. accuracy, appropriateness, register, specificity), or that the validity and/or degree of granularity of our objective lexical measures are moot. Whatever the case may be, there appears to be a need for measures targeting other aspects of lexical performance if lexical

[2] The (strong) correlations found can therefore also be interpreted as a positive indication of concurrent validity.

development in writing performance is to be captured by means of quantitative measures in the course of short-term intensive programs.

7. Conclusions

Even though complexity of form and structure is considered critical to measuring and describing L2 performance, L2 proficiency and L2 development, linguistic complexity is poorly defined in SLA and its sub-disciplines, including L2 writing research. We have argued that it is important to define complexity independently from related notions such as difficulty, development, proficiency, and L2 quality in order to avoid circular reasoning. Specifically, we investigated the potential of complexity as one of the possible axes for characterizing L2 development and L2 writing quality, quantifiable mainly in terms of the constituents of linguistic units and the relationships between such constituents. Even though we must obviously be cautious when drawing conclusions about the overall validity of complexity measures as measures of L2 writing development or L2 writing quality on the basis of the dataset analyzed here, our analyses do indicate that complexity measures can capture changes in L2 writing ability and quality over time, including over relatively short periods of time such as the ones afforded by typical EAP courses. Thus the pessimism that we referred to at the beginning of the paper, about the (im)possibility of developing progress-sensitive measures for charting gains by learners on short-term courses such as the EAP course of the learners in this study, this pessimism seems to be at least partly unfounded. Not only the holistic ratings but also more than half of the quantitative complexity measures in this study were able to capture growth in writing proficiency in the course of one 4-month semester.

Although the measures of linguistic complexity used in this study suggest ways of dealing with the challenge of measuring short-term gains in L2 writing proficiency, which should be of interest to researchers and practitioners concerned with identifying appropriate complexity measures for their specific contexts and ends, there remains the compelling question as to which measure(s) might be considered the best measure(s) of linguistic complexity. Clearly, the answer to this question first requires clarity about the purpose of complexity measurement in a given study: is the complexity measure to serve as an index of L2 development, as a diagnostic of L2 writing quality/ability, or as

tool to investigate linguistic complexity as such? In addition, the answer will require a larger-scale examination of the validity and reliability and a more careful consideration of the practicality of the various measures than was possible within the scope of this study. To this end, both more longitudinal complexity studies are needed over larger periods of time, and in which difference and variation occupy a central role, as well as a broader conceptual framework, such as that offered by dynamic or complex systems theory (Larsen-Freeman & Cameron 2008; Verspoor et al 2011).

Such disclaimers notwithstanding, our findings demonstrate the importance of rejecting the idea of a one-size-fits-all measure of L2 complexity. Rather, a sufficiently wide range of judiciously chosen complexity measures should be calculated in order to get a comprehensive picture of L2 complexity development, given its multidimensional, multilayered and non-linear nature. Our data yielded no significant development for some of the most popular complexity measures in the L2 literature, such as the ubiquitous subordination measures and the measures of lexical diversity and richness D and G. This corroborates the preliminary evidence cited by Ortega (2012) that subordination measures may not be adequate to gauge L2 complexity in all contexts and under all circumstances, and that they may actually be inadequate when dealing with advanced learners and language samples that tend toward the synoptic end of the stylistic continuum (as writing by nature often tends to do). This would mean that a set of at least two complexity measures is needed: one for measuring complexity in dynamic styles, typically at lower levels of proficiency, and one that captures complexity in synoptic styles, which are typically found in the writings of learners at the upper-intermediate and advanced levels of L2 proficiency.

This study also found that the complexity measures that show development over time do not necessarily coincide with the measures that correlate well with more holistic perceptions of writing quality. This finding first underscores the fact that linguistic complexity does not exhaustively capture L2 writing quality or ability but is merely one of its dimensions, along with accuracy, fluency, coherence, eloquence, and so forth. Second, this finding raises the problem of the possibility of 'halo effects' in subjective ratings of language production. Halo effects were first introduced in the field of language testing by Yorozuya & Oller 1980 who defined it as "a tendency for judges to assign similar scores across the various scales ... For instance, a judge rating an

interviewee high on, say, the Vocabulary scale might also assign a high rating on Grammar and each of the other scales quite independently of the constructs supposedly underlying the scales. This kind of judgmental bias could be called a halo effect — a kind of spillover across scales causing them to be more strongly correlated with each other" (p.136). The halo effect has also been observed in the evaluation of L2 productions (e.g., Engelhard 1994; Knoch, Read & von Randow 2007; Kozaki 2004; Malvern et al 2004; Tonkyn 2012), with a rating of one performance feature or area influencing that of another so that above, or below, average performance in one domain of writing (as measured objectively) was not perceived as such by the raters, probably under the influence of other features of the writing. For instance, idiomatic, eloquent and accurate production of relatively short and simple sentences may appear more complex than it actually/objectively is (e.g., due to the use of rote-learned multiword expressions). There may also be cases where complexity and accuracy are confused, with high levels of the latter masking low levels of the former, or vice versa. Finally, relatively sophisticated content may also have an unduly positive effect on overall writing ratings, or on complexity-related ratings, regardless of the objective linguistic complexity of the language produced. On the other hand, syntactically complex language may not be identified as such by raters if it involves undue repetition of structures, or is couched in relatively short sentences, or is formulated with 'unsophisticated' lexis. Complex language might also not be recognized if it is felt to be imprecise, obscure or irrelevant to the topic or task at hand. Finally, linguistically complex language may be masked if it is inarticulate or formulated laboriously in relatively non-fluent or non-idiomatic ways. Raters may need training to discern complexity within inaccurate, ineloquent and/or short or long writing productions, and to distinguish complexity from accuracy and sophisticated content (Knoch et al 2007). And if time and other resources permit, simultaneous ratings of complexity, accuracy and other aspects of writing production (e.g. coherence, eloquence) by one rater should be abandoned in favour of separate ratings by different raters, or by a single rater reading the essays repeatedly, focusing on one performance area or feature at the time.

References

Alderson, J.C. (2000). Testing in EAP: Progress? Achievement? Proficiency? In G.M. Blue, J. Milton, & J. Saville (Eds.), *Assessing English for Academic Purposes* (pp. 21-47). Bern: Peter Lang.

Bachman, L. (2005). *Statistical Analysis for Language Assessment.* Oxford: Oxford University Press.

Bulté, B. (2007). Measure for Measure: Why Type/Token ratio based measures are not valid to assess lexical complexity/richness as a dimension of language proficiency. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, Accuracy and Fluency in Second language Use, Learning and Practice* (pp. 27-35). Wetteren: Universa Press.

Bulté, B. (2013). The development of complexity in second language acquisition. A dynamic systems approach. Unpublished doctoral dissertation, University of Brussels (VUB), Brussels, Belgium.

Bulté, B., & Housen, A. (2012). Defining and Operationalising L2 Complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds), *Dimensions of L2 Performance and Proficiency - Investigating Complexity, Accuracy and Fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins.

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*, *3(18)*, 277-298.

Byrnes, H., Maxim H., & Norris, J.M. (2010). Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment [Monograph]. *Modern Language Journal*, *94* (Suppl. s1).

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

Crystal, D. (1997). *The Cambridge Encyclopedia of Language.* Cambridge: Cambridge University Press.

Cumming, A. (1995). Fostering writing expertise in ESL composition instruction: Modeling and evaluation. In D. Belcher, & G. Braine (Eds.), *Academic writing in a second language* (pp. 375-397). Norwood, NJ: Ablex Publishing Co.

Dahl, Ö. (2004). *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24(2)*, 197-222.

Ehret, K., & Szmrecsanyi, B. (in press). An information-theoretic approach to assess linguistic complexity. In R. Baechler, & G. Seiler (Eds.), *Complexity and Isolation*. Berlin: de Gruyter.

Ellis. R. (2008). *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Ellis. R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics* 30(4), 474-509.

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

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement* 31(2), 93–112.

Foster, P., & Tavakoli, P. (2009). Lexical diversity and lexical selection: a comparison of native and non-native speaker performance. *Language Learning* 59(4), 866-896.

Given, L.M. (2008). *The Sage encyclopedia of qualitative research methods*. Los Angeles, CA: Sage Publications.

Givon, T. (2009). The Genesis of Syntactic *Complexity*: Diachrony, Ontogeny, Neuro-Cognition, Evolution. Amsterdam: John Benjamins.

Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique*. Dordrecht: Reidel.

Halliday, M.A.K. (1998). Things and relations: Regrammaticising experience as technical knowledge. In J.R. Martin, & R. Veel (Eds.), *Reading Science: Critical and Functional Perspectives on Discourses of Science* (pp. 185-235). New York: Routledge.

Hartshorn, K.J., Evans, N.W., Merrill, P.F., Sudweeks, R.R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly, 44*, 84-109.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range [Computer software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Housen, A., & Kuiken F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, *30(4)*, 461-473.

Housen, A., Kuiken, V., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (Eds), *Dimensions of L2 Performance and Proficiency - Investigating Complexity, Accuracy and Fluency in SLA* (pp. 1-20). Amsterdam: John Benjamins.

Huddleston, R., & Pullum, G. (2002). *The Cambridge grammar of the English language*. New York: Cambridge University Press.

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning*, *63* (Suppl. 1), 87-106.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language Complexity: Typology, Contact, Change*. Amsterdam, Philadelphia: Benjamins.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12(1)*, 26-43.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing, 21(1)*, 1–27.

Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking: the effect of mode. In P. Robinson (Ed.), *Second language task complexity: researching the cognition hypothesis of language learning and performance* (Task-based language teaching, 2) (pp. 91-104). Amsterdam: John Benjamins.

Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, I. Vedder, & G. Pallotti (Eds.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research* (EUROSLA monographs series, 1) (pp. 81-99). [S.l.]: European Second Language Association.

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

Larsen-Freeman, D. (2012). In A. Mackey and S. Gass (Eds.), *Handbook of Second Language Acquisition* (pp. 73-88). New York: Routledge.

Larsen-Freeman, D., & Cameron, L. (2008). *Complex Systems and Applied Linguistics*. Oxford: Oxford University Press.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, *27*, 123-134.

Leaver, B. L., & Shekhtman, B. (2002). Principles and practices in teaching superior-level language skills: Not just more of the same. In B. L. Leaver, & B. Shekhtman (Eds.), *Developing professional-level language proficiency* (pp. 3-33). New York: Cambridge University Press.

Lennon, P. (1995). Assessing short-term change in advanced oral proficiency: Problems of reli- ability and validity in four case studies. *ITL Review of Applied Linguistics*, 75–109.

Levelt, W.J. (1989). *Speaking: From intention to articulation*. Cambridge: MIT Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

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.

McWhorter, J. (2001). The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2/3),125-167.

McWorther, J. (2011). *Linguistic Simplicity and Complexity: Why Do Languages Undress?* Boston/Berlin: Walter De Gruyter.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds), *Language Complexity: Typology, Contact, Change* (pp. 23-42). Amsterdam: John Benjamins.

Mitchell, M. (2009). *Complexity – A Guided Tour*. Oxford: Oxford University Press.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63(1)*, 59-81.

Norris, J. M., & Ortega, L. (2009). Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. *Applied Linguistics*, *30(4)*, 555–578.

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

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann, & B. Szmrecsanyi (Eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact* (pp. 127-55). Berlin: de Gruyter.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30(4)*, 590–601.

Polio, C. (2001). Research methodology in second language writing: The case of text-based studies. In T. Silva, & P. Matsuda (Eds.), *On second language writing* (pp. 91-116). Mahwah, NJ: Erlbaum.

Politzer, R. L., & McGroarty, M. (1985). An exploratory study of learning behaviors and their relationship to gains in linguistic and communicative competence. *Tesol Quarterly*, *19(1)*, 103-123.

Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. *Modern Language Journal*, *89*(1), 3–18.

Rimmer, W. (2006). Measuring grammatical complexity: the Gordian knot. *Language Testing*, *23(4)*, 497-519.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics 22 (1)*, 27–57.

Sampson, G., Gil, D., & Trudgill, P. (2009). *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Serrano, R., Tragant, E., & Llanes, A. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review*, *68(2)*, 138-163.

Skehan, P. (1998). A Cognitive Approach to Language Learning. Oxford: Oxford University Press.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*, 1-14.

Skehan, P. (2009a). Lexical performance by native and non-native speakers on language learning tasks. In B. Richards, H.M. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary Studies in First and Second Language Acquisition: The Interface between Theory and Application* (pp. 107-124). London: Palgrave Macmillan.

Skehan, P. (2009b). Modelling Second Language Performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, *30*, 510-532.

Skehan, P. and Foster, P. (1997) Task type and processing conditions as influences on foreign language performance. *Language Teaching Research* 1, 3, 185-211.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183-205). Cambridge: Cambridge University Press.

Skehan, P., & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 199-220). Amsterdam: John Benjamins.

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.

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, *18*, 103-118.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307-336.

Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency: Examining instructed learners' short-term gains.' In A. Housen, F. Kuiken, & I. Vedder (Eds.) *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 221-246). Amsterdam: John Benjamins.

Verspoor, M., Lowie, W., & van Dijk, M. (2008). Variability in Second Language Development From a Dynamic Systems Perspective. *Modern Language Journal*, *92(2)*, 214-231.

Verspoor, M., & Sauter, K. (2000). *English Sentence Analysis: an introductory course*. Amsterdam: John Benjamins.

Verspoor, M., Schmidt, M., & Xu, J. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21(3),* 239-263.

Vyatkina, N. (2012). The Development of Second Language Writing Complexity in Groups and Individuals: A Longitudinal Learner Corpus Study. *Modern Language Journal*, *96(4)*, 576-598.

Westaway, G., Alderson, J.C., & Clapham, C.M. (1990). Directions in Testing for Specific Purposes. In J.H.A.L. de Jong, & D.K. Stevenson (Eds.), *Individualizing the Assessment of Language Abilities* (pp. 239-256). Clevedon: Multilingual Matters.

Wette, R. (2010). Evaluating student learning in a university-level EAP unit on writing using sources. *Journal of Second Language Writing*, *19(3)*, 158-177.

White, E. (1994). *Teaching and assessing writing.* (2nd ed.). San Francisco: Jossey-Bass Publishers.

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.

Yorozuya, R., & Oller, J.W. Jr. (1980). Oral proficiency scales: Construct validity and the halo effect. *Language Learning, 30(1)*, 135-153.

Received: January 16, 2015

Accepted: June 11, 2015

Published: September 28, 2015