L2 developmental measures from a dynamic perspective

Verspoor, Marjolijn; Lowie, Wander; Wieling, Martijn

Published in:
Learner Corpora Researchs Meets Second Language Acquisition

DOI:
10.1017/9781108674577.009

Publication date:
2021

L2 Developmental Measures from a Dynamic Perspective

Marjolijn Verspoor, Wander Lowie, and Martijn Wieling

1 Introduction

One of the challenges for research into Second Language Acquisition (SLA) is to measure language proficiency in an objective, consistent, and reliable way. Traditionally, the focus of the measurements has been on the accuracy of use in different language domains, and proficiency has been determined by the number of errors in receptive language tests (reading and listening), errors in language production (writing and speaking), and vocabulary knowledge and use. Although this is still common practice in language testing, researchers since the 1970s have emphasized that accuracy is only half the story and that evaluation of the quality of second language (L2) use is at least equally important. Since the 1970s, there has been a quest to find the best "yardstick" (Larsen-Freeman 1978) to measure the quality of L2 use objectively. Since observations of language use are crucial in finding this ideal yardstick, SLA research has benefited considerably from corpus-based research in trying to establish measures that can help trace L2 development and help determine objective proficiency measures. For instance, many examples are available of studies that used techniques to measure syntactic complexity in second language writing automatically from learner corpora (Lu 2010; Vyatkina 2012).

Recently, studies from a complex dynamic systems theory (CDST) perspective – which starts from the assumption that development is nonlinear – have questioned whether a stable "yardstick" is what we should be after. A CDST perspective to second language development (De Bot et al. 2007; Larsen-Freeman & Cameron 2008) holds that different components or *subsystems* of language may need to develop before others, and various subsystems may interact differently with each other over time. In other words, it may be possible that L2 beginners will improve on the lexicon first, then sentence constructions, and then perhaps the lexicon again.

From a CDST perspective, the only way to see such dynamic development is to conduct longitudinal, individual case studies with enough data points in which various subsystems of the linguistic system are plotted and traced. The most important reason for the necessity of longitudinal studies is that the results from group studies cannot be individualized to the personal development of the members of that group. Although generalizations from group studies are valid for measurements at one moment in time, this type of generalization is problematic for measurements over time, as the dynamic change of variables cannot be assumed to be identical for different individuals. This has been referred to as the "ergodicity problem" (see Lowie & Verspoor 2019). Consequently, if we are interested in the process of development, we will have to make use of individual case studies. And although a wide variety of (relatively small) learner corpora is available and most of these include longitudinal data of second language use (see, for instance, the LONGDALE Corpus as introduced in Meunier 2016), these corpora do not include dense data of individual language use over time as required by CDST analyses. The exploration of the process of second language development is therefore based on a limited number of studies using small learner corpora.

So far, studies tracing individuals based on small and specific data sets reveal that learners show variability (intra-individual changes over time) in each of the subsystems studied and also show numerous interactions over time between different subsystems (cf. Verspoor et al. 2008). In a few small group studies, some general trends concerning the interaction between lexical and syntactic variables have been established through computer simulation (cf. Lowie et al. 2011). In an attempt to explore such developmental patterns in a cross-sectional study, Verspoor et al. (2012) worked with holistically scored texts (ranging from 1 to 5) to represent phases in the developmental process from absolute beginner to intermediate. Each text was coded for a great number of variables (or yardsticks) representing subsystems of the language, and indeed it was found that in different phases the subsystems developed differently, suggesting that development is nonlinear and difficult to predict. To confirm the findings from the cross-sectional study of Verspoor et al. (2012), the current paper will use a longitudinal corpus with dense measurements to trace the development of 22 similar learners in a similar high school context over one academic year with up to 23 data points per learner. The current study indeed confirms that there is nonlinear growth in all variables, as each learner shows a great amount of intra-learner variability, and that there is an abundant amount of inter-learner variation, as no individual learner develops in exactly the same manner. However, using a generalized additive model (GAM) – an approach that is ideal for analyzing nonlinear change over time in iterated learning experiments – we can detect a general trend that shows clear nonlinear patterns for lexical and syntactic measures, suggesting that the fixed yardstick metaphor may be best replaced with one of "a bundle of twigs."

2 Finding an Index of Development

The field of applied linguistics – especially the field of second or foreign language development – has long benefited from corpus research in a quest to find the best predictors of language development. As early as the 1970s, both Hakuta (1976) and Larsen-Freeman (1978) called for a suitable SLA Index of Development. Based on predictors in L1 writing development, Hunt (1970) suggested the use of the T-unit. In subsequent studies such as Larsen-Freeman and Strom (1977) and Larsen-Freeman (1978), it was indeed found that in English as an L2, the average length of error-free T-units differs among developmental levels at the group level. In a comprehensive study 15 years later, Wolfe-Quintero et al. (1998) also found the best fluency measures to be T-unit length, error-free T-unit length, and clause length. However, these measures only represent a grand sweep of development at the group level, which is not necessarily the same for all the individuals in the group, as many factors may differentially affect the characteristics of writing products of L2 learners. For example, Wolfe-Quintero et al. (1998) and Ortega (2003) pointed out that variation in writing products across learners may occur when writers are compared across different tasks, in addition to the fact that learners from different first languages may have different problems with the L2. Moreover, individual differences, especially language aptitude, are known to have a strong effect on L2 development (Sparks et al. 2008). Recently, a large number of L2 developmental studies have also been discussed in terms of the dimensions of complexity, accuracy, and fluency (CAF), each of which can be operationalized with a variety of measures. In spite of the appealing subdivision in dimensions of language proficiency, Norris and Ortega warned against a universal CAF yardstick, because "it is illusory to think that what we are measuring in CAF is some kind of universal construct that can be applied across all possible learners and contexts" (2009, 575). In line with a CDST perspective, they claimed that especially complexification is variable and that such variability represents a fruitful side of development.

One of the main tenets of a CDST perspective is that variability (intra-individual change over time) is functional and inherent in the developmental process. The learner needs to select the best option from the many different forms he or she is trying out, so at certain times in the developmental process, more variability means more learning (Verspoor & Van Dijk 2012). Basically, to develop in language and learn something new, learners will have to try out different forms to begin with and only after sufficient iteration will they eventually settle for one form or the other. However, because of a lack of attentional resources, a form that may seemingly have been settled may become unstable again when the learning starts to concentrate on another new type of form. This process leads to variability and sometimes developmental peaks in many different subsystems of language. As development is an individually owned process and no individual develops in exactly the same manner (cf. Chan et al. 2015), variation (inter-individual differences at one point of time) is to be expected.

One excellent example of this variable, wave-like process with developmental peaks is the development of negative constructions in a 13-year-old Spanish learner of L2 English, originally reported on by Cancino et al. (1978) and later analyzed from a CDST perspective by Verspoor et al. (2008). This learner had been in the US for less than three months when his language development was traced for 10 months using data from elicitation interviews and free response data. As replicated in Figure 9 below, Cancino et al. (1978) plotted all verb phrases containing a negative construction to see if the L2 learner showed patterns from four developmental phases similar to L1 learners of English, starting with phase one with *No-V* constructions (*No singing song*), to the second phase with *Don't V* constructions (*I don't hear; He don't swim*), to the third phase with *Aux-negative* constructions (*You can't tell her*), and finally ending with adult-like forms in phase four with *Analyzed do* constructions (*One night I didn't even have the light*).

There are several nonlinear patterns noticeable in Jorge's development of these constructions. First of all, even at the very beginning he uses all four constructions, but not many of the two advanced phases. He uses the *No* and *Don't* constructions equally at first, but then there is a peak of *No* constructions at data point 3, which disappear quickly around data point 7, but remain at a very low level until data point 12. The *Don't* constructions are partially target forms that are overgeneralized here and show a developmental peak at data point 7. To evaluate the significance of this peak, we ran a Monte Carlo analysis, which calculates the chance of finding an instance in time (like a peak) by comparing the actual data to 5,000 permutations of randomized

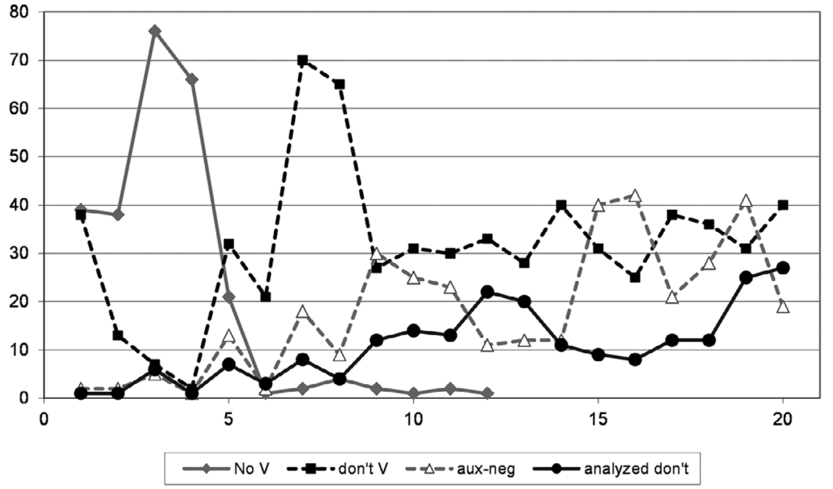

Figure 9 Jorge's (13-year-old Spanish learner of L2 English) development of negative constructions (with permission, Verspoor et al., 2011)
Note. The graph shows the percentage of correct use (y*-axis) over time in biweekly recordings* (x*-axis).*

sequences of the measurements observed. In these data, the chance of finding this peak in a random order turned out to be less than 5 percent. The *Aux-negative* constructions begin to develop more around data point 9 and by the end of the chart, there is a good mix of constructions that seem to be used in a target-like manner: the *No* constructions and the relative overuse of *Don't* constructions have disappeared from his repertoire. This example shows that this L2 learner had a developmental pattern very similar to L1 learners of English, but more importantly it demonstrates that such a longitudinal analysis also shows how intricate, variable, and jumpy the actual process is, with overlapping phases. This type of analysis can only be done on individuals, as group trajectories would average out all the peaks and dips (which learners have at different times). This also shows the relevance of multiple dense measurements in learner corpora.

Over the past 10 years there have been various longitudinal studies in the same vein, examining the writing development of one to 10 participants over the course of 10 months–13 years, and from absolute beginners to advanced learners. The findings can be summarized as follows: every study so far has shown variability in almost every variable traced, and especially beginners, like Jorge in the example given, seem to show peaks of overuse in various target or non-target-like constructions (Spoelman & Verspoor 2010; Verspoor et al. 2008). The studies have shown that there is a great degree of variability in each learner in almost every measure from broad to specific (Verspoor et al. 2017). We can also see that lexicon and syntax may compete even at advanced stages (Caspi & Lowie 2013; Penris & Verspoor 2017), and that similar learners, even identical twins in a similar context, will take different developmental paths (Chan et al. 2015; Tilma 2014; Vyatkina 2012). The strong point of these longitudinal studies is that they show in detail how each individual develops over time and how variables interact over time within the same individual. However, they do not lend themselves to generalizations about general trends at the group level.

While developmental findings of individual cases cannot be generalized to groups of learners, the results from group studies cannot be individualized, as they cannot reveal the nonlinear and variable paths that individual learners tend to follow. For example, as Verspoor et al. (2011) and Larsen-Freeman (2006) found, when the trajectories of learners are averaged out, the resulting line does not look like any one individual learner, as the insightful peaks and dips are smoothed away and even the direction of development may vary within the group. Verspoor et al. (2012) wanted to explore whether such variability and nonlinear behavior could also be detected in a cross-sectional study in which they controlled for inter-learner variation caused by known predictors, such as L1, age, aptitude, and task. Their corpus consisted of 437 writing samples from a homogeneous group: Dutch learners of English as a foreign language between the ages of 11 and 14 with similar scholastic aptitude scores. Each text received a holistic rating for the quality of their writing, resulting in groups of texts at five levels from beginner to intermediate. Each of these texts was coded for 64 CAF measures. The authors found several good predictors to discriminate consistently between proficiency levels, most of which had already been recognized in the literature. Noticeable about these measures was that they were all rather "broad" in that they averaged over a large number of instances (Guiraud index), they were clustered (all dependent clauses combined, all chunks combined,[1] all errors combined), or they were very frequently occurring constructions in the language (simple present versus other tenses). In other words, these measures were likely to show significant differences between all levels –

[1] "Chunks" are also commonly referred to as "formulaic sequences" or Conventional Ways of Saying Things (CWOSTs). Here we mean any type of fixed word combinations.

and therefore suggest linear growth – because they involved frequently occurring phenomena.

The data confirmed the working hypothesis that groups of learners move from the simplest and most frequent constructions to more complex and less frequent ones. For example, beginners, who rely on their L1 to a great extent, used mostly simple sentences in mainly the simple present tense. As the proficiency level increased, the language became somewhat more complex with an increase in all complexity, lexical, and accuracy variables. At the higher proficiency levels, all measures had improved with more complex and more accurate constructions at all levels. The total number of dependent clauses, total number of chunks, number of present or past finite tenses, and the type–token ratio of words occurring in the text were the strongest discriminators.

In addition to these rather linear trends in broad measures, there was also clear nonlinear development from one level to the next. At different proficiency levels, there were signs of overuse of different forms. At the lowest level, all simple constructions were overused, but at the third level, the present perfect and progressive showed a significant rise in the chart, accompanied by a peak in verb use errors. In addition, some (groups of) variables showed a significant difference between two consecutive levels once only, suggesting that particular aspects of the language were focused on at particular proficiency levels. Between levels 1 and 2, there was a significant difference in six variables (schematic chunks, fixed chunks, particles, most frequent words, lexical errors, and mechanical errors), which were mainly lexical in nature. Between levels 2 and 3, there was a change in seven variables (decrease in simple sentences and increase in complex sentences, adverbial clauses, non-finite clauses, partially schematic chunks, particular complement constructions, and spelling), which were mainly syntactic in character. Between levels 3 and 4, there was a mixture of changes: some syntactic measures (finite relative clauses), some lexical measures (fixed phrases), and some accuracy measures (verb use errors). Between levels 4 and 5, mainly lexical changes took place (particles, compounds, and fixed phrases) and very few syntactic changes were found.

To summarize, the cross-sectional data suggested that absolute beginners (between levels 1 and 2) are especially engaged in learning words. Then the learners seem to focus more on syntactic complexity (between levels 2 and 3), which continues between levels 3 and 4, but is then mixed with lexical measures. After changes in syntactic constructions, there is a focus again on lexical matters (between levels 4 and 5). Assuming that L2 learners go through these levels consecutively, albeit with some variability, Verspoor et al. (2012) suggested that it would be very useful to follow similar learners over time to check whether these patterns indeed occur longitudinally as suggested by this cross-sectional study.

The aim of the current paper is therefore to examine whether findings in L2 development based on this cross-sectional study can hold for individual learners over time. In other words, can findings from the group be generalized to the individual and vice versa? From a CDST perspective, we expect intra-learner variability and inter-learner variation, as each learner will follow his own developmental path, so the findings of one learner cannot be generalized to the group, nor can the findings of the group be generalized to the individual. However, as Molenaar and Campbell (2009) have pointed out, generalization to the wider population is possible if we find similar developmental paths in highly similar individuals. In line with these observations, we would expect to find some of the hypothesized general developmental findings from the cross-sectional study to apply to most of the individual learners in the current study.

To do so, learners similar to those in Verspoor et al.'s (2012) cross-sectional study are traced over the course of one academic year to see if their general proficiency develops from lower to higher levels, if learners show similar patterns of development in that they show variability in all measures and show developmental jumps in some (even if these jumps will occur at different points in time for different learners), and whether the group as a whole shows nonlinear patterns in lexical and syntactic development in the sense that one may develop before the other.

3 A Longitudinal Case Study

In this longitudinal multiple case study, we traced the development in L2 writing of 22 Dutch learners of English over 23 weeks. These learners were similar to those in the cross-sectional (CS) study by Verspoor et al. (2012) (from now on referred to as the CS study). They were 12 to 13 years old, they had similar levels of scholastic aptitude as measured by scores on a standardized scholastic aptitude test (CITO scores), and they were in a similar school setting (bilingual education). The one difference between the current longitudinal and the CS study is that in the current study, the writing tasks covered different topics, since it would be impossible to ask these young learners to write about the same topic every week as can be done in a CS study. As in the CS study, all learner data were first anonymized and then rated by a team of trained judges on proficiency levels from

1 to 5. Then each sample was analyzed on a number of syntactic and lexical variables.

For both the CS study and this longitudinal study, independent scoring sessions took place. With a team of eight experts the weakest and strongest writings were determined in the particular data set and given scores from 1 to 5. The cross-sectional data set contained texts by absolute beginners, who had not had any exposure to English before starting high school, and students at the beginning of the third year, who had had bilingual education for two years. The scores given in both the CS set and the current study ranged from 0 to 5. The least proficient (some English) received a 1 and the most proficient a 5. As these two data sets were rated independently from each other, the scores may be in a similar range but are not the same in an absolute sense.

3.1 Participants

The participants in our study were 22 Dutch learners of English who started secondary school shortly before the onset of data collection. These learners attended the same school in a small town in the north of the Netherlands and were of approximately the same age (12 or 13). The learners had enrolled in an English–Dutch bilingual stream, in which at least 50 percent of all classes (from History to Mathematics) were taught through English in a Content Language Integrated Learning (CLIL) setting. To be allowed into the bilingual stream, students were interviewed and selected on motivation and scholastic aptitude. This school setting and the pervasiveness of English in the Dutch environment affords rather massive exposure to English during the period of observation. The learners varied somewhat in the number of English classes they had had prior to starting high school. Lowie and Verspoor (2019) show in a regression analysis that in this homogeneous group neither motivation nor aptitude were predictors in proficiency gains.

3.2 Materials, Procedure, and Analyses

For the purpose of our study, students were asked to produce a short piece of writing on a topic decided on by the teacher every week, which yielded 23 longitudinal samples for each individual. Writing was done digitally on a school computer. The topics related to their experiences at school and in daily life, from "My first month at school" to "Christmas carols" and "The May break." The data collection for this project still continues, but the data reported here were gathered between November 2015 and May 2016. Due to incidental absences, most learners had missed two or three writing sessions, leading to a total of 388 writing samples.

The following texts are examples written by Student 22 at the beginning and end of the data collection session.

Student 22, week 3: I like the first week at school the most, because I like playing games and we playing games in the building. First we doing team sports in the Gym building. I like the American Football the most. . . .

Student 22, week 22: Vlieland is a wonderful island with friendly inhabitants. Our hotel was at the coast and we came from the harbour to the hotel by a TukTuk (A kind of car). The hotel was nice and there were seagulls everywhere. When we came back to the mainland, I almost fell of the boat (Oops . . .). This was at the end of the holiday.

The holistic grading procedure of these writing samples was the same as in the CS study. First the students' writing samples were anonymized and fully randomized for student and sequence of writing. Ten experienced raters were trained until agreement was reached on the holistic scoring of a subset of samples from the data on a five-point scale, with 1 representing the weakest and 5 the strongest piece of writing in terms of overall proficiency (Note, however, that these ratings are based on the relative weak and strong samples within the corpus, so the current numbers 1–5 do not exactly match those in the CS). The focus was on the complexity and the fluency of the writing samples. After the training session, in which the team of raters created their own benchmarks, the remainder of the 388 samples was rated by three raters independently. All samples with more than one point difference among the raters were reassessed by two other raters. After this procedure the rater reliability was assessed by calculating an Intraclass Correlation Coefficient (ICC) on absolute agreement (two-way mixed model). The resulting ICC was 0.78, which indicates a good reliability. Then the holistic score for each text was calculated as the average of three ratings. Lowie and Verspoor (2018) checked for a possible effect of topic on the writing quality of the same data set by calculating average ratings for all topics. This evaluation showed a gradual increase of the scores on the topics over time (ranging from an average rating of around 2.1 for the early writings to 2.9 for the later writings). After correcting for the increasing trend, none of the scores for the topics seemed to deviate from the expected score and there was no reason to delete any of the topics from the data set. Average text length for the topics was 95 words and varied between 82 and 125 average words per text, with a gradual increase toward the later samples. In addition to the global ratings, the writing samples were analyzed on a number of syntactic and lexical complexity measures using TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity) (Kyle 2016; Lu 2010).

To allow for a comparison to the CS study, syntactic development in this study was operationalized as the mean length of T-unit (MLTU) and lexical development as Guiraud. These measures have shown to be robust developmental measures in both cross-sectional studies and longitudinal studies (cf. Bulté 2013). MLTU is a global measurement of syntactic complexity, and Guiraud is a measure of lexical richness corrected for text length that is a reliable measurement of the learner's productive lexicon (van Hout & Vermeer 2007).

3.3 Analyses

Regression analyses were run to test whether students had improved significantly in writing proficiency over time and which analytical measures correlated significantly with gains in proficiency.

To test for developmental peaks, each measure for each learner was tested using the model described by Verspoor et al. (2011) but using an R-script calculating permutations. Basically, the model uses 5,000 randomized iterations to test whether the maximal distance between the lowest and the highest score is random or not. If the chance of finding the same distance is less than 5 percent, the peak is considered significant, meaning it is not a random effect and therefore suggests a developmental peak.

To see if general patterns existed among the 22 learners, we used GAM (Wood 2006, 2017) as our analysis method. This approach is an application of linear mixed effects regression but allows for the analysis of potentially nonlinear patterns over time. The GAM analysis is very suitable to analyze individual (nonlinear) patterns of iterative development in time series, as each next step in development is predicted on the basis of the previous step, and no linear development is assumed (Winter & Wieling 2016). Since the models created with GAM analyses do not include easily interpretable coefficients, the best way to interpret the model fit is through visualization.

4 Results

4.1 Proficiency Gains

Lowie and Verspoor (2019) carried out a group analysis of the mean rating of the first two samples and the last two samples in the writing

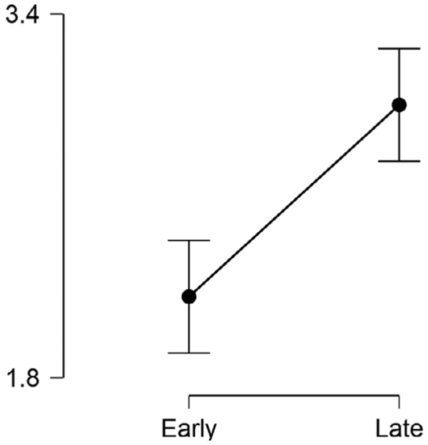

Figure 10 Average holistic rating (five-point scale) of first two (Early) and last two (Late) measurements of the writing samples of the 22 learners
(With permission, Lowie & Verspoor 2018)

corpus (Figure 10). This analysis showed that the group of learners had significantly higher average holistic scores at the end (Wilcoxon signed rank = 20.50; $p < 0.01$; Effect size (matched rank biserial correlation): r = 0.84). In other words, the group significantly improved in writing ability. As far as the complexity measures are concerned, a regression analysis indicated that only two lexical measures, *mean length of words* and *Guiraud* were significant predictors for the overall ratings.

4.2 MLTU

Figure 11 shows the individual patterns of the 22 learners' MLTU per text over time. The final graph shows the general trend by means of a GAM. Visual inspection reveals that no learner shows identical development; some learners show peaks early on (Learner 4), some in the middle (Learner 8), some toward the end (Learner 21), and some show no real single peaks at all (Learner 13). For each learner, all peaks were tested for significance over time using Monte Carlo simulations with 5,000 iterations. The analyses showed that for MLTU the peak showed a significant trend for Learner 12 (Mean = 7.7; Variance = 7.1. 342 out of 5,000 iterations yielded values equal to or larger than the critical value 11.45417; $p = 0.07$) and for Learner 17 (Mean = 6.2; Variance = 2.1. 258 iterations ≥ 8.6875: 258; $p = 0.05$). The individual plots have been scaled per individual by applying a

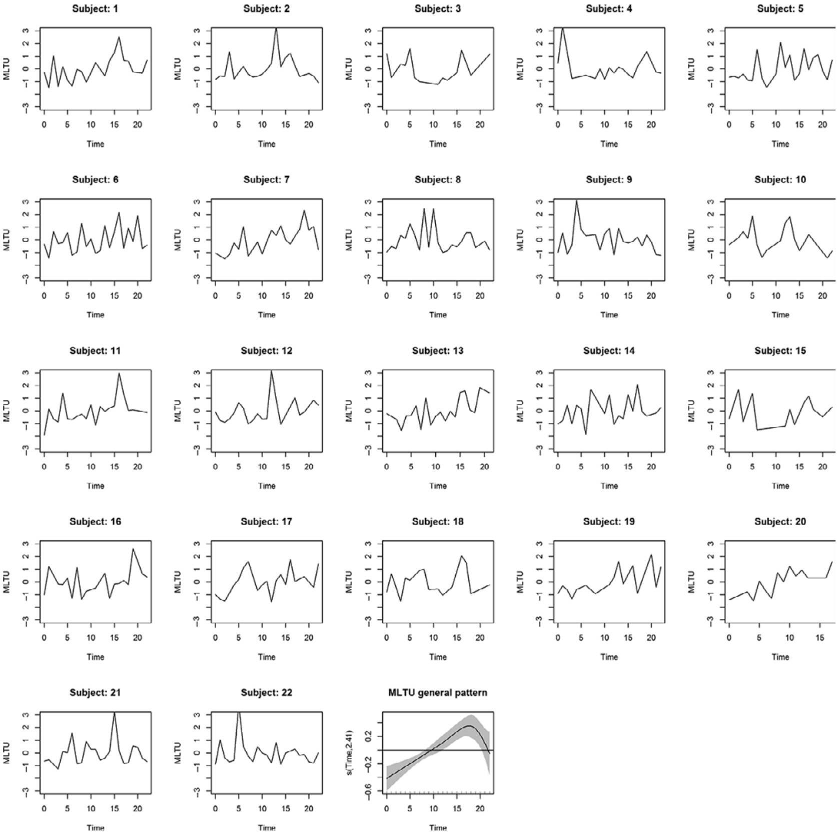

Figure 11 MLTU for individual learners over time in weeks (x*-axis), scaled per individual by z-transformation*
Note. The final graph shows a general MLTU pattern (GAM) for the group trend.

z-transformation. The GAM plot (last picture) shows a general trend for the group, taking individual variation into account by including individual scores as a random factor. These data show an upward trend, with a sharp decrease at the very end. The GAM analysis shows that the MLTU pattern deviates significantly from 0 ($p < 0.001$).

4.3 Guiraud

Figure 12 shows the individual patterns of the 22 learners' Guiraud per text over time, scaled per individual by using z-transformations. The final graph shows the general trend as the outcome of the GAM

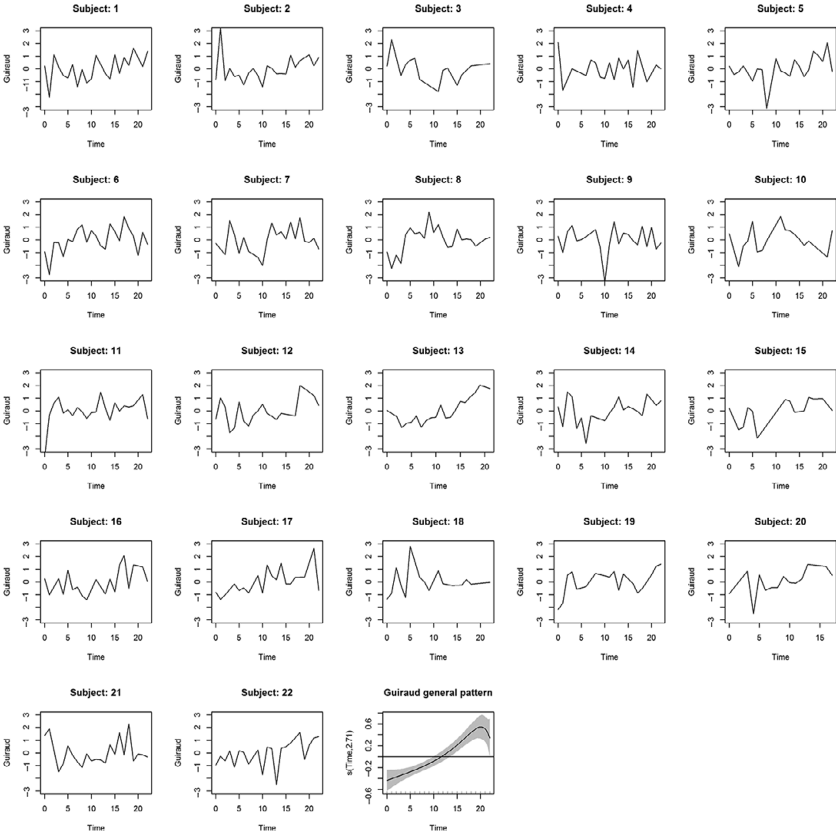

Figure 12 Development of Guiraud for individual learners over time in weeks (x-axis), scaled per individual with z-transformations (y-axis)
Note. The final graph shows the overall development for Guiraud resulting from a GAM analysis.

analysis. Much like the MLTU graphs, visual inspection of the Guiraud patterns seems to indicate that no learner develops in the same way. However, Monte Carlo analyses (simulations with 5,000 iterations) showed that none of the peaks reached significance. The GAM plot shows a general upward trend for the group, with a decrease at the very end.

4.4 MLTU versus Guiraud

Figure 13 compares the overall development of MLTU (on the left) and the Guiraud (in the middle) based on GAMs, as already shown in

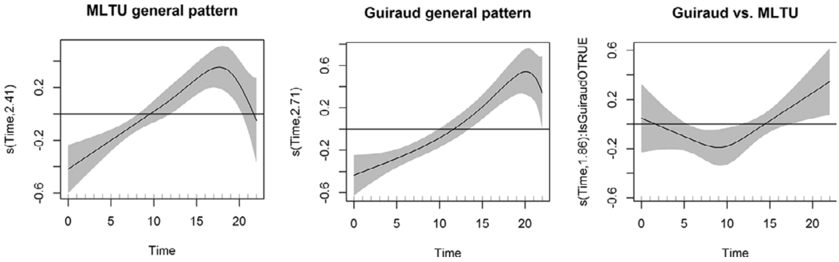

Figure 13 General effects over time in weeks for the MLTU measure (left; p < 0.001), the Guiraud measure (middle; p < 0.001), and their difference (right; p = 0.02)
Note. The measures were standardized for each subject using z-transformations for comparability of variables (y-axis).

Figures 11 and 12. Both analyses show a comparable pattern with an increase at first, which reduces at the end. The comparison between the MLTU and Guiraud (on the right) shows that especially between week 5 and week 15 the Guiraud scores are lower than those for MLTU, while toward the end Guiraud is higher than MLTU. This difference turned out to be significant ($p = 0.02$).

5 Discussion

In our continuing quest to find a common yardstick in L2 developmental studies that can make use of objective corpus research tools, freely available as in Kyle (2016) and Lu (2010), we compared the findings of a cross-sectional study (Verspoor et al. 2012) with those in a longitudinal multiple case study to explore whether group trends can be found in all the data with extensive variation among learners. The learners in the CS study and the current multiple case longitudinal study were very similar in L1, age, aptitude, and learning context. The topics of the writing tasks differed in the longitudinal study, and there is no doubt that the topics affected the individual learners differentially over time, but analyses showed that the average holistic ratings, once the incline over time had been corrected for, were rather similar. Inferential statistics showed that the learners progressed significantly over time, not only as assessed by holistic scores, but also in terms of syntactic complexity (MLTU) and lexical richness (Guiraud). Therefore, the first research question, i.e. whether the learners' general proficiency level would increase, can be answered positively: the group's writing samples were of a significantly higher level at the end. The observation that the quality of writing increased for all learners in this context is not surprising with this selective group of learners who had at least 15 hours of English exposure at school per week.

The second question about variability can also be answered somewhat positively. Like the learners in all studies from a CDST perspective, the learners in this study all showed variability in all measures (holistic, MLTU, and Guiraud), but only three (near) significant developmental peaks were found, all for the MLTU and none for the Guiraud. Visual inspection of the individual graphs did not show any obvious patterns that the learners in this study had in common.

However, in the current study, we also wanted to see if we could detect common nonlinear patterns for the measures in the group. The question was whether the lexicon and syntax develop synchronously or whether one develops before the other, as shown by Caspi (2010) and suggested by Verspoor et al. (2012). Using GAM analyses – which include iterative learning and variability in its algorithms – we were indeed able to detect nonlinear patterns in syntactic development (MLTU) and lexical development (Guiraud). For our purposes, the significantly different developmental patterns for syntax and lexicon are especially important. Figure 13, in which the values were transformed into z-scores per individual, suggests that the MLTU has higher values before Guiraud does. This finding does seem to confirm the findings in the CS study in that between levels 2 and 4 (approximately the comparable levels between the two datasets), there was a change in mainly syntactic variables after which lexical changes would take place, pointing to different growth patterns for different components or subsystems in the language. The regression analysis also showed that measures at the lexical level (Guiraud and mean length of word) were the only predictors for proficiency gains expressed in holistic scores. This suggests that at different phases in language development the subsystems may show different developmental patterns. These patterns vary among individuals, which is shown by the dispersion displayed by the bandwidth in Figure 13. The bandwidth is rather large at the initial stages, then tends to reduce over time, but dramatically increases toward the end of the data collection. The learners' development first converges, but then diverges again. This means that our metaphor for an index of development may indeed need to change from the static yardstick to the dynamic idea of a bundle of interacting twigs.

The current study also shows that as useful as cross-sectional studies can be to find general patterns in development, it is not until we look at individuals over time that we can see the real intricacies of the actual process, and that it may be less predictable and more chaotic than we might have been able to imagine.

An important implication of this finding is that we need more studies that allow us to track the process of development, i.e. longitudinal case studies with dense measurements over a long period of time. Both the density of the measurements and the duration of the period of observation depend on the expected amount of change, which may occur at different time scales for different learners and for different variables. This means that for some variables (or aspects of language development) we may have to measure over the period of a lifespan while for others we may have to focus on milliseconds (Plat et al. 2018). To quantify development, we should no longer aim to use one optimal yardstick for all learners and all timescales, but rather use a bundle of twigs. The corpus tools for efficient longitudinal analyses of learner language are available for the written modality (Kyle 2016). However, process-based research into second language development is dependent on learner corpora that consist of longitudinal data of individual learners and that are measured with the density and the duration that is relevant for the focus of the study. Currently, hardly any corpora exist that would allow for CDST analyses. The study of the process of second language development would therefore strongly benefit from the availability of additional corpora, starting from collections of written data that can be extended to corpora of spoken language.

References

Bulté, B. (2013). The Development of Complexity in Second Language Acquisition: A Dynamic Systems Approach. Unpublished Ph.D. Dissertation, University of Brussels.

Cancino, H., Rosansky, E., & Schumann, J. (1978). The acquisition of English negatives and interrogatives by native Spanish speakers. In E. M. Hatch (ed.), *Second Language Acquisition: A Book of Readings*, 207–230. Rowley, MA: Newbury House.

Caspi, T. (2010). *A Dynamic Perspective on Second Language Development*. Groningen: University of Groningen.

Caspi, T. & Lowie, W. M. (2013). The dynamics of L2 vocabulary development: A case study of receptive and productive knowledge. *Revista Brasiliera de Linguistica* 13(2), 437–462.

Chan, H., Verspoor, M., & Vahtrick, L. (2015). Dynamic development in speaking versus writing in identical twins. *Language Learning* 65(2), 298–325.

De Bot, K., Lowie, W. M., & Verspoor, M. H. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition* 10(1), 7–21. https://doi.org/10.1017/S1366728906002732

Hakuta, K. (1976). A case study of a Japanese child learning English as a second language. *Language Learning* 26(2), 321–351.

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly* 4(3), 195–202.

Kyle, K. (2016). Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Unpublished Ph.D. dissertation, Georgia State University.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly* 12 (4), 439–448.

(2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics* 27(4), 590–619.

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

Larsen-Freeman, D. & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning* 27(1), 123–134.

Lowie, W. M. & Verspoor, M. H. (2019). Individual differences and the ergodicity problem. *Language Learning* 69(S1), 184–206.

Lowie, W. M., Caspi, T., Van Geert, P., & Steenbeek, H. (2011). Modeling development and change. In M. H. Verspoor, K. De Bot, & W. Lowie (eds.), *A Dynamic Approach to Second Language Development: Methods and Techniques*, 22–122. Amsterdam & Philadelphia: Benjamins.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Meunier, F. (2016). Introduction to the LONGDALE Project. In E. Castello, K. Ackerley, & F. Coccetta (eds.), *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*, 123–126. Berlin: Peter Lang.

Molenaar, P. C. M. & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science* 18(2), 112–117.

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.

Penris, W. & Verspoor, M. (2017). Academic writing development: A complex, dynamic process. In S. Pfenniger & J. Navracsics (eds.), *Future Research Directions for Applied Linguistics*, 215–242. Bristol: Multilingual Matters.

Plat, R., Lowie, W., & de Bot, K. (2018). Word naming in the L1 and L2: A dynamic perspective on automatization and the degree of semantic involvement in naming. *Frontiers in Psychology* 8(2256). https://doi.org/10.3389/fpsyg.2017.02256

Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2008). Early first-language reading and spelling skills predict later second-language reading and spelling skills. *Journal of Educational Psychology* 100(1), 162–174.

Spoelman, M. & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics* 31(4), 532–553.

Tilma, C. (2014). The Dynamics of Foreign versus Second Language Development in Finnish Writing Unpublished Ph.D. dissertation, Rijksuniversiteit Groningen/University of Jÿvaskyla.

Van Hout, R. & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (eds.), *Modelling and Assessing Vocabulary Knowledge*, 93–115. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511667268

Verspoor, M. & Van Dijk, M. (2012). Variability in a dynamic systems theory approach to second language acquisition. In C. Chapel (ed.), *The Wiley-Blackwell Encyclopedia of Applied Linguistics*. Hoboken, NJ: Blackwell Publishing.

Verspoor, M., De Bot, K., & Lowie, W. (eds.) (2011). *A Dynamic Approach to Second Language Development: Methods and Techniques (Vol. 29)*. Amsterdam & Philadelphia: John Benjamins Publishing.

Verspoor, M., Lowie, W., Chan, H. P., & Vahtrick, L. (2017). Linguistic complexity in second language development: Variability and variation at advanced stages. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle* 14(1). https://doi.org/10.4000/rdlc.1450

Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal*, 92(2), 214–231.

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

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal* 96(4), 576–598.

Winter, B. & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution* 1(1), 7–18.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Honolulu, HI: University of Hawaii Press.

Wood, S. (2006) *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press

(2017). *Generalized Additive Models: An Introduction with R, 2nd Edition*. Boca Raton, FL: CRC press.