


RESEARCH ARTICLE 

Age effects in spoken second language vocabulary attainment beyond the critical period

Kazuya Saito 

University College London, London, UK
Corresponding author. Email: k.saito@ucl.ac.uk

(Received 15 September 2021; Revised 23 June 2022; Accepted 07 September 2022)

Abstract

The current study set out to examine to what degree age of acquisition (AOA), defined as a learner's first intensive exposure to a second language (L2) environment, mediates the final state of postpubertal, spoken vocabulary attainment. In Study 1, spontaneous speech samples were elicited from experienced Japanese users of English ($n = 41$) using storytelling and interview tasks. The samples were analyzed using a range of corpus- and rater-based lexical measures and compared to the speech of inexperienced Japanese speakers ($n = 40$) and native speakers of English ($n = 10$). The results showed that most experienced L2 learners tended to demonstrate nativelike proficiency for relatively easy lexical dimensions of speech (i.e., richness), but that AOA appeared to play a key role in predicting the ultimate attainment of relatively difficult lexical dimensions (i.e., appropriateness). In Study 2, the findings were successfully replicated with experienced L1 Polish users of English ($n = 50$).

Introduction

Research suggests that a strong relationship exists between learners' ultimate L2 speaking attainment and their age of acquisition (AOA), defined as the age at which intensive exposure to the target language begins. Studies have consistently shown that the earlier L2 learners are exposed to the target language, the more nativelike performance they can ultimately attain (e.g., Abrahamsson & Hyltenstam, 2009; Birdsong & Molis, 2001; Flege et al., 1995; Flege et al., 1999; Granena & Long, 2013; Johnson & Newport, 1989; Patkowski, 1990). This has led scholars to discuss if there is an optimal window of opportunity for L2 acquisition (for a comprehensive review see Muñoz & Singleton, 2011). Some have labeled this time frame as the "critical period" (e.g., Lenneberg, 1967). What remains controversial in this discussion is to what degree and how AOA mediates the final state of L2 speech acquisition when the target language is learned *after* puberty (e.g., Birdsong, 2005 vs. DeKeyser & Larson-Hall, 2005), and when the attainment concerns *lexical* (rather than phonological) aspects of L2 speech (cf. Lahmann et al., 2016). Drawing on Crossley's computational modeling of spoken L2 vocabulary use (Crossley et al., 2015; Kyle & Crossley, 2015; Salsbury et al.,

2011), the current study took a first step toward scrutinizing this topic in the context of two groups of high-proficiency bilinguals immersed in an L2-speaking environment after puberty (> 16 years).¹

Background

Age effects

It has been widely observed that L2 learners tend to demonstrate rapid and substantial improvements in their L2 proficiency following their arrival in a target language speaking country. Such improvement is thought to be possible if learners practice the language with a wide range of interlocutors in various social settings on a regular basis (Flege & Liu, 2001). In contrast, it is known that the final outcomes of learning after years of immersion (i.e., ultimate attainment) are subject to a great deal of individual variation and are linked to several underlying factors. Factors explored to date have included the linguistic distance between L1 and L2 structures (Best & Tyler, 2007), aptitude (DeKeyser et al., 2010), the quality and quantity of L2 input (Jia & Aaronson, 2003), cognitive aging (Birdsong, 2005, 2006; Hakuta et al., 2003), motivation (Moyer, 2014), and ethnic identity (Gatbonton et al., 2011). Among these variables, the timing of a learner's first intensive exposure to the target language (i.e., AOA) is perhaps the most widely researched, particularly when it comes to L2 oral proficiency.

As far as early arrivals are concerned (e.g., AOA < 16 years), there is substantial evidence that the degree of L2 ultimate attainment and incidence of nativelikeness is strongly related to AOA; that is, L2 learners with earlier AOA profiles are more likely to reach advanced L2 proficiency than those with later AOA profiles. One proposed explanation for this effect is that earlier AOA allows L2 learning to occur in the context of a less developed L1 system (Flege et al., 1995, 1999). However, scholars have continued to debate precisely which factors influence the degree of success in L2 speech learning when immersion takes place *after* the passing of the critical period. In what follows, two competing theoretical accounts of second language acquisition (SLA) are reviewed. Following Flege (2003, p. 11), the two different positions were labeled “maturational” and “nonmaturational and developmental.”

Maturation accounts

Maturation refers to the reduction of cerebral plasticity around puberty (Lenneberg, 1967). Although there are many divergent views amongst proponents of the maturational account, they have a common belief that this neurocognitive decline (observed in the critical and sensitive period) only drives the “younger is better” phenomenon up to puberty. This would suggest that AOA is unrelated to ultimate attainment in late bilinguals (> 16 years). Scholars who take on a “maturation account” of SLA believe that early and late L2 ultimate attainment are essentially *different* phenomena separated by the passing of an optimal, or critical, period for language learning (e.g., Abrahamsson, 2012; Abrahamsson & Hylténstam, 2009; DeKeyser et al., 2010; Granena & Long, 2013; Johnson & Newport, 1989; Patkowski, 1990). From this point of view, younger L2

¹Scholars have adopted different cut-off points between early and late learners. For example, Abrahamson and Hylténstam (2009) put the upper bound of the critical period at the age of 12. Following the original empirical study that proposed the idea of a discontinuity between early and late L2 learning (Johnson & Newport, 1989), late L2 learners in the current study were defined as those who had first become immersed in an L2 country after the age of 16.

learners are more likely to attain nativelike proficiency because they have more prolonged exposure to language during the critical period and are able to benefit from robust implicit learning abilities that aid in the acquisition of language through exposure alone. As maturational changes take place in the brain, however, it is thought that learners' access to such implicit learning mechanisms gradually declines (i.e., a robust AOA effect on early bilingualism). After the passing of the critical period, such implicit abilities may be limited to very few exceptional learners regardless of their age-related profiles (Link et al., 2013). Here, the relationship between AOA and acquisition is hypothesized to be nonsignificant (i.e., a discontinuous AOA effect on late bilingualism). The final attainment of late L2 acquisition could be rather determined by a range of factors related to individual differences in language analytic abilities (DeKeyser et al., 2010) and motivation (Moyer, 2014).

Nonmaturation accounts

Other scholars have taken the sharply contrasting stance that the AOA-attainment function can be observed throughout the life span because both early and late bilingualism are the *same* phenomenon. From this point of view, there is no domain-specific, optimal, or critical period for language learning. Instead, it is believed that L2 learners maintain the capacity to learn a new language after puberty as long as they are provided with ample opportunities for input and output (e.g., Best & Tyler, 2007 for Perceptual Assimilation Model-L2; Birdsong, 2005 for Cognitive Aging Hypothesis; Bundgaard-Nielsen et al., 2011 for Vocab Model; Flege, 2018 for Speech Learning Model; Kachlicka et al., 2019 for Auditory Precision Hypothesis). As such, AOA is believed to serve as a strong predictor of early *and* late bilingualism.

Empirical evidence

The role of AOA in late and postpubertal L2 ultimate attainment (learners' intensive exposure to an L2 > 16 years) is an ideal arena for assessing these two competing views on SLA. Maturation accounts predict "discontinuity in the AOA-proficiency link" resulting from a fundamental and qualitative change in learning potential after the mid-teens (DeKeyser & Larson-Hall, 2005, p. 97). This position ascribes the *absence* of age effects to the passing of the critical period. The nonmaturation position suggests that "a linear monotonic decline of learning over the [AOA] spectrum, with age effects continuing past the point at which maturation has ceased" (Birdsong, 2005, p. 115). Proponents of this view all assume the *presence* of age effects, although they have different perspectives on what explains the AOA-proficiency link in adulthood (e.g., L1-L2 interference, perceptual/cognitive aging, sociopsychological individual differences).

When it comes to L2 speech, research findings do not strongly support either position. While some studies have demonstrated strong age effects for early but not late arrivals (e.g., Granena & Long, 2013; Patkowski, 1990), others have failed to replicate these discontinuities (e.g., Flege et al., 2006). This discrepancy could be ascribed to methodology-based ceiling effects—in both groups of studies, L2 speech was elicited using highly controlled tasks (e.g., word and sentence reading), where participants are allowed to carefully monitor their performance by drawing on their explicit metalinguistic knowledge. Such tasks could have allowed most participants to perform at a satisfactory level and may not have elicited sufficiently varied speech to capture the relationship between different levels of AOA and proficiency attainment. Given that L2 learners' speech is more targetlike when elicited from controlled

(compared to free-speaking) tasks (Major, 2008), it remains controversial to what extent such “monitored performance” accurately reflects learners’ ability to carry out meaningful, interactive, and automatic communication in the real world (Piske et al., 2011; see Johnson & Newport, 1989 vs. Birdsong & Molis, 2001 for similar findings and arguments in L2 grammar attainment).

Indeed, the ability to use accurate, fluent and complex language while speaking *spontaneously* is instrumental to becoming a functional L2 user (De Jong et al., 2012), but is difficult to master relative to other dimensions of language (e.g., vocabulary, grammar; Granena & Long, 2013). Though limited in number, studies have begun to show that the final quality of early (AOA < 16 years) and late (AOA > 16 years) bilinguals’ L2 speech (i.e., accentedness) is strongly tied to AOA. This relationship is especially strong when speech is elicited using spontaneous speaking tasks, which require learners to attend to various dimensions of language while conveying an intended message (e.g., Derwing & Munro, 2013; Hopp & Schmid, 2013).

Most of the discussion surrounding age, experience, and L2 speech attainment thus far has been concerned with segmental/suprasegmental aspects of speech (e.g., Saito, 2013) and listeners’ judgments of global accentedness of speech (e.g., Derwing & Munro, 2013). However, less research has examined ultimate attainment for lexical dimensions of L2 speech. This could be considered a weakness in the current literature considering that a growing number of scholars have emphasized the multifaceted nature of L2 speech proficiency—that is, the ability to use pronunciation, vocabulary, and grammar in an accurate, fluent, and sophisticated manner (De Jong et al., 2012). Importantly, some scholars have argued that the AOA-proficiency link could be mediated by the differences in the manner in which and amount of learning that takes place for different linguistic dimensions; assuming that the acquisition of L2 phonology (segmentals and suprasegmentals) is more difficult than that of vocabulary, the presence of age effects could be observed in the former but not the latter (Granena & Long, 2013).

To obtain a full-fledged understanding of the mediating role of AOA in late bilingualisms, more research is needed to examine the extent to which the emerging findings in L2 phonology are generalizable to L2 vocabulary attainment. Although some previous studies have concerned the role of AOA in late attainers’ *comprehension* of L2 vocabulary, the findings have remained mixed arguably because of their substantially different operationalization of lexical knowledge (e.g., Abrahamsson & Hyltenstam, 2009 for lexical inferencing vs. Hellman, 2011 for vocabulary size). Drawing on a recent framework for spoken vocabulary use (i.e., appropriateness and richness; Crossley et al., 2015), the current study took a first step toward examining the mediating roles of AOA in the *production* aspects of late L2 vocabulary attainment.

Second language vocabulary attainment

Modeling spoken L2 lexical proficiency

Vocabulary is a crucial component of language comprehension and production (Jiang, 2000). Most research attention in the field of L2 vocabulary has focused on delineating the *receptive* vocabulary sizes of learners (2–3K frequent word families for beginner L2 speakers; 24K frequent word families for L1 speakers; Webb & Nation, 2017). In terms of acquisition, it has been shown that receptive vocabulary development is strongly linked to the amount of received input, which explains why many learners can achieve nativelike vocabulary size after a great deal of L2 immersion experience (e.g., Hellman, 2011).

Contrastively, very few studies have explored what comprises *productive* vocabulary knowledge (Koizumi, 2012). On the one hand, word frequency may not necessarily capture the complex nature of L2 productive vocabulary learning, as advanced L2 speakers do not necessarily use more infrequent words while speaking (Crossley et al., 2019). On the other hand, Crossley and colleagues have proposed, developed, and refined methods for computationally modeling L2 learners' spoken vocabulary use (Crossley et al., 2015; Kyle & Crossley, 2015; Salsbury et al., 2011). Within this framework, the lexical dimensions of L2 speech are analyzed from two different perspectives: appropriateness and richness.²

The first dimension, appropriateness, refers to whether words are used in a contextually appropriate manner and with the correct assignment of morphological markers. Traditionally, appropriateness has been operationalized in terms of binary accuracy, which is measured by tallying the number of lexical choice errors and morphological errors per clause (for a review see Skehan, 1998). More recently, scholars have begun to emphasize the importance of error gravity in the assessment of appropriateness, given that not all lexicogrammar errors have the same degree of impact on successful communication, and that different *types* of errors can influence native speakers' global ratings of L2 oral proficiency (Révész et al., 2016). Various types of subjective judgment methods have been developed to address the latter point. For example, one common approach is to have linguistically experienced coders evaluate the lexical quality of transcripts based on global accuracy rubrics, such as the acceptability of multiword units (Crossley et al., 2015), weighted lexicogrammar accuracy (Foster & Wigglesworth, 2016), and semantic and morphosyntactic accuracy (Saito, 2019).

The second dimension, lexical richness, comprises a range of lexical features related to "the depth and breadth of lexical knowledge available to speakers" (Kyle & Crossley, 2015, p. 759). Two oft-used indices of lexical richness are word frequency (how often certain words are used in major corpus data) and word range (how widely words are used in diverse contexts). The assumption is that learners who can use less frequent and more narrowly occurring words have a more sophisticated and richer lexical repertoire (Laufer & Nation, 1995). Richness has been also operationalized as the ability to access words that are more abstract (e.g., Salsbury et al., 2011, for concreteness and imageability), semantically specific (e.g., Crossley et al., 2009, for hypernymy), and more polysemous (e.g., Crossley et al., 2010, for sense relations).

Empirical evidence

Recently, an increasing number of studies have shown that late L2 learners can substantially improve the appropriateness and richness dimensions of their L2 vocabulary use during the initial stages of immersion (length of residence [LOR] < 1 year) (e.g., Mora & Valls-Ferrer, 2012; Tavakoli, 2018). For example, Crossley and colleagues longitudinally analyzed the lexical richness of ESL learners' L2 speech development over a period of 1 year, finding that within the first 4 months, vocabulary use became more abstract due to the increased use of more hypernymic and less concrete words

²This literature review focuses on spoken vocabulary research and has deliberately avoided mentioning L2 writing research (which has also used a similar paradigm). Spoken vocabulary is different from written vocabulary as the former does not allow learners to refer to texts for decoding and forces them to resort to other cues (e.g., visuo-gestural; for a comprehensive review on vocabulary research in L2 speaking and writing, see Koizumi, 2012).

(Crossley et al., 2009; Salsbury et al., 2011). In addition, the learners seemed to become more aware and capable of using the peripheral meanings of various lexical items (Crossley et al., 2010).

Surprisingly, the literature on the spoken vocabulary attainment of experienced L2 learners (LOR > 5 years) relative to inexperienced L2 speakers and native controls is far more limited. To my knowledge, there have been only two empirical studies: Lindqvist, Bardel, and Gudmundson (2011) and Bartning, Lundell, and Hancock (2012). Lindqvist et al. examined the lexical richness (word frequency) of high-level L1 Swedish speakers learning L2 French ($n = 7$) and L2 Italian ($n = 10$). The authors found that high-level speakers demonstrated significantly higher levels of lexical richness in their spoken production than less proficient ones, but that their performance still differed from native-speaking controls. Bartning et al. also examined the spoken morphological accuracy of 20 late experienced Swedish learners (LOR > 5 years) of L2 French. They similarly showed that participants' accuracy performance was significantly better than inexperienced learners (LOR < 2 years; $n = 10$), but worse than native controls ($n = 10$). The findings of these small-scale studies need to be replicated with a larger sample size.

There is also little and limited research on the role of AOA in spoken vocabulary attainment. The existing literature has exclusively focused on L2 attainers' lexical *comprehension* proficiency but using a wide range of outcome measures (e.g., judgment, recognition, and recall tasks) and resulting in mixed findings. Some studies have found strong age effects (younger is better) for early bilinguals but not for late bilinguals (e.g., Abrahamsson & Hyltenstam, 2009; Granena & Long, 2013; Spadaro, 2013). Others have found the significant influence of AOA in L2 vocabulary size attainment even after puberty (Hellman, 2011). However, when it comes to the role of AOA in L2 lexical *production* proficiency (the focus of this study), there have been only two empirical studies—Lahmann et al. (2016) and Saito (2015). Lahmann et al. (2016) failed to find any significant effects for AOA on the lexical aspects of L2 oral proficiency attainment with early bilinguals ($n = 102$ German speakers of English with AOA 7–17 years). This could be ascribed to the fact that the authors analyzed just a single component of L2 lexical proficiency—lexical richness (i.e., not appropriateness). Saito (2015), however, examined the predictive power of AOA for lexical appropriateness and richness with 88 late Japanese-English bilinguals (AOA 16–35 years), and two comparison groups ($n = 10$ inexperienced Japanese speakers of English and $n = 10$ native speakers of English). Although the study did not find a significant AOA-acquisition relationship, this was ascribed to methodological factors—that is, the speech samples were elicited using a single, simple task (picture description), and the length of each speech token was very short ($x < 50$ words) (insufficient for the purpose of robust vocabulary analysis, i.e., > 100 words suggested by Koizumi & In'nami, 2012).

Study 1

Considering the literature reviewed in the preceding text, several preliminary conclusions can be drawn about postpubertal speech development. First, extensive L2 experience (LOR > 5 years) makes a positive impact on the lexical aspects of late learners' speech attainment, even if they arrive in an L2 speaking environment after puberty. Second, very few late learners can attain nativelike proficiency. Third, the role of age in postpubertal L2 lexical proficiency attainment remains unclear—while age effects have been clearly observed with early bilinguals (AOA < 16 years), it is still open to debate whether such effects exist for late bilinguals (AOA > 16 years). On the one hand, it has

been argued that the degree of late L2 attainment is unrelated to AOA due to the passing of the critical period around puberty (i.e., maturation accounts). On the other hand, it has been suggested that the relationship between AOA and L2 proficiency should remain significant among late bilinguals, as they use learning processes that are like those used by early bilinguals (i.e., nonmaturation accounts).

Research questions

To move the research agenda forward, I examined the role of AOA in determining the attainment of lexical appropriateness and richness in L2 speech. In the context of 41 experienced speakers of English (i.e., L2 attainers; AOA 18–37 years; LOR = 6–34 years), Study 1 took a first step toward exploring two important research questions:

1. To what degree do postpubertal L2 speakers ultimately enhance the lexical aspects of their oral proficiency after years of immersion relative to the two baseline groups of L1 and inexperienced L2 speakers?
2. What is the nature of the relationship between AOA and spoken vocabulary attainment?

An investigation into whether AOA affects spoken vocabulary attainment in late bilinguals is especially important as it may help resolve a long-held debate between two opposing schools of thought, that is, maturation (e.g., DeKeyser & Larson-Hall, 2005) versus nonmaturation accounts (e.g., Birdsong, 2005). While there have been a number of similar large-scale AOA studies in L2 phonological attainment (e.g., Derwing & Munro, 2013; Flege et al., 2006; Hopp & Schmidt, 2013; Saito, 2015), the current project was the first investigation into AOA effects in L2 vocabulary attainment.

Method

Participants

Attainers ($n = 41$)

Following the standards of previous age-related L2 studies, late Japanese attainers were defined as those who had arrived in the United States after 18 years of age and had resided there for more than 6 years (e.g., Birdsong & Molis, 2001; Johnson & Newport, 1989 for similar decisions). Some scholars have argued that a minimum LOR of 10 years is necessary if ultimate attainment is to be assessed rather than rate effects (e.g., DeKeyser, 2013). In this study, a decision was made to use 6+ years as the cutoff point in accordance with recent LOR research. There is both cross-sectional and longitudinal evidence that a large amount of learning takes place within the first 2 to 3 years of immersion followed by a relatively stable state (e.g., Derwing & Munro, 2013); and that this trend may be especially clear when analyses focus on the lexicogrammar and fluency aspects of L2 speech (e.g., Saito, 2015).

Given that Japanese communities are relatively small across the United States ACS Demographic and Housing Estimate (2020), recruiting participants from a single city was impossible. To feature as many late attainers as possible, an “inclusive” approach was adopted. Electronic flyers were first created that specified the necessary conditions for participating in the current project (i.e., AOA > 16 years, LOR > 6 years), and posted on several community websites for Japanese residents in the United States. Interested participants contacted the researcher and were invited to join a single remote research

session, where learner information and speech recordings were collected using the video-conferencing tool Google Hangouts. Approximately 80 Japanese residents participated in this initial phase of the study.

To ensure that the study only included participants who had ample opportunities to use L2 English (rather than L1 Japanese), and who had reached a relatively stable state of L2 performance (without dramatic change that could typically occur at the initial stage of L2 acquisition), the decision was made to include only motivated and regular L2 users. This additional scrutiny was crucial to eliminate certain L2 users who regularly used L1 Japanese. In addition, including nonhabitual L2 users may have confounded the primary focus of the current study, that is, the extent to which AOA predicts late L2 learners' vocabulary attainment, because these users' L2 English proficiency could be unrelated to their length of residence in the United States (entailing much room for improvement).

Only participants who rated their perceived frequency of L2 English as very frequent according to questionnaire responses ($M = 5.5$: 1 = very infrequent, 6 = very frequent) were featured in the study. This led to the retention of 41 participants (6 males, 35 females) as "attainers," all of whom reported that their main language of communication at work and/or at home was English. The methodology used here (i.e., self-report) was derived from the Language Experience Questionnaire (Saito, 2015, 2019) and the Language Contact Profile (Freed et al., 2004). For a similar screening process for identifying ultimate attainers, interested readers can refer to Abrahamsson and Hyltenstam (2009), Flege et al. (1995), and Saito (2015).

The participants' mean age at the time of the project was 40.1 years ($Range = 26$ – 53 years). All participants had arrived in the United States after the age of 18 ($M = 22.8$ years: $Range = 18$ – 37 years) and had been living there for an extensive period ($M = 15.5$ years: $Range = 6$ – 34 years). Prior to their arrival in the United States, the participants had studied L2 English in Japan for 7 to 9 years from junior high school to university. All the participants can be considered highly educated. In terms of the quality of L2 English education, they reported that instruction was typically provided using a grammar translation approach (very typical of Japanese English-as-a-Foreign-Language classrooms). No participants had prior experience attending English immersion schools (i.e., subject classes taught in L2 English).

Japanese controls (n = 40)

To provide a benchmark for assessing the attainers' lexical attainment, a total of 40 inexperienced Japanese learners (25 males, 15 females) were recruited at a university in Tokyo, Japan. They were college students with an average age of 19.6 years. Like the Japanese attainers, the Japanese controls had spent 6 years learning English in Japan (Grade 7–12). However, they lacked any experience traveling, living, or studying abroad (LOR = 0 years). Not surprisingly, they reported having few opportunities to use L2 English outside of required English lessons at the university (3 hours per week).

English controls (n = 10)

To provide a point of comparison for the upper limit/nativeness of the attainers' lexical proficiency, a total of ten native speakers of Canadian English (5 males, 5 females) were recruited in Ontario, an exclusively English-speaking area of Canada. These participants were considered as native speakers of General American English. All the English controls were postgraduate students at a university ($M_{age} = 27.7$ years), had grown up in English-speaking families (at least one of their parents was an L1 speaker of English), and reported that they used English exclusively on a day-to-day basis.

Data collection setup

As for the attainers, the data were collected through Google Hangouts. Online data collection was adopted out of necessity because the attainer participants resided in various cities across the United States and it was impossible to visit each of them and conduct face-to-face recordings. In contrast, the data of the Japanese and English controls was collected in a quiet room at a university in Japan and another university in Canada. Efforts were made to minimize the confounding effects of differing data-collection methods in the study (online vs. face-to-face). First, the speech data was transcribed for the purpose of vocabulary (rather than phonological) analyses. Second, the online sessions were administered individually with a researcher to help participants understand the procedure and monitor their performance throughout.

Participants (i.e., attainers, Japanese and English controls) were matched in terms of educational background (highly educated). The attainers differed from the Japanese controls in terms of the amount of naturalistic English use ($M = 15.5$ vs. 0 years), and from the English controls in terms of first language (Japanese vs. English).

Speech stimuli

To elicit sufficiently long speech samples and capture speaking proficiency from multiple angles, two different speaking tasks were employed: a storytelling task (Saito, 2019) and an oral interview task (e.g., Crossley et al., 2015). In the storytelling task, the participants first familiarized themselves with an eight-frame cartoon (1 min) and then explained the sequence of the events. The cartoon depicted the following scene: On a corner of a busy street in some metropolitan city, a man and a woman bumped into each other carrying the same suitcase and ended up swapping suitcases. In the interview task, the participants received a card from a researcher containing an assigned topic (i.e., *What was the hardest and toughest change in your life?*) and a set of possible discussion points. After 1 minute of preparation, they were asked to speak for approximately 2 minutes. Finally, the researcher asked one or two follow-up questions about the content of their speech (e.g., *What did you learn from the experience?*); for copies of the materials used in the study, see Supporting Information-A).

According to Skehan's (1998) taxonomy, these two tasks were assumed to require different kinds of cognitive resources. The storytelling task (formal, narrative, structured, less familiar information) was assumed to push learners to prioritize producing accurate language without demanding much conceptualization (i.e., what to say). In this study, therefore, the storytelling task was a more structured, *accuracy-driven* task. In contrast, it was supposed that the interview task (informal, personal, less structured, more familiarity of information) would stimulate L2 speakers' conceptualization ability to a great degree, resulting in potentially more complex, sophisticated language (at the expense of accurate language). Thus, the interview task was a less structured, *complexity-driven* task.

To control for the effects of pronunciation quality on the L2 vocabulary analyses, all recordings were transcribed and cleaned by removing filled pauses (e.g., "ah," "eh," "oh," "um"). The researcher familiarized themselves with a range of problematic pronunciation features typical of Japanese speakers of English (Saito, 2013). In the case of unclear speech, the researcher transcribed what the talker intended to say as inferred from context (e.g., *life* pronounced as *rife* was still spelled as *life*). For a similar methodological decision, see Crossley et al. (2015).

To ensure the robustness of the vocabulary analyses, efforts were made to elicit sufficiently long samples for rating ($x > 100$ words; Koizumi & In'nami, 2012). All the experienced Japanese attainers and English controls easily passed this threshold. Given that the inexperienced speakers were recruited to proxy the lower end of late L2 learners' lexical proficiency, it was unsurprising that some of them demonstrated difficulty producing more than 100 words. To examine the lexical characteristics of low-level L2 speech (relative to advanced L2 speech), all the files were included without any screening. As a result, the duration of all transcripts varied widely for both the storytelling task ($M = 161.9$ words, $Range = 61\text{--}424$ words) and the interview task ($M = 333.6$ words, $Range = 72\text{--}939$ words).

Appropriateness analyses

Traditionally, many scholars analyzed L2 learners' abilities to choose semantically appropriate words with accurate morphosyntax markers by counting the number of erroneous instances in a certain sentence unit (e.g., Yuan & Ellis, 2003 for per clause). However, some scholars have argued that such dichotomous analyses of individual words (correct or incorrect) fail to capture the potential difference in degree of impact each word has on communicative adequacy (Foster & Wigglesworth, 2016). As a remedy, holistic judgment approaches have been proposed wherein linguistically trained raters evaluate the overall (rather than word-by-word) appropriateness of transcripts using a 4-point scale (1 = *entirely accurate*; 2 = *minor errors*; 3 = *serious errors*; 4 = *very serious errors hindering meaning conveyance*) (Foster & Wigglesworth) or a 6-point scale (1 = *minimum accuracy*, 6 = *maximum accuracy*) (Crossley et al., 2015).

Saito, Webb, Trofimovich, and Isaacs (2016) and Saito, Trofimovich, and Isaacs (2017) proposed, validated, and refined the expert judgment approach for assessing the phonological, lexical, and morphosyntactic accuracy of L2 speech. As for lexical appropriateness (the focus of the current study), linguistically trained raters first receive training on how to evaluate the appropriateness of words in context within a range of L2 speech transcripts (for training scripts, see Supporting Information-B). Here, the training clarifies that lexical appropriateness is different from phonological accuracy (i.e., the correct pronunciation of individual sounds) and morphosyntactic accuracy (i.e., the correct assignment of morphological markers). For each transcript, an overall rating is assigned using a moving slider (recorded on a 1,000-point scale). The ends of the continuum are labeled "many inappropriate words" (0 points) and "consistently appropriate" (1,000 points). After they have fully understood the concept of lexical appropriateness (which is distinguishable from phonological and morphosyntactic accuracy), they practice the procedure with three transcripts, receive feedback from a trained research assistant, and then move onto the main analyses.

Saito et al. (2017) examined the relationship between the judgments of 10 expert raters with MA degrees in applied linguistics and the objective analyses of 40 L2 speech transcripts. The results showed that the raters' appropriateness scores were significantly associated with the actual number of lexical errors in transcripts ($r = .50$). More importantly, the expert raters' appropriateness judgments more strongly correlated with the overall comprehensibility of the L2 samples than the objective error analyses did ($r = .68$ vs. $.52$). Not only has the expert judgment method been used in a range of L2 contexts (e.g., Ruivivar & Collins, 2018) but it has also been found to capture how aspects of lexical accuracy in L2 speech develop over time (Saito & Hanzawa, 2018).

Raters

A total of five native speakers of English (2 males, 3 females) were recruited from an American university to assess the lexical appropriateness of the samples. Following Isaacs and Thomson (2013), the raters were considered as “experts” because all were MA students in the department of linguistics and reported extensive L2 English teaching experience ($M = 5.5$ years, $Range = 3-9$ years).

Procedure

The transcripts were presented to the raters in a randomized order using the MATLAB software. A moving slider was used to assess the lexical quality of each token. If the slider was placed at the leftmost end of the continuum, labeled with a frowning face (indicating “nontargetlike”), the rating was recorded as 0. If the slider was placed at the rightmost end of the continuum, labeled with a smiley face (indicating “targetlike”), the rating was recorded as 1,000. As operationalized in previous studies (Saito et al., 2017), a trained researcher first provided the raters with a detailed explanation on what characterized lexical accuracy (see Supporting Information-B). Next, the raters practiced the rating procedure with three transcripts that were not included in the main dataset. For each transcript, they were asked to justify their rating decision, and received feedback from a trained research assistant to ensure that they correctly understood and applied the rubrics without confusion. Finally, they proceeded to rate all 91 transcripts (41 attainers + 40 Japanese Controls + 10 English Controls). The raters assessed the storytelling task on Day 1 and the interview task on Day 2. Each individual session took approximately 60–70 minutes.

Interrater reliability

According to the results of a postrating questionnaire (9-point scale), all raters indicated that they clearly understood the rating category ($M = 8.9$, $Range = 8-9$). Cronbach’s alpha analyses indicated a relatively strong agreement between raters’ assessments ($\alpha = .88$ for storytelling; $.87$ for interview), suggesting that individual variance among the ratings was minor. Because of this, all raters’ scores were averaged to generate a single mean score for each sample (for a similar methodology, see Ruivivar & Collins, 2018).

Richness analyses

Following Crossley’s L2 vocabulary use framework, the transcripts were analyzed for lexical richness using several indices from the Tool for the Automated Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015). These indices are assumed to represent three broad categories of lexical richness: (a) frequency (i.e., the use of infrequent words); (b) n-gram (i.e., the frequency and association of multiword units); and (c) abstractness (i.e., the use of less concrete words).

Frequency

This index represents the relative frequency of word use in various social and academic settings (based on large spoken and written corpora). More advanced L2 learners tend to rely less on frequent words and make greater efforts to incorporate infrequent words in their speech (e.g., Crossley et al., 2009). For each transcript,

frequency was calculated by averaging the values of all content words with reference to a frequency list of 51 million words in the SUBTLEX corpus—a corpus of subtitles in American movies and TV series (Brysbaert & New, 2009). For example, “ponder” (56 counts) and “eavesdrop” (40 counts) occur less frequently than “think” (137,261 counts) and “listen” (27,784 counts) in the SUBTLEX database.³ Raw frequency scores were first averaged. As recommended by Kyle and Crossley (2015), these scores were then logarithmically transformed to control for Zipfian effects common with word frequency lists (i.e., intensive use of the first 3,000 to 4,000 word families). Larger SUBTLEX scores indicate the use of more infrequent words within a given transcript. Thus, a transcript with higher SUBTLEX scores is assumed to represent more advanced L2 lexical proficiency.

N-gram

This index refers to the relative frequency and association of n-word combinations in a reference corpus. While n-gram has been considered as one measure of collocation (for a comprehensive review, Gablasova et al., 2017), the predictive power of raw n-gram frequency scores for L2 lexical proficiency remains unclear. This is because raw frequency scores do not distinguish between semantically and structurally complete sequences (e.g., “go” and “to”) and random co-occurrences of incomplete lexical items (e.g., “go” and “he”).⁴ To capture the strength of meaningful associations (i.e., greater than chance), a range of adjusted measures have been devised. In the current study, Bigram Frequency was calculated based on the spoken subset of the Corpus of Contemporary American English (COCA; Davies, 2010). Following Gablasova et al.’s recommendation, a mutual information (MI) score was calculated by dividing the bigram frequency scores by the frequency of the random co-occurrence of the two words. The ratio value of all the two-word combinations was logarithmically transformed for each text. MI bigram scores have been found to simulate native speakers’ recognition and production of formulaic sequences (Ellis et al., 2008) and native speakers’ judgments of overall L2 oral proficiency (Saito & Liu, 2022; Tavakoli & Uchihara, 2020). Further, more advanced L2 learners tend to combine words that are less frequent, more abstract, and more complex in their writing (Kyle & Crossley, 2016) and speech (Kyle & Crossley, 2015). Here, higher MI scores represented higher L2 lexical proficiency.

Abstractness

This index refers to native speakers’ subjective judgments of word properties in relation to the degree of abstractness. To this end, perceived concreteness (i.e., the degree to which words are relevant to here-and-now concepts) was calculated from the MRC psycholinguistics database (Coltheart, 1981) and two recently added databases (Brysbaert et al., 2014; Kuperman et al., 2012). Total scores were calculated by dividing the sum of the psycholinguistic norm scores by the number of words that were assigned psycholinguistic norm scores. Lower scores corresponded to the use of less concrete (i.e., more abstract) words per sample.

³All the examples in the “Methods” section derive from the transcripts analyzed in the current study.

⁴The data presented here was produced by TAALES. According to the manual, all the bigram/trigram counts are normalized per million words: (actual frequency / number of words in sub-corpus)*1000000.

Results

Attainers versus Japanese/English controls

The first objective of the statistical analyses was to explore what characterized experienced Japanese learners' (attainers) vocabulary use (appropriateness, frequency, n-gram, and abstractness) relative to the inexperienced Japanese and English controls. By comparing the Attainers and Japanese Controls, I aimed to illustrate how much interlanguage development had taken place from the Attainers' presumed starting point (i.e., inexperienced Japanese learners' performance). Further, I focused on the differences between Attainers and English Controls to gauge the extent to which the Attainers had reached nativelike L2 performance (i.e., high-level ultimate attainment). Descriptive statistics of raw scores are summarized in Supporting Information-C and -D.

It should be noted that the sample size of English Controls was small ($n = 10$) relative to Japanese Controls ($n = 40$) and Japanese Attainers ($n = 41$). To detect statistical significance, therefore, a set of conservative nonparametric tests (Mann–Whitney U tests) were performed with four different lexical factors (Appropriateness, Frequency, Specificity, Abstractness) as dependent variables, and group as the independent variable with three levels (Japanese Controls, Attainers, English Controls), for each of the two tasks (storytelling, interview). Because three comparisons were made (i.e., Japanese vs. English Controls, Japanese Controls vs. Attainers, Japanese Attainers vs. English Controls), alpha was set to .0125 (Bonferroni corrected). The results of the analysis are summarized in Table 1.

The results showed that Japanese and English Controls' performance differed significantly for all lexical dimensions in the storytelling task ($p < .0125$). Interestingly, their n-gram performance was comparable in the interview task ($p = .186$). This indicates whereas spoken L2 vocabulary learning can be generally characterized as the more appropriate use of more infrequent and abstract words, the use of n-gram could be task-specific and serve as a crucial predictor of nativelikeness in the more formal and structured task (storytelling) but not in the more informal and less structured task (interview).

Table 1. Summary of nonparametric tests for three different group comparisons (Japanese vs. English Controls, Japanese Controls vs. Attainers, and Japanese Attainers vs. English Controls)

		A. Storytelling		B. Interview	
		Z	p	z	p
Japanese vs. English Controls	Appropriateness	-4.851	< .001*	-4.309	< .001*
	Frequency	-2.838	.005*	-2.003	.011*
	N-gram	-4.729	< .001*	-1.323	.186
	Abstractness	-2.595	.009*	-3.364	.001*
Japanese Controls vs. Attainers	Appropriateness	-6.919	< .001*	-4.581	< .001*
	Frequency	-1.134	.257	-2.868	.004*
	N-gram	-7.085	< .001*	-1.563	.118
	Abstractness	-5.923	< .001*	-5.274	< .001*
English Controls vs. Attainers	Appropriateness	-4.152	< .001*	-4.104	< .001*
	Frequency	-2.325	.020	-0.142	.887
	N-gram	-0.142	.887	-0.261	.794
	Abstractness	-1.851	.064	-0.664	.507

* $p < .0125$ (Bonferroni corrected).

Next, I examined the extent to which Attainers could improve their spoken L2 vocabulary performance after years of immersion relative to Japanese Controls. Whereas the two groups were distinguishable in terms of appropriateness and abstractness in both tasks ($p < .0125$), more advanced L2 speakers appeared to use more infrequent words than the control group only when they were induced to engage in more conceptualization (in the interview task). Finally, the comparison of Japanese Attainers versus English Controls revealed similarities in frequency, n-gram, and abstractness ($p = .020-.887$) but significant differences in appropriateness ($p < .001$).

Taken together, the comparison analyses led to three overall observations: (a) spoken L2 vocabulary development can be observed in terms of appropriateness, frequency, and abstractness (Japanese vs. English Controls); (b) the appropriateness and abstractness of spoken L2 vocabulary improves with increased immersion experience (Japanese Controls vs. Attainers); and (c) highly experienced L2 speakers may attain nativelike abstractness but not appropriateness.⁵

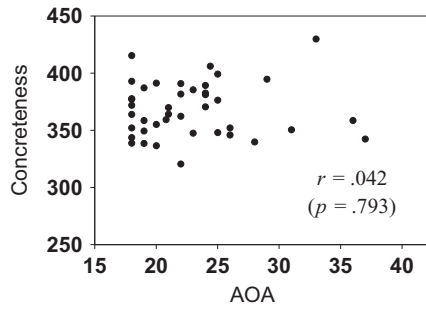
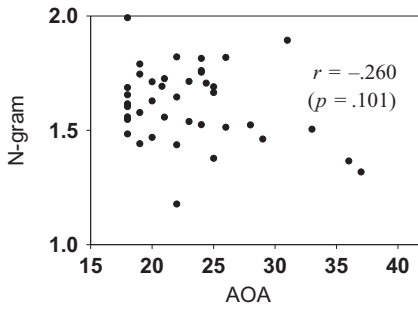
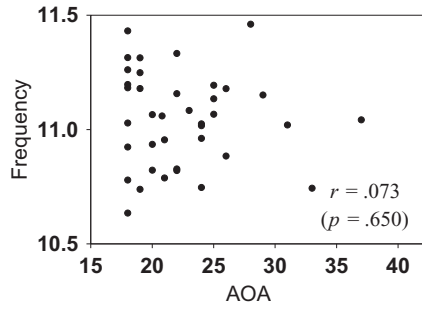
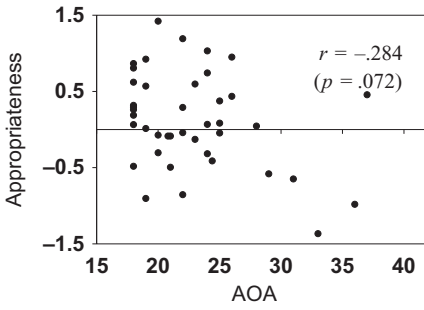
Age of acquisition versus attainment

The next objective of the analyses was to examine the correlations between the attainers' ($n = 41$) lexical performance and their age of arrival in the United States (varying between 18 and 37 years). According to the results of Kolmogorov–Smirnov tests, the participants' lexical attainment scores (appropriateness) were normally distributed without any outliers ($p > .05$). Similarly, Kolmogorov–Smirnov tests did not find their age of arrival and length of residence profiles to significantly deviate from a normal distribution ($p > .05$). Thus, Pearson correlation analysis was conducted to check the significance and strength of the associations between the lexical proficiency scores (appropriateness, frequency, n-gram, and abstractness) and AOA profiles, with alpha set to .0125 (Bonferroni corrected). As shown in [Figure 1](#), the participants' lexical appropriateness scores were significantly associated with AOA profiles in the interview task ($r = -.429, p = .005$). Yet, the correlation between appropriateness and AOA was weak in the storytelling task ($r = -.284, p = .072$). For frequency, n-gram, and abstractness, however, the AOA-appropriateness functions were substantially far from statistical significance ($p > .1-.8$).

Some scholars have argued that AOA effects could be confounded with a range of experience factors (i.e., younger arrivals could practice more than older arrivals; Flege et al., 1995). Thus, to separate any influence of these factors from the effects of AOA, I conducted a series of partial correlation analyses controlling for the two confounding variables: length of residence (the number of years in the United States) and daily L2 use (the frequency of L2 English use). Like many L2 speech studies (e.g., Flege et al., 1995), participants' self-report scores (on a 6-point scale: 1 = very infrequent, 6 = very frequent) were used to index the frequency of L2 use daily at the time of the project. With LOR and L2 use factored out, the relationship between the attainers' AOA and lexical appropriateness scores remained

⁵Notably, the mean age of testing among the attainer group (40.1 years for Japanese Attainers; 35.7 years for Polish Attainers) was considerably higher than for the control groups (19.6 years for Japanese Control; 27.7 years for English Control). In age-related literature, although few studies have identified the predictive power of chronological age for the phonological accuracy aspects of L2 speech attainment (e.g., Flege et al., 1995), there is some evidence for a potential link between age of testing and L2 fluency attainment (Lahmann et al., 2017). Future studies can further explore this topic by including a range of L2 learners and attainers with a range of age profiles.

A. Storytelling Task



B. Interview Task

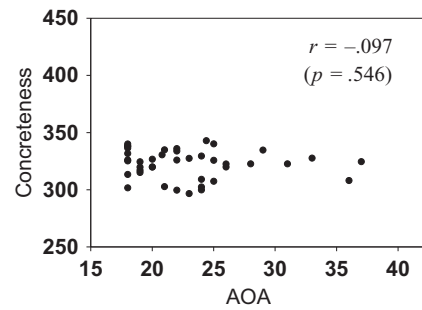
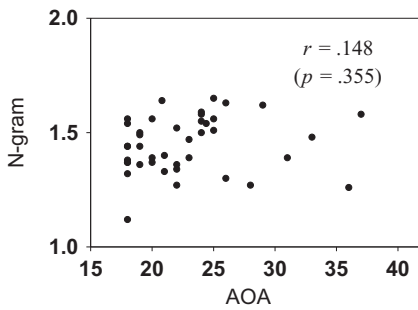
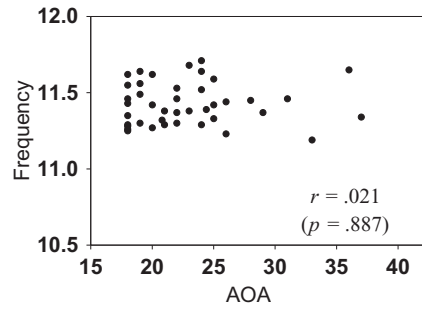
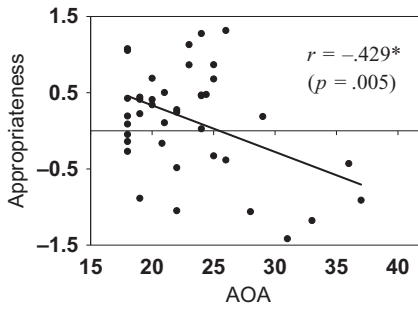


Figure 1. Spoken vocabulary proficiency scores plotted as a function of AOA.

* $p < .0125$.

significant in the interview task ($r = -.409, p = .010$), and remained nonsignificant in the storytelling task ($r = -.269, p = .098$).

Study 2

To provide the conceptual setup for the current study, the first part of the analyses in Study 1 examined the lexical characteristics of L2 attainers' speech relative to those of L1 speakers and inexperienced L2 speakers. The overall findings (i.e., L2 attainers' speech being distinguishable especially by accuracy rather than richness) concurred with Lindqvist et al. (2011) and Bartning et al. (2012). As the focus of the paper, the analysis was then conducted on the mediating role of AOA in L2 spoken vocabulary attainment. It was found that AOA was associated with appropriateness, a relatively difficult dimension of spoken L2 vocabulary proficiency, when speech was elicited using a less structured, complexity-driven task (i.e., interview). To examine the generalizability of this key finding, the analysis was replicated with a different group of L2 attainers: $n = 50$ Polish speakers of English (AOA = 17–32 years; LOR = 6–19). Given that the primary objective, novelty, and controversy of the paper lay in the AOA-acquisition link, Study 2 concerned only the analyses of late L2 attainers' lexical performance in the interview task in particular. It did not further pursue the replication of the baseline data (L1 speakers, inexperienced L2 speakers) as the findings (L2 attainers vs. L1 and L2 Controls) were already in line with Lindqvist et al. (2011) and Bartning et al. (2012).

Participants

The speech data came from a larger project focused on collecting speaking samples from 500+ L2 English speakers with varied L1 backgrounds in London, UK. The current study used spoken samples produced by a total of 50 Polish speakers ($M_{age} = 35.7$ years). The Polish participants engaged in a range of speaking tasks including the storytelling and interview tasks used in Study 1. Given that the AOA-acquisition link was found only in the interview, this task was chosen for further investigation. Interview data was transcribed and submitted to the spoken vocabulary analyses. Speech data from the other tasks (including storytelling) were not further analyzed. Given that one methodological limitation of Study 1 was the unequal number of males and females in the attainer group ($n = 6$ males, 35 females), efforts were made to reach a gender balance in the population ($n = 25$ females, 25 males). These speakers were considered to be late L2 attainers given that they met the same criteria adopted in Study 1: (a) they arrived in the United Kingdom after puberty ($M_{AOA} = 23.1$ years; $Range = 17$ – 32 years); (b) they had resided in the United Kingdom for at least six years ($M_{LOR} = 11.9$ years; $Range = 6$ – 19 years); (c) their primary language of communication at work and/or at home was English; and (d) they rated their perceived percentage of daily L2 English use as relatively frequent ($M = 73.7\%$; $Range = 56$ – 100%). The relationship between the *phonological* dimensions of their L2 proficiency (assessed using multiple comprehension and production tasks) and a range of individual difference factors (e.g., L2 use, aptitude, motivation, awareness) has been reported in other publications (e.g., Saito et al., 2022).

Data-collection setup

Whereas the data in Study 1 was collected from 50 Japanese attainers using a video-conferencing tool, the speech of the 50 Polish attainers in Study 2 was recorded in face-

to-face settings. As such, the analyses allowed us to test the generalizability of findings (i.e., age effects in L2 vocabulary attainment) across two different elicitation conditions (online vs. face-to-face).

Speech stimuli

The same interview task and prompts in Study 1 were used to elicit the participants' speech. The data collection took place individually in tandem with a trained research assistant in a quiet room at a university in London. The speech was recorded with a Roland-05 audio recorder, set at 44.1 kHz sampling rate and 16-bit quantization. The speech tokens were transcribed using the same procedures as in Study 1 ($M_{length} = 551.5$ words; $Range = 248-1,242$ words).

Appropriateness and richness analyses

A total of three native speakers of English (2 males, 1 female) were recruited in the United Kingdom for the appropriateness analysis. All had obtained a university degree in applied linguistics, had extensive experience in L2 English teaching ($M = 6.5$ years, $Range = 1-15$ years), and had conducted similar kinds of lexical analysis in the past. As in Study 1, the raters assessed the 50 transcripts for lexical appropriateness using the MATLAB-based software. According to the results of a postrating questionnaire (9-point scale), all raters reported a very clear understanding of the category ($M = 9$). Because the raters' assessments demonstrated strong agreement (Cronbach $\alpha = .95$), ratings were averaged across raters to generate a single rating per speaker.

For the richness analyses, the participants' transcripts were analyzed for frequency (SUBTLEX frequency), n-gram (Bigram MI), and abstractness (MRC concreteness ratings) using TAALES (Kyle & Crossley, 2015).

Results

Descriptive statistics of the participants' spoken L2 vocabulary scores (appropriateness, frequency, n-gram, and abstractness) are presented in Supporting Information-E. According to the results of the normality tests (Kolmogorov-Smirnov), the richness scores (frequency, n-gram, and abstractness) did not significantly differ from a normal distribution ($D = .056-.105$, $p = .596-.994$). Yet, significant deviation was observed for the appropriateness scores ($D = .236$, $p = .006$). Participants' AOA and LOR profiles followed normal distribution ($D = .112, .105$, $p = .511, .598$). Thus, to examine the AOA-proficiency link, Pearson correlation analysis was performed for frequency, n-gram, and abstractness scores, while Spearman correlation analysis was performed for appropriateness. Alpha level was set to .0125 (Bonferroni corrected). As visually summarized in Figure 2, AOA was not significantly associated with frequency ($r = -.185$, $p = .199$), n-gram ($r = -.135$, $p = .348$), or abstractness ($r = .258$, $p = .071$). Conversely, a significant link was found between participants' AOA and appropriateness ($r = -.406$, $p = .003$). To check for the presence of any confounding effects based on immersion experience, a partial correlation analysis was performed on participants' ranked AOA and appropriateness scores while controlling for LOR and percentage of reported L2 use. The results showed that AOA remained a significant predictor of appropriateness ($r = -.416$, $p = .003$).

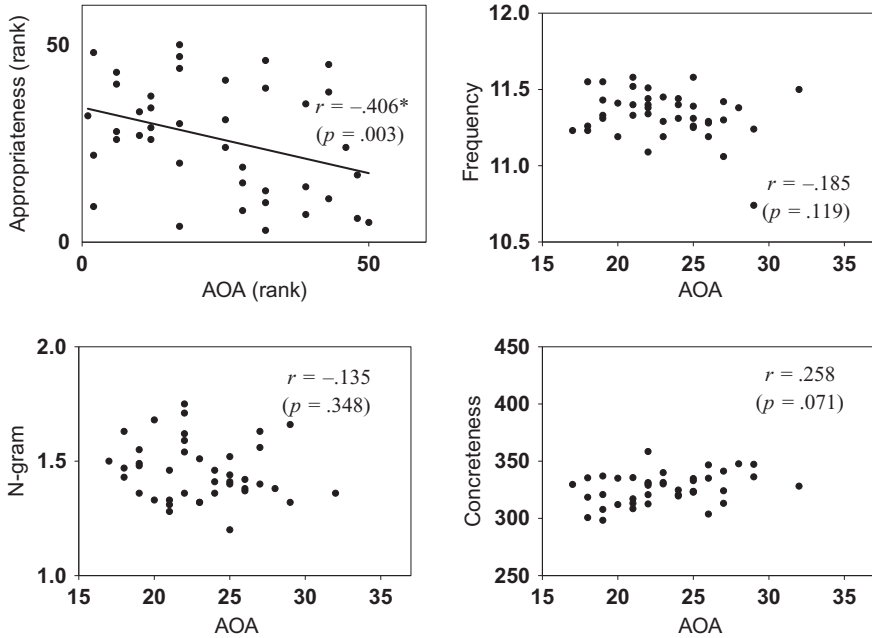


Figure 2. Spoken vocabulary proficiency scores plotted as a function of AOA.

* $p < .0125$.

Discussion

Building on the extensive literature addressing the effects of AOA on the phonological proficiency of late bilinguals (e.g., Derwing & Munro, 2013), the current study set out to examine whether AOA influences the lexical aspects of L2 speech attainment. Phonological factors were controlled for by transcribing and comparing the spontaneous speech of 41 experienced Japanese speakers (AOA 18–37 years; LOR = 6–34 years) and 50 experienced Polish speakers (AOA = 17–32 years; LOR = 6–19) elicited using more and less structured tasks (storytelling, oral interview). Drawing on Crossley’s computational modeling of spoken L2 lexical proficiency, the transcripts were analyzed for two major dimensions: appropriateness (the accurate use of words in contexts) and richness (the use of more infrequent, collocational, and abstract words).

Overall, the results showed (a) that experienced learners demonstrated nativelike attainment in the richness dimension of their spoken L2 vocabulary use (frequency, n-gram, and abstractness); (b) that their appropriateness was subject to a great deal of individual variation; and (c) that the incidence of high-level L2 speech (appropriateness *and* richness) was determined by AOA profiles (18–37 years for Japanese; 17–32 years for Polish). The strength of the correlation coefficients in the current study ($r = -.429$ in Study 1; $r = -.406$ in Study 2) were comparable to the previous findings in L2 morphosyntax and phonology (i.e., moderate effect size). According to Qureshi’s (2016) meta-analysis of L2 morphosyntax attainment research ($n = 20$ primary studies), the overall effects of AOA on early and late L2 learners’ grammaticality judgment outcomes were $Zr = -0.40$ (i.e., $r = -.038$). Similar association strengths were found in various areas of L2 phonological attainment (Saito, 2015 for $-.3$ to $-.4$ among $n = 88$ late L2 attainers).

The findings regarding AOA and the different aspects of lexical proficiency (i.e., significant age effects for accuracy rather than richness) can be interpreted with reference to two contrastive views on the role of age in late L2 speech learning, summarized again here. On the one hand, the findings are not compatible with the strong version of the maturational account of SLA that assumes that postpubertal L2 learning is unrelated to AOA. On the other hand, the findings align with the non-maturational account, which predicts that a significant AOA-acquisition link remains present in late L2 vocabulary attainment.

The presence of age effects for the appropriateness (but not richness) dimension of spoken lexical proficiency is worth further scrutiny. These effects could be related to the differential learning difficulty inherent to the three dimensions. Notably, developing lexical appropriateness is thought to be more difficult than developing richness, especially with tasks that push participants to attend to complexity rather than accuracy (interview; Skehan, 1998). This was borne out in the current study, where most of our experienced participants were able to reach nativelike proficiency for collocational and abstract word use (i.e., no group differences between Japanese Attainers and English Controls) but not for appropriateness (i.e., significant gaps between Japanese Attainers and English Controls).

There is ample evidence that L2 learners have more difficulty developing appropriateness than richness. Many L2 learners' lexical repertoires grow to a great degree after they have experienced short-term immersion (Salsbury et al., 2011) and/or long-term foreign language education (Saito, 2019), and some may reach nativelike vocabulary size and depth of word knowledge (Hellman, 2011). Despite ample exposure experience, however, few L2 speakers can attain the ability to choose word combinations in not only a contextually appropriate but also a nativelike fashion, especially when these words are embedded in more extemporaneous, paragraph-length texts (e.g., Foster et al., 2014). Indeed, expert raters' judgments of advanced L2 spoken vocabulary proficiency is mainly determined by appropriateness rather than richness (Crossley et al., 2015).

Taken together, it can be tentatively stated that AOA is a significant predictor of the acquisition of relatively difficult aspects of L2 lexical attainment (i.e., appropriateness) under complexity-driven (rather than accuracy-driven) task conditions. This "difficulty argument" has also been made in the L2 pronunciation literature. For example, Saito (2013) found that many of the late Japanese attainers could achieve nativelike proficiency of English /r/ by redeploing already existing articulatory configurations in their L1 phonetic systems (i.e., degree and rate of tongue retraction). However, age effects were particularly strong for the acquisition of new articulatory parameters specific to the L2 English phonetic system (i.e., labial, alveolar, and pharyngeal constrictions). This argument also finds support at a broader level, as many studies have shown that late L2 learners can attain advanced-level and nativelike L2 fluency (Trofimovich & Baker, 2006), but that AOA matters for the reduction of foreign accentedness (Flege et al., 2006), and the development of segmental and prosodic accuracy—relatively difficult aspects of L2 oral proficiency development (Saito, 2015).

It is important to note that three similar studies failed to find that AOA played a significant role in late L2 attainers' lexical proficiency (i.e., Abrahamsson & Hyltenstam, 2009; Granena & Long, 2013; Spadaro, 2013; but see Hellman, 2011). These conflicting findings could be ascribed to the different foci of the analyses. These three studies highlighted participants' lexical *comprehension* proficiency, using a range of word and collocation judgment tasks. The current study explored the use of vocabulary in L2 attainers' spontaneous speech *production*. This raises the possibility that AOA effects can be more clearly observed when attainers are asked to *produce*

(rather than *comprehend*) language. This interpretation might provide support for Birdsong's (2006) arguments about the age-related decline in motor abilities and working memory and their impact on L2 attainment. To further examine this topic, future studies can not only measure late L2 attainers' lexical comprehension and production abilities but also their motor sequence abilities (e.g., Thompson et al., 2015 for aging effects in audio-motor integration).

Future directions

The presence of a significant AOA-acquisition link in late L2 vocabulary attainment supports the nonmaturation view that AOA similarly determines the degree of success in both early and late L2 acquisition. One crucial theoretical implication of this perspective is that late learners could continue to learn a new language by drawing on the same mechanisms used in successful L1 acquisition, as long as they use their L2 on a daily basis. Here, it is important to point out that scholars have extensively discussed precisely what kinds of bilingual, perceptual-cognitive, and sociopsychological factors underlie such lifelong age constraints in both early and late bilingualism (see Mayberry & Kluender, 2018 for a comprehensive overview).

For example, one view is that the attainment of early bilinguals becomes more nativelike because they are less affected by the interaction between the L1 and L2 systems. In other words, the earlier a learner arrives, the less the L1 is developed, and the more targetlike their L2 ability becomes (i.e., bilingual effects; Best & Tyler, 2007; Flege, 2018). Another explanation relates to perceptual-cognitive aging. From this perspective, late L2 learning takes place in a common linguistic space, wherein the L1 system is fully developed, and is subject to age-related changes in cognitive functioning caused by decreased brain size, working memory, processing speed, and attentional control (i.e., cognitive effects; Birdsong, 2005, 2006; Hakuta et al., 2003). Similarly, there is emerging evidence that precise perceptual acuity gradually declines throughout adulthood, and that this ability determines the extent to which L2 learners can make the most of each opportunity for input (see Saito et al., 2020 for auditory perception effects). Finally, it has been reported that the quantity and quality of L2 input which late bilinguals are able to access is relatively limited (Bialystok, 1997, for sociopsychological effects). Differing from early bilinguals, who are immersed in the target language community from the onset of L2 learning, late bilinguals can choose whether to have more interaction opportunities in their L1 or L2 (Derwing & Munro, 2013); or their L2 use may be limited to certain contexts (e.g., the home; Jia & Aaronson, 2003).

One promising future direction concerns the inclusion of multiple individual difference measures to further disentangle why an earlier AOA benefits late L2 speech acquisition. To this end, such future studies are strongly encouraged to (a) inspect not only L2 but also L1 performance to reveal the degree of L1-L2 attrition (e.g., Baker et al., 2008); (b) scrutinize the perceptual-cognitive profiles of participants in various age groups (i.e., e.g., Darcy et al., 2016 for executive functions; Saito et al., 2020, for auditory perception; Linck et al., 2013 for working memory); and (c) survey participants' motivation, personality, and communicative orientations toward L1/L2 use and the target language community (e.g., Derwing & Munro, 2013). Once more evidence becomes available, we will have a better picture of how bilingual, perceptual-cognitive, and sociopsychological factors differentially impact language acquisition at different stages of life.

Another important future direction is concerned with the reexamination of this study's methodological decisions and their potential impact on the findings. Drawing on

Crossley's computational modeling of spoken L2 vocabulary use, appropriateness was operationalized using a holistic, rater-evaluated measure based on a myriad of various assumptions or dimensions of transcribed oral narratives that are perceptible to raters. Richness was operationalized using a set of corpus-based measures based on individual words and word combinations used by speakers. Although AOA was significantly associated with appropriateness rather than richness, we have yet to know precisely (a) what aspects of vocabulary use the raters attended to during their appropriateness judgments and (b) how and why such lexical elements could be amenable to age-related decline. Here, I call for further research to examine the linguistic correlates of raters' "lexical appropriateness" judgments especially when they assess the quality of highly experienced L2 speakers' speech transcripts. To this end, we may need to wait for future empirical studies to develop and validate more objective, specific, and fine-grained measures of lexical appropriateness at different ability levels (cf. Saito & Liu, 2022 for the use of collocational association [MI] to distinguish between low-to-intermediate L2 speakers' lexical appropriateness).

Conclusion

The results of current study suggest that (a) L2 learners can substantially improve various dimensions of their L2 lexical proficiency (appropriateness, richness) as a function of increased L2 experience; (b) L2 learners can demonstrate nativelike proficiency for relatively easy lexical dimensions of speech (i.e., richness); and (c) AOA plays a key role in high-level L2 lexical attainment, especially when it comes to difficult dimensions with much room for improvement (i.e., appropriateness). These findings echo the nonmaturational position of SLA, which holds that AOA-related differences can be observed not only in the phonological but also in the lexical dimensions of L2 speech attainment throughout the life span. The current study adds that extensive experience with the L2 may allow learners to produce abstract words in a manner indistinguishable from native speakers. In the long run, however, AOA may still determine the extent to which L2 learners can attain high-level proficiency in both lexical appropriateness and richness.

Acknowledgment. This study was funded by the Grant-in-Aid for Scientific Research Japan (No. 26770202), Economic and Social Research Council (ES/S013024/), and Leverhulme Trust (RPG-2019-039). I gratefully acknowledge the following team members who helped with data collection and analyses at various stages of the project: Yuka Akiyama, Kokoro Muramoto; Takumi Uchihara; Konstantinos Macmillan; Sascha Kroeger; Viktoria Magne; Kotaro Takizawa; and Magdalena Kachlicka.

Data Availability Statement. The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at osf.io/pg9ua

Competing Interests. The author declares no competing interests.

References

- Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition*, 34, 187–214.
- Abrahamsson, N., & Hyltenstam, K. (2009). Age of acquisition and nativelikeness in a second language—listener perception vs. linguistic scrutiny. *Language Learning*, 59, 249–306.
- ACS Demographic and Housing Estimates (2020). 2017 *American Community Survey*. United States Census Bureau. <https://data.census.gov/cedsci/table?d=ACS%205-Year%20Estimates%20Data%20Profiles&table=DP05&tid=ACSDP5Y2017.DP05>

- Baker, W., Trofimovich, P., Flege, J. E., Mack, M., & Halter, R. (2008). Child-adult differences in second-language phonological learning: The role of cross-language similarity. *Language and Speech*, 51, 316–341. <https://doi.org/10.1177/0023830908099068>
- Bartning, I., Lundell, F. F., & Hancock, V. (2012). On the role of linguistic contextual factors for morpho-syntactic stabilization in high-level L2 French. *Studies in Second Language Acquisition*, 34, 243–267.
- Best, C., & Tyler, M. (2007). Nonnative and second-language speech perception. In O. Bohn & M. Munro (Eds.), *Language experience in second language speech learning: In honour of James Emil Flege* (pp. 13–34). John Benjamins.
- Bialystok, E. (1997). The structure of age: In search of barriers to second language acquisition. *Second Language Research*, 13, 116–137.
- Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 109–127). Oxford University Press.
- Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning*, 56, 9–49.
- Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second language acquisition. *Journal of Memory and Language*, 44, 235–249.
- Brybaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977–990.
- Brybaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904–911.
- Bundgaard-Nielsen, R., Best, C., & Tyler, M. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, 33, 433–461.
- Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33, 497–505.
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, 59, 307–334.
- Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, 41, 721–744. <https://doi.org/10.1017/S0272263118000268>
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60, 573–605.
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, 36, 570–590.
- Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25, 447–464.
- Darcy, I., Mora, J. C., & Daidone, D. (2016). The role of inhibitory control in second language phonological processing. *Language Learning*, 66, 741–773. <https://doi.org/10.1111/lang.12161>
- 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.
- DeKeyser, R., Alfi-Shabta, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, 31, 413–438.
- DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning*, 63, 52–67. <https://doi.org/10.1111/j.1467-9922.2012.00737.x>
- DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 88–108). Oxford University Press.
- Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A seven-year study. *Language Learning*, 63, 163–185.
- Ellis, N. C., Simpson-Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42, 375–396. <https://doi.org/10.1002/j.1545-7249.2008.tb00137.x>
- Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. In A. Meyer and N. Schiller (Eds.), *Phonetics and phonology in language comprehension and production, differences and similarities* (pp. 319–355). Mouton de Gruyter.

- Flege, J. E. (2018). It's input that matters most, not age. *Bilingualism: Language and Cognition*, 21, 919–920.
- Flege, J., & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition*, 23, 527–552. <https://doi.org/10.1017/S0272263101004041>
- Flege, J., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, 34, 153–175.
- Flege, J., Munro, M., & MacKay, I. R. A. (1995). Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, 97, 3125–3134.
- Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /l/ and /l/ accurately. *Language and Speech*, 38, 25–55.
- Flege, J., Yeni-Komshian, G., & Liu, S. (1999). Age constraints on second language acquisition. *Journal of Memory & Language*, 41, 78–104.
- Foster, P., Bolibaug, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition*, 36, 101–132.
- Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, 36, 98–116.
- Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition*, 26, 349–356.
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67, 155–179.
- Gatbonton, E., Trofimovich, P., & Segalowitz, N. (2011). Ethnic group affiliation and patterns of development of a phonological variable. *The Modern Language Journal*, 95, 188–204.
- Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, 29, 311–343.
- Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological Science*, 14, 31–38.
- Hellman, A. B. (2011). Vocabulary size and depth of word knowledge in adult-onset second language acquisition. *International Journal of Applied Linguistics*, 21, 162–182.
- Hopp, H., & Schmid, M. (2013). Perceived foreign accent in first language attrition and second language acquisition: The impact of age of acquisition and bilingualism. *Applied Psycholinguistics*, 34, 361–394.
- 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. <https://doi.org/10.1080/15434303.2013.769545>
- Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics*, 24, 131–161.
- Jiang, N. (2000). Lexical representation and development in a second language. *Applied linguistics*, 21, 47–77.
- Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of ESL. *Cognitive Psychology*, 21, 60–99.
- Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, 192, 15–24. <https://doi.org/10.1016/j.bandl.2019.02.004>
- Koizumi, R. (2012). Vocabulary and speaking. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley-Blackwell.
- Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40, 554–564.
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44, 978–990.
- Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49, 757–786. <https://doi.org/10.1002/tesq.194>
- Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12–24.
- Lahmann, C., Steinkrauss, R., & Schmid, M. S. (2016). Factors affecting grammatical and lexical complexity of long-term L2 speakers' oral proficiency. *Language Learning*, 66, 354–385.

- Lahmann, C., Steinkrauss, R., & Schmid, M. S. (2017). Speed, breakdown, and repair: An investigation of fluency in long-term second-language speakers of English. *International Journal of Bilingualism*, 21, 228–242.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307–322.
- Lenneberg, E. H. (1967). The biological foundations of language. *Hospital Practice*, 2, 59–67.
- Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, 63, 530–566.
- Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *IRAL-International Review of Applied Linguistics in Language Teaching*, 49, 221–240.
- Major, R. (2008). Transfer in second language phonology: A review. In J. Hansen Edwards & M. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 63–94). John Benjamins.
- Mayberry, R. L., & Kluender, R. (2018). Rethinking the critical period for language: New insights into an old question from American Sign Language. *Bilingualism: Language and Cognition*, 21, 886–905.
- Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, 46, 610–641.
- Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics*, 35, 418–440.
- Muñoz, C., & Singleton, D. (2011) A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, 44, 1–35. <https://doi.org/10.1017/S0261444810000327>
- Patkowski, M. (1990). Age and accent in a second language: A reply to James Emil Flege. *Applied Linguistics*, 11, 73–89.
- Piske, T., Flege, J., MacKay, & Meador, D. (2011). Investigating native and non-native vowels produced in conversational speech. In M. Wrembel, M. Kul, & K. Dziubalska-Kořaczak (Eds.), *Achievements and perspectives in the acquisition of second language speech: New Sounds 2010* (pp. 195–205). Peter Lang.
- Qureshi, M. A. (2016). A meta-analysis: Age and second language grammar acquisition. *System*, 60, 147–160.
- Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, 37, 828–848.
- Ruivivar, J., & Collins, L. (2018). The effects of foreign accent on perceptions of nonstandard grammar: A pilot study. *TESOL Quarterly*, 52, 187–198.
- Saito, K. (2013). Age effects on late bilingualism: The production development of /ɹ/ by high-proficiency Japanese learners of English. *Journal of Memory and Language*, 69, 546–562.
- Saito, K. (2015). The role of age of acquisition in late second language oral proficiency attainment. *Studies in Second Language Acquisition*, 37, 713–743.
- Saito, K. (2019). To what extent does long-term foreign language education help improve spoken second language lexical proficiency? *TESOL Quarterly*, 53, 82–107.
- Saito, K., & Hanzawa, K. (2018). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. *Language Teaching Research*, 22, 398–417.
- Saito, K., & Liu, Y. (2022). Roles of collocation in L2 oral proficiency revisited: Different tasks, L1 vs. L2 raters, and cross-sectional vs. longitudinal analyses. *Second Language Research*, 38, 531–554.
- Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing as an anchor of post-pubertal second language pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language*, 115, 104–168.
- Saito, K., Macmillan, K., Kroeger, S., Magne, V., Takizawa, K., Kachlicka, M., & Tierney, A. (2022). Roles of domain-general auditory processing in spoken second-language vocabulary attainment in adulthood. *Applied Psycholinguistics*, 43, 581–606.
- Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38, 439–462.
- Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition*, 38, 677–701.

- Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, 27, 343–360.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.
- Spadaro, K. (2013). Maturation constraints on lexical acquisition in a second language. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude and ultimate language attainment* (pp. 113–152). Benjamins.
- Tavakoli, P. (2018). L2 development in an intensive Study Abroad EAP context. *System*, 72, 62–74.
- Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, 70, 506–547.
- Thompson, E. C., White-Schwoch, T., Tierney, A., & Kraus, N. (2015). Beat synchronization across the lifespan: Intersection of development and musical experience. *PLoS One*, 10, e0128839. <https://doi.org/10.1371/journal.pone.0128839>
- 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.
- Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.
- 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. <https://doi.org/10.1093/applin/24.1.1>