The Dynamics of English Writing Development in Advanced Chinese Learners

Hou, Junping

Chapter 3

An exploratory study into the dynamics of Chinese L2 writing development

This chapter is a slightly edited version of Hou, J., Verspoor, M., & Loerts, H. (2016). An exploratory study into the dynamics of Chinese L2 writing development. *Dutch Journal of Applied Linguistics*, *5*(1), 65-96.

Chapter 3 | 38

3.1 Introduction

The current study was inspired by a puzzling question concerning the development of English as a second language (L2) in China. Driven by the need to perform well on different national English exams, Chinese students spend many hours learning English. The most influential and prestigious English tests in China are the National Matriculation English Test (NMET), which is one part of the College Entrance Examination (CEE) for senior high school students, and the College English Test (CET) for non-English major university students at two different levels: the mandatory CET-4 level and the higher CET-6 level. Both the CEE and CET are taken by the entire student population, resulting in the highest numbers of students in the world taking these tests every year. For both the students and their schools, the results are of utmost importance in academic career and rankings. Students at university level, however, often complain that they progress little during their required English courses. We would like to find out if the complaint is justified.

The most common way of teaching English in China is to help students memorize fixed expressions and grammar rules in addition to practicing testing skills. Xu (2010) interviewed groups of Chinese college students about their English proficiency and their attitudes towards English after they finished their English courses at university. She argued that the main goal of English classes in China is to help the students pass the tests and not to supply a supportive English communicative environment, nor to foster a positive attitude towards English, nor to stimulate extra exposure to the language. Such practice may explain why many Chinese students are able to achieve high scores on English tests, but remain weak in productive skills. Adamson and Xia (2011) argued that the English receptive proficiency of Chinese non-English major university students has greatly improved over the last decade (from 1992 to 2003), but that their writing ability has remained rather stagnant. This is surprising as Xu (2010) showed that her participants quickly lost parts of their English abilities after university and argued that this was due to a lack of exposure to the language after university. The present study focuses on students still exposed to input during English courses at university with 5 hours a week of English learning. It can thus be assumed that despite the test driven teaching methods, there is still an

adequate amount of English exposure and writing practice to improve these students' English proficiency. The current study therefore seeks to explore if the English productive proficiency of these non-English majors as shown through written texts develops during their respective college courses, and if so in which variables. Ellis (2005:42) argued that it is important to examine free as well as controlled production in assessing learner's L2 proficiency and Schmid et al. (2011) point out that free production data can give us insight in the full range of the linguistic repertoire.

3.2 Theoretical background

Our study is inspired by both dynamic systems theory and L2 developmental studies from a usage-based perspective. Larsen-Freeman (1997), De Bot (2008) and Verspoor, de Bot and Lowie (2011) have argued that L2 development may be regarded as a dynamic process of change and have applied principles from Dynamic System Theory (DST) to the study of L2 development. In DST a complex system such as language is defined as a group of entities, sub-systems or variables that are interconnected, continuously interact, self-organize and coordinate as a whole. The dynamics of such iteration cause change to be non-linear with a great deal of variability (within systems) and variation (among systems). Systems do, however, tend to go to preferred states, called attractors. Therefore, language development is markedly complex and elusive, involving numerous dimensions that develop at varied and often non-linear rates, meaning that L2 learning is much more fluid and dynamic than any single set of descriptors can capture.

A usage-based perspective on language is basically a dynamic perspective (Langacker, 2000). It holds that there are no innate structures specific to language, that language is learned through experience and that one of the strongest factors in development is frequency of use (iteration in DST terms). Language is intrinsically linked to general cognitive processes (interconnected variables in DST terms) and intrinsically symbolic through form-meaning mappings (coordination in DST terms), constituted by a structured inventory of linguistic constructions, i.e., conventionalized form-meaning pairings used for communicative purposes (emergence and attractors in DST terms). Form-meaning pairings exist at different levels of

complexity and abstraction, comprising concrete and more abstract classes of items, and complex combinations of concrete and abstract pieces of language, so there is no rigid separation between lexis and grammar (cf. Langacker, 2000, 2008; Tomasello, 2003; Ellis, 2002). Language development is seen as a slow, gradual, piecemeal and bottom-up process, proceeding from an initial heavy reliance on concrete items (item-based) to more abstract linguistic schema in an implicit and inductive process (self-organization in DST terms).

A dynamic usage-based (DUB) perspective on language development is usually interested in finding out which different variables of the language develop and how they interact over time. Several studies have recently been conducted within this framework. The focus in the current paper is to explore which variables in language may develop at different levels of proficiency and is a partial replication of Verspoor et al. (2012) with a cross-sectional design.

In DST inspired studies examining change over time in individual learners, Larsen-Freeman (2006) designed a repeated task study on five intermediate Chinese learners of English in their oral and written production, measuring the emergence of complexity, accuracy and fluency (CAF) with both broad and specific measures. She concluded that as a learner uses language her language resources also change. Moreover, none of the learners showed similar patterns in the CAF graphs. Spoelman and Verspoor (2010) tracked a Dutch learner's acquisition of written Finnish from beginner to high intermediate over three years by examining CAF measures and concluded that the interactions of variables in systems showed 'classic jumps, transitions, and non-linear' development. Verspoor, Lowie and Van Dijk (2008) traced the development of an advanced learner, and showed that even at this advanced level there were different variability patterns and interactions among variables. By modeling the developmental trajectories of four learners, Caspi (2010) showed that three of the four learners developed their lexical complexity before lexical accuracy and their syntactic complexity before syntactic accuracy. In other words, the lexicon develops before the syntax and complexity before accuracy. This developmental order makes perfectly good sense, because a learner first has to have words to make sentences longer and more complex and (s)he first has to try things out

before they become accurate.

In a DST inspired study to explore which variables are most likely to change at different proficiency levels, Verspoor et al. (2012) assessed L2 learners' written texts at 5 proficiency levels from beginner to high intermediate with over 40 complexity and accuracy variables. They pointed out that there were five broad measures based on averages of a great number of instances that showed almost linear progression or regression across proficiency levels: fewer simple sentences (representing sentence complexity), fewer present tenses (representing verb phrase complexity), fewer errors (representing accuracy), increased type-token ratio (representing lexical diversity) and increased chunk use (total number of authentic expressions). More specific measures, which tapped into less-frequently occurring constructions such as perfect or progressive verb phrases or specific types of dependent clauses, almost all showed non-linear development, variation, and changing relationships. However, in general, they found that learners at the lower levels developed in lexical measures, then in syntactic measures and finally in lexical measures again, especially at the chunk level. Such findings were in line with those of several other longitudinal studies. Bulté (2013), who examined a sub-group of the Verspoor et al. (2012) participants longitudinally, concludes that the increase in L2 complexity at group level is fairly linear, but there is a high degree of variability at the level of individual learners.

The present study explores if Chinese university students, who themselves complain that their English L2 proficiency does not improve during the English courses, develop and, if so, whether their proficiency level affects the variables in which they do develop. From the longitudinal studies we may conclude that almost all specific constructions show non-linear development in individuals, variation among individuals, and changing relationships among the variables, as one would expect from a dynamic usage-based perspective. From the cross-sectional studies, we also may expect that at different proficiency levels different variables tend to change. Lower level students tend to develop in lexical measures and more advanced students in syntactic measures. Finally, both the longitudinal and cross-sectional studies have shown that even short writing samples can be useful in assessing general proficiency, and a cross-sectional study of

samples at different proficiency levels can give worthwhile insights into L2 developmental patterns at a very general level, which can then later be traced longitudinally.

3.3 Method

The main goal of the present study is to see if the Chinese university students' complaint that their language proficiency does not develop further during their university courses is justified. If they do develop, the question is whether the students, at three different proficiency levels, develop in the same variables. The following questions are addressed:

1. Do the three different groups show development over time in holistic scores?
2. Do the three different groups show development over time in the same or different sub-components of the holistic scores?
3. Do the three different groups show development over time in complexity measures?
4. Do the three different groups differ in development over time as far as the type of complexity measures is concerned?

3.3.1 Design

The present study has a pre-posttest design to explore the development of three different groups of students over the course of their respective educational programs.

3.3.2 Participants

Group 1 (n = 23) consists of senior school students around the age of 18 from one of the top schools in Xi'an (the largest city in Northwest China). These students usually score very high on the CEE (College Entrance Examination), which includes an English test. Students attending this school are usually highly motivated learners with specific aims for their future and most, if not all, of them want to be admitted to the best universities in China.

Group 2 (n = 8) and Group 3 (n = 18) are students at a top university in Xi'an of approximately 20 years old, which only accepts top students in CEE performance. All students take an English proficiency exam after they enter university and are divided into different classes according to their scores: The band 1 Class (our Group 2) aims to pass the CET-4 and the band 2 Class

(our Group 3) aims to pass the higher level of CET-6. Because it is important for this university to have a high CET pass-ratio, it allows repeated attempts and therefore offers courses of 18 months rather than the more traditional two-year (24 months) courses. Even though English is not the major for our participants, they have to spend a relatively large amount of time on English in order to graduate, and this burden often leads to a rather negative attitude toward learning English.

All three groups had dedicated teachers who wanted to improve their students' writing (which is a component of the different tests).

3.3.3 Instructional intervention

During their three-year high school program, students in Group 1 were asked to write English practice test papers with topics related to the CEE at home. Except for giving a general score and marking errors, the teacher gave little feedback on the students' writings but asked them to compare their samples with the examples given in the textbook.

During the 18 months course of Group 2 and 3 at university level, the teachers applied a learning-by-writing method, and spent time on practicing writing following CET directions with a Fawcett's Structure on excellent English composition (a. a clear thesis statement; b. four sentences that develop the thesis statement; c. details, facts, and examples to develop each paragraph; d. a logical order of paragraphs). Most assignments had argumentative topics, sometimes model essays or those written by class members were discussed in class. Besides grading the writings, the teachers gave rather detailed feedback on errors and brief evaluations and suggestions. Also, the university highest level students were occasionally asked to give peer feedback to each other.

3.3.4 Data collection

During their respective courses, the students wrote many texts as assignments in the course, which were all collected by the researcher. An effort was made to keep the genre consistent for all the assignments as far as feasible in this quasi-experimental setting to ensure comparability and consistency of the data (Ortega, 2003). The high school students (Group 1) wrote letters and argumentative essays, and the two university groups

(Group 2 and 3) wrote argumentative essays for all the collected writings. For the current study, only the two first and the two last assignments were used, which resulted in a final dataset containing a total of 196 (4 x 49) writing samples. The texts were scored holistically and analyzed automatically on a number of lexical and syntactic complexity measures.

3.3.5 Holistic rating measures

Holistic rating is a method of evaluating a composition based on its overall quality. Holistic rubrics are the quickest way to score papers in any content area, requiring a rater to read a paper only once (Urquhart & McIver, 2005).

L2 writing is, however, more than using linguistic constructions at different levels; it is a communicative activity that can be assessed at many different levels: the ability to convey meaning, the ability to convince, the ability to structure ideas, and so on. To gauge differences in proficiency levels one can focus on complexity, accuracy and fluency (CAF) measures. According to Skehan (2009), CAF in writing constitutes behavioral performance and reflects language proficiency. Successful performance has often been characterized as containing more advanced language leading to complexity, a concern to avoid errors leading to high accuracy if this is achieved, and the capacity to produce texts without interruption, resulting in greater fluency. CAF measures have been used across various language domains with different tools, ranging from holistic ratings to quantifiable measures (frequencies, ratios, formulas) of general or specific linguistic properties of L2 production so as to 'obtain more precise and objective accounts of an L2 learner's level within each (sub)-dimension of proficiency' (Housen & Kuiken, 2009, p.464; see also Ellis & Barkhuizen, 2005; Iwashita, Brown, McNamara & O'Hagan, 2008; Polio, 2003; Wolfe-Quintero, Inagaki & Kim, 1998). CAF measures have also been used both as performance descriptors for the oral and written assessment of language learners as well as indicators of learners' proficiency underlying their performance (Housen, Kuiken & Vedder, 2012). Furthermore, they may be 'differentially developed by different types of learners under different learning conditions, and for measuring progress in language learning' (Housen & Kuiken, 2009, p.562). All CAF measures have generally shown improvement as proficiency increases in most SLA studies.

It may be argued, however, that traditional CAF measures do not reflect some other important aspects of general proficiency, especially at the higher levels, such as idiomaticity and coherence. The L2 can be rather complex and accurate, but may still sound awkward and non-native-like (Pawley & Syder, 1983; Smiskova-Gustafsson, 2013). Idiomaticity should not be confused with the term 'idioms'. An idiom is a very conventionalized combination of words with a figurative meaning and there are thousands of idioms that occur frequently in all languages. Idiomaticity may include the use of idioms, but also includes the use of expressions that are somewhat less conventionalized than typical idioms. It refers to native-like selection of expressions (Pawley & Syder, 1983) and that which goes beyond rules and words (Fillmore, Kay, and O'Connor, 1988). Features of idiomaticity can be found at different levels, ranging from discourse to phrase levels. One of the general characterizations of idiomaticity is that it consists of knowing what situations and phenomena require which expressions—although alternatives are normally conceivable (Warren, 2005). Verspoor et al. (2012) demonstrated that idiomaticity, operationalized as the use of chunks, is a strong discriminator between proficiency levels. Smiskova-Gustafsson, Verspoor and Lowie (2012) showed that learners who had more exposure to the language expressed themselves less awkwardly and especially used longer conventionalized ways of saying things (CWOSTs). Therefore, we argue that idiomaticity, defined as the use of native-like chunks, formulaic sequences and conventionalized ways of saying things, may be seen as a separate construct, as accuracy may capture incorrect grammar, but not awkward ways of expressing certain notions.

Another feature not explicitly included in traditional CAF measures is coherence. It may be seen as a sub-construct of fluency in that a more fluent text usually flows better, but because in the current study one group of students are supposed to have worked on coherence for their CET-6 exam, we decided to include it as a separate construct in the holistic scoring rubric. Coherence is assumed to be a higher level skill and refers to the logical connections in a text. According to De Beaugrande and Dressler (1981), coherence provides a continuity of sense by making the relevance of and relations between different concepts clear, and it is generally associated with how well the discourse can be interpreted (Anderson, 1995). Coherence is

different from cohesion. Cohesion is the specific lexical or grammatical linking within a text that holds a text together. Graesser, McNamara and Louwerse (2003) state that coherence refers to the representational relationships of a text in the mind of a reader or listener, whereas cohesion refers to the cues in the text that help the reader to build a coherent representation (Foltz, 2007). Cohesive devices are important in text for connecting ideas with topics (Graesser et al., 2003) and in speech for allowing indications of coherence in a message, and providing interlocutors with a means to interpret messages (Tanskanen, 2006). In writing, if one paragraph coherently leads on to another paragraph, the main idea of each paragraph will logically connect and make sense, which makes the reader feel that the writing flows (Weigand, 2009). For our raters, we defined a coherent text as a text in which the ideas flow smoothly from one sentence to the next and one paragraph to the next so that readers can easily understand the ideas that the writer wishes to express.

Assuming that higher level students were not only different in complexity and accuracy measures, but also in the way they formulate their ideas, both in terms of idiomaticity and coherence, the current study used 5 general rubrics, created and refined during a pilot assessment procedure: complexity, accuracy, fluency, idiomaticity and coherence (CAFIC), each of which could be scored from 1 to 5 (see Appendix 2).

3.3.6 Analytic measures

Since holistic rating studies have typically found few proficiency differences within a single program level (Kern & Schultz, 1992) and since it may also be more difficult for raters to make fine-grained judgments that sufficiently discriminate individual performance within such truncated samples comprising fairly small proficiency differences (Ortega, 2003), we also used analytic measures to discover finer-grained differences to complement the global holistic scoring. Because we used automated analyses, we limit our focus to complexity at the lexical and syntactic levels.

The lexicon can be seen as a linguistic sub-system of language and lexical complexity is one of the sub-constructs of lexical proficiency and in automated analyses usually refers to a few different dimensions: diversity, richness, and sophistication (Wolfe-Quintero et al., 1998). It has been

recognized as an important construct in L2 research, as it is directly related to the learner's ability to communicate effectively in written form (Lu, 2010). On the whole, lexical complexity has been shown to be a good predictor of a learners' general language proficiency (e.g., Zareva, Schwanenflugel & Nikolova, 2005) and an essential indicator of the quality of their writing (e.g., Laufer & Nation, 1995). L2 lexical proficiency cannot be evaluated solely on vocabulary (Crossley & McNamara, 2009). Lu (2012) tested 25 lexical metrics in the spoken language of Chinese learners of English testing for the TEM-4, a test for English majors at university and therefore probably at a more advanced level than the levels of the students in the current study. He found a relatively small number of metrics to correlate well with proficiency levels.

The lexical sub-system interacts with the syntactic sub-system. More sophisticated words (e.g., nominalized verbs) may affect the syntax, but syntax is usually analyzed as a separate construct. Syntactic complexity measures derived from L2 writing samples can be useful indices of the writers' overall proficiency in the target language (Lu, 2010). Syntactic complexity is manifest in L2 writing in terms of how varied and sophisticated the production units or grammatical structures are (Foster & Skehan, 1996; Ortega, 2003; Wolfe-Quintero et al., 1998). Researchers have conducted cross-sectional studies to investigate between-proficiency differences in syntactic complexity of second language production (e.g., Bardovi-Harlig & Bofman, 1989; Ferris, 1994; Henry, 1996; Larsen-Freeman, 1978). Longitudinal studies have also been conducted to track the learners' developmental changes in syntactic complexity of second language production over an extended period of time (e.g., Ishikawa, 1995; Ortega, 2000; Stockwell & Harrington, 2003; Polio, 2001; Bulté, 2013; Verspoor et al., 2008; Spoelman & Verspoor, 2010; Penris & Verspoor, 2017). A large number of different measures have been proposed for characterizing syntactic complexity in L2 writing at the sentence and clause level, usually in terms of length of sentences, clauses and phrases or degree of subordination and coordination, each with their own advantages and disadvantages. The operationalization of different complexity measures has also been debated among many researchers because they may depend on the purpose of the research, proficiency level of the learners and text level.

According to the meta-analyses by Wolfe-Quintero et al. (1998) and Ortega (2003), six syntactic complexity measures have been used most frequently: mean length of sentences, T-unit (i.e., the shortest allowable unit of language), and clause; mean number of T-units per sentence, of clauses per T-unit and of dependent clauses per clause. However, Byrnes (2009) argues that in line with Halliday, Mathiessen and Yang (1999), linguistic development proceeds from: (i) mostly parataxis (i.e., coordination) to (ii) hypotaxis (i.e., subordination), to (iii) language with much higher levels of lexical density and more complex phrases (as opposed to more clauses). Especially the third level is not really tapped into with the commonly used L2 developmental measures. Penris and Verspoor (2017) show that at the advanced levels of academic writing a finite verb ratio (number of finite verbs /number of words) and average word length may better reflect such internal complexity within clauses.

Because we are especially interested in variables that show change at different levels of proficiency, we decided to explore not only the measures that proved most robust in previous studies, but all the metrics of lexical richness (density, sophistication and variation) and syntactic complexity provided in the automated, web-based linguistic analysis tool 'Synlex Analyzer' (Lu, 2010), even though many of them may partially overlap (see Appendix 3). For the higher level university groups, however, which appeared to regress on almost all the measures in Synlex, the texts were also hand coded for finite verb ratio, the use of non-finite verb constructions and average word length because the possibility was that this group had moved from subordination to language with more nominalizations and the use of other non-finite constructions.

3.3.7 Holistic rating procedures

After a two-hour training, 8 raters (5 native speakers and 3 non-natives) rated the texts in groups of 2. The raters knew they were rating texts for a study but were not aware of the data collection background, and the 196 essays were ordered randomly. One rater's scores were omitted because his scores were consistently different from the other group members, one rater left for private reasons during the rating, so there were 3 raters in each rating group in the final analysis.

An example of a holistic rating of one learner's writing can be found in Table 3-1 (for writing samples see Appendix 4).

Table 3-1. Sample of holistic rating scoring on one learner's writings by one group of raters

S1	R 1	R 2	R 3	Consensus Score	Computed Score	Sum Consensus Score	Sum Computed Score	**Average Consensus Score**	**Average Computed Score**
Pre-1	14	15	11	13	13.33				
Pre-2	14	9	12	12	11.67	Pre=25	25	**12.5**	**12.5**
Post-1	9	11	9	10	9.67				
Post-2	10	13	12	11	11.67	Post=21	21.34	**10.5**	**10.67**

The procedure yielded several rating results: for each rater we collected the scores on each rubric and the total for each text. For each group of raters we collected the sum of each rubric as well as the total, and average and total scores on each rubric as well as a group consensus score, which was determined by the groups after they had first individually scored each text. All scores were statistically analyzed with SPSS (Version 23) for their reliability, correlations and differences within and between groups.

3.3.8 Analytic scoring procedures

The same texts that were used for the holistic scores were submitted to the automated, web-based linguistic analysis tool 'Synlex' created by Lu (2010). The tool was designed for studies in advanced second language proficiency research and was developed and evaluated by using college-level second language writing data from the Written English Corpus of Chinese Learners (Wen, Wang & Liang, 2005). Experimental results showed that the system achieved very high reliability on unseen test data from the corpus (Lu, 2011). This is another important reason for our choice for our participants were also advanced students in China's senior school and university. Lu's analyzer produces 48 lexical and syntactic complexity measures: 25 Lexical Complexity metrics, 9 sub-scales of Syntactic structure and 14 sub-scales of Syntactic Complexity indices. Lu (2010, 2012) compared the measures and one aim was to find the most robust measures that showed linear development. Our study aims to explore which linguistic variables develop at different stages of proficiency and will use all these measures in order to

be able to make an as finely grained analysis as possible. The subsequent sections will explain the lexical and syntactic measures used in this study in more detail.

3.3.8.1 Lexical measures

According to Lu (2011), the lexical D value is calculated in a cleaned text file converted to the CHAT (Codes for the Human Analyses of Transcripts) format (MacWhinney, 2000) and subjected to three *vocd* analyses, the average of which served as the D value of the sample. To calculate the other measures, the system takes the cleaned text file, which is first tagged for parts-of-speech (POS) by means of the Stanford tagger (Toutanova, Klein, Manning & Singer, 2003). Every token in the language sample is assigned a label such as adjective, adverb, etc. The POS-tagged sample is in turn lemmatized using MORPHA (Minnen, Carroll & Pearce, 2001), a robust morphological analyzer for English that returns the lemma and inflection of a word, given the word form and its part of speech. Next, the lemmatized sample is processed by a python script, which computes the values of the measures. The lexical complexity metrics were divided into three dimensions of lexical richness (Read, 2000): density, sophistication and variation. For each of these dimensions several partially overlapping sub-measures were calculated (see Appendix 5 for complete list).

Lexical Density (LD) refers to the ratio of the number of lexical words to the total number of words in a text, i.e., the proportion of semantically full words (or lexical words) as opposed to function words and operationalized as the ratio of the number of lexical words to the number of words.

Lexical Sophistication (LS) measures the proportion of relatively unusual or advanced words in the learner's text (Read, 2000, p.203). It includes 5 different overlapping sub-measures, each calculated somewhat differently. The principle of defining LS is that it counts words, lexical words, and verbs as sophisticated if they are not on the list of 2,000 most frequent words generated from the British National Corpus (Leech, Rayson & Wilson, 2001). Lexical sophistication-I is the ratio of the number of sophisticated lexical words (tokens) to the total number of lexical words (tokens). Lexical sophistication-II is the ratio of the number of sophisticated word types to the total number of word types. Verb sophistication-I is

computed as the ratio of the number of sophisticated verb types to the number of verbs. Verb sophistication-II counts the square of the number of sophisticated verb types to the number of verbs. Corrected VS-I counts the number of sophisticated verb types to the root of twice the number of verbs in a text.

Lexical Variation refers to the range of a learner's vocabulary as displayed in his or her language use and affected by text length. Several measures have been developed to alleviate this problem: NDW, TTR and Verb Diversity. NDW means the number of different words in the text. NDW-50 refers to the number of word types in the first 50 words. NDW-ER50 refers to the average word types in the expected 10 independent random subsamples of 50 words with a random starting point. NDW-ES50 means the average word types of 10 independent random subsamples which consists of consecutive sequences of 50 words with a random starting point. TTR is the ratio of the number of word types to the number of words in a text (Templin, 1957); MS TTR (50) reduces the sample size problem which is computed by dividing a sample into successive segments of a given length and then calculating the average TTR of all segments. Corrected TTR, Root TTR, Bilogarithmic TTR and Uber index are different calculations to make TTR more specific and reliable. There are also 9 other measures based on TTR: Verb variation-I, Squared VV1, Corrected VV1; Lexical word variation, Verb variation-II, Noun variation, Adjective variation, Adverb variation, and Modifier variation.

3.3.8.2 Syntactic measures

The Synlex system uses the Stanford parser (Klein & Manning, 2003) for syntactic complexity analysis. The parsers generally require the input text to be segmented into individual sentences (with one sentence per line) and each sentence to be tokenized and part-of-speech (POS) tagged, counting the frequency of the following 9 structures in the text: words, sentences, verb phrases, clauses, T-units, dependent clauses, and complex T-units and Coordinate phrases. Adjective, adverb, noun, and verb phrases are counted in coordinate phrases and complex nominals as well. Then it computes the 14 syntactic complexity indices of the text into 5 types.

Type 1 consists of the mean length measures of a clause, sentence, and T-unit; Type 2 consists of clauses per sentence; Type 3 consists of clauses per T-unit, complex T-units per T-unit, dependent clauses per clause and dependent clauses per T-unit; Type 4 consists of coordinate phrases per clause, coordinate phrases per T-unit, and T-units per sentence; Type 5 consists of complex nominals per clause, complex nominals per T-unit, and verb phrases per T-unit (see Appendix 5).

3.3.9 Statistical analyses

The assumption of normality was checked by using Shapiro-Wilks's tests ($p > 0.05$) for each combination of levels of the two independent variables (i.e., 'group' and 'time of testing'). Paired samples *t*-tests were used to test for significant differences between pre- and posttest writings when normality of the data could be assumed. In case normality was violated, Wilcoxon Signed-Rank tests were used to compare performance of the groups on the pre- and the posttest. Non-parametric Kruskall-Wallis H Tests with subsequent Wilcoxon Rank-Sum tests were administered to compare the three groups on their pretest and posttest writings.

3.4 Results

This section will first describe the results comparing the groups and their pre- and posttest writings on holistic CAFIC scores and subsequently discusses the differences (gains or losses) between the lexical and syntactic complexity measures at the pre- and posttests.

3.4.1 Results holistic scoring

The level of agreement between the different raters was assessed by calculating Intraclass Correlation Coefficients (ICC) (Howell, 2009). There were 5 criteria of measurements (CAFIC measures) rated by 3 raters in 3 groups of participants with 4 texts (2 as pretest and 2 as posttest). Average ICC measures showed excellent interrater reliability on the overall CAFIC score (ICC(1,3) = .799). High reliability rates were also found between the raters for the sub-scores complexity, fluency, idiomaticity, and coherence (all ICCs(1,3) > .636) and a fair interrater reliability score was obtained for accuracy (ICC(1,3) = .494) allowing us to use the average scores of the three raters for further analysis.

Inspection of Q-Q plots, values of skewness and kurtosis and the Shapiro-Wilk test for normality suggested most (sub-)measures of holistic proficiency approximated the normal distribution in both the pre- and posttest writings (all *SW*s > .882; *p*s > .117) allowing for a comparison of the pre- and posttest data using Paired-Samples *t*-tests. Those measures marked with an #-sign in the Tables below, however, showed deviations from normality (*SW*s < .921; *p*s < .071) and required the use of non-parametric tests to assess significance.

A Wilcoxon Signed-Rank Test revealed that, when averaging across groups, there were no significant differences between the pre- and posttest writing overall CAFIC scores (Z = -.879; p = .379). Table 3-2 presents the results of the overall holistic ratings (CAFIC) before and after the instructional intervention for the three groups separately. As can be seen, no significant differences were found within any of the three groups.

Table 3-2. Differences in overall CAFIC score within the groups

Group	Pretest writings	Posttest writings	Paired Samples *t*-test	
(*n*)	Mean/Median (SD/IQR)	Mean/Median (SD/IQR)	*T/Z*	*Sig.*
Group 1 (23) [#]	10.50 (2.70)	11.00 (2.83)	-1.495	.135
Group 2 (8)	11.37 (1.64)	12.23 (2.28)	-1.363	.215
Group 3 (18)[#]	14.25 (3.57)	13.42 (3.87)	-.937	.349

[#] As these data did not follow a normal distribution (*SW*s < .898, *p*s < .053), Wilcoxon Signed-Rank Tests were used and medians and interquartile ranges are provided instead of means and standard deviations. *. Difference is significant at the .05 level (2-tailed).

Table 3-3 shows the differences between the groups. As we had different numbers of participants in the different groups and the course spans were different between Group 1 (30 months) on the one hand and Groups 2 and 3 (18 months) on the other, and as not all datasets were normally distributed, we ran Kruskal-Wallis tests to compare the groups and used Wilcoxon Rank-Sum Tests with Bonferroni corrections of alpha for the number of comparisons.

A Kruskal-Wallis H test revealed significant differences between the three groups in both the pretest writings, χ^2 (2) = 24.02, p < .001, as well as the posttest writings, χ^2 (2) = 20.57, p < .001. Post-hoc comparisons revealed

a significantly higher mean of ranks in Group 3 as compared to groups 2 and 1 in the pretest (see Table 3-3 for details). When comparing posttest writings, significant differences were only found between Groups 1 and 3.

Table 3-3. Differences between groups (Wilcoxon Rank-Sum post-hoc comparisons with Bonferroni corrected *p*-values)

Groups	Mean Ranks	*Z*	*Sig.*	Effect size (r^2)
Pre G1-pre G2	14.13 – 21.38	-1.943	.156	.12
Pre G1-pre G3	13.39 – 30.72	-4.600	000**	.52**
Pre G2-pre G3	7.19 – 16.31	-2.811	.015*	.30*
Post G1-post G2	14.30 – 20.88	-1.764	.234	.10
Post G1-post G3	13.59 – 30.47	-4.483	.000**	.49*
Post G2-post G3	9.50 – 15.28	-1.782	.225	.12

(Note: Pre G1=pretest of Group 1; Post G1=posttest of Group 1, etc.)

**. Difference is significant at the .001 level (2-tailed); large effect size. *. Difference is significant at the .05 level (2-tailed); medium effect size.

Since the students were put into the different groups depending on their L2 levels as reflected by their grades, we performed an analysis on the difference between the senior high school group (posttest) and the college groups (pretest) to see whether there were any significant differences between them (Table 3-4).

Table 3-4. Differences between crossed groups (Wilcoxon Rank-Sum Tests)

Groups	Mean Ranks	*Z*	*Sig.*	Effect size (r^2)
Post G1-pre G2	14.85 – 19.31	-1.199	.230	.05
Post G1-pre G3	13.80 – 30.19	-4.350	.000**	.46**

**. Difference is significant at the .001 level (2-tailed); large effect size.

Wilcoxon Rank-Sum Tests showed that the senior high-school group scored significantly lower in their posttest (*Mdn* = 11.00) than the advanced college group in their pretest (*Mdn* = 14.25), $Z = -4.350$, $p < .001$, $r^2 = 0.46$).

Table 3-5 shows the gains or losses in the sub-components of the CAFIC rubric. Although there were no differences within groups in the computed total scores, the sub-components in the rubric were tested to see if

they developed differently within each group. Table 3-5 shows that almost all sub-scores for both Groups 1 and 2 increased somewhat.

Table 3-5. Average pre- & posttest scores of separate CAFIC criteria

CAFIC items	Group 1		Group 2		Group 3	
	Pre-	Post-	Pre-	Post-	Pre-	Post-
Complexity	1.86	2.05	2.29	2.42	*2.92*	*2.81*
Accuracy	2.32	2.36	2.17	2.38	2.69	2.85
Fluency	1.85	1.96	2.13	2.56	*3.09*	*3.06*
Idiomaticity	1.99	2.10	2.38	2.48	2.61	2.67
Coherence	2.17	2.25	*2.42*	*2.40*	2.74	2.74
Total	**10.18**	**10.71**	**11.38**	**12.23**	**14.06**	**14.12**

*. In italics measures that decreased from pre to posttest.

A Paired Sample *t*-test (Table 3-6) indicated that in Group 1, 'complexity' improved significantly ($t(22) = -2.77$; $p < .05$; $r^2 = .26$) and 'idiomaticity' showed an increasing trend, ($t(22) = -1.777$; $p = .089$; $r^2 = .13$). A small increase in 'fluency' scores was found for Group 2, but this increase failed to reach significance ($t(7) = -2.012$; $p = .084$). Group 3 showed no significant differences between pre- and posttest scores.

Table 3-6. Differences between pre- & posttest in separate CAFIC criteria (Paired sample *t*-test)

Group	Values	Complexity	Accuracy	Fluency	Idiomaticity	Coherence
Group 1	*m*	.189	.037	.080	.108	.000[#]
	sig.	**.011***	.723	.351	.089	.354
Group 2	*m*	.124	.209	.438	.104	-.255[#]
	sig.	.538	.277	.084	.622	.932
Group 3	*m*	-.330[#]	.080[#]	-.038	.055	.000[#]
	sig.	.130	.347	.711	.596	.243

[#] As these data did not follow a normal distribution (*SW*s < .921, *p*s < .071), Wilcoxon Signed-Rank Tests were used and median differences are provided instead of means.

*. Difference is significant at the .05 level (2-tailed). In bold, measures that are significant; in shaded box, measures that show a strong trend.

To summarize, the three groups did not significantly improve during their respective courses on the summed up holistic scores. When comparing across the groups, only a few significant differences were found: between

Group 1 and Group 3 both at the pretest and the posttest , and between Group 2 and Group 3 at the pretest. These differences were to be expected as the university groups were streamed based on a college entrance proficiency test. As far as sub-components in the CAFIC rubric were concerned, Group 1 did improve in complexity and idiomaticity and Group 2 in fluency. Group 3 did not improve in any sub-component of the CAFIC rubric.

3.4.2 Results analytic scores

In this part, we focus on the differences (gains or losses) between the scores of the lexical and syntactic complexity measures as obtained by means of the Synlex (Lu, 2010) at the pre- and posttests as well as several hand coded measures for Group 3. Table 3-7 summarizes the findings discussed below. The statistical details of the analyses (mean and median differences) can be found in Appendix 5.

Group 1 increased in lexical density but decreased in lexical sophistication (LS-I). Group 1 significantly increased in 5 of the 10 general lexical variation metrics (NDW,CTTR, RTTR, Uber and NDW-ER50), 3 verb variation metrics (VV1, SVV1, CVV1), lexical word variation (LV) and in adverb variation (AdvV). As far as general syntactic measures is concerned, their sentences, clauses and T-units became longer and they used more coordinate phrases, complex nominal, verb phrases per T-unit.

Like Group 1, Group 2 increased in lexical density and decreased in measures of lexical sophistication-I (LS1). However, Group 2 did not progress in any of the lexical variation measures, and actually decreased significantly in lexical sophistication-II (LS2), lexical word variation (LV), a verb variation measure (VV2), noun variation (NV), and adverb variation (AdvV). In the syntactic measures there were trends towards significant increases in length of T-units and clauses and towards significant increases in the use complex noun phrases and number of verbs per T-unit.

Unlike Group 1 and 2, Group 3 showed no differences in any of the lexical density or general lexical variation measures except in lexical sophistication-II (LS2). Meanwhile, they showed significant increases in some sub measures of verb variation, adjective variation and modal verb variation, pointing to development in the more subtle areas of language

Table 3-7. Significant development (*p*-values) in lexical and syntactic complexity measures of each group.

Complexity	Categories	Indices	Measures		Group 2	Group 3
Lexical Complexity	Lexical Density	Lexical Density	LD	.086 ↑	.028* ↑	
	Lexical Sophistication	Lexical Sophistication	LS1	.058 ↓	.006** ↓	
			LS2		.020* ↓	.008 ↑
			VS1			
			VS2			
			CVS1			
	Lexical Variation	NDW	NDW	.001** ↑		
			NDWZ-50			
			NDW-ER50	.046* ↑		
			NDW-ES50			
		TTR	TTR			
			MSTTR-50			
			CTTR	.001** ↑		
			RTTR	.000** ↑		
			LogTTR			
			Uber	.009** ↑		
		Verb Diversity	VV1	.001** ↑		
			SVV1	.000** ↑		
			CVV1	.000** ↑		
		Lexical Word Diversity	LV	.021* ↑	.011* ↓	
			VV2		.055* ↓	.034* ↑
			NV		.008** ↓	
			AdjV			.006** ↑
			AdvV	.049* ↑	.060 ↓	
			ModV			.016* ↑
Syntactic Complexity	Syntactic Complexity Indices	Length of Production	MLS	.000** ↑		.000** ↓
			MLT	.005** ↑	.083 ↑	.000** ↓
			MLC	.003** ↑	.079 ↑	
		Sentence Complexity	C/S			.000** ↓
		Subordination	C/T			.000** ↓
			CT/T			.001** ↓
			DC/C			.001** ↓
			DC/T			.000** ↓
		Coordination	CP/C	.006** ↑		.011* ↑
			CP/T	.007** ↑		.035* ↑
			T/S			
		Particular Structures	CN/C	007** ↑	.002** ↑	
			CN/T	.011* ↑	.006** ↑	.001** ↓
			VP/T	.016* ↑		.014** ↓

**. Difference is significant at the .01 level (2-tailed); *. Difference is significant at the .05 level (2-tailed).

use. Surprisingly, in Group 3, almost all syntactic measures decreased significantly. All length measures (except the mean length of clause) and all subordination measures and all specific measures went down. Because the measures in Synlex may not capture complexity at the higher levels with nominalizations and other non-finite constructions (Biber & Gray, 2010), we calculated the average word-length and hand-counted the number of finite verbs to calculate the finite verb ratio (Penris & Verspoor, 2017) and non-finite verbs, of each higher level student's pre and post writings and then calculated the difference between pre- and posttest writings (Table 3-8).

Table 3-8. Extra measures for Group 3.

	Pretest	Post test	*M*	*SE*	*Sig*
Average word length (all words)	5.57	5.56	-0.16	0.38	.673
Finite verb ratio (finite verbs / words)	.12	.13	.013	.009	.157
Nonfinite verbs (non-finite verbs / words)	.09	.10	.014	.004	.002**

**. Difference is significant at the .01 level (2-tailed); *. Difference is significant at the .05 level (2-tailed).

To summarize, both Group 1 and Group 2 made some progress in some syntactic measures, while students from Group 3 in fact made no progress in language complexity as measured in Synlex. Still, rather than simply concluding that the writers in Group 3 regressed or attrited, we suspect that the learners were moving towards a more academic style because of their subtle changes in the use of modals, different verbs, and non-finite constructions. The only increase to be found was the number of coordinate phrases per clauses, explaining why the mean length of clause did not go down. The hand counted measures (Table 3-8) (average word length, finite verb ratio, and non-finite verb ratio) showed a significant increase only in the use of non-finite verbs (effect size $r^2 = .39$).

3.5 Discussion and conclusion

The goal of the study was to see if university students, especially those at the highest levels, make any progress at all in their English proficiency during their rather intensive course program. The second question was if the groups, which were at different proficiency levels, developed in different linguistic

variables. The first two and last two writing texts of each learner in three groups of learners in different course programs (senior high school, lower university and higher university English programs) were analyzed holistically and analytically in a pre-post design.

The holistic scoring rubric included 5 categories: complexity, accuracy, fluency, idiomaticity and coherence (CAFIC). There were positive relationships between the total scores of the pre- and posttest and sub scores indicated that the different criteria in the CAFIC rubric measured one factor (i.e., writing proficiency). Between groups, there were differences between the pretest of Group 1, consisting of senior high school students, and the posttest s of Groups 2 and 3, consisting of university students. This difference was to be expected because Groups 2 and 3 both had about 5 years more English instruction at the time of testing. Another difference found was between the pretests of Groups 2 and 3, with higher scores for Group 3, which can be explained by the fact that these groups were streamed at the beginning of their university career based on their English proficiency scores. Interestingly, during their respective instructional programs of 30, 18 and 18 months with an average of 5 hours of instruction per week, the three different cohorts improved little during their respective courses when scored holistically. None of the three groups made significant progress in their summed up scores during the course of their study. Group 1 did, however, progress in two of the separate rubrics: a significant increase in complexity and a trend towards an increasing score in idiomaticity, suggesting there were some changes made during the course. Group 2 showed a trend in fluency, so over time they were able to write more. Group 3 showed no trends in any of the separate rubrics. It is difficult to believe that the groups--Group 3 in particular--did not make any or much progress over the course of 18 months with five hours a week of English lessons. It may also be the case that, as Kern and Schultz (1992) suggest, holistic scores may not be fine-grained enough to discriminate within a single program level.

The analytic scores did show some differences between the groups, suggesting that at different levels of proficiency different variables develop. Group 1 had significant gains or strong trends in measurements of lexical density and lexical variation, but not in lexical sophistication. This senior school group also gained significantly in all syntactic measures except

subordination ones. Taken together these differences suggest that Group 1 developed in both the lexical domain (with relatively more lexical words and variation, but not more sophistication) and in the syntactic domain, both in general measures and three specific ones. We may thus conclude that the analytic scores indicate development in proficiency for Group 1.

Group 2 gained in lexical density, but regressed in sophistication and in most diversity measures. Group 2 gained significantly or showed a strong trend in length measures and particular structures, but also no gain in subordination. Taken together these differences suggest that Group 2 did not develop much in the lexicon and they even showed a significant decrease on some measures, but did develop somewhat in syntactic complexity over time.

Group 3 did not show progress on a lot of the measures, but did show significant gains in word sophistication and some very specific lexical diversity measures. Especially the use of more modals, and to a lesser extent more verb variation, suggests the ability to make more nuanced distinctions in statements made. As Engber (1995:151) points out 'the advanced writer, on the other hand, may be retrieving items that meet precise specifications from an already adequate lexical base, resulting in less variety but perhaps relatively high quality writing'.

We may thus conclude that although the holistic scores confirmed progress in proficiency only for Group 1, and partly for Group 2 and not at all for Group 3, the analytical scores showed clear progress for Group 1, clear progress for Group 2 and very subtle, if any, progress for Group 3. We may also conclude that as expected from a DUB perspective, learners at different stages of proficiency make progress in different variables of the language. Group 1 made progress especially in lexical variation and some syntactic measures, Group 2 mainly in syntactic measures and Group 3 in subtle linguistic areas that seemed to move towards a more academic style. However, at the same time there seemed to be a ceiling effect in their acquisition.

These findings have some implications for both research and teaching. As far as research is concerned, the findings indicate that holistic scores are indeed rather broad and for fine-grained analyses we should complement them with the analytical measures. However, none of the 48 analytical

measures have shown improvement for all three levels of proficiency, indicating that there is not one single measure that works for all, not even the most general ones that have proven robust in the literature (such as sentence length), especially not if we want to include higher levels of writing. Also, as Byrnes (2009) and Penris and Verspoor (2017) suggest, at the higher levels we need to include measures that go beyond subordination with finite verbs and include metrics such as average word length (preferably for nouns and other lexical words), nonfinite verb ratios (which reflect complexification in the form of nominalization and other non-finite constructions), which should be easy to include in Lu's Synlex program. More importantly, the findings suggest that if researchers use only one or two measures to measure development, they may not find anything as it may not be the particular structure developing at that particular level of proficiency. In line with a DUB perspective, we should assume that different variables develop at different moments in time. Using all the metrics in Lu's Synlex Analyzer has given us the opportunity to explore where changes took place.

For teaching, the implication is that despite an enormous investment in time and energy to learn English in China, improvement is slow. This finding is not so surprising as Penris and Verspoor (2017) showed that even a very advanced Dutch major of English, whose L1 is similar to English, and who had a great deal more exposure to English in their academic courses and had daily exposure to English in the media, also took about four years to acquire a truly academic writing style. Still, the Chinese students' complaint that they do not improve much at the tertiary level should be taken seriously. In DST terms they have reached an attractor state in their English proficiency and it would take a strong force to help them develop further. We assume the tests that they have to prepare for are similar in the type of questions, but differ only in level aimed at, which makes the curricula in preparing for these tests more of the same, which might be demotivating. As Xu (2010) pointed out the lessons are not geared to supply a supportive English communicative environment, nor to foster a positive attitude towards English, nor to stimulate extra exposure to the language. It might thus be an idea to change the curriculum in such a way that more exposure to the language is given, especially with texts or videos in the area of study of these students, helping students realize that a different language has

information they could use because language is learned through experience and one of the strongest factors in development is frequency of use (Ellis, 2002).

Of course, there is a strong limitation to the present study. First of all, the groups were small and from highly selective institutions, so no generalizations beyond these groups should be made. But even more importantly, as Jarvis, Grant, Bikowski and Ferris (2003) have pointed out, the group level represents only an average and not all learners of the same level behave in the same or a similar way. Van Dijk, Verspoor and Lowie (2011) additionally point out that group average development presents a picture of development that does not characterize any of the individual learning processes, so a simple pre- and posttest design study can only provide a general picture in what linguistic variables students make progress. The present data reveal considerable variation and the data do not show how 'individual learners follow different developmental paths that in many cases do not coincide with the observed mean group trends' (Bulté, 2013:358). Especially in Group 3 there was a great deal of variation among the students, some who progressed quite a bit in several measures and others who did not progress at all. Our follow up study will therefore take a closer look at them to study this advanced group in more detail. Moreover, by opting for purely objective measures and counting them automatically we ignored other interesting aspects of language development such as the development of idiomaticity and coherence. Because the measures in Synlex may not capture complexity at the higher levels with nominalizations and other non-finite constructions (Biber & Gray, 2010), we calculated the average word-length and hand-counted the number of finite verbs to calculate the finite verb ratio and non-finite verbs, of each higher level student's pre- and posttest writings and then calculated the difference between pre- and posttest writings. Still, rather than simply concluding that the writers in Group 3 regressed or attrited, we suspect that the learners were moving towards a more academic style because of their subtle changes in the use of modals, different verbs, and non-finite constructions.