Research Methods in Applied Linguistics 1 (2022) 100011

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Is it possible to measure word-level comprehensibility and accentedness as independent constructs of pronunciation knowledge?

Takumi Uchihara

Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

ARTICLE INFO

Keywords:
Comprehensibility
Accentedness
Vocabulary knowledge
Spoken vocabulary

ABSTRACT

The current study aims to explore the validity of measuring comprehensibility versus accentedness of L2 words as a construct of word pronunciation knowledge. Two research questions were addressed by investigating (a) the interrelationships among four listener-based measures (comprehensibility, accentedness, intelligibility, processing time) and (b) the relative contribution of linguistic features of L2 speech (segmental, word stress, rhythm, fluency) to comprehensibility and accentedness ratings. Nineteen native speakers of English rated L1 Japanese speakers' productions of 37 English words elicited through a picture naming task for comprehensibility and accentedness. Two expert raters were recruited to complete a timed dictation task from which measures of intelligibility (orthographic transcription of L2 words) and processing time (how fast raters can initiate word transcription) were derived. The analyses of rating responses and relationships among listener-based measures showed that the current results were consistent with previous L2 speech studies measuring comprehensibility and accentedness at the sentence or paragraph level (Derwing and Munro, 2009). Three linguistic measures (segmental, word stress, rhythm) were significantly related to comprehensibility and accentedness ratings. The length of words (number of syllables) significantly predicted comprehensibility but not accentedness, indicating that longer words were easier to understand than shorter words. These findings provide initial evidence supporting the partial independence of comprehensibility and accentedness when L2 speech is measured at the word level. This study provides methodological implications for L2 vocabulary research and suggests using a word-level comprehensibility measure as an additional tool to gauge the employability of L2 words in real-life spoken communication.

Introduction

The goal of second language (L2) vocabulary teaching is not only to assist learners in acquiring new words but also to develop the ability to use L2 words in real-life communication. To ensure successful spoken communication, it is essential for oral production of L2 words to be sufficiently accurate and understandable to listeners. Teaching spoken vocabulary is also important given that research has shown a gap in learners' vocabulary knowledge between spoken and written forms (Milton & Hopkins, 2006). In particular, learners in instructional settings where spoken input outside of the classroom is limited tend to have weaker knowledge of spoken vocabulary compared to written vocabulary (Uchihara & Harada, 2018). Given the growing attention to spoken vocabulary knowledge (Matthews, 2021), research has increasingly explored how learners incidentally acquire L2 words from exposure to spoken input, such as listening to academic lectures (Dang et al., 2021), watching television (Peters & Webb, 2018), listening to songs (Pavia et al., 2019),

E-mail address: tuchihar@aoni.waseda.jp

https://doi.org/10.1016/j.rmal.2022.100011
Received 4 December 2021; Received in revised form 29 March 2022; Accepted 30 March 2022
2772-7661/© 2022 Elsevier Ltd. All rights reserved.

and listening to teacher talk (Jin & Webb, 2020). This line of research has suggested that learners improve the ability to recognize the spoken form of L2 words encountered while engaging in comprehension-based activities. However, the majority of vocabulary studies have tended to measure the receptive knowledge of spoken form or form-meaning connection via multiple-choice or translation tests. Not much attention has been directed toward the productive knowledge of spoken form or the knowledge of word pronunciation. Although pronunciation is regarded as one of the important aspects of word knowledge (Nation, 2013), the way to operationalize and measure the construct of word pronunciation has not been established yet nor fully discussed in the L2 vocabulary literature.

In the L2 pronunciation literature, there has been a considerable amount of discussion regarding the construct definition and measurement of pronunciation accuracy (Saito & Plonsky, 2019). In L2 pronunciation instruction, nativelike accuracy has long been prioritized and assessed in classrooms. However, attaining nativelike pronunciation is not realistic nor ideal for learners studying English as a foreign language (EFL), since English has been increasingly used as an international language among non-native speakers with different L1 backgrounds (Levis, 2020). Alternatively, L2 speech assessment and teaching have shifted their focus toward speech intelligibility and comprehensibility (Derwing & Munro, 2015). Intelligibility is defined as listeners' actual understanding of L2 speech measured through a range of listening tasks, including listener transcription of heard utterances, responses to true/false statements, and perception of nonsense sentences (Kang et al., 2018). Comprehensibility, often distinguished from intelligibility, refers to listeners' perceived ease or difficulty of understanding L2 speech. Accentedness (or linguistic nativelikeness) is defined as listeners' judgments of how different L2 speech sounds from the expected language variety. These two constructs are measured through listeners' scalar ratings of L2 speech, using numerical point scales (e.g., 1 = *no accent*, 9 = *heavily accented*; 1 = *easy to understand*, 9 = *hard to understand*). Studies have supported the partial independence of intelligibility, comprehensibility, and accentedness (Derwing & Munro, 2009; Huensch & Nagle, 2021; Munro & Derwing, 1995a, 1995b, 2020), most of which have focused on exploring the relationship between comprehensibility and accentedness (Isaacs & Trofimovich, 2012; Saito et al., 2016; Trofimovich & Isaacs, 2012; see Saito, 2021 for a meta-analytic review supporting the independence between comprehensibility and accentedness). Given that comprehensibility is an intuitive and easy-to-use measure as a general metric of listener's understanding of utterances (Martin, 2020), researchers have utilized scalar ratings of comprehensibility for gauging the effectiveness of instructional approaches for pronunciation learning (e.g., Zhang & Yuan, 2020) and examining the roles of individual differences in L2 speech development (e.g., Saito et al., 2019).

Given the accumulated evidence for the validity of a comprehensibility rating measure in the L2 pronunciation literature, adopting a word-level comprehensibility measure might be a practical solution to the issue concerning the paucity of existing spoken vocabulary measures and provide insight into different aspects of word pronunciation knowledge. Introducing a listener-based approach to measuring word pronunciation accuracy aligns with the ongoing proposal for developing vocabulary tests measuring word knowledge directly relevant to the ability to employ L2 words in real-life communication (Kremmel & Schmitt, 2016). However, whether the listener-based measure of comprehensibility can be applied to evaluating the pronunciation of individual words remains underexplored. In fact, the majority of pronunciation studies have based comprehensibility measures on listeners rating elicited productions of short sentences (e.g., 5.9 words in Munro & Derwing, 1995b), 20 to 30 s excerpts (e.g., Trofimovich & Isaacs, 2012), and entire speaking performances (e.g., 32.1 to 408.7 s in Suzuki & Kormos, 2020). The current study therefore set out to explore the possibility that listener-based pronunciation measures can be applied to vocabulary testing by (a) establishing the independence of comprehensibility and accentedness previously confirmed in L2 pronunciation research and (b) examining which linguistic properties of L2 speech (segmental, word stress, fluency, rhythm) contribute to comprehensibility and accentedness of individual words.

Adopting the vocabulary-pronunciation interdisciplinary approach, this research has a great potential to add to the existing repertoire of tools to measure spoken vocabulary knowledge beyond the receptive or productive knowledge of form-meaning connections (knowledge of mapping L2 forms and L1 meanings). Measuring word-level comprehensibility allows vocabulary researchers to gauge whether and to what extent the spoken forms of words are sufficiently accurate and employable with the aim to achieve communicative success.

Measuring spoken vocabulary knowledge

The majority of earlier studies tended to measure vocabulary learning in written form. Reliance on written measures is not surprising because the main source of learning in question has been written input and output through reading written texts (Uchihara et al., 2019; Yanagisawa et al., 2020) or engaging in form-focused activities such as writing sentences or compositions (Webb et al., 2020). Recently, many researchers have shifted their main focus to the role of spoken input as a source of vocabulary learning such as listening to academic lectures (Dang et al., 2021), watching television (Peters & Webb, 2018), listening to songs (Pavia et al., 2019), and listening to teacher talk (Jin & Webb, 2020). This line of research has suggested that learners improve the ability to recognize the spoken form of L2 words encountered while completing meaning-focused activities without explicitly drawing learners' attention to vocabulary. However, earlier studies have tended to measure the receptive knowledge of spoken form or form-meaning connection through having learners identify target words that appeared during listening activities (Pavia et al., 2019), selecting the first language (L1) translations cued by the spoken L2 forms (Peters & Webb, 2018), and writing L1 translations cued by the spoken L2 forms (Dang et al., 2021). In contrast, little attention has been directed toward the productive knowledge of spoken form or the knowledge of word pronunciation. A few studies elicited the spoken form of L2 words (via L1-to-L2 translation or picture naming tasks) and measured word pronunciation accuracy by counting the number of mispronounced phonemes (Barcroft & Sommers, 2005), identifying misplacement of word stress (Bürki, 2010), or deriving a listener's judgement of how nativelike the pronounced words sounded (Kang et al., 2013). What appears missing in the L2 vocabulary literature is an in-depth discussion and justification for the choice of pronunciation measures from a listener's perspective. In fact, there has been a call for developing vocabulary tests gauging

the degree to which word knowledge tested can be employed in real-life communication (Kremmel & Schmitt, 2016; Schmitt et al., 2020). Although it is agreed that form-meaning connection is one of the most important aspects of word knowledge (Nation, 2013), from the perspective of word employability, the pronunciation of a word whose form and meaning are already mapped needs to be further enhanced to the level that the articulated form is sufficiently accurate and comprehensible to listeners (Uchihara et al., 2021). Adopting the listener-based perspective in testing spoken vocabulary does not only contribute to such an ongoing suggestion for testing lexical employability but also aligns with the widely accepted notion in the L2 speech literature that the listener's ease of comprehension should be prioritized over nativelike accuracy in L2 pronunciation assessment (Derwing & Munro, 2015; Levis, 2020). This paper therefore integrates a listener perspective to define word pronunciation knowledge as the knowledge of form-meaning connection with sufficient phonological accuracy to produce the spoken form of words comprehensible or (nativelike-sounding) to listeners.

Partially independent constructs of comprehensibility and accentedness

Since Munro and Derwing's (1995a) seminal study, L2 pronunciation research has supported the partial independence between comprehensibility and accentedness (e.g., Huensch & Nagle, 2021; Munro & Derwing, 2020). Research has shown that comprehensibility and accentedness are significantly associated with relatively large correlations ranging from .74 to .90 or higher (Crowther et al., 2018; Isaacs & Thomson, 2013; Isbell et al., 2019; Saito et al., 2016; Trofimovich & Isaacs, 2012). Despite the strong links between the comprehensibility and accentedness ratings, evidence for the separation between the two constructs has been documented in the L2 pronunciation literature. First, although the average correlation between comprehensibility and accentedness is found to be relatively strong, the data of individual listeners show considerable variation in the strengths of the correlations. Munro & Derwing (1995a) found a positive and significant relationship between accentedness and comprehensibility for 17 out of 18 listeners. However, the strength of the relationship varied considerably across the listeners with correlation coefficients ranging from .41 to .82. Similarly, Huensch & Nagle (2021) using a mixed-effects modeling analysis reported a significant variation in the accent-comprehensibility links across listeners. Huensch and Nagle attributed this finding to listeners' individual differences in the degree to which they associate the presence of foreign accent with perceived processing difficulty. Second, the degree of raters' severity is found to be different between comprehensibility and accentedness. Listeners tend to be more lenient in evaluating comprehensibility, whereas they are likely to be harsher in rating the degree of foreign accent (Derwing & Munro, 1997, 2009). The resulting distribution of rating responses for accentedness generally becomes skewed with more ratings assigned toward the heavy-accent end and for comprehensibility with more ratings assigned toward the easy-to-understand end. Third, intelligibility (listeners' actual understanding of L2 speech) is more closely related to comprehensibility than accentedness (Derwing & Munro, 2009). The way to operationalize the construct of intelligibility remains contentious, resulting in various approaches to capturing listeners' actual understanding of L2 words or utterances (Kang et al., 2018). One of the most commonly adopted ways to measure intelligibility is to have listeners orthographically transcribe L2 speech (e.g., Huensch & Nagle, 2021; Derwing & Munro, 1997; Munro & Derwing, 1995a). In Derwing & Munro's (1997) study with L2 speakers with different L1 backgrounds, comprehensibility and intelligibility (r = .54) appear to be more closely associated than accentedness and comprehensibility (r = .45) or accentedness and intelligibility (r = –.46). Huensch and Nagle (2021) found a significant association between intelligibility and comprehensibility, but no significant association between accentedness and intelligibility while comprehensibility ratings and other covariates (e.g., L2 proficiency) were statistically controlled for. However, perfectly intelligible utterances are not necessarily judged as the most comprehensible speech, indicating that intelligibility and comprehensibility are closely related but different constructs (Munro & Derwing, 1995a).

The perspective of listeners' processing load also accounts for the partial independence of comprehensibility and accentedness. When listeners rate L2 utterances for comprehensibility and accentedness, the processing cost indicated by response-time measures significantly predicts listeners' comprehensibility but not accentedness judgments (Ludwig & Mora, 2017; Munro & Derwing, 1995b), implying that the two constructs can be distinguished through a reaction-time measure. Processing times represent the degree to which retrieval of meanings (or messages) conveyed through L2 speech is effortless (or effortful), reflecting listeners' perception of the ease or difficulty in comprehending L2 speech. However, processing difficulty is not necessarily related to how listeners perceive the degree to which L2 speech deviates from nativelike forms (Munro & Derwing, 1995b). Finally, exploration of linguistic correlates with comprehensibility and accentedness can reveal the partial independence of the two constructs. This line of research aims to reveal the extent to which various linguistic properties of L2 speech (e.g., segmental errors, word choice, discourse) inform listeners' decision in assigning comprehensibility and accentedness rating scores (Saito, 2021). To judge how comprehensible L2 speech is, listeners appear to pay attention to multiple linguistic dimensions including phonological, temporal, lexical, and grammatical features, whereas for accent judgements, they focus on a limited range of linguistic properties such as segmental and prosodic accuracy (e.g., Crowther et al., 2018; Huensch & Nagle, 2021; Isaacs & Trofimovich, 2012; Munro & Derwing, 1995a; Saito, 2021; Saito et al., 2016; Suzuki & Kormos, 2020; Trofimovich & Isaacs, 2012). Accordingly, listeners' judgements of comprehensibility and accentedness involve different attentional weight to a number of linguistic features in a way that listeners rely more heavily on phonological aspects for the accent judgement, while they attempt to collect as much linguistic information as possible to arrive at the comprehensibility judgement (Saito, 2021; Saito et al., 2016).

Word-level comprehensibility and accentedness

As reviewed above, research has increasingly documented that comprehensibility and accentedness are related but distinct constructs of L2 pronunciation proficiency (e.g., Derwing & Munro, 2009, 2015). However, findings of earlier studies have been predomi-

nately based on speech samples elicited at the sentence or passage level. The extent to which comprehensibility and accentedness can be measured as independent constructs for individual words remains underexplored. Martin (2020) is one exception measuring comprehensibility and accentedness for individual words in order to evaluate the effectiveness of homework-based pronunciation training for learners of L2 German. Forty-nine participants completed a word reading and paragraph reading task before and after 10 weeks of pronunciation training. Eight native speakers rated the elicited recordings of words and paragraphs (20 s excerpts) on a 9-point scale of accentedness and comprehensibility. Cronbach's alpha for word-level productions was comparable to paragraph-level production, exceeding .80 for accentedness (α_{word} = .85–.92, $\alpha_{paragraph}$ = .83–.90) and comprehensibility (α_{word} = .89–.95, $\alpha_{paragraph}$ = .89–.95). Due to the redundancy of the results with the same pattern of statistical significance and effect sizes found for the two rating measures, only results for comprehensibility scores were reported. A larger effect of training was found for the word-level measure than the paragraph-level measure. Martin (2020) explained the word-versus-paragraph difference in the size of training effect in light of listener's attention to linguistic features. Because fewer distracting dimensions such as lexical and syntactic aspects were available at the word level, listeners might end up focusing on phonological information in the word ratings. The word-level comprehensibility measure might therefore function as an alternative metric of a linguistically targeted pronunciation measure. Given that training effects in general tend to be clearer when pronunciation learning is assessed through linguistically targeted measures (e.g., segmental accuracy) than through global rating measures (Saito & Plonsky, 2019), the comprehensibility measure in Martin's study might have made the training effect more salient for the word level than the paragraph level. However, Martin's (2020) study raises a potential issue for the validity of a word-level comprehensibility measure. Given that "distracting dimensions" present in the ratings of connected speech provide linguistic clues for listeners to distinguish comprehensibility from accentedness (e.g., Trofimovich & Isaacs, 2012), the remaining linguistic features—phonological properties—could be predominant sources of linguistic information at the word level that listeners focus on to assign both accentedness and comprehensibility scores. The resulting correlation between accentedness and comprehensibility ratings (although the correlation was not reported in Martin's study) could be much higher at the lexical level than found in earlier studies adopting sentence- or paragraph-level measures (e.g., $r > .90$). Another possible issue is that the word-level comprehensibility judgement might no longer serve as a global measure, hence interchangeable with a linguistically targeted specific measure. A most likely candidate is a segmental measure, considering that replacing or deleting L2 sounds might significantly change the meaning of a word, having a negative impact on listeners' perception of word pronunciation. If the global measure of comprehensibility is replaceable with a segmental measure or the distinction of comprehensibility and accentedness is not maintained at the word level, we might run the risk of missing the intended target construct and misusing the comprehensibility rating score as an indicator of speakers' accentedness or segmental accuracy.

The current study

The current study set out to address the issue that no research has systematically investigated the validity of word-level measures of comprehensibility and accentedness. It is not surprising that previous L2 pronunciation studies have mainly used sentence- or paragraph-level stimuli because this is the way in which messages are most often communicated in real-life conversation. The motivation for this study rather stemmed from the lack of attention to the spoken form of L2 words and the limited repertoire of spoken vocabulary measures available in the domain of L2 vocabulary research. In fact, previous L2 vocabulary studies rarely pay attention to different aspects of word pronunciation (apart from segmental and word stress accuracy, see Nation, 2013, p. 65) or justify the choice of pronunciation measures with a listener perspective considered. If different aspects of L2 speech (i.e., comprehensibility and accentedness) observed in previous L2 pronunciation studies are also confirmed at the lexical level, such findings should advance the understanding of different dimensions of L2 word pronunciation and inform methodological choice in measuring L2 spoken vocabulary knowledge.

Therefore, the aim of the current study was to investigate the validity of the word-level comprehensibility measure by re-examining the extent to which comprehensibility and accentedness are partially independent when the pronunciation of individual words is targeted. For this purpose, the data gathered as a part of the large-scale vocabulary training study conducted by Uchihara (2020: Study 1) was re-analyzed to investigate the interrelationships among intelligibility, comprehensibility, processing time, and accentedness of Japanese speakers' productions of English words. This study also attempted to determine the relative contributions of a range of linguistic features of L2 speech (segmental, word stress, rhythm, fluency) to word-level ratings of comprehensibility and accentedness. This study was guided by the following research questions:

1 To what extent are comprehensibility, accentedness, intelligibility, and processing time related to one another when they are measured at the word level?
2 To what extent are four linguistic features of L2 speech (segmental, word stress, rhythm, fluency) related to comprehensibility and accentedness when they are measured at the word level?

For the first research question, the following findings were expected on the basis of previous L2 speech studies: (a) distinctive listener rating behavior (more lenient for comprehensibility and harsher for accentedness) (Derwing & Munro, 2009), (b) a stronger relationship between intelligibility and comprehensibility, but a weaker (or lack of) relationship between intelligibility and accentedness (Huensch & Nagle, 2021; Munro & Derwing, 1995a), (c) a considerable variability in the relationship between comprehensibility and accentedness across listeners (Derwing & Munro, 1997; Huensch & Nagle, 2021; Munro & Derwing, 1995b), and (d) a stronger relationship of processing time with comprehensibility than with accentedness (Ludwig & Mora, 2017; Munro & Derwing, 1995b). However, these hypotheses were based on previous studies measuring L2 speech at the sentence or paragraph level. The distinctive relationships among the four speech measures found in earlier studies may not persist at the word level. In particular, due to the limited

availability of linguistic information in the production of individual words (Martin, 2020), listeners might experience difficulty in differentiating comprehensibility from accent ratings, resulting in the two global measures being indistinguishable (e.g., much stronger correlation between accentedness and comprehensibility, little variation in the correlation across listeners, the same rating behavioral pattern). For the second research question, it was predicted that (a) the production of words with fewer segmental errors would be more comprehensible and nativelike with a larger effect expected for accentedness (Saito, 2021), (b) production of words with correct stress placement would be more comprehensible and nativelike (Saito et al., 2016; Trofimovich & Isaacs, 2012), (c) production of words with appropriate English stress patterns (emphasizing stressed vowels and reducing unstressed vowels) would be more comprehensible and nativelike (Trofimovich & Isaacs, 2012), and (d) optimally fluent production of words would be more comprehensible and nativelike with a larger effect expected for comprehensibility (Munro & Derwing, 2001; Pinget et al., 2014; Saito, 2021; Suzuki & Kormos, 2020). Alternatively, due to the limited linguistic information available at the word level (Martin, 2020), listeners might rely heavily on the most conspicuous feature of L2 speech (i.e., segmental errors), which might solely contribute to both comprehensibility and accentedness judgements to the same degree. If the word-level comprehensibility rating is informed by the quality of various linguistic sources (e.g., segmental and word stress accuracy) rather than segmental accuracy alone, it is expected that additional linguistic measures other than a segmental measure would jointly contribute to the listener judgement of comprehensibility, which would eliminate the possibility that a word-level comprehensibility measure is interchangeable with a segmental measure.

Method

Overview of the study

The source of the data analyzed in the current study came from the vocabulary training study by Uchihara (2020: Study 1). In their previous study, 75 Japanese EFL university learners studied 40 low-frequency, concrete English words in a paired-associates learning condition where they received repeated exposures to the spoken word forms with the meanings conveyed through pictorial information. Learners completed a word production task (i.e., picture naming) three times (pretest, immediate posttest, and delayed posttest) and their production of words was evaluated by native speakers of English. The current study revisited a portion of the data from 12 Japanese learners producing 37 words at immediate posttests.[1] Production of target words was elicited via a picture naming task (i.e., recall of spoken word forms cued by pictures) immediately after participants completed the vocabulary training program. A total of 307 speech samples excluding items that participants failed to recall were assessed through a timed dictation task by two native speakers of English for intelligibility and processing time and rated for comprehensibility and accentedness by an additional panel of 19 native speakers of English. Unlike earlier studies (e.g., Munro & Derwing, 1995a; Huensch & Nagle, 2021), the two raters providing the data of intelligibility and processing time were different from the 19 raters who evaluated comprehensibility and accentedness. All data were initially planned to be collected from the same listeners for all pronunciation measures (intelligibility, processing time, comprehensibility, and accentedness), but due to the spread of the worldwide pandemic, data collection was suspended after two raters completed all listening sessions. Resuming online data collection involving a timed dictation task to measure intelligibility and response time was not feasible, and consequently a follow-up data collection focused on scalar ratings of comprehensibility and accentedness. Due to this methodological decision, smaller associations between the data derived from a timed dictation task (intelligibility and processing time) and scalar ratings (comprehensibility and accentedness) might be expected in comparison to earlier pronunciation studies. The speech data were coded linguistically for segmental accuracy (phonemic accuracy), word stress accuracy (stress placement accuracy), rhythm (vowel duration ratio), and fluency (articulation rate).

Participants

Speakers. Twelve Japanese university L2 English students with a mean age of 20 years (5 females, 7 males) participated in this study (see Table 1 for information about participants). According to the updated Vocabulary Levels Test (Webb et al., 2017), all students had considerable knowledge of the form-meaning connections of the most frequent 1000 and 2000 word families (M = 29/30, SD = 1 for both frequency levels), and weak-to-moderate knowledge of the 3000 (M = 26/30, SD = 3), 4,000 (M = 23/30, SD = 4), and 5000 (M = 19/30, SD = 5) frequency levels. One participant (ID 11) started learning English earlier than others and stayed in America for one month, but his overall vocabulary knowledge was within the expected range for Japanese EFL learners (within 2 *SDs* of the VLT scores).

Listeners. Nineteen native speakers of English (12 females, 7 males) evaluated Japanese speakers' productions of individual words for comprehensibility and accentedness. All participants spoke a variety of North American English (11 from Canada, 8 from America). Their familiarity with Japanese-accented English was moderate (1 = *not familiar at all*, 6 = *very familiar;* M = 4.4, SD = 1.5). Eight participants had taught English to L2 learners in various contexts (e.g., teaching primary and secondary school students, teaching conversational English at private companies), and eight participants reported having some knowledge about English or Japanese linguistics through taking university courses. They had no hearing problems.

[1] This study focused on 12 speakers for the sake of feasibility (i.e., having raters listening to and linguistically coding 37 items produced by each speaker). The minimal number of participants was based on one of the first studies by Munro & Derwing (1995a) (N = 10). Later, Munro & Derwing (2020) reanalyzed the same data set using a mixed-effects modeling, the same approach adopted in the current study. The 12 participants were randomly selected from a pool of 75 speakers within the range of two standard deviations of the mean for the VLT scores.

Table 1
Information about Japanese EFL speakers.

Participant ID	Age of testing	Age of learning	Overseas experience	VLT (overall)
01	19	12	No	130
02	19	12	No	130
03	20	12	No	134
04	18	10	No	127
05	21	12	No	118
06	20	12	No	140
07	18	12	No	104
08	19	12	No	126
09	23	12	No	110
10	20	9	No	135
11	21	3	1 month (USA)	141
12	18	12	No	113

Note. VLT = Vocabulary Levels Test (Max = 150).

Two additional native English raters, both speakers of North American English (one female from America, one male from Canada), completed a timed dictation task from which data were derived for measuring intelligibility and processing time (for a detailed description about the timed dictation task, see Intelligibility and Processing Time below). They were considered expert raters having extensive language teaching and speaking assessment experiences targeting learners with different L1 backgrounds in various countries (e.g., Korea, China, Canada). Their familiarity with Japanese-accented English was moderate (2 and 3 in response to 1 = *not familiar at all*, 6 = *very familiar*). They had no hearing problems. The two raters completing a timed dictation task were considered more experienced than the 19 listeners assessing comprehensibility and accentedness in that beside ample L2 teaching experience, the two raters were doctoral students engaging in a number of linguistics-related research projects (but not specializing in L2 speech research). Given that research shows no strong evidence of a substantial difference in raters' behaviors between linguistically experienced and novice listeners (Isaacs & Thomson, 2013), potential effects of listeners' background on their ratings were considered minimal in the current study.

Materials

In the previous study (Uchihara, 2020: Study 1), 40 low-frequency, concrete English words were selected as target items. Japanese speakers heard target words recorded by a native English speaker while presented with pictures conveying the meanings of the words (target words and visual stimuli are available in Supplementary Material). The participants were asked to learn as many of the words as possible without being given any pronunciation instruction. Immediately after listening to the words, participants were asked to produce the words corresponding to the same pictures shown on the computer screen twice orally. If participants did not remember a word, they were instructed to move to the next item. Their speech was recorded with a TASCAM DR-05 audio recorder and digitized into a wav format (44.1 kHz sampling rate with 16-bit quantization). One out of two productions per word (i.e., a speech sample without fillers or self-corrections during articulation) was selected and stored in an individual sound file, with peak intensity normalized using Praat (Boersma & Weenink, 2019). The pretest result showed that three words (i.e., *clover, chandelier, escalator*) were considered known to 12 Japanese speakers prior to the experiment. Therefore, speech samples for the current study were based on the production of the remaining 37 words.[2] Prior to data collection, issues with clarity of visual stimuli and testing procedures were resolved through a pilot study with 20 university students with a similar learning background. Data for pilot study participants were not included in the main data analysis.

Comprehensibility and accentedness ratings

The researcher arranged a virtual meeting with the 19 native speakers individually for rating sessions. A total of 307 spoken words elicited through the picture naming task were played once and rating responses were recorded using an online experiment builder, Gorilla (Anwyl-Irvine et al., 2020). As operationalized in the existing literature (e.g., Derwing & Munro, 1997), raters were asked to familiarize themselves with the target words prior to the rating sessions. As such, it was ensured that listeners' comprehensibility and accentedness evaluations would not be confounded with the familiarity effects (i.e., the listeners' likelihood to become more familiar with the content of speech materials and thus more lenient about their assessments as they engage in more exposure). The current study also followed earlier pronunciation studies (e.g., Trofimovich & Isaacs, 2012) wherein the presentation of speech stimuli was randomized across listeners so that potential familiarity effects were minimized (for a discussion of familiarity effects, see Munro & Derwing, 2020, p. 295). Listeners received a brief description of two pronunciation criteria (see Appendix A)—accentedness (1 = *no accent*, 9 = *extremely strong accent*) and comprehensibility (1 = *easy to understand*, 9 = *extremely difficult to understand*)—and went

[2] The current study focused exclusively on unknown words indicated by the pretraining picture-naming test in order to simulate the context of vocabulary learning research (which often targets unknown words) and examine the extent to which word-level pronunciation measures would be reliable and valid in such specific contexts.

through a practice set of 12 items, three of which were produced by native speakers of English. The researcher confirmed that all 19 raters understood the rating procedure and assigned a rating of 1 (*no accent* and *easy to understand*) to samples produced by native English speakers. The practice trial was followed by a main rating session in which listeners evaluated speech samples with interim breaks when necessary.

Intelligibility and processing time

Two expert raters completed a timed dictation task programmed using PsychoPy (Peirce, 2007). In this task, raters listened to each of the speech samples and typed the spelling of the word they heard as fast as possible. Recordings were played only once. Intelligibility score per rater was derived from transcription accuracy (1 = accurate, 0 = inaccurate) with minor misspellings considered accurate (e.g., *chisle, camelieon, ladel*). Processing time (in milliseconds) was defined as the time lapse between the onset of the audio recording and the first keystroke on the computer keyboard. The processing-time measure indicated the degree to which listeners' retrieval of meanings (or messages) conveyed through L2 speech is effortless or effortful (Ludwig & Mora, 2017; Munro & Derwing, 1995b). The current study followed the standard procedure of including data from only those words that were transcribed correctly by two raters via a timed dictation task (Munro & Derwing, 1995b). This procedure was necessary because data from incorrectly transcribed words would be confounded by construct-irrelevant variables. Before completing the rating task, raters completed a practice set of 15 samples representing varying pronunciation qualities (not included in the main dataset). The rating session was implemented individually in the researcher's office.

Linguistic coding

Linguistic features of spoken target words were coded and analyzed for segmental accuracy, word stress, rhythm, and fluency. The choice of the four linguistic features was motivated by earlier studies analyzing speech samples beyond the word level (Isaacs & Trofimovich, 2012; Saito et al., 2019; Saito et al., 2016; Suzuki & Kormos, 2020; Trofimovich & Isaacs, 2012).

Segmental accuracy. Following the definition of segmental errors by Saito et al. (2019), the current study adopted the following three categories for measuring segmental accuracy: (a) unintelligible interlanguage forms, (b) Japanese-like but intelligible pronunciation, and (c) accurate and intelligible L2 English pronunciation. Spoken words with serious errors to the extent that they compromise word intelligibility (e.g., substituting initial /b/ with /k/ in *binoculars*) were categorized under (a).[3] Spoken words with a lack of effort to pronounce L2 sounds (e.g., substituting L1 counterparts and inserting extra vowels within consonant clusters) were categorized under (b). The rest of samples pronounced accurately with an effort to produce L2 sounds (but not necessarily nativelike) were categorized under (c). The researcher and a native Japanese-speaking teacher who had extensive English language teaching experience in EFL and ESL programs independently coded 100 speech samples (not included in the main dataset). A Cohen's kappa analysis confirmed high inter-coder agreement (k = .963). After disagreements were resolved through discussion, the remaining speech samples were coded by the researcher.

Word stress accuracy. Following the definition of word stress errors by Trofimovich & Isaacs (2012), the current study adopted three categories for measuring word stress accuracy: (a) flat or missing primary stress, (b) stress misplacement (e.g., *TREAD-mill* spoken as *tread-MILL*), and (c) correct primary stress placement. The researcher and the same rater who coded segmental errors independently coded 100 speech samples (not included in the main dataset). A Cohen's kappa analysis confirmed high inter-coder agreement (k = .967). After disagreements were resolved through discussion, the remaining speech samples were coded by the researcher.

Rhythm. Following the definition of rhythm by Trofimovich & Isaacs (2012), it was measured as vowel duration ratio (i.e., duration ratio of unstressed to stressed vowels). In English, successful reduction of unstressed vowels in duration is one of the key characteristics determining acquisition of rhythm and more advanced L2 pronunciation proficiency (Trofimovich & Baker, 2006). Using Praat, the duration (in milliseconds) of stressed and unstressed vowels was measured manually between two cursors placed at the onset and offset of voicing in each vowel. The ratio of unstressed to stressed vowels was calculated by dividing the duration of unstressed vowels by that of stressed vowels (when multiple unstressed vowels were available, average duration was calculated).

Fluency. The current study operationalized fluency as the word articulation rate to measure speed fluency, one of the triad measures widely adopted in L2 speech research (speed, breakdown, and repair fluency; e.g., Suzuki & Kormos, 2020; Suzuki et al., 2021). Since the speech data were based on production of individual words, breakdown and repair fluency measures were not measurable in this study. Using Praat, the duration of word production (in milliseconds) was measured manually between two cursors placed at the onset and offset of spoken words. The production duration was divided by the number of syllables per word.

Data analysis

In response to the first research question regarding the interrelationships among four listener-based measures (comprehensibility, accentedness, intelligibility, and processing time), three mixed-effects models were analyzed with: (a) accentedness and comprehensibility as predictor variables and intelligibility as the outcome variable, (b) accentedness and intelligibility as predictor variables and

[3] The relationship between segmental accuracy and intelligibility (orthographic transcription) appeared linear: the production of words coded accurate (c) was transcribed accurately 100% of the time, the production of words coded accurate but with Japanese L1 sounds (b) was transcribed accurately 98% of the time, and the production of words coded inaccurate (a) was transcribed accurately 70% of the time.

comprehensibility as the outcome variable, and (c) accentedness and comprehensibility as predictor variables and processing time as the outcome variable. The rationale for conducting mixed-effects modeling analyses (e.g., instead of simple correlation analyses, see Munro & Derwing, 1995a) is that the analysis allows for taking account of systematic variability due to the individual speakers, words, and listeners in a single model, considered an improved statistical technique to examine the relationship between L2 speech measures (Huensch & Nagle, 2021; Nagle & Huensch, 2020; Munro & Derwing, 2020). The intelligibility model (a) was built following Munro & Derwing (2020) predicting that accurately pronounced words whose phonetic forms are perceived to be easy to understand and nativelike-sounding would help native listeners decode the spoken words efficiently and transcribe them accurately (Huensch & Nagle, 2021). For the comprehensibility model (b), it was expected that the spoken forms of words which sound nativelike (i.e., matching or approximating representations of words in native listeners' long-term memory) and recognized correctly (i.e., accurate decoding of individual sounds) would reduce the listeners' cognitive load in accessing and retrieving the word meanings, thereby increasing the degree of ease in which the spoken words are understood (Floccia et al., 2009). The processing-time model (c) was built on the theoretical basis that when the L2 phonological form of a word matches (or approximates) the corresponding phonological representation stored in the listeners' mental lexicon, their processing speed during word recognition would be accelerated (Ludwig & Mora, 2017). Such representation-matched L2 spoken forms might be more easily understood and perceived to be more nativelike were predicted to have a positive impact on listeners' processing times. However, a stronger link between comprehensibility and processing time was expected (Munro & Derwing, 1995b).

For intelligibility, all word responses were binary coded with spoken words transcribed correctly by two raters coded as accurate. For processing time, the raw data was log-transformed and averaged to yield a single score per speaker. For the first model with binary intelligibility data, a generalized model was built with by-speaker and by-word random intercepts fitted. For the second model with comprehensibility ratings, a mixed effects model was built with by-speaker, by-word, and by-listener random intercepts fitted. For the third model with processing time, a mixed effects model was built with by-speaker and by-word random intercepts fitted. All predictor variables (except intelligibility score, which was dummy-coded) were grand-mean centered and statistical assumptions for regression analysis (normality, homoscedasticity, collinearity) were confirmed. A total of 5,827 observations were available for data analysis except for the analysis on processing time (number of observations = 4,788) to answer the first research question.

In response to the second research question regarding the relationship between linguistic features of L2 speech (segmental, word stress, rhythm, fluency) and listener-based constructs (comprehensibility and accentedness), a mixed-effects modeling analysis was conducted for accentedness and comprehensibility separately, with by-speaker, by-word, and by-listener random intercepts fitted. In each model, predictor variables included segmental accuracy (dummy-coded), word stress accuracy (dummy-coded), vowel duration ratio (grand-mean centered), and fluency (grand-mean centered) with the number of syllables as a covariate (grand-mean centered). Before conducting the main analyses, the validity for the three categories for coding segmental accuracy and word stress accuracy was examined. For word stress accuracy, the different types of errors (flat vs. misplacement) did not have significant effects on either accentedness or comprehensibility, and therefore two types of data were combined to yield a binary single score for word stress accuracy (0 = inaccurate stress placement, 1 = accurate stress placement). Because a number of samples were not codable for vowel stress measures due to deletion of target vowels and significant changes to syllable structures, 474 samples were identified as missing data, resulting in a total of 5,353 observations available for data analysis to answer the second research question. Descriptive statistics of all speech measures and intercorrelations among them were provided in Appendix B and Appendix C, respectively (the raw data for speech measures are available at https://osf.io/nc7yp/).

Results

Relationships among accentedness, comprehensibility, processing time, and intelligibility

Prior to conducting mixed-effects modeling analyses to answer research questions, Cronbach's alpha was computed to examine the interrater consistency for accentedness and comprehensibility scores. The alpha for both accentedness (α = .942, 95% CI [.932, .951]) and comprehensibility (α = .964, 95% CI [.958, .970]) were comparable to the average interrater consistency reported in Saito's (2021) meta-analysis of paragraph-level measures of accentedness (α = .909, 95% CI [.864, .954]) and comprehensibility (α = .896, 95% CI [.871, .920]). The Pearson correlation between comprehensibility and accentedness ratings based on the observation date (5,827 observations) was .665 ($p < .001$), indicating that the two constructs were moderately associated with each other. However, the frequencies of listeners' rating responses showed a distinctive pattern. Fig. 1 indicates that for accentedness ratings, listeners' responses tend to be evenly distributed or clustered around the strongly-accented end of the scale (Categories 6, 7, and 8), whereas listeners tend to be more lenient with comprehensibility ratings and their responses were clustered around the easy-to-understand end (Categories 1, 2, and 3). These rating patterns for word-level measures were consistent with findings of earlier studies for sentence- or paragraph-level measures (Derwing & Munro, 2009).

In answer to the first research question regarding the interrelationships among four listener-based measures (comprehensibility, accentedness, intelligibility, processing time), three sets of mixed-effects modeling analyses were conducted. First, a generalized mixed-effects model analysis was conducted to examine the extent to which accentedness and comprehensibility predict intelligibility score. Table 2 summarizes the results of the intelligibility model with odds ratios (*OR*) computed by exponentiating the log odds of each fixed effect. There was a significant negative association between comprehensibility and intelligibility ($B = -0.35$, $OR = 0.70$, $p < .001$). Since lower ratings of comprehensibility meant that the pronunciation of words was easier to understand (1 = *easy to understand*, 9 = *extremely difficult to understand*), the result of *OR* indicated that the odds of a target word transcribed accurately increased by 30% when the word-level comprehensibility increased by one category. In contrast, accentedness ratings (1 = *no accent*, 9 = *extremely*

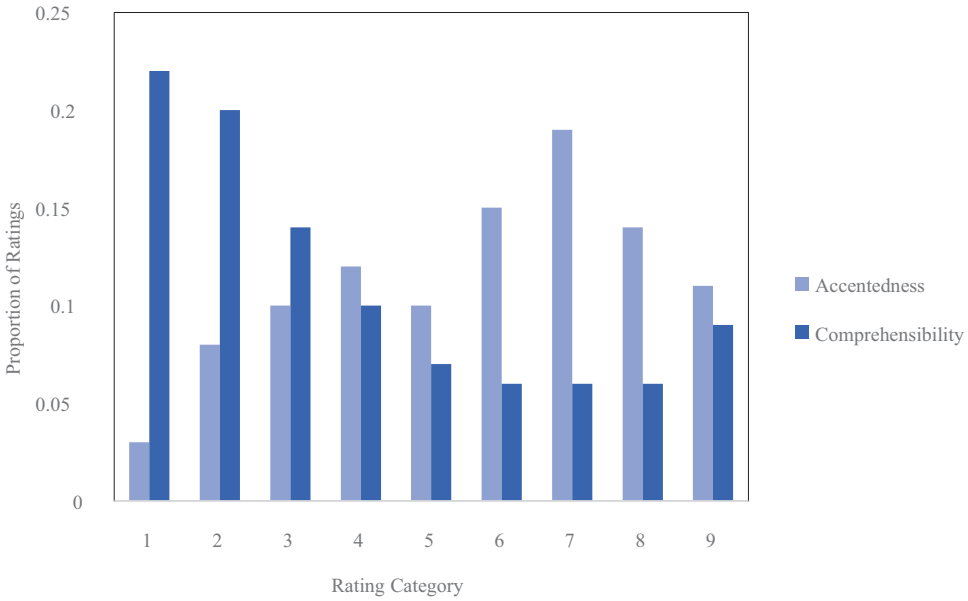

Fig. 1. Frequency of listeners' rating responses for accentedness and comprehensibility.

Table 2
Summary of generalized mixed-effects model fit to the binary intelligibility score.

Fixed effects	*B*	*SE*	*OR*	95% CI	*z*	*p*
Intercept	3.52	0.67	33.79	[9.13, 125.08]	5.27	< .001
Comprehensibility	–0.35	0.02	0.70	[0.67, 0.74]	–14.71	< .001
Accentedness	–0.04	0.03	0.96	[0.91, 1.02]	–1.30	.192
Random effects	*SD*	Variance				
By-speaker intercepts	1.69	2.86				
By-word intercepts	2.54	6.43				

Note. OR = odds ratio, calculated by exponentiating the log odds of each fixed effect, or exp(*B*).

Table 3
Summary of mixed-effects model fit to comprehensibility ratings.

Fixed effects	*B*	*SE*	95% CI	*t*	*p*
Intercept	4.95	0.25	[4.46, 5.44]	19.90	< .001
Accentedness	0.71	0.05	[0.62, 0.81]	14.70	< .001
Intelligibility	–1.44	0.06	[–1.56, –1.32]	–23.60	< .001
Random effects	*SD*	Variance			
By-speaker intercepts	0.28	0.08			
By-word intercepts	0.43	0.18			
By-listener intercepts	0.95	0.90			
By-listener slopes: Accentedness	0.21	0.04			

Note. Random parameter correlation (by-listener intercepts–by-listener slopes: accentedness) = .207.

strong accent) were not significantly associated with intelligibility scores ($B = –0.04$, $OR = 0.96$, $p = .192$). The intelligibility model with two predictors (comprehensibility and accentedness) explained 7% of the variance in intelligibility scores with only fixed effects considered (marginal $R^2 = .07$) and the variance explained by the model increased up to 76% once the random effects were considered (conditional $R^2 = .76$).

Table 3 reports the comprehensibility model. There was a significant positive relationship between accentedness and comprehensibility ($B = 0.71$, $p < .001$), indicating that a word rated as one category more nativelike was estimated to increase the word's comprehensibility by 0.71 category. Following Huensch & Nagle's (2021) procedure, by-listener random slopes for accentedness were added to the model in order to examine the extent to which the accent-comprehensibility association varies across listeners. Including by-listener slopes for accentedness significantly improved model fit $\chi^2(2) = 224.76$, $p < .001$, suggesting a significant variation in the relationship between comprehensibility and accentedness across listeners. Intelligibility was a significant predictor of comprehensi-

Table 4
Summary of mixed-effects model fit to processing time.

Fixed effects	*B*	*SE*	95% CI	*t*	*p*
Intercept	0.064	0.034	[–0.002, 0.130]	1.88	.067
Comprehensibility	0.020	0.002	[0.017, 0.023]	12.52	< .001
Accentedness	0.009	0.002	[0.006, 0.013]	5.87	< .001
Random effects	*SD*	Variance			
By-speaker intercepts	0.069	0.005			
By-word intercepts	0.163	0.027			

Table 5
Summary of relationships between linguistic features and global constructs of comprehensibility and accentedness.

Fixed effects	*B*	*SE*	95% CI	*t*	*p*
Comprehensibility model					
SA (1): L1-like	–2.40	0.09	[–2.56, –2.21]	–26.40	< .001
SA (2): accurate	–3.61	0.12	[–3.84, –3.37]	–30.06	< .001
Word stress accuracy	–0.53	0.11	[–0.74, –0.31]	–4.85	< .001
Vowel duration ratio	–0.38	0.08	[–0.55, –0.22]	–4.51	< .001
Articulation rate	–0.63	0.83	[–2.26, 1.00]	–0.76	.449
Number of syllables	–0.38	0.16	[–0.69, –0.07]	–2.43	.019
Random effects	*SD*	Variance			
By-speaker intercepts	0.50	0.25			
By-word intercepts	0.46	0.21			
By-listener intercepts	0.85	0.72			
Accentedness model					
SA (1): L1-like	–1.16	0.08	[–1.33, –0.10]	–13.77	< .001
SA (2): accurate	–2.48	0.11	[–2.70, –2.27]	–22.31	< .001
Word stress accuracy	–0.36	0.10	[–0.56, –0.16]	–3.57	< .001
Vowel duration ratio	–0.16	0.08	[–0.32, –0.01]	–2.07	.039
Articulation rate	0.79	0.76	[–0.71, 2.28]	1.03	.304
Number of syllables	–0.06	0.12	[–0.30, 0.18]	–0.45	.654
Random effects	*SD*	Variance			
By-speaker intercepts	0.50	0.25			
By-word intercepts	0.60	0.36			
By-listener intercepts	0.86	0.74			

Note. Comprehensibility model: marginal R^2 = .19; conditional R^2 = .42. Accentedness model: marginal R^2 = .10; conditional R^2 = .36. SA = segmental accuracy. For SA, the reference category was inaccurate segmental pronunciation. Comprehensibility: accurate vs. L1-like (B = –1.21, SE = 0.09, 95% CI [–1.38, –1.04], t = –14.16, p < .001); Accentedness: accurate vs. L1-like (B = –1.32, SE = 0.08, 95% CI [–1.48, –1.17], t = –16.66, p < .001).

bility scores, indicating that words that were typed out accurately were likely to increase the word's comprehensibility by 1.44. In the comprehensibility model, the fixed effects alone explained 49% of the variance in comprehensibility scores (marginal R^2 = .49), and the additional random effects increased the explained variance up to 68% (conditional R^2 = .68).

Lastly, to examine the relationship between listener-based measures (comprehensibility and accentedness) and processing time, the mixed-effects model for processing time was analyzed and the result was summarized in Table 4. Both comprehensibility and accentedness were significantly predictive of processing time. Processing time appeared more strongly associated with comprehensibility (B = 0.020, t = 12.52, p < .001) than accentedness (B = 0.009, t = 5.87, p < .001), indicating that a word rated as one category more comprehensible and nativelike was estimated to expedite the first keystroke 20 and 9 ms faster respectively. According to Plonsky & Oswald's (2014) effect-size benchmarks (small: r = .25, medium: r = .40, large: r = .60), the correlation between processing time and comprehensibility (r = .390) was considered medium in comparison to a small effect for the correlation between processing time and accentedness (r = .290). The mixed effects model with the fixed effects alone explained 5% of the variance in processing time (marginal R^2 = .05) and the additional random effects increased the explained variance up to 51% (conditional R^2 = .51).

Linguistic correlates of word-level comprehensibility and accentedness

In answer to the second research question regarding the linguistic correlates of listener-based constructs, Table 5 reports a summary of relationships between linguistic predictors (segmental, word stress, rhythm, fluency) and listener-based speech measures (comprehensibility, accentedness). Results of segmental accuracy and word stress accuracy show that the production of words with fewer segmental and stress placement errors was likely to be perceived as more comprehensible and nativelike. Results of vowel duration

ratio and articulation rate were rather unexpected, showing that faster word production was not significantly related to comprehensibility or accentedness, and words pronounced with more nativelike rhythm (reducing unstressed vowels and emphasizing stressed vowels) were perceived to be less comprehensible and more heavily accented. The number of syllables was not significantly related to accentedness but a significant predictor of comprehensibility, indicating that the production of longer words was likely to be more comprehensible. The comprehensibility model explained 19% of the variance in comprehensibility with the fixed effects alone (marginal R^2 = .19) and 42% of the variance was explained when the random effects were added (conditional R^2 = .42). The accentedness model explained 10% of the variance in accentedness with the fixed effects alone (marginal R^2 = .10) and 36% was explained when the random effects were considered (conditional R^2 = .36).

Discussion

In answer to the first research question regarding the interrelationships among four listener-based measures (comprehensibility, accentedness, intelligibility, processing time), the current data showed that (a) comprehensibility and accentedness ratings were associated, but the strength of the association significantly varied across listeners, (b) listeners' rating response pattern was distinctive for comprehensibility (more lenient) and accentedness (more strict), and (c) compared to accentedness, comprehensibility was more closely associated with intelligibility (transcription of L2 spoken words) and processing time (reaction-time measure). These findings based on word-level stimuli are consistent with previous studies measuring L2 pronunciation at the sentence or paragraph level (Derwing & Munro, 1997, 2009; Huensch & Nagle, 2021; Munro & Derwing, 1995a, 1995b). Based on these findings altogether, the current study supports the view that the two global constructs measured for individual words are related yet partially independent and suggests that word-level comprehensibility can be measured as a proxy of learners' word pronunciation knowledge, distinct from whether the production of words approximates nativelike pronunciation. Results of linguistic correlates of comprehensibility also confirm that the word-level comprehensibility measure can serve as a global pronunciation measure because listeners relied on different aspects of L2 speech (segmental, word stress, stress timing features) rather than focusing exclusively on a single linguistic feature.[4]

In answer to the second research question regarding the relationship between linguistic features of L2 speech (segmental, word stress, rhythm, fluency) and listener-based constructs (comprehensibility and accentedness), first, the results of segmental and word stress accuracy supported the findings of earlier studies (Crowther et al., 2018; Saito, 2021; Saito et al., 2016; Suzukida & Saito, 2019; Trofimovich & Isaacs, 2012). The pronunciation of words with fewer phonemic errors were more easily understood and perceived more nativelike than the Japanese-like pronunciation of English words or the pronunciation of words with more serious errors that compromised word intelligibility. Correctly placing primary stress on stressed vowels in word pronunciation increased the chance that the spoken forms of words were perceived more comprehensible and nativelike. One notable difference between the current study (word-level) and earlier studies (sentence- or paragraph-level, see Saito, 2021) is that segmental accuracy was observed to play a major role in the comprehensibility judgement when measured at the word level, as evidenced by a relatively larger increase in comprehensibility from inaccurate to accurate pronunciation of individual words (B = –3.61) contrasting with the result for accentedness (B = –2.48). In rating sentence- or paragraph-level pronunciation, listeners could use a wide range of linguistic information (e.g., lexical, syntactic, and discursive features) as well as background or world knowledge to compensate for their lack of understanding caused by segmental errors in context (Saito et al., 2016). In rating word-level pronunciation, listeners may not be able to use such compensation strategies and the segmental sound could be the primary source of information available for listeners to use in order to arrive at the correct meaning of spoken words. Due to listeners' increased attention to segmental features, segmental accuracy might have a relatively stronger impact on comprehensibility compared to accentedness ratings.

The results for vowel duration ratio and articulation rate were unexpected. It was initially predicted that pronouncing words with an appropriate English stress-timed rhythm (emphasizing stressed vowels and reducing unstressed vowels) and faster articulation of words would increase comprehensibility and nativelikeness. However, as for vowel duration ratio, the direction of the relationship with global measures was opposite, indicating that the production of words with nativelike rhythm was perceived to be less comprehensible and more heavily accented. At the word level, it is possible that native listeners do not pay as much attention to vowel duration as to other acoustic features such as vowel quality to determine the degree to which words are pronounced with nativelike stress timing (Zhang & Francis, 2010). Japanese learners who can quickly learn to produce nativelike durational patterns are likely to overuse durational cues (and underuse spectral cues) to realize the difference between stressed and unstressed vowels (Lee et al., 2006). Accordingly, the nativelike (or near nativelike) use of durational patterns without the reduction of vowel quality (using centralized schwa-like vowels) might not have had a positive influence on native listeners' perception of L2 speech. This explanation does not fully account for why duration measures negatively related to listeners' judgements and should be considered speculative. The differential impacts of vowel quality and duration on listener perception of L2 speech needs to be further investigated in future research. Additionally, the finding that fluency was not significantly associated with comprehensibility (and accentedness) needs further exploration. Again, the length of speech samples could be considered a factor causing these conflicting results. At the paragraph level, breakdown fluency (e.g., frequent silent pauses) in addition to speed fluency (e.g., articulation rate) are strong predictors of L2 speech comprehensibility (Suzuki & Kormos, 2020). However, at the word level, the fluency measure used in this study was solely

[4] It is important to note that the "global" pronunciation measure indicates that a listener-based measure reflects multiple linguistic features of L2 speech, and the term should be not confused with the "general" pronunciation measure. To assess learners' general proficiency, we need to consider real-life communication where speakers produce not only individual words accurately but also a longer stretch of words, sentences and paragraphs, using target-like rhythmic and intonational patterns in a contextually appropriate manner.

indicated by articulation rate. It is also not appropriate to assume that word-level articulation rate is equivalent to paragraph-level fluency, because the former primarily underlies the quality of spoken form knowledge (how accurately phonological information is encoded and represented), while the latter reflects a wider array of cognitive processes such as conceptual, morphological, and syntactic processing (Kormos, 2006). What is clear from this finding is that pronouncing words faster may not necessarily help listeners comprehend the meaning of spoken words better.

Lastly, the number of syllables was significantly associated with comprehensibility but not with accentedness, indicating that the production of longer words was likely to be perceived more comprehensible yet not necessarily more nativelike. This finding supports the view that word-level comprehensibility and accentedness are partially independent constructs. For accent judgement, native listeners can reliably and quickly capture the degree of foreign accent through listening to a fragment of L2 speech (Munro et al., 2010), and therefore their perception of accentedness might not be further assisted with extra information from longer words. In contrast, the comprehensibility judgement reflects listeners' experience of ease or difficulty in processing semantic information of words and utterances (Ludwig & Mora, 2017; Munro & Derwing, 1995b), requiring listeners to collect as much linguistic information as possible to arrive at the meaning of spoken words (Saito et al., 2016). Even if part of a longer word was mispronounced, listeners could compensate their lack of word-meaning comprehension with the aid of remaining segmental clues, whereas they could not use such a compensation strategy effectively when listening to shorter words.

Conclusion

The current study provided initial evidence supporting the partial independence of comprehensibility and accentedness as global constructs of L2 pronunciation knowledge when L2 speech is measured at the word level. It was also confirmed that a word-level comprehensibility measure is not replaceable with a segmental measure or a measure of any given linguistic feature of L2 speech. This section first notes several limitations of the current study with suggestions for future research, followed by discussion regarding methodological implications for measurement of L2 spoken vocabulary knowledge.

Some limitations are worth bearing in mind when interpreting the current findings. First, the number of native-speaking raters (n = 2) who performed a timed dictation task for measuring intelligibility and processing time was limited, and the two raters did not contribute to the rating data for comprehensibility and accentedness. Individual data points per listener for intelligibility and processing time therefore could not be used to analyze the relationship with accentedness and comprehensibility ratings or further explore the degree of between-listeners variability in the link between dictation-based measures (intelligibility and processing time) and scalar rating measures (accentedness and comprehensibility). Future studies recruiting a greater number of raters all of whom contribute to producing four speech measures can, for example, allow researchers to examine the variability in the link between comprehensibility and intelligibility across listeners (Huensch & Nagle, 2021). Finally, the sample size for L2 speakers was limited. The small number of L2 speakers (N = 12) was focused in relation to the large sample for target items (N = 37) in order to account for item-specific variations (i.e., some words are easier to pronounce than other words). A challenge in applying the listener-based measures into the domain of vocabulary research is if eliciting speech samples from a large number of speakers using a large number of target items, the resultant number of samples increases considerably, and workload for listeners to rate all speech samples becomes overwhelming. Future studies may prioritize the sample size of speakers over that of target items in order to examine the replicability of the current findings (e.g., 40 speakers, 10 words).

The current study suggests that the use of a word-level measure of comprehensibility has important implications for L2 vocabulary research. The findings of the current study contribute to methodological improvement by adding another possible measure of spoken vocabulary knowledge to the existing battery of lexical measures. The significance of this study lies in its interdisciplinary approach to integrate a listener's perspective into measurement of the quality of spoken form of L2 vocabulary. This approach is consistent with the ongoing proposal for the need to measure employability of L2 words (Kremmel & Schmitt, 2016; Schmitt et al., 2020). The ultimate goal of vocabulary teaching is to ensure that learners develop the ability to employ L2 words, and the use of a word-level measure of comprehensibility caters to this need. In order to be successful in oral communication, it is crucial to develop not only knowledge of form-meaning connection but also the ability to produce L2 words in a way that the production of words is sufficiently comprehensible to listeners. However, researchers should not be discouraged from using a word-level measure of accentedness to assess the spoken forms of words. What matters, instead, is that we should not conceptualize the knowledge of spoken forms as a monolithic "pronunciation accuracy" or "target-like accuracy" but consider what aspects of word pronunciation knowledge learners are expected to develop and demonstrate, which would inform the choice of specific word-level pronunciation measures. For example, a word-level measure of comprehensibility, rather than accentedness, may be more appropriate to capture a gradual increase in the knowledge of spoken forms in the paradigm of incidental vocabulary acquisition (e.g., learning as a by-product of listening to teacher talk, songs, and academic lectures, see Uchihara et al., 2019 for a review of incidental word learning activities). It may be possible to use both accentedness and comprehensibility measures to gauge learning gains for learners completing word-focused activities (e.g., learning words using word lists, flashcards, sentence writing, and repetition techniques, see Webb et al., 2020 for a review of word-focused activities).

Although the current study suggests the practical value of word-level pronunciation measures in vocabulary assessment, it is important to note that such measures may not be suitable for indicating learners' general pronunciation proficiency. Spoken communication in real-life situations requires not only the ability to pronounce individual words accurately but also the ability to produce a longer stretch of words using rhythmic and intonational patterns in a contextually appropriate manner. With this caveat in mind, future studies of spoken vocabulary learning are encouraged to incorporate listener-based pronunciation measures in addition to

other existing vocabulary measures in order to indicate the extent to which words whose meanings are mapped to L2 forms can be sufficiently accurate and employable in real-life oral communication.

Declaration of Competing Interest

The Author declares that there is no conflict of interest.

Acknowledgment

This research was supported by the Language Learning Dissertation Grant Program (Grant No.: R5370A13) and a Waseda University Grant for Special Research Projects (Project number: 2021C-BARD01107101). I would like to thank Yui Suzukida, Dru Sutton, and Michael Karas for their constructive feedback on the previous version of the manuscript. I also thank Takeshi Hattori and Shuhei Kudo for their assistance with data collection.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rmal.2022.100011.

Appendix A. Descriptions of accentedness and comprehensibility constructs provided for listeners

Word	Explanation
Accentedness	This refers to how much a speaker's speech is influenced by his/her native language and/or is colored by other non-native features.
Comprehensibility	This term refers to how much effort it takes to understand what someone is saying. If you can understand (the word produced by a speaker) with ease, then the speaker is effortless to understand. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker is effortful/difficult to understand.

Appendix B. Descriptive statistics for pronunciation measures ($N = 12$)

	M	*SD*	[95% CI]	Range
Accentedness	5.87	0.78	[5.38, 6.37]	4.38–6.91
Comprehensibility	4.01	1.00	[3.37, 4.64]	2.53–5.39
Intelligibility	0.81	0.15	[0.71, 0.91]	0.56–1.00
Processing time	1.51	0.43	[1.24, 1.78]	0.99–2.16
Articulation rate	0.27	0.02	[0.26, 0.29]	0.24–0.32
Vowel duration ratio	0.94	0.12	[0.87, 1.02]	0.76–1.15
Segmental accuracy	0.97	0.25	[0.81, 1.13]	0.64–1.28
Word stress accuracy	0.87	0.14	[0.78, 0.96]	0.56–1.00

Appendix C. Intercorrelations among pronunciation measures ($N_{speaker} = 12$; $N_{word} = 37$; $N_{rater} = 19$)

	1	2	3	4	5	6	7
1. Accentedness							
2. Comprehensibility	.665						
3. Intelligibility	−.302	−.460					
4. Processing time	.290	.390	N/A				
5. Articulation rate	−.066	−.014	−.038	−.061			
6. Vowel duration ratio	.110	.125	−.082	.086	.050		
7. Segmental accuracy	−.433	−.530	.457	−.294	.106	−.192	
8. Word stress accuracy	−.159	−.227	.173	−.013	.090	−.231	.256

Note. Pearson's *r* was computed based on the non-aggregated observation data (5,827 observations). The correlation between processing time and intelligibility was not calculated because the processing time was available only for speech samples considered accurate by two raters (intelligibility).

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. 10.3758/s13428-019-01237-x.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*(3), 387–414. 10.1017/S0272263105050175.

Boersma, P. & Weenink, D. (2019). *Praat: doing phonetics by computer* [computer program]. Version 6.1.07. http://www.praat.org./.

Bürki, A. (2010). Lexis that rings a bell: On the influence of auditory support in vocabulary acquisition. *International Journal of Applied Linguistics, 20*(2), 206–231. 10.1111/j.1473-4192.2009.00246.x.
Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition, 40*(2), 443–457. 10.1017/S027226311700016X.
Dang, T. N. Y., Lu, C., & Webb, S. (2021). Incidental learning of single words and collocations through viewing an academic lecture. *Studies in Second Language Acquisition*, 1–29. 10.1017/S0272263121000474.
Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1–16. 10.1017/S0272263197001010.
Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(4), 476–490. 10.1017/S026144480800551X.
Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.
Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguistic Research, 38*(4), 379–412. 10.1007/s10936-008-9097-8.
Huensch, A., & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Munro and Derwing (1995a). *Language Learning, 71*(3), 626–668. 10.1111/lang.12451.
Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135–159. 10.1080/15434303.2013.769545.
Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*(3), 475–505. 10.1017/S0272263112000150.
Isbell, D. R., Park, O. S., & Lee, K. (2019). Learning Korean pronunciation. *Journal of Second Language Pronunciation, 5*(1), 13–48. 10.1075/jslp.17010.isb.
Jin, Z., & Webb, S. (2020). Incidental vocabulary learning through listening to teacher talk. *The Modern Language Journal*. 10.1111/modl.12661.
Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review, 20*(6), 1259–1265. 10.3758/s13423-013-0450-z.
Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning, 68*(1), 115–146. 10.1111/lang.12270.
Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum.
Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly, 13*(4), 377–392. 10.1080/15434303.2016.1237516.
Lee, B., Guion, S. G., & Harada, T. (2006). Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals. *Studies in Second Language Acquisition, 28*(3), 487–513. 10.1017/S0272263106060207.
Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation, 6*(3), 310–328. 10.1075/jslp.20050.lev.
Ludwig, A., & Mora, J. C. (2017). Processing time and comprehensibility judgments in non-native listeners' perception of L2 speech. *Journal of Second Language Pronunciation, 3*(2), 167–198. 10.1075/jslp.3.2.01lud.
Martin, I. A. (2020). Pronunciation can be acquired outside the classroom: Design and assessment of homework-based training. *The Modern Language Journal*. 10.1111/modl.12638.
Matthews, J. (2021). Aural vocabulary knowledge. In H. Mohebbi, & C. Coombe (Eds.), *Research questions in language education and applied linguistics* (pp. 439–443). Springer Texts in Education.
Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *The Canadian Modern Language Review, 63*(1), 127–147. 10.3138/cmlr.63.1.127.
Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. 10.1111/j.1467-1770.1995.tb00963.x.
Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*(3), 289–306. 10.1177/002383099503800305.
Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*(4), 451–468. 10.1017/S0272263101004016.
Munro, M. J., & Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation, 6*(3), 283–309. 10.1075/jslp.20038.mun.
Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication, 52*(7), 626–637. 10.1016/j.specom.2010.02.013.
Nagle, C. L., & Huensch, A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation, 6*(3), 329–351. 10.1075/jslp.20009.nag.
Nation, I. S. P. (2013). Learning vocabulary in another language (2nd ed.). Cambridge University Press.
Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Studies in Second Language Acquisition, 41*(4), 745–768. 10.1017/S0272263119000020.
Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1), 8–13. 10.1016/j.jneumeth.2006.11.017.
Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 40*(3), 551–577. 10.1017/S0272263117000407.
Pinget, A. F., Bosker, H. R., Quené, H., & de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing, 31*(3), 349–365. 10.1177/0265532214526177.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. 10.1111/lang.12079.
Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*. 10.1002/tesq.3027.
Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning, 69*(3), 652–708. 10.1111/lang.12345.
Saito, K., Suzukida, Y., & Sun, H. (2019). Aptitude, experience, and second language pronunciation proficiency development in classroom settings: A longitudinal study. *Studies in Second Language Acquisition, 41*(1), 201–225. 10.1017/S0272263117000432.
Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*(2), 217–240. 10.1017/S0142716414000502.
Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition, 19*(3), 597–609. 10.1017/S1366728915000255.
Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53*(1), 109–120. 10.1017/S0261444819000326.
Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition, 42*(1), 143–167. 10.1017/S0272263119000421.
Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal, 105*(2), 435–463. 10.1111/modl.12706.
Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*. 10.1177/1362168819858246.

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), 1–30. 10.1017/S0272263106060013.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*(4), 905–916. 10.1017/S1366728912000168.

Uchihara, T. (2020). *The effects of spoken input on learning the spoken forms of second language words: Studies of frequency of exposure, acoustic variability, and mode of input*. London, ON, Canada: University of Western Ontario Unpublished doctoral dissertation.

Uchihara, T., & Harada, T. (2018). Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly, 52*(3), 564–587. 10.1002/tesq.453.

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2021). The effects of talker variability and frequency of exposure on the acquisition of spoken word knowledge. *Studies in Second Language Acquisition*. 10.1017/S0272263121000218.

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning, 69*(3), 559–599. 10.1111/lang.12343.

Webb, S., Sasao, Y., & Oliver, B. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics, 168*(1), 33–69. 10.1075/itl.168.1.02web.

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal, 104*(4), 715–738. 10.1111/modl.12671.

Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition, 42*(2), 411–438. 10.1017/S0272263119000688.

Zhang, R., & Yuan, Z. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*, 1–14. 10.1017/S0272263120000121.

Zhang, Y., & Francis, A. (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *Journal of Phonetics, 38*(2), 260–271. 10.1016/j.wocn.2009.11.002.