PAUSE FOR THOUGHT (GROUPS): NON-NATIVE PAUSING BEHAVIOR AND EASE OF PROCESSING OF L2 SPEECH

Sadi Phillips, Indiana University

Alejandra Aguilar Perez, Indiana University

Hannah Alt, Indiana University

Isabelle Darcy, Indiana University

Features of prosodic phrasing in English are difficult to acquire even for advanced learners. One notably difficult prosodic feature is "thought-grouping", or pausing to delineate meaningful word groups. Learners may split clauses unpredictably or skip prosodic boundaries, creating “run-on” sentences. Either pausing pattern may impede processing of non-native speech, but it is unclear how much they impact processing difficulty. We used a tone detection task as an indirect measure of how split and run-on sentences impact processing difficulty of L2 speech. Thirty-four native English listeners responded to short tones semi-randomly inserted in sentences spoken with target-like or non-target-like thought-grouping. Listeners also judged each sentence as true or false, ensuring they processed them for meaning (not listening for tones alone). Tone detection reaction times (RTs) were compared in three pausing conditions: Original (pause at clause boundary); Run-on (pause absent); Split (additional pause mid-clause). As predicted, non-target-like pausing increased processing difficulty as evidenced by RTs in Run-on or Split conditions being slower than in the Original condition. No clear difference emerged between Run-on and Split conditions. These findings provide corroborating evidence with a processing task that non-target-like pausing results in increased processing difficulty. We discuss the implications of these findings for language pedagogy.

Cite as: Phillips, S., Aguilar Perez, A., Alt, H., & Darcy, I. (2022). Pause for thought (groups): non-native pausing behavior and ease of processing of L2 speech. In J. Levis & A. Guskaroska (eds.), *Proceedings of the 12th Pronunciation in Second Language Learning and Teaching Conference*, held June 2021 virtually at Brock University, St. Catharines, ON. https://doi.org/10.31274/psllt.13355

INTRODUCTION

Clusters of words delineated by pauses are sometimes termed ‘thought groups.’ To comprehend speech, a listener’s brain follows the rhythmicity of speech, including pausing behavior (Bourguignon et al., 2013). Thought groups as a unit of speech are important for processing and encoding information, to the extent that target-like thought grouping leads to better information retention as well as listener engagement (Hahn, 2004). Pausing behavior and prosody more generally can to some extent be a characteristic of individual speakers, but suprasegmental features such as pausing behavior are often acquired later and with greater difficulty by L2 English learners,

compared to native speakers. In addition, suprasegmental features tend to get less attention than other language components such as grammar and vocabulary in ESL or EFL classrooms, so learners may not even be aware of the existence of thought groups or their effect on speech. While there is no universally accepted definition of exactly what constitutes a thought group, we know that these units are important for processing and understanding speech. We use the term 'thought group' to indicate a unit of speech bounded by pauses, akin to the way that Pawley & Syder (2000) consider a 'clause' to be a unit of speech, despite the somewhat nebulous nature of both terms.

But thought groups themselves are not the only factor in this equation, evidenced by the fact that native speakers frequently experience disfluencies in spontaneous speech that can aid native listeners in predictive speech strategies (Bosker et al., 2014). So, how do non-native pausing disfluencies differ from those of native speakers? There are many ways to categorize pauses, such as by location, filled or silent, length, and frequency. Both a different distribution of pauses and increased frequency of pauses is characteristic of non-native speech (De Jong, 2016). When target-like pausing behaviors are altered, the result is that speech may be more difficult to understand.

Pauses that are misplaced, excessive in number, or overly lengthy (in any combination) can make speech more difficult to understand. There are striking differences in the characteristics of pauses that occur between clauses and within clauses, even in native speech (Pawley & Syder, 2000), and in general, within-clause pauses are seen as more disruptive and less fluent. Pause location can play a significant role in perceived L2 fluency (Kahng, 2018). That is, pauses occurring within clauses appear to reflect a lower level of fluency, as perceived by native listeners, whereas pauses between clauses are less informative about fluency and reflect conceptual planning time (De Jong, 2016). Pausing behavior in L2 speech can also be an indicator of proficiency, as it may reflect L2 linguistic knowledge and skills (Kahng, 2020), although more general prosodic patterns in an individual's L1 and L2 are indeed related.

Mid-clausal pauses can also be connected to processing difficulty on the listener's part. A few studies (Sanderman & Collier, 1997; Lege, 2012) have demonstrated links between pausing behavior and listener processing difficulty, but less well-studied is what pause characteristics affect processing difficulty. That is, while we know that processing difficulty and pausing behavior are linked, to what extent pause characteristics affect process difficulty has not been fully explored. In the body of work on English prosody, relatively few studies have been devoted to L2 prosody, and fewer still to the relationship between thought grouping (an umbrella term for various definitions of thought groups) and processing difficulty. This study sheds light on this relationship by investigating what types of pauses most affect processing difficult in native listeners. In doing so, we hope that the results can be used to draw attention to the importance of these features and aid in the acquisition of target-like pausing behaviors.

Research Question

Our research question asks whether non-target like thought grouping patterns in L2 speech increase processing difficulty for native listeners, and if yes, how?

METHODS

To answer this question, we operationalize "increased processing difficulty" by measuring reaction time (RT) in a dual-task paradigm. We ask participants to determine if a sentence is true or false (Task 1) while they simultaneously monitor for a tone that may or may not occur at any point in the sentence (Task 2). Since the participant has a finite amount of cognitive resources to devote to these tasks, a slower response time on Task 2 is hypothesized to indicate that fewer resources are available to devote to Task 2 due to the need for resources to complete Task 1. Thus, an increased difficulty of Task 1 (by artificially inserting non-target-like pauses into the stimulus) is hypothesized to lead to a slower RT in Task 2. Because we keep the semantic content of sentences constant across conditions, and only vary the pause location, any differences in processing ease across conditions (as seen via the tone detection task) should be due to the pause location manipulation. Comparing the RTs of tone detection in two different conditions of inauthentic pause location to original, target-like sentences (see below) allows us to evaluate how much non-target-like thought grouping slows sentence processing (i.e., increases processing difficulty). This study was preregistered on the Open Science Framework (OSF) on December 11, 2020. All data, analysis, and materials are available at

https://osf.io/bafsw/ (https://osf.io/bafsw/?view_only=1893804c0398437987511556bb9171cb).

Participants

Thirty-seven volunteers participated, with three excluded (one due to a reported hearing loss, two due to technical malfunctions during data collection). Five participants reported having a neurological disorder, but all reported it as treated. Their data were included in the final analysis. Thus, our total N = 34. Listeners were all native speakers of English (23 females, 11 males) who lived most of their lives in the US. This included 31 monolinguals and 3 early bilinguals. The average age of the participants was 31 years (*SD* = 3.9). Fourteen participants reported knowing at least 1 foreign language, but none reported a high proficiency in it. Participants reported moderate levels of familiarity with L2 accented speech, with an average score of *M* = 5.4 (*SD* = 2.9, range 1-10) on a 10-point Likert scale (1 = no familiarity, 10 = extremely familiar).

Materials

The experimental stimuli were 30 multiclausal (15 true, 15 false) statements with unfilled pauses occurring in one of three ways: original (200 ms), split (610 ms), or run-on (<80 ms). Word frequency is the same across conditions. Additionally, 40 filler sentences (20 true, 20 false) were created. All stimuli were recorded on a Zoom ZH1 recorder by a highly proficient non-native speaker of English (L1 French), phonetically trained to produce the pauses as required. All sentences were manually extracted from the audio file. Table 1 outlines the three different pause locations in the three conditions as well as an example of a true/false statement.

Table 1

Pause location and example for the three experimental conditions.

Condition	Pause location	Example sentence
Original	Authentic (The pause is at the clause boundary)	Glass can be broken [pause] especially if it falls on concrete.
Run-on	Absent (there is no pause between the two clauses)	Glass can be brokenespecially if it falls on concrete.
Split	Inauthentic (there is an additional pause occurring mid-clause, in addition to the pause at the clause boundary)	Glass can be broken [pause] especially if it [pause] falls on concrete.

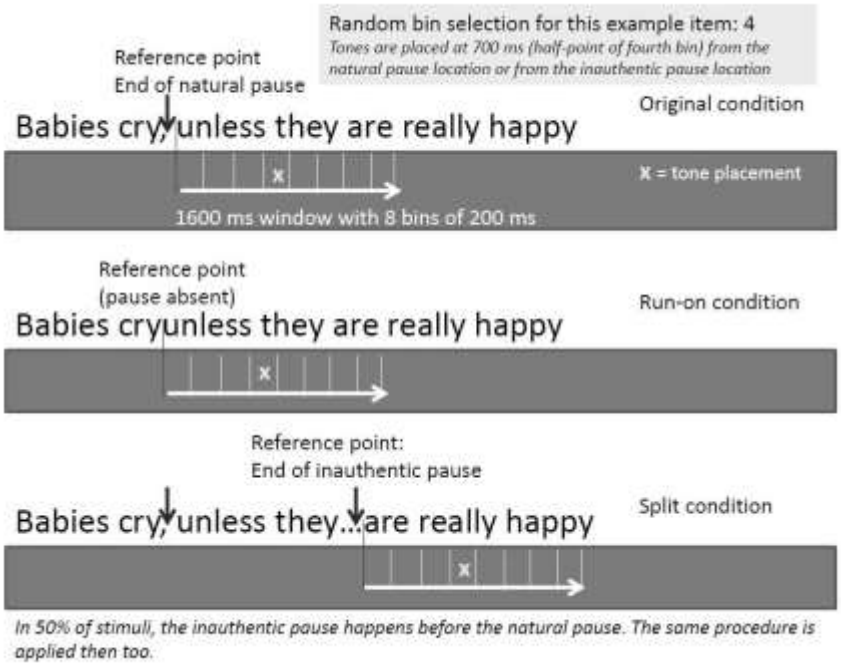
Original sentences (condition O) contained the authentic pause produced at the clause boundary. This pause also served as the reference for the timing of the tonal insertion. Run-on sentences (condition R) were characterized by the absence of any pause, such that the two clauses are not prosodically separated. Finally, split sentences (condition S) contained an inauthentic pause which broke up a prosodic group. Half the experimental sentences contained this pause before the reference pause, half after it. In total, there were 30 triplets of experimental sentences, varied according to 3 conditions of pause placement (O, R, S). This resulted in a total of 90 individual recorded audio files. All target sentences had a tone. All experimental sentences were piloted by 5 English native speakers to ensure all would pause at the natural (authentic) location in oral production.

The experimental stimuli were split into three lists. Each list contained one sentence from each experimental triplet: 1) run-on condition 2) split condition 3) original condition. Additionally, half of the experimental sentences were true; the other half were false. These lists were counterbalanced following a 3×3 Latin Square design, such that each list had one version of each experimental sentence, the same number of items in each condition, as well as the same number of true and false items. For each condition, this resulted in 10 datapoints per list. Each participant was randomly assigned to one list, producing 30 experimental datapoints (10 per condition). The 40 fillers were the same in each list.

A 200ms 1000Hz tone was inserted at various points in the sentences. Figure 1 shows insertion placement of the tone at semi-random locations after the original pause for condition O, after the removed pause, for condition R, and after the inserted pause for condition S. All experimental sentences as well as 50% of fillers contained a tone.

Figure 1

Semi-random tone insertion in experimental sentences for each condition.



*A 200 ms bin (from 1 to 8) was randomly chosen for each triplet, and the tone was inserted at the midpoint of that bin after the original pause's reference point, or in the S condition, after the end of the inauthentic pause.

Procedures

All testing was conducted remotely, using Psychopy (Peirce et al., 2019) via the Pavlovia platform, as well as Qualtrics. Participants met individually via video conference with a researcher. A strict testing protocol was followed to ensure consistency. Participants completed a background questionnaire, a volume calibration and headphone check to ensure they were in a favorable environment for the perception task, and the tone detection task. The entire procedure took approximately 50 minutes.

To control volume consistency across participants, we used step 1 from a series of tests via HearingTest.Online (Pigeon, 2013). Participants matched the volume of the sound of rubbing their hands to that of the calibration file played on their computer and kept this volume for the remainder of the experiment. Following this calibration, all participants completed a headphone check, a brief beep detection task in each ear, to ensure their environment allowed them to hear faint tones before completing the main experiment.

Tone Detection Task

The tone detection task consisted of two simultaneous tasks: the tone detection itself, and a true/false decision task (to ensure that listeners process the sentences for meaning). During a trial, participants heard a statement that was either true or false and had to 1) press a key if a tone was present, and then 2) decide on the sentence's truth value. Each participant was randomly assigned to one of three experimental lists (see section Materials).

A practice round (10 trials) with RT and accuracy feedback familiarized participants with the two parallel tasks and emphasized the importance of responding quickly if a tone is detected. The 'J' key was designated for 'True', the 'F' key for 'False'. The spacebar was assigned to the tone detection, ensuring similarity across participants' computer setups. While the tone response emphasized speed, the true/false was not timed. If no true/false response was recorded, the trial timed out after 10 seconds. Reaction time, accuracy of tone detection, and true-false judgement were recorded for analysis. This portion of the experiment took approximately 20 minutes.

Data Analysis

The measures collected were tone detection RT, tone detection accuracy, and true/false judgment accuracy. Tone detection accuracy was scored as 1 or 0 and expressed as hits or misses, and false alarms or correct rejections, respectively. True/False accuracy was measured using the same 1/0 scoring system, expressed as a percent error. The tone detection data for the test conditions and fillers were first checked to ensure that the 34 included participants did the task correctly and detected tones when present with sufficient accuracy. A predetermined threshold of 50% hits or correct rejections for each trial did not result in the removal of any participants.

RTs were computed in ms from the onset of the tone until the participant's button press. This means that on trials without tone, no RT is reported. For trials with tones, only hits (=correct tone detection) were included to compute the RT. Any RTs stemming from a False Alarm (when a sentence did not contain a tone, n = 3) were discarded. Similarly, any detection response given prior to the tone or shorter than 200 ms was discarded, being too fast and unlikely to represent actual responses to the tone in the current sentence. They are more likely either a wrong press or a delayed response to a preceding sentence. All datapoints from the training items (n=310) are also discarded, leaving a total of 1620 datapoints included in the analysis. Finally, extremely long RT values that are beyond 2 SD from the mean RT for a given participant were trimmed to the mean $\pm$ 2 SD for that participant.

To further neutralize inherent RT differences among participants due to their operating systems for instance (given our remote testing situation), we z-scored the RT using the Original condition as the baseline. So, the z-scores give us a participant-specific estimate of how much faster or slower someone is on the R and S condition compared to their RT on the O condition. A positive z-score would indicate the participant is slower and has higher RTs compared to the mean, while negative z-scores indicate the participant is faster and has lower RTs compared to the mean.

We first examine whether listeners correctly processed the sentences. After excluding items as described above, overall error rate was computed for true (n = 814) and false (n = 806) sentences. The overall error rate was 3.6% (see Table 2), indicating that participants were accurately processing the sentences for meaning. However, the error rate was higher for true sentences (*M* = 4.5, *SE* = 0.7) than false ones (*M* = 2.8, *SE* = 0.6), *t*(1549.2) = 1.8, *p* = .036 (one-tailed), 95% CI [-0.0 – 0.035], Hedges' g = (0.028 – 0.045) / 0.189 = 0.09. This effect is small but seems consistent across test conditions. We will therefore evaluate the tone detection RTs and z-scores separately for true and false sentences.

Table 2

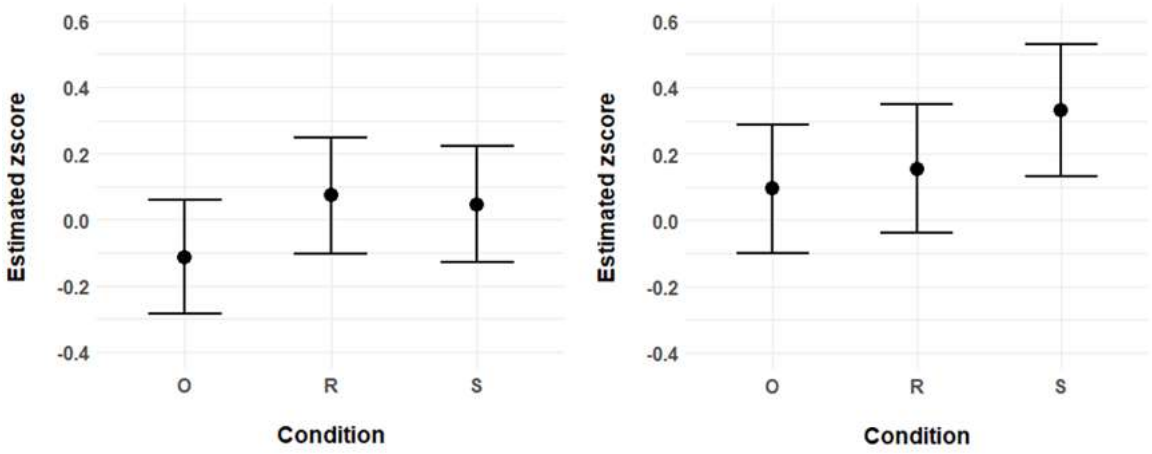
Error percentage by condition for true vs. false sentences.

% Error by sentence type	Filler	Original	Run-on	Split	Overall
FALSE	4.0%	1.1%	2.3%	1.1%	2.9%
TRUE	2.9%	4.0%	5.7%	8.6%	4.2%

We next checked the accuracy of tone detection by sentence type. Tone detection accuracy is also expressed as % error (misses and false alarms). On average, the mean error rate across all sentences is 2.6% (M_{false} = 2.4%, M_{true} = 2.9%). Since tone detection was overall highly accurate, providing enough data points for reliable latency analysis, we now examine the z-scored RT of tone detection in each of the three test conditions. Figure 2 presents the mean estimate of the z-score in each condition separately for true and false sentences.

Figure 2

Mean z-score of RTs and 95% CI in each of the test conditions (O, R, S) for true sentences (left panel) and false sentences (right panel).



The results show that tones in Original sentences are detected faster than in Split or Run-on sentences (for true and false sentences). The z-score for R and S is higher compared to O (the reference for z-scoring), especially for true sentences. This means that participant-internally, RTs are slower for R and S compared to O items. A linear mixed model fit by maximum likelihood was run in R (with the package *lme4*) on z-scores, declaring varying intercepts for participants and items. Assumptions of normality of error terms and of homogeneity of variance were checked for true and false sentences separately. Residual plots showed mild to moderate deviations from homogeneity of variance, and Shapiro-Wilk tests indicated moderate non-normality. Residuals for the item intercepts were normally distributed, but for the overall model and for the participant intercepts, the Shapiro Wilk tests were significant ($w > .9$ for all tests, $p < .05$ for all tests). However, according to Schielzeth et al. (2020), linear mixed models are remarkably robust even to moderate/severe violations of distributional assumptions. They conclude that effects of such violations on both random and fixed effects are surprisingly small, although the worst outcomes were in the case of an underlying bimodal distribution, which is not the case for our data. As such, we decided to cautiously proceed with interpreting the model as is. Fixed factors were condition (O, R, S) and list (1, 2, 3) as well as their interaction. An analysis of variance was run using the package *car*, reporting Type III Wald chi-square tests. There was a main effect of condition ($\chi^2(2, N = 34) = 8.55$, $p = .014$), no effect of list, but an interaction between list and condition ($\chi^2(2, N = 34) = 8.02$, $p = .018$). This indicates that the size of the difference between conditions is modulated by list. The same effects and interaction emerged for false sentences in a parallel model. When examining the z-score differences by list (across sentence types), the data show that items in List 2 produced a higher z-score ($M_R = .18$, $M_S = .42$, vs. $M_O = .01$) than items in Lists 1 and 3: the means for R and S items hovered around .05-.08. There is no obvious reason why items in List 2 are more consistent across conditions and sentence types.

Overall, the effects we found are modest, and they appear to be modulated by intervening factors such as list, which remain to be investigated in more depth. However, these results suggest that non-target-like thought groups can reduce the ease of sentence processing.

DISCUSSION

As predicted, non-target-like pausing resulted in slower tone detection RTs in comparison to target-like pausing, even for a highly effective non-native speaker. Participants' RTs were slower in both the Split and Run-on conditions in comparison to the Original condition. This is the case for true sentences as well as false sentences, but mostly for Split sentences. Because the sentences are otherwise the same and only the pausing differs, the data suggest that pausing behavior moderately impacts processing difficulty. Given the increased processing load required to repair sentences with non-target like pausing, we hypothesize that the cognitive resources left available to detect the tone result in a slower RT.

Overall, listeners were highly accurate in their ability to distinguish true and false sentences, suggesting that they processed the meaning of the items accurately. Accuracy errors are slightly higher in the case of true sentences. This may indicate that participants needed to reprocess the entirety of the true sentences prior to making a judgment, whereas the false sentences may have been transparently false from the first clause. It may be the case that in sentences which are falsified

in the beginning, listeners are then able to free up cognitive resources because the accuracy task is completed early in the sentence. Conversely, for true sentences, the listener must process the entirety of the sentence to ensure its truthfulness, leaving relatively fewer cognitive resources available (in comparison to false sentences) for the tone detection task.

Despite having a modest sample size, we found that non-target-like thought-grouping increases native listeners' effort while processing the sentences.

What we have operationalized in this study as processing difficulty may be a component measure of comprehensibility, a term which is often used as a global measure but rarely in a systematic, standardized way. Comprehensibility is a nebulous term in the literature in much the same way as the term "clause," (as it is argued by Pawley & Syder [2000]). That is, we seem to have a prototypical understanding of what the term refers to, but exactly what instantiations are members of that category (and which are not) is less clear. Comprehensibility is by nature an indirect measure, and the component mechanisms which contribute to the overall measure of "comprehensibleness" are not well understood. While our results cannot be said to indicate overall comprehensibility, we hypothesize that our measure of processing difficulty could give insight into one facet of comprehensibility.

If processing difficulty is indeed a component of comprehensibility, this could have important implications for language pedagogy. These results lend tentative support to the importance of the teaching of prosodic features such as thought grouping. While both native and non-native speakers demonstrate speech disfluencies, the characteristics of non-native disfluencies seem to be qualitatively different. These disfluencies may have a subtle but measurable impact on processing difficulty for native listeners. As such, even advanced non-native speakers may benefit from attention to thought grouping. In terms of future research, we hope to see an increase in studies investigating thought groups and thought group instruction, with a variety of methodologies. In our own research, a more robust sample size would help obtain clearer effects. Despite the small effect size, our task allows us to obtain a measure of processing load during processing. However, as a follow-up study, it may be useful to correlate a more traditional measure of comprehensibility with our measure of processing difficulty to provide support for the relationship between the two. This could shed light on the potential for processing difficulty to be measured as a subcomponent of comprehensibility, and be used to better understand the impact of non-target-like thought grouping on native listeners.

ACKNOWLEDGMENTS

We want to thank the following people for their contributions to this project: our fellow team members Danny Graff and Lucas Derry; Brian Rocca and Mike Iverson, for their help with statistical analysis in R; Francis Tyers, for his help with python scripting; Ryan Lidster and Nora McNamara, for their presentation feedback; and finally, the L2 Psycholinguistics Lab at IU for their continued moral support.

ABOUT THE AUTHORS

Sadi Phillips is a PhD student of Second Language Studies at Indiana University. Her primary research focus is phonology, specifically the Mental Lexicon. Email: sadiphil@iu.edu

Alejandra Aguilar Perez is a graduate of the TESOL & Applied Linguistics MA program at Indiana University. A Fulbright ETA grantee in 2018, she is interested in L2 pedagogy, L2 pragmatics instruction, and instructional design for adult language learners. Email: aleaguil@iu.edu

Hannah Alt is a graduate of the IU TESOL & Applied Linguistics MA program and a current PhD student at Miami University of Ohio. She is interested in user experience design and optimizing integration of technological tools for language learning and teaching. Email: althc@miamioh.edu

Isabelle Darcy is a Professor in the Department of Second Language Studies at Indiana University (USA). Her research focus is speech perception and word recognition in first and second language, the acquisition of second language phonology, and effective pronunciation instruction. Email: idarcy@indiana.edu

REFERENCES

Bourguignon, M., De Tiege, X., de Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., ... & Jousmäki, V. (2013). The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human brain mapping*, *34*(2), 314-326.

Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. H. (2014). The perception of fluency in native and nonnative speech. *Language Learning*, *64*(3), 579-614.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113-132.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201-223.

Kahneman, D. (1973). *Attention and effort.* Prentice-Hall, Inc.

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569-591.

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics*, *41*(2), 457-480.

Lege, R. F. (2012). The Effect of Pause Duration on Intelligibility of Non-Native Spontaneous Oral Discourse.

Pawley, A., & Syder, F. H. (2000). The one-clause-at-a-time hypothesis. In *Perspectives on fluency*. University of Michigan Press.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

Pigeon, S. (2013). https://hearingtest.online/. Accessed 2021-09-28.

Sanderman, A. A., & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech*, *40*(4), 391-409.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. S., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*, 1141-1152.