Expanding the scope of L2 intelligibility research

Intelligibility, comprehensibility, and accentedness in L2 Spanish

Charles L. Nagle and Amanda Huensch
Iowa State University | University of Pittsburgh

This study investigated relationships among intelligibility, comprehensibility, and accentedness in the speech of L2 learners of Spanish who completed a prompted response speaking task. Thirty native Spanish listeners from Spain were recruited through Amazon Mechanical Turk to transcribe and rate extracted utterances, which were also coded for grammatical and phonemic errors, and speaking rate. Descriptively, although most utterances were intelligible, their comprehensibility and accentedness varied substantially. Mixed-effects modeling showed that comprehensibility was significantly associated with intelligibility whereas accentedness was not. Additionally, phonemic and grammatical errors were significant predictors of intelligibility and comprehensibility, but only phonemic errors were significantly related to accentedness. Overall, phonemic errors displayed a stronger negative association with the listener-based dimensions than grammatical errors. These findings suggest that English-speaking learners of Spanish are not as uniformly intelligible and comprehensible as FL instructors might believe and shed light on relationships among speech constructs in an L2 other than English.

Keywords: intelligibility, comprehensibility, accentedness, L2 Spanish

1. Introduction

Twenty-five years ago, Munro and Derwing (1995) demonstrated that comprehensibility and accentedness were distinct, listener-based constructs whose relationship to intelligibility varies across listeners. In their study, comprehensibility was strongly aligned with intelligibility, and within-listener correlations ranged from medium to large. In contrast, the relationship between accentedness and

https://doi.org/10.1075/jslp.20009.nag | Published online: 7 July 2020
Journal of Second Language Pronunciation 6:3 (2020), pp. 329–351. issn 2215-1931 | e-issn 2215-194x

intelligibility was generally weaker and more variable. Since that study, second language (L2) speech research has experienced a theoretical and methodological renaissance centered on the three constructs. For instance, over the past few years, a significant body of scholarship has emerged on the linguistic correlates of comprehensibility and accentedness across multiple speaking tasks (Crowther et al., 2018) and target languages (e.g., Bergeron & Trofimovich, 2017; O'Brien, 2014). Yet, most of this work has concentrated on L2 English, and work that has addressed other L2s has focused on comprehensibility as the primary construct of interest. What is needed, then, is a return to intelligibility, comprehensibility, and accentedness in L2s other than English and in different contexts of learning.

The context of the original studies was English as a Second Language (ESL) in Canada, whereas our focus is on Spanish as a Foreign Language (FL) in the United States. Applying constructs generated in the ESL context to the FL context brings with it a series of conceptual questions related to if and how the constructs need to be adapted. For example, ESL speakers need to be able to communicate with members of the local community so that they can fulfill their immediate needs, which means that local listeners are an appropriate evaluation group. In contrast, FL learners are studying the L2 out of personal and/or professional interest and may not come into contact with proficient L2 speakers other than their instructor during the first few years of FL study. Thus, for FL learners, the question of "Intelligible and comprehensible to whom?" is less straightforward, given that the group of native speakers with whom they might interact is largely imaginary until they study or live abroad. Moreover, FL learners likely envision themselves interacting with a range of native speakers in the US and abroad, which further complicates defining a valid reference group for intelligibility, comprehensibility, and accentedness evaluations. On a more theoretical level, relationships among the constructs may depend on L1-L2 pairings, such that we might expect a slightly different portrait to emerge for L2 Spanish, at least in terms of the magnitude of the attested relationships.

L2 Spanish seems like a logical starting point for expanding the scope of intelligibility, comprehensibility, and accentedness research. Spanish is an important world language. In the US context in particular, it is the most frequently studied FL, both in K-12 (approximately 7.3 million learners, representing 70% of K–12 FL learners; American Councils, 2017) and post-secondary (approximately 1.4 million learners, representing 50% of higher education FL learners; ACTFL, 2018) settings. This fact is not surprising since Spanish is the second most spoken language in the US with approximately 38 million speakers (American Community Survey, 2015). We also find Spanish to be an interesting case since in our experience, many FL Spanish instructors seem to believe that L1 English-speaking learners of Spanish are completely intelligible, and that their intelligibility is not

impacted by pronunciation. By investigating intelligibility, comprehensibility, and accentedness in FL learners of Spanish, such claims can be tested and insights into the generalizability of Munro and Derwing (1995) to new L2s and contexts can be gained.

Overall then, revisiting intelligibility, comprehensibility, and accentedness in FL Spanish has the potential (1) to enhance the validity and generalizability of findings by generating parallel evidence in a new research and learning context and by using more sophisticated statistical techniques, which have become widely available in recent years; (2) to begin laying a methodological and conceptual framework for extending intelligibility research to a greater variety of FLs, including less-commonly-taught languages; (3) to shed light on listeners' perception of L2 Spanish speech, which has practical value for FL Spanish instructors and language program directors.

2. Background

In a series of seminal studies, Munro, Derwing, and colleagues (Derwing & Munro, 1997; Munro & Derwing, 1995; Munro et al., 2006) provided evidence of the partial independence of three dimensions of speech: intelligibility (actual understanding of an utterance), comprehensibility (effort required to understand an utterance), and accentedness (the extent to which pronunciation deviates from an expected pattern/norm). Participants in the 1995 study were advanced ESL learners living in Canada and studying at university, whose speech was elicited via a picture description task. Utterances extracted from their narrations were presented to native speakers of English from the local context, who transcribed them and rated their comprehensibility and accentedness. Results indicated that most utterances were transcribed accurately, comprehensibility ratings were somewhat positively skewed, and accentedness ratings were somewhat negatively skewed. Critical findings from that work included evidence that comprehensibility was more related to intelligibility than accentedness and that even some utterances rated as strongly accented were nevertheless transcribed with perfect accuracy. These results provided empirical evidence that being accented was not synonymous with being difficult to understand, and they laid the foundation for a shift in pronunciation research and teaching away from accent reduction toward a focus on comprehensibility and intelligibility (Levis, 2005). To further explore the relationship among these speech dimensions, the authors conducted additional studies with L2 English learners in Canada focusing on the potential impact of speaker L1 (Derwing & Munro, 1997) and listener L1 (Derwing & Munro, 2013; Foote & Trofimovich, 2018; Munro et al., 2006). As

in the 1995 study, accentedness, intelligibility, and comprehensibility emerged as partially independent speech dimensions.

Another component of the 1995 and 1997 studies was to investigate the extent to which linguistic features (e.g., phonemic errors, grammatical errors, speech rate) were correlated with the global speech dimensions in an effort to better understand which factors might underlie judgements and/or have an impact on intelligibility. Results indicated that linguistic features were more likely to be related to accentedness/comprehensibility ratings than intelligibility scores, but there was a great deal of interlistener variation in the attested relationships. For instance, in the 1995 study, only 28% of listeners showed significant correlations between phonemic errors and intelligibility, versus 44% and 78% for comprehensibility and accentedness, and there were fewer significant correlations across the board in the 1997 study. Subsequent work examining a greater variety of linguistic predictors has shown that pronunciation and lexicogrammatical features contribute to listener judgments in L2 English (e.g., Crowther et al., 2016; Isaacs & Trofimovich, 2012; Saito et al., 2017; Trofimovich & Isaacs, 2012), German (O'Brien, 2014), French (Bergeron & Trofimovich, 2017), and Japanese (Saito & Akiyama, 2017), but these studies have focused exclusively on comprehensibility and accentedness.

A survey of literature dealing with the FL Spanish context indicates an emphasis on accentedness (i.e., goals of sounding nativelike) and a heavy reliance on read speech. One line of inquiry in this area has examined speaker and listener characteristics that affect ratings of foreign accent (e.g., George, 2017; Schoonmaker-Gates, 2015). For example, Schoonmaker-Gates (2015) manipulated the Voice Onset Time (VOT) length of segments in read speech to determine if VOT had an impact on accentedness judgements. Results from her study indicated that both native and non-native speaker listeners are sensitive to VOT as a marker of foreign accent. Another body of work has examined the extent to which phonetics instruction facilitates gains in pronunciation, as determined by listener ratings or through acoustic comparison of learner productions to a native speaker baseline (e.g., Kissling, 2013; Lord, 2005, 2008). In her survey of Spanish FL instructors, Huensch (2019) observed a tension in instructors' responses, insofar as they seemed to prioritize intelligible speech as an important learning goal while also valuing nativelike accuracy (see also, Nagle et al., 2018). On the one hand, an emphasis on accentedness in the literature and in the classroom can be important given that more accented speech may be perceived as less grammatical (Ruivivar & Collins, 2018) and may be associated with negative evaluations of intelligence, successfulness, and other markers of social status (Fuertes et al., 2012). On the other hand, an emphasis on accentedness alone has been questioned, for instance, by Kissling (2013) whose conclusion references Derwing and

Munro's work and asks "whether accentedness is in fact worthy of future study" (p. 737), arguing that "the most interesting research in the future will balance measures of… accentedness, comprehensibility and intelligibility" (p. 737).

A handful of studies have focused on comprehensibility instead of or in addition to accentedness in the Spanish FL context (e.g., McBride, 2015; Nagle, 2018; Schairer, 1992). For example, Schairer (1992) compared comprehensibility ratings to phonetic analysis of speech samples from English L1 learners of Spanish and concluded that learners' productions of vowels (avoiding reduction to schwa and diphthongization of stressed vowels) best predicted comprehensibility scores. More recently, McBride (2015) had listeners rate speech samples for comprehensibility and pleasantness and additionally asked open-ended questions about what made the samples sound accented or difficult to understand. Issues with fluency and intonation surfaced as the features that had the greatest impact on comprehensibility ratings. Ultimately, little to no FL research has focused on intelligibility either independently or in conjunction with comprehensibility and accentedness. Addressing this gap, the following research questions guided the current study:

Research Questions

1. To what extent are intelligibility, comprehensibility, and accentedness related to one another in beginner L2 Spanish speech?
2. To what extent do linguistic features (i.e., phonemic errors, grammatical errors, speech rate) predict the intelligibility, comprehensibility, and accentedness of beginner L2 Spanish speech?

3. Method

3.1 Participants

3.1.1 *Speakers*

Participants ($n = 19$, five men) were recruited from second to fifth semester Spanish courses at a large public university. In their responses to a language background questionnaire (to be made available on IRIS; Marsden et al., 2016) all participants indicated English as their native language, and when asked about their weekly language use, reported using English a majority of the time: 90–100% ($M = 96\%$, $SD = 3\%$). Participants had a mean age of 23 ($SD = 11$, *range* = 18–65) and were majoring in a variety of non-language-related subjects (e.g., Political Science, Biomedical Sciences, Chemistry, Business).

3.1.2 *Listeners*

Following Nagle's (2019) procedure and recommendations, listeners ($n = 30$, 23 men) were recruited from Spain using Amazon Mechanical Turk (AMT). Table 1 provides a summary of listener characteristics based on listeners' responses to a language background questionnaire. Listeners self-assessed their proficiency in English and Spanish using 9-point Likert scales (1 = *extremely low proficiency*, 9 = *extremely high proficiency*). On average, they judged themselves to be highly proficient in Spanish and moderately proficient in English, though they reported minimal English use on a daily basis. They also self-evaluated their level of familiarity with non-native Spanish (1 = *not at all familiar*, 9 = *extremely familiar*), indicating a moderate level of familiarity with non-native speech. They reported interacting with non-native speakers on a monthly or daily basis in both personal and professional contexts. Half had training in linguistics, and a third reported some form of teaching experience. This general listener profile arguably represents the type of listener with whom many FL learners are likely to interact, namely, native listeners who have studied multiple languages and who are reasonably familiar with non-native speech.

Table 1. Summary of listener characteristics

	M (SD)	*Range*
Age	31.63 (8.22)	18–48
Age of onset L2 English	6.67 (2.80)	0–12
Global English proficiency	7.00 (1.36)	4–9
Global Spanish proficiency	8.88 (0.31)	8–9
Percent daily English use	13.87 (13.32)	0–50
Familiarity L2 Spanish	6.33 (2.02)	2–9
Interactions with L2 speakers:	Never: 3 Monthly: 14 Daily: 7 More than daily: 6	
Context of L2 interactions:	Personal: 7 Professional: 7 Both: 14	
Linguistic training:	Yes: 16	
L2 teaching experience:	Yes: 11	

3.2 Materials

Speech data were elicited via a prompted response modeled on the NCSSFL-ACTFL Can-Do Statements: *¿Qué haces en tu tiempo libre?* (What do you do in your free time?). Speaker recordings were transcribed in CLAN following CHAT conventions and checked by the second researcher. Two utterances representing

full phrases minus any initial hesitations such as *uh* were selected from each speaker to be used as stimuli for the AMT rating task, for a total of 38 utterances. Utterances ranged between 4–17 words and 2.14–18.63 seconds with a mean length of 9.47 words (*SD* = 3.90) and 8.26 seconds (*SD* = 4.66). The CLAN transcripts were converted into Praat TextGrids, and segmented TextGrids were used to extract the utterances. The scale peak function in Praat set to 99dB was used to create files of approximately equal loudness for the listening task. Pilot testing with three native speakers indicated that listeners were able to successfully complete the transcription and rating task.

3.3 Procedure

Speaker recording sessions were held individually in a quiet room. After completing the informed consent process, listeners completed a variety of tasks related to a larger project on L2 Spanish learning. For the speaking task used in the current study, participants were instructed to speak for approximately 1 minute in response to the question, *¿qué haces en tu tiempo libre?* Participants were given a few moments to think about their answers before responding. Speakers were compensated with a US$ 20 Amazon gift card.

We used geographic filtering in AMT to recruit online listeners from Spain[1] to transcribe and rate the utterances. After completing a background questionnaire (to be placed on IRIS), listeners were asked to transcribe and rate the 38 utterances presented in a random order while wearing headphones. The task began with instructions and two practice items before continuing to the main task. For each item, listeners pressed play when they were ready to hear the utterance. The task interface required listeners to listen to the complete utterance before having 45 seconds to provide a transcription and their ratings. Listeners were instructed to write down exactly what they heard and then to rate the comprehensibility and accentedness of each utterance using 100-point sliding scales. Figure 1 is an image of the online AMT rating interface. At the end of the experiment, listeners were asked to rate how well they understood the constructs and the difficulty of the task. They also had the opportunity to provide additional open-ended comments

1. AMT allows for geographic filtering by country but not by specific regions within countries. Thus, although we attempted to control for dialect influences on ratings using this filtering option, we would like to acknowledge that there are multiple varieties of Spanish spoken within Spain, which is typically divided into two major dialect zones: north/central and southern. We asked participants to indicate the city in which they had been born. Twenty-two listeners were born in central or northern Spain (e.g., Madrid, Segovia, Valencia), six in southern Spain (e.g., Sevilla, Murcia), one in Caracas, Venezuela, and one in Lisbon, Portugal. Although one listener indicated that he was born in Portugal, he nonetheless reported Spanish as one of his L1s.

on the task and rating interface. Listeners spent an average of 32 minutes on the task ($SD = 7.69$) and were compensated US$ 4 for their participation, in line with the US federal minimum wage at the time of listener recruitment ($ 7.25/hour).

Preliminary inspection of the transcription and rating results indicated that listeners understood the constructs (on a 100-point scale with 100 being "I understood it very well", Accentedness, $M = 91$, $SD = 17$; Comprehensibility, $M = 93$, $SD = 11$) and found the task relatively easy to complete (on a 100-point scale with 100 being "Very easy to complete", $M = 77$, $SD = 23$). The comprehensibility and accentedness data were submitted to reliability analysis using two-way, consistency, average-measure intraclass correlation coefficients (ICC). Results of this analysis indicated excellent reliability for both constructs: for comprehensibility, ICC = .97, 95% CI = [.95, .98] and for accentedness, ICC = .97, 95% CI = [.96, .99].

Figure 1. Amazon Mechanical Turk rating interface

3.4 Analysis

3.4.1 *Data coding*

In line with Munro and Derwing (1995), the 38 utterances used in the listening tasks were coded for phonemic and grammatical errors. Phonemic errors were defined as any deletion, insertion, or substitution of a phoneme clearly interpretable as a Spanish phoneme different from the correct one (e.g., ['ko.ɾo] 'chorus' vs. ['ko.ro] 'I run', ['mi̯a.ɾo] [no translation] vs. ['mi.ɾo] 'I watch"). Errors in word stress placement and inappropriate vowel reduction were also included (e.g., ['me.nəs] vs. ['me.nos], ['be.ɾe] vs. [be.'ɾe]). Grammatical errors (e.g., number, gender, preposition use) in each utterance were also counted (e.g., la_{FEM} $restaurante_{MASC}$, $yo_{1stSING}$ $habla_{3rdSING}$). Phonemic and grammatical errors were coded by a native Spanish research assistant and checked by the first author. Speech rate was operationalized as the number of syllables per second speaking time (i.e.,

excluding pauses; this measure has also been referred to as articulation rate and avoids confounds with measures of pausing [De Jong et al., 2013]). To determine utterance length, syllables were counted manually by two coders based on the audio and transcriptions of the 38 utterances. An inter-rater reliability analysis conducted using two-way, agreement, average-measures ICC on the independent coding from two raters on the 38 utterances was high, ICC = 0.99, 95% CI = [.99, .99]. Utterance duration was calculated automatically from the segmented TextGrids (250ms silent pause cutoff, De Jong & Bosker, 2013) using a Praat script.

Transcriptions provided by the listeners were compared to those created by the authors after careful listening and coded for exact word matches. Misspellings (including lack of accent marks, which some listeners did not use) were not considered deviations. Trivial errors such as phonemic and grammatical regularizations (e.g., *telanovela* transcribed as *telenovela* 'soap opera'; *yo habla* transcribed as *yo hablo*) were also coded. Two coders separately completed the coding for 10 of the 30 listeners ($n = 380$ utterances). Inter-rater reliability (two-way, agreement, average-measure ICC) for the exact match (ICC = .99, 95% CI = [.99, .99]) and trivial error (ICC = .91, 95% CI = [.88, .92]) codings was excellent. Therefore, one coder completed the coding of the transcriptions for the remaining 20 listeners. From the coded transcriptions, an intelligibility score was calculated by summing the exact word matches and trivial errors and dividing by the total number of words.

3.4.2 *Mixed-Effect Models*

Mixed-effects models were fit in R version 3.6.1 (R Core Team, 2019) using the lme4 package (Bates et al., 2014). The following covariates were included in all models to control for their relationship with the dependent variables.

- Speaker-level covariates: age of onset L2 Spanish and amount of time learning L2 Spanish.
- Listener-level covariates: age, age of onset L2 English, self-estimated global proficiency in L2 English, daily English use, familiarity with non-native speech, and previous teaching experience.
- Utterance-level covariates: Number of syllables, mean silent pause duration (computed over the utterance), local speech rate (i.e., articulation rate, computed over the utterance), number of corrections per utterance, and number of repetitions per utterance.

All continuous predictors were z-scored, and for the categorical teaching experience variable, the baseline value was set to zero (i.e., no previous teaching experience). By-speaker and by-listener random effects were fit. All models included random intercepts for both groupings, with additional by-listener random slopes

fit for fixed effects of interest, as described below. Likelihood ratio tests were used to compare models and evaluate fit, and QQ plots were used to check the assumption that model residuals were normally distributed. For intelligibility, we opted to fit models to the more lenient intelligibility metric that did not penalize trivial errors.

4. Results

As displayed in Figure 2, most utterances were transcribed with perfect accuracy, comprehensibility ratings were distributed throughout the 100-point scale, and accentedness ratings were skewed toward moderately to strongly accented.

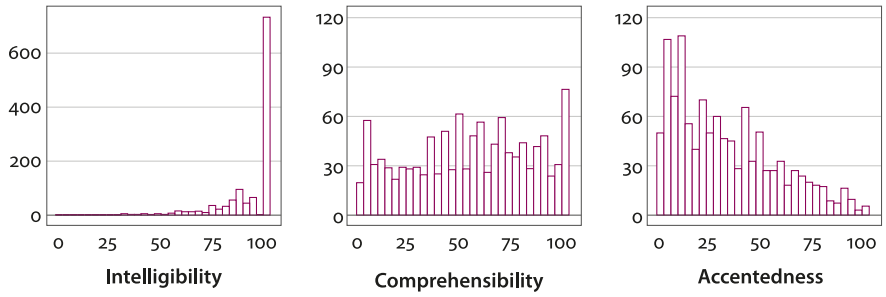

Figure 2. Distribution of intelligibility scores (transformed to a 100-point scale for the sake of display) and comprehensibility and accentedness ratings

4.1 Relationships among intelligibility, comprehensibility, and accentedness

To evaluate the first research question related to relationships among intelligibility, comprehensibility, and accentedness, separate models were fit to the intelligibility and comprehensibility data. Preliminary models fit to the continuous intelligibility variable revealed that residuals significantly deviated from normality. Attempts to bring residuals closer to normality by transforming the data were unsuccessful. Therefore, the continuous measure was recoded into a binary measure where scores < .90 were assigned a value of 0 and scores ≥ .90 were assigned a value of 1. A cutoff of .90 was selected on the basis of previous literature indicating intelligibility rates of .90 to 1 for native speaker utterances. A generalized model, which does not impose the same assumption with respect to normality of model residuals, was then fit to the binary measure.

The primary predictors of interest in this model were the z-transformed comprehensibility and accentedness scores. Generalized models output log-odds,

which can be transformed into odds ratios through exponentiation. On the odds ratio scale, a ratio less than 1 indicates that the predictor reduces the probability of an intelligible transcription, whereas a ratio greater than 1 indicates that the predictor enhances the probability of an intelligible transcription. As reported in Table 2, the association between intelligibility and comprehensibility was statistically significant. Utterances that were rated as more comprehensible were far more likely to be transcribed intelligibly. More precisely, an utterance rated as one unit more comprehensible (1 *SD* above the mean) on the z-scored comprehensibility scale would be 3.29 times more likely to be transcribed intelligibly, an utterance with a comprehensibility score of 2 (2 *SD* above the mean) would be 6.58 times more likely to be transcribed intelligibly, and so forth. In contrast to the significant positive association between intelligibility and comprehensibility, the relationship between intelligibility and accentedness missed significance. A number of covariates, however, emerged as significant predictors. With respect to listener-level covariates, listeners who were older on average and who reported more experience with non-native speech were more likely to transcribe utterances intelligibly. Finally, with respect to utterance-level covariates, utterances containing a greater-than-average number of repetitions and utterances with a greater-than-average silent pause duration were more likely to be intelligible, whereas longer utterances (i.e., utterances containing a greater-than-average number of syllables) were less likely to be intelligible. Including by-listener random slopes for comprehensibility resulted in a singular fit, suggesting overfit. Therefore, the random effect was not retained. By-listener random slopes for accentedness were not tested since the fixed effect missed significance.

Inspection of the comprehensibility model residuals showed a normal distribution. Thus, the comprehensibility models were fit to the original variable on the 100-point scale. This model contained z-scored intelligibility and accentedness predictors and speaker-, listener-, and utterance-level covariates. Including by-listener random slopes for intelligibility and accentedness significantly enhanced model fit ($\chi^2(5) = 39.22$, $p < .001$), suggesting that there was significant between-listener variation in the strength of the association between both predictors and comprehensibility. As shown in Table 3, there were positive relationships between intelligibility and comprehensibility and between accentedness and comprehensibility. Utterances that were more intelligible and less accented – on the 100-point scale with higher scores indicating a more targetlike accent – were also more comprehensible. Contrasting with the intelligibility results showing positive effects for pause length and repetitions, the only significant covariate for comprehensibility was self-corrections. Utterances containing a greater-than-average number of self-corrections were rated as less comprehensible. Regarding the by-listener random effects, there was comparatively more variance in the relationship between

Table 2. Summary of generalized mixed-effects model fit to intelligibility scores

Fixed effects	*Odds ratio*	*95% CI*	*p*
Intercept	4.76	[2.53, 8.96]	<.001
Comprehensibility	3.29	[2.53, 4.27]	<.001
Accentedness	.80	[.62, 1.02]	.07
Speaker-level covariates			
Age of onset L2 Spanish	1.49	[.85, 2.62]	.17
Learning time	1.31	[.74, 2.34]	.36
Listener-level covariates			
Age	1.44	[1.09, 1.89]	<.01
Age of onset L2 English	.88	[.67, 1.17]	.39
L2 English proficiency	.85	[.64, 1.13]	.26
Daily English use	1.02	[.78, 1.33]	.91
Familiarity L2 speech	1.37	[1.04, 1.81]	.03
Teaching experience: Yes	1.24	[.68, 2.26]	.48
Utterance-level covariates			
Speech rate	1.29	[.86, 1.92]	.22
Mean silent pause duration	1.35	[1.04, 1.76]	.02
Number of corrections	1.24	[.88, 1.75]	.23
Number of repetitions	1.48	[1.09, 1.99]	.01
Length (syllables)	.56	[.40, .88]	.001
Random effects			
By-speaker intercept	1.28		
By-listener intercept	.25		

Note. All continuous predictors were transformed into z-scores.

accentedness and comprehensibility than intelligibility and comprehensibility, as evidenced by the greater *SD* for the former (5.60 for accentedness vs. 3.62 for intelligibility).

Residuals for the accentedness models were mostly normal, except at the upper end where they were slightly larger than expected. Despite this minor deviation from normality, the distribution of accentedness model residuals was deemed sufficiently normal to proceed with the linear models on the original 100-point accentedness scale. The effects reported in Table 4 confirm findings documented in the intelligibility and comprehensibility models, namely a marginally significant negative relationship with intelligibility – more intelligible

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

Fixed effects	*Estimate*	*95% CI*	*p*
Intercept	55.90	[47.58, 64.23]	<.001
Intelligibility	6.95	[5.16, 8.75]	<.001
Accentedness	9.30	[6.66, 11.94]	<.001
Speaker-level covariates			
Age of onset L2 Spanish	−.19	[−5.86, 5.48]	.95
Learning time	1.62	[−4.01, 7.25]	.57
Listener-level covariates			
Age	−3.28	[−8.40, 1.84]	.21
Age of onset L2 English	.61	[−4.50, 5.71]	.82
L2 English proficiency	4.04	[−1.43, 9.50]	.15
Daily English use	.35	[−4.78, 5.47]	.90
Familiarity L2 speech	−3.87	[−9.11, 1.38]	.15
Teaching experience: Yes	−5.13	[−16.06, 5.79]	.36
Utterance-level covariates			
Speech rate	.82	[−1.68, 3.32]	.52
Mean silent pause duration	.84	[−.51, 2.18]	.22
Number of corrections	−3.19	[−5.36, −1.01]	.004
Number of repetitions	1.58	[−.39, 3.55]	.12
Length (syllables)	−.09	[−2.41, 2.23]	.94
Random effects	*SD*		
By-speaker intercept	11.85		
By-listener			
Intercept	13.40		
Intelligibility	3.62		
Accentedness	5.60		

Note. All continuous predictors were transformed into z-scores.

utterances were rated as more accented – and a positive relationship with comprehensibility – utterances that were rated as more comprehensible were rated as less accented. With respect to model estimates, there was a far stronger relationship between comprehensibility and accentedness (*estimate* = 8.22) than between intelligibility and accentedness (*estimate* = −1.14). With respect to covariates, utterances spoken at a faster-than-average pace were rated as significantly less accented.

Table 4. Summary of Mixed-Effects Model Fit to Accentedness Ratings

Fixed effects	*Estimate*	*95% CI*	*p*
Intercept	34.56	[27.10, 42.02]	<.001
Intelligibility	−1.14	[−2.21, −.06]	.04
Comprehensibility	8.22	[6.75, 9.70]	<.001
Speaker-level covariates			
Age of onset L2 Spanish	.53	[−3.28, 4.35]	.78
Learning time	1.21	[−2.56, 4.99]	.53
Listener-level covariates			
Age	.63	[−4.93, 6.19]	.82
Age of onset L2 English	.09	[−5.59, 5.78]	.97
L2 English proficiency	−2.98	[−8.81, 2.85]	.32
Daily English use	−.60	[−6.05, 4.86]	.83
Familiarity L2 speech	4.93	[−.72, 10.59]	.09
Teaching experience: Yes	−7.07	[−18.93, 4.80]	.24
Utterance-level covariates			
Speech rate	4.29	[2.16, 6.43]	<.001
Mean silent pause duration	1.04	[−.16, 2.23]	.09
Number of corrections	.69	[−1.22, 2.59]	.48
Number of repetitions	−.86	[−2.59, .88]	.33
Length (syllables)	−.72	[−2.73, 1.28]	.48
Random effects	*SD*		
By-speaker intercept	7.75		
By-listener intercept	13.62		

Note. All continuous predictors were transformed into z-scores.

4.2 Phonemic and grammatical errors

To answer the second research question concerning relationships between phonemic and grammatical errors and intelligibility, comprehensibility, and accentedness, a model was fit to each global speech dimension including the z-scored error variables as predictors as well as their interaction term. Phonemic errors were negatively related to intelligibility (*odds ratio* = .55, 95% CI = [.41, .75], $p < .001$), which shows that utterances containing more errors were less likely to be intelligible. Surprisingly, the relationship between grammatical errors and intelligibility was positive (*odds ratio* = 1.39, 95% CI = [1.07, 1.80], $p = .02$), which would suggest

that utterances containing more errors were *more* likely to be intelligible. Because detailed follow-up analyses suggested that this was in fact not the case, we will not discuss this finding further.[2] The phonemic × grammatical errors interaction term was not significant (*odds ratio* = 1.15, 95% CI = [.71, 1.85], $p = .57$). Including by-listener random slopes for the error terms resulted in a singular fit, so those effects were not retained.

Models fit to the comprehensibility and accentedness data included intelligibility as a covariate, which allowed for the estimation of the phonemic and grammatical error predictors while controlling for the overall intelligibility of the utterance. For comprehensibility, utterances containing more phonemic errors were rated as significantly less comprehensible (*estimate* = −4.45, 95% CI = [−6.65, −2.24], $p < .001$), as were utterances containing more grammatical errors (*estimate* = −3.97, 95% CI = [−5.83, −2.11], $p < .001$). As illustrated by the magnitude of the estimates, phonemic errors had a stronger negative effect on comprehensibility than grammatical errors did. As in the intelligibility model, the interaction term failed to reach significance (*estimate* = .37, 95% CI = [−3.05, 3.79], $p = .83$). Including the error terms as by-listener random effects did not significantly improve model fit (for phonemic errors, $\chi^2(1) = .35$, $p = .55$; for grammatical errors, $\chi^2(2) = .84$, $p = .66$). This suggests that relationships between the error categories and comprehensibility were relatively consistent for the individual listeners sampled in this study. For accentedness, only phonemic errors reached significance (*estimate* = −2.87, 95% CI = [−4.79, −.95], $p = .003$), demonstrating that utterances containing more phonemic errors were rated as more accented. The model containing by-listener random slopes for phonemic errors resulted in a singular fit, so the random effect was not retained.

2. To probe this finding, we fit a zero-one inflated beta regression model. This type of model is advantageous because it fits a separate model to the inflated values at one, which resolves the problematic residuals in the linear model. At the same time, one principal limitation is that this model, as implemented in the glmmTMB package, only accepts one random effect grouping and thus cannot simultaneously estimate the by-speaker and by-listener random effects in the present study. Thus, we fit two models, one with by-speaker random effects and another with by-listener random effects. In both models, all significant effects from the generalized model remained significant, save grammatical errors. In the by-speaker random effect model, grammatical errors was no longer significant (*estimate* = −.02, *SE* = .05, $p = .70$), and in the by-listener model, it remained significant, but the coefficient was negative (*estimate* = −.11, *SE* = .04, $p = .007$), indicating that utterances containing more grammatical errors were less intelligible, as expected.

5. Discussion

5.1 Intelligibility, Comprehensibility, and Accentedness

In the present study, we found a strong, positive association between comprehensibility and intelligibility, a nonsignificant relationship between accentedness and intelligibility, and strong, positive alignment between comprehensibility and accentedness. These results largely fall in line with Munro and Derwing's (1995) original findings, except that whereas they reported a fairly even spread of accentedness scores and comprehensibility scores skewed toward easier to understand, we found the opposite. In our study, accentedness was skewed toward moderately to strongly accented, and comprehensibility scores were distributed throughout the 100-point scale. This difference is likely due to proficiency differences in the two samples: advanced ESL speakers in Munro and Derwing (1995) versus novice to intermediate L2 Spanish learners in our study (see also Derwing & Munro, 1997).

We attempted to test for individual, listener-based variation in relationships among the three constructs through the specification of by-listener random effects. The intelligibility models either did not converge, or they demonstrated a singular fit, which indicates that we were not able to estimate a unique slope for each individual listener in our 30 listener sample. However, the inability to model this variation should not be taken as evidence that it does not exist. In contrast, we were able to incorporate by-listener random slopes for intelligibility and accentedness into the model of comprehensibility. The model-estimated standard deviations for those terms indicated greater variability in the relationship between accentedness and comprehensibility than in the relationship between intelligibility and comprehensibility, reinforcing the view that the latter two constructs are more closely aligned with one another. Thus, in some sense, we were able to replicate using more sophisticated modeling techniques the within-listener correlations that Munro and Derwing (1995) and Derwing and Munro (1997) carried out.

5.2 Phonemic and grammatical errors

Our results diverge somewhat from Munro and Derwing (1995) and Derwing and Munro (1997) with respect to relationships between phonemic and grammatical errors and intelligibility, comprehensibility, and accentedness. Whereas Munro and Derwing (1995) found that most listeners demonstrated significant correlations between both error categories and accentedness, we found no significant relationship between accentedness and grammatical errors. Furthermore,

whereas they found that only about 50% of listeners showed significant correlations between the two types of errors and comprehensibility, we found that both types of errors were associated with lower overall comprehensibility and that incorporating by-listener random effects did not enhance model fit, which would suggest that the effect was relatively uniform across the listeners in our sample. Finally, Munro and Derwing (1995) reported relatively few significant correlations between errors and intelligibility (less than 30% for any error type), but we found that phonemic and grammatical errors showed a strong negative relationship with intelligibility. One possible explanation is proficiency differences between the participants in the current study and those in Munro and Derwing (1995). However, our findings also differ from Derwing and Munro (1997), whose speakers more closely resembled our own participants. Overall, they found fewer significant correlations in the 1997 study, but grammar scores showed the strongest relationship to all three constructs, at least in terms of the number of listeners showing a statistically significant correlation. Again, this contrasts somewhat with our finding that phonemic errors were most consistently associated with the listener-based constructs. A final result worth mentioning is that none of the models showed a significant interaction among phonemic and grammatical errors. We intuitively thought that utterances containing more overall errors and more error types would substantially degrade comprehensibility beyond the effects of the individual error categories. However, in the current study that does not seem to be the case. Thus, the relationship between errors and speech dimensions appears to be additive instead of multiplicative.

5.3 Other factors

One of the strengths of the present approach is that through modeling we were able to account for a wide variety of speaker-, listener-, and utterance-based influences on intelligibility, comprehensibility, and accentedness, while also controlling for correlations among the predictors themselves. Typically, researchers focus on variation in one facet (e.g., speakers or listeners), while limiting variation in the others to mitigate potential confounding factors. Though methodologically sound, the reality of communication is that the intelligibility, comprehensibility, or accentedness of any stretch of speech necessarily arises out of the complex interaction of speaker, listener, and stimulus features. Thus, we opted to embrace all three facets of the data, prioritizing phonemic and grammatical errors as predictors while also investigating speaker- and listener-based background variables and utterance-level properties.

Two listener-level covariates were shown to enhance intelligibility: age and familiarity with L2 speech. The effect of age is somewhat surprising and to our

knowledge has not been attested in the literature. Perhaps older listeners were more attentive during the task and therefore were able to transcribe utterances more accurately. For now, we leave this as an open question for future research. Our finding that listeners who reported more familiarity with L2 Spanish speech tended to transcribe it more accurately but not rate it as more comprehensible or less accented fits with previous research documenting similar effects (e.g., Kennedy & Trofimovich, 2008). Thus, it seems that familiarity with L2 speech may help listeners understand precisely what the speaker is trying to say, but it does not necessarily reduce processing effort or alter listeners' perceptions of the speaker's accent.

Three utterance-level covariates also emerged as significant predictors of intelligibility: silent pauses, repetitions, and length. Silent pauses and repetitions were positively related to intelligibility, whereas utterance length demonstrated a negative relationship. Intuitively, these findings make sense. Longer pauses and repetitions may have helped listeners sort out precisely what the speaker was saying, boosting intelligibility. In contrast, longer utterances were probably more difficult to remember, and as a result, more difficult to transcribe accurately. Although Munro and Derwing (1995) did not find any significant correlations with utterance length, two methodological differences can account for our significant finding. First, whereas Munro and Derwing (1995) carried out separate correlations between utterance length and the listener-based measures, we integrated utterance length into our models alongside an array of other factors, which arguably allowed us to arrive at more reliable estimates of each individual predictor while controlling for the effects of the other predictors in the models. Second, they defined utterance length as number of words, whereas we operationalized it as number of syllables, which resulted in a greater overall range for the predictor.

Relationships between the covariates and comprehensibility and accentedness were far more limited. Corrections seemed to impair comprehensibility, insofar as utterances containing a greater-than-average number of corrections were rated as less comprehensible. Notably, when phonemic and grammatical errors were entered into the comprehensibility model, the effect of corrections was no longer significant, suggesting that errors may have in fact prompted self-corrections, leading to the observed effect. With respect to accentedness, the only significant covariate was speech rate. Derwing and Munro (1997) reported that 23% of listeners showed significant correlations between speech rate and accent ratings. Previous research also suggests that speech is least accented at rates above 4 syllables per second, at least for English (Munro & Derwing, 2001). In the present study, most utterances were spoken at a slower rate of 3.49 syllables per second ($SD = .61$) excluding pauses, or 2.50 syllables per second ($SD = .80$) with

pauses. This could explain why utterances spoken at a faster-than-average pace were rated as less accented in this study.

5.4 Adapting listener-based constructs to a new research context

Working in an ESL context, Munro and Derwing (1995) originally defined intelligibility, comprehensibility, and accentedness in reference to local listeners and local speakers. In other words, the constructs were designed to capture speakers' ability to make themselves understood to a group of listeners with whom they might reasonably interact on a daily basis in their personal and professional lives. Since Munro and Derwing's original work, the constructs have taken on a life of their own and have been applied to different varieties of English (Kang et al., 2018) and different L2s, including German (O'Brien, 2014), French (Bergeron & Trofimovich, 2017), Spanish (Nagle, 2018), and Japanese (Saito & Akiyama, 2016), though most of the L2-other-than-English work has focused on comprehensibility and accentedness. Given how far the constructs have travelled, it seems like the right time to reflect upon any necessary adaptations that might need to take place in order to conduct intelligibility, comprehensibility, and accentedness research in a learning and teaching context that is in many ways radically different from the context in which the constructs were initially defined and measured.

One of the most important issues for FL research is precisely who should evaluate FL learners, since during the first few years of FL study, most, if not all, FL learners will spend a majority of their time interacting with one another and their instructor. In the present study, we opted to recruit online raters from Spain using geographic filtering in AMT. This strategy gave us access to a large pool of potential raters while controlling for some of the variability associated with different dialects of Spanish. Nevertheless, this approach has its limitations. For instance, it is unclear exactly how many participants had been exposed to Peninsular varieties of Spanish, and how many of them would envision themselves interacting with speakers of those varieties in the future. Thus, although the general listener profile could be considered ecologically valid in that many FL learners will likely interact with native listeners who are proficient in multiple languages, somewhat familiar with L2 speech, and interact with L2 speakers in different contexts, there may have been a mismatch between the variety of Spanish that participants had learned and the varieties of Spanish that listeners spoke and with which they were familiar. Due to this potential mismatch, some listeners may have assigned harsher accentedness scores, which could explain why the accentedness data in this study were skewed toward the more accented end of the continuum. A full discussion of methodological choices in rater selection for FL learners is beyond the scope of this paper, but one alternative would be to recruit raters from the dialects to

which learners have been exposed through their instructors and course materials, which would ensure greater parity with respect to the FL varieties that speakers and listeners use.

Despite this limitation, the overall score distributions in the present study suggest that listeners found these FL Spanish speakers to be highly intelligible, moderately comprehensible, and moderately to strongly accented. Consequently, though learners were generally intelligible, they were far from uniformly comprehensible, a finding that calls into question the tacit belief that English-speaking learners of Spanish have few intelligibility and comprehensibility issues and that these issues are not related to pronunciation. In fact, phonemic errors were a far stronger predictor than grammatical errors in all three models. Given these findings, it would be advantageous for future research to continue to investigate the intelligibility, comprehensibility, and accentedness of FL speakers of varying proficiency and in various L2s, adopting broader definitions of intelligibility whenever possible. Ultimately, this research can help bridge the gap between ESL and FL pronunciation research while also providing actionable information that can help FL instructors decide what to prioritize in their courses.

Funding

The work reported here was funded by an ISU Social Sciences Seed Grant to the first author, and by a Language Learning Early Career Research Grant, a USF Creative Scholarship Grant, and a USF Nexus Initiative Award to the second author.

Acknowledgements

We would like to thank the participants and our research assistants, especially Aneesa Ali and Bianca Pinkerton

References

ACTFL. (2018). *Enrollments in languages other than English in United States institutions of higher education, summer 2016 and fall 2016: Preliminary report*. https://www.mla.org/content/download/83540/2197676/2016-Enrollments-Short-Report.pdf

American Community Survey. (2015). *Detailed languages spoken at home and ability to speak English for the population 5 years and over: 2009–2013*. https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html

American Councils for International Education. (2017). *The National K-12 Foreign Language Enrollment Survey Report*. https://www.americancouncils.org/sites/default/files/FLE-report-June17.pdf

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). _lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1.-7. CRAN.R-project.org/package=lme4

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, *50*, 547–566. https://doi.org/10.1111/flan.12285

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility. *Journal of Second Language Pronunciation*, *2*, 160–182. https://doi.org/10.1075/jslp.2.2.02cro

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, *40*, 443–457. https://doi.org/10.1017/S027226311700016X

De Jong, N.H., & Bosker, H.R. (2013). Choosing a threshold for silent pauses to measure second language fluency. Paper presented at DiSS, Stockholm.

Derwing, T.M., & Munro, M.J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*, 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T.M., & Munro, M.J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*(2), 163–185. https://doi.org/10.1111/lang.12000

Foote, J.A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, *74*, 253–278. https://doi.org/10.3138/cmlr.2017-0011

Fuertes, J.N., Gottdiener, W.H., Martin, H., Gilbert, T.C., & Giles, H. (2012). A metaanalysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology*, *42*, 120–133. https://doi.org/10.1002/ejsp.862

George, A. (2017). Effects of listener and speaker characteristics on foreign accent in L2 Spanish. *JSMULA*, *5*, 127–148.

Huensch, A. (2019). Pronunciation in foreign language classrooms: Instructors' training, classroom practices, and beliefs. *Language Teaching Research*, *23*, 745–764. https://doi.org/10.1177/1362168818767182

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*, 475–505. https://doi.org/10.1017/S0272263112000150

Kang, O., Thomson, R.I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension: Measuring intelligibility in varieties of English. *Language Learning*, *68*, 115–146. https://doi.org/10.1111/lang.12270

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*, 459–489. https://doi.org/10.3138/cmlr.64.3.459

Kissling, E.M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal*, *97*(3), 720–744. https://doi.org/10.1111/j.1540-4781.2013.12029.x

Levis, J.M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly 39*, 369–377. https://www.jstor.org/stable/3588485

Lord, G. (2005). (How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania*, *88*, 557–567. https://doi.org/10.2307/20063159

Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals*, *41*, 365–379. https://doi.org/10.1111/j.1944-9720.2008.tb03297.x

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). Routledge.

McBride, K. (2015). Which features of Spanish learners' pronunciation most impact listener evaluations? *Hispania*, *98*, 14–30. https://doi.org/10.1353/hpn.2015.0001

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*, 451–468. https://doi.org/10.1017/S0272263101004016

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*, 111–131. https://doi.org/10.1017/S0272263106060049

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal*, *102*, 199–217. https://doi.org/10.1111/modl.12461

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation*, *5*(2), 294–323. https://doi.org/10.1075/jslp.18016.nag

Nagle, C., Sachs, R., & Zárate-Sández, G. (2018). Exploring the intersection between teachers' beliefs and research findings in pronunciation instruction. *The Modern Language Journal*, *102*(3), 512–532. https://doi.org/10.1111/modl.12493

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech: L2 learner assessments. *Language Learning*, *64*, 715–748. https://doi.org/10.1111/lang.12082

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org/

Ruivivar, J., & Collins, L. (2018). Nonnative accent and the perceived grammaticality of spoken grammar forms. *Journal of Second Language Pronunciation*, *5*(2), 269–293. https://doi.org/10.1075/jslp.17039.rui

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, *3*, 199–217. https://doi.org/10.1075/jslp.3.2.02sai

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*, 439–462. https://doi.org/10.1093/applin/amv047

Schairer, K. E. (1992). Native speaker reaction to non-native speech. *The Modern Language Journal*, *76*, 309–319. https://doi.org/10.1111/j.1540-4781.1992.tb07001.x

Schoonmaker-Gates, E. (2015). On voice-onset time as a cue to foreign accent in Spanish: Native and nonnative perceptions. *Hispania*, *98*, 779–791. https://doi.org/10.1353/hpn.2015.0110

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

Address for correspondence

Charles L. Nagle
Iowa State University
Department of World Languages and Cultures
505 Morrill Road
3102G Pearson Hall
Ames, IA 50011
USA

cnagle@iastate.edu
https://orcid.org/0000-0003-2712-2705

Co-author information

Amanda Huensch
University of Pittsburgh
Department of Linguistics

amanda.huensch@pitt.edu
https://orcid.org/0000-0001-6229-3521

Publication history

Date received: 10 February 2020
Date accepted: 28 May 2020
Published online: 7 July 2020