＜論文＞

The cloze test and overall, clausal, and subclausal indices of grammatical complexification in second language essays

NAKAMURA Daisuke

Abstract

In this study, we investigated the relationships between overall, clausal (coordination and subordination), and subclausal grammatical complexity in essays written by 54 adult Japanese EFL learners, with special reference to the developmental trajectories, from coordination through subordination to phrasal elaboration, that are identified in the literature. We also clarified whether using the cloze test with the fixed ratio deletion method and the exact-word scoring method to measure learners' proficiency could capture such trajectories. The results indicated that, while the cloze test did not correlate with any indices of grammatical complexity, overall complexity measures had positive correlations with subordination and phrasal elaboration indices, while subordination and phrasal elaboration had negative correlations. This suggests that, as a whole, the learners in this study were in transition to the later, advanced stage of the trajectories. Further analyses, employing general, but fine-grained, measures of phrasal complexity, revealed that the cloze test significantly correlated with these fine-grained measures, which indicates that the cloze test could capture learner variability within a particular developmental stage.

Keywords : *second language writing, grammatical complexity, the cloze test, the developmental trajectories*

Introduction

In general, second language learners lengthen their text at the interclausal level, such as a sentence or a terminable or T-unit (Hunt, 1965), as their proficiency increases. Ortega's (2003) research synthesis revealed that differences in length of about 4.5 words per sentence or 2 words per T-unit cause statistically significant differences between learners with different levels of proficiency. This indicates that there are positive correlations between proficiency and complexity in terms of interclausal length. The story is not so simple, however, since learners can complexify text in various ways : by adding subordinations or elaborating linguistic items at an interclausal or phrasal level, for instance. Thus, investigations into the development of grammatical complexity in writing have employed diverse indices of grammatical complexity, such as clauses per T-unit, to capture such diversity and have found general developmental trajectories of grammatical complexity that move from complexity by coordination through subordination to phrasal elaboration (see Norris & Ortega, 2009 and Wolfe-Quintero, Inagaki, & Kim, 1998 for reviews). This general trend can be studied either cross-sectionally (by measuring learners' proficiency using some independent measures and investigating their relations to the three dimensions of grammatical complexity) or longitudinally (by tracking increases and decreases in the dimensions across several discrete time points as in Nakamura, 2019), although the majority of studies are of the former type (Ortega, 2003).

Recent studies, however, especially some of those in the special issues of *Journal of Second Language Writing* (Connor-Linton & Polio, 2014 ; Vyatkina, 2015), have shown counterexamples to the general developmental trajectories described above. These mixed results for the developmental trajectories partly arise from ambiguities in their interpretation. Related to this, the relationships between overall complexity of L2 learners' writing and clausal and subclausal complexity are also vague, since some

日本女子体育大学（講師）

authors have included both types of measures but do not clarify their direct relationships (e.g. Lu & Ai, 2015). Moreover, different studies have employed different proficiency measures, such as course levels or rater judgements, each of which has its own limitations (Thomas, 1994). Surprisingly, none of the research in this literature has used the cloze test, which taps into integrative knowledge of lexico- grammatical, semantic, and discourse aspects by requiring students to fill in blanks in text (Fotos, 1991).

The aims of this study were therefore to appraise (a) the relationships between overall, clausal (coordination and subordination), and subclausal complexity, with special attention to the developmental trajectories, and (b) whether or not the cloze test can capture proficiency differences in such developmental trajectories in a cross-sectional study. We used an automated analyser, the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) (Kyle, 2016).

Developmental Trajectories, the Theoretical Underpinnings, and Counterexamples

Wolfe-Quintero et al. (1998) and Norris and Ortega (2009), from reviews of large numbers of past studies, identified the following general developmental course of grammatical complexity. Development of grammatical complexity 'begins' with the production of *fragments* (Ishikawa, 1995). After learners start to produce *main clauses* and reach the 'intermediate' levels, they expand their main clauses by adding *coordination* (Bardovi-Harlig, 1992) and *subordinations*. At 'advanced' levels, *phrasal elaborations* come to the fore (Norris & Ortega, 2009; Wolfe-Quintero et al., 1998).

Based on Halliday's Systemic Functional Grammar (e.g. Halliday & Martin, 1993), Norris and Ortega (2009) explained the transition from coordination through subordination to phrasal elaboration in terms of the development of the organisation of ideas. Initially, ideas are organised around *parataxis* (e.g. coordination); they are then expanded by *hypotaxis* (e.g. subordination). Further development induces learners to use *grammatical metaphors* (e.g. nominalisation). Norris and Ortega (2009) cited the following examples from Halliday and Martin (1993, pp.31-41): "(1) Darwin thought that species gradually became more complex. (2) Darwin's gradual rise to mounting complexity…" (pp.562-563). The contents expressed by the main and dependent clauses in (1) are in (2) condensed into one noun phrase by nominalisation.

Recent studies, however, have provided evidence against the general developmental trajectories. For example, Bluté and Housen (2014) analysed ESL learners' essays in an intensive English course (the Michigan State University Corpus, or the MSU corpus, Connor-Linton & Polio, 2014) and compared the grammatical complexity of the first essays with that of the last essays (written 4 months after the first ones) written by learners with proficiency at intermediate－advanced levels. The study found that, while coordination, phrasal complexity, and overall complexity increased, complexity by subordination remained the same.

Lu and Ai (2015) compared L2 learners with different first languages who were at intermediate and advanced levels with native speakers of English in large corpuses. While learners at the intermediate level (B2 according to the Common European Framework or CEF) produced less coordination, or fewer subordinations and phrasal elaborations, than the native speakers, those at the advanced (C1) level produced more coordination than the native speakers, at least on some measures. In other words, the advanced groups showed a greater amount of coordination than the less proficient groups.

These concurrent increases of coordination and phrasal elaborations (Bluté & Housen, 2014), or reverse relationships between complexity by coordination and learners' proficiency levels (Lu & Ai, 2015), seem to be counterexamples to the general developmental trajectories, from coordination through subordination to phrasal elaboration, found in earlier studies.

Ambiguities in the Developmental Trajectories

One of the explanatory factors for such apparent

discrepancies is that there is ambiguity in the developmental trajectories in relation to learners' proficiency in quantitative terms. It is very difficult to determine learners' proficiency levels in terms of discrete points : if they score 500 points on Test of English as a Foreign Language (TOEFL), can we identify their proficiency as being at an intermediate level? In other words, each dimension of grammatical complexity cannot correspond to a particular point in L2 proficiency measured by a specific measure. Therefore, the general developmental trajectories can be investigated only in relative terms or by somewhat artificially dividing learners into relatively different proficiency groups. At least the following developmental scenarios can then be inferred for cross-sectional studies : (1) as proficiency increases, the number and/or the mean length of the relevant variable (e.g. subordinations) also increases. This should be verified by either (a) positive correlations between proficiency and relevant complexity measures or (b) positive main effects of the former on the latter (this can, for example, be conducted by ANOVAs, with learners divided into different proficiency groups by mean or median split methods). Nevertheless, one ambiguity arises here : is it presupposed that, as proficiency increases, so does overall complexity? In other words, where can overall complexity fit within the developmental trajectories from coordination through subordination to phrasal elaboration?

Another ambiguity is the relationship between the dependent variables : (2a) there should be negative correlations among them ; (2b) there are no such correlations among them. The former prediction captures the possibility that learners complexify text along one particular dimension at the expense of the others, while the latter prediction is that there are no such competitive relationships. Wolfe-Quintero et al. (1998, p.73) and Bluté and Housen (2014) took the former interpretation, although Wolfe-Quintero et al. (1998) admitted nonlinearity in such a way that each dimension would exhibit an inverse U or Ω shaped development with initial increases followed by later decreases. Nevertheless, the second interpretation is also logically possible. For example, compare *Sam often becomes sick in bed when he talks to his boss* with *Because Sam has generalised anxiety disorder, he often becomes sick in bed when he talks to his boss*. The first sentence consists of two clauses per T-unit, and each clause contains six words. By contrast, in the second sentence, the average number of words per clause remains the same (six), while the number of clauses increases (to three). That is, as the length of the T-unit increases, the number of dependent clauses per T-unit increases while the length of the clauses is unchanged. In other words, complexity by subordination increases, but complexity by phrasal elaboration remains the same.

Since the general developmental trajectories were derived from the results of heterogeneous studies where different indices for grammatical complexity and proficiency measures were employed in both predictions, key aspects to consider, and thus other possible explanatory factors for the contradictory results, are the diversities in grammatical complexity and learners' proficiency measures.

Measures of Grammatical Complexity

A huge number of indices of grammatical complexity have been employed in the writing literature, using the Complexity－Accuracy－Fluency (CAF) framework as a basis (e.g. Housen & Kuiken, 2009 ; Housen, Kuiken & Vedder, 2012) to track the trajectories from coordination through subordination to phrasal elaboration. Recent development of automated analysers such as the Coh-Metric (e.g. McNamara, Crossley, & McCarthy, 2010), the Syntactic Complexity Analyser (SCA) (Lu, 2010), and the TAASSC (Kyle, 2016) render analyses much easier and more reliable. Typical measures used in the literature are length-based measures such as Mean Length of Sentence (MLS) and T-unit (MLT) and ratio-based measures, especially of subordinations, such as clauses per T-unit (C/T) (see Bluté & Housen, 2012 ; Norris & Ortega, 2009 ; Ortega, 2003 ; Wolfe-Quintero et al., 1998). Wolfe-Quintero et al. (1998) identified C/T and dependent clauses

per clause (DC/C) or per T-unit (DC/T) as the best developmental indices for grammatical complexity[1]. Norris and Ortega (2009) suggested the following different types of metrics, each corresponding to different dimensions of the developmental trajectories:

1. "complexity via subordination…measured by any metric with clause (subordinate or dependent clause) in the numerator"
2. "overall or general complexity…measured by any length-based metric with a potentially multiple-clause unit of production in the denominator"
3. "subclausal complexity via phrasal elaboration …measured by mean length of clause" (pp.561-562).

Norris and Ortega argued that measures of overall grammatical complexity, such as MLT, should be employed since they may be able to capture variability even if none of the specific measures can do so. In addition to these measures, they suggested a further two types of metrics: one for complexity by coordination (e.g. coordination index, Bardovi-Harlig, 1992), if the learners involved have the proficiency of beginners, and another for "complexity defined as the variety, sophistication, and acquisition timing of forms produced" (p.562) (e.g. index of productive syntax, Scarborough, 1990).

Different researchers, however, have measured grammatical complexity using (slightly) different types of complexity indices, which leads to their proliferation in the literature. Obviously, this introduces several problems.

First, Norris and Ortega (2009) claimed that if some indices (e.g. C/T and DC/C) showed huge overlaps in their variance, these indices were redundant or worse from a psychometric point of view (since redundancy causes multicollinearity for regression analyses). This can be resolved, however, by choosing one of them for multivariate analyses after the correlation analyses, as Norris and Ortega (2009) suggested, since whether or not multiple measures for a particular dimension (e.g. complexity by subordination) exhibit redundancy can only be discovered after statistical analyses.

Second, researchers have differed in the treatment of non-finite clauses. Some (e.g. Bluté & Housen, 2014) have included both finite and non-finite verb phrases as instances of clauses, while others (e.g. Lu & Ai, 2015; Nakamura, 2019; Wolfe-Quintero et al., 1998) have treated only finite clauses as clauses. For example, *I stopped to smoke* can be analysed either as two clauses (one main clause plus one dependent clause) or as one clause (one main clause). The former increases the number of dependent clauses while keeping the number of independent clauses the same. This discrepancy in the treatment of non-finite verb phrases might create some apparent counterevidence for the general developmental trajectories (Vyatkina, 2012; Wolfe-Quintero et al., 1998). In this paper, following Wolfe-Quintero et al.'s (1998) suggestions, only finite clauses containing overt subject(s) and finite verbs were counted as clauses. This is consistent with Wolfe-Quintero et al. (1998) and was chosen because we analysed the data in the present study using the SCA components (Lu, 2000) of the TAASSC (Kyle, 2016), which includes indices identified by Wolfe-Quintero et al. (1998) as the best complexity measures.

Third, as Norris and Ortega (2009) recommended, many recent studies have included some length-based indices of overall complexity but have not clarified their relationships to complexity measures along the three dimensions of the developmental trajectories. For example, Bluté and Housen (2014) measured both types of indices across different time points, and, indeed, both increased across time. Nevertheless, we cannot conclude that shared variability underlies these increases. In addition to this, it is unclear whether, as proficiency increases, so does overall complexity. Therefore, the present paper evaluates the relationships between overall complexity and the three complexity dimensions along the developmental trajectories.

Learners' Proficiency

In order to track the developmental trajectories

of grammatical complexity described above, we needed some measures for learners' proficiency. Second language acquisition (SLA) researchers have tried to capture learners' proficiency by employing five types of measures : impressionistic judgment, institutional status, in-house assessment, and standardised tests (Thomas, 1994, 2006) or (holistic) judgements by raters (Ortega, 2003). As Ortega's (2003) research synthesis demonstrated, the majority (19 out of the 21 studies on college students' writing) of the writing research studies employed course or university level, or rater judgement, as independent measures of learners' proficiency[2] ; meanwhile, other, more fine-grained or objective measures, such as TOEFL or cloze tests, have rarely been employed. Ortega (2003) showed that only two studies used standardised tests, and these tests were different. Moreover, none of the researchers employed a cloze test, which is a type of in-house assessment (Thomas, 1994), with relation to analyses of the grammatical complexity in learners' writing. Nevertheless, this considerable reliance on coarse-grained (course level) or subjective measurements (rating) poses serious problems.

First, since different universities set different standards for placing learners at different levels or promoting them to higher classes (Thomas, 1994), inferring learners' proficiency based solely on course levels (e.g. Vyatkina, 2012) or in conjunction with classroom teachers' evaluations (e.g. Kormos, 2011) is a rather coarse-grained method. For example, in Neary-Sundquist's (2014) study, learners' proficiency was determined to be at novice—intermediate level in the first semester, although the learners were placed into different classes in later semesters on the basis of the results of a placement test (see also Vyatkina, 2012). In Ortega's (2003) metanalysis, 11 of the 21 writing studies were of this type.

Second, another popular method of measuring learners' (writing) proficiency is by rater judgements, and 8 of the 21 studies in Ortega's review are of this type. In this methodology, experienced raters judge the quality of learners' writing according to certain rubrics (Brown, 2012), and studies investigate which grammatical complexity measures can explain the variance in such rating scores and how effectively they do so. For example, in Lu and Ai (2015), students' proficiency levels were gauged by raters according to the CEF writing descriptors'[3]. Such subjective ratings, however, are not always consistent with learner's proficiency when measured using objective methods. This can be seen in the mixed results for the consistency between rater judgement and grammatical complexity in writing (or speech) measured by objective indices : while some studies (Magnan, 1988 ; McNamara, 1990) showed affinities, others (Bluté & Housen, 2014 ; Douglas, 1994 ; Iwashita, Brown, McNamara, & O'Hagan, 2008 ; Kuiken & Vedder, 2014 ; Crossley & McNamara, 2014) showed discrepancies between the two methods. Such discrepancies could be partly due to (a) multiple types of raters, (some emphasising grammar while others give more importance to content, for example) (Eckes, 2008 ; Schaefer, 2008) ; (b) differences in the language backgrounds of raters (Johnson & Lim, 2009) ; (c) raters' lack of adherence to directions (raters might judge essays as a whole even when they are asked to rate each aspect of writing separately, using analytic rubrics, for example) (Davis & Kondo-Brown, 2012).

How, then, can we avoid the potential pitfalls in measuring learners' proficiency independently from indices for grammatical complexity? One way is to use standardised proficiency tests such as TOEFL (e.g. Bluté & Housen, 2014). One of the merits of these tests is their comparability with other studies because they are open to public use or function "as a recognizable benchmark, enhancing generalizability of the research results…. [T]he content of standardized tests is available for public scrutiny, and their validity is subject to ongoing investigation" (Thomas, 1994, p.324). In Bluté and Housen's (2014) analyses of the MSU corpus, the participants' proficiency was measured by TOEFL, and the majority of them scored below 500 points. Such expensive and time-consuming tests are not options for many researchers (Tremblay, 2011), however,

and such additional burdens in terms of both time and cost cannot be placed on classroom SLA research in particular because they could skew the contents and/or schedule of the class. Another solution is the use of a cloze test (Brown, 1980). In contrast to standardised tests, cloze tests can be implemented in a short time without excessively disturbing the class administration ; some of the weaknesses of standardised tests can thus be overcome. Because in-house assessment in general differs in form and content, however, it is not possible to generalise the results into the population in other studies (Thomas, 1994).

Since the cloze test requires various types of linguistic knowledge (vocabulary, grammar, semantic, and discourse knowledge (Hanania & Shikhani, 1986)), it is an integrative test. Its (methodological) validity and reliability have been extensively investigated (e.g. Bachman, 1985 ; Brown, 1980), and it is thus still widely used today in the SLA literature (e.g. Gaillard & Tremblay, 2016 ; see also Brown (2012) and Tremblay (2011) for recent reviews). Surprisingly, the cloze test has been neglected with relation to grammatical complexity in writing. Tremblay (2011) found that 16 out of 53 studies published in major SLA-related journals employed either cloze tests or C-tests as independent tests of learners' proficiency. For example, Fotos (1991) compared the effects on L2 learners' essay writing of TOEFL with those of a cloze test created using a fixed ratio deletion method. In the fixed ratio deletion method, every nth word is deleted and the results are marked by the exact-word scoring method, where for each blank, they have to think of the same word as the one in the original text. She found that TOEFL and the cloze test together explained 63% of the variance in the essay scores, while TOEFL alone explained 48% of the variance. Furthermore, they had a partial correlation at $r =$ .45 level, suggesting that the cloze test had "construct validity as measurements of English language proficiency" (p.331). Similarly, Hanania and Shikhani (1986) demonstrated that the cloze test, in conjunction with a standardised test, could explain the small but additional variance in composition scores that could not be explained by the standardised test alone. Furthermore, there were significant partial correlations between the cloze test and composition when the effects of the standardised test were 'partialled out', suggesting that the cloze test could gauge aspects of learners' (writing) proficiency that the standardised test could not capture. One limitation, which we tried to overcome in the study described in the present paper, is that in both Fotos's (1991) and Hanania and Shikhani's (1986) studies, the quality of the essays and, thus, the learners' writing proficiency were judged by raters. As described above, the reliability of rater judgement is controversial and cannot be taken at face value.

In summary, cloze tests have not been investigated in relation to more micro measures of grammatical complexity ; they have been investigated only in relation to TOEFL or other standardised test scores and rating scores on essays (Fotos, 1991 ; Hanania & Shikhani, 1986), class levels, and self-rated listening and speaking proficiency (Gaillard & Tremblay, 2016). As is indicated by the mixed results on the relationship between rating and more objective measures of grammatical complexity in particular, and other dimensions of L2 development in general, the reliability of rater judgement remains controversial at best. Furthermore, considering the fact that the cloze test is a more objective and integrative test and, thus, a measure for learners' general proficiency and not specifically their writing skills, it is worth investigating whether or not the cloze test can capture the general developmental trajectories indexed by the more objective indices for grammatical complexity. The second aim of this study was therefore to investigate this possibility.

Research Questions

1. What is the relationship between overall, clausal (coordination & subordination), and phrasal complexity?
2. Can the cloze test capture the developmental trajectories from coordination through subordina-

tion to phrasal elaboration?

Method

Participants

54 Japanese EFL learners (47 male, 7 female; mean age＝20.79) served as the participants. The majority of these were third-year students, and all of them were majoring in science and technology at a private university in Japan. They had been learning English for at least six years before entering university and had participated in non-mandatory classes on essay writing, where the emphasis was on structural knowledge (e.g. how to organise a thesis statement) and the content of English essays (e.g. what kinds of supporting evidence should be provided for the thesis) but not on grammar. No placement tests were administered; rather, students simply had the option of taking these classes if they so wished. Aims and procedures of the present study were explained in the first class and they agreed to participate in the study. A cloze test and a final essay writing task were set at, respectively, the beginning and the end of the semester. They were told that both were irrelevant for their grades.

Instruments

Cloze test. A passage for the cloze test in the present study ("*Modern Writing : Be Brief!*") was taken from a reading textbook for students at novice－intermediate level (Takeuchi, Yabukoshi, Sumi, & Ueki, 2012). We deleted every sixth word (the fixed ratio deletion method), and we gave one point only for an exact answer (the exact-word scoring method), as in Fotos (1991)[4]. The total number of blanks was 63. We gave students 15 minutes to complete the test in the first class.

Writing task. In the final class, students were asked to write an essay on a PC on 'an ideal writing class'. Use of the spell checker in Microsoft Word was permitted. Although no specific time limits were set, most students completed the essay within an hour.

Grammatical Complexity Measures

All indices of grammatical complexity were taken from, and their analyses were run by, the L2 SCA component (Lu, 2010) of the TAASSC (version 1.1) developed by Kyle (2016; see also Nakamura (2019) for how to calculate these):

1. Overall complexity:
   a. Mean Length of Sentence (MLS)
   b. Mean Length of T-unit (MLT)
2. Complexity by coordination:
   a. T-units per Sentence (T/S)
3. Complexity by subordination:
   a. Clauses per T-unit (C/T)
   b. Dependent Clauses per T-unit (DC/T)
   c. Dependent Clauses per Clause (DC/C)
4. Complexity by phrasal elaboration:
   a. Mean Length of Clause (MLC)

Results

We ran Shapiro－Wilk tests over the data to check the normality of the distributions. We removed by pair-wise deletions outliers, defined as those who scored more or less than 3 SDs from the means, in the following analyses. We repeated the deletions until no outliers were identified. Table 1 presents the descriptive statistics on the independent and dependent measures resulting from these data (we have provided Tables 2-9 and Appendix A as Online Supplementary Materials, available from https://www.researchgate.net/publication/337020806_Online_Materials_Nakamura_2020).

First, to confirm that the learners significantly differed in their L2 proficiency, we created more and less proficient learner groups using the mean-split method and analysed their differences in the cloze test using t and Mann-Whitney tests. As Table 1 illustrates, the more proficient learners were able to fill approximately 10 blanks more than their less proficient counterparts; they therefore clearly differed in their proficiency as measured by the cloze test ($t(52)=9.46$, $p<.01$). Table 1 also shows, however, that the more proficient learners did not outperform their less proficient counterparts

Table 1 *Means, SDs, and Statistical Differences*

Index	Overall	High	Low	*T* or *U*
Cloze test	15.54 (6.60) (N＝54)	20.74 (4.80) (N＝27)	10.33 (3.10) (N＝27)	*T* (52)＝9.46 *p*＜.01
MLS	17.46 (3.30) (N＝54)	17.54 (3.24) (N＝27)	17.38 (3.43) (N＝27)	*T* (52)＝.176 n.s.
MLT	14.88 (2.48) (N＝50)	14.94 (2.65) (N＝24)	14.81 (2.38) (N＝26)	*U*＝298.5 *Z*＝－.26, n.s.
T/S	1.13 (.11) (N＝52)	1.12 (.14) (N＝26)	1.14 (.12) (N＝26)	*U*＝320.5 *Z*＝－.32, n.s.
C/T	1.51 (.22) (N＝52)	1.50 (.24) (N＝26)	1.53 (.20) (N＝26)	*U*＝300.5 *Z*＝－.69, n.s.
DC/T	.52 (.23) (N＝54)	.53 (.26) (N＝27)	.52 (.21) (N＝27)	*U*＝360.5 *Z*＝－.69, n.s.
DC/C	.33 (.01) (N＝54)	.33 (.11) (N＝27)	.32 (.09) (N＝27)	*T* (52)＝.261 n.s.
MLC	9.98 (1.59) (N＝53)	10.17 (1.54) (N＝26)	9.81 (1.64) (N＝27)	*T* (51)＝.832 n.s.

Note. Missing cells and outliers were removed from the analyses so that the number in each cell differs. MLS＝Mean Length of Sentence ; MLT＝Mean Length of T-unit ; T/S＝T-units per Sentence ; C/T＝Clauses per T-unit ; DC/T＝Dependent Clauses per T-unit ; DC/C＝Dependent Clauses per Clause ; MLC＝Mean Length of Clause.

on any indices of grammatical complexity (all *p*s＞.05). The correlation analyses, presented in Table 2, confirmed this ; none of the complexity measures exhibited significant correlations with the scores on the cloze test.

Second, global complexity, measured by MLS and MLT, showed significant positive correlations with the majority of the complexity by subordination and complexity by phrasal elaboration measures : MLS correlated with C/T (*rho*＝.595), DC/T (*rho*＝.325), DC/C (*r*＝.540), and MLC (*r*＝.434) ; MLT correlated with C/T (*rho*＝.369), DC/T (*rho*＝.320), and MLC (*rho*＝.519).

Third, the subordination indexes positively correlated with each other : C/T with DC/T (*rho*＝.855), CT with DC/C (*rho*＝.753), and DC/T with DC/C (*r*＝.978). Such high correlations (*r*＜.7) might indicate that these measures, in fact, measured the same thing. Finally, the phrasal elaboration index showed a negative correlation with all the subordination measures : MLC with C/T (*rho*＝－.374), DC/T (*rho*＝－.375), and DC/C (*rho*＝－.375).

To assess the contributions of complexity by subordination and complexity by phrasal elaboration to the overall complexity, we conducted step-wise regression analyses. For the regression analyses on MLS as a dependent variable, we entered all subordination and phrasal elaboration measures as independent variables. We excluded DC/T and DC/C from the analyses, either because of multicollinearity (VIF＞10 and tolerance statistics＜.2, see Field (2013)) or because their contributions to the model were not significant. Table 3 provides the final model. As can be seen from Table 3, MLC alone explained 17% of the variance in MLS, and adding C/T increased the amount of variance explained (by an additional 35%), leaving MLC and C/T together significantly accounting for 52% of the variance in MLS. Furthermore, Table 4 illustrates that MLC and C/T had similar impacts on MLS

(compare the βs).

Similar pictures emerged from the regression analyses on MLT. Since the scores for MLT did not show normal distributions, we log-transformed them; we then entered all subordination and phrasal elaboration measures into the analyses. Again, we excluded DC/T and DC/C from the regression analyses because of multicollinearity. Table 5 illustrates that MLC accounted for 22% of the variance, and adding C/T increased the amount of variance explained (an additional 44%), leaving MLC and C/T together accounting for 66.6% of the variance in MLT at a significant level. As can be seen in Table 6, MLC and C/T had similar impacts on MLT, as in the case of MLS.

Interim Discussion

The present study found that coordination, indexed by the number of T-units per sentence (T/S), did not increase with the overall complexity (MLS, MLT). This can be seen in the lack of a significant correlation between them. Instead, L2 learners lengthened the text either by increasing the number of subordinations (C/T, DC/C, DC/T) or by lengthening a clause (MLC), but not by doing both. Surprisingly, however, none of the variables exhibited significant correlations with the cloze test. This is attested to by the lack of significant correlations between them and by the lack of significant effects for the cloze test. Does this mean that the cloze test does not have the capacity to discriminate between the grammatical complexity of learners? This is not the case, since there were significant differences between the scores of the more proficient learners and the scores of the less proficient learners, grouped by the mean-split methods, such that the cloze test itself was able to capture proficiency differences. Before concluding, we must consider the possibility that the cloze test can measure differences in the subcomponents of a dimension: phrasal elaboration. Recent investigations by Crossley and McNamara (2012) and McNamara et al. (2010), for instance, employed the number of words before the main verb as one of the grammatical complexity measures (see also Vyatkina (2012)).

Fine-Grained Analyses

The TAASSC (Kyle, 2016) provides a huge number of fine-grained general indices (32 for clausal, and 132 for phrasal, complexity)[5]. As in the main analyses, we first identified and removed from the subsequent analyses all outliers, defined as those scoring beyond the mean +/− 3 SDs, and checked the normality of the distributions using Shapiro−Wilk tests. Only the following 3 phrasal complexity indices (out of the 132) showed significant correlations with the cloze tests and are thus reported here[6]:

1. Dependents[7] per direct object (DDO)
2. Dependents per direct object (no pronouns)[8] (DDONP)
3. Dependents per nominal subject (no pronouns) (DNSNP)

Generally, a dependent here means a modifier of a nominal.

Table 7 presents the descriptive statistics, and the results of the correlational analyses are provided in Table 8 (see also Appendix A, for descriptive statistics for the overall, high, and low proficiency groups on these measures).

As Table 7 demonstrates, there are significant medium sized correlations between the cloze test and the nominal complexity indices: the cloze test correlated with DDO ($r = .282$, $p < .05$), DDONP ($r = .331$, $p < .05$), and DNSNP ($rho = .289$, $p < .05$). The high correlations between DDO and DDONP ($r < .9$) signify that both, in fact, measured the same aspects of nominal complexity. As in the main analyses, we log-transformed DNSNP because it did not exhibit a normal distribution. We then conducted simple regression analyses over DDONP and DNSNP. The results indicated that the cloze test accounted for approximately 9% and 14%, respectively, of the variance of DNSNP and DDONP at a statistically significant level ($p < .05$ and $p < .01$) (see Tables 8 and 9). This suggests that more proficient learners added more modifying elements to the lexical subject and direct object nouns than learners who were

less proficient, and that the cloze test could capture these slight but fine-grained differences in phrasal elaboration.

Discussion

RQ 1 : Overall Complexity and the Developmental Trajectories

In the present study, we investigated the relationship between overall, clausal (coordination and subordination), and phrasal complexity with regards to the developmental trajectories, from coordination through subordination to phrasal elaboration, in essays written by L2 learners. First, learners either increased the amount of subordination or lengthened a clause when they complexified text in terms of its length ; this was demonstrated by significant correlations of the scores for MLS and MLT with those for the subordination measures (C/T, DC/C, and DC/T) and with those for the phrasal elaboration measure (MLC). Nevertheless, these two options could not coincide but were in competition, as the negative correlations between the subordination and the phrasal complexity indices indicated.

By contrast, coordination had no positive or negative correlations with the other measures. It should be noted that there are two interpretations of the general developmental trajectories : the three variables, coordination, subordination, and phrasal elaboration, could have negative correlations or they could have no correlations. The present results showing negative correlations between subordination and phrasal elaboration and no correlation for coordination partially confirm both interpretations. One possibility is that participants in the present study had already passed the novice learner stage, where they would have complexified text in terms of coordination, and were on a transition path from the intermediate to the advanced stages, as demonstrated in the results of the regression analyses where C/T and MLC had a similar magnitude of effects. Thus, a tentative general picture is as follows : when learners are at a particular developmental stage, they complexify text in only one relevant dimension (e.g. coordination for the novice learner stage) ; however, negative correlations between the relevant dimensions (coordination and subordination for the transition from the novice to the intermediate proficiency level) emerge when they are in transition to the next stage (similar results were found in Nakamura's study (2019)).

Only Bluté and Housen (2014) seem to give a reliable counterexample. In their study, as overall complexity (measured by MLS and MLT) increased, so did coordination and phrasal elaboration, during a semester-long intensive English programme. Although the contradiction cannot be immediately resolved, methodological differences should be noted. First, the definition of a clause differs : Bluté and Housen (2014) included both finite and non-finite clauses, while non-finite clauses were excluded in the present study (see the brief discussion in Vyatkina (2012)). Second, and most importantly, their study was a longitudinal study, while the present study was cross-sectional in nature. Third, as described above, Bluté and Housen (2014) did not conduct correlational analyses among overall, clausal, and subclausal complexity indices ; there is, therefore, a possibility that coordination and phrasal elaboration did not share variance. Since the majority of studies have not investigated the relationship between general, clausal, and subclausal complexity (e.g. Lu & Ai, 2015)[9], one direction for future research arises from the necessity of including general complexity measures as well as indices for the three dimensions, as Norris and Ortega (2009) pointed out, and of investigating the direct relationship between them.

RQ 2 : The Cloze Test and the Developmental Trajectories

We also investigated in this study whether or not proficiency as measured by the cloze test captures the developmental trajectories of grammatical complexity from coordination through subordination to phrasal elaboration. The cloze test was not able to capture proficiency differences in the developmental trajectories. The lack of any significant

correlations between the cloze test and any of the indices of grammatical complexity seemed to show this (Table 2), and the lack of any correlations between the cloze test and overall complexity is consistent with the findings presented in Ortega (2003) : differences in length of approximately 4.5 words per sentence or 2 words per T-unit correspond, with statistical significance, to differences in L2 proficiency. Both more proficient learners and less proficient learners produced a sentence of 17 words and a T-unit of 15 words. Does the lack of any statistically significant relationships between them mean that the cloze test is an unreliable measure of proficiency?

Considering that the reliability and validity of the cloze test have been disputed (Brown, 2013 ; Tremblay, 2011) and tested (Bachman, 1985 ; Brown, 1980 ; Gaillard & Tremblay, 2016), it is not the case that the cloze test itself cannot capture learners' proficiency. Indeed, high and low proficiency groups in the present study, created by the mean-split method on the cloze test scores, exhibited statistical significance (Table 1). One possibility, then, is that all learners in the present study were at the intermediate—advanced level, as revealed by the negative correlation between complexity by subordination and that by lexical sophistication (Table 2) and by the indication that MLC and C/T had similar impacts (compare the βs in Tables 4 and 6) despite the fact that we always entered MLC into the models first in the stepwise regression analyses (Tables 3 and 5). Because of the lower variance in the test scores, the cloze test might not capture individual differences across the broader developmental trajectories. In contrast, the cloze test could capture more micro-level differences, that is, variance within one developmental stage of the trajectories, the advanced stage, in such a way that learners complexified text along a particular dimension of lexical complexity, that is, noun phrases. This is revealed by the positive correlations between the cloze test and the fine-grained, general measures of grammatical complexity (the average number of dependents per subject and direct object noun). Similar findings were obtained by Crossley and McNamara (2014), who found that learners increased the average number of modifiers per noun phrase from the beginning to the end of the academic term. MLC cannot capture this proficiency difference, because complexity at all phrasal levels, such as indirect object noun phrases, is entered into the analyses, and any differences in a particular 'type' are 'partialled out' during calculations.

Conclusion

In the study described in the present paper, we investigated the relationships between overall, clausal, and subclausal complexity, specifically paying attention to the developmental trajectories from coordination through subordination to phrasal elaboration identified in the literature. We also studied whether or not differences in L2 proficiency measured by a cloze test could capture such developmental trajectories. While the cloze test did not correlate with any of the general 'coarse-grained' measures for these three different phases of the trajectories, either the amount of subordination increased or the subclauses were lengthened as learners complexified the text by lengthening sentences or T-units. Therefore, the developmental trajectories themselves were observed. In addition, the cloze test correlated with general but 'fine-grained' measures of phrasal elaboration, suggesting that it can capture differences at a more micro level, at least within a particular phase of development.

Footnotes

1 Wolfe-Quintero et al. (1998) also suggested verb phrases per clause for potential global measures, as well as 22 potential indices for fine-grained analyses of grammatical complexity.

2 This is also a general tendency in SLA research outside writing (see Thomas (1994) and Tremblay (2011)).

3 As Lu and Ai (2015) admitted, in their study learners' proficiency and nationalities were confounded, which introduced another problem : their effects could not be

separated.

4 One exception is that replacement of the restrictive relative clause *which* for *that* was permitted.

5 As noted in Nakamura (2019), reasons for selecting these indices are that nominalisation is "the hallmark of advanced L2 capacity" (Norris & Ortega, 2009).

6 Additionally, two indices (Undefined Dependents per Clause and Dependents per Clause (standard deviation)), which, it is argued, measure clausal complexity and variety, respectively, showed significant correlations with the cloze test ($rho = .338$, $p < .01$, and $rho = .329$, $p < .05$, respectively). Since Kyle (2016) counted both finite and non-finite clauses as clauses, however, these were removed from the discussions in order to be consistent with the rest of the analysis.

7 Dependency relations among words were parsed by the Stanford Neural Network Dependency Parser. For example, in a subject noun phrase, "*the linguist…, the* is parsed as a dependent and *linguist* as a governor of the dependency relation 'determiner'" (see Kyle (2016), Chapter 2).

8 Kyle (2016) explained the treatment of pronouns as follows:

> Noun phrases in English can consist of pronouns, and except in very rare cases, pronouns do not take direct dependents (relative clauses being an exception). Due to the potential for pronouns as phrases to skew counts of dependents, TAASSC includes two versions of each index. (p.57)

9 Some studies have investigated these relationships. For example, Vyatkina (2012) found positive correlations between MLS and subordination, and negative correlations between MLS and coordination. In her study on the development of grammatical complexity in German, however, both subordination and coordination measures included both inter- and intra-clausal connectors.

References

Bachman, L.F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19* (3), 535-555.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*, 390-395.

Brown, J.D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal, 64* (3), 311-317.

Brown, J.D. (2012). Developing rubrics for language assessment. In J.D. Brown (Ed.), *Developing, using, and analyzing rubrics in assessment with case studies in Asian and Pacific languages* (pp.13-32). Honolulu, HI: University of Hawai'i Press.

Brown, J.D. (2013). My twenty-five years of cloze testing research: So what? *International Journal of Language Studies, 7* (1), 1-32.

Bluté, B. & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp.21-46). Amsterdam, The Netherlands: John Benjamins.

Bluté, B. & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42-65.

Connor-Linton, J. & Polio, C. (2014). Comparing perspectives on L2 writing; Multiple analyses of a common corpus. *Journal of Second Language Writing, 26*, 1-9.

Crossley, S.A. & McNamara, D.S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35* (2), 115-135.

Crossley, S.A. & McNamara, D.S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66-79.

Davis, L. & Kondo-Brown, K. (2012). Assessing student language performance: Types and uses of rubrics. In J.D. Brown (Ed.), *Developing, using, and analyzing rubrics in assessment with case studies in Asian and Pacific languages* (pp.169-182). Honolulu, HI: University of Hawai'i Press.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11* (2), 125-144.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25* (2), 155-185.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Thousand Oak, CA: Sage Publications.

Fotos, S.S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning, 41* (3), 313-336.

Gaillard, S. & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning, 66* (2), 419-447.

Halliday, M.A.K. & Martin, J.R. (1993). *Writing science: Literacy and discursive power*. London, UK: Falmer Press.

Hanania, E. & Shikhani, M. (1986). Interrelationships among three tests of language proficiency: Standardized ESL, cloze, and writing. *TESOL Quarterly, 20* (1),

97–109.
Housen, A. & Kuiken, F. (2009). (Eds.). Complexity, accuracy, and fluency in second language acquisition. A special issue. *Applied Linguistics, 30* (4).
Housen, A., Kuiken, F., & Vedder, I. (2012). (Eds.). *Dimensions of L2 performance and proficiency : Complexity, accuracy, and fluency in SLA*. Amsterdam, The Netherlands : John Benjamins.
Hunt, K. (1965). *Grammatical structures written at three grade levels*. NCTE Research report No. 3. Champaign, IL, USA : NCTE.
Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4* (1), 51–69.
Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency : How distinct? *Applied Linguistics, 29* (1), 24–49.
Johnson, J.S. & Lim, G.S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26* (4), 485–505.
Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*, 148–161.
Kuiken, F. & Vedder, I. (2014). Rating written performance : What do raters do and why? *Language Testing, 31* (3), 329–348.
Kyle, K. (2016). *Measuring syntactic development in L2 writing : Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Unpublished doctoral dissertation. Georgia State University, Atlanta, GA.
Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15* (4), 474–496.
Lu, X. & Ai, H. (2015). Syntactic complexity in college-level English writing : Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16–27.
Magnan, S.S. (1988). Grammar and the ACTEL oral proficiency interview : Discussion and data. *Modern Language Journal, 72* (3), 266–276.
McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7* (1), 52–75.
McNamara, D.S., Crossley, S.A., & McCarthy, P.M. (2010). Linguistic features of writing quality. *Written Communication, 27* (1), 57–86.
Nakamura, D. (2019). Development of clausal and subclausal grammatical complexity and their relationship to overall length in second language writing over a year. *Bulletin of Japan Women's College of Physical Education, 49*, 45–57.
Neary-Sundquist, C. (2014). The development of complexity in a learner corpus of German. In J. Connor-Linton & L.W. Amoroso (Eds.), *Measured language : Quantitative studies of acquisition, assessment, and variation* (pp.213–227). Washington, DC : Georgetown University Press.
Norris, J.M. & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA : The case of complexity. *Applied Linguistics, 30* (4), 555–578.
Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency : A research synthesis of college-level L2 writing. *Applied Linguistics, 24* (4), 492–518.
Scarborough, H. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*, 1–22.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25* (4), 465–493.
Takeuchi, O., Yabukoshi, T., Sumi, M., & Ueki, M. (2012). *Reading stream : Pre-intermediate*. Tokyo, Japan : Kinseido.
Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning, 44* (2), 307–336.
Thomas, M. (2006). Research synthesis and histography : The case of assessment of second language proficiency. In J.M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp.279–298). Amsterdam, The Netherlands : John Benjamins.
Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research : "Clozing" the gap. *Studies in Second Language Acquisition, 33*, 339–372.
Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals : A longitudinal learner corpus study. *Modern Language Journal, 96* (4), 576–598.
Vyatkina, N. (2015). New developments in the study of L2 writing complexity : An editorial. *Journal of Second Language Writing, 29*, 1–2.
Wolfe-Quintero, K., Inagaki, S., & Kim, H.Y. (1998). *Second language development in writing : Measures of fluency, accuracy, & complexity*. Honolulu, HI : University of Hawai'i Press.

(2019年 9 月 4 日受付
2019年12月12日受理)