The Modern Language Journal

Development of Comprehensibility and its Linguistic Correlates: A Longitudinal Study of Video-Mediated Telecollaboration

Yuka Akiyama
Georgetown University
Kazuya Saito
Birbeck College, University of London

Abstract

This study examined whether 30 learners of Japanese in the United States who engaged in a semester-long video-based eTandem course made gains in global language comprehensibility, that is, ease of understanding (Derwing & Munro, 2009), and what linguistic correlates contributed to these gains. Speech excerpts from Week 2 and 8 of tandem interactions were retrieved and later assessed subjectively and objectively for global comprehensibility and its linguistic correlates (lexical appropriateness, lexical richness, speech rate, and morphological accuracy) in a pre/posttest sample design. The results revealed that, although the group made significant gains in vocabulary and some gains in grammar, improvement in overall comprehensibility was subject to considerable individual variability. According to a follow-up cluster analysis and discriminant analysis, increase in speech rate was the strongest predictor of those individuals who improved comprehensibility. The findings suggest that telecollaborative interaction may promote the development of vocabulary and, to some extent, grammar, but that significant gains in comprehensibility come mostly from the fluency trait of speech rate and may require longer interactional intervention. The findings have implications for the design of telecollaboration that supports second language learning.

Keywords: telecollaboration/eTandem; comprehensibility; fluency; vocabulary; L2 interaction; corrective feedback; longitudinal

Introduction

Internet-mediated communication technology has made it possible to connect with people who do not share the same physical space. Accordingly, interacting with the target language community, which used to be achieved via physical mobility (e.g., study abroad), is now possible in a contact zone (Kern, 2014). Many language practitioners nowadays incorporate such online intercultural communication in order to enhance language teaching and learning. Thus, Belz (2003) defines telecollaboration as “institutionalized, electronically mediated intercultural communication under the guidance of a languacultural expert (i.e., a teacher) for the purpose of foreign language learning and the development of intercultural competence” (p. 2).

Among many possible approaches to telecollaboration, one of the most commonly adopted models is *eTandem* (Cziko, 2004), in which a pair of learners with different first languages team up and help each other learn their respective languages by making “native speaker voices a central part of the language learning experience" (O'Rourke, 2007, p. 42). In this particularly autonomous yet collaborative learning set-up, learners are expected to work for mutual understanding via negotiation for meaning (Long, 1996) and to provide corrective feedback, which is considered the "central overtly pedagogical element of a tandem partnership" (Little et al., 1999, p. 39).

Despite the increasing interest in telecollaboration, however, very few studies have documented the development of oral proficiency at a macro level. This is in contrast with previous studies that have analyzed the longitudinal development of specific linguistic features such as address forms (Belz & Kinginger, 2003) and modal particles (Belz & Vyatkina, 2008). Such microgenetic analysis is particularly conducive to advancing second language acquisition (SLA) research. However, to bring about curricular changes, we may need to take a broader perspective and provide empirical support that telecollaboration can offer opportunities to develop language abilities that are essential for engaging in the kind of intercultural exchanges that are deemed vital in the age of globalization.

Pursuing that larger interest, this study took a longitudinal approach to examining a linguistic construct that is essential for successful communication – comprehensibility, that is, ease of understanding (Trofimovich & Isaacs, 2012). Based on eTandem participants' performance in video-mediated interaction, the study analyzed how specific linguistic correlates (lexical appropriateness, lexical richness, speech rate, and grammatical accuracy) developed over a semester and how the change contributed to the development of global comprehensibility. Finally, we explored the linguistic profile of those participants who made significant gains in comprehensibility, focusing on the degree of development in each of the four linguistic correlates and participants' initial proficiency level.

Interaction and Second Language Development

From a theoretical perspective, the effectiveness of telecollaboration for SLA can be explained based on the premise of the interaction hypothesis (e.g., Long, 1996). The main tenet of the theory states that adult SLA takes place when language learners negotiate for meaning using conversational moves such as confirmation checks, comprehension checks, and clarification requests, and modify their output in the face of communication breakdowns (Mackey, 2012). The theory also focuses on the role of corrective feedback (e.g., recasts) in facilitating learners' noticing the gap between their interlanguage and the target language. Accordingly, the theory is grounded in the centrality of comprehensibility, specifically whether a lack thereof leads to communication breakdowns and/or triggers an interlocutor's provision of corrective feedback.

Previous studies on second language (L2) interaction have repeatedly found that negotiation for meaning and corrective feedback occur in relation to meaning consultation, often concerning lexical items (e.g., Ellis, 1995). This is because lexical items carry meaning and because unfamiliar words can easily be substituted or defined in isolation (Pica, 1994). Mackey, Gass, and McDonough (2000) found that L2 learners are likely to notice and repeat interlocutors' recasts on lexical and phonological errors because they have "more potential to seriously interfere with understanding" (p. 493) than do morphosyntactic errors. In other words, conversational participants pay particular attention to linguistic items that can interfere with comprehensibility.

Acquisitional benefits of L2 interaction have been found in numerous studies (for meta-analyses, see Lyster & Saito, 2010; Mackey & Goo, 2007). Such traditional interactional studies have investigated the effect of short-term interactional treatment (e.g., 10 minutes) on specific language features such as nouns (de la Fuente, 2002) and question formation (Mackey, 1999) in a controlled environment (e.g., in a lab using a highly structured task) in order to control for various factors that may influence acquisition. Aside from study abroad literature that claims benefits for the development of fluency (e.g., Segalowitz & Freed, 2004) and vocabulary (e.g., Dewey, 2008) on the basis of increased interactional opportunities, only a few studies have examined the effect of long-term L2 interaction on the development of global language abilities in a less controlled setting (see Bueno–Alastuey, 2011; Payne & Whitney, 2002). Considering the accumulated research evidence that L2 interaction indeed promotes language learning, now may be the time to return to the original idea of negotiation for meaning and ask how L2 interaction helps learners develop a performance ability that is crucial for successful communication.

Comprehensibility

Comprehensibility is a global construct that is subjectively experienced and is considered essential for performing successful communication. Comprehensibility is defined as listeners' perception of how easy or difficult it is for them to understand L2 speech (Derwing & Munro, 2009). Whereas many L2 speakers and their teachers tend to see nativelikeness as their ideal goal (Tokumoto & Shibata, 2011), few adult L2 speakers can actually pass for native speakers (Abrahamsson & Hyltenstam, 2009). Thus, many L2 education researchers (e.g., Levis, 2005) have emphasized the importance of setting realistic goals, such as developing comprehensibility rather than focusing on reducing accent, for the purpose of successful communication.

In the research literature, two global constructs for describing L2 speech—comprehensibility and accentedness—have typically been measured via native speakers' scalar judgments. These studies have found that comprehensibility and accentedness are partially interrelated yet essentially different (Saito, Trofimovich & Isaacs, 2015, 2016; Trofimovich & Isaacs, 2012;). While accentedness is typically influenced only by pronunciation accuracy (especially at a segmental level), comprehensibility is generally associated with a range of variables spanning several dimensions of fluency (Derwinget al., 2004), vocabulary (Saito et al., 2015), and grammar (Derwing, Rossiter, & Ehrensberger–Dow, 2002).

In one study that examined the longitudinal development of comprehensibility, Derwing, Munro, and Thomson (2008) looked at how two groups of Canadian learners of English as a second language (ESL) with Chinese and Slavic languages as their first languages (L1s) developed comprehensibility over 2 years. The authors found that only the Slavic language group, who had more exposure to English outside the classroom and who thus improved fluency, increased comprehensibility. The finding suggests that fluency may be one of the key constructs for comprehensibility development and that adult L2 learners have the capacity to continue improving their overall comprehensibility as a function of increased input and interaction. That is, given a rich environment for interactional opportunities, which often is challenging in a foreign language setting, classroom-based learners can improve comprehensibility.

The Current Study

The current study examined the change in comprehensibility and its linguistic correlates of 30 American learners of Japanese who engaged in a semester-long conversational exchange with native-speaking partners via *Google Hangout*. The interactional treatment was left naturalistic with focus on form reinforced via recast training (see the subsequent section on recast training) based on major principles of eTandem that emphasize autonomy, reciprocity, and corrective feedback (Cziko, 2004; Little et al., 1999). The following research questions guided the study:

RQ1. To what extent does Japanese L2 learners' speech change in terms of comprehensibility and its linguistic correlates (i.e., vocabulary, fluency, and grammar) over the course of a semester-long eTandem project?

RQ2. What is the linguistic profile of learners who made significant gains in comprehensibility?

Methods

Participants

According to previous relevant literature, L2 learners' initial proficiency has been identified as a crucial affecting variable for various dimensions of instructed SLA (Shintani, Ellis, & Li, 2013) as well as the effectiveness of telecollaborative interaction (e.g., Ryder & Yamagata–Lynch, 2014). In light of that fact, this study carefully selected participants with various demographic backgrounds in order to represent a wide range of proficiency levels in L2 Japanese. A total of 30 learners of Japanese in six American universities participated in this eTandem project, either as students enrolled in a one-credit course (19 students from one university) or as volunteer exchange partners aside from their regular classroom instruction (11 students from five universities). The participants ranged from beginners (those who had just completed the first-year Japanese course as the project was launched) to advanced learners (e.g., heritage speakers who might even pass for native speakers), where some were L1 speakers of English/Chinese, while others were bilingual speakers of English and Chinese/Japanese. Their instructional level (i.e., the highest-level course they had taken) correlated with oral performance ($r = .47, p = .009$), as measured by an elicited imitation test (EIT)[1] (Ortega et al., 2002). Table 1 provides more details on the participants.

TABLE 1
Characteristics of Japanese Learners

Years of instruction	*M* (*SD*)	2.33 (1.15)
	Min–Max	1–6
EIT scores (Max: 120)	*M* (*SD*) Min–Max	88.67 (20.95) 44–119
Age	*M* (*SD*)	20.50 (1.46)
	Min–Max	19–26
Gender	Male	14
	Female	16
L1 background		English (21)
		English & Japanese (4)
		English & Chinese (3)
		Chinese (2)

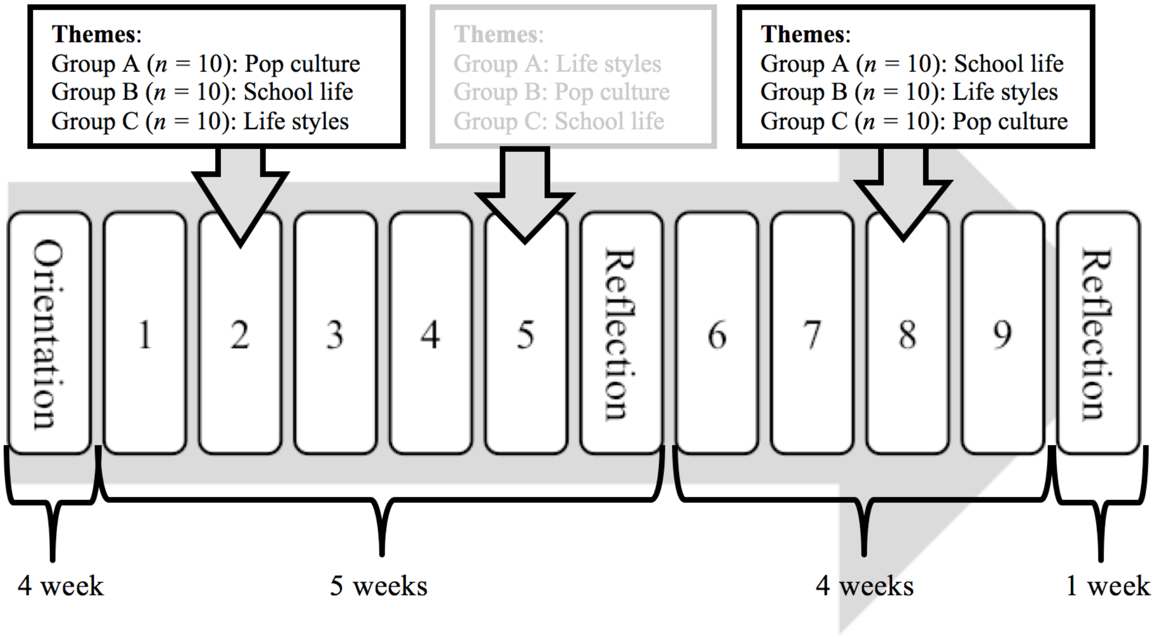
The majority of the participants reported they felt comfortable using videoconferencing tools. Only two participants had used them for L2 learning purposes. Some had friends in Japan, but they had never used videoconferencing tools for communication except for two participants who had done an online language exchange in the past. As for classroom instruction and L2 exposure outside the classroom, the majority reported that the classroom was the main venue for learning Japanese, that instruction focused on language forms over culture, and that they had little access to Japanese speakers outside the classroom.

The Japanese learners' partners were 30 native speakers of Japanese (14 males and 16 females; mean age = 20.56 ranging from 19 to 26), 27 of whom were studying English at three Japanese universities. The other three students were recent graduates from one of the three universities who wanted to keep studying English outside school. The majority of the participants had never studied abroad or experienced extensive exposure to learners of Japanese. None reported extensive use of English. Several participants had participated in a face-to-face and online language exchange project, yet none reported any formal language teaching experience, indicating that these participants' awareness of learner errors in Japanese, which may influence their error correction behavior, was minimal.

Project Set-up

Figure 1 summarizes the overall schedule for the eTandem project. It consisted of four weeks of orientation, two reflection sessions, and nine weekly *Google Hangout* interactions. The orientation at the beginning of the semester continued until the Japanese universities began their fall semester. During the four weeks of orientation, the American participants engaged in various types of training/workshops such as learning how to use *Google Hangout* and how to provide recasts (see the subsequent section on *recast training*). The participants in Japan also received the same training once their semester started.

FIGURE 1
Overall Schedule of the eTandem Project


Themes:
Group A (n = 10): Pop culture
Group B (n = 10): School life
Group C (n = 10): Life styles
Themes:
Group A: Life styles
Group B: Pop culture
Group C: School life
Themes:
Group A (n = 10): School life
Group B (n = 10): Life styles
Group C (n = 10): Pop culture
Orientation
1
2
3
4
5
Reflection
6
7
8
9
Reflection
4 week
5 weeks
4 weeks
1 week

The main tool for communication was *Google Hangout*, a video-conferencing tool that offers potential affordances of multimodal interaction (e.g., text chat, screen sharing). Participants were not explicitly taught how to use these multimodal features, although some were familiar with their use. Due to the time difference between Japan and the United States, which made it difficult to meet as a class, each tandem pair engaged in a weekly session using their own computers at home and according to their individual schedules.

Each hour-long Google Hangout session was divided into English and Japanese. Participants started the conversation in English with a five-minute free conversation (i.e., talking about random things such as their highlight of the week and things they did at school) followed by a 25-minute task-based conversation using visuals (see the subsequent section on *tasks*). After the 30 minutes, they switched languages and followed the same sequence in Japanese. Examination of the video interaction data revealed that some pairs occasionally spent longer than 30 minutes in one language; however, the majority of the participants followed the guidelines and stayed within the one-hour time limit.

Tasks

We decided to employ a type of information exchange tasks (see O'Dowd & Ware, 2009 for the taxonomy of telecollaboration tasks) called *visual-based conversation,* following the suggestion by Lee (2002), who found that two-way exchange of information on real-life topics that are theme-based and minimally structured helped students recycle ideas and reinforce language skills. In the current study, each participant was asked to find two visuals that would represent the theme of the week (e.g., Google images, their own pictures): one for Japan and the other for the United States. They were also asked to prepare two discussion questions for each visual image. For instance, if the theme of the week was *pop culture*, they might choose a visual of a Japanese idol group for the Japanese visual and Hollywood movies for the American visual and then come up with two questions for each visual. This type of open-ended task would require language learners to use various functional skills such as describing, narrating, and expressing opinions (Lee, 2002), and to negotiate for meaning (Doughty & Pica, 1986).

Note that, although the participants were instructed to choose a visual that matched their proficiency level and the topics of interaction were controlled and counter-balanced (see Figure 1), the visuals that each participant used varied. For instance, for the theme of *school life*, one participant chose a visual of his college dining hall, while another student chose a visual of college graduation (see Appendix A for sample excerpts). Therefore, it is possible that participants' oral performance was influenced by the learners' choice of visuals, especially regarding lexical richness (see Vercellotti, 2015 for similar arguments). However, we decided not to distribute the same visual to all the learners because controlling the tasks is neither ecologically valid nor ideal for an autonomous, longitudinal learning set-up like eTandem. In addition, based on the major principles of autonomy and reciprocity in eTandem, we considered it crucial for eTandem participants to take responsibility for their own learning and learn from each other by selecting visuals they thought represented each other's cultures.

Recast Training and Amount of Interactional Feedback Provided

The current study is based on a precursor study that investigated the relationship between learner beliefs and actual error correction behavior in eTandem (Akiyama, 2016). In the study, participants were trained to provide six types of corrective feedback (based on Lyster & Ranta, 1997) on their partner's erroneous utterances. Examination of the interaction data in the precursor study revealed that the participants used only three types of corrective feedback: recasts, explicit correction, and clarification requests, with recasts consisting of more than half of the error correction instances. Participants' perception data, which were collected three times throughout the semester, also revealed that the majority of the participants chose to provide recasts by the middle of the semester because recasts were considered the most intuitive, easiest, and least intrusive way to provide correction.

Accordingly, in the current study we trained both Japanese and American participants to provide recasts only on errors that they thought would hinder successful communication. In other words, the participants mainly attended to meaningful interaction with occasional focus on serious errors that would substantially hinder comprehensibility. To record the frequency and types of feedback provision, participants were required to submit an error correction log that is based on Mackey's (2006) notion of a learning journal (see Appendix B).

Analysis of the error correction log confirmed that native speakers indeed provided corrective feedback (see Appendix C for self-reported frequency of error correction by native speaking partners on grammar, vocabulary, and pronunciation). Specifically, each recast-trained native speaker provided an average of 5.64 instances of correction per session. Of a total of 1524 corrective feedback instances, almost half pertained to vocabulary and about one third to grammar. Pronunciation received the least amount of corrective feedback. This indicates that the Japanese native speaking partners mostly focused on correcting lexical errors, namely items that often carry meaning and are easy to correct.

Speech Analyses

Selection of Speech Data

Based on a precursor study (Saito & Akiyama, in press) that took an experimental approach to examining the effect of eTandem interaction on the *off-line* performance of English learners in Japan (i.e., English speech by the Japanese native speakers who interacted with the Japanese learners in this study) in a pre/posttest design, this study examined the *on-line* performance of the Japanese learners by analyzing the actual speech of the video interaction. This focus was motivated by moves to emphasize context-sensitive, ecologically valid accounts of SLA findings (Atkinson, 2002).

Two speech excerpts were selected for analysis for each Japanese learner. Thus, a total of 60 speech excerpts were analyzed (i.e., pre- and posttest excerpts from 30 participants). The pretest excerpts came from Hangout Session 2 (T1) while the posttest excerpts came from Hangout Session 8 (T2). There were about six to seven weeks between T1 and T2. The excerpts were taken when the learners were describing their first visual (out of two), which is usually when they were engaged in the least amount of turn-taking. This decision was intended to ensure that the speech excerpts would represent *learners*' rather than *interlocutors*' speech data, as interlocutors' contribution in conversation or/and interactivity may influence raters' judgment (cf. Derwing et al., 2004, who found that comprehensibility rating was not different between monologic and dialogic tasks). While comprehensibility was measured using the first 30 seconds of the one-minute speech excerpts, the full length (i.e., one minute) was used for measuring linguistic correlates. The shorter excerpts were used for comprehensibility rating because it is a subjectively perceived construct that can be judged more intuitively than its linguistic correlates (Derwing & Munro, 2009). The longer excerpts were used for the analysis of the linguistic correlates given that more data is required to conduct detailed linguistic analysis (Saito et al., 2016). To make sure that L2 learners' speech would add up to 30 seconds and one minute, we only counted the Japanese learners' conversational contribution towards the amount of time used for analysis, excluding the native speakers' contribution. Therefore, some of the speech excerpts exceeded 30 seconds/one minute, with the longest speech being 47 seconds for the former and 1.5 minutes for the latter. In order to equalize the quality of speech excerpts as much as possible, we normalized the speech excerpts for peak amplitude.

Subjective and Objective Speech Measures

While we measured comprehensibility via subjective human rater judgment, its linguistic correlates were measured both subjectively and objectively. We used the objective coding data to validate the subjective human judgment, since this is the first study investigating the relationship between comprehensibility and its correlates of L2 Japanese. The subjective and objective measures of the linguistic correlates are shown in Table 2.

TABLE 2
Subjective and Objective Measures of the Linguistic Correlates

	Human Rater Judgment (subjective measure)	Linguistic Coding of Transcribed Data (objective measure)
Lexical appropriateness	LexApp 0 = poor, 1000 = consistently appropriate	LemmaError Lemma inappropriateness (error ratio ranging from .0–1.0)
Lexical richness	LexRich 0 = few simple words, 1000 = varied vocabulary	Variation/Sophistication 1. Variation via Guiraud's index 2. Sophistication via JLPT Level 1–2 vocabulary (number of Level 1–2 words used)
Speech rate	SpeechRate 0 = too fast/slow, 1000 = optimal	Moras Number of moras per minute
Morphological accuracy	MorphAcc 0 = poor, 1000 = excellent	MorphError Morphological inappropriateness (error ratio ranging from .0–1.0)

Note. JLPT (Japanese Language Proficiency Test)

Subjective Human Rater Judgments of Comprehensibility and the Linguistic Correlates

Adapting procedures from Saito et al. (2015), four expert L1 Japanese speakers rated speech excerpts from the video-based interactions. In line with the definition of expert raters in Isaacs & Thomson (2013), these raters were graduate students in linguistics at a university in the United States, where they had received extensive training on pronunciation, vocabulary, and grammar analysis. They also reported extensive teaching experience prior to the project (*M* = 4.25 years, *SD* = 3.77, min = 1.5: max = 9.5 years).

For each rating session, the raters received a thorough explanation of the rated categories (see Appendix D for the rubrics) and the rating procedure and then evaluated five practice samples not included in the subsequent analysis. For each practice sample, they were asked why they had made their decisions and then received feedback to ensure that the rated categories were understood and applied appropriately. The raters then proceeded to rate 60 excerpts, presented to each rater in a unique random order. The rating was carried out a custom software, Z-Lab (Yao, Saito, Trofimovich, & Isaacs, 2013), developed using commercial software package (MATLAB 8.1, The MathWorks Inc., Natick, MA, 2013), and the raters used a free-moving slider on a computer screen to assess each category. When the slider was placed at the leftmost (negative) end of the continuum, labeled with a frowning face, the rating was recorded as 0; when it was placed at the rightmost (positive) end of the continuum, labeled with a smiley face, it was recorded as 1000. Except for the frowning and smiley faces and accompanying brief verbal descriptions for the endpoints of each category, the scale included no numerical labels or marked intervals (see Appendix D for the onscreen labels). The slider was initially placed in the middle of each scale, and the raters were told to use the full range of the scale. They were also told that even a small movement of the slider might represent a fairly large difference in the rating. A 1000-point sliding scale thus allowed raters to make fine-grained judgments for each linguistic category without being tied to discrete-point labels typical of Likert scales. The rating was conducted in two days: comprehensibility on Day 1 and its linguistic correlates on Day 2. While they were allowed to listen to a 30-second speech segment only once for comprehensibility, they had the option to listen to the one-minute speech repeatedly until they felt satisfied with their linguistic coding.

After the two days of rating, the raters took an exit questionnaire that asked them to assess the extent to which (a) they understood the four linguistic correlates (1 = *I did not understand at all,* 9 = *I understood this concept well*) and (b) they could comfortably and easily rate the linguistic correlates (1 = *very difficult,* 9 = *very easy and comfortable*) on a Likert scale. All raters demonstrated a relatively high level of understanding and comfort/ease, although one of the four raters' (Rater B) level of understanding and comfort was somewhat lower than the other raters' (see Appendix E).

Cronbach alpha revealed that the four expert raters were consistent in rating comprehensibility (α = .82), LexApp (α = .85), LexRich (α = .87), SpeechRate (α = .87), and MorphAcc (α = .86). These reliability indexes exceeded the benchmark value of .70–.80 (Larson–Hall, 2010). We decided to include Rater B's data in the analysis, after we checked that the results of Cronbach alpha did not change significantly when this rater's data were excluded. The four raters' scores, therefore, were averaged to generate a single score for each of the four linguistic correlates in T1 and T2 for each participant.

Objective Coding of Transcribed Speech Samples for the Linguistic Correlates

The one-minute-long speech excerpts (30 from Hangout Session 2 and 30 from Hangout Session 8) were transcribed by the first author and then verified for accuracy by a second transcriber. Building on previous literature (e.g., Yuan & Ellis, 2003), lexical appropriateness was measured by lemma inappropriateness (henceforth LemmaError). LemmaError is defined as the ratio of contextually and conceptually inappropriate words (including English substitutions) over the total number of words[2]. All inappropriately-used words (e.g., *gomikan* [for *gomibako*, ゴミ箱*, garbage bin*]) and English substitutions (e.g., *freedom* [for *jiyu:*,自由]) were counted as lemma errors.

Based on the description of the analytic rubric that was used for the subjective human rating of lexical richness, "varied and sophisticated uses of Japanese vocabulary" (Appendix D), we coded for both lexical *variation* (henceforth Variation) and *sophistication* (henceforth Sophistication). Variation was measured by the Guiraud's index, which is calculated by dividing the total number of different words (i.e., types) by the square root of the number of tokens (types/√tokens). Guiraud's index offers the advantage that it is robust against the varying length of texts or data sets, as in our data (Vermeer, 2000).

In turn, choosing an appropriate measure of linguistic sophistication turned out to be a major challenge for L2 Japanese data, due to the dearth of lexical sophistication research with this target language. Some relevant studies (e.g., Hatakeyama, 2014) suggest that the use of Level 2 vocabulary, as indicated by the Japanese Language Proficiency Test (JLPT), is correlated with the oral proficiency interview (OPI) ratings of the American Council on the Teaching of Foreign Languages (ACTFL). Hatakeyama (2014) found that the use of JLPT Level 2 vocabulary, which consists of abstract words that mostly originate in Chinese (i.e., *kango*), characterizes the noun usage of those who are rated Intermediate-Mid and higher on the OPI, with Level 2 nouns accounting for about 30% of noun usage. Accordingly, the current study defined lexical richness as the number/token frequency of JLPT Level 1–2 words. Following Hatakeyama (2014), the analysis focused on content words (i.e., nouns, verbs, adjectives, adverbs) and excluded function words, proper nouns, unknown words, expressive words (e.g., backchanneling), and adjectives that are used in conjunction with other verbs (see Appendix F for included/excluded categories). *Reading Tutor* (http://language.tiu.ac.jp), available in the online version of this article, was used to classify those content words into JLPT vocabulary levels.

For speech rate, the number of moras per minute was chosen (henceforth Moras). While syllables are often counted to measure fluency (e.g., Derwing et al., 2004), we chose moras as the unit of analysis because they are considered the basic timing unit that indicate phonological length in Japanese (Warner & Arai, 2001). Repetitions, reformulations, and replacements were counted towards the total number of moras. Fillers (e.g., *ah, un un, etto*) were excluded from the analysis.

Finally, we defined morphological inappropriateness (henceforth MorphError) as the ratio of morphological errors over the total number of words. These errors were related to the conjugations of verbs/adjectives/nouns such as *te*-form and derivational forms like *haya-ku* (*quickly*, adverb) versus *haya-i* (*quick*, adjective), tense/aspect, voice, modality, particles, and transitivity.

For inter-rater reliability, the 60 transcripts were first coded by a trained coder; then another trained coder re-coded 30 randomly-chosen transcripts (50%). The inter-rater reliability was 86.67% for LemmaError, 90.00% for Variation, 86.67% for Sophistication, 96.67% for Moras, and 93.33% for MorphError. Items that the two raters did not agree on were negotiated until they reached an agreement.

Results

Change in Comprehensibility

Table 3 shows the descriptive statistics for comprehensibility scores, as measured subjectively by the four expert raters. First, in order to examine whether the 30 participants improved comprehensibility between T1 and T2, a paired-sample *t*-test was calculated. The test revealed no significant effect of time, $t(29) = 1.114$, $p = .275$, $d = .15$, indicating that the group's mean comprehensibility score did not improve over the course of the project. Additionally, a univariate analysis of covariance (ANCOVA) was utilized to examine whether the *t*-test result held true even when learners' proficiency differences (as measured with an elicited imitation test, the covariate) were held constant. The ANCOVA test also revealed that the learners' improvement in comprehensibility was not statistically significant, Wilks' Lambda (.985), $F(1,28) = .419$, $p = .523$, partial $\eta^2 = .015$. This indicates that learners' initial proficiency was not related to how much improvement they made in comprehensibility.

TABLE 3
Subjective Human Rater Judgment of Comprehensibility

	M	*SD*	Min	Max	Mean Difference T2–T1	95% Confidence Interval for Difference		Cohen's *d*
						Lower	Upper	
T1	584.03	246.12	175.00	986.00	39.00	−32.62	110.62	.15
T2	623.03	248.05	153.00	992.50				

Change in the Linguistic Correlates

Next, we examined the extent to which Japanese L2 learners' speech changed in terms of the four correlates of comprehensibility as measured subjectively: LexApp, LexRich, SpeechRate, and MorphAcc. Table 4 shows the descriptive statistics for each linguistic construct (Figure 2 depicts the same information visually.)

TABLE 4
Subjective Human Rater Judgment of Linguistic Correlates (*N* = 30)

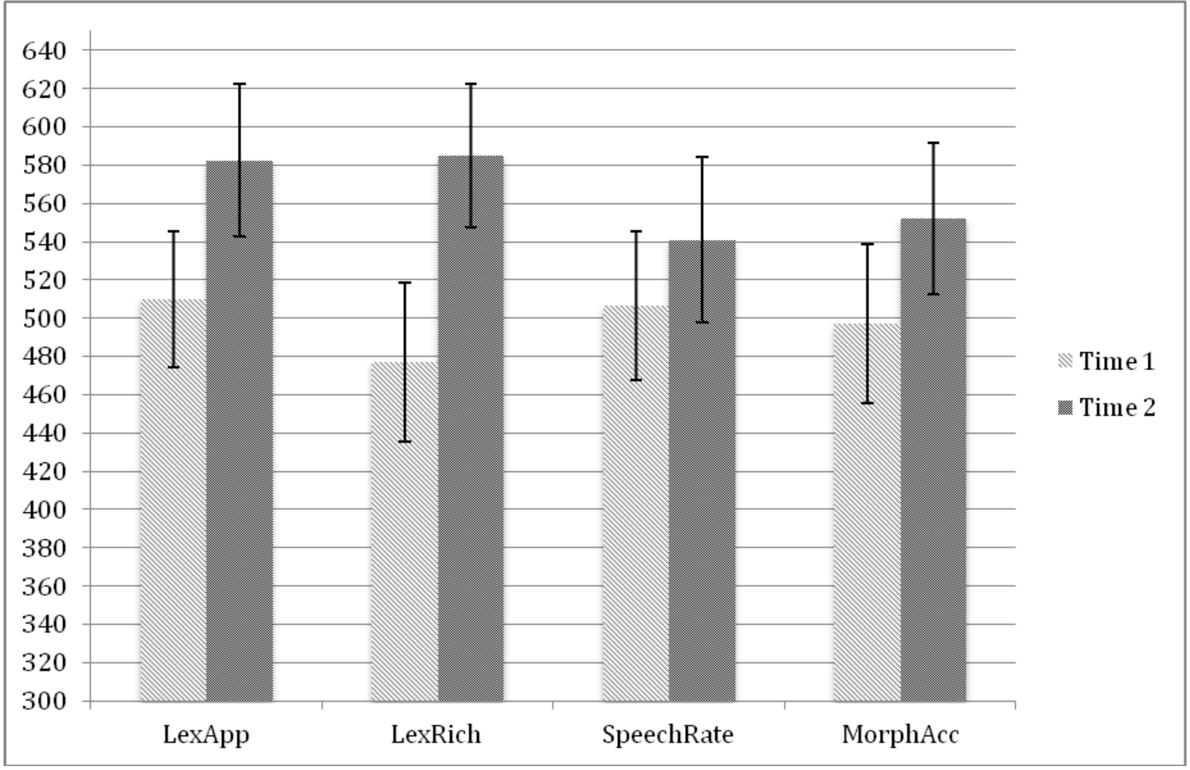
	Time	*M*	*SD*	Mean Difference T2–T1	95% Confidence Interval for Difference		Cohen's *d*
					Lower	Upper	
LexApp	1	510.13	194.48	72.34	12.58	132.11	.35
	2	582.48	219.29				
LexRich	1	477.13	227.60	107.658	56.86	158.46	.50
	2	584.79	204.98				
SpeechRate	1	506.44	212.83	34.52	−13.95	82.98	.15
	2	540.96	237.44				
MorphAcc	1	497.28	226.65	54.83	−3.37	113.02	.25
	2	552.10	217.91				

A one-factor, between-subjects multivariate analysis of variance (MANOVA) with repeated measures was conducted to examine the effect of time on a combination of four dependent variables (DVs). No extreme univariate outliers were identified. Evaluation of the homogeneity of variance-covariance matrices (Box's M), error variances (Levene's test), linearity, nonmulticollinearity, and normality assumptions underlying MANOVA did not reveal any substantial anomalies for the resulting sample of 30 cases, and the *a priori* alpha level was thus set at $p < .05$.

The MANOVA indicated statistically significant group differences on the combined DVs according to Wilks' Lambda (.590), *F*(1,26) = 4.523, *p* = .007, partial η^2 = .410. Follow-up univariate ANOVAs were conducted to examine where the statistically significant differences existed. Since the assumption of sphericity was violated, we used Greenhouse–Geisser for interpretation. The univariate test results with the Bonferroni adjustment (α = .0125) indicated a statistically significant improvement for LexRich, *F*(1,29) = 18.787, $p < .001$, *d* = .50, partial η^2 = .393, with the mean difference between T1 and T2 being 107.66. Although LexApp was not statistically significant, *F*(1,29) = 6.129, *p* = .019, *d* = .35, partial η^2 = .174, it demonstrated a relatively large mean difference and effect sizes. Similarly, although MorphAcc was not statistically significant, *F*(1,29) = 3.712, *p* = .064, *d* = .25, partial η^2 = .113, the mean difference and the effect sizes were relatively large. SpeechRate, in contrast, improved the least with the mean difference of 34.52, *F*(1,29) = 2.122, *p* = .156, *d* = .15, partial η^2 = .068.

In summary, although LexRich was the only construct that improved at the statistically significant level after the Bonferroni adjustment, LexApp also improved to a noticeable extent with the 95% confidence intervals of the pre- and post-scores barely overlapping each other. In contrast, neither SpeechRate nor MorphAcc improved substantially with their confidence intervals overlapping as shown in Figure 2.

FIGURE 2
Development of Linguistic Correlates between T1 and T2: Rater Judgment



Following the analysis of subjective ratings, we examined the change in the five linguistic correlates that were measured objectively: LemmaError, Variation, Sophistication, Moras, and MorphError. As the data except for Variation were found to be nonnormally distributed, a series of Wilcoxon signed-rank tests with the Bonferroni adjustment were used. Table 5 shows the descriptive statistics including the percentiles. The *α* level was set at *p* = .01 after the Bonferroni adjustment.

TABLE 5
Objective Coding of Transcribed Data for Linguistic Correlates (*N* = 30)

	Time	*M*	*SD*	Percentiles			Mean Difference T2–T1	95% Confidence Interval for Mean Difference	
				25th	50th	75th		Lower	Upper
LemmaErorr	1	.044	.033	.016	.034	.060	−.015	−.029	−.001
	2	.029	.040	.00	.017	.032			
Variation	1	4.63	.86	3.92	4.65	5.13	.322	.083	.561
	2	4.95	.76	4.56	4.91	5.28			
Sophistication	1	3.43	3.87	.00	2.50	5.25	1.63	.526	2.74
	2	5.07	3.45	3.00	4.00	7.00			
Moras	1	151.00	48.15	124.25	140.00	175.50	26.53	−4.20	57.27
	2	177.53	73.81	120.00	159.50	235.00			
MorphError	1	.022	.024	.00	.017	.036	−.008	−.017	.00
	2	.014	.017	.00	.011	.022			

The results revealed a very similar pattern to that of subjective rater judgment. Variables regarding lexical richness, namely Variation and Sophistication, improved at the statistically significant level, $Z = -2.687$, $p = .007$ and $Z = -2.620$, $p = .009$, with a medium effect size of $r = .49$ and .48, respectively. LemmaError also improved at the statistically significant level, $Z = -2.561$, $p = .01$, $r = .47$. MorphError, on the other hand, did not improve as much as the three lexical variables, although it demonstrated some improvement with a medium effect size ($Z = -1.769$, $p = .077$, $r = .32$). Similar to the analysis of subjective ratings, Moras (i.e., speech rate) improved the least ($Z = -1.351$, $p = .117$, $r = .25$).

Next, to examine the effect of initial proficiency, correlation analyses were performed between the participants' initial proficiency (as measured with the EIT) and gains in the linguistic correlates. Considering the small sample size, we decided to run correlation analyses instead ofa MANCOVA. The results showed that Moras was the only construct whose gain score correlated with the initial proficiency level ($r = .376$, $p = .041$). The others did not correlate, whether rated subjectively or objectively: LexApp, $r = .231$, $p = .219$; LexRich, $r = .062$, $p = .747$; SpeechRate, $r = .233$, $p = .214$; and MorphAcc, $r = -.203$, $p = .282$ for subjective ratings and LemmaError, $r = -.043$, $p = .823$; Variation, $r = .049$, $p = .795$; Sophistication, $r = -.194$, $p = .305$; and MorphError, $r = .122$, $p = .521$. This indicates that learners' initial proficiency was not related to how much improvement they made in the linguistic correlates of comprehensibility.

In summary, similar patterns were observed between the subjective and objective ratings. Namely, (a) vocabulary improved the greatest, especially lexical richness, (b) grammar improved but not at the statistically significant level, and (c) fluency was the construct that improved the least from T1 to T2. The initial proficiency level was not related to the degree of improvement.

The Linguistic Profile of Comprehensibility Improvers

While the group as a whole ($N = 30$) did not demonstrate a statistically significant gain in comprehensibility, a wide range of gain scores was observed. Since the study sought to examine individual learners' trajectories in developing the linguistic correlates and how the development contributed to comprehensibility gains, we used a two-stage cluster analysis procedure: hierarchical cluster analysis and the k-means analysis (Byrne, 1998) to classify the sample population into homogenous groups based on the comprehensibility gain scores.

A hierarchical cluster analysis using Ward's method was employed to determine the optimum solution for the number of clusters. Based on the result of the dendogram in Appendix G, we defined the number of clusters at three and ran a cluster analysis using the the k-means method. A one-way ANOVA confirmed the presence of substantial mean differences in comprehensibility gain scores between the three groups, $F(2,27) = 144.467$, $p <. .001$. The three groups were labeled: Minus (eight learners whose comprehensibility score decreased from T1 to T2), No Change (15 learners whose comprehensibility did not change), and Plus (seven learners who made comprehensibility gains). Table 6 shows the three groups' mean and gain scores in comprehensibility.

TABLE 6
Descriptive Statistics of Comprehensibility Scores by the Three Groups

Groups	Time	*M*	*SD*	Min	Max	Mean Difference T2–T1	95% Confidence Interval for Difference		Cohen's *d*
							Lower	Upper	
Minus	1	706.25	265.94	281.30	986.00	−197.31	−246.55	−148.08	−.75
(*n* = 8)	2	508.94	260.40	153.50	869.25				

No change	1	574.43	247.63	246.30	906.30	36.77	11.45	62.09	.15
(*n* = 15)	2	611.20	251.58	273.75	953.75				
Plus	1	464.93	174.65	175.00	701.00	313.86	240.62	110.62	1.89
(*n* = 7)	2	778.79	156.76	604.50	992.50				

Next, we used a discriminant function analysis to examine the linguistic profile of those 7 participants whose comprehensibility improved significantly. Discriminant analysis is a statistical method that is used to predict group membership from a set of predictors (Tabachnick & Fidell, 2013). In this study, we used the gain scores of the four linguistic correlates (i.e., LexApp, LexRich, SpeechRate, and MorphApp) as predictors of those whose comprehensibility scores decreased/remained the same/increased. The gain scores of the objective coding were not used because the data did not meet the required assumptions for running the statistical analysis.

Evaluation of the homogeneity of variance–covariance matrices (Box's M), error variances (Levene's test), linearity, nonmulticollinearity, and normality assumptions underlying discriminant analysis did not reveal any substantial anomalies for the resulting sample of 30 cases, and the α level was thus set at $p = .05$. However, equality of the cell size was not met, as our sample population was divided into groups of seven, eight, and fifteen. Nonetheless, we decided to proceed with the analysis, first, considering the exploratory nature of the study and, second, because we interpreted the results by focusing on descriptive statistics rather than relying on null hypothesis significance testing (Norris, 2015).

The test of equality of group means showed that improvement in SpeechRate was the only statistically significant variable ($p = .004$) that contributed to the prediction of the membership. The *p* values for the other variables were: LexApp gain ($p = .063$), LexRich gain ($p = .052$), and MorphError gain ($p = .839$), respectively. Although the three variables were not likely to contribute to the prediction of group membership, they were retained for further analysis, considering that the *p* values of LexRich and LexApp gain scores were close to the statistically significant level.

The structure matrix in Table 7 shows two discriminant functions that were used to predict the group membership. An overall statistically significant effect was found for the combined functions (1 and 2), Wilks' Lambda = .491, $\chi^2(8, N = 30) = 18.143$, $p = .02$. The second function on its own did not provide additional statistically significant predictions, Wilks' Lambda = .852, $\chi^2(3, N = 30) = 4.088$, $p = .252$. This indicates that Function 1, which is best represented by SpeechRate ($r = .815$), together with Function 2, which is best represented by LexApp ($r = .701$), accounted for 51% of the variance between groups.

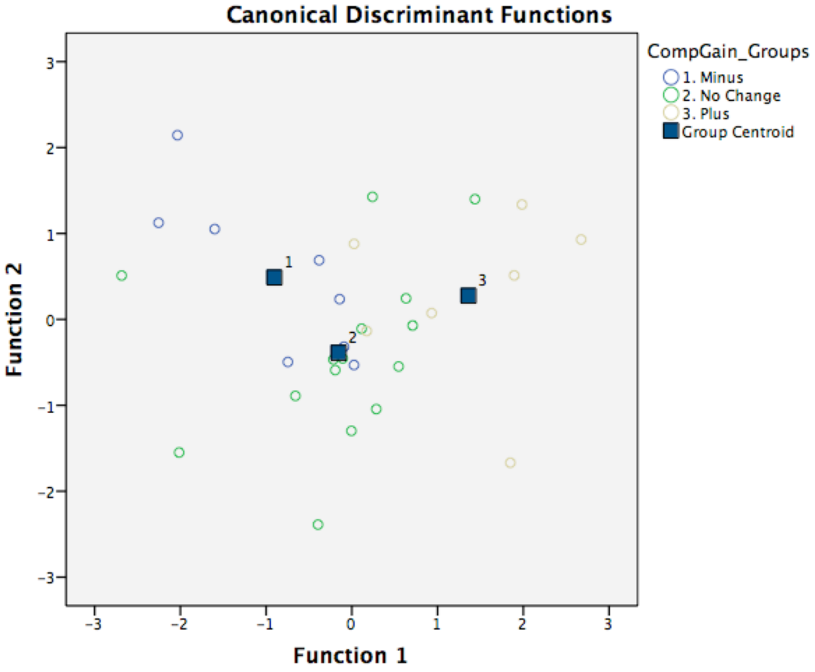
TABLE 7
Structure Matrix

	Function	
	1	2
SpeechRate gain score	.815*	−.148
LexRich gain score	.524*	.496
LexApp gain score	.439	.701*
MorphAcc gain score	.104	.171*

*. Largest absolute correlation between each variable and any discriminant function.

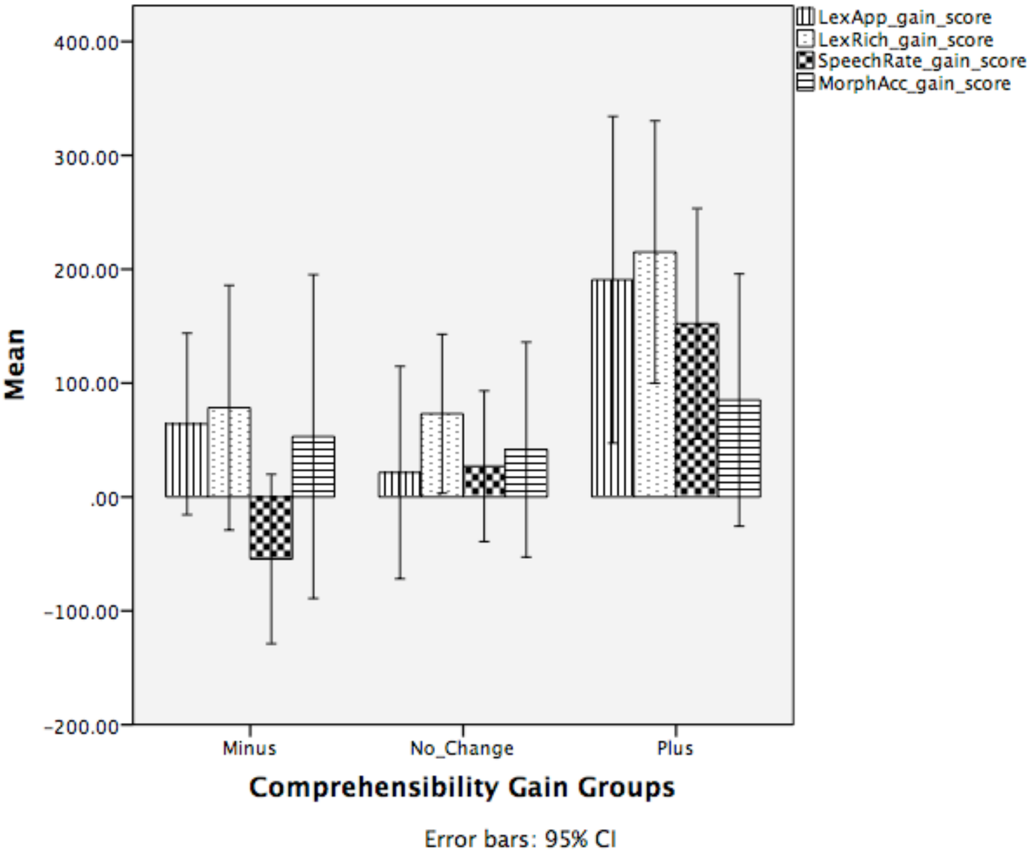
Figure 3 displays the individual cases and group centroids (average values for each group) in relation to the two discriminant functions: (a) from left to right, Function 1 clearly distinguishes between all three groups, and much more so between Group 3 (Plus) and the other two; (b) from top to bottom, Function 2 additionally distinguishes between Groups 1 and 2, but less so between Group 3 and the other two groups. This indicates that those who made significant gains in comprehensibility (i.e., Plus Group) and the other two groups were discriminated mostly based on SpeechRate gain score (Function 1), while the Minus Group and the No Change Group were differentiated additionally by the gains in vocabulary, especially lexical appropriateness (Function 2).

FIGURE 3
Predicting Comprehensibility Gains: Cases and Group Centroids for Two Discriminant Functions



Finally, the descriptive statistics were examined in comparison with the result of the discriminant analysis. The three groups' average gain scores in the four linguistic correlates are presented in Figure 4. The graphic representation revealed that (a) the Plus Group made great improvement in all four linguistic correlates, especially in LexApp and LexRich, and (b) SpeechRate gain score seems to be the key correlate that differentiated the three groups. The confidence intervals as well as the effect sizes (see Appendix H for the descriptive statistics) confirm the result of the discriminant analysis.

FIGURE 4
Linguistic Correlates' Mean Gain Scores by Comprehensibility Groups



To summarize the result section, this study examined whether 30 learners of Japanese who engaged in a semester-long video-based eTandem intercultural exchange made global language gains in comprehensibility and its linguistic correlates. We found that the group did not make significant gains in comprehensibility, although vocabulary scores improved significantly, as did grammar to some extent. In contrast, development of fluency was not observed. The study also examined the linguistic profile of seven learners who made significant gains in comprehensibility. The result of discriminant analysis revealed that these improvers were the ones who made gains with respect to all aspects of the linguistic correlates, not only vocabulary and grammar but also fluency. Participants' initial proficiency was not related to how much gains they made in comprehensibility and its linguistic correlates.

Discussion

The findings of this study suggest that improving comprehensibility entails enhanced performance in a wide range of linguistic correlates. In the current study, limited improvement in fluency prevented the group from improving their comprehensibility, while vocabulary improved, as did grammar to some extent. In other words, improved fluency may be a prerequisite for developing comprehensibility. Considering the finding in study abroad literature that an abundant amount of meaningful interaction often facilitates the development of fluency and vocabulary (see Llanes, 2011, for a review), it may be the case that one semester of regular classroom instruction and weekly telecollaborative interactions does not amount to the quantity and quality of meaningful L2 interaction that is required for the development of fluency, while it is sufficient for the development of vocabulary, which is often the target of corrective feedback (as in this study) and which often develops faster than other linguistic constructs (Vercellotti, 2015).

If fluency development takes a long time before it exhibits significant improvement (e.g., Trofimovich & Baker, 2006), research that examines the development of a global construct like comprehensibility needs to span over a long period of time (i.e., greater than six to seven weeks). Hence, the findings of the current study shed light on the importance of conducting longitudinal studies, as suggested by Ortega & Byrnes (2008). Derwing et al. (2008), for instance, investigated the longitudinal development of fluency and comprehensibility of 32 ESL learners over 22 months. The study did not find a significant improvement of comprehensibility in the Chinese group who did not have much contact with English outside the classroom instruction. If a 22-month observation in the ESL context was not long enough to see the improvement of comprehensibility, how long will it take for foreign language learners with regular classroom instruction and occasional telecollaborative interactions (nine weekly interaction sessions equals nine hours in this study) to develop comprehensibility?

Future studies may answer this question by examining foreign language learners' interactional opportunities afforded not only by telecollaboration but also by classroom instruction (see Lewis & O'Dowd, 2016 for a similar argument). For instance, researchers can utilize tools such as the Language Contact Profile (Freed et al., 2004) and the Social Interaction Questionnaire (Dewey, Belnap, & Hillstrom, 2013) that are frequently used in study abroad research to examine the quantity and quality of L2 input that participants experience in and outside the classroom. Ethnographies and case studies (e.g., Negueruela–Azarola, 2011) may also provide researchers with a better idea of what to realistically expect from video-based telecollaboration and reveal how comprehensibility influences qualitative aspects of telecollaborative exchanges and how specific aspects of interaction in telecollaboration could draw learners' focus toward certain aspects of their production that are related to comprehensibility.

This study also revealed the minimal effect of participants' initial proficiency on how much improvement they made in comprehensibility and its linguistic correlates. There may be various reasons for this, including ceiling effects (i.e., the advanced learners did not have much room for improvement) and task effects (i.e., individual participants' task selection influenced the degree of change in the linguistic constructs under investigation). It is also possible that patterns regarding the relationship between participants' initial proficiency level and gain scores were obscured because this study measured change in macro-level constructs that consist of many sub-components. For instance, an elementary level learner may improve adjective conjugation (relatively simple structure introduced earlier in the curriculum), while an advanced learner may make gains in the intransitive versus transitive verb contrast (a complex system even for advanced learners). That is, each participant's development was relative to what they could perform at their proficiency level. When there is no constant dependent variable across different proficiency levels, the proficiency effect may not seem so crucial. Accordingly, future studies may examine both macro (e.g., morphological accuracy, lexical appropriateness) and micro constructs (e.g., verb conjugations, counters) to reveal the complex picture of comprehensibility development and its relationship with initial proficiency level.

Lastly, although it is not the focus of this study, it is important to acknowledge the potential impact of video mediation, because the conversational participants might have interacted differently via video-mediated interaction than they would have in face-to-face conversation. For instance, while video mediation allows participants to utilize the affordances of combining audio, video, and text (Jauregi & Bañados, 2008), it limits interactants' capabilities regarding eye contact, gestures, and pointing (Kern, 2014). Although this study did not find statistically significant improvement in comprehensibility in the video-mediated context, it is possible that the participants' off-line performance may have improved (see Saito & Akiyama, in press). By extension, future studies may compare the modes of communication (i.e., video vs. face-to-face tandem projects) (e.g., Canto, Jauregi, & van den Bergh, 2013) and how the modality factor influences the development of comprehensibility.

In summary, on the bases of these findings, we call for studies that take a holistic, longitudinal approach to tracing learners' development of comprehensibility. Accordingly, we may examine learners' performance during telecollaboration (i.e., situated performance, as in this study) and outside telecollaborative interaction (i.e., off-line performance, as in Saito & Akiyama, in press) and how their performance is related to classroom instruction and affordances of video mediation.

Pedagogical Implications

Practitioners considering incorporating eTandem into a language curriculum need to be aware that such intercultural interactions with native speakers may need to span over a longer period of time before learners can make improvements along various linguistic dimensions (considered in this study) and subsequently develop their global performance in comprehensibility. This is important to note because, as any language teacher knows, intercultural settings like eTandem could turn out to be a demotivating experience for L2 learners if their limited comprehensibility prevented them from communicating smoothly. Thus, practitioners may raise learners' awareness regarding what comprehensibility development involves (i.e., improvement not only in vocabulary but also in grammar and fluency), the long-term nature of the endeavor, and how comprehensibility could be negotiated with their partner.

Another pedagogical implication pertains to task design. In this study, the participants engaged in open-ended interactional tasks that imposed little time pressure. Consequently, the participants may not have been motivated to complete the task within a certain time limit. Employing tasks that are more controlled and require task completion in a limited time might have encouraged the participants to speak at a faster pace and increase the intensity of input/output, facilitating fluency development. However, as O'Dowd and Ware (2009) stated, too much focus on task completion may result in psychological pressure to rush through tasks simply for the sake of completion and reduce the opportunities to negotiate meaning to the fullest. Therefore, practitioners may need to be eclectic in choosing such tasks that are in accordance with the goals of the interactional intervention.

Practitioners are also advised to incorporate and integrate eTandem into the regular language instruction to the extent possible rather than considering it a supplemental activity (see O'Dowd, 2011 for the same argument). In this study, eTandem was an add-on course, which could not be incorporated into regular classroom instruction for various socioinstitutional reasons (see O'Dowd & Ritter, 2006). The lack of contingent interactions between the practitioner and participants may have prevented the learners from taking full advantage of the telecollaborative opportunities. Thus, it seems crucial to establish a mutually beneficial, reciprocal relationship between classroom instruction and telecollaborative interaction. For instance, upon observing the lack of fluency development, practitioners may provide explicit teaching of fluency (e.g., De

Jong & Perfetti, 2011). In turn, practitioners can consider telecollaboration as a venue where such explicit training can be enacted in a real-life context.

As suggested by Kern (2014), eTandem practitioners may need to be aware of the positive and negative affordances of video mediation and set a realistic goal because telecollaboration is not a magic bullet. There are things video-mediated interaction can and cannot do for language development. Practitioners may hold an orientation for the learners and discuss the positive (e.g., tools for enhancing the multimodal environment such as text chatting) and negative affordances (e.g., technical issues like transmission delays) of video mediation.

Finally, contrary to many practitioners' general belief that telecollaboration should be implemented after achieving a certain proficiency level, this study revealed that learners' proficiency level was not related to the degree of change in comprehensibility. Accordingly, it seems that even elementary-level students can improve their ability to negotiate for meaning, as long as appropriate scaffolding is provided. Thus, although eTandem is perhaps one of the most autonomous forms of telecollaboration in that instructors' involvement is limited to minimal, in the end, the role of instructors is crucial in ensuring that learners receive the resources they need for developing the ability to engage in successful communication.

Limitations

Some limitations are worth bearing in mind when interpreting the present findings. First, since the current study does not employ a control group, we cannot claim that the improvement observed is due to the eTandem treatment. Rather, this study is focused on whether learners' performance in video-based interaction improved over time. Readers are advised not to assume a causal relationship between the eTandem treatment and the participants' language development.

Second, although this study was careful to ensure that the assumptions to run inferential statistics were met to the greatest extent possible, the sample size of 30 is relatively small for running statistical analyses. Consequently, it is possible that the statistical power to detect significant differences was reduced. Although we ensured the validity of the statistical analysis by resorting to descriptive statistics as well, future studies may employ a larger sample size to validate the findings of this study.

Third, although the topics of interaction at T1 and T2 were controlled and counter-balanced, the visuals that each participant used varied. More controlled tasks (e.g., use of the same visuals) might have led to different results, although this might have reduced the ecological validity of the study. Relatedly, our findings were exclusively based on one speaking task (a visual-based conversation), although L2 oral proficiency has often been measured via a range of tasks with different levels of complexity, argumentativeness, and formality (De Jong et al., 2012). It would be intriguing for future studies to adopt multiple tasks to capture the multifaceted nature of L2 comprehensibility development (Derwing et al., 2004).

Finally, due to the lack of studies on L2 Japanese on this topic, we did not examine pronunciation (e.g., segmentals, word stress, intonation) as the linguistic correlates of comprehensibility. However, as previous studies have shown (e.g., Derwing & Munro, 2009), these constructs contribute greatly to comprehensibility. Thus, future studies may examine what makes L2 Japanese speech comprehensible to provide a more encompassing picture.

Conclusion

This study took a longitudinal approach in investigating comprehensibility development of Japanese learners who engaged in a semester-long video-mediated eTandem project. The results revealed that the group as a whole did not improve global comprehensibility, mainly due to the lack of improvement in fluency. Considering that the original idea of telecollaboration was to provide an alternative to study abroad (i.e., physical mobility), which is well-known for its effectiveness in improving fluency, it is ironic in a sense that the findings of this study highlight the differences, rather than similarities, between study abroad and telecollaboration. However, the findings of this study should be used as a guide when creating a curriculum that incorporates a telecollaborative component to create a virtual study abroad experience for the students. One question we may pursue along this line of reasoning is "What is it that video-based telecollaboration can do that physical mobility cannot do, and vice versa, for developing a linguistic skill that is required for intercultural communication?"

Acknowledgements

We thank the journal editor, Heidi Byrnes, and *MLJ* reviewers as well as Lourdes Ortega, John Norris, Yasser Teimouri, and Brandon Tullock for their helpful input and feedback on the content of this manuscript. We are also grateful to the students for participating in the project, Michael Ferreira and João Telles for the opportunity to conduct the study, and Ze Shan Yao for his technical help. This study was funded by the ITEL Teletandem Project awarded to the first author and by the Grant-in-Aid for Scientific Research in Japan (No. 26770202) awarded to the second author.

Notes

1. EIT provides a quick estimate of learners' global L2 proficiency. We chose the present test because it had been proven to correlate highly with other criterion measures such as the ACTFL Oral Proficiency Interview (Ortega et al., 2002). Test takers are asked to repeat what they hear after a short pause. The possible score range is from 0 to 120.
2. The definition of a "word" in this study follows that of *Reading Tutor* (http://language.tiu.ac.jp)*,* the program that was used for analyzing lexical richness.

References

Akiyama, Y. (2016). Learner beliefs and corrective feedback in telecollaboration: A longitudinal investigation. *System.*

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*, 249–306.

Atkinson, D. (2002). Toward a sociocognitive approach to second language acquisition. *Modern Language Journal*, *86*, 525–545.

Belz, J. A. (2003). Linguistic perspective on the development of intercultural communicative competence in telecollaboration. *Language Learning & Technology*, *7*, 68–117.

Belz, J. A., & Kinginger, C. (2003). Discourse options and the development of pragmatic competence by classroom learners of German: The case of address forms. *Language Learning*, *53*, 591–647.

Belz, J. A., & Vyatkina, N. (2008). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. *Language Learning & Technology*, 12(3), 33–52.

Bueno–Alastuey, M. C. (2011). Perceived benefits and drawbacks of synchronous voice-based computer-mediated communication in the foreign language classroom. *Computer Assisted Language Learning*, *24*, 419–432.

Byrne, D. (1998). *Complexity theory and the social sciences*. London: Routledge.

Canto, S., Jauregi, K., & van den Bergh, H. (2013). Integrating cross-cultural interaction through video-communication and virtual worlds in foreign language teaching programs: Is there an added value? *ReCALL*, *25,* 105–121.

Cziko, G. (2004). Electronic tandem language learning (eTandem): A third approach to second language learning for the 21st century. *CALICO Journal*, *22*, 25–39.

De Jong, N. H., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61,* 533–568.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34.

De la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in receptive and productive acquisition of words. *Studies in Second Language Acquisition*, *24*, 81–112.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 476–490.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, *29*, 359–380.

Derwing, T. M., Rossiter, M. J., & Ehrensberger–Dow, M. (2002). "They speaked and wrote real good": Judgements of non-native and native grammar. *Language Awareness*, *11,* 84–99.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 655–679.

Dewey, D. P. (2008). Japanese vocabulary acquisition by learners in three contexts. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *15*, 127–148.

Dewey, D. P., Belnap, R. K., & Hillstrom, R. (2013). Social network development, language use, and language acquisition during study abroad: Arabic language learners' perspectives. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *22,* 84–110.

Doughty, C., & Pica, T. (1986). "Information gap" tasks: Do they facilitate second language acquisition? *TESOL Quarterly*, *20*, 305–325.

Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics*, *16*, 409–435.

Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition*, *26*, 349–356.

Hatakeyama, M. (2014, March). *Corpus analysis of vocabulary by OPI proficiency levels: Interviews from Database of Japanese Language Learners Conversation (DJLLC)*. Poster session presented at the meeting of the American Association of Japanese Teachers, Philadelphia, PA.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*, 135–159.

Jauregi, K., & Bañados, E. (2008). Virtual interaction through video-web communication: A step towards enriching and internationalizing language learning programs. *ReCALL*, *20*, 183–207.

Kern, R. (2014). Technology as pharmakon: The promise and perils of the Internet for foreign language education. *Modern Language Journal*, *98*, 340–357.

Larson–Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge/Taylor & Francis.

Lee, L. (2002). Enhancing learners' communication skills through synchronous electronic interaction and task-based instruction. *Foreign Language Annals*, *35*, 16–24.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 369–377.

Lewis, T., & O'Dowd, R. (Eds.). (2016). *Online intercultural exchange: Policy, pedagogy, practice*. New York: Routledge/Taylor & Francis.

Little, D., Ushioda, E., Appel, M. C., Moran, J., O'Rourke, B., & Schwienhorst, K. (1999). *Evaluating tandem language learning by e-mail: Report on a bilateral project* (Occasional Paper No. 55). Dublin: Centre for Language and Communication Studies.

Llanes, À. (2011). The many faces of study abroad: An update on the research on L2 gains emerged during a study abroad experience. *International Journal of Multilingualism, 8,* 189–215.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition: Second language acquisition* (pp. 413–468). New York: Academic Press.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, *19*, 37–66.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32,* 265–302.

Mackey, A. (1999). Input, interaction, and second language development. *Studies in Second Language Acquisition*, *21*, 557–587.

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, *27,* 405–430.

Mackey, A. (2012). *Input, interaction, and corrective feedback in L2 learning*. Oxford: Oxford University Press.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*, 471–497.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford: Oxford University Press.

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(S1), 97–126.

Negueruela–Azarola, E. (2011). Changing reasons as reasoning changes: A narrative interview on second language classroom motivation, telecollaboration, and the learning of foreign languages. *Language Awareness*, *20*, 183–201.

O'Dowd, R. (2011). Intercultural communicative competence through telecollaboration. In J. Jackson (Ed.), *The Routledge handbook of language and intercultural communication* (pp. 342–358). New York: Routledge/Taylor & Francis.

O'Dowd, R., & Ritter, M. (2006). Understanding and working with 'failed communication' in telecollaborative exchanges. *CALICO Journal*, *23*, 623–642.

O'Dowd, R., & Ware, P. (2009). Critical issues in telecollaborative task design. *Computer Assisted Language Learning*, *22*, 173–188.

O'Rourke, B. (2007). Models of telecollaboration (1): eTandem. In O'Dowd (Ed.), *Online intercultural exchange: An introduction for foreign language teachers* (pp. 41–61). Tonawanda, NY: Multilingual Matters.

Ortega, L., & Byrnes, H. (2008). *The longitudinal study of advanced L2 capacities*. New York: Routledge.

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October). *An investigation of elicited imitation tasks in crosslinguistic SLA research.* Paper presented at the Second Language Research Forum, Toronto.

Payne, J. S., & Whitney, P. J. (2002). Developing L2 oral proficiency through synchronous CMC: Output, working memory, and interlanguage development. *CALICO Journal*, *20,* 7–32.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, *44*, 493–527.

Ryder, L. Z., & Yamagata–Lynch, L. (2014). Understanding tensions: Activity systems analysis of cross-continental collaboration. *CALICO Journal*, *31*, 201–220.

Saito, K., & Akiyama, Y. (in press). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning*.

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*. Advance online publication. doi: 10.1093/applin/amv047

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*, 217-240.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2015). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness and sense relations. *Studies in Second Language Acquisition, 38*. Advance online publication. doi: 10.1017/S0272263115000297

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 173–199.

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, *63*, 296–329.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.

Tokumoto, M., & Shibata, M. (2011). Asian varieties of English: Attitudes towards pronunciation. *World Englishes, 30*, 392–408.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28,* 1–30.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*, 905–916.

Vercellotti, M. L. (2015). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics,* 1–23. Advance online publication. doi:10.1093/applin/amv002

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*, 65–83.

Warner, N., & Arai, T. (2001). The role of the mora in the timing of spontaneous Japanese speech. *The Journal of the Acoustical Society of America*, *109*, 1144–1156.

Yao, Z., Saito, K., Trofimovich, P., & Isaacs, T. (2013). *Z-Lab.* Retrieved August 15, 2013, from https://github.com/ZeshanYao/Z-Lab

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*, 1–27.

APPENDIX A

Speech Excerpts and Sample Scores via Subjective Judgment

The following two excerpts demonstrate how their comprehensibility is different depending on the ratings of the linguistic correlates and how the same topic (i.e., life styles) can be enacted using different visuals.

ID102_T2 on Life Styles

この写真はアムトラクの電車の中です。アムトラクは、アメリカには、アメリカには、アメリカには大きい市の外であまり人の電車がない。けど、アムトラクは市から市まで人の電車を乗る電車。アムトラク、アムトラク他にも、他にないと思う。ニューヨークに行く時、アムトラクで行く。

English translation

This picture is the inside of the Amtrak train. Amtrak is, in America, in America, in America, there are not many people trains outside big cities. But, Amtrak is the people’s train that *rides trains from a city to a city. Amtrak, other than Amtrak, there are none, I think. When I go to New York, I go by the Amtrak.

Comprehensibility	Lexical Appropriateness	Lexical Richness	Morphological Accuracy	Speech Rate
273.75	295.50	290.00	280.23	205.50

ID106_T1 on Life Styles

日本にはごみかん、ごみかん？ごみかんは、あまりないと思います。ゴミの箱。ばこ？ばぐ。ばこ、はあまりないと思う。アメリカはアメリカはたくさんありますけど、日本にいた時には、電車に、乗って後で、乗った後で。びんが・・・。(partner: でも、駅にゴミ箱ない？). けど、飲み物機しかないと思います。

English translation

In Japan, garbage *can, garbage *can? There aren’t so many garbage *cans, I think. Garbage box. Box? *Bix. There aren’t many boxes, I think. In America, America has a lot of them, but when I was in Japan, train, after I get on the train, after I got on the train, the *can… (partner: But aren’t there garbage boxes in the train station?). But, they only have vending machines.

Comprehensibility	Lexical Appropriateness	Lexical Richness	Morphological Accuracy	Speech Rate
540.00	431.75	408.75	602.25	497

APPENDIX B
Sample Error Log

COMPLETE THIS SECTION AFTER ENGLISH INTERACTION (5 minutes)
(1) How often and what types of errors did you correct?
(2) If you remember some example errors, write them down as well.

Grammar	11	2 (partner's grammar did not have many overt errors, but sentences were simple)	● ● ● ● ● ●
Vocabulary (Word choice)	11111	5	● Said story to mean about yourself ● Overtime got replaced for all the time ● ● ● ●
Pronunciation	111111111	9	● pronunciation of "introduce" ● vei bik = very big ● hanbaga = hamburger ● ● ●

COMPLETE THIS SECTION AFTER JAPANESE INTERACTION (5 minutes)

Pronunciation	My partner	Me	In the visuals	Yes, new	No, I've heard of it	No, knew it
1. Long vs short sounds are hard in context		x	x			x
2. uh could sound like wa (topic marker)		x				x
3. English stress accent comes through		x				x
4.						
5.						
Grammar	My partner	Me	In the visuals	Yes, new	No, I've heard of it	No, knew it
1. Wanted to overuse da/desu		x				x
2. Casual vs. polite forms was confusing		x				x
3. Overused past tense		x			x	

4. Relative clauses		x			x	
5. mou vs. mada is confusing		x			x	
Vocabulary	My partner	Me	In the visuals	Yes, new	No, I’ve heard of it	No, knew it
1. Tried to use literal words for an idiom	x			x		
2. sukoshi jikan for sugu ni (soon) (awkward phrasing)	x				x	
3. owarimashou for jikan desu ne	x				x	
4. had no idea how to say “anachronistic”		x		x		
5. Japanese word for common		x			x	
Content	My partner	Me	In the visuals	Yes, new	No, I’ve heard of it	No, knew it
1. bukatsudou		x			x	
2. Family	x	x			x	
3. What the university is like	x	x			x	
4. Anime	x	x				x
5. General favorite subjects		x		x		

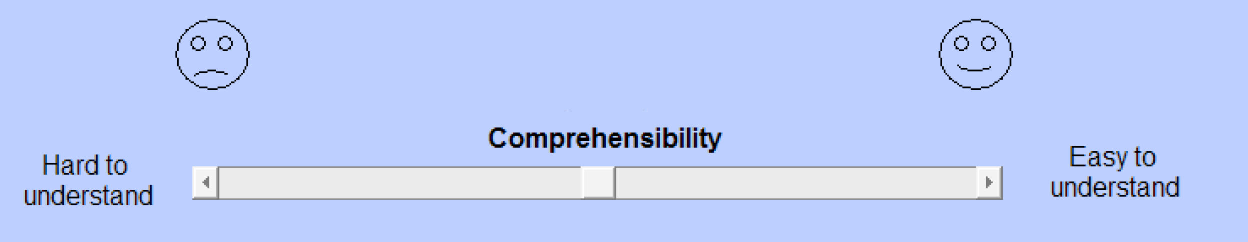
APPENDIX C
Rated Categories

Self-Reported Number of Interactional Feedback Episodes per Session (*N* = 30)

	Session	1	2	3	4	5	6	7	8	9	Total
Grammar	Sum	70	67	75	61	68	53	61	47	62	564
	M	2.33	2.23	2.50	2.03	2.27	1.77	2.03	1.57	2.07	2.09
	SD	2.92	2.36	2.27	2.04	2.23	1.63	2.08	1.22	1.53	2.03
Vocabulary	Sum	58	94	93	101	69	83	64	91	55	708
	M	1.93	3.13	3.10	3.37	2.30	2.77	2.13	3.03	1.83	2.62
	SD	1.95	2.71	2.14	1.99	1.80	1.65	1.31	1.59	1.37	1.83
Pronunciation	Sum	50	29	31	24	13	31	21	27	26	252
	M	1.67	.97	1.03	.80	.43	1.03	.70	.90	.87	.93
	SD	3.10	1.43	1.33	.96	.86	1.65	.99	.92	.82	1.34
ALL	Sum	178	190	199	186	150	167	146	165	143	1524
	M	5.93	6.33	6.63	6.20	5.00	5.57	4.87	5.50	4.77	5.64
	SD	5.16	4.68	4.54	3.68	2.88	3.11	2.85	2.58	2.82	3.59

APPENDIX D
Rated Categories

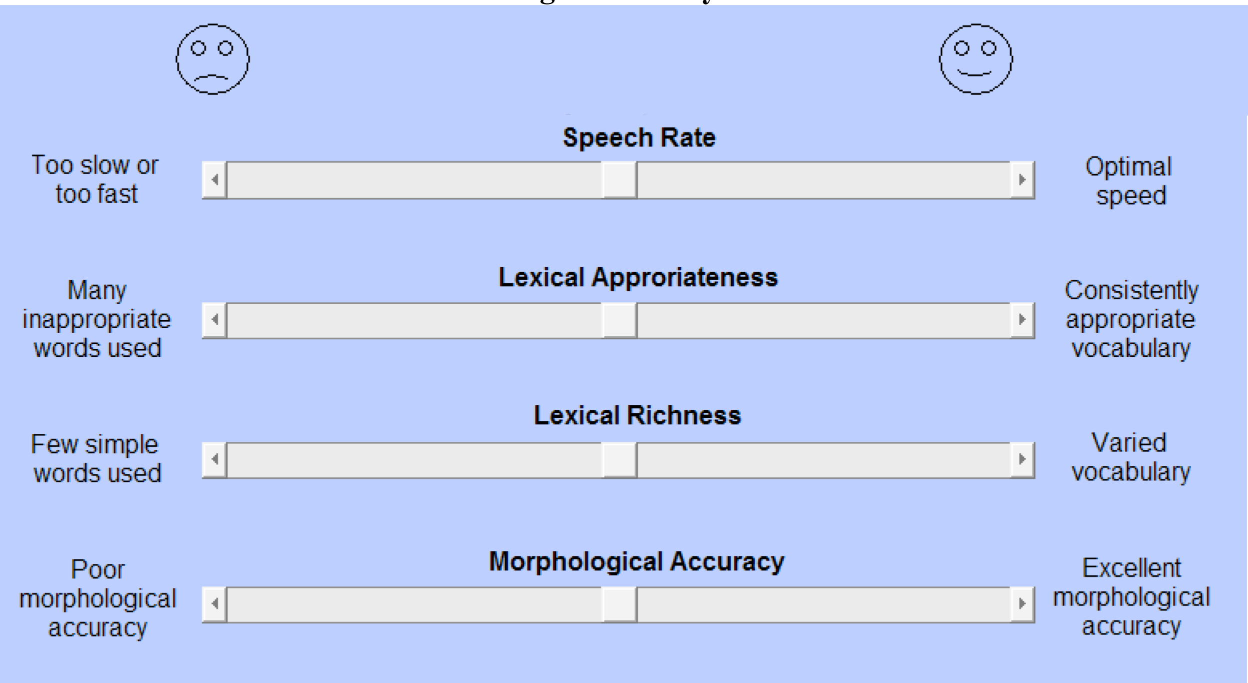
Global Analysis



Comprehensibility

This term refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.

Linguistic Analyses



Speech rate

Speech rate is simply how quickly or slowly someone speaks. Speaking very quickly can make speech harder to follow, but speaking too slowly can as well. A good speech rate should sound natural and be comfortable to listen to.

Lexical appropriateness

This dimension refers to the appropriateness of the vocabulary words used by the speaker. If the speaker uses incorrect or inappropriate words, including words from the speaker's native language, lexical accuracy is low. On the other hand, lexical accuracy is high if the speaker has all the lexical items required to accomplish the speaking task and does so using frequently-used and/or precise lexical expressions.

Lexical richness	This dimension also refers to the vocabulary used by the speaker. What is important here, however, is how sophisticated this vocabulary is, taking into account the demands of the speaking task. If the speaker uses a few simple, unnuanced words, the speech lacks lexical richness. However, if the speaker's language is characterized by varied and sophisticated uses of Japanese vocabulary, the speech is lexically rich.
Morphological accuracy	This rubric refers to the number of morphological errors related to the conjugations of verbs/adjectives/nouns such as *te*-form and derivational forms like 早く vs. 早い, tense/aspect, voice such as "causative" and "causative passive," modality such as "〜てはいけない," particles, and transitivity over the total number of words.

APPENDIX E

Level of Understanding and Comfort in Subjective Rating by the Four Expert Raters for the Four Linguistic Correlates (Maximum score = 9)

Raters		LexApp	LexRich	SpeechRate	MorphAcc	ALL
A	Understanding	8	9	9	9	8.75 (.50)
	Comfort	7	9	9	8	8.25 (.96)
B	Understanding	7	7	6	7	6.75 (.50)
	Comfort	7	7	4	7	6.25 (1.50)
C	Understanding	9	9	9	9	9.00 (.00)
	Comfort	8	6	9	8	7.75 (1.26)
D	Understanding	8	8	6	9	7.75 (.58)
	Comfort	9	8	8	9	8.50 (.58)
ALL		7.88 (.83)	7.88 (1.13)	7.50 (1.93)	8.25 (.96)	

APPENDIX F
Included and Excluded Categories in Analyzing Lexical Sophistication

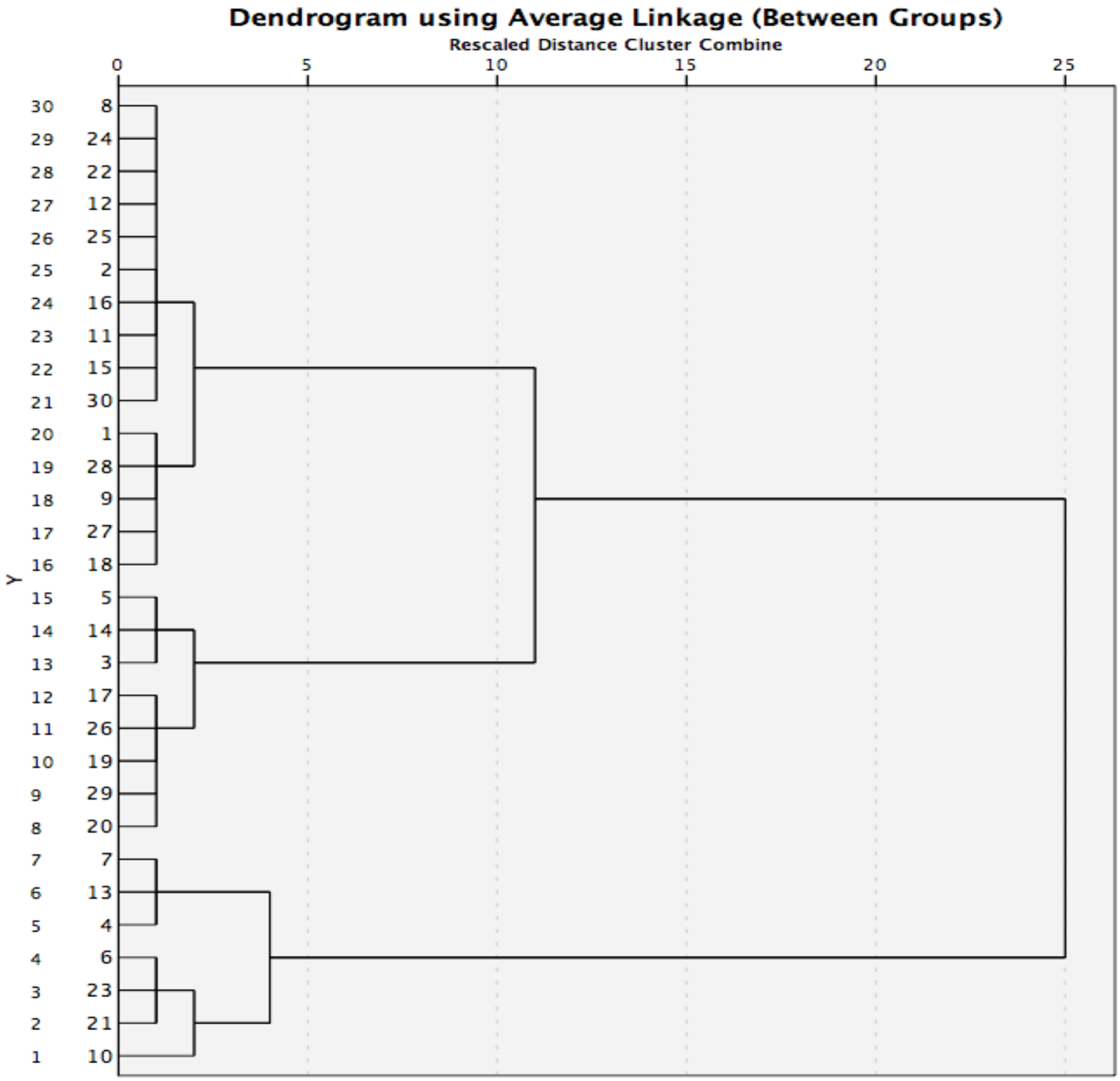
Included Categories of Words

Parts of Speech	Examples
Nouns	• 名詞 (漢字含む２字以上): 自分, 友達, 先生, 大学 • 名詞 B (平仮名のみ): あと, うち, とこ • 名詞 C (漢字 1 字): 人,姓,国,年,本 • 副詞可能: 今, 最近, 前, きょう • サ変名詞：仕事, 勉強, 話, 教育
Adjectives	• 形容動詞: 好き, 必要, 大変, だめ • 形容詞: 多い, 難しい, 大きい, 悪い • 形容詞 B (平仮名のみ): いい, すごい, よい, よろしい
Verbs	• 動詞: 思う, 言う, 見る, 来る, 行く • 動詞 B (平仮名のみ): する, ある, いう, なる
Adverbs	• 副詞: 特に, 全然, 一応, 少し • 副詞 B (平仮名のみ): そう, まあ, こう, ちょっと, やっぱり

Excluded Categories of Words

Categories	Examples
Unknown words	ニート, ノムヒョン, ビジャ
Expressive words	はい, えー, うん
Proper nouns	ワールドカップ, 羅生門, オリコン
Names of institutions	国民党, 朝日新聞, JR
Names of people	太郎, マサ, 田中
Names of places	日本, 韓国, アメリカ
Nai-adjectives	問題, 違い, 申し訳
Negative modals	ない, ん, ぬ, まい
Adjectives that are used in conjunction with other verbs	にくい, やすい, がたい, づらい

APPENDIX G
Dendrogram



APPENDIX H
Descriptive Statistics of Linguistic Correlate Scores by the Three Groups

Groups	Linguistic Correlates	T	*M*	*SD*	Mean Difference T2–T1	*SD*	Cohen's *d*
Decrease (*n* = 8)	LexApp	1	498.06	244.24	64.13	95.35	.28
		2	562.19	211.31			
	LexRich	1	508.22	260.02	78.41	128.55	.35
		2	586.63	184.76			
	SpeechRate	1	471.13	263.15	−54.47	88.96	.22
		2	524.22	211.67			
	Morph	1	607.31	201.43	53.09	170.06	−.28
		2	552.84	191.95			
No change (*n* = 15)	LexApp	1	536.95	195.93	21.45	168.46	.10
		2	558.4	240.06			
	LexRich	1	498.95	246.25	73.10	126.04	.31
		2	572.05	229.23			
	SpeechRate	1	513.05	236.88	27.03	119.54	.18
		2	554.65	234.16			
	Morph	1	489.08	238.15	41.60	170.45	.10
		2	516.12	278.58			
Increase (*n* = 7)	LexApp	1	466.46	139.41	190.79	155.12	1.13
		2	657.25	193.24			
	LexRich	1	394.86	140.2	215.14	124.53	1.25
		2	610.00	198.72			
	SpeechRate	1	493.36	187.14	152.25	109.38	.42
		2	578.5	217.81			
	Morph	1	428.36	133.51	85.14	119.80	.86
		2	580.61	211.82			