Compressing learner language: an information-theoretic measure of complexity in SLA production data

Short title: Compressing learner language

Katharina Ehret (University of Freiburg)

Benedikt Szmrecsanyi (KU Leuven)

Corresponding author:
Benedikt Szmrecsanyi
Department of Linguistics
Blijde-Inkomststraat 21
PO Box 03308
B-3000 Leuven
Belgium
+32 16 3 24823
benszm@kuleuven.be

Abstract

We present a proof-of-concept study that sketches the use of compression algorithms to assess Kolmogorov complexity, which is a text-based, quantitative, holistic, and global measure of structural surface redundancy. Kolmogorov complexity has been used to explore cross-linguistic complexity variation in linguistic typology research, but we are the first to apply it to naturalistic SLA data. We specifically investigate the relationship between the complexity of L2 English essays and the amount of instruction the essay writers have received. Analysis shows that increased L2 instructional exposure predicts increased overall complexity and increased morphological complexity, but decreased syntactic complexity (defined here as less rigid word order). While the relationship between L2 instructional exposure and complexity is robust across a number of L1 backgrounds, L1 background does predict overall complexity levels.

Keywords: Kolmogorov, complexity, information theory, compression, writing, learner corpus, proficiency

Acknowledgments

The first-named author is grateful to the *Cusanuswerk* (Bonn) for a generous PhD scholarship. Both authors acknowledge financial support from the Freiburg Institute for Advanced Studies. Thanks for helpful feedback go to the audience of the 2014 colloquium on "Cross-linguistic aspects of complexity in SLA" (VUB Brussels), as well as to Christy Thanh Linh Ha, Benedikt Heller, Lourdes Ortega, Amir Zeldes, two anonymous reviewers, and the editors. The usual disclaimers apply.

1. Introduction

The past decade has seen a flurry of research and publication activity throwing doubt on the long-held belief in theoretical linguistics that all languages are essentially equally complex (see e.g. Kusters 2003; Sampson 2009 for discussion). One of the primers was a lead article in the journal *Linguistic Typology* in which John McWhorter suggested that creole languages tend to have simpler grammars than older languages "by virtue of the fact that they were born as pidgins, and thus stripped of almost all features unnecessary to communication" (McWhorter 2001:125). Following this claim, much recent theoretical work on language complexity takes a functional-typological (e.g. Miestamo, Sinnemäki & Karlsson 2008), contact linguistic (e.g. Kortmann & Szmrecsanyi 2012), and/or sociolinguistic (e.g. Trudgill 2011) perspective. The new consensus is that human languages and dialects of the same language can and often do differ in their complexity.

In contrast to theoretical linguists, applied linguists have been in the business of measuring complexity for a longer while. According to Ortega (2012), the quest for interlanguage complexity measures is guided by the following objectives: "(a) to gauge proficiency, (b) to describe performance, and (c) to benchmark development'' (128). That said, the SLA community lacks a commonly accepted construct definition of complexity (Bulté & Housen 2012:22): the terms "complexity" and "complex" are often used to describe disparate concepts including, among others, complexity of linguistic structures but also L2 acquisition difficulty. This is why the development of a clear-cut definition of complexity and of common metrics that would improve comparability across different studies remains a desideratum. Bulté & Housen (2012), for instance, propose an elaborate taxonomic framework to define the multifaceted nature of L2 complexity, which ranges from the distinction between absolute and relative complexity at the top of the taxonomic ladder to the various (sub)levels of language (e.g. clausal syntax, derivational morphology) at the bottom.

There is also a lively debate about the triad complexity – accuracy – fluency (CAF), their definition, and their relation to L2 development and proficiency. Recent publications advocate disentangling the CAF triad from the notion of language development and the acquisition of linguistic structures (Norris & Ortega 2009; Pallotti 2009:593,599): complexity measures should be used to independently describe linguistic performance, rather than being defined by the time of their appearance in L2 development (Pallotti 2009:593–594, 599).

Another major point of controversy concerns measures and operationalizations of complexity. Popular complexity metrics in the SLA literature include various measures of unit length (e.g. T-unit length), subordination frequency measures, and measures of the frequency of "complex" forms (see Ortega 2003 for a review). There is, however, a sense in the applied linguistics community that measures of this kind are problematic. Among other things, they suffer from what Ortega (2012:128) calls "concept reductionism", and they yield inconclusive (Bulté & Housen 2012:34) and/or misleading (Biber, Gray & Poonpon 2011) findings. This is in part because it is often not clear what the existing metrics actually measure and whether and to what extent they map onto the underlying theoretical constructs (Norris & Ortega 2009:560; Bulté & Housen 2012:26–28). A case in point is the fact that the majority of L2 research applies only a small number of (mainly syntactic and lexical complexity) metrics which tend to be interpreted as indicator for L2 complexity and development in general (Bulté & Housen 2012:34). On the other hand, many such metrics reflect merely one of many dimensions of the theoretical constructs that they are meant to operationalize (Norris & Ortega 2009:560). Alternatively, some existing metrics arer inadequate in that they are based not on one but on several linguistic subdomains, e.g. by mixing syntax and morphology at the level of operationalization (Bulté & Housen 2012:29). Apart from such interpretational and definitional issues, complexity metrics are often redundantly applied in SLA research: in much extant research, equivalent or very similar metrics are utilized, while some areas of linguistic complexity such as (derivational) morphology are rarely considered, if at all (Pallotti 2015; Bulté & Housen 2012; Norris &

Ortega 2009).

Against this backdrop, we present a measure of language complexity that bridges the gap between theoretical linguistics and applied linguistics: KOLMOGOROV COMPLEXITY (Kolmogorov 1965; Li & Vitanyi 1997). This measure is inspired by work in language typology and information theory and defines the complexity of a text as proportional to the length of the shortest algorithm that can generate that text. Thus, unlike the plethora of measures used in the theoretical literature (e.g. the number of rules a grammar specifies -- see McWhorter 2001), Kolmogorov complexity is a usage-based measure because it gauges the text complexity, in the parlance of Pallotti (2015), of production data (see Szmrecsanyi 2015 for more discussion). At the same time, it is arguably less reductionist than customary measures in the SLA literature, for what takes center stage are not selected aprioristically properties of texts (e.g. subordination, length of units) or single facets of multi-dimensional constructs (e.g. inflectional complexity as representative metric of morphological complexity), but – more holistically – the predictability of upcoming text based on previously seen text. On the technical side, Kolmogorov complexity can be approximated using file compression programs: text samples that can be compressed efficiently count as linguistically simple. As we shall show, this method may be combined with various distortion techniques to calculate measures of morphological and syntactic complexity specifically.

We are not the first to use the compression technique as a measure of linguistic complexity (see Juola 1998; Juola 2008; Sadeniemi et al. 2008; Ehret & Szmrecsanyi 2016). What these previous studies share in common is that they all study more or less artificial parallel corpus databases, where propositional content is constant: Juola (2008) investigates complexity in Bible translations; Ehret & Szmrecsanyi (2016) study Bible translations and translations of Carroll's *Alice's Adventures in Wonderland*; Sadeniemi et al. (2008) measure the complexity of translations of the European Constitution. By contrast to this previous work, the methodological innovation that we introduce in this paper is the application of Kolmogorov complexity to non-parallel naturalistic texts that constitute the primary data source in SLA and learner language research (but we hedge right at the outset that the texts analyzed in SLA research are often rather short, which is a bit of a problem for the technique; see Section 5 for more discussion). We will specifically measure Kolmogorov complexity in the *International Corpus of Learner English* (ICLE) (Granger, Dagneaux & Meunier 2002; Granger 2003), which samples learner essays from students of 11 different mother tongue backgrounds and with differential exposure to English language instruction. Our specific research question is a pseudolongitudinal one: we are interested in whether increased exposure to English language and writing instruction (which we view as a rough but by no means perfect proxy for more proficiency at later stages of L2 development) translates into increased Kolmogorov complexity scores of students' essays., as one would reasonably expect (see Bulté & Housen 2014:45 and the literature cited there).[1]

This paper is structured as follows. In Section 2, some background on information theory is provided. Section 3 explains the methodology. In Section 4, we present the empirical findings. Section 5 offers a discussion and conclusion.

2. Information theory and Kolmogorov complexity

Information theory is concerned with the concept and measurement of 'information' (van der Lubbe

[1] In this connection, we do acknowledge – as pointed out by a reviewer – that in certain styles and registers, simple structures may be more appropriate and native-like than (over-)complicated ones. In terms of the student essays subject to study in the present paper we feel justified, however, in following the standard SLA view that "the ability to produce more linguistically complex oral or written texts reflects increasingly more developed and mature capacities to use the second language" (Ortega 2012:127).

1997:1). In his landmark (1948) paper, Shannon derived a quantitative measure of information, *Shannon entropy,* and defined 'information' as referring to the unpredictability which is involved in the selection of a message from a possible set of messages: the information content of a message is directly related to its unpredictability or unexpectedness, i.e. a message is informative to the extent that it is not predictable or known in advance.

Kolmogorov complexity is a related measure, but in contrast to Shannon entropy, Kolmogorov complexity measures the information content of a string of symbols or text sample, not of a set of possible messages (Li & Vitanyi 1997:521–525). More precisely, the information content of a string is measured as the length of the shortest possible description of this string (Juola 2008:92; Sadeniemi et al. 2008:191; Li et al. 2004:3252). To exemplify, let us assume that the complexity of the two strings in (1) needs to be measured. Both strings count 20 characters, yet the length of the shortest possible description of string (1a) is 10 times *ab*, which is compact, whereas the length of the shortest possible description of the random character sequence in (1b) is the string itself, counting a full 20 different characters. In terms of Kolmogorov complexity, then, string (1b) is more complex than string (1a).

(1) a. abababababababababab

b. ag!73kjrq4#tmn0e1y5z

For mathematically non-trivial reasons, Kolmogorov complexity is actually incomputable (Kolmogorov 1965; Li & Vitanyi 1997), but its upper bounds can be approximated through adaptive entropy estimation methods. As it happens, modern file compression programs such as gzip use a variant of adaptive entropy estimation that approximates Kolmogorov complexity. In fact, the Kolmogorov complexity of a given text file is the file size of the compressed version of this file (Li et al. 2004; Ziv & Lempel 1977). Text compression algorithms such as gzip compress text strings by describing new strings on the basis of previously encountered (sub)strings, and so measure the amount of information and redundancy in a given text string (Juola 2008:93). In this paper, we will use the gzip program to assess linguistic complexity by measuring the information content, and hence predictability, of text samples (more specifically, of learner essays). Essays in our dataset which can be compressed comparatively well, i.e. efficiently, count as linguistically comparatively simple; essays which are less compressible count as linguistically more complex.[2] For more information – such as e.g. the type of features that contribute to compression efficiency – we refer the reader to the detailed discussion in Ehret (2016).

How can Kolmogorov complexity be interpreted linguistically? Compression algorithms are, of course, agnostic about form-function pairings and other deep linguistic relationships of this kind, but they do capture the recurrence of linguistic structures and (ir)regularities. Therefore, Kolmogorov complexity is a quantitative measure of complexity which restricts attention to *formal aspects* (as opposed to function or meaning) of an entire text (as opposed to individual features). Kolmogorov complexity is thus a text-based, holistic, and global measure of structural surface redundancy. In this paper, it measures complexity on the lexical, morphological, and syntactic level. The basic idea behind Kolmogorov complexity is that language is more complex the less predictable it is.

3. Methods and data

3.1. General method

We use gzip (GNU zip, Version 1.2.4., http://www.gzip.org/) to approximate the relative

[2] The compressibility of text samples is always compared to data points within a given dataset. Single compression ratios are therefore always to be seen relative to the other ratios in a dataset.

informativeness of our text samples by measuring overall Kolmogorov complexity as well as, after distortion, Kolmogorov complexity at the morphological and syntactic level. In the present paper we will refer to this method as the *compression technique* (see Ehret 2016 for extended discussion).

Overall Kolmogorov complexity is measured fairly straightforwardly by taking two measurements for each text sample: the file size (in bytes) before compression and the file size after compression. We subsequently subject the file size pairings to regression analysis in order to discard the correlation between the two measurements.[3] The resulting *adjusted overall complexity score*s (regression residuals, in bytes) are used to rank texts in terms of complexity: higher scores in the positive range are indicative of overall higher linguistic complexity; lower scores analogously indicate lower complexity.

Inspired by Juola (Juola 1998; 2008), morphological and syntactic complexity is indirectly measured by distorting the text samples prior to compression. Syntactic distortion is performed by deletion of 10% (this is the customary percentage used in the literature; see Juola 1998; Juola 2008; Kettunen et al. 2006; Sadeniemi et al. 2008) of all *word tokens* in each text file before applying the compression technique. This procedure leads to the disruption of word order regularities. Syntactically complex texts, i.e. texts with a comparatively fixed word order (such as e.g. Standard English texts -- see Dryer 2013), are badly affected and their compressibility is compromised. Languages with relatively free word order, on the other hand, are little affected by this procedure, as they lack syntactic interdependencies that could be compromised. Comparatively bad compression ratios after syntactic distortion thus indicate comparatively high syntactic complexity. Consider (2b), which is the distorted version of (2a) from which the second occurrence of *seen* was deleted. In (2a), a compression algorithm could have reduced file size by replacing the second occurrence of the sequence *would have seen* by a reference to the first occurrence. In (2b), this compression is not as elegantly possible because the two sequences are not identical any more.[4]

(2) a. he would have seen us, and we would have seen him

b. he would have seen us, and we would have ___ him

Rigid word order (as opposed to free word order) is thus defined as being syntactically complex. This may seem a bit counterintuitive at first glance because Kolmogorov complexity is related to predictability – should not, therefore, rigid word order be more predictable than free word order? Recall here though that Kolmogorov-based syntactic complexity is measured indirectly, as we assess the effect of distortion on predictability in a text. If a text is comparatively less predictable after distortion, the text must be considered syntactically complex. Therefore, rigid word order counts as Kolmogorov-complex from a technical point of view. In a theoretically responsible perspective, we acknowledge that this view of syntactic complexity does not necessarily overlap with customary views in SLA. It is, however, compatible with conceptions in the theoretical literature (e.g. McWhorter 2001), where grammars that constrain users' grammatical choices, including word order choices, typically count as more complex than grammars that do not constrain choices. We also hasten to add that the method could, of course, be adapted to measure syntactic complexity more directly: Martens (2011), for example, utilizes a related measure to assess regularity in treebanks, based on syntactic annotation. Notice though that applications such as Martens' are not really text-based but rather annotation-based – but text-basedness is the motto in the present paper.

To measure Kolmogorov complexity at the morphological level, we delete at random 10% of all

[3] The file sizes before and after compression are correlated such that the bigger an uncompressed text file is to start with, the bigger will be its corresponding compressed file, all other things being equal.

[4] The exact sequences that are compressed depend on the context they occur in. As a rule, gzip compresses the longest possible character sequence that can be matched, i.e. *would have seen* in our example. For a detailed discussion of how gzip works see Ehret (2016).

characters in each text file. This creates new word forms, which negatively affects the compressibility of morphologically simple texts which, on the whole, have fewer word forms than morphologically complex texts. By contrast, morphologically complex texts exhibit overall a relatively large amount of word forms in any case, and are thus not affected as much by this kind of distortion.[5] Therefore, after distortion, comparatively bad compression ratios indicate low morphological complexity. Consider (3b), which is not as compressible as (3a) because the last word token, *r_se*, cannot be replaced by a reference to previous occurrences of *rose*.

(3) a. Rose is a rose is a rose is a rose

b. Rose is a rose is a rose is a r_se

On a more technical note, to calculate local (morphological/syntactic) complexity, we take the size in bytes of the original undistorted file and the size in bytes of the syntactically / morphologically distorted file. On the basis of these values we calculate two complexity scores: the *morphological complexity score*, defined as − *m* / *c*, where *m* is the compressed file size after morphological distortion and *c* is the compressed file size before distortion[6]; and the *syntactic complexity score*, defined as *s* / *c*, where *s* is the compressed file size after syntactic distortion and *c* the file size before distortion. Interpretationally, the measures of morphological complexity that we will report in this paper indicate the extent to which words exhibit comparatively many word forms and/or derivational forms (as opposed to being invariant). Syntactic complexity, in Kolmogorov terms, is about the extent to which word order is rigid – which we here define as “complex” – as opposed to free (which we define as “simple”).

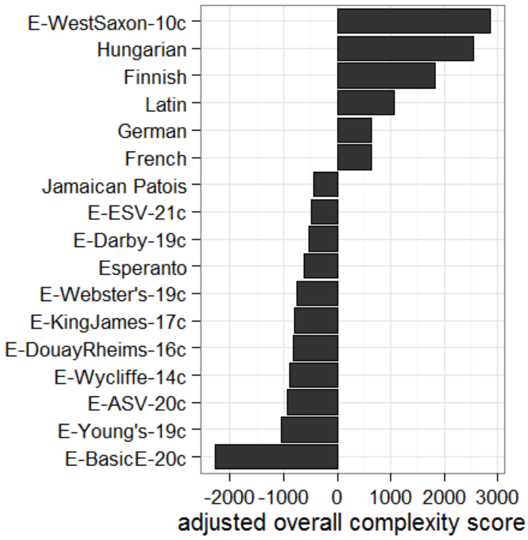

Figure 1: Overall Kolmogorov complexity (*x*-axis) of Bible texts (*y*-axis). Negative complexity scores indicate below-average complexity; positive scores indicate above-average complexity (adapted from Ehret & Szmrecsanyi 2016:Figure 1).

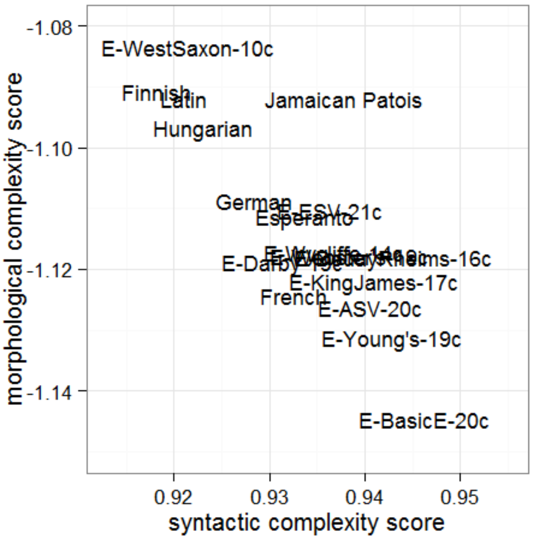

Figure 2: Morphological complexity (*y*-axis) by syntactic complexity (*x*-axis) (adapted from Ehret & Szmrecsanyi 2016:Figure 2).

[5] For an analysis of distorted texts and word forms see Ehret (2016).

[6] We invert the sign of the morphological complexity score so that higher scores indicate higher complexity.

To instill confidence in the construct validity of Kolmogorov complexity, Figure 1 and 2 summarize Ehret & Szmrecsanyi's (2016) measurements of Kolmogorov complexity in a parallel text database containing translations of the Gospel of Mark into a number of languages, including historical varieties of English. According to Figure 1, the three most complex translations are the West Saxon (Old English), Hungarian, and Finnish versions; Jamaican Patois and Esperanto are rather non-complex, and so are most modern translations into English. The overall least complex data point in Figure 1 is the Basic English translation of the Bible. Basic English is a simplified variety of English designed by Charles Kay Ogden as, among other things, an aid to facilitate teaching of English as a foreign language (Ogden 1934). A distortion-based analysis of Kolmogorov complexity at the syntactic and morphological levels (Figure 2) shows that the morphologically most complex languages in the sample are West Saxon, Finnish, and Latin (in that order), while the syntactically most complex data point (i.e. the text in which word order is most rigid) is the Bible in Basic English. It is, we believe, fair to say that the compression technique thus assesses complexity in a way that is consistent with the intuitions that most linguists have about the languages at hand. The task before us, then, is to apply the compression technique to learner essays as detailed in the next session.

3.2. Data source

We use the *International Corpus of Learner English* (ICLE), Version 1.1 (Granger, Dagneaux & Meunier 2002; Granger 2003). ICLE is a corpus of written learner English (or EFL), and contains both argumentative and literary essays composed by intermediate to advanced learners from different L1 backgrounds such as Bulgarian, Czech, Dutch, Finnish, French, German, Polish, Russian, Spanish, and Swedish. The corpus totals about 2.5 million words of running text. All components of the corpus were designed according to the same guidelines. Metadata about a number of learner variables (including time spent in an English speaking country, time spent studying English at school, time spent studying English at university) as well as task variables (topic of the essay, length, argumentative versus literary, timed versus untimed, exam conditions versus use of reference tools) are available. A big advantage of ICLE is its comparatively large size (see Paquot & Granger 2012:132) – Kolmogorov complexity measurements do not work well with small texts (see Section 5 for discussion).

3.3. Defining the dataset

Because the technique requires comparatively large text samples, we will aggregate over groups of learners: individual learner essays in ICLE are, alas, too short for taking reliable Kolmogorov measurements. Our aggregation endeavor is guided by our interest in the relationship between the complexity of the learner essays and the amount of previous instruction in English that the essay writers have received. We thus utilize a pseudolongitudinal research design (Gass & Selinker 2001:32–33), in that we investigate the growth, maintenance, and possibly decline of complexity in a data source that can be taken to represent different proficiency levels (see also Bestgen & Granger 2014:30). In this spirit, we categorize ICLE essays into six different groups according to the amount of L2 instructional exposure, i.e. the number of years of instruction in English at school and/or university. We also restrict attention to argumentative essays, which constitute the largest part of the data, because of the fairly homogenous content of these texts.

The essays were categorized as follows. First, the number of texts for every possible combination of years spent studying English at school and university was surveyed. In some cases, data is very sparse or the available data comes from learners who have not attended university-level language classes; for example, there are only four essays from learners who have studied English for one year at school but have not studied English at university at all (we refer the reader to the corpus manuals for more information on the corpus design and sampling principles). In order to obtain a representative sample for each L2 instructional exposure group, such borderline cases were excluded. Thus, the range of years spent studying English at school was restricted to 4–9 years and the range of years at university to 1–5 years (see Ehret 2016 for a more detailed description).

Table 1: L2 instructional exposure groups by years of instruction in English at school and university. The total number of years, number of texts, sentences, and words are provided for each group.

group	school (yrs.)	university (yrs.)	total yrs.	# texts	# sentences	# words
I	4-6	1-2	5-8	340	12,531	230,054
II	4-6	3	7-9	345	13,644	238,590
III	4-6	4-5	8-11	464	16,792	303,233
IV	7-9	1-2	8-11	533	17,285	335,091
V	7-9	3	10-12	262	9,328	171,762
VI	7-9	4-5	11-14	253	8,765	169,237

On the basis of the number of years spent studying English at school and university, six groups of essays were distinguished according to L2 instructional exposure to English of their writers (see Table 1). The aim was to minimize group overlap. The most advanced groups – the groups sampling essays from learners with the highest amount of L2 instructional exposure in English – are groups V and VI, while groups I and II are the groups with the least amount of L2 instructional exposure. Groups III and IV both represent intermediate levels of L2 instructional exposure; the learners in these groups have received the same amount of instruction in English, but not in the same type of setting (school vs. university).[7] We hedge that the groups are probably not entirely homogeneous and that there might be individual differences between learners of the same level, as we cannot easily control for e.g. differences in the learning context between/within countries.

3.4. Data processing

All essays in particular groups were combined into one single text file (hence each group has its own single text file), which was then fed into gzip. We used scripts to apply the compression technique with N = 1, 000 iterations and subject to random sampling: in each iteration, 10% of the sentences per text were randomly sampled. We sampled the same percentage of sentences rather than, say, the same percentage of words because in this manner syntactic interdependencies remain intact.[8] This sampling and iteration procedure serves three purposes: (1) taking multiple measurements increases reliability, (2) random sampling ensures the comparability of the different texts because it keeps sample size constant, (3) taking random samples of a constant size from differently sized texts yields more representative results than fixed-size samples which cover only a certain part of a text (see Ehret 2016 for a detailed discussion).

In each iteration the overall Kolmogorov complexity of the data is established, and the adjusted complexity scores are calculated per iteration. Subsequently, we calculate *average adjusted overall complexity scores*, which take the arithmetic mean across all iterations. See Ehret (2016) for details such as the mean uncompressed and compressed file sizes on whose basis the average adjusted

[7] Note that the group labels are not to be understood as a proficiency level index, and that years spent learning at school are not considered more important than years spent learning at university.

[8] Although this results in unequal sample sizes in terms of the exact number of words considered, experiments (see Ehret 2016) have shown that despite this caveat the variable "percentage of sentences" delivers better results than "percentage of words" or "percentage of characters".

overall complexity scores reported in this paper were calculated. To assess morphological and syntactic complexity, we apply the distortion and compression script with N = 1, 000 iterations, which yields the morphological and syntactic complexity score for each group for each iteration of the script. We then take the arithmetic means across iterations to calculate the *average morphological* and *syntactic complexity score*, respectively.

All statistics reported in this paper were conducted in R, version 3.2. R: A language and environment for statistical computing. Developing Core Team 2008. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

4. Results

Section 4.1. explores complexity differentials according to the compression technique as a function of previous exposure to English language instruction in ICLE as a monolithic whole. Furthermore, we explore how Kolmogorov-based complexity measurements relate to more traditional SLA metrics. Section 4.2. investigates the role that L1 background plays.

4.1. Complexity as a function of L2 instructional exposure: the big picture

We begin by discussing the complexity of the essays produced by learners with different levels of L2 instructional exposure on the overall, syntactic, and morphological tiers. We take the liberty in this subsection to ignore L1 background of the essay writers for the time being, not because we consider it irrelevant but because this abstraction is accompanied by a considerable gain in clarity and fewer data sparsity issues. This clarity is intended to set the stage for a more accountable analysis of the role that L2 background plays in section 4.2.

In terms of overall complexity, the results exhibit the theoretically expected relationship between L2 instructional exposure and complexity: essays by more advanced learners tend to be more Kolmogorov-complex than essays by less advanced learners, as Figure 3 demonstrates. An informal Pearson's correlation test indicates that overall complexity correlates significantly with the amount of L2 instructional exposure (r = 0.85, p = 0.034).[9] A closer look reveals that essays in groups V and VI are overall the most complex texts, coming from learners who studied English for ten to fourteen years in total and who therefore have the highest level of L2 instructional exposure in the dataset. Essays in group I, on the other hand, are overall the least complex texts – written by learners who have studied English for only about five to eight years in total. The texts in groups II, III and IV are below-average complex; the ranking within this subgroup is opposite to expectations. But for the more advanced learners (groups IV through VI), the pseudolongitudinal ranking does seem to point to a progression from less complex production to more complex production as L2 instructional exposure increases. We note also that there seems to be quite a quantum jump, in terms of Kolmogorov complexity, between Group IV (8-11 years of total instruction) and Group V (10-12 years of instruction).

[9] We caution that this correlation test must be taken with a grain of salt, as it strictly speaking treats the groups as an index of amount of exposure and not as a factor that represents different levels of exposure. Because of the overlap between groups III and IV, it implicitly assumes that school years are "more important" than university years.

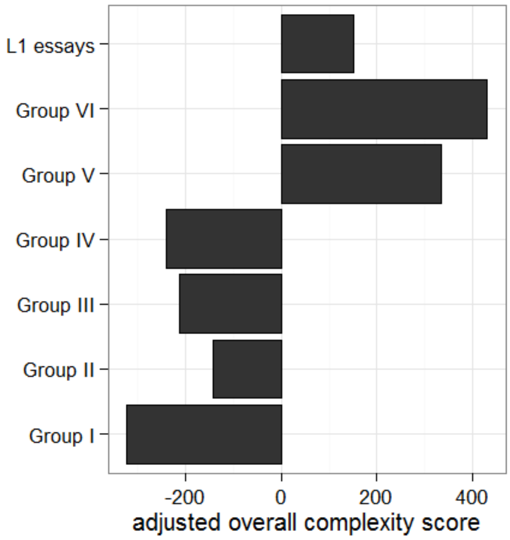

Figure 3: Overall Kolmogorov complexity (*x*-axis) by L2 instructional exposure group in ICLE (y-axis; Group VI: most L2 instructional exposure to English; Group I: least L2 instructional exposure) and in reference L1 essays (LOCNESS). Negative complexity scores indicate below-average complexity; positive scores indicate above-average complexity.

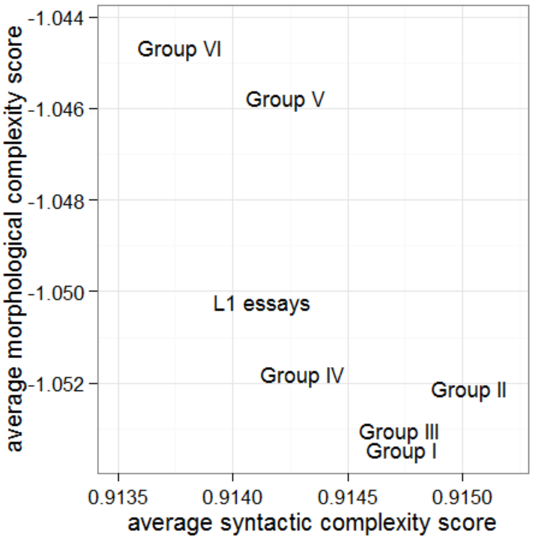

Figure 4: Morphological complexity (*y*-axis) by syntactic complexity (*x*-axis) in ICLE and in reference L1 essays (LOCNESS). Data points represent L2 instructional exposure groups according to level of previous L2 instructional exposure (Group VI: most L2 instructional exposure to English; Group I: least L2 instructional exposure).

Figure 4 plots morphological complexity against syntactic complexity, once again distinguishing between the six exposure groups. In general, we note a trade-off between morphological and syntactic complexity which is statistically reflected in a negative correlation coefficient (r = - 0.82, p = 0.023). The texts produced by learners with the most L2 instructional exposure, groups V and VI, are rated as the most complex morphologically, which indicates that essays written by these learners exhibit comparatively many word forms and/or derivational forms. In terms of syntactic complexity, though, group V and VI essays are *least* complex. Recall along these lines that we defined the "complex" pole of the syntactic complexity continuum as capturing word order rigidity. Hence group V and VI essays exhibit comparatively variable word orders. Now, compared to group V and VI essays, all other groups are morphologically considerably less complex but syntactically more complex. Essays in group II are the syntactically most complex essays followed, in decreasing order of syntactic complexity, by groups III, I and IV. The morphologically simplest essays we find in group I, where L2 instructional exposure is lowest. Observe here that more L2 instructional exposure predicts more morphological complexity: the two-tailed Pearson's correlation coefficient for morphological complexity and L2 instructional exposure is r = 0.89 (p = 0.018).

In terms of Kolmogorov complexity, how does the performance of the ICLE learners compare to that of native speakers? For benchmarking purposes, we also took complexity measurements in essays written by native speakers of English, drawing on the American subsection (149,574 words of argumentative essays written by American university students) of the Louvain Corpus of Native English Essays (LOCNESS, see https://www.uclouvain.be/en-cecl-locness.html). LOCNESS is designed as ICLE's control native corpus, which perfectly serves our purposes. The LOCNESS benchmarks are plotted into Figures 3 and 4. As can be seen, in terms of overall complexity (Figure 3) the L1 essays tend toward the complex end of the continuum, as expected: they receive a positive overall adjusted complexity score and are ranked as being more complex than most ICLE groups except for groups V and VI. As to morphological and syntactic complexity, the L1 essays are likewise located between the ICLE I-IV cluster and the ICLE V-VI cluster. In summary, then, native essays tend to be more complex overall and more complex morphologically, but less complex syntactically than most learner essays. It is only very advanced learners, with upwards of 10 years of formal instruction, who produce essays that surpass native essays – some of whom are written by college students in their first or second year of study – in terms of overall and morphological Kolmogorov complexity.

That syntactic complexity should *de*crease with increasing L2 instructional exposure (data points depicted in Figure 4: $r = -0.83$, $p = 0.042$), and that native essays are less complex syntactically than many learner essays may at first glance seem counterintuitive. Keep once again in mind, however, that word order rigidity is defined as complex in the present study's approach, and word order non-rigidity as simple. So we may speculate that more proficient learners use, for one thing, non-SVO word order patterns such as e.g. inversion of the type *rarely have we been more astonished* versus *we have rarely been more astonished* more readily than less proficient learners (in a similar way, Klein & Perdue 1997 argue that less advanced learners heavily rely on word order to convey grammatical information). Secondly, our findings can be seen as supporting claims by Biber et al. (2011) who show that the measure of complexity commonly used in writing development studies, namely the degree of clausal embedding, does not capture well the complexity of advanced writing proficiency (Biber, Gray & Poonpon 2011:10–12; see also Biber & Gray 2010). On the contrary, Biber et al. (2011) find that clausal embedding is a feature of conversational language which is acquired in early stages of language development. Later stages of proficiency are instead characterized by a higher degree of phrasal complexity and greater range of lexico-grammatical combinations such as finite complement clauses (e.g. *I think that [. . .]*) (Biber, Gray & Poonpon 2011:29–32). How is this related to morphological and syntactic complexity as measured with the compression technique? Clausal embedding is concerned with the degree of subordination and thus with syntactic complexity. But it could be argued that an increasing use of different lexico-grammatical patterns increases morphological complexity. For instance, according to Biber et al., the majority of *that*-clauses in spoken language occur with only four different verbs (Biber, Gray & Poonpon 2011:31). In the context of Kolmogorov complexity, this means that a text with few verbal patterns is easily compressible and hence morphologically simple. A text with many different verbal patterns, on the other hand, will be more difficult to compress and thus morphologically more complex. Considerations like these may explain the finding that more complexity in writing is not necessarily accompanied by more *syntactic* complexity in particular.

Table 2: Correlations between complexity measures (Pearson correlation coefficients). Significance codes: *significant at $p < 0.05$ under Bonferroni correction. $\text{MCI}_{\text{verbs}}$/ $\text{MCI}_{\text{nouns}}$: morphological complexity indices (Brezina & Pallotti 2015; Pallotti 2015); LDTTRc: lexical diversity, type-token ratio (content words); SYNLE: left embeddedness, words before main verb (mean); SYNNP: number of modifiers per noun phrase (mean); DRNP: noun phrase density, incidence; DRPVAL: agentless passive voice density, incidence.

	overall Kolmogorov complexity	**morphological Kolmogorov complexity**	**syntactic Kolmogorov complexity**
$\text{MCI}_{\text{verbs}}$	0.96 *	0.92	-0.61
$\text{MCI}_{\text{nouns}}$	0.55	0.43	0.14
LDTTRc	0.53	0.58	-0.40
SYNLE	0.24	0.23	0.13

SYNNP	0.83	0.88	-0.85
DRNP	0.78	0.69	-0.28
DRPVAL	-0.07	-0.14	0.04

In an applied linguistics perspective, Kolmogorov complexity is an exotic bird that differs from commonly utilized SLA metrics. To test its validity, we now move on to exploring the extent to which the measurements we report above relate to more traditional SLA complexity metrics. To address this question, we took additional measurements for each of the six ICLE data points (Groups I through VI) displayed in Figures 3 and 4.[10] Subsequently, we calculated Pearson correlation coefficients, as shown in Table 2. Due to the comparatively small number of observations, there is only one coefficient that is statistical significant under Bonferroni correction. Nonetheless, inspecting the top correlations is instructive. Overall Kolmogorov complexity correlates best with morphological complexity in the verbal domain ($\mathrm{MCI}_{\mathrm{verbs}}$) à la Brezina & Palotti (2015), followed by the number of modifiers per noun phrase (SYNNP), and noun phrase density (DRNP). Morphological Kolmogorov complexity likewise correlates robustly with $\mathrm{MCI}_{\mathrm{verbs}}$ à la Brezina & Palotti (2015), followed again by SYNNP and DRNP. We thus see that there is a good deal of overlap between overall and morphological Kolmogorov complexity, an issue that is discussed in Ehret (2016). Syntactic Kolmogorov complexity is inversely correlated with SYNNP, which is another way of saying that a decreasing number of modifiers per noun phrase predicts more rigid word order (which is what syntactic Kolmogorov complexity measures). Syntactic Kolmogorov complexity is also inversely correlated with $\mathrm{MCI}_{\mathrm{verbs}}$.

4.2 Does L1 background matter?

The analysis in the preceding section has sketched the big picture, ignoring the L1 backgrounds of the essay writers. But it is clear that this procedure is simplistic, and may raise concerns that the relationship between Kolmogorov complexity uncovered in the previous section is some sort of spurious aggregation effect. Therefore, the aim of this section is to complement the analysis offered in the previous section by establishing whether (1) the complexity of essays depends on L1 background, and (2) whether the correlation between previous L2 instructional exposure and complexity that we see in ICLE as a whole (see previous section) survives this section's distinction between a number of different L1 backgrounds sampled in ICLE.

We address these questions by taking a closer look at complexity variance in essays by writers with four different L1 backgrounds: German, French, Italian and Spanish. In terms of data quantity, these are the best-documented L1 backgrounds in ICLE, and so mitigate to some extent data sparsity concerns. But even so, we hedge that these ICLE subcomponents are relatively small (see Table 4, Appendix for details). Some of the L2 instructional exposure groups for particular L1 backgrounds are covered by merely two texts, and for the L1 French IV group we do not even have data at all. Coverage of the French, Italian, and Spanish subsets is particularly unbalanced in regard to the distribution of data across the six exposure groups. So the findings will have to be taken with a grain of salt.

Table 3: Kolmogorov complexity in the 4-part German/French/Italian/Spanish sub-dataset, by L1 background and level of previous L2 instructional exposure (Group VI: most L2 instructional exposure

[10] $\mathrm{MCI}_{\mathrm{verbs}}$ and $\mathrm{MCI}_{\mathrm{nouns}}$ were calculated using the online tool at http://corpora.lancs.ac.uk/vocab/analyse_morph.php. The other measures (LDTTRc, SYNLE, SYNNP, DRNP, DRPVAL) were calculated using the online tool at http://cohmetrix.com. As these tools have text limits, we analysed random samples of sentences drawn from the big text files that were used to calculate Kolmogorov complexity. Also because of the text limits, the measurements as well as the resulting correlation coefficients must not be over-interpreted.

to English; Group I: least L2 instructional exposure). Texts are ranked by decreasing complexity.

a. adjusted overall complexity score		b. average morphological complexity score		c. average syntactic complexity score	
German VI	5.9940	German V	-0.9412	French I	0.9239
German V	4.3750	German VI	-0.9418	Italian IV	0.9223
Italian III	4.3328	German II	-0.9426	French VI	0.9216
German III	4.1014	French VI	-0.9426	Spanish II	0.9208
German IV	3.4731	French II	-0.9430	French IV	0.9201
Italian VI	3.3337	German IV	-0.9432	French III	0.9199
Italian V	2.6388	German III	-0.9434	Spanish I	0.9198
German II	2.1794	German I	-0.9443	Spanish VI	0.9195
Spanish III	2.1003	French III	-0.9448	French II	0.9190
German I	1.7605	Spanish III	-0.9449	Spanish III	0.9189
French III	0.3569	French IV	-0.9450	German I	0.9188
French II	-0.0394	Italian VI	-0.9459	Spanish V	0.9188
Spanish IV	-0.0683	French I	-0.9459	German V	0.9184
Spanish VI	-0.2475	Spanish VI	-0.9459	Spanish IV	0.9181
Italian II	-1.0418	Italian III	-0.9460	Italian V	0.9181
Italian IV	-2.0201	Italian V	-0.9461	German II	0.9181
French IV	-3.1943	Italian IV	-0.9462	Italian I	0.9181
Spanish I	-3.4242	Spanish IV	-0.9471	German IV	0.9179
French VI	-3.4417	Spanish I	-0.9485	German III	0.9178
Spanish V	-3.6846	Italian II	-0.9519	Italian VI	0.9175
Spanish II	-4.4872	Spanish V	-0.9523	Italian II	0.9170
French I	-5.5880	Spanish II	-0.9557	Italian III	0.9169
Italian I	-7.4086	Italian I	-0.9628	German VI	0.9162

We applied the compression technique to each L1 background x L2 exposure subset with *N* = 1, 000[11] iterations and random sampling of 10% of the sentences per text and iteration; as before, we compressed (i) undistorted, (ii) morphologically distorted, and (iii) syntactically distorted versions of the texts. The results are displayed in Table 3. It is clear that L1 background clearly matters: L1 German essays tend to be more Kolmogorov-complex in terms of overall complexity (Table 3a) and morphological complexity (Table 3b) than essays from other L1 backgrounds. French essays, on the other hand, tend to be syntactically more complex than essays from other L1 backgrounds (Table 3c). Indeed simple linear regression modeling on the basis of the data displayed in Table 3 shows that L1 background is overall the more potent predictor compared to L2 instructional exposure. In a linear regression model predicting adjusted overall complexity scores (Table 3a), the adjusted R^2 value for L1

[11] For computational reasons the number of iterations is limited to 1,000.

background is 33.5% while the adjusted R^2 value for L2 instructional exposure is 14.3%; in a linear regression model predicting average morphological complexity scores (Table 3b), the adjusted R^2 value for L1 background is 29.9% while the adjusted R^2 value for L2 instructional exposure is -0.92%; and in a linear regression model predicting average syntactic complexity scores (Table 3b), the adjusted R^2 value for L1 background is 30.7% while the adjusted R^2 value for L2 instructional exposure is -11.8%. But that being said, within each L1 background we do tend to see the expected relationship between L2 instructional exposure and complexity, at least in terms of overall complexity (Table 3a); the morphological and syntactic complexity measurements (Tables 3b-c) are quite noisy.

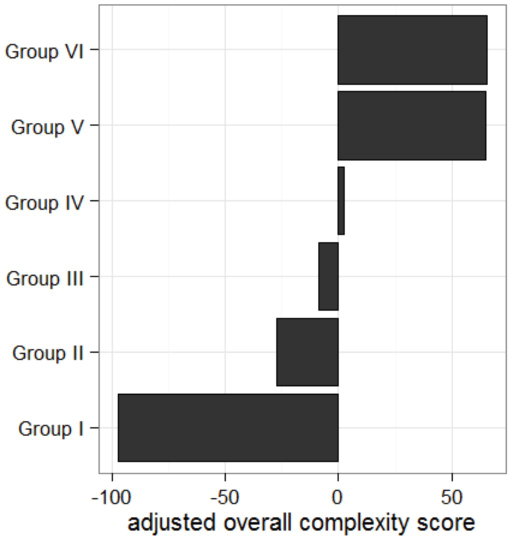

Figure 5: Overall Kolmogorov complexity (*x*-axis) by L2 instructional exposure group in the L1-German sub-component of ICLE (y-axis; Group VI: most L2 instructional exposure to English; Group I: least L2 instructional exposure). Negative complexity scores indicate below-average complexity; positive scores indicate above-average complexity.

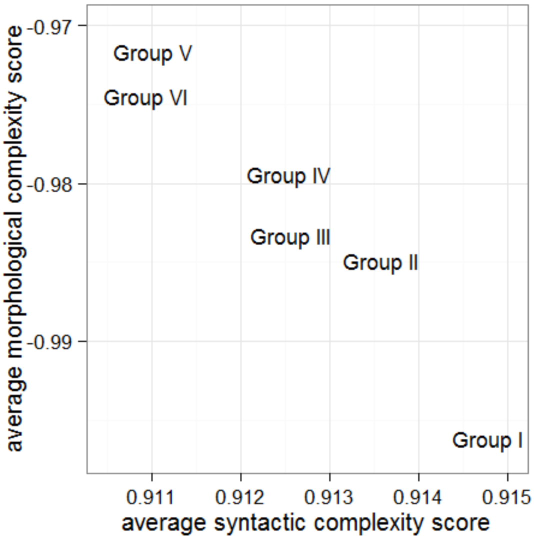

Figure 6: Morphological complexity (*y*-axis) by syntactic complexity (*x*-axis) in the L1-German component of ICLE. Data points represent L2 instructional exposure groups according to level of previous L2 instructional exposure (Group VI: most L2 instructional exposure to English; Group I: least L2 instructional exposure).

We reiterate here that the 4-part German/French/Italian/Spanish subset suffers from data sparsity issues: for example, the French and Spanish ICLE components are particularly unbalanced. Hence caution should be exercised when interpreting this particular corner of the ranking. So in an attempt to obtain relatively reliable measurements, we now restrict attention to one particular L1 background – L1 German learners of English – which happens to be the best documented (the amount of data per L2 instructional exposure group ranges from approximately 11,000 to 50,000 words)[12] and most balanced subset, and will thus serve as a control group of sorts. Figures 5 and 6 display the results. In terms of overall Kolmogorov complexity, we find a neat and consistent relationship between L2 instructional exposure and the complexity of the essays produced. Not only does overall complexity significantly correlate with L2 instructional exposure ($r = 0.96$, $p = 0.001$) as it does in ICLE as a whole (Figure 3), but in the L1 German subset the six groups (I, II, III, IV, V, VI) are perfectly ranked according to L2 instructional exposure. Recall that in the ranking based on ICLE as a whole some less advanced groups exhibit more complexity than the more advanced groups III and IV. We conclude, then, that this could well be an effect of the mixing of various L1 backgrounds. The morphological and syntactic

[12] Essays per group: I = 25; II = 22; III = 24; IV = 107; V = 80; VI = 85.

complexity patterns in ICLE-German (Figure 6) are also fairly similar to the findings based on the complete ICLE, though once again the pseudolongitudinal developments in the L1 German subset are more consistent and less noisy: morphological complexity steadily increases as L2 instructional exposure increases, while syntactic complexity (i.e. word order rigidity) tends to decrease with increasing L2 instructional exposure, with the exception of groups V and V1.

By way of an interim summary, we conclude that (a) L1 background matters, but that (b) the relationship observed in Section 4.1. between essay complexity and L2 instructional exposure persists across German learners, and seems to be quite robust across other L1 backgrounds.

5. Discussion and conclusion

Drawing inspiration from information theory and the literature on cross-linguistic complexity variation, we have sketched in this paper a complexity measure previously not utilized in SLA research: KOLMOGOROV COMPLEXITY*,* which defines the complexity of a text as being proportional to the length of the shortest algorithm that can (re-)generate that text. The compression technique used in this paper specifically approximates Kolmogorov complexity by compressing texts using the gzip program.

The basic idea behind Kolmogorov complexity is that language (in fact, text) should count as more complex the less predictable it is. Kolmogorov complexity is thus a text-based, holistic and global measure of structural surface redundancy. It is related to the notion that more varied and/or diverse language should count as more complex (Bulté & Housen 2014:45), a view that is especially popular in lexical complexity research (Jarvis 2013). Along the way, we showed how to combine the instrument with distortion techniques to measure not only the overall complexity of texts, but also complexity in morphology and syntax.

Thus we investigated essays written by advanced learners of English, as sampled in the *International Corpus of Learner English* (ICLE) (Granger, Dagneaux & Meunier 2002; Granger 2003), which covers essays from students of numerous L1 backgrounds with different levels of exposure to English language instruction. We drew on a pseudolongitudinal research design (Gass & Selinker 2001:32–33) to explore the relationship between essay complexity and the amount of instruction that the learners in ICLE had received.

The analysis showed that increased L2 instructional exposure predicts increased overall complexity and increased morphological complexity, but decreased syntactic complexity. Recall that in our study design, the polarity of the syntactic complexity axis is set such that rigid word order counts as complex, so the fact that we measure a decrease in syntactic complexity, i.e. a decrease in word order rigidity, basically means that learners use more varied word order patterns as they receive more instruction. This is in line with the customary view (discussed above) that varied language should count as complex. The positive relationship between overall and morphological Kolmogorov complexity and L2 instructional exposure is robust across a number of L1 backgrounds, although L1 background is a good predictor of overall complexity. We saw, for example, that German learners of English tend to produce more Kolmogorov-complex essays than French, Italian, or Spanish learners of English.

The finding that increased L2 instructional exposure – and thus, by reasonable inference, proficiency (though we acknowledge that many factors contribute to a learners' degree of proficiency) – should correlate with more complex written production is certainly not unexpected. But that the compression method is able to pick up this relationship demonstrates that the method works as advertised. Now, why do we need yet another measure of complexity in SLA materials? Kolmogorov complexity is not the magic bullet that will solve all problems that beset CAF research (Norris & Ortega 2009). However, the measure does have a number of advantages over more traditional complexity measures utilized in the extant CAF literature. First and foremost, Kolmogorov complexity is a more holistic notion than e.g. unit length measures, thanks to its radical text-basedness: it is not based on the recurrence of arbitrarily selected features but on texts (and their predictability) as a whole. Because of this inherently holistic nature, Kolmogorov complexity is well-suited for capturing the complex, multi-dimensional nature of L2 complexity. That being said, the compression method can be flexibly combined with various distortion techniques to measure complexity on particular linguistic levels (Ehret 2014 demonstrates how distortion even enables the analyst to study the Kolmogorov complexity contribution of particular morphs and constructions). Finally, we note that Kolmogorov complexity as a holistic, quantitative measure of text complexity is both more economical to obtain and arguably more objective than e.g. subjective proficiency/complexity ratings of learner texts by expert evaluators.The compression technique can thus serve both as an independent analysis tool and as a complementary diagnostic in research. In proficiency testing contexts, it could potentially be used along with assessment questionnaires to benchmark proficiency.

The main limitations of Kolmogorov complexity include the following. For one thing, the current state of the technology is such that the compression method does not work well with short texts (that is, texts that count less than 1,000 words). Complexity measurements obtained through compression are more robust and representative if they are based on larger datasets. Furthermore, the data should be equally distributed across the samples to be assessed, i.e. the samples should be of the same size. As for minimum sample size, experiments with parallel corpora show that the Gospel of Mark, which counts around 15,000 words in the English Standard version, is sufficiently large, but the Lord's Prayer with only about 50 words, for instance, is not. Another limitation is that Kolmogorov complexity scores are inherently relative – they are hard to interpret in absolute terms, and are really meaningful only when seen in the context of the rankings in which they are presented. It is also clear that the method's rather unitary/holistic view on complexity makes it harder to study discrete subdimensions of (complexity) development, although of course the analyst can marshal selective distortion techniques, as we have seen in this paper.

This is a proof-of-concept study, and hence much remains to be done in future research. Our ICLE-based findings must, for reasons of data availability and distribution, be considered tentative. As always, further exploration with a larger dataset and learners with more L1 backgrounds are needed. It would also be desirable to correlate Kolmogorov complexity scores with subjective expert ratings of essay complexity and/or with more objective ratings of L2 learners' proficiency (as available, e.g., in The International Corpus Network of Asian Learners of English, http://language.sakura.ne.jp/icnale/), for the sake of corroborating the construct's validity.

Appendix

Table 4: Data coverage by L2 instructional exposure (sub)groups: total number of texts, sentences, and words are provided for each group.

L1 background	group	# texts	# sentences	# words
German	I	25	595	11,491
	II	22	635	11,926
	III	24	697	13,267
	IV	107	2,753	51,605
	V	80	1,981	37,031
	VI	85	2,714	50,894
French	I	2	100	1,344

	II	77	2,742	46,719
	III	134	4,423	80,605
	IV	*nil*	*nil*	*nil*
	V	4	138	2,247
	VI	3	114	1,674
Italian	I	3	54	1,503
	II	6	117	3,093
	III	27	634	14,559
	IV	2	66	987
	V	9	232	5,169
	VI	34	906	20,083
Spanish	I	14	437	8,876
	II	2	60	1,342
	III	6	188	3,890
	IV	86	2,806	55,650
	V	6	181	3,666
	VI	18	858	16,766

References

Bestgen, Yves & Sylviane Granger. 2014. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing* 26. 28–41. doi:10.1016/j.jslw.2014.09.004.

Biber, Douglas & Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9. 2–20.

Biber, Douglas, Bethany Gray & Kornwipa Poonpon. 2011. Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? *TESOL Quarterly* 45(1). 5–35. doi:10.5054/tq.2011.244483.

Brezina, Vaclav & Gabriele Pallotti. 2015. Morphological complexity tool, available from http://corpora.lancs.ac.uk/vocab/analyse_morph.php.

Bulté, Bram & Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken & Ineke Vedden (eds.), *Dimensions of L2 Performance and Proficiency: Investigating Complexity, Accuracy and Fluency in SLA*, 21–46. Amsterdam: Benjamins.

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

Dryer, Matthew S. 2013. Order of Subject, Object and Verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/81.

Ehret, Katharina. 2014. Kolmogorov complexity of morphs and constructions in English. *Language Issues in Linguistic Technology* 11. 43–71.

Ehret, Katharina. 2016. An information-theoretic approach to language complexity: variation in naturalistic corpora. Freiburg PhD dissertation.

Ehret, Katharina & Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler & Guido Seiler (eds.), *Complexity, Isolation, and Variation*, 71–94. Berlin, Boston: De Gruyter. http://www.degruyter.com/view/books/9783110348965/9783110348965-004/9783110348965-004.xml (2 August, 2016).

Gass, Susan M. & Larry Selinker. 2001. *Second language acquisition: an introductory course*. 2nd ed. Mahwah, N.J: L. Erlbaum Associates.

Granger, Sylviane. 2003. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly* 37(3). 538–546.

Granger, Sylviane, Estelle Dagneaux & Fanny Meunier (eds.). 2002. *The International Corpus of Learner English: Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Jarvis, Scott. 2013. Capturing the Diversity in Lexical Diversity: Lexical Diversity. *Language Learning* 63. 87–106. doi:10.1111/j.1467-9922.2012.00739.x.

Juola, Patrick. 1998. Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics* 5(3). 206–213.

Juola, Patrick. 2008. Assessing linguistic complexity. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language Complexity: Typology, Contact, Change*, 89–107. Amsterdam, Philadelphia: Benjamins.

Kettunen, Kimmo, Markus Sadeniemi, Tiina Lindh-Knuutila & Timo Honkela. 2006. Analysis of EU Languages through Text Compression. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo & Tapio Pahikkala (eds.), *Advances in Natural Language Processing*, 99–109. (Lecture Notes in Artificial Intelligence 4139). Heidelberg: Springer-Verlag Berlin.

Klein, Wolfgang & Clive Perdue. 1997. The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research* 13. 301–347.

Kolmogorov, Andrej N. 1965. Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii* 1(1). 3–11.

Kortmann, Bernd & Benedikt Szmrecsanyi. 2012. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: de Gruyter.

Kusters, Wouter. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.

Li, Ming, Xin Chen, Xin Li, Bin Ma & Paul M. B Vitányi. 2004. The similarity metric. *IEEE Transactions on Information Theory* 50(12). 3250–3264.

Li, Ming & Paul M. B Vitanyi. 1997. *An introduction to Kolmogorov complexity and its applications*. New York: Springer.

Lubbe, J. C. A. van der. 1997. *Information theory*. Cambridge [England] ; New York: Cambridge University Press.

Martens, Scott. 2011. Quantifying linguistic regularity. KU Leuven PhD dissertation. http://ccl.kuleuven.be/~scott/quantifyingLinguisticRegularity.pdf.

McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 6. 125–166.

Miestamo, Matti, Kaius Sinnemäki & Fred Karlsson (eds.). 2008. *Language complexity: typology, contact, change*. (Studies in Language Companion Series v. 94). Amsterdam, Philadelphia: Benjamins.

Norris, J. M. & L. Ortega. 2009. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics* 30(4). 555–578. doi:10.1093/applin/amp044.

Ogden, C. K. 1934. *The system of Basic English*. New York: Harcourt.

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

Ortega, Lourdes. 2012. Interlanguage complexity: A construct in search of theoretical renewal. In Bernd Kortmann & Benedikt Szmrecsanyi (eds.), *Linguistic Complexity: Second Language*

Acquisition, Indigenization, Contact. Berlin: De Gruyter.

Pallotti, G. 2009. CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics* 30(4). 590–601. doi:10.1093/applin/amp045.

Pallotti, Gabriele. 2015. A simple view of linguistic complexity. *Second Language Research* 31(1). 117–134. doi:10.1177/0267658314536435.

Paquot, Magali & Sylviane Granger. 2012. Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics* 32. 130–149. doi:10.1017/S0267190512000098.

Sadeniemi, Markus, Kimmo Kettunen, Tiina Lindh-Knuutila & Timo Honkela. 2008. Complexity of European Union Languages: A Comparative Approach. *Journal of Quantitative Linguistics* 15(2). 185–211.

Sampson, Geoffrey. 2009. A linguistic axiom challenged. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 1–18. Oxford: Oxford University Press.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27. 379–423.

Szmrecsanyi, Benedikt. 2015. Recontextualizing language complexity. In Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman & Hubert Cuyckens (eds.), *Change of Paradigms -- New Paradoxes: Recontextualizing Language and Linguistics*, 347–360. (Applications of Cognitive Linguistics [ACL] 31). Berlin, Boston: De Gruyter Mouton.

Trudgill, Peter. 2011. *Sociolinguistic typology : social determinants of linguistic complexity*. Oxford, New York: Oxford University Press.

Ziv, Jacob & Abraham Lempel. 1977. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory* IT-23(3). 337–343.