Evaluating the Impacts of Accent and Semantic Context on Listening Effort

by

Avanti Khare

****

Submitted in partial fulfillment

of the requirements for

Honors in the Department of Neuroscience

UNION COLLEGE

June, 2023

ABSTRACT

KHARE, AVANTI Evaluating the impacts of accent and semantic context on listening effort. Department of Neuroscience, June 2023.

ADVISOR: Chad Rogers

Nonnative accented speech is associated with increased listening effort for English-monolingual listeners, even if the speech signal is intelligible. Semantic context is a global characteristic of English phrases that quantifies the degree to which words communicate a cohesive idea. Previous research suggests that semantic context may be used as a helping factor during speech perception in adverse conditions. The current work examines the relationship between speaker accent and semantic context using global semantic anomalies. Participants performed a randomly prompted recall task during lists of varying semantic context levels recorded by native and nonnative-accented speakers. Results are discussed in terms of two frameworks for understanding listening effort and speech processing: the Ease of Language Understanding Model and the Effortfulness Hypothesis, which differ in their prediction of an interaction between speaker accent and semantic context. A significant interaction was found between Speaker Accent and Context Group, lending support for the Effortfulness Hypothesis.

Evaluating the impacts of accent and semantic context on listening effort

In my sophomore year of college, I completed the most difficult chemistry course that I would ever have to take. On one particularly cold winter day, the heating system in the classroom suddenly turned on halfway through the lecture and fully muffled the professor's voice from where I sat. I only realized later in the day when I tried to complete the accompanying practice problems that I could not remember any of the content from the lecture, likely because I had strained so hard to hear the professor in the first place.

Speech perception often takes place in adverse listening conditions that interfere with the speech signal (Mattys et al., 2012). An adverse condition (Mattys et al., 2012) is any manipulation of the speech signal associated with increased listening demand and listening effort when compared with speech perception in optimal settings (e.g., no environmental noise, healthy native-accented speaker). Listening effort refers to the amount of energy that a listener actually expends during the cognitive tasks involved for speech perception in adverse conditions, independent of the actual listening demand required to process the target speech signal (Pichora-Fuller et al., 2016). In clear speech, speech perception is carried out through the listener's ability to match segments of the incoming speech signal to internal representations of lexical items, which becomes disrupted during speech perception in adverse conditions (Peelle, 2018).

Adverse conditions for speech perception can be divided into source degradations, which are adverse conditions that are intrinsic to the speaker, and environmental degradations, which are adverse conditions that are intrinsic to the environment (Mattys et al., 2012). Nonnative accent is an example of an adverse condition that is considered a source degradation (Mattys et al., 2012). Nonnative-accented speech is characterized by the presence of systematic deviations from native language norms (Van Engen & Peelle, 2014). Previous research has shown that the

behavioral consequences of listening to nonnative-accented speech mirror the behavioral consequences studied when listeners encounter other adverse conditions for speech perception, such as reduced intelligibility (Gass & Varonis, 1984; Munro & Derwing, 1995; Bent & Bradlow, 2003; Burda et al., 2003; Ferguson et al., 2010; Gordon-Salant et al., 2010a,b), comprehensibility (Anderson-Hsieh & Koehler, 1988; Major et al., 2002), and processing speed (Munro & Derwing, 1995; Floccia et al., 2009).

More broadly, speech perception in adverse conditions has been linked to decreased performance on tasks of downstream processes such as memory, suggesting that the detrimental effect of increased listening demand is at the level of perception (Peelle, 2018). The downstream effect of increased listening demand on memory was demonstrated in two experiments by Rabbitt (1968) where participants were asked to recall strings of digits that were either noise-masked or presented in the clear. Rabbitt (1968) showed poorer recall for the strings of noise-masked digits, even in masking conditions where participants were able to immediately recall items after presentation. In a second experiment, participants were presented with a spoken eight-digit list, where half of the list was masked in noise and the other half of the list was presented in quiet and found that participants showed poorer recall for the first half of the list when the second half of the list was presented in noise compared to quiet. Rabbitt (1968) concluded that the cognitive resources that were diverted to speech perception in the noise-masked condition were diverted at the cost of cognitive resources needed to encode the previously encountered information in long-term memory.

Semantic context is a helping factor for speech perception in adverse conditions that listeners use to decrease listening effort (McCoy et al., 2005). Semantic context can be defined as the degree to which the words in a given phrase are dependent on the words that come before

them (Miller & Selfridge, 1950). Semantic context can be manipulated to varying degrees, ranging from manipulating the last word of a sentence (e.g., "He mailed the letter without a stamp" is considered a semantically congruent phrase, while "He mailed the letter without a lamp" is considered a semantically anomalous phrase) to the construction of a more global sentence-wide semantic anomaly. Statistical approximations to English are an experimental construct developed by Miller and Selfridge (1950) that simulates the semantic context structure of written English, but remain semantically anomalous. These statistical approximations to English range on a scale of zeroth to ninth order of approximation where the zeroth order of approximation represents the lowest degree of contextual constraint, while the ninth order of approximation represents excerpts of standard written English. The zeroth order of approximation provided listeners with no semantic context structure to predict the next word in the phrase, while the ninth order of approximation provided listeners with all of the semantic context that is provided in standard written English. The first, second, third, fourth, fifth, and seventh orders of approximation represent degrees of contextual constraint that lie between the zeroth and ninth orders of approximation.

The zeroth order of approximation was developed using a random selection of the most common words present in the English language based on a dictionary developed by Thorndike and Lorge (1944). The first order of approximation was developed by scrambling words taken from word lists at higher orders of approximation. The second, third, fourth, fifth, and seventh orders of approximation were developed using a norming study that employed a method similar to the children's game "telephone". Participants were presented with a starting phrase containing the same number of words as the degree of contextual constraint (e.g., three starting words for a third order approximation to English). The participant completed a sentence using the provided

starting phrase. The first word after the starting phrase was saved and the rest of the sentence was dropped, including the first word of the starting phrase. This procedure was repeated with other normative participants until a string of the desired length was obtained. The resulting phrase shows a level of semantic context that varies with the number of starting words used in the phrase construction because each successive word is determined only by the context of the preceding word. Miller and Selfridge (1950) held that this method approximated language to the extent that language is associative, with greater preceding semantic context generating a strong approximation to standard English. The ninth order of approximation was developed by taking direct excerpts from current fiction or biography.

Using Miller and Selfridge (1950)'s materials, McCoy et al., (2005) examined whether increased approximations to language (i.e., increased semantic context) reduced the impact of adverse conditions on listening in adults with and without age-related hearing loss. Participants included a young adult group, an older adult group with normal hearing for their age and an older adult group with age-related hearing loss. All participants listened to continuous strings representing all zeroth through ninth order statistical approximations to English (Miller & Selfridge, 1950). While listening, participants performed a running recall task where the audio was stopped after a random interval of words, and participants recalled the last three words they heard, starting with the most recently presented word.

McCoy et al. (2005) found that performance in the high context conditions (second through ninth orders of approximation) was similar between both the better hearing and hearing loss groups, while the hearing loss group performed significantly worse on the running recall task in the low context conditions (zeroth and first orders of approximation). McCoy et al. (2005) concluded that the added perceptual effort required for successful recall by participants with

hearing loss was sufficient to affect memory performance, especially in the low context condition. Additionally, the higher level of semantic context provided by the higher orders of approximation may have operated to facilitate the recognition of the target words either by increasing their likelihood or by decreasing the number of potential lexical possibilities as the words were being heard. The progressive reduction of semantic context in the second through seventh orders of approximation and elimination of the zeroth and first orders of approximation as a helping factor for speech perception may have contributed to the significant difference in recall performance between the better hearing and hearing loss groups in the low context condition.

McCoy et al. (2005) defined the effortfulness hypothesis, which states that the extra effort that a hearing-impaired listener must expend to achieve perceptual success of degraded speech may come at the cost of processing resources that might otherwise be available for encoding the content of the speech signal in long-term memory. The Ease of Language Understanding model developed by Rönnberg et al. (2008; 2013) could also potentially explain the increase in listening effort associated with speech containing low semantic context. The Ease of Language Understanding (ELU) model describes the relationship between working memory, defined as a limited capacity system for storing and processing information, and the conditions in which people understand language in the presence of signal distortions (Rönnberg et al., 2013). The ELU model assumes that an incoming speech signal is first bound to an internal representation of a word in an episodic buffer. If the speech signal matches an internal representation in semantic long-term memory, then the person is able to successfully comprehend the speech signal. If there is not a close enough match between the speech signal and an internal representation, explicit working memory processes are invoked in order to reach

successful comprehension, including the processing of semantic context. The ELU model predicts that signal distortion of the speech signal will increase listening effort, but it does not make any prediction about changes in memory for the speech signal relating to signal distortion. Additionally, the ELU model does not make any predictions about the potential role of semantic context in immediate speech perception, as only single words are evaluated against internal lexical representations.

Figure 1 shows a hypothesized pattern of recall accuracy scores for the native-accented speaker and nonnative-accented speakers that would be in support of the effortfulness hypothesis (McCoy et al., 2005). If the effortfulness hypothesis (McCoy et al., 2005) supports the pattern of results found in the running recall experiment, an interaction between speaker accent and approximation level is expected where the highest difference in recall accuracy between the native and nonnative accented stimuli will be found in the lowest context condition and the lowest difference in will be found in the highest context condition. Additionally, the interaction between speaker accent and context level is expected to be more pronounced in the recall prompt for the third-to-last word than the recall prompt for the most recently presented word.

Figure 2 shows a hypothesized pattern of recall accuracy scores for the native-accented speaker and nonnative-accented speakers that would be in support of the ELU (Rönnberg et al. 2008; 2013). If the ELU (Rönnberg et al. 2008; 2013) more accurately explains the pattern of results found in the running recall experiment than the effortfulness hypothesis, a main effect of accent on recall showing a native-accent advantage for recall accuracy over the nonnative-accents is expected but not a main effect of context on recall. Because nonnative accent is classified as an adverse condition for speech perception (Mattys et al., 2012), and McCoy et al. (2005) have demonstrated that the combination of an adverse condition (age-related hearing loss)

and reduced semantic context negatively impacts running recall performance, we would expect to observe a similar pattern of results as the results found by McCoy et al. (2005) when substituting hearing loss with accent. We hypothesize that the running recall experiment will support the effortfulness hypothesis (McCoy et al., 2005) through an interaction between speaker accent and approximation level such that the greatest differences between accuracy for native and nonnative-accented speech will be present at the smallest context level and the least difference in accuracy for native and nonnative-accented speech will be present at the largest context level.

It remains an open question as to how the processing challenges associated with comprehending both a nonnative accent and semantically anomalous sentence construction would impact the performance of downstream cognitive processes such as memory. The present experiment followed a 2 (Speaker Accent: Native, Nonnative) × 3 (Recall Prompt: 1, 2, 3) × 4 (Context Group: XS, S, M, L) within-subjects design. The current study followed the experimental procedure used by McCoy et al. (2005) and original stimulus sets using the statistical approximations to English included in Miller and Selfridge (1950) spoken by a native English-accented speaker, a Hindi-accented speaker, and a Chinese-accented speaker. The native and nonnative-accented stimulus sets were tested for intelligibility and judgment of speaker accentedness during a prior norming experiment.

Experiment 1

Experiment 1 was a norming experiment conducted to determine word-level intelligibility scores for the words from the selected lists from Miller & Selfridge (1950) for each of the three speakers: a Native English-accented speaker, a Hindi-accented speaker, and a Chinese-accented speaker. In order to ensure that English-monolingual listeners perceived the accents of these

speakers faithfully, and to obtain estimates of word-level intelligibility, we first performed a norming study to validate the set of recordings.

Method

Participants

Thirty English-speaking participants were recruited using the online participant recruitment platform Prolific (www.prolific.co) from a population of English monolingual users based in the United States. Participants were compensated monetarily at a rate of $12 per hour.

Stimuli

The stimuli consisted of 208 unique words taken from Miller and Selfridge (1950) across all orders of approximation (0-9). All stimuli were recorded by a female native English-accented speaker, a female Hind-accented speaker, and a female Chinese-accented speaker using Audacity (Audacity, Seattle, WA). Stimuli were leveled for sound intensity using Adobe Audition (Adobe Inc., San Jose, CA).

Procedure

All testing was completed remotely and asynchronously using the Gorilla Experiment Builder (www.gorilla.sc).

Participants completed a pre-experimental survey containing questions about the listeners' auditory environment and prior language experience. Participants were asked to wear headphones, maximize the window in which they were completing the experiment, and turn off all competing auditory stimuli (e.g., music, television). Participants then self-reported whether they had actually performed the previously described actions to ensure that they were prepared for participation. The pre-experiment questionnaire also contained questions about the participant's language experience, adapted from the Language Experience and Proficiency

Questionnaire (LEAP-Q; Marian et al., 2007) including self-identification as monolingual or bilingual, identification of all acquired languages and their dominance, self-reported auditory and written exposure to languages other than English, and self-reported speaking dominance for languages other than English. Answers to these questions were used as a manipulation check to confirm that all participants were monolingual English speakers.

Participants' main task was to identify single words presented in one of the three speaker accents. Participants listened to a target word after a fixation cross was presented for 100 milliseconds. Participants recorded the most recently presented word immediately after its presentation. Participants rated their confidence in the accuracy of their response after every word presented on a scale of 0-100 with zero representing that the participant was not confident at all that the answer that they reported was the word they heard and 100 representing that the participant was completely confident that the word they reported was the word that they heard. At the end of each block of recordings, participants rated the accentedness of the speaker on a scale of 0-100, with zero representing that the speaker had a heavy accent and 100 representing that the speaker sounded like a native English speaker.

Participants completed a brief practice phase to familiarize themselves with the experiment and reporting procedures prior to completing the experimental trials. The practice trial list consisted of five words recorded by the same female native English speaker as that in the native English-accented experimental stimuli. Participants provided recall and confidence measures after every presented word as well as an accentedness rating after the presentation of the whole list. Participants were not exposed to stimuli recorded by the Hindi-accented speaker or the Chinese-accented speaker during the practice phase.

Participants heard the stimulus recordings blocked by speaker, with two blocks of 69 words per speaker and one block of 70 words per speaker. The order of presentation of each block was counterbalanced across participants using a Balanced Latin Square.

Results

All scores provided by two participants were excluded from the final analysis, one who self-reported as bilingual in the pre-experimental questionnaire and a second participant who indicated that they have routine exposure to languages other than English in the pre-experimental questionnaire.

Figure 3 shows the mean intelligibility across the Native English-accented, Chinese-accented, and Hindi-accented speakers. Overall, the Native English-accented speaker received a higher mean intelligibility rating than both the Chinese-accented and Hindi-accented speakers. To confirm this pattern of data, intelligibility scores for each speaker averaged across subjects were submitted to a one-way repeated measures ANOVA which revealed a significant effect of Speaker Accent, $F(2, 54) = 40.82$, $p<.001$, $\eta^2_p = .60$. Post-hoc t-tests applying the Bonferroni correction for familywise error rate revealed that trials with the native-English speaker ($M = .88$, $SD = .07$) had significantly higher recall accuracy than trials with the Hindi-accented speaker ($M = .77$, $SD = .07$), $t(26) = 5.67$, $p<.001$, and the Chinese-accented speaker ($M = .70$, $SD = .10$), $t(26) = 8.93$, $p<.001$. Trials with the Hindi-accented speaker had significantly higher accuracy than trials with the Chinese-accented speaker, $t(26) = 3.26$, $p= .006$.

Figure 4 shows the mean confidence rating for correctly recalled trials across the three speaker accents. The Native English-accented speaker received a greater mean confidence rating for correct trials than both the Chinese-accented and Hindi-accented speakers. To confirm this pattern, average confidence scores over correct trials averaged across subjects were submitted to

a one-way repeated measures ANOVA which revealed a significant effect of Speaker Accent, $F(2, 54) = 35.03$, $p<.001$, $\eta^2_p = .54$. Post-hoc t-tests applying the Bonferroni correction for familywise error rate revealed that trials with the native-English speaker ($M = 94.64$, $SD = 6.20$) had significantly higher correct trial confidence scores than trials with the Hindi-accented speaker ($M = 88.40$, $SD = 9.48$), $t(26) = 5.87$, $p<.001$, and the Chinese-accented speaker ($M = 86.03$, $SD = 9.36$), $t(26) = 8.10$, $p<.001$. Trials with the Hindi-accented speaker did not have significantly higher correct trial confidence scores than trials with the Chinese-accented speaker.

Figure 5 shows the mean accentedness rating for each of the three speakers, provided after each block of trials. The Chinese-accented and Hindi-accented speakers received higher accentedness scores compared to the Native English-accented speaker. To confirm this relationship, average accentedness scores averaged across subjects were submitted to a one-way repeated measures ANOVA which revealed a significant effect of Speaker Accent, $F(1.36, 36.73) = 32.01$, $p<.001$, $\eta^2_p = .54$. Post-hoc t-tests applying the Bonferroni correction for familywise error rate revealed that trials with the native-English speaker ($M = 23.36$, $SD = 37.48$) had significantly higher recall accuracy than trials with the Hindi-accented speaker ($M = 70.21$, $SD = 21.59$), $t(26) = -6.80$, $p<.001$, and the Chinese-accented speaker ($M = 71.89$, $SD = 21.40$), $t(26) = -7.05$, $p<.001$. Trials with the Chinese-accented speaker did not have significantly higher overall confidence scores than trials with the Hindi-accented speaker.

Discussion

In summary, recall accuracy for the native English-accented speaker was significantly higher than recall accuracy for the Hindi-accented and Chinese-accented speakers. Confidence scores for correct trials were significantly higher for the native English-accented speaker than for the Hindi-accented and Chinese-accented speakers. Additionally, accentedness ratings were

significantly lower for the native English-accented speaker than accentedness ratings for the Hindi-accented and Chinese-accented speakers. The previously described results confirm the adverse impact of nonnative accent on speaker intelligibility (Van Engen & Peelle, 2014; Gass and Varonis, 1984; Munro and Derwing, 1995; Bent and Bradlow, 2003; Burda et al., 2003; Ferguson et al., 2010; Gordon-Salant et al., 2010a,b). Most crucially, the results of Experiment 1 confirm that the actual speaker characteristics for both the native and nonnative-accented speakers matched the predicted characteristics, particularly accentedness rating. The average word-level intelligibility scores calculated from the target words in Experiment 1 were used as a covariate to assess the impact of word-level intelligibility on recall accuracy in Experiment 2.

Experiment 2

Following the successful validation of the native-accented and nonnative-accented stimuli developed for Experiment 1, Experiment 2 aimed to assess the potential downstream effects of accentedness and semantic context on memory. Additionally, Experiment 2 tested whether the Effortfulness Hypothesis (McCoy et al., 2005) or the ELU model (Rönnberg et al., 2008; 2013) more accurately explained the pattern of results under speaker accent, an adverse condition that is intrinsic to the speaker. Methods, stimuli, and hypothesized results were pre-registered with the Open Science Framework and can be accessed at https://osf.io/b5rzp/.

Method

Participants

Forty-eight online participants were sourced through the online participant recruitment platform Prolific (www.prolific.co). Participants were screened according to the following criteria: Age (18-30 years), nationality, place of most time spent before turning 18, English-speaking monolingual status, self-identification as neurodiverse, history of dyslexia, history of

head injury, history of mild cognitive impairment, history of hearing difficulties, and history of cochlear implant.

Stimuli and Materials

The stimuli consisted of 16 strings, each 15 words in length, taken from Miller and Selfridge (1950). All strings combined represented four general degrees of contextual constraint from smallest to largest order of approximation, with two lists each of zeroth and first order of approximation (extra small context group; XS), two lists each of second and third order of approximation (small context group; S), two lists each of fourth and fifth order of approximation (medium context group; M), and two lists each of seventh and ninth order approximations to English (large context group; L).

All stimuli were recorded by a female Chinese-accented speaker, a female Hindi-accented speaker, and a female native English-accented speaker using Audacity (Audacity, Seattle, WA). Stimuli were leveled for sound intensity using Adobe Audition (Adobe Inc., San Jose, CA). All recorded stimuli were the same as the recordings used in Experiment 1.

Procedure

Participants completed the same listening environment questionnaire and language history questionnaire as the questionnaires used in Experiment 1. All testing was completed remotely and asynchronously using the Gorilla Experiment Builder (www.gorilla.sc).

Figure 6 summarizes the procedure for each trial of the running recall experiment. Participants were instructed to listen carefully to each word list as it was presented and to be prepared for recall at any moment. The presentation of words in each list was randomly stopped for recall after the passage of 5, 7, 8, 12, or 15 words, which was referred to as lag. When prompted by the appearance of three asterisks on their computer screen, participants recalled the

last three words that were presented. This reporting procedure was the same for each trial. Each participant was presented all 16 word lists with the speaker accent and order of presentation of the word lists counterbalanced between participants. Lag for each word list was randomized between participants.

Results

All scores provided by two participants were excluded from the final analysis, one who self-reported as bilingual in the pre-experiment questionnaire and one who did not self-report as bilingual but indicated an ability to speak more than one language in the following questionnaire items, which gave a remaining number of 46 participants.

Pre-Registered Analyses

Accuracy at Recall Prompt 1, Recall Prompt 2, and Recall Prompt 3 across all context conditions and speaker accents is given in Figure 7. The greatest difference in recall accuracy across all context and speaker conditions was found between Recall Prompt 1 and Recall Prompt 3 (Figure 7). Within each recall prompt, the greatest difference in recall accuracy between the native and nonnative-accented speakers was found at the XS context condition (Figure 7).

To confirm the statistical reliability of these results, recall accuracy scores were submitted to a 2 (Speaker: Native, Nonnative) × 4 (Context Group: XS, S, M, L) × 3 (Recall Prompt: 1, 2, 3) repeated-measures ANOVA conducted using JASP (JASP Team, 2023). This revealed a main effect of Speaker Accent (Native, Nonnative), $F(1, 45)= 7.88$, $p= .007$, $\eta^2_p = 0.15$. Post-hoc t-tests applying the Bonferroni correction revealed increased performance on Native-English accented ($M = 0.87$, $SE = 0.02$) trials over Nonnative-accented trials ($M = .82$, $SE = 0.02$), $t(45) = 2.81$, $p < .001$. Secondly, the repeated measures ANOVA revealed a significant main effect of Recall Prompt (1, 2, 3), $F(1.49, 45)= 6.67$, $p= .005$, $\eta^2_p = 0.13$. Post-hoc t-tests

applying the Bonferroni correction revealed significantly higher accuracy on Recall Prompt 1 (M = 0.89, SE = 0.02) compared to Recall Prompt 3, (M = 0.81, SE = 0.03), $t(45) = 3.66$, $p<.001$). Lastly, a main effect of Context Level (XS, S, M, L) was significant, $F(2.28, 45)= 30.11$, $p<.001$, $\eta^2_p = 0.40$. Post-hoc t-tests applying the Bonferroni correction revealed increased recall accuracy across all three recall prompts in the Small (M = 0.86, SE = 0.02), Medium (M = 0.90, SE = 0.02), and Large (M = 0.89, SE = 0.02) conditions when compared to the Extra Small condition, (M = 0.73, SD = 0.03), all t's > 8.93, $p<.001$.

The three-way repeated measures ANOVA revealed a significant Accent × Context interaction $F(2.72, 45)= 4.84$, $p= .004$, $\eta^2_p = 0.10$. Post-hoc t-tests applying the Bonferroni correction revealed significantly lower performance in the Nonnative XS (M = 0.66, SE = 0.04) condition compared with the Native XS (M = 0.80, SE = 0.03), Native S (M = 0.89, SE = 0.03), Native M (M = 0.93, SE = 0.03), Native L (M = 0.88, SE = 0.02), Nonnative S (M = 0.83, SE = 0.03), Nonnative M (M = 0.88, SE = 0.03), and Nonnative L (M = 0.91, SE = 0.02) conditions, all t's < 0.14, $p<.001$. Recall performance in the Native XS condition was lower compared to the Native M (M = 0.93, SE = 0.03) and Nonnative L (M = 0.91, SE = 0.02) conditions, all t's < -3.36, $p<.026$.

A significant Recall Prompt × Context interaction was revealed by ANOVA, $F(4.21, 45) = 2.90$, $p= .021$, $\eta^2_p = 0.06$. Post-hoc t-tests applying the Bonferroni correction revealed significantly lower performance at Recall Prompt 1 XS (M = 0.77, SE = 0.03) compared to Recall Prompt 1 S (M = 0.924, SE = 0.02), Recall Prompt 1 M (M = 0.94, SE = 0.01), Recall Prompt 1 L (M = 0.91, SE = 0.03), and Recall Prompt 2 L all t's < -3.78, $p<.012$. Recall Performance at Recall Prompt 1 S was significantly higher than recall performance at Recall Prompt 2 S (M = 0.80, SE = 0.03), $t(45) = 3.40$, $p<.05$. Recall Prompt 1 refers to the most

recently presented word, which is held in working memory, according to the Effortfulness Hypothesis (McCoy et al., 2005) and the ELU (Rönnberg et al. 2013). Therefore, accuracy at Recall Prompt 1 was used to measure intelligibility for all preceding words in the list.

Recall Prompt 2 and Recall Prompt 3 refer to items that have been encoded into long-term memory that may reveal a downstream effect of speaker accent and semantic context on recall. Lower recall performance was observed at Recall Prompt 2 XS (M = 0.77, SE = 0.03) compared to Recall Prompt 1 S (M = 0.92, SE = 0.02), Recall Prompt 1 M (M = 0.94, SE = 0.20), Recall Prompt 2 M (M = 0.92, SE = 0.03), Recall Prompt 1 L (M = 0.91, SE = 0.03) and Recall Prompt 2 L (M = 0.91, SE = 0.03), all t's < 3.93, p<.007. Recall Performance at Recall Prompt 3 XS was significantly lower than recall accuracy at Recall Prompt 1 S, Recall Prompt 2 S (M = 0.80, SE = 0.03), Recall Prompt 3 S (M = 0.86, SE = 0.03), Recall Prompt 1 M, Recall Prompt 2 M, Recall Prompt 3 M (M = 0.85, SE = 0.04), Recall Prompt 1 L, Recall Prompt 2 L, Recall Prompt 3 L (M = 0.86, SE = 0.03), all t's < -3.93, p<.007. Performance at Recall Prompt 2 S was significantly lower than performance at Recall Prompt 1 M and Recall Prompt 2 M, all t's < -3.44, p<.042.

Accuracy as a Function of Word-Level Intelligibility and Word Frequency

We attempted to fit a linear mixed effect model to the data collected in the running recall experiment to perform a more focused examination of the impacts of nonnative accent on listening effort while statistically controlling for pre-experimental differences in speaker intelligibility. Linear mixed effect modeling was conducted using the *lme4* (Bates et al., 2015), *lmerTest* (Kuznetsova et al., 2017), and *afex* (Singmann et al., 2015) packages in R (R Core Team, 2023). The following section describes a directed assessment of the relationship between recording intelligibility, word frequency, and recall accuracy using linear mixed effect modeling.

Word-level intelligibility scores, based on mean intelligibility of all recordings presented in Experiment 1 were assigned to each target word in Experiment 2 at Recall Prompt 1, Recall Prompt 2, and Recall Prompt 3 for further analysis. Due to experimenter error, 21 stimulus recordings did not receive word-level intelligibility scores during Experiment 1. The intelligibility scores assigned to these recordings were hypothesized intelligibility scores computed by averaging the intelligibility ratings for all other words in the list intelligibility rating over order of approximation for each of the three speakers. Values for word frequency, number of phonemes, number of syllables, and orthographic length were assigned to each target word to be used as covariates in future analysis. Scores for all word-level covariates for each target word except word-level intelligibility were obtained using the English Lexicon Project database (Balota et al., 2007).

Figure 8 shows the relationship between word-level intelligibility and recall accuracy across all recall prompts for each speaker accent and context group condition. The potential effect of word-level intelligibility scores and word frequency on recall accuracy for target words was used to fit a linear mixed effect model to the present data. Context group, speaker accent, recall order, intelligibility, word frequency, and accuracy scores across Recall Prompt 1, Recall Prompt 2, and Recall Prompt 3 were submitted to a linear mixed effect model. A likelihood-ratio test indicated that the model including both intelligibility and word frequency was a better fit for the data than the model that did not include them, $\chi^2(2) = 167.24$, $p < .001$. Additionally, Figure 8 shows a significant positive correlation between intelligibility and recall accuracy using Pearson's r for the native-accented speaker in the XS context group, $r(44) = 0.54$, $p < .001$, and the S context group, $r(42) = 0.31$, $p = .048$. A significant positive correlation was also found using

Pearson's r between intelligibility and recall accuracy for the nonnative-accented speakers for both the XS context group, $r(44)= 0.65$, $p<.001$, and the S context group, $r(42)=0.41$, $p=.005$.

Because Recall Prompt 1 is thought to measure the intelligibility of the target word and all preceding words in the list, word-level intelligibility was compared to accuracy scores at Recall Prompt 1. Context group, speaker accent, word-level intelligibility, and accuracy scores at Recall Prompt 1 were submitted to a linear mixed effect model to assess the effect of word-level intelligibility on recall accuracy Recall Prompt 1. A likelihood-ratio test indicated that the model including intelligibility was a better fit for the data than the model that did not include it, $\chi^2(1) = 73.16$, $p < .001$. The potential effect of word-level intelligibility on the downstream effects of context and speaker accent observed in Recall Prompt 2 and Recall Prompt 3 was also assessed. Context group, speaker accent, word-level intelligibility, and accuracy scores at both Recall Prompt 2 and Recall Prompt 3 were submitted to a linear mixed effect model. A likelihood-ratio test indicated that the model including intelligibility was a better fit for the data than the model that did not include intelligibility, $\chi^2(1) = 58.54$, $p <.001$.

Figure 9 shows the relationship between word frequency across all recall prompts for each speaker accent and context group condition. A likelihood-ratio test indicated that the model containing word frequency for target words at Recall Prompt 1 was a better fit for the data than the model without it, $\chi^2(1) = 17.40$, $p<.001$. The potential effect of word frequency on the downstream effects of context and speaker accent observed in Recall Prompt 2 and Recall Prompt 3 was also assessed. Context group, speaker accent, word frequency, and recall accuracy scores at both Recall Prompt 2 and Recall Prompt 3 were submitted to a linear mixed effect model. A likelihood-ratio test indicated that the model containing word frequency for target words at Recall Prompt 2 and Recall Prompt 3 was a better fit for the data than the model

without it, $\chi^2(1) = 20.01$, $p<.001$. Additionally, Figure 9 shows a significant positive correlation between word frequency and recall accuracy using Pearson's r for the nonnative-accented speakers in the NN XS group, $r(44) = 0.31$, $p=.033$, and the NN M group, $r(41) = 0.4$, $p=.008$. Context group, speaker accent, word frequency, and accuracy scores at Recall Prompt 1 were submitted to a linear mixed effect model to assess the effect of word frequency on performance at Recall Prompt 1.

General Discussion

In summary, Experiment 1 was conducted to confirm the relationship between speaker accent and intelligibility for single words that would be combined into the stimulus word lists used in Experiment 2. We found lower overall intelligibility in both nonnative-accented speakers than the native-accented speaker. Additionally, the nonnative-accented speakers did not significantly differ in intelligibility between each other. We were also able to confirm that the nonnative-accented speakers that we chose from the community had a significant nonnative accent through finding significantly higher accentedness ratings for the nonnative-accented speakers than the native-accented speaker. The nonnative-accented speakers did not significantly differ in accentedness rating between each other.

Experiment 2 was conducted to assess the potential downstream effect of speaker accent and context group on recall using the procedure used by McCoy et al. (2005). We found higher recall accuracy on Recall Prompt 1 than Recall Prompt 3 over all speaker accent and context conditions, suggesting the presence of a downstream impact of speaker accent and semantic context on recall. Additionally, we found lower performance in the Nonnative XS group than the Native XS group across all recall prompts, suggesting that eliminating any potentially helpful

effect of semantic context made the effect of speaker accent on recall more pronounced (Figure 5).

The pattern of results for Experiment 2 support the Effortfulness Hypothesis proposed by McCoy et al. (2005) through lower recall accuracy in the Nonnative XS accent-context group across all recall prompts compared with the Native XS accent-context group. The previously described results confirm both that nonnative accent can be considered an adverse condition for speech perception and that semantic context can be categorized as a helping factor for successful speech perception in adverse conditions.

While the decreased recall accuracy in the Nonnative XS accent-context group compared with the Native XS accent-context group across all recall prompts does not lend support to the ELU model (Rönnberg et al., 2008; 2013), it is still worth discussing in terms of the implication of semantic context on helping with recall accuracy. The ELU model (Rönnberg et al., 2008; 2013) describes speech perception as a process where individual words enter an episodic buffer and are matched against internal lexical representations of those words. Processing the total semantic context structure of the speech signal becomes relevant in the ELU model when the initial process of binding the incoming speech signal to internal lexical representations fails and explicit working memory processes are invoked to help with this process (Rönnberg et al., 2013). The ELU model (Rönnberg et al., 2008; 2013) does not support the pattern of results found in Experiment 2 because speaker accent and context level both contribute to differences in recall accuracy both in immediate recall measured through Recall Prompt 1, and through delayed recall for earlier presented words measured through Recall Prompt 2 and Recall Prompt 3 (Figure 7). This pattern of results suggests that processing of the total semantic context present in the speech

signal occurs concurrently with processing of individual words in both clear and adverse conditions for speech perception.

Word-level Intelligibility and Recall Accuracy

Through linear mixed effect modeling, we found that word-level intelligibility is a significant contributor to the impact of speaker accent, context group, and recall prompt on recall accuracy in Experiment 2 (Figure 8). More broadly, this finding suggests that nonnative-accented speakers may be put at a disadvantage when placed in public speaking situations, as English-monolingual listeners may have to expend more effort to understand them. Previous research has demonstrated that English-monolingual listeners are able to adapt to nonnative-accented speech after less than one minute of exposure to a nonnative-accented speaker (Clarke & Garrett, 2004). Clarke and Garrett (2004) also found that varying the semantic structure to reduce listener expectation (e.g., varying the part of speech at the end of the sentence from consistently containing nouns) did not significantly affect listeners' ability to adapt to nonnative-accented speech. While measuring speaker normalization was not a goal of Experiment 2, the results of Clarke and Garrett (2004) relating to the time course of normalization to nonnative-accented speech provide an interesting avenue to further explore the relationship between speaker accent and semantic context using prolonged exposure to nonnative-accented speech.

Word Frequency and Recall Accuracy

We found that word frequency was a significant contributor to the impact of speaker accent, context group, and recall prompt on recall accuracy in Experiment 2 using linear mixed effect modeling. We found a positive relationship between word frequency and recall accuracy such that words with high word frequency have a higher recall accuracy score than words that have a lower word frequency (Figure 9). This relationship between word frequency and recall

accuracy was especially pronounced in the XS and S context groups across the native and nonnative-accented speakers.

The observed variation in word frequency in the XS context group may be a result of the method used to construct the lists in the XS context group. The XS context group spans the zeroth and first orders of approximation defined by Miller and Selfridge (1950). The lists developed for the zeroth order of approximation contained a random sample of words taken from a dictionary of the 30,000 most common words in the English language published by Thorndike and Lorge (1944). It may be possible that the frequencies of the selected words in the XS context condition have changed in the time since Thorndike and Lorge (1944) was published and may not reflect current day word frequencies.

Limitations and Future Research

A limitation of the current study is that the word lists used in Experiment 2 may not reflect current day word frequencies, especially in the XS context condition. Future replication of the word list construction procedure used by Miller and Selfridge (1950) to develop their statistical approximations to English would be valuable to a future replication of this experiment as well as any other future work examining the relationship between semantic context and listening effort.

An additional limitation of the current study is that the recall responses provided by participants were scored according to an identical orthographic match with the target word. This method of scoring is highly restrictive, as it does not account for homophones and minor misspellings. A more dynamic scoring method is warranted for the results of Experiment 2 because the target words are presented entirely auditorily. A future reassessment of the recall accuracy data from Experiment 2 could utilize a dynamic scoring tool such as Ponto (Kessler,

2009), a computer software used to evaluate children's spelling, to construct a clearer picture of the impact of speaker accent and semantic context on recall accuracy.

References

Adobe Audition [Computer software]. (2023). Adobe Inc., San Jose, CA.

Audacity [Computer software]. (2021). Retrieved from https://www.audacityteam.org/download/

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language learning*, *38*(4), 561-613.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445-459.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600-1610.

Bradlow, A. R., and Bent, T. (2008). Perceptual adaptation to non-native speech. Cognition 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Burda, A. N., Hageman, C. F., Scherz, J. A., & Edwards, H. T. (2003). Age and understanding speakers with Spanish or Taiwanese accents. *Perceptual and Motor Skills*, *97*(1), 11-20.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647-3658.

Floccia, C., Butler, J., Goslin, J., and Ellis, L. (2009). Regional and foreign accent processing in English: can listeners adapt? J. Psycholinguist. Res. 38, 379–412. doi: 10.1007/s10936-008-9097-8

Ferguson, S. H., Jongman, A., Sereno, J. A., & Keum, K. A. (2010). Intelligibility of foreign-accented speech for older adults with and without hearing loss. *Journal of the American Academy of Audiology*, *21*(03), 153-162.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language learning*, *34*(1), 65-87.

Gordon-Salant, S., Yeni-Komshian, G. H., & Fitzgibbons, P. J. (2010). Recognition of accented English in quiet by younger normal-hearing listeners and older listeners with normal-hearing and hearing loss. *The Journal of the Acoustical Society of America*, *128*(1), 444-455.

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., & Schurman, J. (2010). Short-term adaptation to accented English by younger and older adults. *The Journal of the Acoustical Society of America*, *128*(4), EL200-EL204.

Incera, S., Shah, A. P., McLennan, C. T., & Wetzel, M. T. (2017). Sentence context influences the subjective perception of foreign accents. *Acta Psychologica, 172*, 71-76. doi:10.1016/j.actpsy.2016.11.011

JASP Team (2023). JASP (Version 0.17.2)[Computer software].

Kessler, B. (2009). Ponto [Computer Software]. Available at http://spell.psychology.wustl.edu/ponto.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL quarterly*, 36(2), 173-190.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals.

Mattys, S. L., Davis, M. H., Bradlow, A. R., et al. (2012). Speech recognition in adverse conditions: A review. *Lang Cogn Process*, *27*, 953–978.

Miller, G. A., & Selfridge, J. A. (1950). Verbal context and the recall of meaningful material. doi:10.2307/1418920

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. doi:10.1080/02724980443000151

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *45*(1), 73-97.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2),204–238.

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and hearing*, *39*(2), 204.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly journal of experimental psychology*, *20*(3), 241-248.

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International journal of audiology*, *47*(sup2), S99-S105.

Rönnberg, Lunner, T., Zekveld, A. ., Sorqvist, P., Danielsson, H., Lyxell, B., Dahlstrom, O., Signoret, C., Stenfelt, S., Pichora-Fuller, M. ., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. Frontiers in Systems Neuroscience, *7*, 31. https://doi.org/10.3389/fnsys.2013.00031

Rönnberg, J., Holmer, E., & Rudner, M. (2019). Cognitive hearing science and ease of language understanding. *International Journal of Audiology*, *58*(5), 247-261.

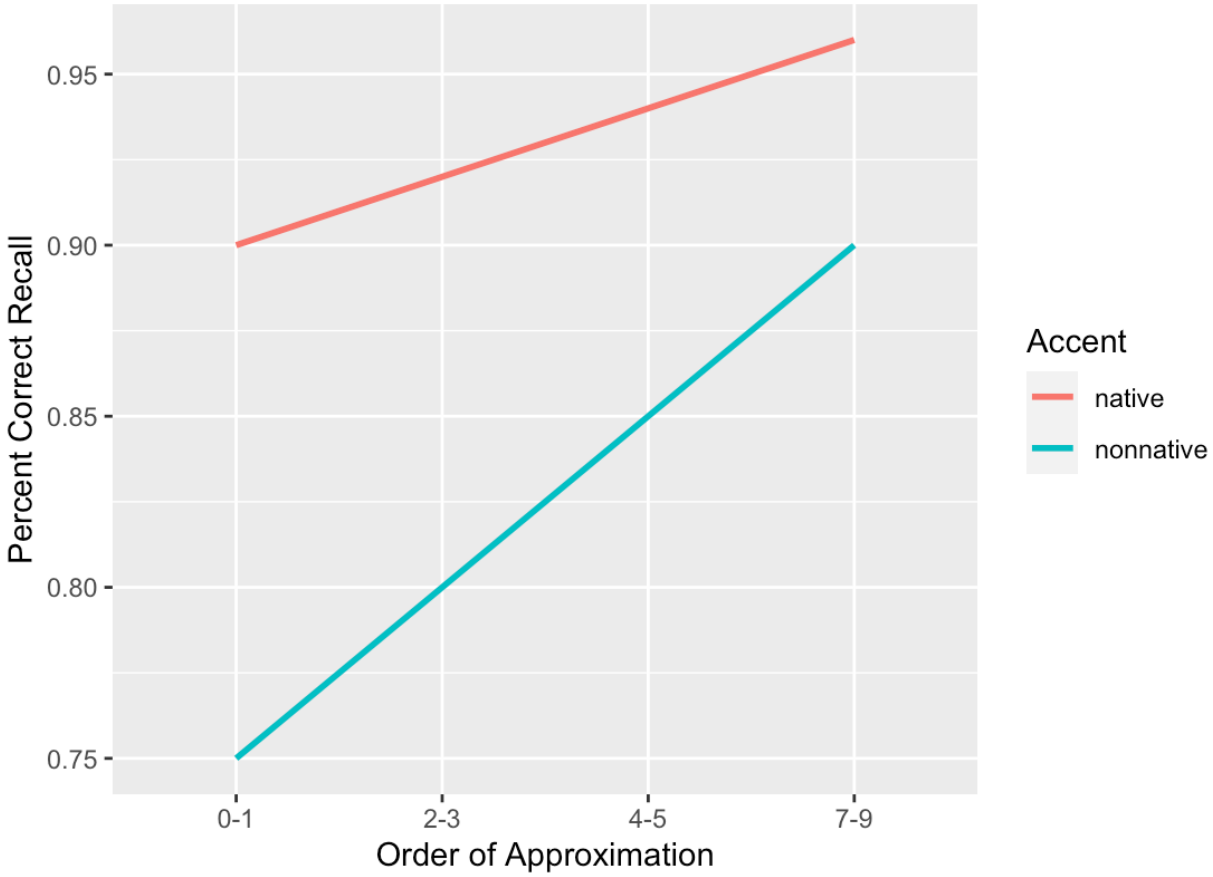
Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2015). afex: Analysis of factorial experiments. *R package version 0.13–145*.

Thorndike, E. L., & Lorge, I. (1944). The teacher's word book of 30,000 words.

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience, 8*, 1-4. doi.org/10.3389/fnhum.2014.00577

Figure 1.

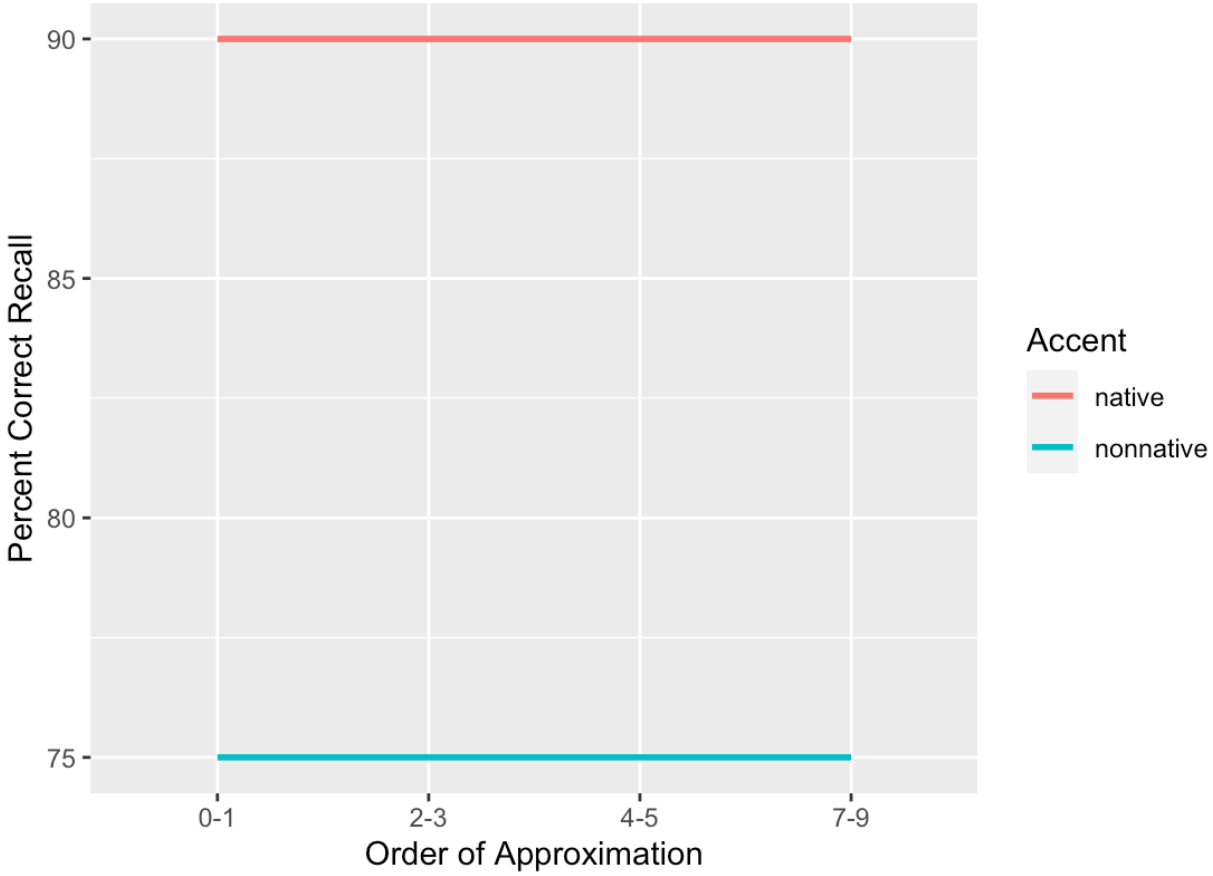
Predicted data pattern for Experiment 2 supporting the effortfulness hypothesis (McCoy et al., 2005).



Note. Order of approximation directly corresponds to context group such that 0-1 order corresponds to the XS context group, 2-3 order corresponds to the S context group, 4-5 corresponds to the M context group, and 7-9 corresponds to the L context group.

Figure 2.

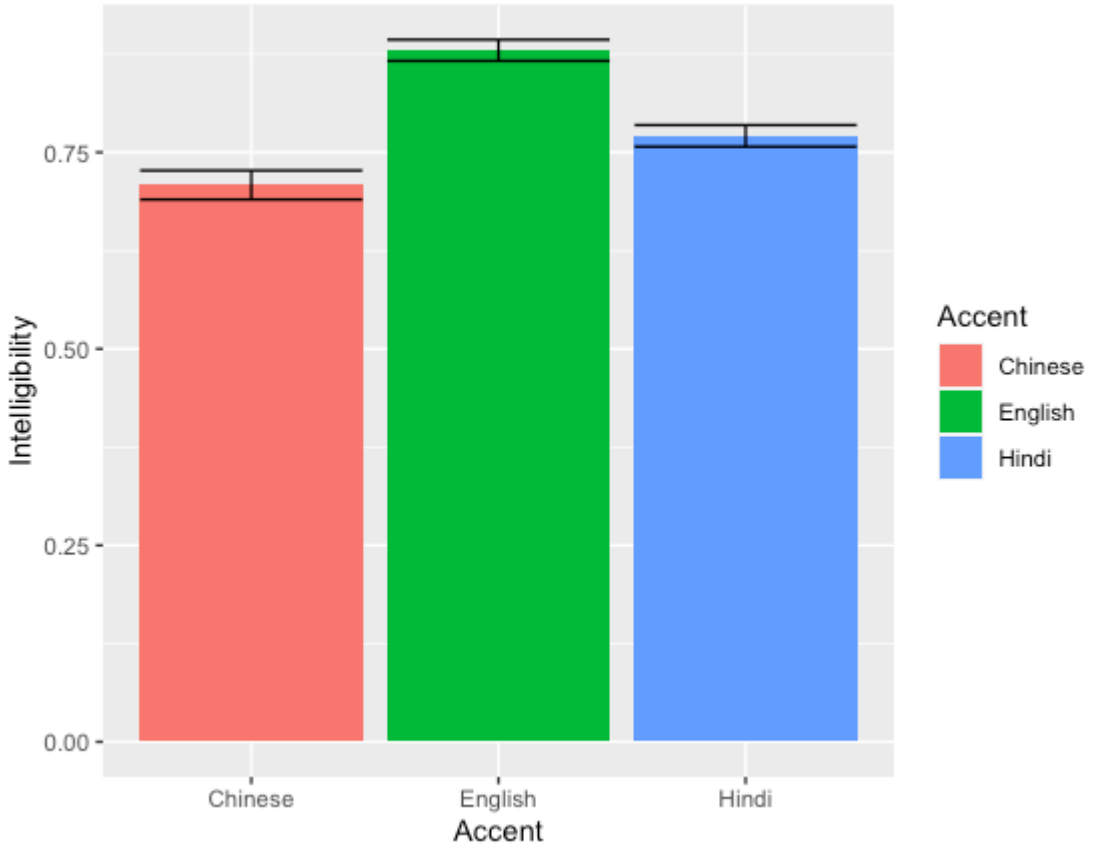
Predicted data pattern for Experiment 2 supporting the ELU (Rönnberg et al., 2008; 2013).



Note. Order of approximation directly corresponds to context group such that 0-1 order corresponds to the XS context group, 2-3 order corresponds to the S context group, 4-5 corresponds to the M context group, and 7-9 corresponds to the L context group.

Figure 3.

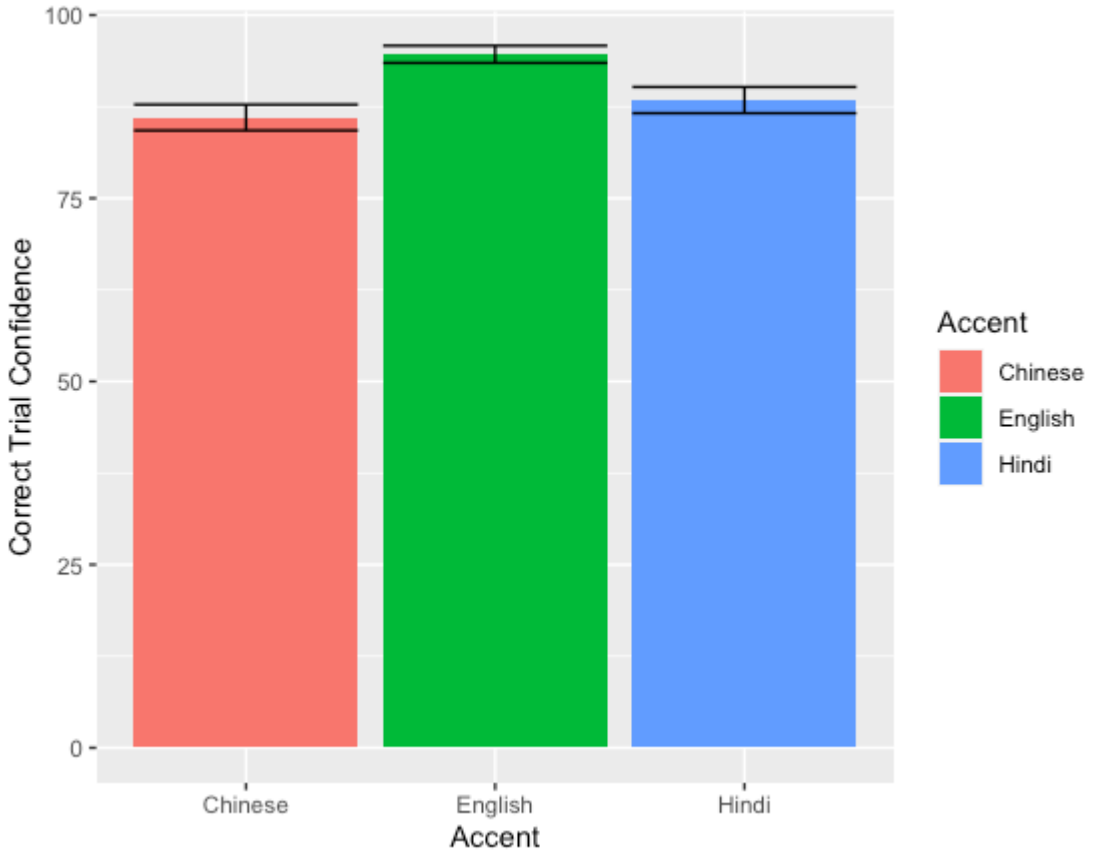
Intelligibility as a function of Speaker Accent.



Note. Error bars represent the 95% confidence interval for all scores.

Figure 4.

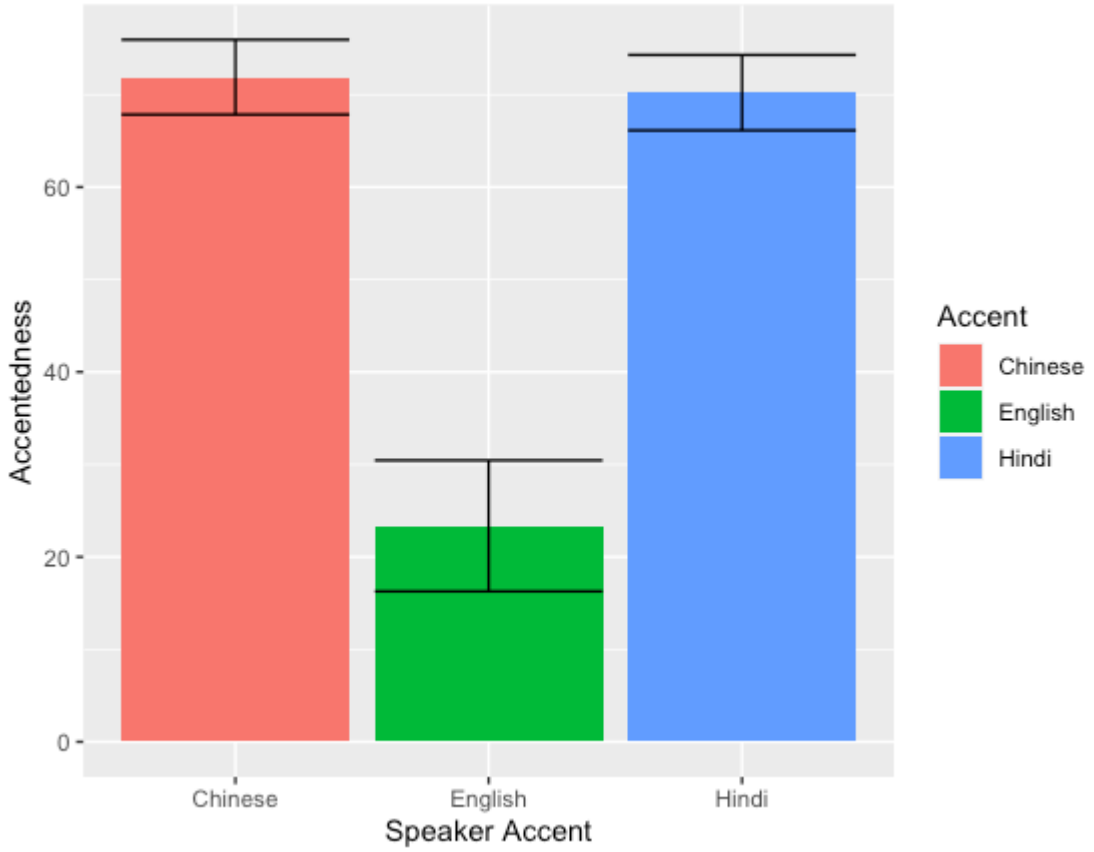
Correct trial confidence as a function of Speaker Accent.



Note. Error bars represent the 95% confidence interval for all scores.

Figure 5.

Accentedness rating as a function of Speaker Accent.



Note. Error bars represent the 95% confidence interval for all scores.

Figure 6.

Diagram describing the procedure of each trial in Experiment 2.

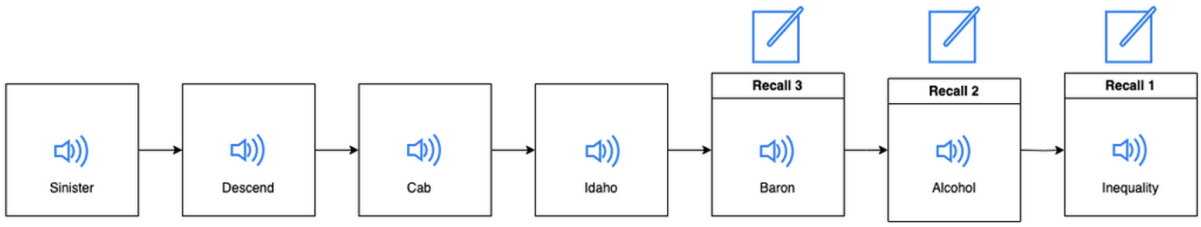
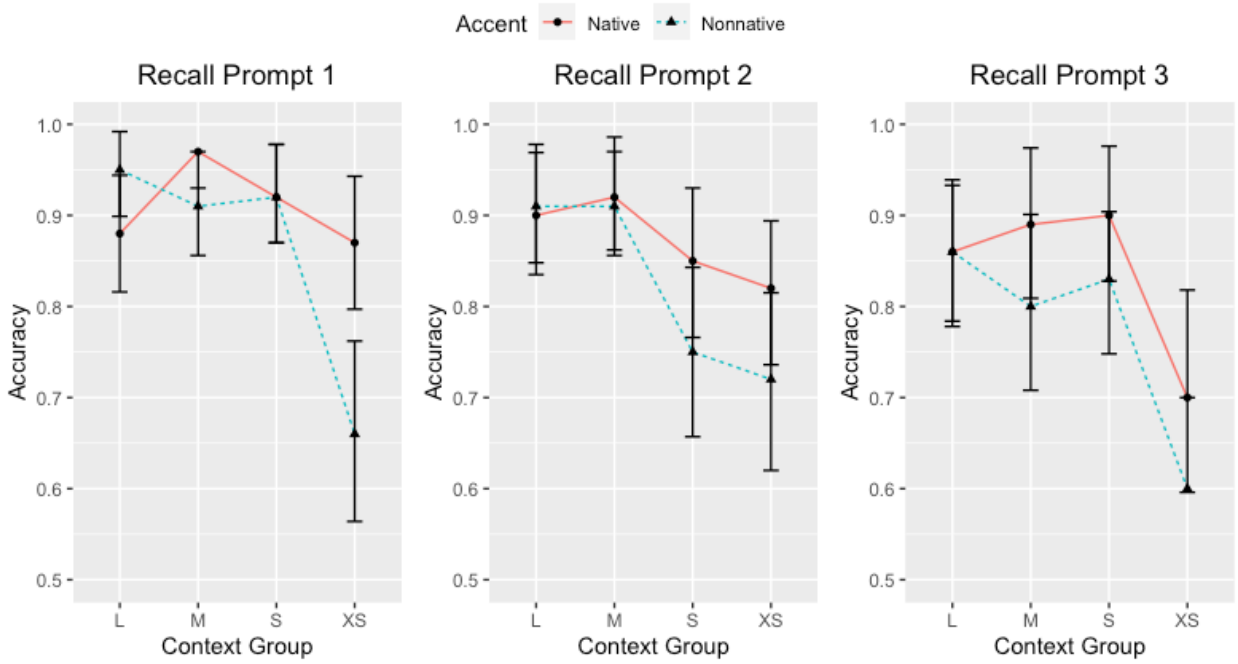


Figure 7.

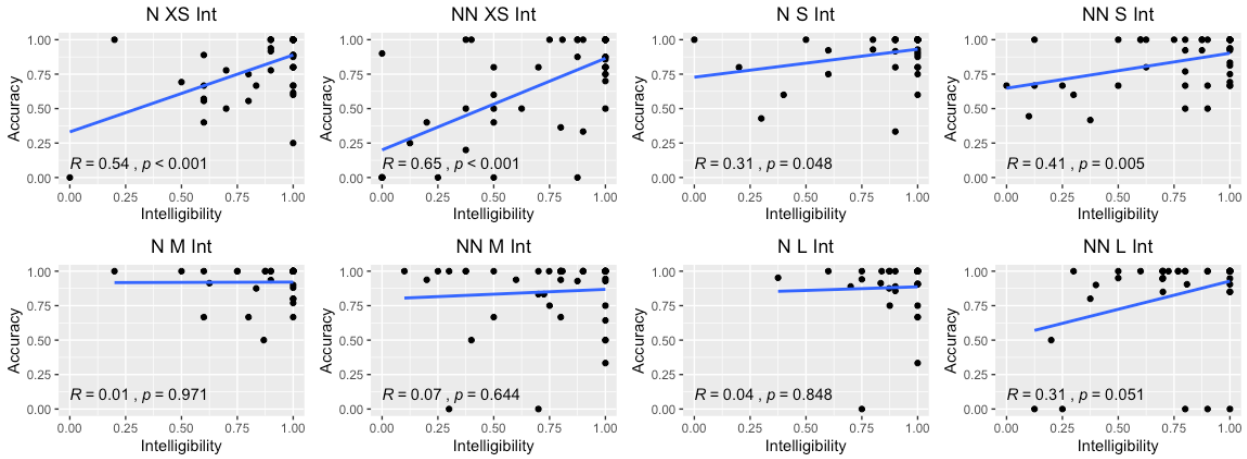
Average accuracy at all recall prompts as a function of Speaker Accent and Context Group.



Note. Error bars represent the 95% confidence interval for all scores.

Figure 8.

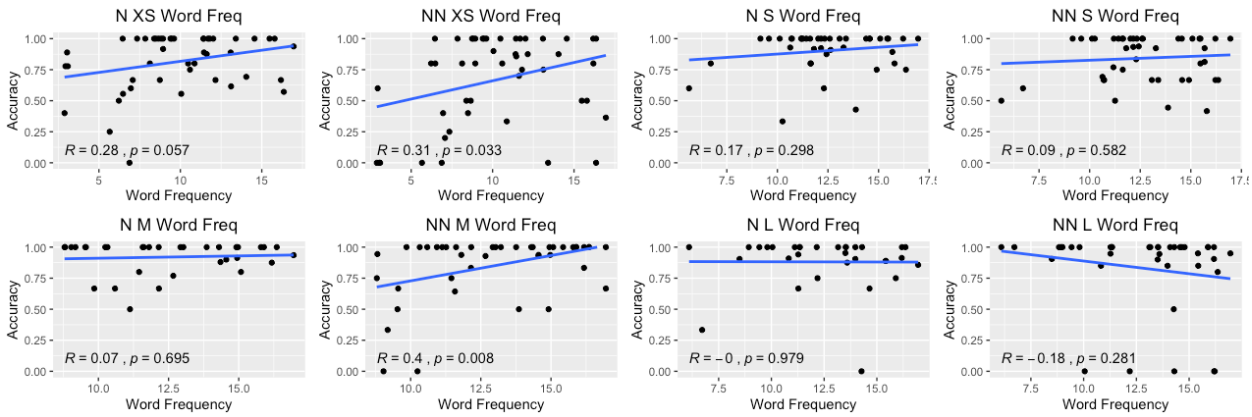
Recall Accuracy as a function of Intelligibility for each Speaker Accent and Context Group condition.



Note. "N XS Int" refers to Native-accent XS context group intelligibility, "NN XS Int" refers to Nonnative-accent XS context group intelligibility, "N S Int" refers to Native-accent S context group intelligibility, "NN S Int" refers to Nonnative-accent S context group intelligibility, "N M Int" refers to Native-accent M context group intelligibility, "NN M Int" refers to Nonnative-accent M context group intelligibility, "N L Int" refers to Native-accent L context group intelligibility, and "NN L Int" refers to Nonnative-accent L context group intelligibility.

Figure 9.

Recall Accuracy as a function of Word Frequency for each Speaker Accent and Context Group condition.



Note. "N XS Word Freq" refers to Native-accent XS context group word frequency, "NN XS Word Freq" refers to Nonnative-accent XS context group word frequency, "N S Word Freq" refers to Native-accent S context group word frequency, "NN S Word Freq" refers to Nonnative-accent S context group word frequency, "N M Word Freq" refers to Native-accent M context group word frequency, "NN M Word Freq" refers to Nonnative-accent M context group word frequency, "N L Word Freq" refers to Native-accent L context group word frequency, and "NN L Word Freq" refers to Nonnative-accent L context group word frequency.