Durational cues to word boundaries in spontaneous speech

Jiaer Tao[1], *Francisco Torreira*[1], *Meghan Clayards*[1 2]

[1]Department of Linguistics, McGill University, Canada
[2]School of Communication Sciences and Disorders, McGill University, Canada
jiaer.tao@mail.mcgill.ca, francisco.torreira@mcgill.ca, meghan.clayards@mcgill.ca

Abstract

We investigated the extent to which durational cues to word boundaries are present in spontaneous speech. Spontaneous speech of North American English was elicited in a production experiment, with target phrases embedded in articles provided to participants. Each pair of target phrases only differed in the placement of word boundaries, e.g. *beef#eater vs. bee#feeder*. We examined the duration of: (1) the pivot consonant at the juncture (e.g. [f] in [bi:fiɚ]), (2) the pre-juncture section (e.g. [bi:] in [bi:fiɚ]), and (3) the post-juncture section (e.g. [iɚ] in [bi:fiɚ]), to see how these durations can signal word boundaries. We found no evidence for word-final lengthening in our study. However, similar to boundary-related lengthening found in laboratory read speech, word-initial lengthening was found in spontaneous speech, which could potentially serve as an important cue to word segmentation.

Index Terms: word segmentation, duration, prosody, speech recognition, boundary-related lengthening

1. Introduction

Spoken language comprehension is a complex process that requires language listeners to segment the continuous speech signal efficiently. Possible cues to word boundaries have been studied for several decades on the basis of laboratory read speech. Aside from phonotactic restrictions [1, 2], metrical structure [3, 4, 5, 6] and syntactic and semantic contexts [7], acoustic phonetic cues, e.g. durational patterns, F0, amplitude contour, and allophonic variation, have been found to be informative in indicating word boundaries when other (e.g. contextual) cues are lacking or ambiguous [7]. In particular, durational patterns (e.g. segmental duration, boundary-related lengthening; aspiration duration; pause; etc.) provide rich information in signaling word boundaries [8, 9, 10, 11, 12, 13].

Speech style also has a large effect on how speech sounds are realized [14]. While lab read speech generally tends to involve hyper-articulation, spontaneous speech is characterized by more reduction and lenition phenomena. For this reason, one may wonder to what extent the acoustic word-boundary cues (i.e. durational patterns) identified in previous research, for the most part based on lab read speech data, are present in everyday spontaneous speech. To date there is little work on this topic, which we hope to address [15].

Regarding durational cues to word boundaries, two boundary-related lengthening phenomena have been well discussed in previous literatures: word-initial lengthening and word-final lengthening. Word-initial lengthening is the word-level manifestation of boundary-related prosodic strengthening. It could be interpreted as the phenomena of domain-initial strengthening (DIS) such as the long post-aspiration of English voiceless stops in word initial positions [17, 18, 19, 20, cf. 16]. Regarding the scope of DIS, previous studies on English suggested that the temporal lengthening effect is strongest at the initial segment and becomes gradually weaker for the following segments [18, 19, 20, 21, 22, cf. 16].

Word-final lengthening is a word-level manifestation of preboundary lengthening [16]. Preboundary lengthening is thought to be the temporal modulation of domain-final phonological units, as controlled by the speakers to help encode prosodic structure [11, 16]. The most well attested case of domain-final lengthening is phrase-final lengthening—the lengthening of the final segments of phrasal prosodic units. It has been reported that the effect size of domain-final lengthening is proportional to the level of the domain in the prosodic hierarchy [11]. The evidence for lengthening at the ends of words is less clear. One previous study did not find strong evidence for the lengthening of word-final units, and instead argued that any observed lengthening can be explained by other mechanisms [11]. For instance, in many cases it's possible that target words were in fact phrase-final; meanwhile, the evidence from non-phrase-final positions was lacking [11]. An additional confounding factor is poly-syllabic shortening [23] whereby a syllable is shorter when it is part of a poly-syllabic word compared to a mono-syllabic word. For example, in Turk & White's study [24, cf. 11], the duration of the sequence *shake* was longer in *shake#downstairs* than in *shakedown#stairs*. This result could be either explained by word-final lengthening (as *shake* was preceding a word boundary in *shake#downstairs*) or by poly-syllabic shortening (*shake* was shorter in *shakedown#stairs* as there were more syllables in the word, compressing the duration of each syllable). This mechanism can be interpreted as an instantiation of durational compression, which states that at each level of representation, the more units that are present, the more the duration of each unit will be compressed [25, for Menzerath's law, see 26].

Our study examined these two boundary-related lengthening phenomena in spontaneous speech. If the results from previous studies of lab speech [e.g. 11] hold for spontaneous speech (i.e. only an effect of word-initial lengthening), we expect word onset consonants to be longer than coda consonants in our study.

A better understanding of the systematic variation of boundary-related durational patterns, especially in spontaneous speech, is important for understanding the interaction of single sound units and utterance-level prosodic structure. It would also provide important information for models of language production, speech recognition and language acquisition.

2. Experiment

We elicited spontaneous speech containing a series of target phrases by asking participants to relay information in a text to

a confederate. Inspired by classic studies [3, 8], 27 pairs of (quasi-) homophonous phrases were compiled differing in word boundary placement, namely, whether the pivot consonant was a coda or an onset, e.g. *beef#eater* vs. *bee#feeder*. Pivot consonants at word boundaries included voiceless stops /p, t, k/, voiced stops /b, d, g/, fricatives /f, s/, nasals /m, n/ and cluster /st/. Each target phrase was embedded in a different article. See Figure 1 for example.

Participants were instructed to read the article silently first, and to explain its content to the confederate. Half the articles were presented in an initial session (e.g. an article about a *beef eater*), and their counterparts (e.g. an article about a *bee feeder*) in a second session one week later. For each session, participants were asked to do the task twice.

Figure 1: *Two of the articles used in our experiment, corresponding to the target phrases 'beef eater' and 'bee feeder'.*

Data from 6 monolingual speakers of North American English (3 females and 3 males) was used. All productions were force-aligned [27], and target segments at the juncture of target phrases were hand-adjusted. Note that the data was not balanced due to the unscripted nature of the task. For certain target phrases, one member of the pair was produced more frequently than the other member by some participants.

We collected 745 tokens from 27 pairs of target phrases. Out of 745 tokens, we excluded 246 tokens from 7 pairs of target phrases which had outstanding pronunciations (e.g. disfluencies) or were not matched for number of segments (e.g. *fork handles vs. four candles*). Additionally, 144 cases exhibiting salient word-boundary cues were excluded, such as categorical allophonic variations of word-final /t, d/s; glottalization when a word-final stop was followed by a word-initial vowel; and short pauses at word boundaries. We did not exclude those post-aspiration tokens of word-initial /p, t, k/s because this well-known allophonic variation phenomenon of English voiceless stops could be argued to be a manifestation of DIS. Whether it is purely a durational modulation is still under discussion, so we kept these tokens and report separate analyses with and without them below.

In the end, a homogeneous group of 355 tokens (baseline of 62%: 133 tokens of coda consonants & 222 tokens of onset consonants) from 20 pairs of target phrases were investigated for the present analysis. All tokens of /g/ were excluded because of glottalization. The number of tokens for each type of pivot consonant are reported in Table 1.

Table 1: *Number of tokens*.

	Voiceless			Voiced		Nasal		Fricative		Cluster
	/p/	/t/	/k/	/b/	/d/	/n/	/m/	/f/	/s/	/st/
N	67	26	21	42	33	30	17	41	56	22

We annotated each token into three sections: (1) the pivot consonant (e.g. [f] in [bi:fiɚ]), (2) the pre-juncture section excluding the pivot consonant (e.g. [bi:] in [bi:fiɚ]), and (3) the post-juncture section excluding the pivot consonant (e.g. [iɚ] in [bi:fiɚ]). Segmental boundary markers for pivot consonants were at the onset and offset of constriction for fricatives (marked by frication) and nasals (marked by drop in amplitude); for stops, the segmental duration included both the duration of closure and the duration from the release to the onset of vocal fold vibration (VOT) for the following vowel. Markers for closure and release were inserted as well.

Absolute durations of the three sections were extracted from each target phrase, and relative duration measures were computed as proportions of whole phrase duration, to control for speech rate. For each token, the presence or absence of an intonational boundary (i.e. ip/IP) at each edge of the target phrase was also noted mainly on the basis of boundary tonal events and perceptibly distinct junctures.

3. Results

3.1. Relative duration of the pivot consonant

We first examined relative duration of the pivot consonant at word boundaries. A linear mixed-effects model [28] of relative duration of the pivot consonant was fit with pivot consonant position (*coda* or *onset*) as the main predictor, and by-participant and by-item intercepts. This model indicated that consonants in onset positions were longer than those in coda positions (position: β = -0.02, t = -4.23, $p < .0001$), consistent with word-initial lengthening effects found in previous research on lab speech (Figure 2). Note that we observed variability among items (SD = 0.04). The lengthening pattern was not consistent across all items. For instance, the relative duration of [b] in *crab eater* was longer than in *cry beater* in many tokens.

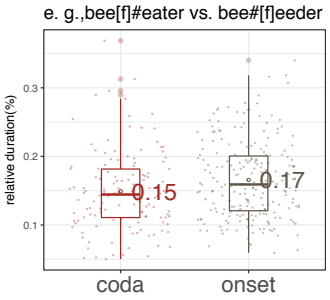

Figure 2: *Pivot consonant at word boundary*.

As mentioned previously, this lengthening pattern might be mainly due to aspiration of voiceless stops in English, especially considering that voiceless stops made up a large

proportion of our data. Therefore, we ran separate models for each type of consonant. The lengthening of word initials was found in the model for voiceless stops (position: β = -0.04, t = -5.38, p < .0001), fricatives (β = -0.02, t = -2.03, p < .05), and nasals (β = -0.03, t = -5.68, p < .0001). However, it was not significant for voiced stops (β = 0.01, t = 1.4, p = .16).

To better understand the difference between voiced and voiceless stops, we examined the durations of closure (cl) and VOT. In Figure 3, the top two plots show the sub-phonemic durational patterns of voiceless stops /p, t, k/; the bottom two plots show the pattern of voiced stops /b, d/. Wilcox tests were used for comparing the mean values between two position groups since the data was not normally distributed. We observed that for voiceless stops, there was no significant difference in closure duration between two position groups (W = 2922, p > .10); however, onset voiceless stops had longer VOT (W = 191, p < .0001). Thus, word-initial lengthening for voiceless stops was arguably an allophonic contrast between aspirated and unaspirated stops. Regarding voiced stops, VOT didn't differ significantly between onsets and codas (W = 358, p > .5); but closure duration was significantly longer when the stop was an onset (W = 455, p < .05).

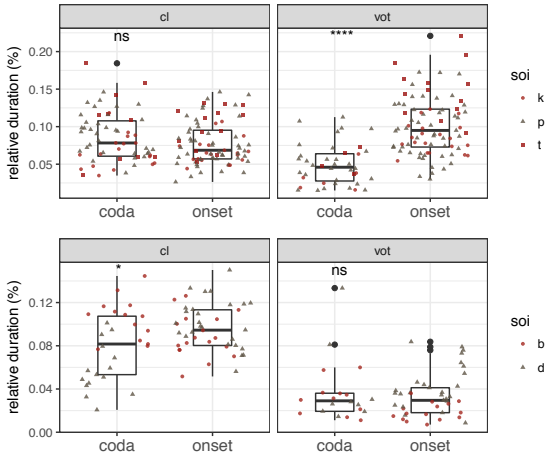

Figure 3: *Sub-phonemic duration patterns of pivot consonant.*

Finally, we ran models with pivot consonant constriction duration (i.e. excluding VOT for stops) as response, and boundary position as the main effect. For the model excluding voiceless stops, boundary position was a significant effect (β = 0.04, t = 3.85, p < .001). For the model including voiceless stops, the effect of position was not significant (β = 0.01, t = 1.8, p = .07).

3.2. Relative duration of the post-juncture section

We observed a lengthening of the post-juncture section when the pivot consonant was a coda, e.g. relative duration of [iɚ] in *beef eater* was longer than [iɚ] in *bee feeder*, as shown in Figure 4. This finding is consistent with word-initial lengthening because there was a word-initial vowel in the section when the pivot consonant was a coda.

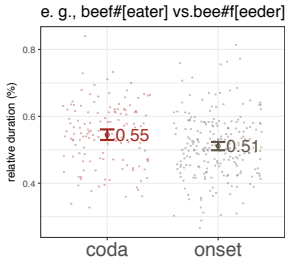

Figure 4: *Post-juncture section.*

However, this finding is also in accordance with another mechanism—poly-segmental shortening, the segmental analog of poly-syllabic shortening, a duration compression effect. There was always one more segment in the post-juncture word when the pivot consonant was in onset position, so that for example, *feeder* has one more segment than *eater*, potentially compressing the relative duration of [iɚ] in *feeder*. To investigate this alternative explanation, we examined our 114 cases with clusters such as *peace talk* vs. *pea stalk*. For instance, if the duration of '*alk*' from *stalk* was shorter than from *talk*, we could argue that there was a pure effect of poly-segmental shortening independent of word-initial lengthening, since the vowel from '*alk*' was not in initial position in either case. However, a statistical comparison suggested no significant difference depending on boundary position for these cases (t = 0.74, p = .50). Thus, while, our sample is relatively small, we did not find strong evidence for the existence of poly-segmental shortening in our study. We argue that the lengthening pattern we found in the post-juncture section was mainly caused by the effect of word-initial lengthening.

This lengthening was statistically significant in a linear mixed-effects model using relative duration of post-juncture section as response, pivot consonant position as the main predictor, and prosodic-phrasing position of the target phrase as a covariate (position: β = 0.12, t = 3.28, p < .01). The model also indicated that this effect of word boundary position exists only when there was an ip/IP boundary to the right of the target phrase (right-ipIP-boundaryT: β = 0.28, t = 7.73, p < .0001). Thus, it may have been due to a boundary related phrase-final lengthening effect expanding the durational space, which enabled us to observe the word-level effect.

3.3. Relative duration of the pre-juncture section

What we would expect from word-final lengthening is longer duration of the pre-juncture section when there was a word-final segment in the section, e.g., [bi:] in *bee feeder* should be longer than [bi:] in *beef eater*. We fit a mixed-effects model using relative duration of pre-juncture section as response, pivot consonant position as the main predictor, and prosodic-phrasing position (whether there was an ip/IP boundary to the left of target phrase) as a covariate. The model indicated that neither pivot consonant position nor prosodic phrasing were statistically significant, though there was a trend for consonant position (position: β = -0.06, t = -1.86, p = .06; left-ipIP-boundaryT: β = 0.031, t = 0.893, p = .37).

3.4. Boundary prediction

In previous sections we have seen that boundary position has a significant influence on the duration of the pivot consonant and the post-juncture section of the phrase, separately. In this

section, we investigated how durational information can assist in the prediction of boundary placement by means of a Random Forest model [29]. This type of model performs classification tasks by using ensemble methods that combine the 'votes' from many individual decision trees. It can also provide an out-of-bag (OOB) error estimate, which is an unbiased estimate bypassing the need of cross-validation, as each tree is constructed using different samples from the data.

We fit a model with the relevant acoustic variables found in previous sections, i.e. relative duration of the pivot consonant, relative duration of the post-juncture section, pivot consonant type, and prosodic-phrasing position (ntree = 1000, mtry = 4). This model achieved an out-of-bag classification accuracy of 64%. In order to assess the contribution of each acoustic cue, we ran a variable importance analysis. This analysis indicated that consonant type was the most important cue (variable importance 0.026); the second important factor was the relative duration of the post-juncture section (0.025); the third important variables were the relative duration of the pivot consonant and prosodic-phrasing position (0.021).

4. Discussion

The experiment investigated whether we observe durational cues to word boundaries in spontaneous speech. Furthermore, we investigated which of two possible mechanisms – word-initial lengthening or word-final lengthening – best explains our data.

4.1. Boundary cues

We found evidence that, even in spontaneous speech, pivot consonants were longer in onset position than coda position. This confirms previous work using lab speech on unscripted, spontaneous speech.

4.2. Word-initial lengthening

The first piece of evidence supporting word-initial lengthening as the source of the juncture effect is the longer duration of the pivot consonant when it was in onset position.

We found that the lengthening of word-initial voiceless stops can mainly be attributed to lengthening of VOT, as there was no closure duration difference between word-initial and word-final consonants. This may be viewed as a case of position-sensitive allophonic variation in English or as initial strengthening, but either way, the word boundary was clearly marked for voiceless stops.

In order to separate this allophonic contrast and durational lengthening in word-initial position, we examined the constriction duration of the other consonants. We found that the closure duration of voiced stops was longer when the stop was an onset than it was a coda. More importantly, similar results were found for fricatives and nasals: the constriction duration was lengthened when the consonant was in word-initial positions. As fricatives and nasals have steady-state acoustic characteristics, and all involve a stable consonantal constriction, there is no doubt that word-initial lengthening existed in these cases.

The second piece of evidence supporting word-initial lengthening lay in the lengthened duration of the post-juncture section when it included a word-initial vowel, e.g. [ɪɚ] from *beef eater* was longer than [ɪɚ] from *bee feeder*. Although poly-segmental shortening could be an alternative explanation, we did not find strong evidence for the existence of this compression effect in our study. As the most well-studied phenomenon of duration compression effect is poly-syllabic shortening, it is possible that the compression effect was subtler on a segmental level. Further work with a bigger sample is needed to confirm the extent to which poly-segmental shortening could be responsible for the differences in post-juncture lengthening.

4.3. Word-final lengthening

We did not find a statistically significant effect of boundary position's influence on the duration of the pre-juncture section. Furthermore, the duration compression effect described above for the post-juncture section could also be a potential explanation for the numerical trend we observed. For instance, the duration of [bi:] in *beef* may have been shorter compared to [bi:] from *bee* because there was one more segment in the word *beef*. We did not have a condition which allowed us to test this more directly however.

4.4. Future work

The OOB estimate provided by our Random Forest model was 64%, suggesting that, while word-initial lengthening may be informative above chance, durational cues alone are not sufficient to segment word boundaries in many instances of spontaneous speech. One issue of the current study is that there was considerable variability among across experimental items. First, due to the unscripted nature of our task, the data was unbalanced. More data is needed to better understand word-specific effects. Furthermore, future research should consider talker familiarity with target phrases as a relevant factor.

5. Conclusions

In the present study, we found evidence supporting word-initial lengthening in spontaneous English. The durational pattern of pivot consonant and post-juncture section both supported the existence of this boundary-related lengthening effect. A model for boundary prediction (*Section* 3.4.) also suggested that word-initial lengthening was an important cue assisting in word-boundary recognition. In addition, we observed that higher prosodic-phrasing structure had a large influence on the realization of the word-level lengthening effect, which is worth further studying. On the other hand, the evidence for word-final lengthening was not strong in our study. However, the fact that we could not find strong evidence of word-final lengthening in spontaneous English in this study does not necessarily lead to the conclusion that this effect does not exist. Our next step will be to run a perception study, aimed at estimating how the extent to which the word-initial lengthening effects identified in this study are perceptually relevant.

6. Acknowledgements

The authors would like to thank our participants for their contribution in this study, thank our research assistant Anastasia Provias for her excellent work. The first author would like to thank Dr. Michael Wagner, Yeong-woo Park, Emily Kellison-Linn and P*reading group members in the Linguistics Department of McGill University for helpful discussions and suggestions.

7. References

[1] J. M. Mcqueen, "Segmentation of Continuous Speech Using Phonotactics," *Journal of Memory and Language*, vol. 39, no. 1, pp. 21–46, 1998.

[2] D. Norris, J. M. McQueen, A. Cutler, and S. Butterfield, "The Possible-Word Constraint in the Segmentation of Continuous Speech," *Cognitive Psychology*, vol. 34, no. 3, pp. 191–243, 1997.

[3] L. A. Taft, *Prosodic constraints and lexical parsing strategies*. University of Massachusetts, Graduate Linguistics Student Association, 1984.

[4] F. C. A. Grosjean and J. P. Gee, "Prosodic structure and spoken word recognition," *Cognition*, vol. 25, no. 1-2, pp. 135–155, 1987.

[5] A. Cutler, and D. Norris, "The role of strong syllables in segmentation for lexical access," *Journal of Experimental Psychology: Human perception and performance*, vol. 14, no.1, pp. 113-121, 1988.

[6] A. Cutler and S. Butterfield, "Rhythmic cues to speech segmentation: Evidence from juncture misperception," *Journal of Memory and Language*, vol. 31, no. 2, pp. 218–236, 1992.

[7] A. Cutler, *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: Mit Press, 2012.

[8] I. Lehiste, "An Acoustic – Phonetic Study of Internal Open Juncture," *Phonetica*, vol. 5, no. 1, pp. 5–54, 1960.

[9] W. M. Christie, "Some cues for syllable juncture perception in English," *The Journal of the Acoustical Society of America*, vol. 55, no. 4, pp. 819–821, 1974.

[10] H. Quené, "Segment durations and accent as cues to word segmentation in Dutch," *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 2027–2035, 1993.

[11] A. E. Turk and S. Shattuck-Hufnagel, "Word-boundary-related duration patterns in English," *Journal of Phonetics*, vol. 28, no. 4, pp. 397–440, 2000.

[12] M. A. Redford and P. Randall, "The role of juncture cues and phonological knowledge in English syllabification judgments," *Journal of Phonetics*, vol. 33, no. 1, pp. 27–46, 2005.

[13] T. Cho, J. M. Mcqueen, and E. A. Cox, "Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English," *Journal of Phonetics*, vol. 35, no. 2, pp. 210–243, 2007.

[14] B. Lindblom, "Explaining Phonetic Variation: A Sketch of the H&H Theory," *Speech Production and Speech Modelling*, pp. 403–439, 1990.

[15] D. Kim, J. D. Stephens, and M. A. Pitt, "How does context play a part in splitting words apart? Production and perception of word boundaries in casual speech," *Journal of Memory and Language*, vol. 66, no. 4, pp. 509–529, 2012.

[16] T. Cho, "Prosodic Boundary Strengthening in the Phonetics-Prosody Interface," *Language and Linguistics Compass*, vol. 10, no. 3, pp. 120–141, 2016.

[17] J. Pierrehumbert and D. Talkin, "Lenition of /h/ and glottal stop," *Gesture, Segment, Prosody*, pp. 90–127.

[18] J. Cole, H. Kim, H. Choi, and M. Hasegawa-Johnson, "Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech," *Journal of Phonetics*, vol. 35, no. 2, pp. 180–209, 2007.

[19] T. Cho and P. Keating, "Effects of initial position versus prominence in English," *Journal of Phonetics*, vol. 37, no. 4, pp. 466–485, 2009.

[20] T. Cho, Y. Lee, and S. Kim, "Prosodic strengthening on the /s/-stop cluster and the phonetic implementation of an allophonic rule in English," *Journal of Phonetics*, vol. 46, pp. 128–146, 2014.

[21] C. C. A. Fougeron and P. A. Keating, "Articulatory strengthening at edges of prosodic domains," *The Journal of the Acoustical Society of America*, vol. 101, no. 6, pp. 3728–3740, 1997.

[22] J. A. Barnes, *Positional neutralization: a phonologization approach to typological patterns*. University of California, Berkeley, CA: Unpublished Ph.D. dissertation, 2002.

[23] I. Lehiste, "The Timing of Utterances and Linguistic Boundaries," *The Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2018–2024, 1972.

[24] A. E. Turk and L. White, "Structural influences on accentual lengthening in English," *Journal of Phonetics*, vol. 27, no. 2, pp. 171–206, 1999.

[25] M. McAuliffe, M. Sonderegger, and M. Wagner, "A system for unified corpus analysis, applied to duration compression effects across 12 languages," in *LabPhon 15, July, Cornell University, United States, poster*, 2016.

[26] P. Menzerath and J. M. de Oleza, *Spanische lautdauer: eine experimentelle Untersuchung, mit 4 Abbildungen. 15 Figuren und 37 Tabellen*, W. de Gruyter & Company, 1928.

[27] M. Mcauliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," Interspeech 2017, 2017.

[28] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1-48, 2015.

[29] C. Strobl, J. Malley and G. Tutz, "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests.," *Psychological Methods*, vol. 14, no. 4, pp. 323, 2009.