東京外国語大学大学院　言語・地域文化研究　第 29 号　2023 年
Language, Area and Culture Studies No.29, 2023

Article

A longitudinal study of fluency and complexity of English spoken interaction by two high school Japanese students

Maksim Tikhonenko

二名の日本人高校生の英語スピーキング「やりとり」における流暢性・複雑性の縦断的な研究

チホネンコ・マクシム

要旨

本研究は，英語スピーキング「やりとり」の流暢性と複雑性の縦断的成長について，日本語母語の高校生 2 名を対象に分析する。分析するデータは，遠隔により行われた英会話レッスンを録画したスピーキングデータである。レッスンは高校 1 年次から 3 年次にかけて毎月 1 回，講師と 30 分間 1 対 1 で行われた。学習者は事前に配布されている教材を用い，レッスンの準備として英作文を執筆した上で，対話に参加した。さらに，最後のレッスンである 20 カ月目から 3 カ月経った 2020 年 9 月に，British Council による Aptis スピーキングテストを受験したところ，そのスコアは 2 名とも，50 点満点中 33 点で，CEFR 評価基準では，B1 レベルであった。

流暢性と複雑性の計測については，ELAN ソフトウェアを用いて、スピーキング音声を文字化したのち，発話の長さとポーズの長さを計測した。その上で，Foster, P., A. Tonkyn, G.Wigglesworth (2000) に基づき，テキストを AS-unit 単位に分割した。本稿では 1，5，10，15，20 課のレッスンのデータを分析した。

流暢性の分析の基準としては，以下の指標を使用した。

1) スピーチ・レート 語数 / 総計時間）
2) 発音レート（語数 / 発話時間）
3) ポーズの割合 総計ポーズ時間 / 発話時間）
4) ポーズの平均数 ポーズ数 / 総計時間）
5) ポーズの平均時間
6) フィラーの割合 フィラー数 / 語数）
7) 繰り返しの割合 繰り返し / 語数）
8) 言い直しの割合 言い直し / 語数）

複雑性の指標は以下の 2 つである。

9) 従属節の平均数（従属節数 /AS-unit 数）
10) 平均発話時間

分析結果として、複雑性については、顕著な成長がみられた。一方で，流暢性においては振れ幅が大きく、1 分当たりのスピーチ･レート，ポーズの割合とポーズの平均時間が、レッスンによって変動しており，向上したとは言いがたい。これは、レッスンのテーマや講師の教え方が影響しているのではないかと考えられるが、複雑性において顕著な成長がみられたことが研究成果として特筆される。

Keywords: second language acquisition, fluency, complexity, longitudinal analysis

1. Introduction

Oral fluency is considered one of the clearest indicators of overall second language proficiency and is commonly regarded as one of the major learning goals (Tavakoli, 2020). As the learning process takes place over time, it is important to examine how indicators of fluency change over time for a better understanding of the relation of oral fluency to overall language proficiency. For a long time, fluency was conceptualized in broad and narrow senses (Lennon, 1990). In the broad sense, fluency stands for overall language proficiency. In the narrow sense, it refers to temporal features of speech that determine how smoothly and rapidly a learner is able to use L2.

Skehan (2003) divides fluency into three dimensions: speed fluency, breakdown fluency, and repair fluency. Speed fluency refers to the rapidity of speech, breakdown fluency refers to the smoothness of speech related to pausing phenomena, and repair fluency is related to the effectiveness of repair strategies used by learners in real-time. Segalowitz (2010) studied fluency from a different perspective and divided it into three notions: cognitive fluency, utterance fluency, and perceived fluency. Cognitive fluency is "the efficiency of operation of the underlying processes responsible for the production of utterances," utterance fluency is "the features of utterances that reflect the speaker's cognitive fluency," and perceived fluency is "the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency". While cognitive fluency is difficult to observe and measure numerically, utterance fluency is easily measured using numerical variables, and most studies focus on this dimension of fluency. However, there is an ongoing discussion on the questions of which factors reflect learners' fluency, and how exactly fluency measures change over time.

This paper is an attempt to clarify these matters using longitudinal data of spoken English by Japanese high school learners. In its first half, it provides theoretical background and primarily draws on studies on fluency and complexity in second language acquisition, especially those featuring longitudinal analyses.

In its second half, this paper provides methodology and results of a preliminary analysis of a small part of English-speaking performance data by two Japanese high school students. The data was taken from a learners' corpus of English that consists of recordings of online spoken English lessons that had continued once a month for 20 months. In the final section, we reflect on the shortcomings of this analysis and possible future developments as well as implications for applying the results of this research in the fields of second language acquisition and evaluation.

2. Literature review

2. 1. Fluency

There are competing points of view on the question of what fluency measures are more characteristic of fluency.

Lennon (1990) examined 12 features of oral performance, pruned and unpruned speech rate, repetitions, self-corrections and filled pauses, percentage of repeated and self-corrected words as a function of unpruned words, ratios of filled and unfilled pauses to a total speaking time, mean length of speech runs between pauses, and three measures relating T-units and the combination of filled and unfilled pauses (percent of T-units followed by a pause; percent of total pause time at all T-unit boundaries; mean pause time at T-unit boundaries). All participants improved over time in terms of speech rate and exhibited a decreased number

of filled pauses per T-Unit. However, Lennon argues that speech rate differences reflected pause time differences. In contrast, self-corrections appeared not to be a good fluency indicator, and Lennon suggested that in the development of L2 fluency it is important to increase one's ability to self-correct in real-time. These results suggest that speech rate and pausing phenomena by themselves are not critical indicators of fluency (Segalowitz, 2010). Kahng's research (2014) mentions that higher-proficiency speakers tend to be better at repairing disfluencies, while lower-proficiency speakers tend to fail at lexico-grammatical repair and consequently abandon the topic.

Cucchiarini et al. (2002) suggest that speech rate and pause frequency are the most important factors in read-aloud speech fluency perception. Riggenbach (1991) also concluded that the central elements of foreign language (L2) fluency are pausing, speech rate, and repairs.

Towell et al. (1996) report that in their study, learners who had spent time in the country of their L2 increased their articulation rate and had longer runs between pauses. However, there was no significant change in phonation/time ratio, and mean pause length did not change over time.

De Jong et al. (2009) obtained automated measures of oral fluency and compared participants' performances in both L1 and L2. Only the number of silent pauses per word did not yield a significant L1-L2 correlations. On the other hand, there was strong evidence that hesitations and speech rate may be characteristic of the way individuals speak in general. De Jong et al. (2009) suggests that the oral variables best reflecting L2 fluency are percentage of silent pauses per word (but not length), words per second speech rate, and percentage of corrections or self-repairs per word. Also, some studies consider breakdown and repair fluency as inseparable phenomena (Williams & Korko, 2019).

Some recent studies investigated pause locations as an important factor in L2 fluency. While pauses in the middle of utterances are believed to reflect disruptions in L2-specific linguistic processing, pauses at clausal boundaries capture breakdowns in conceptualization-related processes (De Jong, 2016; Tavakoli, 2011). Yan et al. (2020) found that whereas L2 speakers at all proficiency levels paused frequently, the pauses produced by higher-proficiency speakers tended to occur at syntactic junctures more often than did pauses produced by lower-proficiency speakers.

2. 2. Complexity

According to Ellis and Barkhuizen (2005), complexity measures are grouped according to the aspect of the language they relate to: 1) interactional, 2) propositional, 3) functional, 4) grammatical, and 5) lexical. They created a reference list of all complexity measures used in previous studies and describe the context of their usage. Interactional complexity includes two measures: the number of turns and mean turn length. While the former may distort the results of analysis if a speaker uses a lot of short sentences, the latter is considered a good metric for interactional complexity. Propositional complexity is represented by the number of idea units encoded. This measure works best when the elicitation task requires learners to communicate re-specified content. Functional complexity is represented by one measure: the frequency of some specific language function. Grammatical complexity is the most studied dimension of complexity, and Ellis and Barkhuizen (2005) describe three measures of it: the amount of subordination, use of some specific linguistic feature, and the mean number of verb arguments. Among them, the use of specific linguistic features is better for analyzing the complexity of students at the elementary level, while the amount of subordination serves

as an effective indicator of complexity for learners who have acquired some of the various subordinating devices. Finally, lexical complexity is defined by such variables as the type-token ratio, which is the total number of different words used divided by the total number of words in the text.

2. 3. Longitudinal studies of CAF

When it comes to longitudinal studies of complexity, accuracy, and fluency (referred to as CAF from now on), their number, especially in the context of Japanese learners of English, is limited. Researchers agree that changes in pause length are usually the most significant. It is also evident that students at different levels of L2 proficiency develop different skills. Factors that influence CAF development also include educational environment, namely differences in the medium of education (online/offline) and environment (classrooms/overseas immersion programs)

Hanzawa (2021) studied L2 fluency development in Japanese learners of English who studied English at a university during the course of 1 year. While within-clause pause duration significantly declined during both semesters, it took an entire academic year to observe significant improvements in between-clause pause length. Virtually no development was observed in articulation rate, between- and within-clause pause frequency, or repair frequency over time.

García-Amaya (2015) conducted a longitudinal analysis of speech rate and the use of filled pauses (FPs) and unfilled or silent pauses (SPs) in the oral production of L2 learners of Spanish in an intensive overseas immersion (OIM) program (6 weeks) and a 15-week L2 "at-home" classroom (AH). The results show a significant increase in the rate of speech over time in the OIM group compared to the AH group. Additionally, the OIM learners show greater use of "disfluencies" over time, namely filled pauses and short silent pauses.

Another study by Maeda (2021) was not longitudinal, but it compared characteristic differences between different CEFR levels. Speaking data from TEAP tests taken by 153 Japanese high school learners of English was analyzed, and the following differences were found. Between below-A2 and A2 levels, speed fluency significantly increased, as well as the number of disfluency markers. Between A2 and B1 increased syntactical complexity and interactional effectiveness were observed. Finally, between B1 and B2 accuracy and lexical complexity increased, while the number of pauses decreased.

2. 4. Studies of interrelations between CAF components

There is no agreement among researchers on the exact model of CAF intercomponent relations. According to Housen et al. (2012) CAF interact in intricate ways and this interaction is sometimes mutually supportive and sometimes competitive. Researchers who believe that human attention and processing capacity are limited see fluency as an aspect of L2 production which competes for attentional resources with accuracy, while accuracy in turn competes with complexity. A rival view is proposed by Robinson (2001, 2003) who claims that learners can simultaneously access multiple and non-competitional attentional pools so that, depending on the conditions imposed by the task, all three components may in principle either jointly increase or decrease in L2 performance.

Housen et al. (2012) propose a scenario where the internalization of more complex L2 structures leads to greater complexity, followed by the modification of the internalized structures (leading to greater accuracy) and, finally, the development of performance control over and consolidation of the acquired structures (result-

ing in more fluent L2 performance)

Yan et al. (2020) divide fluency features into holistic (speech rate, articulation rate, number of silent pauses) and fine-grained (mean length of run, juncture pause rate, repair success rate) features and argue that they represent two different fluency dimensions, among which fine-grained fluency measures, measured manually, are more characteristic of fluency.

3. Research Questions

This paper will analyze in terms of the following three research questions:

a) Which fluency and complexity variables change over 20 months of spoken English lessons, and how do they change at different stages of the development of fluency?

b) How are fluency and complexity variables related to each other in the long term?

c) What factors are most characteristic of fluency and complexity in the long term?

4. Methodology

4. 1. Description of the project

This paper investigates changes in the fluency and complexity of English performance by high school learners.

A research team from Tokyo University of Foreign Studies organized an educational project in collaboration with Sankei Human Learning and two prefectural high schools. High school students took classes in spoken English and had free conversations with teachers from the Philippines using Zoom. Each lesson lasted about 25 minutes. 32 students started the lessons in the first year, November 2018, and completed the 20th lesson in the 3rd year, July 2020.

Students used a textbook developed by researchers affiliated with Mochizuki laboratory at Tokyo University of Foreign Studies. Topics featured in each lesson are provided in the following list, and topics of lessons analyzed in this paper are given in bold and underlined.

1) **Introduce the Meaning of Your Name in English**
2) Talk about Japanese Food and Your Favorite Food
3) Talk about Extracurricular Activities
4) Introduce Your Hometown to Foreigners
5) Going on a Trip to Nagano
6) Introduce Japanese Shrines and Temples
7) Going to Parties
8) Getting Sick
9) Smartphones: “Convenient”, or “Weakening Our Brains”?
10) Experience of a Homestay
11) Social Issues and Viewpoints of Minorities
12) Talk About Your Future Path
13) Guiding Foreign Tourists around Your Region or Local Area
14) Diet and Health

15) Studying Abroad and Going on an Internship

16) Study Abroad in Japan

17) The Difficulties of Japanese

18) Coronavirus and Social Changes

19) How to get into a university

20) How to Coexist with AI

Lessons varied in their difficulty levels, which might have affected the performances of students. The first important factor that contributes to the difficulty level is the percentage of high CEFR-level vocabulary in each lesson. The second factor is how familiar each student was with a topic of a given lesson. For example, in general, students reported that they struggled with lessons such as lesson 6 (Introduce Japanese Shrines and Temples) or 11 (Social Issues and Viewpoints of Minorities), which contained sophisticated vocabulary that did not match with students' levels; topics such as number 10 (Experience of a Homestay) were reported to be difficult for those students who had not had such experience and therefore found it difficult to discuss.

Information on students' backgrounds, motivations, and goals was collected and put into students' profiles. A textbook created by the research team was used during lessons and contained topics from level A1 (self-introduction) to B1 (A.I.).

Before each lesson, students read materials distributed via Moodle and wrote short essays on the topic of the lesson. Each lesson lasted 25 minutes, during which students read texts, learned key expressions, and, more importantly, had free conversations with teachers.

After each lesson students filled in surveys providing information about their impressions. Also, teachers provided feedback on the students' performance, and native speakers of English provided feedback on essays written by students. Finally, students were able to watch supplementary videos about vocabulary and grammar created by the team members.

Spoken data was recorded and a learner corpus was created.

Three months after the project ended, all students took the Aptis speaking test. Their oral performance was recorded and added to the corpus.

4. 2. Learner Corpus of Japanese High School Learners of English

Videos of lessons were transcribed using ELAN software by trained undergraduate research assistants, and the length of utterances was measured. Transcription only included utterances and fillers from free conversation parts. Lessons included multiple sections in which students were reading the textbook, repeated big chunks of words or sentences after a teacher, or practiced pronunciation. Such sections could easily distort the results of the research and therefore were excluded from the corpus. There were also special cases of ambiguous sections, mostly featuring answers to questions in the Preparation section, in which some students were reading prepared texts, while others were speaking freely and used prepared texts only for reference. In such cases, it is sometimes difficult to judge whether a student reads a text or speaks freely, so in general assistants followed the following rules while transcribing texts: if a student used her or his notes as a basis for the answer for more than 50% of the time, this section would not be considered free conversation. However, if a student occasionally looks at a note, but mostly gazes in other directions while producing utterances,

these utterances would still be considered free conversation. Nevertheless, each case of such nature required subjective evaluation.

Along with utterances, silent pauses longer than 500 ms were also marked and measured for length. Although in most recent L2 studies pause threshold of 0.25 ms or even 0.20 ms are used (Tavakoli, 2020), the threshold of 500 ms was proposed by Foster et al (2000), and it is a realistic threshold to use in manual tagging of pauses in large amounts of spoken data.

4. 3. Participants

Two students were selected as subjects of analysis of fluency and complexity according to the following criteria:

1) students achieved an Aptis speaking score of 33, which is the average score for the whole participant group. It corresponds to the B1 level of language proficiency in CEFR.
2) students had consistently submitted homework and stayed motivated about participating in the project over a long time.
3) both students applied for a follow-up study and are taking an additional course in spoken English, which will help to continue researching their performances on a longer scale.

4. 4. Measures of fluency and complexity

To measure speed fluency, two measures were used: **pruned speech rate** and **articulation rate**. These measures are among the most used in CAF studies (Tavakoli, 2020) and are easy to measure manually.

Different studies use different definitions of speech rate and articulation rate, often making speech rate a function of the number of syllables. However, since this study only used manually calculated measures, it was decided to measure both as a function of the number of words. So, in this paper, the **pruned speech rate** is the overall number of pruned (i.e. without fillers, self-corrections, and repetitions) words divided by total time (speaking time including pauses). The **articulation rate** is the number of pruned words divided by phonation time (speaking time excluding pauses).

For measuring breakdown fluency, four variables were used: **pause time ratio**, **mean pause length**, **number of pauses per minute**, and **frequency of filled pauses**. Pause time ratio is a percentage of time spent on pausing, and it is calculated as total pause time divided by total time. Mean pause length is calculated as total pause time divided by the number of pauses. The number of pauses per minute is the total number of pauses divided by the total time. Finally, the frequency of filled pauses is the number of fillers divided by the overall number of unpruned words.

Variables of repair fluency used in this paper are the following: **frequency of partial or complete repetitions** (the number of repetitions divided by the number of unpruned words), and **frequency of self-corrections** (the number of self-corrections divided by the number of unpruned words). As Tavakoli (2020) mentions, repair fluency measures require working with unpruned data, so unlike speed fluency measures, this category of variables works with number of unpruned words.

Finally, to measure syntactical complexity, two variables were used: **mean AS-unit length** (number of words divided by the number of AS-units) and **subordinate clause ratio** (the number of clauses divided by the number of AS-units. Mean AS-unit length (or mean utterance length) is widely used in measurements of

syntactic complexity as a general measure that tap global, overarching complexity constructs (Bulté, Housen, 2012). As a measure of subordination, we chose the ratio of subordinate clauses to the number of AS-units, which is considered an effective indicator of complexity for those learners who have acquired some of the various subordinating devices (Ellis, 2005).

In studies of L2 complexity, it is also quite common to use grammatical measures of complexity such as the use of some specific linguistic features (e.g. auxiliary verbs or conjunctions). However, in the data samples used in this research, the number of such features was not especially large to conduct a proper statistical analysis.

This paper uses AS-units as a unit of spoken speech, and this notion will be discussed in the next section.

4. 5. AS-units

An AS-unit is a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either (Foster et al. 2000, p. 365). This unit inherits characteristics of speech including its syntactic, semantic, and prosodic features. It has proved to be a suitable and authentic unit of analysis for spoken data, and most recent L2 fluency and complexity studies use it in their analysis (Tavakoli, 2020, p. 49).

An independent clause is any clause that includes a finite verb: 'Turn left', 'I take a different way'

A sub-clausal unit doesn't include a verb but can be elaborated to a full clause: 'Oh, poor woman', 'Yes'.

Foster et al. (2000) propose several levels of application of AS-units. Level 1 is used for the full analysis of data, and it includes everything except untranscribable data. Level 2 is used with highly interactional data, which can yield a high proportion of minimal units, whose inclusion can distort the perception of the performance. On this level, one-word utterances (yeah, okay, right) are excluded from analysis as well as echo responses:

(1) A: 'I think two years'
B: 'Two years'

Finally, level 3 is for use in special cases where analysis of non-fragmentary AS-units is required. It is used by researchers who need to look at what the performer can do in the production of relatively 'complete' units. Along with units excluded at level 2, it also excludes the next units: a) verbless elliptical AS-units that ellipses elements of interlocutor speech:

(2) A: What is your mother tongue then?
B: Arabic Arabic.

b) AS-units involving substituting of clause, predicate:

(3) 'yes, I think so';

and c) one or two-word greetings and closures.

This paper deals with highly interactional data and students produced a lot of short utterances and echo responses, especially at earlier stages of the project, and that could distort the results of the analysis. Therefore, it was decided that level 2 would be used for this research. Additionally, two-word greetings were also

excluded from the analysis, as their production stayed the same at all stages of the project and did not reflect the difference in proficiency between different lessons.

The most difficult cases of defining AS-units in speakers' performances were cases of long monologues with numerous pauses. In such cases, sometimes it was difficult to define whether an utterance after a pause longer than 500 ms is a continuation of the previous statement and a part of a complex sentence, or simply a new AS-unit. In such cases, we considered not only the syntactic structure of given utterances but also intonational patterns. If a turn before a pause featured rising intonation, the next turn was considered a continuation of an AS-unit and a complex sentence.

4. 6. Analysis of data and its conventions

Because of time limitations, it was not possible to analyze each student's data from all 20 lessons, so it was decided to look at differences in performance at certain points in time. Therefore, this paper analyzes performances in lessons number 1, 5, 10, 15, and 20.

In order to calculate the variables of fluency and complexity, transcribed texts together with lengths of utterances and pauses were transferred to Microsoft Excel worksheets and divided into AS-units. Then, the following variables were counted manually: number of AS-units, number of pruned words (excluding fillers, self-corrections, and repetitions), number of fillers, number of repetitions, number of self-corrections and false starts, number of native language intrusions, number of clauses.

A filler or a filled pause is a non-semantic word that is used to fill gaps in speech. Words like 'Ah', 'umm', 'えっと', and laughter would be considered one filler when counting.

A word or phrase that is repeated the same way is considered a repetition, and every word in that phrase is counted as one repetition. Partial repetitions of single words are also counted as repetitions:

(4)　J007: **{Ame…}** America (**1 repetition**, 1 word).

However, if repetition is clearly a part of stylistic choice by a speaker and not a disfluency, it will be counted as a separate word:

(5)　J007: {ah}, **sorry sorry** (1 AS-unit, **2 words**, 1 filler)

Any phrase in which any element is later corrected would be considered a self-correction, and each word in that phrase is counted as one self-correction:

(6)　J007: {**I went to** ah} I have been to {mm} Korea (**3 self-corrections**, 5 words, 2 fillers).

False starts would also be counted as self-corrections.

The definition of a clause is a difficult task, especially in highly interactive discourse. This paper bases its definition of a clause on Foster's (2000) definition. As was noted earlier, an independent clause is minimally a clause including a finite verb. At the same time, a subordinate clause consists minimally of a finite verb or non-finite verb element plus at least one other clause element (subject, object, complement, or adverbial). This leads us to some interesting applications.

In other words, if an AS-unit does not include a verb, it will be a clauseless AS-unit:

(7)　J007 (20th lesson): easily (1 AS-unit, **0 clauses**)

However, if a verb is present, an AS-unit will consist of at least one clause however short it is:

(8) J007 (20th lesson): {um}, I'm fine (1 AS-unit, 1 clause)

It was concluded that this way of application reflects the complexity of students' English at the beginning stages of the project better: although clauseless AS-units were observed at all stages of the project, they made up a large percentage of all AS-units in lessons number 1 and 5. However, after getting some experience in speaking, students increased the number of verbs even in the shortest AS-units.

Following Foster et al. (2000), auxiliary verbs such as "have" in the present perfect tense, or modal verbs such as "can" and "may" were not considered a base for a distinct clause, so AS-units such as (9) contain only 1 clause:

(9) J007 (20th lesson): so I **can communicate** with (+) them (1 AS-unit, **1 clause**)

However, if an AS-unit contains two normal verbs connected using "to", and the second one has at least one clause element, the AS-unit is thus divided into two clauses:

(10) J007 (20th lesson): uh I **want :: to learn** Germans (1 AS-unit, **2 clauses**)

After counting the variables, Microsoft Excel formulas were used to calculate the measures defying fluency and complexity and compared between different times and students. The results are discussed in the next section.

5. Results

The results of the analysis of English-speaking performance by two students are divided into respective sections. We transformed the numbers of lessons into pure numerical values and measured the correlation coefficients of all variables calculated in this study. However, due to the small number of data points, statistical analysis can be inaccurate and distorted and requires careful qualitative analysis. Correlation coefficients were calculated using Microsoft Excel.

The results of statistical analysis are illustrated in the tables in each section.

5. 1. Speed fluency

Table 1

Results of the analysis of speed fluency

	J003		J007	
	Speech rate (words/min)	Articulation rate (words/min)	Speech rate (words/min)	Articulation rate (words/min)
1st month	62.41	85.75	48.06	61.95
5th month	99.68	144.76	54.27	78.46
10th month	32.03	67.16	36.77	56.85
15th month	54.72	88.07	53.31	67.38
20th month	51.23	77.31	53.91	64.77

It is visible in the tables that changes in the speed fluency were not linear, however, both participants showed similar trends. Relatively high pruned speech rate and articulation rate were observed in the first month, and they also increased in the 5th month of the project. It is also worth noting that changes in articulation rate during this period were slightly bigger in both students.

However, in the 10th month, both variables significantly decreased in both students, after which they increased again in the 15th month. In the case of student J003, the speech rates in the analyzed lessons never reached the same levels as in the 1st month, whilst student J007 was able to reach the same levels. Finally, the results in the 15th and 20th months did not differ significantly.

The following table 2 illustrates the results of the statistical analysis of linear correlations between variables. Significant coefficients are in bold in all tables containing the results of statistical analysis.

Table 2

Correlation coefficients of analysis of speed fluency

	Number of a lesson	Speech rate J003	Articulation rate J003	Speech rate J007	Articulation rate J007
Number of a lesson	1				
Speech rate J003	-0.44	1			
Articulation rate J007	-0.41	**0.97**	1		
Speech rate J007	0.23	**0.65**	0.55	1	
Articulation rate J007	-0.13	0.93	**0.95**	**0.76**	1

According to Garcia-Amaya (2015), the rate of speech increased linearly for overseas immersion program learners over time and remained stable for ‘at-home’ foreign language classroom students.

However, the learning curve of the two students we observed is rather a sigmoid curve, so the results of the current analysis are different from both scenarios described in the paper by Garcia-Amaya, and therefore, statistical correlations between the number of a lesson and other variables should be calculated as a multivariable function, whilst linear correlation coefficient may not be a very good statistical tool for analyzing speed fluency in this research.

Despite that, it is important to note two facts. Firstly, the speech rate and articulation rate of each student show a very strong and strong correlation between each other (0.97 for J003 and 0.76 for J007). Also, the speech rates of J003 and J007 show a strong positive correlation (0.65), while the articulation rates of J003 and J007 show a very strong correlation (0.95), which proves that these variables developed similarly in the two students.

5. 2. Breakdown fluency

The following two tables show changes in four variables of breakdown fluency for each student.

Table 3

Results of the analysis of breakdown fluency, student J003

	J003				
	Pause ratio	Mean pause length, s.	Mean pause length, st. deviation	Pauses per minute, pause/min	Fr. of filled pauses, %
1st month	37.39%	1.01	0.51	16.13	22.06%
5th month	45.22%	1.37	0.98	13.64	4.90%
10th month	109.7%	1.88	1.96	16.7	24.75%
15th month	60.96%	1.15	0.72	19.82	13.30%
20th month	50.9%	1.03	0.46	19.65	13.04%

Table 4

Results of the analysis of breakdown fluency, student J007

	J007				
	Pause ratio	Mean pause length, s.	Mean pause length, st. deviation	Pauses per minute, pause/min	Fr. of filled pauses, %
1st month	28.92%	1.35	0.78	9.94	25.74%
5th month	44.56%	1.19	0.72	15.57	23.90%
10th month	54.61%	1.55	0.78	13.67	26.97%
15th month	26.38%	0.89	0.54	14.02	25.50%
20th month	20.14%	0.89	0.4	11.36	27.20%

Two out of four breakdown fluency measures as well as in the case of speed fluency showed a similar pattern; however, it is quite different from what we observed in the speed fluency.

In both cases pause ratio and mean length of pauses were higher in the 10th month. However, after the 10th month in lessons 15 and 20, both measures showed decreases, meaning that overall breakdown fluency increased during these lessons. Furthermore, similarly to the situation with speed fluency, the differences in performance between the 15th and the 20th months were less significant.

On the other hand, changes in the number of pauses per minute did not show any recognizable pattern, and in the case of student J003, they were quite minor.

The frequency of fillers barely changed in student J007 and changed unpredictably in student J003, which shows that this parameter is not a good predictor of fluency.

Table 5

Correlation coefficients of analysis of breakdown fluency

	Num. of a lesson	Pause ratio, J003	Mean pause length, J003	Num. of pauses per min., J003	Fr. of fillers, J003	Pause ratio, J007	Mean pause length, J007	Num. of pauses per min., J007	Fr. of fillers, J007
Num. of a lesson	1								
Pause ratio, J003	0.22	1							
Mean pause length, J003	-0.11	**0.89**	1						
Num. of pauses per min., J003	0.82	0.11	-0.34	1					
Fr. of fillers, J003	-0.17	0.54	0.325	0.17	1				
Pause ratio, J007	-0.43	**0.67**	0.93	-0.65	0.21	1			
Mean pause length, J007	-0.67	0.51	**0.69**	-0.62	0.63	**0.8**	1		
Num. of pauses per min., J007	0.05	0.31	0.54	-0.32	-0.55	0.56	-0.03	1	
Fr. of fillers, J007	0.56	0.49	0.12	0.65	0.66	-0.17	0.06	-0.55	1

As measures of pausing phenomena did not change linearly as well as in the case of speed fluency, it may be futile to measure correlations between the number of a lesson and the values of each measure. However, analysis of correlations between different measures in each student, and between the same measures of different students reveal some useful insights. First, the mean pause length and pause ratio show very strong correlations in the data of both students (0.89 and 0.8 for J003 and J007 respectively). Secondly, pause ratio and mean pause length showed strong correlations between both students' values (0.67 and 0.69 respectively). There were also other strong correlation coefficients (0.65 between mean student J003's number of pauses and J007's frequency of fillers; -0.65 between J007's pause ratio and J003's number of pauses per minute; 0.63 between J003's frequency of fillers and J007's mean pause length), which are difficult to explain. They may be distortions due to the small number of data points, so in the future, these correlations should be examined with more data available.

5. 3. Repair fluency

The following table illustrates the results of the analysis of repair fluency variables.

Table 6

Results of the analysis of repair fluency

	J003		J007	
	Repetition frequency, %	Self-correction frequency, %	Repetition frequency, %	Self-correction frequency, %
1st month	2.20%	8.08%	5.15%	4.42%
5th month	2.30%	6.56%	0.00%	3.77%
10th month	9.41%	5.94%	6.74%	4.87%
15th month	10.73%	7.94%	6.02%	7.45%
20th month	7.73%	8.21%	7.88%	4.88%

As the table illustrates, the changes in both measures in both cases were insignificant or unpredictable, sometimes the number was not enough to get accurate results, and thus we may conclude that these parameters are not good predictors of fluency, at least between levels A2 and B1.

Table 7

Correlation coefficients of analysis of repair fluency

	Number of a lesson	Repetition fr. J003	Self-correction fr. J003	Repetition fr. J007	Self-correction fr. J007
Number of a lesson	1				
Repetition fr. J003	**0.77**	1			
Self-correction fr. J003	0.28	-0.02	1		
Repetition fr. J007	**0.62**	**0.68**	0.37	1	
Self-correction fr. J007	0.52	0.78	0.36	0.46	1

The table shows some strong correlations, for example between the number of a lesson and repetition frequencies in students J003 and J007 (0.77 and 0.62); between repetition frequencies of both students (0.68), and between repetition frequency of student J003 and self-correction frequency of student J007. While the first three may indicate that repetitions showed some signs of change in relation to the number of lessons, the last one is difficult to explain qualitatively. All relations need to be examined with a broader data set.

5. 4. Syntactic complexity

The next graph provides information about the changes in syntactic complexity over the course of the project.

Table 8

Results of the analysis of syntactic complexity

	J003			J007		
	Subordinate clause rate, clause/ AS-unit	Mean AS-unit length, words	Mean AS-unit length, SD	Subordinate clause rate, clause/ AS-unit	Mean AS-unit length, words	Mean AS-unit length, SD
1st month	0.81	4.24	2.17	0.74	3.78	2.46
5th month	0.83	4.87	2.47	0.56	3.38	2.5
10th month	1.3	5.09	3.04	0.72	4.2	3.78
15th month	1.15	6.64	3.96	1.31	6.08	3.4
20th month	1.37	7.68	4.92	1.34	6.36	3.64

The results allow predicting that there may be a trend for an increase of both complexity measures through time despite occasional decreases in some months (subordinate clause rate in student J003 was lower in the 15th month compared to the 10th month; both variables slightly decreased in student J007 in the 5th month in comparison to the 1st month). In the 20th month, both students had complexity measures 1.5-2 times larger than in the 1st month, which apparently indicates that complexity has significantly improved over the 20 months.

Table 9

Correlation coefficients of analysis of syntactical complexity

	Number of a lesson	Subordinate clause ratio J003	Mean AS-unit length J003	Subordinate clause ratio J007	Mean AS-unit length J007
Number of a lesson	1				
Subordinate clause ratio J003	**0.87**	1			
Mean AS-unit length J003	**0.97**	**0.75**	1		
Subordinate clause ratio J007	**0.86**	**0.64**	0.91	1	
Mean AS-unit length J007	**0.92**	0.73	**0.95**	0.99	1

As both variables changed linearly, correlation coefficients may be useful to measure in relation to the number of a lesson. Both variables show very high levels of correlation with the number of a lesson. Each variable also shows high and very high correlations between the two students' data.

6. Discussion

After conducting a longitudinal analysis of fluency and complexity measures, we can now attempt to explain possible reasons for such patterns in the long-term development of fluency and complexity.

It is necessary to note, however, that due to a small number of data and sample points the results of this analysis by no means can be generalized to the whole group of participants at the current stage. Also, currently, it is impossible to take into consideration some possible external factors (such as the influence of lessons outside the project) that might have influenced the results, so we cannot state that changes in speaking performance that we observed are a result of taking this course in particular.

The first notable thing is the fact that only complexity increased significantly and in a more or less linear way. This is also accompanied by the fact that with each month lessons were getting more difficult, so the increase of complexity took place despite and together with the increasing difficulty of tasks, although, unfortunately, we do not have a metric to measure it.

The next notable observation is that the development of those fluency measures that changed significantly and followed apparent patterns took place in stages and was not linear. Combining it with the development of complexity, we can argue that the two dimensions of CAF were in relation to each other.

The topics of the 1st and 5th lessons were relatively easy ('Introduce the meaning of your name in English' and 'Going on a trip to Nagano' respectively), because they featured words and concepts that were more or less familiar to learners. Also, students could not produce complex long sentences, as both statistical and qualitative analysis shows. Considering these two facts, it is easy to explain the higher speed fluency: apparently, simpler utterances took less time to produce, and the small number of long sentences and low amount of interaction led to less pausing, which is also confirmed by the breakdown fluency measures. In other words, in the 1st lesson, students produced short simple utterances, which they could articulate easily and quickly, thus increasing their speed fluency. Many of such utterances were clauseless AS-units, as can be seen in the next example:

(11) 1st lesson, J007:

Teacher: Which among the seasons that you like best?

J007: {hmm} **winter**

However, as we can see in the results of the breakdown fluency analysis, by the 5th month the amount of interaction between teachers and students increased, which possibly led to the increased amount of pausing. It may be illustrated by the following example:

(12) 5th lesson, J003:

{the} the most impressive travel experience for me is **(pause 4.384 sec)** visiting Hokkaido // **(pause 1.282 sec)** {when I was} **(pause 1.672 sec)** er **(pause 0.897 sec)** when I was 6 years.

Despite the increased amount of pausing, speed fluency measures were significantly higher in the 5th month. It is especially important that the articulation rate increased more significantly than the speech rate because this variable does not depend on amount of pausing, which indicates that it was exactly the speed of articulation, which grew up.

This may imply that the first five months of the longitudinal online course of English was the period of the most extensive growth of fluency, however, this statement needs to be confirmed by analysis of all lessons between the two analyzed in this paper.

Speed fluency in the 10th month experienced a significant drop in comparison to the 5th month. It is possible to argue that the reason might be the continuously growing complexity and increased amount of interaction, which started to require too much processing capacity. In other words, it may be concluded that complexity started increasing at the expense of fluency. However, it is also possible that the reason is that both students in the 10th lesson engaged in conversations too complicated for their respective levels of English, so they could not express themselves and produced large amounts of pausing and slow speech. Also, the topic itself ('Experience the homestay') was quite difficult because high school students had not had such an experience, so it was hard for them to speak about it. It is illustrated by the following example of a section that was hard for a student in the 10th lesson.

(13) J003, 10th lesson:

there are maple oh, なんだろう **(pause 4.48 sec)** maple, oh **(pause 4.195 sec)** eh, that is **(pause 2.51 sec)** on the bread // ohhh **(pause 14.07)** oh, なんだろ **(pause 1.17 sec)** {uh, very} **(pause 1.51 sec)** {sweet} um, very sweet // **(pause 1.66 sec)** uh (pause 1.48 sec) {eh} that is {taken by eh} taken **(pause 1.76 sec)** of **(pause 0.595 sec)** maple tree //

Lessons 10 and 15 are characterized by visible increases in speed fluency and decreases in pause ratio and pause length in comparison with the earlier lessons. This may indicate that students started to feel more comfortable with the increased complexity of performance and interactions. On the other hand, complexity did not stop increasing, which means that the process of language acquisition became more balanced.

However, in the 20th month, the fluency of both participants did not differ much from lesson 15, although there is an increase in complexity. The slight drop in fluency in the 20th lesson in participant J003 may be explained by the overall difficulty of the topic, 'Coexisting with AI', which featured a large number of words defined as belonging to levels B2 and higher in the CEFR framework. In other words, there is a possibility that fluency may have reached a ceiling before lesson 20 or that it depends on the difficulty of tasks, so a more detailed analysis of speaking performance between the 15th and 20th month is required.

These findings are in accordance with Maeda (2000): increases in fluency tend to be characteristic

between A1 and A2 levels, as well as increases in the number of disfluencies, while increases in complexity are more characteristic of the B1 level.

Finally, some measures **did not change**, and therefore cannot be considered good fluency measures in this research. These measures are two measures of breakdown fluency, namely **the number of pauses per minute and frequency of fillers**, and all measures of repair fluency, namely **the frequency of repetitions and self-corrections**. This result matches some other reports stating that frequencies of fillers and hesitations (i.e. repairs and repetitions) are rather individual characteristics of the way individuals speak in general (Segalowitz, 2010; De Jong et al. 2009).

7. Conclusion

After conducting a longitudinal analysis of L2 English performance by two Japanese high school students we may conclude the following.

1) Analysis of speaking data by two students taken from 20 online lessons of spoken English, which were conducted once a month and lasted about 25 minutes each, shows that the students significantly increased their complexity, which manifested in increases of two complexity measures, **subordinate clause rate** and **mean AS-unit length**.
2) On the other hand, fluency measures changed between lessons in a non-linear way, and students experienced increasing and decreasing fluency.
3) Fluency measures in which we observed significant changes were both measures of speed fluency, namely **speech rate**, and **articulation rate**, and two measures of breakdown fluency, namely **pause ratio** and **mean pause length**. Speed rate fluency was high at the beginning, then experienced a dramatic drop in the middle of the project and started increasing again until the end of the project. At the same time, breakdown fluency started to decrease from the very beginning, which may be explained by the increasing amount of interaction. Then, apparently, after students reached some level of control over their L2 in later lessons, it increased.
4) These patterns may be explained by the fact that increasing complexity required much mental effort to process, and complexity increased at the expense of fluency.
5) Finally, some fluency measures appeared not to be good predictors of fluency. These are two breakdown fluency measures related to the number of pauses, namely **the number of pauses per minute** and **frequency of fillers**, and all measures of repair fluency, namely **frequencies of repetitions** and **self-corrections**.

8. Implications for future studies

In this particular study, we met some restrictions, mostly in regard to time and manpower for more elaborate analysis, and could not use some useful methods that are now common in the field of CAF studies.

First and foremost, the dataset used in this paper is too small to conduct a more precise statistical analysis and generalize its results. Therefore, it is important to analyze speakers' performance in all lessons of the project, and also to use data from more students. This will allow us to conduct a more precise statistical analysis, which can help us understand the nature of interrelations between CAF variables better and determine patterns of CAF changes at different levels for students on different levels of CEFR. In addition, multivariable

statistical analysis should be conducted.

The next limitation of this paper is that the analysis of speed fluency mostly deals with measures taken manually and uses words as a basic unit for analysis. However, automatic measuring of speed fluency, especially articulation rate, based on such phonetic units as syllables can bring more insights into the nature of L2 performance and fluency.

Another phenomenon that can be effectively measured using automatic tools is pausing. First, more recent studies use lower thresholds for pauses being as low as 0.2 sec. Due to complications of manual tagging of pauses, it was difficult to measure pauses shorter than 0.5 sec, so using automatic tools is necessary for more effective research. Another important aspect of pausing that we were not able to analyze is pause location. According to different studies (Tavakoli, 2020; Tavakoli, 2011; Kahng, 2014), a key distinction between fluent and dysfluent speech is that in the former pauses mostly occur on clause boundaries, while in the latter many pauses occur in mid-clause positions. Speakers at earlier stages of L2 acquisition may need more mid-clause pauses to deal with the demands of the speech production processes. In the data used for this research we have seen multiple cases of mid-clause pausing throughout all stages of the project and measuring their proportions would provide very valuable data.

This paper only used two variables of complexity, although as the analysis showed, complexity was the dimension of CAF that increased the most in this project. In the future, we would like to use more metrics of complexity to allow a more sophisticated and broad analysis of this dimension.

Additionally, this paper does not deal with accuracy, the third important aspect of CAF. It was decided that accuracy at this point of level acquisition did not show significant increases, but would require conducting a time-consuming analysis, preferably by native speakers. However, more detailed observations reveal that accuracy might have increased at least at some points in the project, so conducting an analysis of accuracy may be helpful in the future.

As for implications for practical use, further research of this learning corpus can lead to useful insights in the field of evaluation of speaking and improvements of teaching techniques in high schools. Analysis of data by a larger number of students with different levels of English might contribute to a better understanding of the connection between certain variables and CEFR level. For instance, variables that certainly improved over time can be given more importance in the evaluation process.

Finally, a preliminary analysis of questionnaires and interviews with participants of the project shows that students reported lesser anxiety levels. Students associated it with the fact that, unlike during lessons in their high school, they had a chance to speak English and had fun doing that. Therefore, further analysis from the point of view of second language anxiety can bring more insights into the importance of speaking tasks in high school English education.

References

Bulté, B., Housen. A. (2012) Defining and operationalizing L2 complexity. In Housen, A., Kuiken, F., Vedder, Ineke. (Eds.), *Dimensions of L2 performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 21-46). Amsterdam/Philadelphia: John Benjamins.

Cucchiarini, C, Strik, H., Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, 111: 2862–2873.

De Jong, N. H. Schoonen, R., & Hulstijn, J. (2009). *Fluency in L2 is related to fluency in L1*. Paper presented at the Seventh International Symposium on Bilingualism (ISB7), Utrecht, The Netherlands.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, 54, 113–132.

Ellis, R., & Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford: Oxford University Press.

García-Amaya, L. (2015, August). A longitudinal study of filled pauses and silent pauses in second language speech. In Lickley, R., Wester, M., Rose, R., Eklund, R. (Eds.), *Proceedings of the 7th Workshop on Disfluency in Spontaneous Speech* (pp. 23–27). University of Edinburgh, Scotland, UK.

Hanzawa, K. (2021) Development of second language speech fluency in foreign language classrooms: A longitudinal study. In *Language Teaching Research*. Online first. https://doi.org/10.1177/13621688211008

Housen, A., Kuiken, F., et al. (2012). Complexity, accuracy and fluency: definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and Proficiency: complexity, accuracy and fluency in SLA* (pp. 1–20). Amsterdam/Philadelphia: John Benjamins.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning,* 64(4) (pp. 809–854). https://doi.org/10.1111/lang.12084

Lennon, Paul. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40(3)*, (pp. 387-417).

Maeda, H. (2021). Speaking performance among high school students at different CEFR levels: A comparative study on the TEAP speaking section.『全国英語教育学会紀要』32, (pp.33-48)

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of non-native speaker conversations. *Discourse Processes,* 14 (pp. 423–441).

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 287–318). Cambridge: Cambridge University Press.

Robinson, P. (2003). The Cognition Hypothesis, task design and adult task-based language learning. *Second Language Studies,* 21(2) (pp. 45–107).

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. New York, NY: Routledge

Skehan, P. (2003). Task-based instruction. *Language Teaching,* 36(1) (pp. 1–14). https://doi.org/10.1017/S026144480200188X

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal,* 65 (pp. 71–79). https://doi.org/10.1093/elt/ccq020

Tavakoli, P., & Wright, C. (2020). *Second language speech fluency: From research to practice*. Cambridge University Press.

Williams, S. A., & Korko, M. (2019). Pause behavior within reformulations and the proficiency level of second language learners of English. *Applied Psycholinguistics,* 40 (pp. 723–742). https://doi.org/10.1017/S0142716418000802

Yan, X., Kim. J., Kim, H. (2020) Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test, *Language Testing*, 38(4) (pp. 485–510). https://doi.org/10.1177/0265532220951508