



ORIGINAL ARTICLE

Testing the role of processing speed and automaticity in second language listening

Bronson Hui*  and Aline Godfroid 

Michigan State University

*Corresponding author. Email: huibrons@msu.edu

(Received 05 July 2019; revised 17 March 2020; accepted 03 April 2020)

Abstract

Second language (L2) listening requires efficient processing of continuing incoming information (Vandergrift & Goh, 2012). Even so, research into individual differences in L2 listening has mostly shed light on the role of linguistic knowledge measured without time pressure (e.g., Mecarty, 2000; Wang & Treffers-Daller, 2017; cf. Vafae & Suzuki, 2020), leaving the role of processing speed and automaticity largely unexplored. To close this gap, we explored the determinants of successful listening using three processing tasks at lexical, syntactic, and propositional levels. Participants were 44 Chinese learners of English. Response accuracy afforded measures of vocabulary size, syntactic parsing skills, and formulation of propositional meaning. Reaction times and the coefficient of variation (Segalowitz & Segalowitz, 1993) afforded processing speed and automaticity measures at each level. We found a hierarchical relationship between different levels of processing, whereby lower-level, lexical effects cascade up and are mediated by propositional comprehension in accounting for listening comprehension. The results highlight the importance of considering processing accuracy and speed at different levels of the linguistic hierarchy to explain variability among L2 listeners. Different from most previous studies, we argue for a need to consider the temporal aspects of processing, along with linguistic knowledge, in modeling L2 listening.

Keywords: automaticity; coefficient of variation; lexical processing; syntactic processing; second language listening

Second language (L2) listening is a difficult skill to acquire (e.g., Vandergrift & Goh, 2012) at least partly because the listener tends to have little control over the speed of the incoming speech stream. This challenge highlights the need for L2 listeners to process the incoming speech stream in a rapid and efficient manner (Vandergrift & Goh, 2012). While researchers have identified determinants of successful L2 listening, such as speech perception, grammar and vocabulary knowledge, working memory capacity, and metacognitive ability (e.g., Vandergrift & Baker, 2015; Wang & Treffers-Daller, 2017), research on L2 listening has commonly relied on measures administered without time pressure (cf. Vafae & Suzuki, 2020). There is thus a

potential misalignment of how linguistic knowledge has been measured in research and how it might be used in real life. For example, Vandergrift and Baker (2015) used tests of meaning recognition without time pressure to measure first language (L1) and L2 vocabulary sizes. While informative, size or accuracy-based measures such as these do not necessarily reveal the extent to which learners can use this linguistic knowledge when under time pressure, which is a typical case for authentic L2 listening (Godfroid, 2020).

One way to test the availability of knowledge for real-time use is by examining learners' processing speed (Andringa, Olsthoorn, van Beuningen, Schoonen, & Hulstijn, 2012) and processing automaticity (Segalowitz, 2010; Segalowitz & Segalowitz, 1993). However, speed and automaticity may not be monolithic constructs. Speed and automaticity may differ for various levels of processing (e.g., lexical, syntactic, and propositional). As such, an investigation into the role of processing speed and automaticity at various levels in L2 listening can shed important light on the subtle relationships between L2 listening components. This line of work can also help put psycholinguistic models of L2 listening to the test.

LITERATURE REVIEW

Contributing factors to L2 listening

Researchers have identified three broadly defined types of knowledge or skill important to L2 listening: (a) speech perception and general cognitive skills such as auditory discrimination skills and working memory (e.g., Vandergrift & Baker, 2015), (b) linguistic knowledge such as grammar and vocabulary (e.g., Mecarty, 2000; Vafae & Suzuki, 2020; Wang & Treffers-Daller, 2017), and (c) metacognition (e.g., Vandergrift, Goh, Mareschal, & Tafaghodtari, 2006).

Vandergrift and Baker's (2015) study included measures of all three types of knowledge or skill, providing a rather comprehensive picture of the relative contributions of each type of knowledge or skill to L2 listening comprehension. The study involved 157 seventh-grade students in French immersion classes in Canada from three cohorts. The authors measured L1-English listening comprehension, L1-English and L2-French vocabulary sizes, auditory discrimination ability, and working memory. Path analysis showed that the general cognitive skills were initially important to process bottom-up auditory information, and then fed into learners' linguistic knowledge, which was more directly related to L2 comprehension. Similarly to other researchers (e.g., Mecarty, 2000; Wang & Treffers-Daller, 2017), Vandergrift and Baker highlighted the contribution of linguistic knowledge to L2 listening.

One limitation in most studies was perhaps that researchers relied rather heavily on measures administered without much time pressure. For example, Wang and Treffers-Dallers (2017) measured vocabulary knowledge using the Vocabulary Size Test. Vandergrift and Baker (2015) used the Peabody Picture Vocabulary Test and its French adaptation, and Mecarty (2000) assessed vocabulary knowledge of participants through a meaning recognition task, and grammatical knowledge through a sentence-completion, multiple-choice task and a grammaticality

judgement task. Caution should be exercised when interpreting these findings because these tasks, which are administered without time pressure, tend to target linguistic knowledge that may or may not be readily available during actual language use (Godfroid, 2020). In other words, untimed, accuracy-based tests provide researchers with relatively little information about the extent to which the measured knowledge is available for use in real-time processing. The ability to point out the meaning of a word without time pressure in a meaning recall task, for example, may or may not entail efficient access to meaning during authentic listening.

L2 listening often places stress on the listener's processor. Unlike in reading where readers can reread parts of the text when comprehension is impeded, listeners usually do not have control over the incoming speech stream (Kim & Godfroid, 2019). In addition, difficulties in perception as a result of differences between the L1 and L2 tend to accumulate and create a ripple effect on word recognition and syntactic parsing (e.g., Broersma & Cutler, 2008; Brown, 2008; Cutler, 2012; Weber & Cutler, 2004). Coupled with listeners' slower processing in the L2, the burden on the processor is certainly substantial, and eventually, communication can break down as a result of "derailments of attention" (Rost, 2014, p. 136). On this account, efficient processing appears to be a necessary condition for successful L2 listening, in that if the listener is able to process information and resolve issues in an efficient manner, they might recover quickly from these lapses, resulting in better listening performance. Therefore, assessment of the availability of linguistic knowledge for efficient processing can be as important as measuring how much knowledge the learner possesses. In the context of investigating factors important to L2 listening, the addition of a temporal component to the vocabulary and grammar measures is desirable. In tandem with accuracy-based measures, reaction time-based tests can reveal how readily available the linguistic knowledge measured is for processing.

While most listening studies did not incorporate a temporal component in their measures, there have been two exceptions. Vafaei and Suzuki (2020) reported significant contributions of vocabulary and syntactic knowledge to general listening skills measured by a standardized listening test. These authors attempted to impose time pressure on their vocabulary and syntactic knowledge measures by limiting the amount of time allotted to participants on each item. For example, in their aural sentence comprehension task, the length of time was limited to 11 s during which participants were auditorily presented a sentence (e.g., "If she were not rich, she could not travel" [Vafaei & Suzuki, 2020, p. 11]) and asked to respond to the question "Can she travel now?" (Vafaei & Suzuki, 2020, p. 11). As acknowledged by the authors, however, the use of paper-and-pencil answer sheets prevented them from claiming that they measured "proceduralized knowledge" (p. 23) that could be accessed quickly in real-time listening.

Another study that included a temporal component was Andringa et al. (2012). The authors recruited 121 native Dutch speakers and 113 L2-Dutch learners who performed a series of tasks tapping into their Dutch listening comprehension, vocabulary size, semantic, grammatical and sentence processing, segmentation ability, working memory, and intelligence. The authors submitted all variables to confirmatory factor analyses. The resulting four latent variables, labeled as linguistic knowledge (e.g., vocabulary size), processing speed (e.g., semantic processing

speed), memory (e.g., working memory measures), and intelligence, were then specified in regression analyses as predictors of listening comprehension, which was the outcome variable. For the Dutch native speakers, linguistic knowledge explained the most variance. Processing speed also accounted for unique variance, albeit negatively. For the L2-Dutch learners, linguistic knowledge again explained the most variance, and intelligence emerged as a significant predictor. Together, these predictors accounted for as much as 96% of variance in L2 listening performance. Processing speed and memory did not account for any unique variance beyond that. The authors concluded that processing speed was a separate construct that can differentiate variation in listening, but only in the case of native listeners. The non-significant result of processing speed in L2 listening in this study seemed to diverge from the idea that listening requires rapid, efficient word retrieval (e.g., Vandergrift & Goh, 2012), although it is worth noting there was very little statistical variance left for processing speed to explain. A further observation is that the authors reduced various levels of processing speed to a single latent variable, labelled *processing speed*, for their analysis. Given that processing speed may not be a unitary construct, and different levels may contribute to L2 listening performance in different ways, there is a need to adopt an approach that can tease apart the different levels of processing, uncovering the unique contribution of each level of processing to L2 listening comprehension.

Taken together, and building on the work by Vafaei and Suzuki (2020) and Andringa et al. (2012), it is high time researchers incorporated time pressure in their measures of linguistic knowledge when investigating the determinants of successful L2 listening. This approach will further bring the use of listening in real-life situations in agreement with how it is measured in research, and hence will add to the external validity of the research. An examination of the contribution of processing speed at each of the different levels of processing offers exciting opportunities for researchers to understand the unique contribution of a given processing level to listening comprehension.

Processing speed and automaticity measures

While reaction time data provide a good processing speed measure, they may not capture all dimensions of fluent language processing (Segalowitz, 2010). One important dimension that can be overlooked is processing automaticity, which represents the cognitive basis of language fluency (Segalowitz, 2010). Intuitively, automaticity is manifested by faster processing (e.g., Hulstijn, Van Gelderen, & Schoonen, 2009). However, the reverse is not necessarily true. Faster processing does not always entail automatic processing because automaticity is conceptualized as qualitatively different from a simple speed-up (i.e., quicker execution of controlled procedures; Hulstijn et al., 2009; Segalowitz, 2010, also see Leow, 2015, for an overview of controlled and automatic processing of the L2). To measure processing automaticity, Segalowitz and Segalowitz (1993) proposed the use of the coefficient of variation (CV). The CV can be derived from processing time and/or reaction time (RT) data obtained from eye tracking and/or behavioral judgement tasks (e.g., Hui, 2020). Specifically, the CV is computed by dividing the standard deviation of all of an individual's RTs by their mean RT ($CV = \frac{SD}{Mean\ RT}$). CV values reveal RT variability in

the same individual's processing while correcting for their processing speed. This variability is then taken as a measure of processing stability—more stable processing corresponds to a smaller CV value, and hence lower RT variability.

Although CV itself indexes only processing variability (stability), Segalowitz and colleagues (e.g., Segalowitz, 2010; Segalowitz & Segalowitz, 1993; Segalowitz, Segalowitz, & Wood, 1998) argued that it can also signal automaticity development when analyzed together with RT data. In particular, the three behavioral signatures of automatization are: first, a decrease in the mean RT (i.e., improving speed); second, a decrease in the CV (i.e., improving stability), and most important, a positive correlation between both (see also Hulstijn et al., 2009, for a review). One implication is that when a positive CV-RT correlation is found, the CV can then be interpreted as a measure of processing automaticity. Using the CV, researchers have investigated automaticity development and individual differences associated with it (Akamatsu, 2008; Elgort, 2011; Hui, 2020; Hulstijn et al., 2009; Lim & Godfroid 2015; McManus & Marsden, 2019; Pili-Moss, Brill-Schuetz, Faretta-Stutenberg, & Morgan-Short, 2019; Rodgers, 2011; Suzuki, 2018). In these cases, the CV often served as a dependent variable while time, language proficiency, and/or treatment were typically the primary predictors of interest.

To investigate automatization in lexical and syntactic processing, for instance, Lim and Godfroid (2015) tested native English speakers ($n = 20$) and Korean learners of English ($n = 40$) using three processing tasks: a semantic classification task, a sentence construction task, and a sentence verification task. Results generally confirmed the presence of all three behavioral signatures of automatization summarized above, in that the native speakers had the lowest mean RTs and CVs, followed by advanced learners, and then intermediate learners. There were also often (though not always) positive and significant correlations between RTs and CVs. In contrast, Hulstijn et al. (2009) did not find a decrease in the CV in Dutch learners of English as the learners progressed through their high school years. Finally, McManus and Marsden (2019) found evidence for grammatical knowledge restructuring leading to more automatic processing (i.e., a decrease in CV) but only for learners who had received explicit instruction and practice in both their L1 and L2.

Other researchers have also investigated the effect of processing automaticity on learning (Elgort & Warren, 2014; Segalowitz & Freed, 2004). In these studies, CV was used as a predictor to index processing automaticity as an individual learner differences measure. In an incidental vocabulary learning study, Elgort and Warren (2014) operationalized the construct of lexical proficiency by including both an explicit vocabulary measure (i.e., a measure of vocabulary size) and an implicit vocabulary measure (i.e., a learner's processing skills as measured by the CV). The researchers found that learners with more automatized processing skills learned more word meanings from reading than those whose processing was less automatic. Elgort and Warren's (2014) work was the first of its kind, highlighting the need to include a processing dimension of lexical skills as a component of word knowledge. This new dimension echoes with Godfroid's (2020) proposed addition of processing automaticity to Nation's (2013) oft-cited, comprehensive framework of vocabulary knowledge. Given such importance of conceptualizing lexical proficiency as knowledge plus processing skills, one issue that remains to be addressed in the automaticity literature is the exact role of such processing skills in actual language

use (e.g., listening). Engaging in this line of work will help to bridge L2 listening research and processing automaticity research, and at the same time, fill important research gaps in both research bases.

Modeling L2 listening

To guide further research into the component processes of L2 listening, researchers have proposed theoretical models that bring together and summarize extant psycholinguistic findings. Cutler's (2012) native listening model, for example, stresses the importance of L1 background in L2 listening. The author suggested that "[n]onnative listening is hard because native listening is easy" (p. 335). The key idea is that L2 listeners rely on their L1 phoneme repertoires in perceiving L2 sounds. Due to the differences between the two sound systems, misperception arises, representing some of the earliest difficulties in L2 listening at the lowest, phonemic level. Of importance, these problems have "a disastrous cumulative effect" (p. 354) on higher level processing such as syntactic parsing, especially when the processor is operating under stressful conditions (Brown, 2008). Listeners may, for example, ignore certain linguistic information such as syntax that is less relevant to comprehension. In the long run, Cutler (2012) argued, such selective listening could result in L2 listeners computing less fine-grained syntactic representations (Clahsen & Felser, 2006). One implication of this account is that the importance of grammatical knowledge in L2 versus L1 listening comprehension may well differ. If the listener can achieve comprehension while ignoring at least some syntactic information, grammatical knowledge and parsing skills may actually play a smaller role in L2 listening.

While Cutler (2012) placed more emphasis on lower level processing, Field's (2013) model of listening consists of both lower level processes (i.e., input decoding, lexical search, parsing) and higher level processes (i.e., meaning construction, discourse construction; see Figure 1). Focusing on bottom-up processes in listening, Field (2013) proposed that acoustic information is processed at phonological, lexical, and syntactic levels before a proposition is understood. This propositional information is then the basis for higher level meaning construction at the discourse level, which results in listeners creating a meaning representation in their memory. Unlike Cutler (2012), Field's (2013) model does not seem to explicitly describe the relative importance of each level of processing. Success and failure at a lower level might cascade up to a higher level, consistent with the view that speech processing is an incremental and hierarchical process; however, there is no one processing level that seems privileged. Propositional comprehension in particular—which is the comprehension of the context independent, literal meaning of an utterance—may be an interesting stage to study because this is where lower level and higher level processes meet during listening. By focusing on propositional comprehension, we can also assess the contributions of different levels of automaticity (lexical, syntactic, propositional) in L2 listening, building on research on automaticity in sentence-level processing (Hulstijn et al., 2009; Lim & Godfroid, 2015). Last but not least, a focus on propositional comprehension will enable us to study the relationship between lower level processes in listening and general listening comprehension, as measured by an independent listening proficiency test.

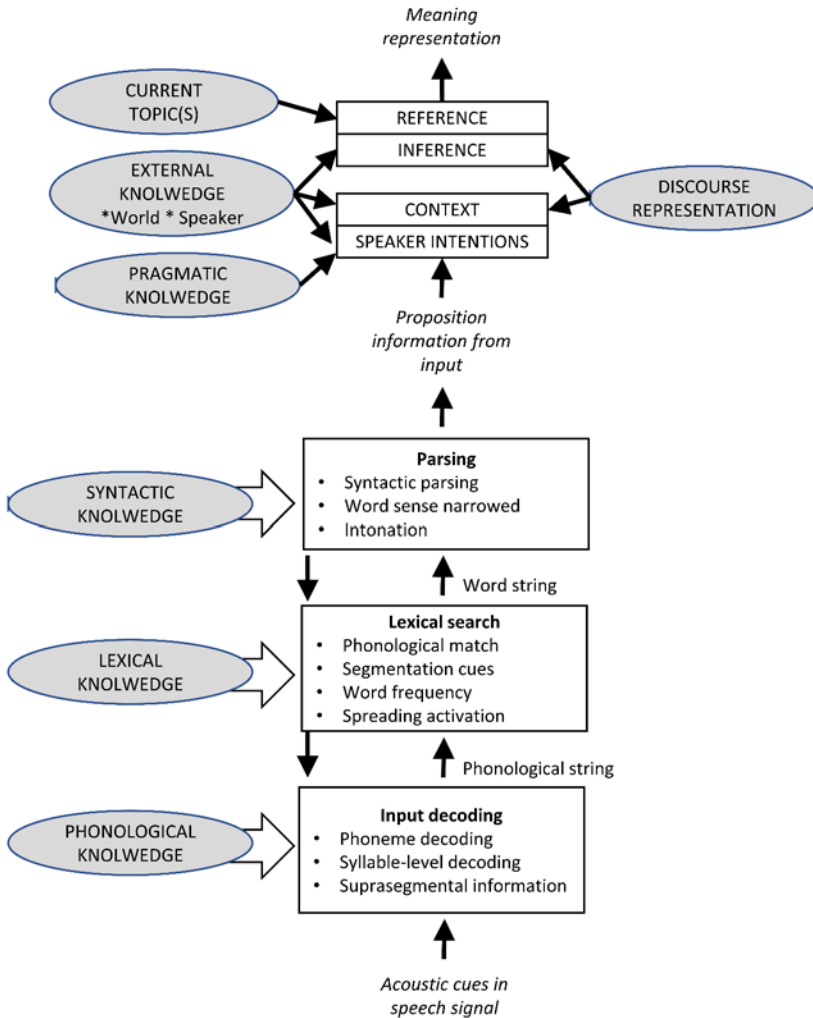


Figure 1. Field's (2013) model of listening.

THE PRESENT STUDY

In the present study, we aimed specifically to investigate the potential relationship between processing speed and automaticity at three levels (lexical, syntactic, and propositional) and general L2 listening performance. This investigation is important because, first, much of the existing work on attributes of L2 listening has relied mostly on measures without time pressure. Use of measures with a temporal component can shed light on the contribution of linguistic knowledge that is available for real-time language processing in L2 listening (as opposed to linguistic knowledge that can only be retrieved offline). Second, the role of processing speed in L2 listening was investigated in only one study (i.e., Andringa et al., 2012). Follow-up investigation is very much needed, especially because there is a need to

Table 1. Biographic information of participants

	Mean	SD	Range
Age	20.16	1.95	18–28
Years of studying English	10.50	3.63	4–18
TOEFL score ^a	84.50	9.98	60–112

Notes: ^aThe TOEFL test is an international, standardized test of English proficiency for learners of English as a foreign language. It has a maximum score of 120. It is widely used for university admission purposes in North America. Three of the participants did not report any standardized English proficiency test scores. Five reported a band score in the International English Language Test System. Conversion was made based on information on the Educational Testing Service website. An average was taken between the upper and lower bound for participants who reported a band score. For example, an IELTS band score of 6 was converted to a TOEFL score of 69.

distinguish processing speed at various levels. Third and finally, given that processing speed alone does not suffice to explain the construct of processing automaticity, including indices of processing accuracy, processing speed (response times), and processing automaticity (CV) will offer a more comprehensive picture of the role of different aspects of processing in L2 listening. To achieve the aims of this study, our primary research question was thus

(RQ1) To what extent do lower level processes at lexical, syntactic, and propositional levels contribute to general listening proficiency?

As secondary research questions, we also investigated the relationship between these three levels. Therefore, our second and third research questions were

(RQ2) To what extent can one level of processes in the model be accounted for by processes at a relatively lower level?

(RQ3) How do these levels of processes interact in their contribution to general listening skills?

METHOD

Participants

We recruited a sample of 44 Chinese undergraduate students enrolled at Michigan State University (39% male and 61% female). The participants were not majoring in linguistics or any language-related areas, including English. Their English proficiency ranged from low-intermediate to advanced, or B1 to C1 levels in the Common European Framework of Reference (Educational Testing Services, 2019). This range reflects the variability in proficiency levels of the international student body at our university. The students' basic biographic information is reported in Table 1. In addition, we included 26 native speakers of English who took part in the three processing tasks (see Tasks and Materials below). The native speakers helped us ensure that our materials worked as envisioned, which was considered important especially given that we adopted all the processing tasks from Lim and Godfroid (2015) and Meara (2010) in the auditory modality for this study. All

participants received either monetary compensation for their time or extra credit for a course in which they were enrolled. Ethical clearance was obtained from the institutional review board according to our university's regulations governing research involving human participants.

Tasks and materials

In the spirit of open science, all instruments have been made available on the Open Science Framework (OSF; <https://osf.io/35vfx/>). There were a total of four tasks: a listening comprehension test, an auditory yes/no RT test, an auditory sentence construction task, and an auditory sentence verification task. For our three processing tasks (i.e., all but the listening comprehension test), we invited a native English-speaking volunteer who had experience in teaching English as a L2 in the United States to record the stimuli. The recording took place in a professional studio. The first author trained the speaker to read aloud the materials while minimizing emotional prosody. To do so, the volunteer had time to practice and record a number of stimuli. The first author and the volunteer listened to these practice recordings together and made sure no unnatural intonation was included. After that, the remaining stimuli were recorded. All sound files contained only the stimuli; that is, there was no silence either before or after the stimulus. We then normalized all sound files in volume.

Listening comprehension test

We used the listening section of a practice test for a general English proficiency test administered by the English Language Center at Michigan State University. The university accepts this test as an assessment of L2 English proficiency for university admission purposes; in that regard the test is on a par with standardized English proficiency tests such as TOEFL. The test is aligned with the Common European Framework of Reference and has been designed to assess general communicative ability in English at the C2 (mastery) level. The test consisted of 40 items, all of which were multiple-choice questions in which one (out of three) options was correct (see Appendix A on OSF for an example item [<https://osf.io/35vfx/>]). All test items were from retired test forms that had been administered to test takers in the past. The test required participants to process speech in (a) short conversations ($k = 8$), (b) longer conversations ($k = 15$), and (c) extended discourse ($k = 17$). The order of test items followed the same order as described above (i.e., short conversations first, followed by longer conversations, and then the extended discourse). Only for the extended discourse was the audio played twice. While the test development manual did not reference any particular theoretical model of listening, the first author, who is an experienced teacher of English as a foreign language, assessed that the items tested comprehension of general and specific ideas as well as inferencing, which was largely in line with Field's (2013) model of meaning construction in listening.

Auditory yes/no RT test

We adapted and computerized Meara's (2010) yes/no test obtained from Lextutor (www.lextutor.ca) for the auditory modality. We instructed the participants to indicate if they knew the meaning of a sound string as quickly and accurately as possible by pressing corresponding yes/no keys on a Cedrus' RB740 response pad (far right key being yes, and far left being no). Given the expected level of proficiency of the participants, only Levels 3 to 5 were used. There were 60 items (40 real words and 20 nonwords) at each level, making up a total of 180 items (see Appendix B on OSF [<https://osf.io/35vfx/>]). Seven nonwords that were deemed confusing when heard by the first author and an English native speaker (e.g., *stace* as a nonword being potentially confused with *stays*) were replaced by other nonwords in another level that was not used in this study. Sixteen (9%) items were excluded as unsatisfactory items (see Data Analysis section).

All items were programmed and presented through Superlab, a piece of software for psycholinguistics research. Each trial started with a fixation cross (+) in black (font size 24) against a white background in the middle of the computer screen for 400 ms. This fixation cross was immediately followed by the presentation of the spoken form of an item through a pair of earphones. After the participant's yes/no response, the next trial started after 400 ms. There was no feedback given. All items were pseudorandomized by Superlab. Response accuracy served as a measure of spoken vocabulary breadth (LEXacc; see the Data Analysis section for scoring method). RTs to individual items were recorded by Superlab from the onset of the spoken stimulus. The mean RT and CV for each participant provided lexical processing speed (LEXrt) and automaticity (LEXcv) measures.

Sentence construction task

We adapted this task from Lim and Godfroid (2015) to the auditory modality. Participants listened to a sentence fragment (e.g., *After some time . . .*) and chose an appropriate, grammatical continuation between two options (e.g., A. *woke* and B. *she*). They did so by pressing corresponding A/B keys on a Cedrus' RB740 response pad (far left key being A, and far right being B). There were a total of 40 items (see Appendix C on OSF [<https://osf.io/35vfx/>]). No items were excluded in the item screening (see the Data Analysis section). Each trial started with a fixation cross (+) presented for 400 ms, followed by the auditory stimulus (i.e., the sentence fragment) and simultaneously, presented in their written forms on the screen, the two possible options for continuing the sentence.¹ The following trial started 400 ms after the participant's response. No feedback was given. All items were pseudorandomized by Superlab, which also registered the RT from the onset of the auditory stimulus. The accuracy of each participant served as a measure of parsing skills (SYNacc). The mean RT and CV for each participant provided syntactic processing speed (SYNrt) and automaticity measures (SYNcv), respectively.

Sentence verification task

This task was also adapted from Lim and Godfroid (2015) to the auditory modality. In this task, participants verified the plausibility of sentences. They did so by pressing corresponding yes/no keys on a Cedrus' RB740 response pad (far right key being yes [plausible], and far left being no [implausible]). Semantic anomalies and structural violations were included in the implausible trials. For example, "*Most theaters in the US have only one chair*" (Lim & Godfroid, 2015, p. 1256) contained a semantic anomaly, while "*My uncle made me for a snowman*" (p. 1256) contained a structural violation at the phrasal level. There were initially a total of 60 items (see Appendix D on OSF [<https://osf.io/35vfx/>]), five of which (8%) were excluded as unsatisfactory items (see the Data Analysis section).

Each trial started with a fixation cross (+) presented for 400 ms, followed by the auditory stimulus. The following trial started 400 ms after the participant's response. All items were pseudorandomized by Superlab. The response accuracy of each participant served as a measure of formulation of propositional meaning (PROacc). The mean RT and CV for each participant provided processing speed (PROrt) and automaticity (PROcv) measures at the propositional level.

Procedure

Data collection took place individually in a quiet lab on campus. We presented an information sheet to participants explaining the overall procedure before we sought their consent to participate. The participant first completed the linguistic background questionnaire. They then took part in the auditory yes/no RT test, before they did the sentence construction task, and then the sentence verification task. Finally, the listening comprehension test was administered. The whole session lasted about an hour. Participants took breaks between tasks.

Data analysis

We first performed accuracy and RT analyses, followed by a participant performance and item screening. We then proceeded to inspect reliability and validity information for our instruments before we conducted our main statistical procedure.

Accuracy data

The first author scored the listening comprehension test according to the answer key provided by the test administrator. We awarded 1 point to each correct answer, for a total of 40 possible points. Each participant had one overall listening comprehension score, which was converted to an accuracy percentage, representing a participant's general listening proficiency (LISTEN). For the auditory yes/no RT test, we first considered proposals on how to score the original yes/no test: the simple hits-minus-false-alarms rule, the correction for guessing formula (Huibregtse, Admiraal, & Meara, 2002), the delta m (Huibregtse et al., 2002), and the index of signal detection (Huibregtse et al., 2002). We also considered the RT approach suggested by Pellicer-Sánchez and Schmitt (2012). Given the lack of consensus as to

the best scoring methodology (Pellicer-Sánchez & Schmitt, 2012) and the small differences found in the comparison of the different formulas (Mochida & Harrington, 2006), we decided to use the simple hits-minus-false-alarms rule in our computation of the accuracy scores. Specifically, we coded correct YES responses to word trials as hits, and incorrect YES responses to nonwords as false alarms. From this, we calculated proportions of hits and false alarms. We then subtracted the proportion of false alarms from that of hits to penalize guessing. Given the similar, possible responses between the auditory yes/no test and the sentence verification task, we also used this hits-minus-false-alarms rule for the sentence verification task for consistency. For the sentence construction task, we used the response accuracy logged by Superlab based on the expected responses in Lim and Godfroid (2015). We awarded 1 point to each correct answer, adding up to an overall score and hence an accuracy rate. Each participant had one accuracy percentage for the listening test (LISTEN) and each of the three processing tasks (LEXacc, SYNacc, and PROacc).

RT and CV data

For the RT data in the processing tasks, we first computed the response times from the offset (end) of the stimuli by subtracting the total stimulus durations from the onset RTs registered by SuperLab. We only used these offset response times to trim the extremely slow RTs (see below). In all other analyses, we used RT from the onset (beginning) of the stimulus as registered by Superlab. For the yes/no RT test, we followed recommendations and previous CV research (Keating & Jegerski, 2015; McManus & Marsden, 2019) to remove RTs faster than 150 ms (from the onset of the stimulus) and slower than 2000 ms (from the offset of the stimulus). For the sentence processing tasks, we consulted the literature to identify an appropriate cutoff value for the upper end, but could not discern a standard practice. For example, in McManus and Marsden's (2019) study, participants had up to 4500 ms after the offset of the stimulus for their responses to be included in the data analysis. Lim and Godfroid (2015) and Hulstijn et al. (2009) defined slow outliers as RTs of more than 3 *SD* above the item mean. Given our design, we arbitrarily decided to use 4500 ms from the offset as the cutoff while keeping 150 ms at the other (fast) extreme. Following Hulstijn et al. (2009), Lim and Godfroid (2015), and McManus and Marsden (2019), only correct responses to YES trials (hits) were analyzed; that is, we analyzed only correct responses to the real words and to plausible trials in the auditory yes/no RT test and the sentence verification task, respectively. The mean RT and CV were computed for each participant for each processing task (LEXrt, SYNrt, PROrt, LEXcv, SYNcv, and PROcv).

Performance and item screening

The purpose of this screening was to ensure that participants were engaged, and the experimental materials functioned as intended. Although the entire procedure was approximately 1 hr, we were aware of the possibility that not all participants would be fully engaged with all tasks throughout the entire procedure. It was also assumed that participants who performed below chance level either did not understand the tasks or did not have a proficiency level sufficient to process the materials in the

experiments. In the listening test, chance level was 0.33, given the three options in each question. In the auditory yes/no RT test and the sentence verification task, chance level was zero, given that accuracy was computed by subtracting the proportion of false alarms (incorrect YES to nonword/implausible trials) from that of hits (correct YES to real-word/plausible trials; see description of Accuracy data above). For the sentence construction task, the chance level was 0.50, because it was a two-alternative forced choice test. When a participant performed below chance level, we coded their data for that task as missing.

In order to ensure that performance variation was not due to the materials, we relied on the native-speaker data to identify any potential problems with an individual item's content and/or recording. Specifically, we excluded items that were shown to be confusing for native speakers. To this end, we arbitrarily decided to use 0.75 accuracy as a cutoff value: an item was considered as unsatisfactory when more than one fourth of the native speakers in our sample responded incorrectly to it. In the auditory yes/no RT test, we excluded 16 items (9%). In the sentence construction task, no items were excluded. For the sentence verification task, 5 items (8%) were excluded.

Reliability and validity of instruments

For the listening test, we obtained a Cronbach's α reliability of .84 with our sample of 44 Chinese undergraduates, using the $\alpha()$ function in the psychometric package in *R* (version 2.2). This reliability was on a par with L2 listening research (Plonsky & Derrick, 2016; median = .77, $k = 38$). For the three processing tasks, we used Cronbach's α and split-half reliability values to assess reliability for accuracy and RT respectively. For the auditory yes/no RT test, we also report a split-half reliability, because many items (Levels 3-4) had a 100% accuracy and hence there was no item-level variance to compute a Cronbach's α . In assessing reliability for the processing tasks, we used pooled data from both learners and native speakers. For the auditory yes/no RT test, we obtained split-half reliability values of .92 and .92 for accuracy and for RT, respectively. Cronbach's α reliability was .75 for accuracy² and split-half reliability was .92 for RT for the sentence construction task. Finally, for the sentence verification task, we obtained a Cronbach's α reliability of .85 for accuracy and a split-half reliability value of .66 for RT.

To assess the extent to which our CV measures tapped into participants' processing automaticity, we relied on Segalowitz's (2010) suggestion that the CV value could be interpreted as an index of processing automaticity only when there is a positive RT-CV correlation. For each processing task, we computed three RT-CV correlations: from the learner data, the native speaker data, and the pooled data from both groups (see Table 2). As we were not able to obtain a positive RT-CV for the sentence construction task, we decided to drop SYNcv as a processing automaticity measure. We will return to this decision in the Discussion.

Statistical procedures

Table 2 summarizes all the nine variables in this study. In the following procedure, only learner data were included. For each variable, we checked the descriptive

Table 2. Overview of variables in the present study

Variable	Variable Name	Construct	Task
Accuracy in listening test	LISTEN	Listening comprehension	Listening test
Accuracy in auditory yes/no RT test	LEXacc	Vocabulary breadth	Auditory yes/no RT test
Mean RT in auditory yes/no RT test	LEXrt	Lexical processing speed	Auditory yes/no RT test
CV in auditory yes/no RT test	LEXcv	Lexical processing stability	Auditory yes/no RT test
Accuracy in sentence construction task	SYNacc	Parsing skills	Sentence construction task
Mean RT in sentence construction task	SYNrt	Syntactic processing speed	Sentence construction task
Accuracy in sentence verification task	PROacc	Formulation of propositional meaning	Sentence verification task
Mean RT in sentence verification task	PROrt	Processing speed at the propositional level	Sentence verification task
CV in sentence verification task	PROcv	Processing stability at the propositional level	Sentence verification task

statistics (see Table 3) and normality of the data using visual inspection of histograms and the Shapiro–Wilk test (α set at .05). Three variables (LISTEN, SYNacc, and PROacc) were negatively skewed, leading us to first reverse the score (subtracting each score by the highest score + 0.1; Field, Miles, & Field, 2012), and then perform a \log_{10} transformation.

To address each research question, we engaged in a backward, step-wise model selection in two stages. In particular, we started with LISTEN as the outcome variable in our Stage 1 model. In the initial model for Stage 1, we included all potential predictors at all levels of processing; that is, all eight predictors at the propositional, syntactic, and lexical levels. At each step of model selection, we removed the predictor that had the highest p value. The final model for Stage 1 had the highest adjusted R^2 value and thus represented the model that explained the most variance in LISTEN performance adjusted for model complexity.

The significant predictor(s) in Stage 1 then became the outcome variable(s) in Stage 2. The new outcome variable(s) were in turn predicted by all variables associated with processing at the lower levels of the listening hierarchy (see Figure 1, syntactic and lexical levels). For instance, if formulation of propositional meaning (PROacc) was found to predict LISTEN in Stage 1, it became a dependent variable in Stage 2. The model selection procedure was identical to that of Stage 1. The same modeling procedure was repeated for all levels of the listening hierarchy until no further lower level predictors were found. These additional models helped set the stage for the final mediation analysis (see below), which integrated the results from the different modeling stages.

Table 3. Descriptive statistics of all four tasks

	<i>M (SD) [95% CI]</i>											
	Listening comprehension			Auditory yes/no RT test			Sentence construction task			Sentence verification task		
	NS	Learner	All	NS	Learner	All	NS	Learner	All	NS	Learner	All
Accuracy (%) ^{a,b}		.78 (.14) [.74, .82]		.85 (.20) [.77, .93]	.57 (.18) [.51, .63]	.66 (.25) [.60, .72]	.96 (.04) [.94, .97]	.83 (.08) [.80, .86]	.88 (.88) [.86, .90]	.93 (.08) [.89, .96]	.71 (.24) [.63, .78]	.79 (.22) [.74, .84]
Mean RT (ms)				910 (97) [870, 949]	1007 (142) [963, 1051]	971 (136) [939, 1000]	1518 (277) [1410, 1630]	2142 (479) [1996, 2288]	1910 (513) [1790, 2030]	2617 (287) [2500, 2730]	2968 (378) [2850, 3080]	2837 (385) [2750, 2930]
CV (<i>SD</i> /Mean RT)				0.27 (0.05) [0.26, 0.30]	0.32 (0.06) [0.30, 0.34]	0.31 (0.06) [0.29, 0.32]	0.30 (0.05) [0.28, 0.32]	0.35 (0.07) [0.33, 0.37]	.033 (0.07) [0.32, 0.35]	0.24 (0.07) [0.22, 0.27]	0.26 (0.07) [0.24, 0.28]	0.25 (0.07) [0.24, 0.27]
Group Differences				<i>t</i> (60) = 3.33, <i>p</i> = .001			<i>t</i> (67) = 3.40, <i>p</i> = .001			<i>t</i> (52) = 0.710, <i>p</i> = .48		
RT-CV Correlation (<i>r</i>)				.27, <i>p</i> = .19	.35, <i>p</i> = .02	.41, <i>p</i> < .001	.34, <i>p</i> = .09	−.20, <i>p</i> = .19	.13, <i>p</i> = .28	.62, <i>p</i> < .001	.57, <i>p</i> < .001	.56 <i>p</i> < .001

^aIn the case of the auditory Yes/No RT test and the sentence verification task, accuracy represented a Hit-minus-False-Alarm score (see Data Analysis).

^bThe accuracy for the listening comprehension test, the auditory Yes/No RT test, and the sentence verification task was reversed and transformed for regression analyses, but numbers presented in this table are the raw score before any transformation. RT, reaction time. CV, coefficient of variation.

We identified and removed outliers for the final model at each stage, which we defined as participants that had a standardized residual larger than 2.5 *SD* from the mean (i.e., outliers in the regression analysis). We report the refitted model as our final model for each stage after outlier removal, where applicable. Assumption checks for all final models are also reported in Appendix E on OSF (<https://osf.io/35vfx/>). Finally, we engaged in a model validation process for our final models using bootstrapped regression (Hamrick, 2019). Specifically, we used the `validate()` function in the `rms` package (version 5.1-3) in R to resample from our original samples with replacement (i.e., to bootstrap). The number of iterations was set at 3000. Each of these 3000 bootstrapped samples was subject to a separate regression analysis, based on which we obtained a distribution for R^2 values (the bootstrapped model fits). Our goal was then to compare our original model fit with this distribution in order to assess the extent to which our original models were overfitted, and hence potentially overestimating our effect sizes (R^2). The possibility of an overestimation needs to be addressed properly because our effect sizes may be the basis of an *a priori* power analysis of subsequent research. An over-optimistic effect size may mislead researchers to plan a study that is underpowered, which can have a serious influence on reproducibility of research in the field (Hamrick, 2019). On top of reporting multiple R^2 and adjusted R^2 values, we also report a corrected R^2 from the result of the validation procedure, which has already been corrected for any overfitting (and, hence, would be the recommended effect size to use for future *a priori* power analyses).

Finally, based on the results of the multiple stages of regression analysis, we engaged in a parsimonious mediation (path) analysis where we specified LISTEN as the outcome variable, the significant predictor(s) from Stage 1 modeling as the mediator(s) and the significant predictors in Stage 2 modeling as primary predictors. We used the `mediate()` function in R's `psych` package (version 1.8.12) to do so. The number of bootstrap resampling was set at 3000 with the width of the confidence interval as .95. In the spirit of Open Science, all raw data (e.g., output text files from Superlab, item-level data for the listening test) and R scripts used in data cleaning, processing, and analysis have been made available on OSF (<https://osf.io/35vfx/>).

RESULTS

Descriptive statistics of all the tasks are presented in Table 3, followed by the correlation matrix of all variables with only learner data in Table 4. Although three accuracy measures (LISTEN, SYNacc, and PROacc) were negatively skewed and hence reversed and \log_{10} transformed for the analyses, we flipped the positive and negative signs in the relevant cases in our reporting to help readers interpret the relationship between the variables in a more straightforward manner.

In the initial model of our Stage 1 analysis, we entered LISTEN as the outcome variable and all eight variables at all three processing levels as predictors. In the final model (Final Model 1), five variables survived the model selection procedure: LEXacc, LEXrt, SYNrt, PROacc, and PROcv. Only PROacc emerged as a significant predictor. The validation process revealed that the original model was overfitted (optimism value = .14). The corrected R^2 using bootstrapping was .30, indicating that 30% of the variance in the listening score was explained by the predictors.

Table 4. Correlation matrix of all variables (learner data)

Correlation coefficient (<i>p</i> value)									
	LISTEN	LEXacc	LEXrt	LEXcv	SYNacc	SYNrt	PROacc	PROrt	PROcv
LISTEN	1	.25 (.11)	−.29 (.07)	−.16 (.31)	.19 (.22)	−.28 (.07)	.45 (.003)	−.33 (.03)	−.23 (.15)
LEXacc		1	.47 (.001)	.15 (.34)	.41 (.007)	.12 (.44)	.53 (<.001)	.13 (.42)	.00 (.98)
LEXrt			1	.36 (.02)	.03 (.84)	.42 (.006)	.05 (.73)	.34 (.03)	.07 (.66)
LEXcv				1	.10 (.54)	−.02 (.88)	.04 (.81)	.11 (.47)	.50 (<.001)
SYNacc					1	−.08 (.59)	.33 (.03)	−.04 (.80)	.09 (.58)
SYNrt						1	.11 (.50)	.38 (.01)	.03 (.83)
PROacc							1	.00 (.99)	−.08 (.61)
PROrt								1	.54 (<.001)
PROcv									1

Notes: LISTEN represents accuracy in the listening comprehension test. LEXacc, LEXrt, and LEXcv represent accuracy, mean RT and CV for the auditory yes/no test, respectively. SYNacc and SYNrt represent accuracy, mean RT and CV for the sentence construction task. PROacc, PROrt, and PROcv represent accuracy, mean RT and CV for the sentence construction task, respectively. Significant correlations (*p* <.05) are **bolded** for easier reference.

Table 5. Summary of the regression models

	Final. Model.1 ($y = \text{LISTEN}$)			Final. Model.2 ($y = \text{PROacc}$)		
	B (95% CI)	p	sR^2	B (95% CI)	p	sR^2
LEXacc	0.29 (−0.08, 0.66)	.14	.01	0.68 (−0.95, −.041)	<.001	.12
LEXrt	−0.24 (−0.61, 0.13)	.21	.004	−0.33 (−0.58, −0.07)	.017	.006
LEXcv						
SYNacc				0.17 (−0.08, 0.42)	.20	.002
SYNrt	0.17 (−0.48, 0.15)	.31	.0009			
PROacc	0.38 (0.03, 0.72)	.038	.04			
PROrt						
PROcv	−0.17 (−0.42, 0.08)	.20	.002			
Multiple R^2 /Adjusted R^2 / Corrected R^2 from bootstrapping	.45 / .37 / .30			.48 / .43 / .35		
Analysis of variance	$F(5, 35) = 5.61, p < .001$			$F(3, 38) = 11.49, p < .001$		

Notes: Since LISTEN, SYNacc, and PROacc were reversed in the transformation, the signs of corresponding coefficients reported here were flipped from the analysis output for more straightforward interpretations. sR^2 represents the unique variance in the outcome variable that can be attributed to the individual predictor (Larson-Hall, 2016; Tabachnick & Fidell, 2001).

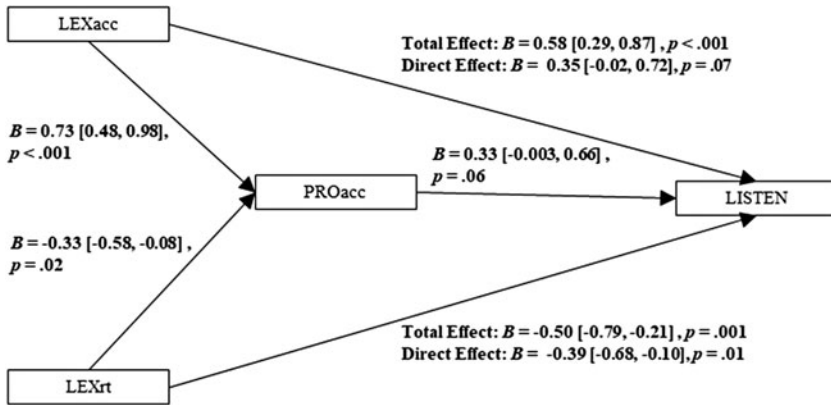


Figure 2. Statistical relationships between components of L2 listening comprehension.

Based on a meta-analysis of multiple regression analyses in L2 research, Plonsky and Ghanbar (2018) found the median unadjusted R^2 value to be .32. Given our unadjusted R^2 value (.45), the effect size we obtained was of medium strength, although, as previously noted, this estimate was likely somewhat over-optimistic. The results indicated that a learner who was more accurate at evaluating propositional meaning of spoken sentences was also a better listener as measured by the listening test. The observed power was at least .90 (as calculated from the conservative, corrected R^2 value of .30). Table 5 presents the model summary.

In our second stage of analysis, we entered PROacc (the only significant predictor in the previous stage) as the outcome variable. In the initial model at this stage, we included all five variables at syntactic and lexical levels, three of which (LEXacc, LEXrt, and SYNacc) remained in the final model (Final Model 2). The two lexical measures emerged as significant predictors. The validation process suggested that the original model was overfitted (optimism value = .13). The corrected R^2 using bootstrapping was .35, and the more liberal unadjusted R^2 value of .48 indicated a medium effect size. The two significant predictors indicated larger spoken vocabulary sizes and faster lexical processing were associated with a higher ability to formulate propositional meaning. The observed power was at least .98 (as calculated from the conservative, corrected R^2 value of .35). We present the model summary in Table 5.

Finally, in the mediation (path) analysis, we specified LISTEN as the outcome variable, PROacc as the mediator (the only significant predictor in the first stage of regression analysis), and LEXacc and LEXrt as the predictors. Figure 2 represents the relationship between the variables. The R^2 value was .40, indicating that 40% of the variance in the outcome was explained by the total effects. All relationships were statistically significant, or approached statistical significance (p values of .06 and .07 for the path from PROacc to LISTEN and that from LEXacc to LISTEN [direct effect], respectively). In addition, the total effects of the two lexical measures (direct effects on LISTEN and indirect effects via PROacc) were larger than the direct effects alone, supporting the conclusion that PROacc mediated the relationships

Table 6. Summary of the mediation analysis

	<i>B</i> (95% CI)	<i>p</i>
Total effects		
LEXacc → LISTEN	0.58 (.29, 0.87)	<.001
LEXrt → LISTEN	−0.50 (−0.79, −0.21)	.001
Direct effects		
LEXacc → LISTEN	0.35 (−0.02, 0.72)	.07
LEXrt → LISTEN	−0.39 (−0.68, −0.10)	.01
Indirect effects		
LEXacc → PROacc	0.73 (0.48, 0.98)	<.001
LEXrt → PROacc	−0.33 (−0.58, −0.08)	.02
PROacc → LISTEN	0.33 (−0.003, 0.66)	.06
<i>R</i> ²	.40	
ANOVA	<i>F</i> (3, 37) = 9.39, <i>p</i> < .001	

Note: Since LISTEN and PROacc were reversed in the transformation, the signs of corresponding coefficients reported here were flipped from the analysis output for more straightforward interpretations.

between LEXacc and LEXrt and the outcome (LISTEN). The model summary is presented in Table 6.

DISCUSSION

In this study, we sought to address the role of processing speed and automaticity in lexical, syntactic, and propositional processes for general L2 listening. Informed by two initial stages of regression analyses, we engaged in a parsimonious mediation (path) analysis to unveil the interrelationships between the components of listening and general listening skills. Results demonstrated the importance of vocabulary size and lexical processing speed for listening, which was mediated (enabled) by a better formulation of meanings at the propositional level.

Mediating role of propositional meaning in a hierarchy of listening processes

With regard to our primary research question, the final, parsimonious mediation (path) analysis highlights the complex interplay of processing at lexical and propositional levels in relation to general listening. This hierarchical pattern makes good sense in the light of Field's (2013) model of listening, in that listeners' lexical processing plays a crucial role in their formulation of propositional meanings, which in turn is the foundation of higher level meaning construction. Specifically, a listener needs to process phonological strings in order to pass on the information to the next level of lexical representations. Once words are successfully recognized (as measured by our yes/no RT test), our data showed that such success contributes to the

formulation of propositional meanings demonstrated in the performance in the sentence verification task. In other words, success at relatively higher levels builds on, and to some extent, depends on success at relatively lower levels. The extent of dependency can be inspected by comparing the total effects (direct and indirect effects) and the direct effects of the two lexical measures (LEXacc and LEXrt) on general listening skills (LISTEN). In both cases, the direct effects were numerically smaller than the total effects ($B = 0.35$ vs. 0.58 and -0.39 vs. -0.50), indicating a clear need (and statistical support) to consider the role of propositional processing as a mediating variable in L2 listening comprehension.

In Field's model, higher level meaning construction starts with propositional information. Listeners consider contexts and the speaker's intention, at the same time drawing on pragmatic and external world knowledge, to make sense of the propositional meanings. The overall meaning is then represented in memory, which we assessed with a general listening skills test. Seen in this light, propositional meaning plays a fundamental role in these higher level comprehension processes. This interpretation is also confirmed by our Stage 1 regression analysis, where formulation of propositions was the only significant predictor of general listening. However, this model was in no way a complete picture of listening; it needed to be considered alongside the results of our Stage 2 modeling. In Stage 2, we identified the importance of lexical processing in formulating propositions. Since lexical items carry meanings, accurate and efficient retrieval of meaning in a time-pressured listening task is a key to listening success, in line with Field's model. Taken together, these indirect effects of lexical processing on general listening, mediated by proposition formulation, support the value of a bottom-up, hierarchical approach in understanding L2 listening.

Potential bypassing of syntactic parsing

Despite the general alignment with Field's (2013) model, we did not find a robust contribution of parsing skills to L2 listening, which somewhat diverges from what one might expect. Indeed, grammatical knowledge has been previously found to be important in L2 listening (e.g., Andringa et al., 2012; Mecarty, 2000; Vafaei & Suzuki, 2020). In terms of task, we used the sentence construction task to measure syntactic parsing skills and syntactic processing speed. This task was somewhat similar to the grammatical processing task used by Andringa et al. (2012), who asked their participants to judge whether sentence fragments could be in the sentence-initial position. In Andringa et al.'s study, grammatical processing accuracy loaded onto the latent variable labeled as *knowledge*, which in turn predicted listening comprehension. The researchers therefore did not inspect the contribution of parsing skills directly, but opted to do structural equation modeling instead. As the factor loading for parsing skills was not reported, we cannot tell what the actual contribution of parsing skills to listening comprehension was.

In Vafaei and Suzuki's (2020) study, the standardized loading of the latent variable syntactic knowledge was .28 and was significant. The authors tested "syntactic structures that potentially play a role in listening comprehension" (Vafaei & Suzuki, 2020, p. 11) through an aural grammaticality judgement and an aural sentence comprehension task. Their selection of the five target structures (e.g., active vs. passive)

was informed by studies in cognitive psychology and corpus research. Therefore, both their tasks and materials were different from ours.

We adopted the sentence construction task from Lim and Godfroid (2015). To restrict higher level semantic processing, the authors of the original study constructed the stimuli to be “as short as possible” (Lim & Godfroid, 2015, p. 1256). When adopting the task, we also thought that it was necessary to have a relatively pure measure of parsing skills, especially when we also had the sentence verification task, which aimed exactly at higher-level (i.e., propositional) processing. Because of this, the experimental materials in this task differed substantially from those in Vafaei and Suzuki’s (2020) study and, perhaps more importantly, from the listening test, at least in terms of syntactic complexity. On the one hand, the sentence construction task contained mostly simple sentences; on the other hand, the listening test, which was designed to assess listening at the C2 level, had sentences of varying degrees of syntactic complexity. In future investigation, we would like to use a measure that could tap into the learners’ parsing skills of more complex sentences, while minimizing higher level semantic analysis. The exclusive use of highly frequent words (to minimize semantic load) in the items in Vafaei and Suzuki’s (2020) study seems to be a step in that direction.

Another potential explanation for the nonsignificance of syntactic parsing in our data is perhaps the proposed, smaller role of syntactic processing in L2 listening (Cutler, 2012). As previously reviewed, Cutler places much emphasis on lower level processing such as speech perception and word recognition in listening. Issues at these levels accumulate and cascade up if lower level processing is not done efficiently. To ease the burden on the processor, listeners may selectively attend to meaning, which is important to comprehension and often carried by lexical items (e.g., Brown, 2008). On this account, perhaps it was of little surprise that we could not find a role for syntactic processing in L2 listening. Successful listening may have depended more on speech perception and good word recognition. At the same time, the idea that L2 users do not always process syntactic information online (Clahsen & Felser, 2006) is heavily based on reading studies, as also rightly pointed out by Cutler (2012). A recent study on listening has revealed that L2 listeners are able to construct full syntactic representations in real time (Fernandez, Höhle, Brock, & Nickels, 2018). Based on the available evidence so far, it may be premature to draw a conclusion about the exact role of syntactic processing in L2 listening. Further research is certainly very much needed.

Role of lexical processing speed

One other important finding is that we captured, for the first time, the association between lexical processing speed and general L2 listening proficiency. Specifically, lexical processing speed significantly predicted both formulation of propositional meanings and general listening. The evidence that we present offers unique, empirical support for the idea that listening requires rapid, efficient word retrieval (e.g., Vandergrift & Goh, 2012), although it differs from results in a previous study by Andringa et al. (2012). Again, the different data analysis approaches might explain some of the differences. In particular, our data analysis procedure enabled us to isolate processing speed specifically at the lexical level. It appears that the

contribution of processing speed varies for native versus L2 listeners (as found in Andringa et al., 2012) but, importantly, it also depends on the level of processing (as found in this study). Of note, our data suggested that processing speed seemed to matter more at the lexical level than higher up the linguistic hierarchy.

Lexical processing speed contributes to listening comprehension in two ways: first, as listeners attempt comprehension, they tend to focus more on meaning, which is often carried by lexical items (e.g., Brown, 2008). In this regard, rapid retrieval of a word's meaning becomes one important factor in listening. Second, lexical processing speed may be a factor in how quickly processing issues at lower levels can be resolved. As discussed in the literature review, difficulties at phonemic levels, for example, can cause a ripple effect on higher level processing, eventually leading to a breakdown of communication (Cutler, 2012). Listeners who demonstrate fast lexical processing may either not have encountered great difficulties at some of these lower levels or they may have resolved these issues in an efficient manner. In both cases, attentional resources can then be diverted to higher level processes that comprehension requires, potentially leading to better listening performance. Given that L2 listeners tend to process information more slowly than native listeners, the role of processing speed (and hence the ability to recover from processing difficulties) is an important one in understanding L2 listening.

If these interpretations are correct, the role of lexical processing speed identified in this study also has implications for teachers, researchers, and language assessment specialists. Language teachers can consider incorporating vocabulary learning activities with some time pressure with the aim to improve their learners' lexical processing speed. As a case in point, Fukkink, Hulstijn, and Simis (2005) trained students with a translation and a cloze sentence task with increasing time pressure, which later led to decreased retrieval time of the lexical items. Such time-pressured exercises and the resulting improvement in lexical processing speed may then have a positive impact on learners' listening skills overall. From the perspectives of researchers and language testers, the present finding also highlights the importance of testing lexical knowledge from a processing perspective, as well as an accuracy-based perspective. Paper-based vocabulary tests have been very useful in providing a general picture of lexical competence; however, they remain silent on the extent to which such knowledge can be processed in a rapid and efficient manner, which is important in real-time communication (Godfroid, 2020).

Coefficient of variation as an automaticity measure in the auditory modality

This study was one of the first to extend the research base on automaticity to the auditory modality (cf. McManus & Marsden, 2019). The current automaticity literature relies heavily on written stimuli, and because of that, we know relatively little about the processing of auditory linguistic information. This study begins to fill this gap by identifying the relationships between automatic processing and listening comprehension. In reading, learners have total control over their reading speed, hence the processor is less likely to be overloaded because readers are able to pause and go back even when their processing automaticity is still developing. In contrast, the continuing speech stream in listening highlights the need for learners to process linguistic information in an automatic manner because they have little control over

the amount of information that floods into the processor. Processing is then more likely to break down during listening when automaticity has not been fully developed yet. To empirically confirm and replicate these relationships, researchers need valid, reliable measures of processing automaticity to further their investigation. We have taken one of the first steps in this direction.

In our study, we carefully chose our three processing tasks. We particularly chose the two sentence processing tasks from Lim and Godfroid (2015) because the researchers were able to find evidence of automatization in written sentence processing in their study. As such, our use of these tasks also served as a replication attempt in a different modality. While we eventually decided to drop the CV measure of the sentence construction task, the other two tasks performed as intended (i.e., the CV was confirmed as a measure of automaticity, which correlated positively with processing speed). A positive CV-RT correlation is considered a precondition for the CV measure in question to be interpreted as a measure of automaticity (Segalowitz & Segalowitz, 1993). Consistent with this, for two of our processing tasks, we found such positive CV-RT correlations in both the learner data and the overall data pooled for both learners and native speakers; in addition, we found differences in CV values between learners and native speakers, although the difference was only numerical in the sentence verification task. Overall, then, the replication attempt was successful.

Regarding why the sentence construction task did not perform as intended, we suspect that modality had an impact on the trial sequence. In Lim and Godfroid, the options for sentence construction were presented on the next screen after the sentence fragment, and so participants were allowed to read the sentence fragment in a self-paced fashion before proceeding to the critical screen with the answer options. This arrangement was not necessary in our case because the options were presented in their written forms simultaneously with the auditory stimulus. We can only suspect that the difference in modality and trial sequence might have played a role in the lack of a significant, positive RT-CV correlation. One way to look into this issue would be to combine the tasks and measures from Lim and Godfroid (2015) and this study and collect processing data in both modalities, using a within-subject design. In sum, much more research needs to be conducted to examine how processing automaticity can be best measured at the sentence level in the auditory modality.

In terms of the limitations of this study, we acknowledge that our approach to the statistical analysis was data driven. The model we present in the mediation analysis was exploratory (we did not specify it *a priori*) and therefore we believe it deserves a replication. Six of our eight variables survived model selection in at least one stage of modeling, indicating not only their important role in listening but also their complex interrelationships. In addition, because we built this study on existing automaticity research, we included only tasks that tap into three levels of processing. As a result, we did not test participants' phonological processing, which, according to Cutler (2012), is a major source of difficulty in L2 listening. We also did not ask the native speakers to complete the listening test because they would likely have performed at ceiling levels. To further clarify the role of native listening and proficiency in nonnative listening, future researchers can sample participants with a wider range of proficiency levels, from intermediate, to advanced and near native, to native speakers. In that case, listening proficiency might be better assessed

through a test that can discriminate a wide range of proficiency levels (vs. ours that is appropriate for advanced L2 learners). In addition, we acknowledge that our CV measure might not be able to capture the many characteristics of processing automaticity although it appears to be one of the few proposed automaticity measures in this area of research. Finally, more confirmatory work needs to be conducted to further elucidate how different aspects of processing at different levels interact during listening, potentially preregistering what statistical analyses will be performed (cf. Marsden, Morgan-Short, Trofimovich, & Ellis, 2018).

Conclusion

L2 listening researchers have often omitted a temporal element in their measures of linguistic knowledge, making studies of L2 listening relatively silent on the role of processing speed and automaticity. We filled this gap using three processing tasks as our measures of linguistic knowledge and skill at lexical, syntactic, and propositional processing levels. Our results supported the hierarchical nature of the listening construct, with success at higher levels building on and, to some extent, depending on success at relatively lower levels. The final, parsimonious mediation analysis indicated that lower-level, lexical processes (spoken vocabulary size and lexical processing speed) support the construction of propositional meanings and, at the same time, contribute directly to general listening. These complex interrelationships revealed by our study also underscore the importance of measuring linguistic knowledge that is available for use under both time pressured and not time pressured conditions.

ACKNOWLEDGEMENTS. This study is supported by the Second Language Studies PhD program at Michigan State University. This manuscript benefited tremendously from valuable feedback from Drs. Nick Ellis, Shawn Loewen, and Patti Spinner, as well as the assistance of Dr. Daniel Reed with identifying experimental materials. The first author would also like to thank his fellow PhD classmates for their input during LLT 861/862 *Advanced Research in Second Language Acquisition* and *Advanced Topics in Second Language Acquisition*, which took place in 2017/2018 at Michigan State University as well as the Daily Writing Slacker group for their support.

Notes

1. As rightly pointed out by a reviewer, participants needed to process information in both the auditory and written modalities for this task. This made the task multimodal, which was different from the other two, auditory processing tasks. Although this may be a limitation of the current task design, and differs from what Lim and Godfroid (2015) did, presenting answer options in the written modality was a necessary concession in order to be able to measure participants' response times.
2. There was one item (3%) that had a 100% accuracy. This item was excluded from the calculation of the Cronbach's α value.

REFERENCES

- Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied Psycholinguistics*, 29, 175–193.
- Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, 62, 49–78.

- Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, **36**, 22–34.
- Brown, G. (2008). Selective listening. *System*, **36**, 10–21.
- Clahsen, H., & Felser, C. (2006). Continuity and shallow structures in language processing. *Applied Psycholinguistics*, **27**, 107–126.
- Cutler, A. (2012). *Native listening*. Cambridge, MA: MIT Press.
- Educational Testing Services. (2019). Performance descriptors for the TOEFL iBT® test. Retrieved from <https://www.ets.org/s/toefl/pdf/pd-toefl-ibt.pdf>
- Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, **61**, 367–413.
- Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, **64**, 365–414.
- Fernandez, L., Höhle, B., Brock, J., & Nickels, L. (2018). Investigating auditory processing of syntactic gaps with L2 speakers using pupillometry. *Second Language Research*, **34**, 201–227.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.
- Field, J. (2013). Cognitive validity. In A. Geranpaye & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge: Cambridge University Press.
- Fukink, R. G., Hulstijn, J., & Simis, A. (2005). Does training in second-language word recognition skills affect reading comprehension? An experimental study. *Modern Language Journal*, **89**, 54–75.
- Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing: expanding Nation's framework. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433–453). New York: Routledge.
- Hamrick, P. (2019). Adjusting regression models for overfitting in second language research. *Journal of Research Design and Statistics in Linguistics and Communication Science*, **5**, 107–122.
- Hui, B. (2020). Processing variability in intentional and incidental word learning: An extension of Solovyeva and DeKeyser (2018). *Studies in Second Language Acquisition*, **42**, 327–357.
- Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, **19**, 227–245.
- Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, **30**, 555–582.
- Keating, G., & Jegerski, J. (2015). Experimental designs in sentence processing research. *Studies in Second Language Acquisition*, **37**, 1–32.
- Kim, K. M., & Godfroid, A. (2019). Should we listen or read? Modality effects in implicit and explicit knowledge. *Modern Language Journal*, **103**, 648–664.
- Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. New York: Routledge.
- Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. New York: Routledge.
- Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, **36**, 1247–1282.
- Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. C. (2018). Introducing registered reports at *Language Learning*: Promoting transparency, replication, and a synthetic ethic in the language sciences. *Language Learning*, **68**, 309–320.
- McManus, K., & Marsden, E. (2019). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics*, **40**, 205–234.
- Meara, P. (2010). *EFL Vocabulary Tests*. Washington, DC: ERIC Clearinghouse.
- Mecarty, F. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, **11**, 323–348.
- Mochida, K., & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing*, **23**, 73–98.
- Nation, I. S. P. (2013). *Teaching & learning vocabulary*. Boston: Heinle Cengage Learning.
- Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring yes-no vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, **29**, 489–509.
- Pili-Moss, D., Brill-Schuetz, K. A., Faretta-Stutenberg, M., Morgan-Short, K. (2019). Contributions of declarative and procedural memory to accuracy and automatization during second language practice. *Bilingualism: Language and Cognition*. Advance online publication.

- Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, **100**, 538–553.
- Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, **102**, 713–731.
- Rodgers, D. M. (2011). The automatization of verbal morphology in instructed second language acquisition. *International Review of Applied Linguistics in Language Teaching*, **49**, 295–319.
- Rost, M. (2014). Listening in a multilingual world: The challenges of second language (L2) listening. *International Journal of Listening*, **28**, 131–148.
- 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.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.
- Segalowitz, N., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, **14**, 369–385.
- Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics*, **19**, 53–67.
- Suzuki, Y. (2018). The role of procedural learning ability in automatization of L2 morphology under different learning schedules: An exploratory study. *Studies in Second Language Acquisition*, **40**, 923–937.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn & Bacon.
- Tanabe, M. (2016). Measuring second language vocabulary knowledge using a temporal method. *Reading in a Foreign Language*, **28**, 118–142.
- Vafae, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*. Advance online publication.
- Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, **65**, 390–416.
- Vandergrift, L., & Goh, C. (2012). *Teaching and learning second language listening: Metacognition in action*. New York: Routledge.
- Vandergrift, L., Goh, C., Mareschal, C. J., & Tafaghodtari, M. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language Learning*, **56**, 431–462.
- Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, **65**, 139–150.
- Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, **50**, 1–25.