This version, February 2016. To appear in Second Language Research (special issue on linguistic complexity).

Morphological complexity in written L2 texts

Vaclav Brezina, Gabriele Pallotti

Introduction

Morphological complexity (MC) is a relatively new construct in L2 studies. Most SLA research to date has focused on lexical and syntactic complexity (for a review see Bulté & Housen, 2012, who report that only six studies out of forty included morphological complexity measures). Despite this lack of attention to MC in SLA - which may be explained by a prevailing focus on English, a language with few inflectional resources - MC plays a crucial role in a full and theoretically adequate description of the language learning process (De Clercq & Housen, this issue), This becomes even more apparent in morphologically rich languages, whose inflectional paradigms have all the properties of complex systems: they consist of many formal elements expressing a number of grammatical functions; the relationships among these forms and functions are complex, too, because they often involve cases where one form realises several grammatical functions (synchretism) or the same grammatical function is realised by several forms (allomorphy). This is why many morphological systems can be said to have high entropy (Ackerman & Malouf, 2013), in that the relationships among different parts of the system cannot be straightforwardly derived from a small set of systematic rules.

Acquiring inflectional morphology in a first or a second language is thus no easy task (DeKeyser, 2005; Lardiere, 2006). Learners must identity at the same time the forms - or, better, morphologically-conditioned formal variations of lexical bases – and their functions, which often realise subtle grammatical meanings, not shared by all languages and that could in principle be expressed by non-morphological means as well (Carstairs-McCarthy, 2010; Housen and Simoens, 2016). In some languages, this means reconstructing highly complex abstract systems, which may take many years, with a number of intermediate stages, characterised by partial and unstable representations of the target grammar (Slabakova 2009).

In order to track this development, it is desirable to have an objective metric to express how the complexity of inflectional paradigms deploys over time, and how it varies across different languages. It is in fact clear that an interlanguage's complexity depends on the one hand on the level reached by the learner, and on the other hand on the complexity of the target language itself (DeKeyser, 2016; Housen & Simoens, 2016).

This article thus aims to propose a metric allowing the comparison of inflectional systems both within a language (developmental varieties) and across languages (comparative interlanguage analysis). Following the 'simple view of complexity' advocated by Pallotti (2015), we will intentionally restrict the scope

of the construct, in order to make it more internally consistent and operationalisable – there is certainly more to interlanguage morphology analysis than what is covered by the present definition, but we believe that it can be a valuable contribution to the growing debate on how to define and measure morphological complexity. First, the term 'complexity' will be applied to structural aspects of linguistic outputs (i.e. linguistic performance) only, defining it as "the number of different elements and their interconnections (i.e. their systematic, organized relationships)" (Pallotti, 2015, p. 120). More specifically, we will define morphological complexity as the diversity of inflectional types of a given word-class – verbs, in the present case. Secondly, the focus will be on inflectional *forms* occurring in written texts, disregarding the complexity of form-meaning relationships, for reasons that will be explained below. A linguistic and mathematical analysis will then lead to the calculation of the Morphological Complexity Index (MCI). This measure will be empirically tested in two case studies on interlanguage morphology in L2 written texts in Italian and English.

In English verbal morphology can be said to be relatively simple, having just three regular inflectional forms (*-ed, -s, -ing*) and a few dozen irregular verbs. Italian is much more complex: there are three conjugation classes, normally distinguished by a theme vowel (TV); all verbs are systematically inflected for tense, aspect and mood (TAM), and all finite forms are also inflected for three persons and two numbers. In some cases, the person/number exponent is concatenated to the TAM exponent, as in *parl-a-v-o* (speak-TV-PST.IPFV-1SG), in others all these grammatical values are fused in a single exponence, as in *parl-ai* (speak-PST.PFV-1sg). This gives rise to over one hundred forms for regular verbs, plus several hundred irregular verbs exhibiting various types of allomorphy of the base and/or of the inflectional endings.

Since morphological complexity is a relatively new construct in SLA research and the MCI is a new way or measuring it, a large emphasis in this article is placed on the conceptual and methodological underpinnings of the definition and operationalisation of these two notions.

Previous research

In the last decades, morphological complexity has been extensively discussed in typological linguistics, focussing mostly on the structural complexity of language systems (e.g. Baechler & Seiler, 2012; Baerman, Brown & Corbett, 2015; Shosted, 2006; Stump & Finkel, 2013). These studies have mainly been concerned with morphological complexity at the level of Saussurean *langue*, comparing, for example, the complexity of inflectional paradigms across different languages or language varieties. This research, while giving important insights in the structure of paradigms and inflected word forms in native languages, is not of direct relevance when analysing L2 data. In fact, both native languages and interlanguages can be seen as 'systems' ideally reconstructed from the observation of concrete texts, or *parole*. However, native languages are relatively stable and often used by large communities, making it possible to publish descriptive grammars, which are the main source for most typological work on morphological complexity. Interlanguages, on the other hand, are by definition unstable, idiosyncratic systems, whose (often variable and probabilistic) 'grammar' can only be inferentially reconstructed from a few observable texts. This is why typological researchers may choose whether to analyse complexity in actual texts or in those ideally stable and abstract 'distillations' called grammars, while SLA researchers can only describe complexity in the texts they observe.

The few existing studies on morphological complexity in L2 learners, in fact, all measure this trait at the level of specific acts of performance, i.e. oral or written texts. For example, Bygate (1996), looking at how one learner's production changed from the first to the second performance of the same task, counted the number of 'verb forms' in the text. Foster (1996) and Foster and Skehan (1996) computed what they called 'syntactic variety' by analysing the range of tense, modality, aspect and voice forms on the verbs used by their learners. Ellis and Yuan (2004, 2005; Yuan, & Ellis, 2003), too, defined 'syntactic variety' as the 'total number of verb forms used in the task. Grammatical verb forms included tense (e.g. simple past, past continuous), modality (e.g. *should, have to*), and voice (e.g. passive voice in the past)' (Ellis & Yuan, 2005, p. 181). In all these studies, morphological complexity was operationalized as the range of verb forms, with only few indications about its practical implementation. For instance, did the authors count verb-form types or tokens - i.e. if two instances of the simple past were found in a text, possibly on different verbs, did this amount to one or two 'verb forms'? Furthermore, counting the number of forms in absolute terms is clearly sensitive to text length, as longer texts are more likely to include more forms. A more sophisticated approach was used by Verspoor et al. (2012) who, among other things, compared broad grammatical categories of verbal forms across proficiency levels. These included tense, aspect, voice and modalised forms and showed some basic discriminatory power between proficiency-based groups of learners. However, due to a broader focus of the paper, the specific implications of these findings for morphological complexity were not explored.

First-language acquisition researchers, too, have been interested in assessing the development of morphological complexity, and a few studies have proposed more sophisticated measures than just counting the absolute number of inflection types. For example, Malvern et al (2004) propose the Inflectional Diversity (ID) index, which is based on their D index of lexical diversity. First, lexical diversity is calculated using the D index applied to all word forms, so that *go, goes* and *went* would represent three different types. Then D is computed again, but this time on verb stems or verb roots: in the first case, *go* and *goes* count as two tokens of the stem *go,* while in the second, *go, goes* and *went* count as three tokens of the lemma 'to go'. ID simply results from subtracting D_{roots} or D_{stems} from $D_{wordforms}$. Malvern et al (2004) show that ID tends to increase in English-speaking children between 18 and 36 months, reaching values of about 2.5 for ID_{stems} and 4 for ID_{roots}. Children learning Spanish, a language with much richer inflectional paradigms, display a more sustained trend, with ID values steadily growing from 18 to 42 months, until they reach about 15 for ID_{roots} and 10 for ID_{stems}. ID is based on D, a measure that is relatively insensitive to text length, and this is certainly an advance with respect to previous approaches. However, this proposal, too, suffers from two serious limitations. Firstly, at least in Malvern et al's account, ID does not differentiate among word classes, so that the inflectional diversity stemming from nouns, verbs, adjectives and determiners is confounded. Secondly, and more importantly, ID is clearly related to D, which means that higher levels of lexical diversity automatically produce lower levels of inflectional diversity. In the extreme case of a text containing only one token for each lexical type, ID will be zero ($D_{wordforms} = D_{stems} = D_{roots}$) regardless of whether these lexemes all contain one and the same inflection or exhibit a wide range of morphological markers.

These problems are clearly identified by Xanthos and Gillis (2010), who propose an alternative measure of inflectional diversity, called Normalized Mean Size of Paradigm, or MSP(S). The approach consists in extracting from a text N

samples of S words, and then calculating the mean number of different inflected forms of a given word-class (e.g. verbs or nouns) for each of these subsamples. Working with subsamples of a standard size eliminates the effects of text length, as with Johnson's (1944) Mean Standardized Type/Token Ratio (MSTTR), and inflectional diversity is directly computed in its own terms, and not subtractively as with ID. While this proposal marks a significant progress with respect to existing measures, it leaves a few unresolved issues. The first is that in a sample of S words the number of tokens of a given word-class, e.g. verbs, may vary. Even though the use of repeated random sampling may limit this effect, the mean size of paradigm in S-word samples is nonetheless conditioned by the mean density of a certain word-class in the whole text. This is a clear confounding variable because samples with a smaller proportion of the word-class of interest will also have a smaller chance of occurrence of different inflected forms. Secondly, the authors are not clear as to what size of S they would recommend. In their first article they report findings for MSP(50) and MSP(500), i.e. based on 50- and 500-word samples, noting that the measure clearly increases with sample size, as is to be expected. In another publication (Xanthos et al., 2011), MSP(50) is used for assessing inflectional diversity in children's speech, while MSP(1,000) is used for analysing caregivers' utterances, and it is not clear how measures based on completely different scales may be compared. Finally, no clear indication is given as regards the way inflectional forms are to be identified and counted in the corpus.

The measure proposed here is an extension of an initial proposal by Pallotti (2015) and is based on the same logic as MSP(S), i.e. calculating inflectional diversity in standardised samples, but it aims to overcome the difficulties described above. The main goal of this article is to present the measure and discuss how it can be applied to L2 data, also critically addressing problems of morphological analysis in interlanguages. In order to do so, the measure will be employed in two case studies on verbal morphological complexity in native and non-native speakers, one on English, a morphologically simple language, the other on Italian, whose verbal paradigms exhibit a much richer array of inflectional endings. In both studies, MCI values will be computed and correlated to other measures of lexical and syntactic complexity. The aim of the case studies is primarily methodological; they were designed to illustrate how variation in morphological complexity can be meaningfully investigated and how the measure of morphological complexity relates to other existing complexity measures used in SLA. The study explores a wide range of factors (including the language of the texts analysed) that are of interest in the study of interlanguages.

The Morphological Complexity Index (MCI)

Linguistic analysis

The Morphological Complexity Index (MCI) is a measure of the average inflectional diversity for the occurrences of a given word class in a text – in this article we will restrict our discussion to verbs. It bears some resemblance with indices of lexical diversity, such as the type-token ratio. A text containing *talk, write, drink* can be said to be lexically more complex than one containing *talk, talk, talk* (or *talk, talking, talks*), at least in the sense of complexity as diversity, i.e. range of forms (Page, 2010). Likewise, a text containing *talk, talking, talks* can be said to be morphologically more complex than one with *talking, talking,*

talking (or *talking, writing, drinking*). The difference is that in the first case one has to do with lexical types and their repetition as tokens, whereas in the second case one considers the range and diversity of inflectional types and tokens. Before presenting how diversity is mathematically computed in our approach, we need to discuss the relatively new notion of 'inflectional type' and its operationalization.

In a language like English, it is often quite easy to identify the inflectional part of a verb, as in the examples given above, where one might say that Ø, *-ing* and *-s* are three different inflectional forms. By the traditional structuralist account, these would be called three morphs, each related to a certain grammatical meaning to produce three different grammatical morphemes. However, the notion of morpheme has been the object of considerable criticism in the last decades, and many morphologists today believe it should be entirely dismissed (for an early discussion, see Anderson, 1992). While it may hold relatively well for concatenative processes like 'adding *-ing* to the lexical base', the notion of morpheme proves to be highly problematic for other types of inflectional operations, like stem change or reduplication; the traditional notion of zero-morph is equally regarded as untenable by many. A more general approach that focuses on 'morphological processes' or 'operations' seems to be less controversial, conceptualising inflectional morphology as a series of manipulations to lexical bases such as adding affixes, changing the lexical base in various ways, selecting a different base, or leaving the base untouched. The more technical term for these manipulations is 'exponence', which covers both cases involving one single operation (or exponent) or multiple operations.[1]

While it is commonplace to theoretically characterize interlanguages as systems governed by their own rules and characterized by their own regularities, in practice it is often difficult to write a 'grammar' of an interlanguage in the same sense as one does with native languages[2]. In particular, the difficulty lies in finding systematic descriptions of how different forms (exponences) express particular functions (grammatical meanings, or 'morphosyntactic property sets' (Stump, & Finkel, 2013), and how these form/function relationships are organised in paradigms. Thus, establishing the grammatical meaning of some inflected interlanguage forms often involves a large amount of conjecture and speculation. For example, can one be sure that a sentence like *John playing in park* encodes progressive aspect, and not just a generic present tense? This form may perhaps be used as default in all temporal and aspectual contexts, as is often the case in certain initial varieties. Also, can one say that it encodes indicative mood, in a system where there seem to be no traces of subjunctive, conditional or other

1 There is not much consensus in today's theoretical and descriptive morphology as regards terms like *root, stem, base, inflection*, but a thorough discussion of the area would require a paper of its own (for an introduction, see Spencer, 2012). For our current purposes, we will use the most neutral term '(lexical) base' to indicate the form of the lexeme undergoing morphological processes. Some lexemes have a variety of base forms, often called 'stems', and morphologists do not agree on whether these should be considered as part of the lexicon, the result of a specific morphological component, or lying somewhere in between. We will occasionally use the term 'stem' to refer to alternative base forms that undergo some further, clearly morphological modifications, e.g. person / number inflection.

2 Most studies on interlanguage grammars have focused on very specific aspects, trying to collect data samples with the highest possible 'data density' (Pienemann, 1998) of the particular structure(s) under examination.

moods? This is the reason why our characterization of morphological complexity will be restricted to the diversity of inflectional *forms*, which already presents a number of methodological challenges, some of which will be discussed below[3].

Having thus defined morphological complexity in terms of the diversity of inflectional *forms*, we need to operationalise the construct of inflectional form, or exponence, in such a way that it can be reliably applied to a variety of texts produced by native and non-native speakers of different languages. Given the difficulty of working on oral morphology, which requires a full phonetic transcription of oral corpora, in this preliminary study we are going to focus on written morphology, as realised in written texts. Hence our operational definition will only describe what happens to *written forms* when they undergo inflectional processes. For the sake of space and simplicity, most exemplification will be carried out on English, but the procedure has been applied so far to German and Italian, too (Pallotti & Brezina, 2015), with extensions to French and Spanish under way.

The basic idea is that any inflected word form can be described in terms of the changes occurring to a base as a result of inflectional processes. In the case of concatenative morphology, the process can be straightforwardly described as adding a graphological string to the lexical base, as in *talk-ed*, with the exponence simply being the appended string. But what about forms involving internal modifications of the base, as in *found*? Treating all these cases as irregular suppletives would dismiss the fact that some of these so-called irregular verbs do follow some 'minor rules' of a kind (Lightner, 1968). These are the remnants of once-productive inflectional processes, which even synchronically do exhibit regular patterns – in this case, *find/found* clearly patterns with *grind/ground, bind/bound* etc., and it does indeed occur that L1 and L2 learners, and even neural networks, occasionally treat these minor rules as productive (Taatgen & Dijkstra, 2003). One would thus want to consider *found, ground, bound* as three tokens of one inflectional type, namely the process turning a base containing an <i> into a word form containing <ou>.

In order to systematically describe all these processes affecting the lexical base, it is necessary to provide clear criteria for the identification of the base and the processes. As regards the identification of the base, one needs to establish whether, in a pair like *find / found,* the base is *find* and *found* the form that underwent the inflectional process, or vice versa. While the question may sound idle in English, in languages like Italian, German or French, where many verbs have a variety of allomorphic bases, it is not always clear what the base and the derived forms should be. The proposed operationalization identifies the default base (DB) as that which, in the target language, is common to most cells in the verb paradigm[4]. Hence, *find* can be said to be the verb's default base because it is

3 Analysing the diversity of forms is, however, not a necessary requirement for our approach and researchers willing to measure the complexity of form/function relationships may well do so, simply computing the average diversity of strings containing exponence/meaning mappings, as in standard morphemic transcriptions like e:PRS.IND.3SG (meaning that the form /e/ encodes third person singular of the present indicative).

4 Clearly, the cells in the paradigm also differ in terms of the frequency of their actual use, so that some forms which are theoretically possible are rarely used in current language. In this article, decisions about the DB were based on the theoretical paradigm and not on token frequencies, as this was considered more practical. It is indeed possible, and certainly a fruituful direction for

used to build six persons of the present tense (including *finds*), the present participle *finding*, the infinitive *to find* and the imperative (9 cells); *found* only covers six persons of the simple past plus the past participle (7 cells). Likewise, the German verb *sprechen* uses the base *sprech-* for four persons in the present tense, six of the subjunctive I, three of the imperative, the infinitive and the present participle (15 cells); the base *sprich-* is used in two persons of the present tense and one of the imperative; *sprach-* appears in six persons of the simple past, *spräch-* is the base for six persons of the subjunctive II, while *-sproch-* is only used in the past participle *gesprochen*. S*prech-* can thus be said to be the default base.

Having defined the default base, inflectional processes may be characterized as changes with respect to this base. Thus, a written word form like *found* can be said to be the result of changing <i> in the base into <ou>. A form like *sprachst* (speak.PST.2SG) can be described as a two-step process turning the <e> of the DB *sprech-* into an <a> and then appending the suffix -st.

We hasten to make clear that we do not claim any historical, psychological or theoretical validity for this way of characterizing inflectional processes, which should be taken as a purely descriptive algorithm. We are well aware of its limitations. In comparison to current theoretical discussions on inflectional morphology, our approach looks rather simple. It treats all inflectional processes as one single exponence, when it is clear that, at least in some cases, these may more appropriately be described in terms of two or more consecutive processes (e.g. stem formation and affixation, as in *sprachst* = (i) forming the stem *sprach-* and (ii) affixing the 2SG ending *-st*), or the simultaneous addition of two or more distinct inflectional endings (e.g. in Italian *parlavo* 'I used to speak' , where -v- encodes past imperfective tense/aspect and -o 1SG person). Furthermore, DBs are identified by reference to the target, native language, which implies some form of comparative fallacy.

However, we also think that the procedure has some strengths, beginning with its simplicity. This means high scoring reliability with little or no room for subjective interpretations, straightforward application to a variety of typologically different languages and implementability by a computer. From a theoretical point of view, it is a systematic application of the item-and-process model of morphological analysis (Hockett, 1954), which is still considered to be valid and effective way of accounting for morphological phenomena across languages. The basic logic is that of the 'edit distance' measures, commonly used in computational linguistics as an objective way of calculating relationships among word forms (Kruskal, 1999), and is thus particularly suited to an approach like ours which involves automatic computation of the index.

It is important to note that our analysis of interlanguage data aims at identifying the array of inflectional forms used by the writers, with no concern with their accuracy in terms of L2 norms. Thus, word-forms like *they finds* or *they finded* or *she find* are analysed as containing the exponences -s, -ed and Ø[5], respectively. The same holds, for instance, for Italian verbs inflected with exponents from the 'wrong' inflection class, as in *scriva*, which consists of the lexical base *scriv-* ('to write') with the *-a* ending typical of first conjugation verbs,

future research, to identify DBs according to their frequency in one or more corpora of language use, which would make the MCI an even more usage-based construct.

5 Our use of Ø for denoting a particular exponence does not imply any commitment to a theoretical notion of zero-morph, but simply means that the process for this particular form consists in reproducing the DB without any change.

while it would have required the *-e* ending of the second conjugation. Furthermore, misspellings in the lexical base, such as *arived*, are ignored when it is clear what morphological exponent is involved (in this case, *-ed*). It is even possible to identify an exponence in non-existent words, such as coinages like *two waker-s* (for *two alarm clocks*) or lexical loans from the L1 (as in *she impast-ed*, after the Italian *impastare* 'to knead'). There are however a few cases for which it is not possible to reach a conclusion as to what should be counted as the exponence, and they are thus excluded from quantitative analysis. Among these are completely unintelligible and unanalysable words, or word forms like *commite*, which could equally be parsed as *commit-e* (with an idiosyncratic rendering of the exponent) or *commite-* Ø (with an idiosyncratic rendering of the base).

Mathematical analysis

The procedure outlined above allows one to identify all exponences of a relevant word class in a text. The MCI is computed by calculating their average diversity. In order to do so, a number of samples n (e.g. 100) of k (e.g. 10) exponences are randomly extracted from the text, ensuring that the same form is not extracted twice in the same sample (sampling without repetition), although it may occur more than once in different samples. For each sample, the number of different exponences (inflectional types) is computed, to arrive at an average within-set diversity score. For each pair of k-exponence samples, across-set diversity is also calculated, by counting, for each pair, how many forms belong exclusively to one of the two sets. The results of both the within-set variety (component a) and between-set diversity (component b) comparisons are averaged. The MCI value consists in adding the mean within-sample average diversity to the mean across-sample average diversity/2, and subtracting 1; hence:

MCI = (within-subset variety + between-subset diversity/2) – 1

Take, for example, a short English text that contains the following 22 exponences (this is an invented example for illustration purposes): Ø, Ø, Ø, ed, ing, s , Ø, Ø, ing, ed, Ø, Ø, Ø, ed, ed, Ø, Ø, Ø, ed, ed, ing, Ø. We first extract two random 10-exponence samples (assuming k=10) and calculate the within subset variety of each and the mean subset variety:

Sample 1: s, Ø, Ø, ing, Ø, Ø, Ø, ing, ed, ed; within-subset variety$_1$ = 4
Sample 2: ed, Ø, Ø, ing, Ø, Ø, ed, ed, ed, Ø; within-subset variety$_2$ = 3
Mean subset variety = (4+3)/2 = 3.5

After this, between-subset diversity is computed by comparing samples 1 and 2. As can be seen, sample 1 has one unique exponence type (s), while sample 2 does not have any. The mean value of between subset diversity thus is 0.5

The MCI (based on two samples) will therefore be:

MCI = (3.5 + 0.5/2) – 1) = 2.75

The theoretical range for MCI calculated with 10-verb samples (MC10) thus goes from a minimum of 0 (1+0-1) to a maximum of 19 (10+20/2-1). The choice of 10-verb samples clearly implies some arbitrariness: besides 10 being a round number, and allowing MCI to be calculated on samples of 21 verbs or more (which roughly correspond to 100-word texts, a reasonable size for many projects

on L1 and L2 acquisition), there are no other special reasons for choosing this value. We are going to conduct in-depth validation studies to assess the effects on MCI of using smaller or larger samples. A preliminary analysis shows that there are very high correlations among different MCI values with samples ranging from 5 to 15 (Pallotti, in press), which suggests that perhaps MCI could be computed on 5-verb samples (MC5), thus allowing analysis of texts containing just 11 verb tokens. Another question to be addressed in future research is whether it might be possible to simplify the measure even further, by just calculating within-set variety (component 'a' in the current formulation) and dispensing with across-set diversity (component 'b'). The resulting measure may be called MC10a (in the case of 10-verb sets) and MC5a (in the case of 5-verb sets).

It is important to underscore the fact that linguistic and mathematical analyses are completely independent. Hence, even if one were to follow a different linguistic analysis in order to identify inflectional types and tokens, both on a general level and for specific interlanguage samples, it would still be possible to calculate the diversity indices following the proposed mathematical procedure.

Computer implementation

The data were analysed using a computer tool developed by the authors of this study (http://corpora.lancs.ac.uk/vocab/analyse_morph.php) that implements the operational definition of morphological complexity by sequentially performing the two levels of analysis. First, the tool carries out a linguistic analysis that identifies the word class of each word in a text (token) and assigns it the dictionary form (headword) using the TreeTagger (Schmid, 1994). Each token is then compared with the headword and its specific inflectional form (exponence) is identified. The results of this automatic linguistic analysis were manually checked for accuracy and all systematic errors were corrected. Accuracy of automatic analysis for native speakers' data was very high from the start for both English (98.18%) and Italian (86.73%).

Second, after the text has been linguistically analysed and exponences have been extracted, the tool computes the MCI by randomly drawing 100 subsets of k exponents from the text and computing the average within- and across-subset MC. In what follows, we are going to present results based on MC10, i.e. the Morphological Complexity Index calculated on 10-verb samples and including both within-set variety and across-set diversity. In addition, we will also briefly discuss the relationship between MC10 and MC5a, another possible operationalization, which is calculated on 5-verb samples and includes only within-set variety; this measure is especially useful for very short texts.

Study 1

Method

The first case study is based on written argumentative essays produced by Dutch university students learning Italian as a foreign language and by native-speaking Italian university students, taken from the project 'Communicative Adequacy and Linguistic Complexity in L2 Writing' (CALC) (Kuiken, Vedder & Gilabert, 2010; Kuiken & Vedder, 2014). Learners' proficiency level ranged from A2 to B2 (Council of Europe, 2001). Both learners and native speakers of Italian produced two short argumentative essays, one about which of three charities

should be funded by a small university grant, the other asking to choose one of three topics for an article to appear on the first page of the monthly magazine of the local newspaper. For the purposes of this study, the two texts were combined for each writer and analysed as one piece, in order to achieve a sufficiently long sample to calculate MC10 for all participants. The two essays were written by the same person on the same day and belong to the same genre, so they can legitimately be considered a homogeneous sample of that person's (inter)language. Table 1 shows the details of the dataset for both the non-native speaker (NNS) and the native speaker (NS) group. On average, NNS produced essays which were 251 words long (*SD* = 54), while NS produced essays which were slightly shorter (*M* = 244, *SD* = 63). The essays by the NNS speakers included on average more verbs (*M* = 51, *SD* = 10.94) than the essays by NS (*M* = 42.78, *SD* = 11.49). This also resulted in a higher density of verbs per 100 words in NNSs' texts (20.32) than in NSs' texts (17.53), showing the problematicity of an approach such as MSP(S) by Xanthos and Gillis (2010), which calculates diversity over N-word samples under the assumption that the density of a given word class remains constant across groups and samples.

Table 1
The Italian corpus

Corpus	No. of texts	Corpus size (words)	Mean text length (*SD*)	Mean verbs/text (*SD*)
NNS	39	9,793	251 (54)	51.00 (10.94)
NS	18	4,384	244 (63)	42.78 (11.49)

In addition, two MCI indexes (MC10 and MC5a) and text measures of lexical (standardised TTR with 100-word samples, based on lemma counts) and syntactic complexity (sentence length) were computed. Participants' overall language proficiency was established by means of a C-test where "learners were asked to complete 100 words in five short texts in which half the letters of every other word had been replaced by blanks" (Kuiken & Vedder, 2014, p. 336).

Results

Table 2 *MC10 in native and non-native speakers, Italian*

Corpus	*M* (*SD*)	*Range*
NS	12.85 (1.42)	10.60 – 15.33
NNS (all)	11.75 (1.93)	7 – 15.30
NNS-high (> 71)	*13.05 (1.15)*	*11.33 – 15.30*
NNS-low (≤ 71)	*10.64 (1.78)*	*7 – 14.30*

Verb morphological complexity was higher in NSS's texts (Table 2), and the difference is statistically significant (Welch t-test: t (44.17) = -2.41, p = 0.02), with a medium effect size (r = 0.341). There is also more variance in learners' data, with a very wide range of scores (7 – 15.30). If learners are divided into two broad proficiency groups, based on whether their C-score was higher (NNS-high) or lower than or equal to (NNS-low) the median C-score value of 71, differences

are even clearer. The mean MCI score was 10.64 in the lower-level learners, 13.05 in the higher-level learners and 12.85 in native speakers (one-way ANOVA: F (2, 54) = 16.03, $p < 0.001$, effect size (ω) = 0.588, large effect). Post-hoc Bonferroni tests established statistically significant differences between NNS-low and NNS-high as well as between NNS-low and NS groups (all $p<0.001$). Data also show that variance was higher among low-proficiency learners. The described distribution of MC10 scores is visualised in Figure 1.

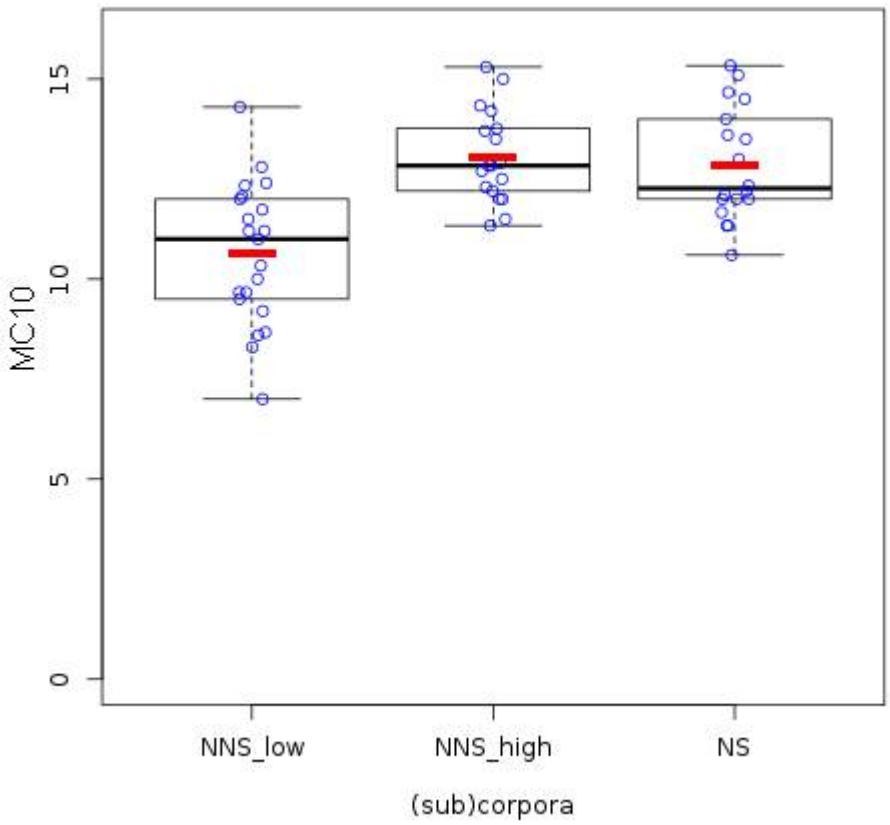

Figure 1. MC10 in learners-low, learners-high and natives.

The correlation between MC10 and C-test results is strong, as can be seen in Figure 2 (r = 0.759, $p < 0.001$, 95% CI [0.584, 0.867]).

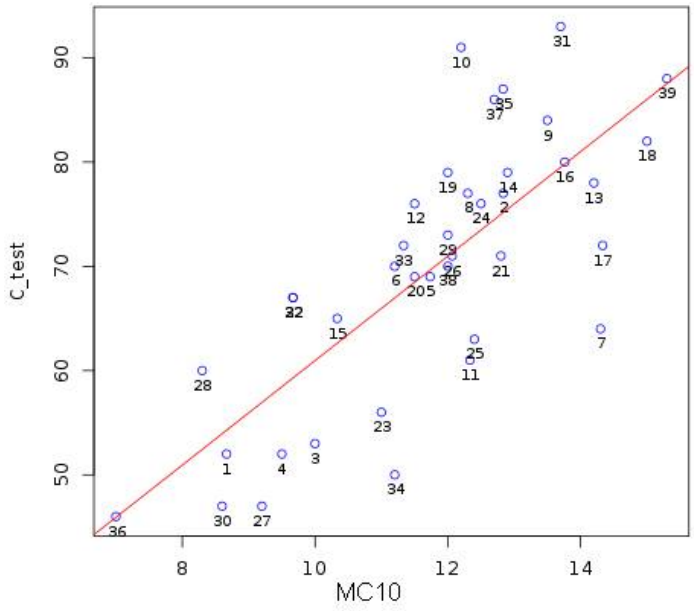

Figure 2. Correlation between MC10 and proficiency (C-test)

In the Italian L2 texts, MC10 was also positively correlated with lexical complexity, as measured by the standardised TTR (r = 0.441, p = 0.005, 95% CI [0.145, 0.664]) and syntactic complexity, as measured by sentence length (r = 0.416, p = 0.008, 95% CI [0.115, 0.646]), with medium effect sizes (r) in both cases. These correlations are visualised in Figure 3.

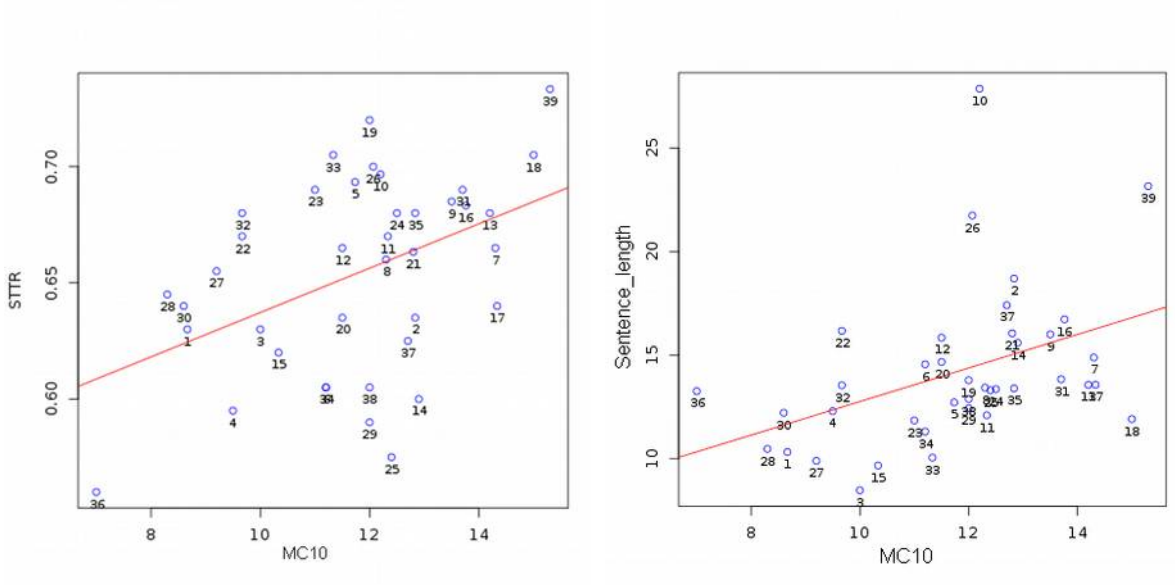

Figure 3. Correlation between MC10 and lexical (left panel) and syntactic (right panel) complexity.

The data also show a strong correlation between MC10 and MC5a both in the NNS (r = 0.762, p <0.001, 95% CI [0.588, 0.869]) and NS (r = 0.542, p = 0.020, 95% CI [0.1, 0.805]) texts. MC5a is an index designed for very short texts, as it is based on samples of just 5 exponents. In addition, MC5a correlates with C-test

scores (r = 0.726, p<0.001, 95% CI [0.533, 0.848]), standardised TTR (r = 0.483, p = 0.002, 95% CI [0.198, 0.693]) and mean sentence length (r = 0.441; p = 0.005, 95% CI [0.146, 0.664]) and distinguishes between NNS-low on the one hand and NNS-high and NS on the other hand (one-way ANOVA: F (2, 53) = 8.22, p < 0.001; effect size (ω) = 0.453; medium effect). We can therefore conclude that all results reported in this section based on MC10 hold true for MC5a as well.

Study 2

Data

The second case study is based on written argumentative essays in English produced by Italian university students, taken from the ICLE corpus (Granger et al., 2002). The learners' proficiency level ranged from B1 to C1 (Council of Europe, 2001). As a comparison group, similar texts produced by native speakers (both British and American) were extracted from the LOCNESS corpus (Granger. n. d.). The essays in both groups (NNS and NS) were written on a number of different topics of general interest (e.g. crime, money, feminism, Britain and Europe). Table 3 provides details about the two data sets.

Table 3
The English corpus

Corpus	No. of texts	Corpus size (words)	Mean text length (*SD*)	Mean verbs/text (*SD*)
NS	40	21,718	543 (103)	110.43 (30.29)
NNS	90	53,068	590 (191)	112.32 (32.48)

On average, the texts from the NNS group were 590 words in length (SD = 191); the NS subset consists of texts with the mean length of 543 (SD = 103). Although longer on average, the essays written by NNSs included approximately the same number of verb forms (M = 112.32, SD = 32) as the essays written by NSs (M = 110.43, SD = 11.49), showing again a difference in the mean verb density (19.04 vs 20.34 verbs/100 words in the NNS and NS corpora, respectively).

In addition to two MCI indexes (MC10 and MC5a), text measures of lexical (standardised TTR with 100-word samples based on lemma counts) and syntactic complexity (sentence length) were computed.

Results

Overall, the MC10 scores for verbs cluster towards the lower end of the scale, with the mean values of 5.89 and 5.86 for the NNS and NS writers respectively (Table 4). The small observed difference between the groups is not statistically significant (Welch t-test: t (60.78) = 0.17, p = 0.867). Interestingly, there is more variance in NS data[6], with the scores ranging between 3.38 and 9.33, than in the NNS data.

6 Although this finding might appear surprising because NS production is expected to be more homogeneous than NNS production, we have to realise that the comparison here is made between NS and proficient NNS; the latter have also very likely had more training on writing argumentative

Table 4
MC10 in native and non-native speakers, English

Corpus	*M* (*SD*)	*Range*
NS	5.86 (1.21)	3.38 – 9.33
NNS	5.89 (0.94)	4.14 – 7.92

The actual distribution of the MC10 values in the two groups can be seen in Figure 4.

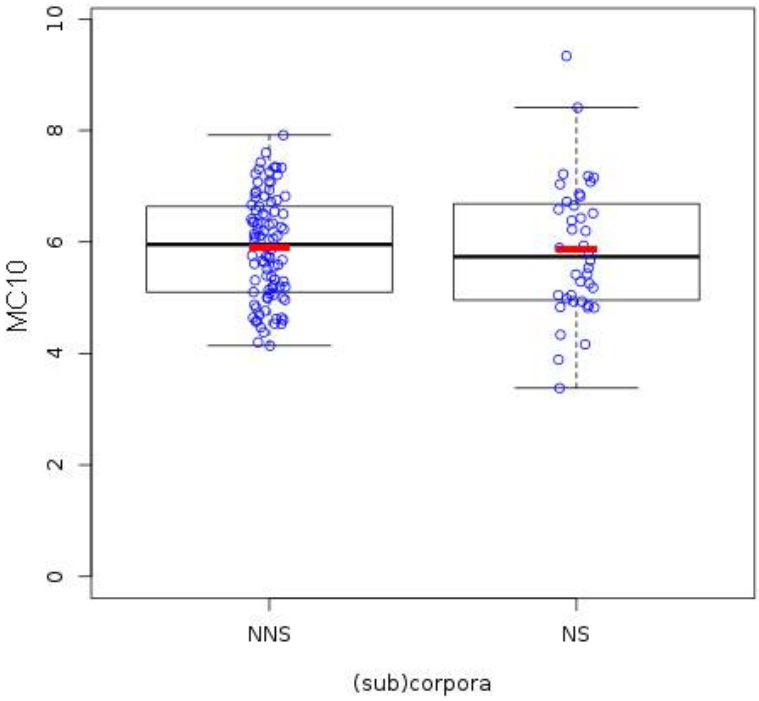

Figure 4. MC10 in learners and natives.

Looking at NNS data, MC10 does not significantly correlate with measures of lexical (standardised TTR with 100-word samples) and syntactic (sentence length) complexity. The correlations are as follows: MC10 and standardised TTR: r = 0.158, p = 0.138, 95% CI [-0.051, 0.353]; MC10 and sentence length: r = 0.112, p = 0.295, 95% CI [-0.098, 0.312] (Figure 5).

essays of this kind during the course of their studies. As a critical review of variation in NS and NNS corpora shows (Gablasova et al., under review), NS production is far from homogeneous and considerations such as corpus representativeness and sampling are crucial for interpretation of the results.

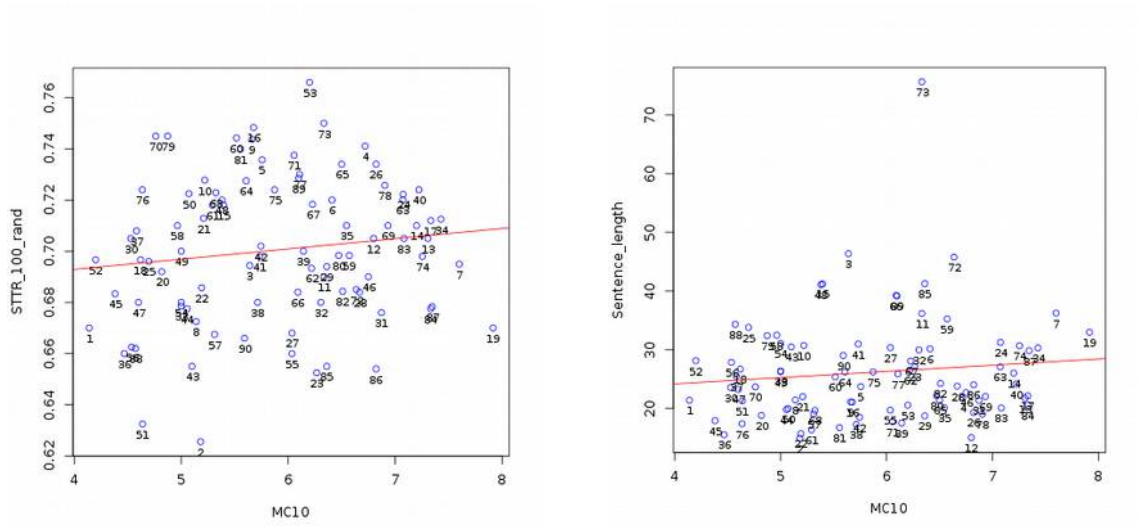

Figure 5. Correlation between MC10 and lexical (left panel) and syntactic (right panel) complexity.

There is a strong correlation between MC10 and MC5a ($r = 0.779$, $p < 0.001$, 95% CI [0.682, 0.849]), indicating the mutual replicability of the two measures.

Discussion

Our findings show that the MCI reflects both the complexity of the verbal inflectional system in a particular target language (Italian and English) as well as the realised complexity of L2 and L1 texts, which may be related to individual stylistic choices and L2 proficiency. As regards the first aspect, in case study 1 (target language Italian) MC10 scores were approximately twice as large as in case study 2 (target language English), with the mean score of 11.75 for L2 speakers and 12.85 for L1 speakers, vis-à-vis mean scores of 5.89 and 5.86 in case study 2. This finding can be directly related to the range of available exponences in Italian, a morphologically complex linguistic system, and English, a morphologically simple(r) language.

However, within the scope of a given target language, and keeping the variable 'text type' constant, the analysed texts showed a range of MCI values. In case study 1, learners' MC10 was related to other measures of performance such as C-test, standardised TTR and sentence length. In contrast, case study 2 did not find any relationship with the available performance measures for L2 learners (standardised TTR and sentence length). In addition, case study 1 found statistically significant differences between different groups of learners, in particular between those with lower and higher proficiency as measured by the C-test. The explanation of the fundamental difference between the results of case study 1 and 2 can be found in a) the range of learners' proficiency levels in the two studies and b) the difference in the development of inflectional competence in a morphologically complex (Italian) and a morphologically simple language (English).

With respect to the first point, in case study 1 learners' proficiency level ranged from A2 to B2, while the texts in case study 2 were written by more proficient learners (B1-C1). We can assume that, especially in a morphologically complex language such as Italian, learners at lower proficiency levels do not utilise the full

repertoire of exponences available in the language system, because some of them have not been fully acquired, i.e. are not available in their written production, which results in lower MCI scores for this sub-group. On the other hand, more proficient learners can choose from a larger set of exponences in the production of their texts and their performance is therefore comparable to that by native speakers. This is corroborated by the fact that no difference was found between high proficiency learners (C-test score over 71) and native speakers in case study 1, and between the whole group of advanced learners (B1-C1) and native speakers in case study 2. We can therefore hypothesise the existence of a threshold beyond which variation in morphological complexity is no longer related to learners' linguistic ability in their L2.

With respect to the second point, we can assume that this morphological threshold may be different for different target languages, as the evidence from the cross-linguistic acquisition of L1 morphology (e.g. Peters 1997) suggests. However, in order to prove this point, data from learners with a wider range of proficiency levels should be collected for English, too, which was not possible for the present project.

Finally, both studies showed a strong relationship between the two versions of the MCI: MC10 and MC5a. MC5a is a simplified operationalization of the construct, based on smaller exponence samples (5 exponences rather than 10). While both MC10 and Mc5a were available for the texts used in the case studies (because the mean number of verbs in the texts was larger than 20), this finding suggests a potential application of MCI to short and very short learner texts. Although operating on a different scale, MC5a yielded results comparable to MC10.

Further validation work should systematically investigate the effects of choosing different k values for sample size - such as MC5, MC10 or other values - on texts of varying length. It should also assess whether within-sample diversity (the 'a' component) can reliably substitute the current measure, based on within- and across-sample diversity indices. Further research is also needed on the correlation between MCI and other textual parameters, such as measures of lexical and syntactic complexity, and between MCI and learners' proficiency level, as measured through a variety of indicators or observed directly in longitudinal studies. This, however, should not be seen as a form of validation, as correlations between a measure and other aspects of linguistic performance or development should be treated as an empirical finding, possibly corroborating or falsifying theories postulating such relationships, and not as an indication of the measure's validity.

In this preliminary study we have presented MCI as a viable construct for assessing morphological complexity in L1 and L2 texts. The measure overcomes some of the shortcomings of previous approaches and, after some further validation work, may become a useful complement to existing indicators of the multi-dimensional construct of linguistic complexity.

References

Ackerman, F & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture, *Language,* 89, 429-464.

Anderson, S. R. (1992). *A–Morphous morphology*. Cambridge: Cambridge University Press.

Baechler, R., & Seiler, G. (2012). Simplification, complexification, and microvariation: Towards a quantification of inflectional complexity in closely related varieties. In A. Ralli, G. Booij, S. Scalise, & A. Karasimos (Eds.), *Morphology and the Architecture of Grammar. On-line Proceedings of the Eighth Mediterranean Morphology Meeting* (pp. 23-41).

Baerman M., Brown, D., & Corbett, G. G. (Eds.). (2015). *Understanding and measuring morphological complexity*. Oxford: Oxford University Press.

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/Philadelphia: Benjamins.

Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In J. Willis, & D. Willis (Eds.), *Challenge and Change in Language Teaching* (pp. 136-146). London: Heinemann.

Carstairs-McCarthy, A. (2010). *The evolution of morphology*. Oxford: Oxford University Press.

Council of Europe (2001). *Common European framework of references for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

DeKeyser, R. (2005). What makes second-language grammar difficult? A review of issues. *Language Learning* 55, Supplement 1, 1-25.

DeKeyser, R. (2016). Of moving targets and chameleons: Why the concept of difficulty is so hard to pin down. *Studies in Second Language Acquisition*, 38, 2.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in second Language acquisition*, *26*(1), 59-84.

Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167-192). Amsterdam: John Benjamins.

Foster, P. (1996). Doing the task better: How planning time influences students' performance. In J. Willis, & D. Willis (Eds.), *Challenge and Change in Language Teaching* (pp. 126-135). London: Heinemann.

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition, 18*(3): 299-324.

Gablasova, D., Brezina, V. & McEnery, T. (under review) Comparing speech in native and non-native corpora: Implications for SLA.

Granger, S. (n.d.). *Louvain Corpus of Native English Essays.* Université catholique de Louvain: Centre for English Corpus Linguistics.

Granger, S., Dagneaux, E., & Meunier, F. (2002). *International Corpus of Learner English*. Version 1.1. Université catholique de Louvain: Centre for English Corpus Linguistics.

Hockett, C. F. (1954). Two models of grammatical description. *Word*, *10*(2-3), 210-234.
Housen, A. & Simoens, H. (in press). Cognitive Perspectives on Difficulty and Complexity in L2 Acquisition. *Studies in Second Language Acquisition*, 38, 2.
Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs, 56*(2), 1–15.
Kruskal, J. (1999) [1983]. An overview of sequence comparison. In D. Sankoff, & J. Kruskal (Eds.), *Time warps, String edits and macromolecules: the theory and practice of sequence comparison* (pp. 1–44). Stanford, CA: CSLI Publications.
Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329-348.
Kuiken F., Vedder I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development. Intersections between SLA and language testing research* (pp. 81–99). Eurosla Monograph series 1. European Second Language Association.
Lardiere, D. (2006). Attainment and acquirability in second language acquisition. *Second Language Research,* 22, 3, 239–242.
Lightner, T. M. (1968). On the use of minor rules in Russian phonology. *Journal of linguistics*, *4*(01), 69-72.
Malvern D., Richards B., Chipere N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave MacMillan.
Page, S. E. (2010). *Diversity and complexity*. Princeton: Princeton University Press.
Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117–134.
Pallotti, G. (in press). "Una nuova misura della complessità linguistica: l'Indice di Complessità Morfologica (ICM)". *Rivista Italiana di Linguistica Applicata (RILA)*.
Pallotti, G. & Brezina, V. (2015). Defining and computing a text's morphological complexity. Paper presented at the workshop on Morphological complexity, Societas Linguistica Europea, Leiden, September 2015.
Peters, A. M. (1997). Language typology, prosody, and the acquisition of grammatical morphemes. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (pp.135-197), vol. 5. Hillsdale, NJ: Erlbaum.
Pienemann, M. (1998). *Language processing and second language development: Processability Theory*. Amsterdam: Benjamins.
Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
Shosted, R. K. (2006). Correlating complexity: A typological approach. *Linguistic Typology*, *10*(1), 1-40.
Slabakova, R. (2009). How is inflectional morphology learned? In L. Roberts, D. Véronique, A. Nilsson, M. Tellier (Eds.), *EuroSLA Yearbook*, 9 (pp. 56-75). Amsterdam: Benjamins.
Spencer, A. (2012). Identifying stems. *Word Structure*, *5*(1), 88-108.
Stump G., Finkel, R. A. (2013). *Morphological typology: from word to paradigm*. Cambridge: Cambridge University Press.
Taatgen, N. A., Dijkstra, M. (2003). Constraints on generalization: Why are past-

tense irregularization errors so rare? *Proceedings of the 25th annual conference of the cognitive science society* (pp. 1146-1151). Mahwah, NJ: Erlbaum.

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

Xanthos A., & Gillis, S. (2010). Quantifying the development of inflectional diversity. *First Language*, *30*(2), 175–198.

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., ... & Korecky-Kröll, K. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language*, *31*(4), 461-479.

Yuan, F., & Ellis, R. (2003). The effects of pre-task and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1), 1-27.