L2 developmental measures from a dynamic perspective

Introduction

SLA research has benefitted a great deal from corpus-based research – usually based on group studies—in trying to establish measures that can help trace L2 development and help determine objective proficiency measures. Since the 1970s, there has been a quest to find the best "yardstick" (Larsen-Freeman, 1978) to measure proficiency objectively. Recently, studies from a complex dynamic systems theory (CDST) perspective—which starts from the assumption that development is non-linear—have questioned whether a stable "yardstick" is what we should be after. A CDST perspective holds that different sub-components of language may need to develop before others, and various sub-components 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 a longitudinal, individual case study with enough data points in which various sub-components of the linguistic system are plotted and traced. So far several studies tracing individuals have been conducted, showing that learners show variability in each sub-component, some with developmental peaks, and various interactions over time between different sub-components (cf. Verspoor, Lowie & Van Dijk 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, Caspi, van Geert & Steenbeek, 2011). In an attempt to explore such developmental patterns in a cross-sectional study, Verspoor, Schmid, and Xu (2012) worked with holistically scored texts (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 representing sub-components of the language and it was found that in different phases, different sub-components indeed developed, suggesting that development is non-linear.

To confirm the findings from this cross-sectional study, the current paper will trace the development of 22 similar learners in a similar context longitudinally over one academic year with up to 23 data points per learner. The current study indeed confirms that there is non-linear growth in all variables as each learner shows a great amount of variability, including peaks, and that there is a great deal of variation as no one 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 show clear non-linear patterns for lexical and syntactic measures suggesting that the fixed yardstick metaphor may be best replaced with one of “a bundle of twigs.”

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, 1976) called for a suitable Second Language Acquisition Index of Development. Based on predictors in L1 writing development, Hunt (1970) suggested the use of a 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 differ among developmental levels at the group level. In a comprehensive study 15 years later, Wolfe-Quintero, Inagaki, and Kim (1998) also found the best fluency measures to be T-unit length, error-free T-unit length, and clause length. However, these measures only hold for development at the group level, not necessarily for individuals, as many factors may affect the characteristics of writing products of L2 learners. For example, Wolfe-Quintero et al. (1998) and Ortega (2003) point 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, Patton, Ganschow, Humbach, & Javorsky, 2009). Recently, a great deal of L2 developmental studies have also been discussed in terms of complexity, accuracy, and fluency (CAF) measures, but Norris and Ortega (2009) warn 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” (p. 575). They point to the need for multivariate analyses, as it is problematic to measure development solely through length measures. Moreover, in line with a CDST perspective, they claim that especially complexification is variable and that such variability represents a fruitful site 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 form of behavior among 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 enough iteration they will eventually settle for one form or the other. However, because of a lack of attentional resources, a form that may seemingly have been settled upon may become unstable again when another new type of form is being worked on by the learner. This process leads to variability and sometimes developmental peaks in many different sub-components of language. As development is an individually owned process and no individual develops in exactly the same manner (cf. Chan, Verspoor, & Vahtrick 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, Rosansky, and Schumann (1978) and later analyzed from a CDST perspective by Verspoor, Lowie, and Van Dijk (2008). This learner had been in the US for less than three months when his language development was traced for 10 months with elicitation interviews and free response data. Cancino et al. (1978) plotted all verb phrases containing a negative construction to see if the L2 learner showed patterns of 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-neg* 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*).

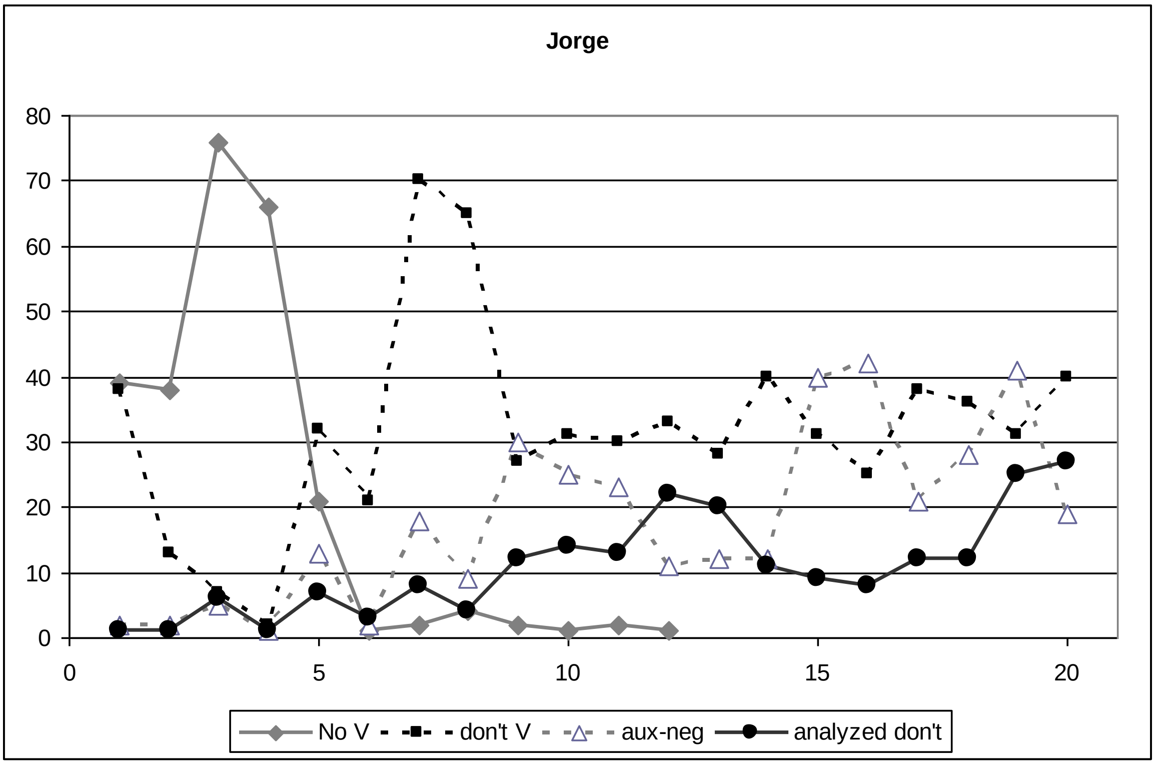

Fig. 1. Jorge's (13-year old Spanish learner of L2 English) development of negative constructions (with permission, Verspoor et al., 2011)

There are several non-linear 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 still remain at a very low level until data point 12. The *don't* constructions, which are partially target forms but overgeneralized here, show a developmental peak at data point 7, which proved to be significant in a Monte Carlo analysis (the chance of this peak being random was less than 5%). 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 that such a longitudinal analysis also shows how intricate, variable and jumpy the actual process is, with overlapping phases. This type of analysis can be done on individuals only, as group trajectories would average out all the peaks and dips (which learners have at different times).

Over the past 10 years there have been various longitudinal studies in the same vein, examining the writing development of one to ten participants over the course of 10 months to thirteen years from absolute beginners to advanced learners. The findings so far 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, Lowie, and Van Dijk, 2008), there is a great degree of variability in each learner in almost every measure from broad to specific (Verspoor, Lowie, Chan & Vahtrick, 2017), lexicon and syntax may compete even at more advanced stages (Caspi & Lowie, 2013; Penris & Verspoor, 2017), similar learners, even identical twins, in a similar context will take different developmental paths (Chan et al., 2015; Tilma, 2014; Vyatkina, 2012). The advantage of these longitudinal studies is that they show in detail how each individual develops over time; however, they do not lend themselves to generalizations or general trends.

On the other hand, typical group studies cannot be generalized to the individual, as they cannot show the non-linear paths that learners tend to follow. For example, as Verspoor et al. (2011) found, if the trajectories of learners are averaged out, the resulting line does not look like any one learner as the insightful peaks and dips are smoothed away. However, Verspoor, Schmid, and Xu (2012) wanted to explore whether such variability and non-linear behavior could also be detected in a cross-sectional study in which variation caused by known predictors, such as L1, age, aptitude, and task, were controlled for. 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 proficiency level, resulting in groups of texts at five levels from beginner to intermediate. Each of these texts was coded for 64 complexity, accuracy, and fluency 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, 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—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 a bit 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 were the strongest discriminators.

In addition to these rather linear trends in broad measures, there was also clear non-linear development from one level to the next. At different proficiency levels, there were signs of overuse of different constructions. 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. Also 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, in 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).

To summarize, the cross-sectional data suggested that absolute beginners (between levels 1 and 2) are especially busy 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. (2102) suggested that it would be very useful to follow similar learners over time to check whether these patterns indeed occur 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 variability and 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) point out, generalization to the wider population is possible if we find similar developmental paths in 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 the Verspoor et al. (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 not at the same time), and whether the group as a whole shows non-linear patterns in lexical and syntactic development in that one may develop before the other.

A Longitudinal Case Study

In this multiple case longitudinal study, we traced the development in L2 writing of 22 Dutch learners of English over 23 weeks' time. 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 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 had different topics, as it would be impossible to ask these young learners to write about the same topic every week. 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.

Participants

The participants in our study were 22 Dutch learners of English who started secondary school at the onset of data collection. These learners were in 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% 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 at secondary school. Lowie and Verspoor (submitted) show in a regression analysis that in this homogeneous group, neither motivation nor aptitude were predictors in proficiency gains.

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 were 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 5-point scale, with 1 representing the weakest and 5 the strongest piece of writing in terms of overall proficiency. (Note, however, that as these ratings are based on the relative weak and strong samples within a given corpus, the current numbers 1-5 do not match exactly with those in the CS.) The focus was on the complexity and the fluency of the written productions. 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 1 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 .78. Then the holistic score for each text was calculated as the average of three ratings.

Lowie and Verspoor (submitted) checked for a possible effect of topic on the writing quality by calculating average ratings for all topics. This evaluation showed a gradual increase in the effect of 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 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 towards 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).

For the current study, syntactic development 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).

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. Basically, the model tests with 5000 randomized iterations 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%, 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 generalized additive modeling (Wood, 2006, 2017) as our analysis method. This approach allowed us to assess potentially non-linear patterns over time, while simultaneously taking into account individual (non-linear) patterns over time (cf. mixed-effects regression as outlined in Winter and Wieling, 2016).

Results

Proficiency Gains

Lowie and Verspoor (submitted) carried out a group analysis of the mean rating of the first two writing samples and the last two writing samples. This analysis showed that the group of learners had significantly higher average holistic scores at the end. 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.

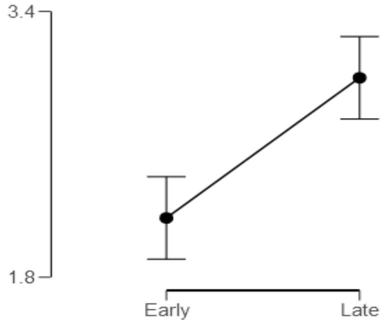

Figure 1. Average first two (Early) – average last two (Late) measurements of the writing samples of the 20 learners (from Lowie and Verspoor, submitted)

MLTU

Figure 2 shows the individual patterns of the 22 learners over time of their MLTU per text. The final graph shows the general trend by means of a GAM. Visual inspection shows that no learner seems to do the same; some learners show peaks early on (learner 4), some in the middle (learner 8), some towards the end (learner 21), and some no real single peaks (learner 13). All peaks were tested for significance for each learner, and for the MLTU learner 12 showed a trend (p = > 0.07) and learner 17 a significant peak (p = > 0.05). The GAM plot shows a general upward trend for the group, with a decrease at the very end.

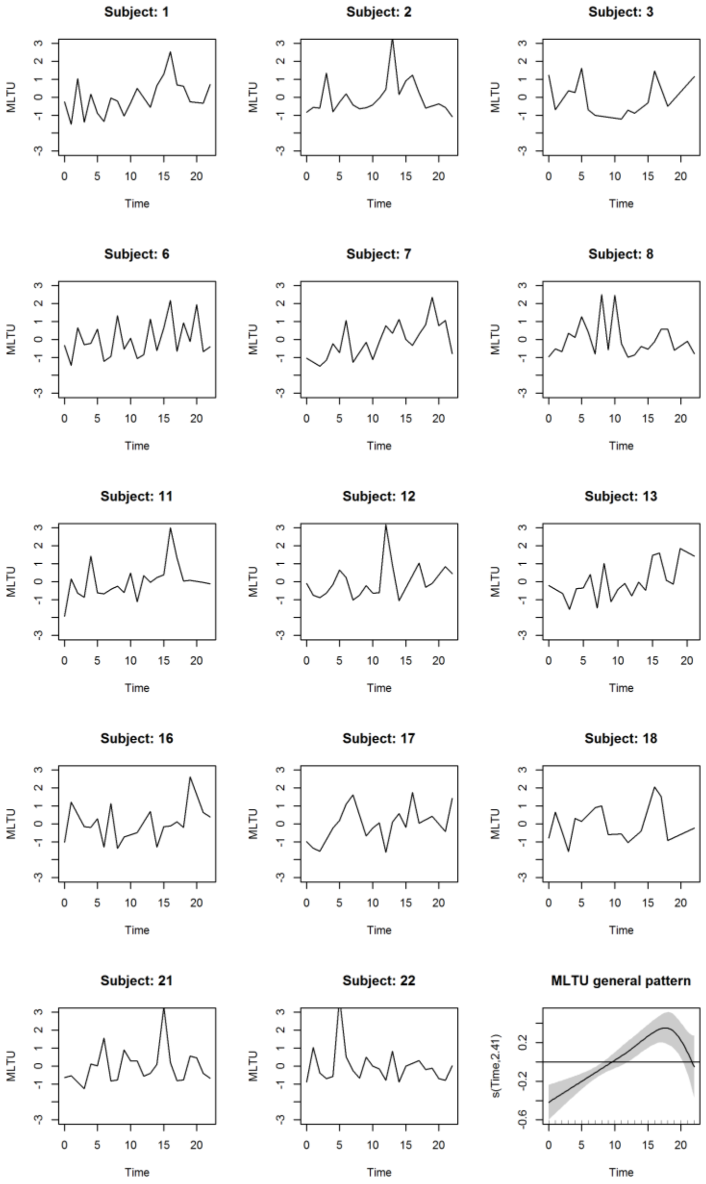

Figure 2: MLTU individuals GAM

Guiraud

Figure 3 shows the individual patterns of the 22 learners over time of their Guiraud per text. The final graph shows the general trend by means of a GAM. Much like the MLTU graphs, visual inspection of the Guiraud patterns shows that no learner seems to do the same. Some learners show peaks or dips early on (learner 2), some in the middle (learner 9), some towards the end (learner 17), and some did not show single peaks (learner 1). All peaks were tested for significance for each learner. The GAM plot shows a general upward trend for the group, with a decrease at the very end.

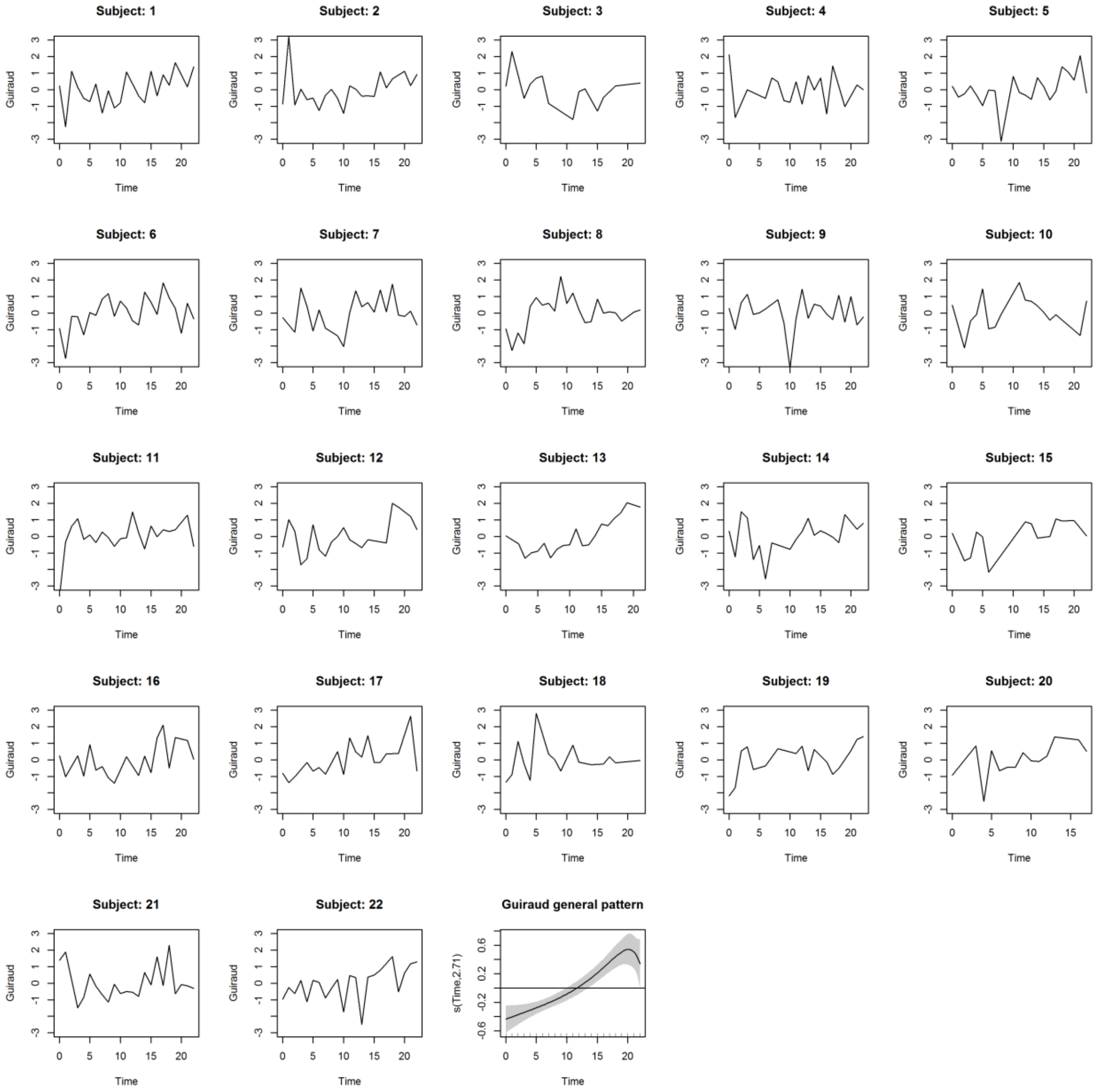

Figure 3: Guiraud GAM and individuals

MLTU versus Guiraud

Figure 4 compares the MLTU (on the left) and the Guiraud GAMS (in the middle), as already shown in Figures 2 and 3. They show a comparable pattern with an increase at first, which reduces at the end. However, the right graph shows that while the Guiraud measure is initially below the MLTU measure, this pattern inverts at the end.

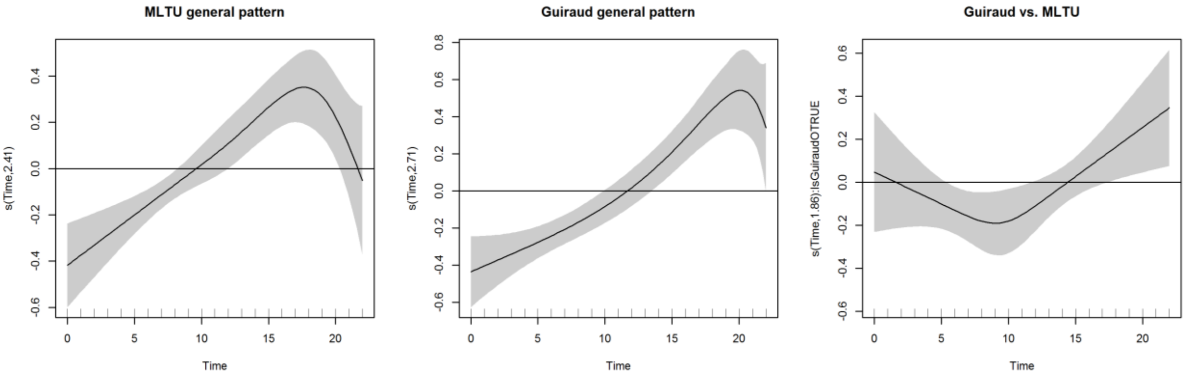

Figure 4. The three graphs show the general effects over time for the MLTU measure (left; p < 0.001), the Guiraud measure (middle; p < 0.001) and their difference (right; p = 0.02). Note that the measures were standardized for each subject (i.e. z-transformed).

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 variable data with a great deal of variation among learners. The learners in the cross-sectional study and the current multiple case longitudinal studies 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 in holistic scores but also for MLTU and 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 per week English exposure at school.

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 3 (near) significant developmental peaks were found, two for the MLTU and one for the Guiraud. Visual inspection of the individual graphs did not show any patterns that the learners might have in common.

However, in the current study, we also wanted to see if we could detect common non-linear 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 (2012) 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 non-linear 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 4, 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 two and three (approximately the level that most of our learners started with), there was a change in mainly syntactic variables after which lexical changes would take place, pointing to different growth patterns for different components or sub-systems 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 proficiency development different subsystems may develop differentially. This means that our metaphor for an index of development may indeed need to change from the static yardstick one 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 a lot messier and creative than we might have been able to imagine.

References

Bulté, B. (2013). *The development of complexity in second language acquisition: A dynamic systems approach (Unpublished PhD Dissertation)*. Brussels: 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.* (pp. 207-230). Rowley, MA: Newbury House.

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).

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

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*, 195-202.

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (*Doctoral Dissertation*).

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

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

Lowie and Verspoor (submitted) A dynamic and longitudinal perspective on Individual Differences.

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* (pp. 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

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* (pp. 215-242). (Second language acquisition; Vol. 109). Bristol ; Tonawanda, NY ; North York, Ontario : Multilingual Matters Ltd.

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.

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 PhD, Rijksuniversiteit Groningen/ University of Jÿvaskyla, Groningen/ Jÿvaskyla.

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

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(14-1).

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, & complexity* (No. 17). University of Hawaii Press.

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

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.