The online advantage of repairing metrical structure: Stress shift in pupillometry

Canaan Breiss (cbreiss@ucla.edu)
UCLA Department of Linguistics, 3125 Campbell Hall
Los Angeles, CA 90095 USA

Jesse A. Harris (jharris@humnet.ucla.edu)
UCLA Department of Linguistics, 3125 Campbell Hall
Los Angeles, CA 90095 USA

Amanda Rysling (rysling@ucsc.edu)
UCSC Department of Linguistics, 1156 High Street
Santa Cruz, CA 95064 USA

Abstract

In this paper we use pupillometry, a non-invasive, naturalistic method of measuring attention and cognitive load, to measure the effect of stress clash (*Chinése shíp*) and its metrical repair (*Chínese shíp*) during auditory sentence processing. We addressed two main research questions. The first question explores whether phonologically-disfavored metrical structures induce processing costs indexed by changes in pupil size. The second investigates whether the application of an optional process of stress retraction called the Rhythm Rule (Liberman & Prince, 1977) ameliorates or compounds any general penalty for stress clash. We find that unrepaired stress clash leads to greater pupil diameter relative to non-clashing sequences, indicating increased attention and cognitive load. We also find that repaired sequences lead to a decrease in overall pupil diameter, indicating facilitation.

Keywords: stress clash; rhythm rule; metrical constraints; lexical processing; pupillometry; auditory sentence processing

Introduction

An extensive literature in both formal phonology and speech science more broadly has demonstrated that the prosodic structure of a sentence — its overall intonation, as well as the placement of phrase-level and word-level prominence or stress — impacts its acceptability. A classic example of the role of prosodic well-formedness is Stress Clash (SC) in North American English, which is dispreferred by the phonology of the language. Stress Clash is a prosodic configuration which occurs when two adjacent syllables within a prosodic phrase both bear primary stress, e.g., *maroón sweáter* (Hayes, 1995).

There are multiple ways to resolve SC in English. In one, stress on the first word of a clashing pair may be retracted to a metrically prominent syllable earlier in that same word by an optional metrical repair process often referred to as the Rhythm Rule (RR; Kiparsky, 1966; Liberman & Prince, 1977), as in *Chinése shíp* → *Chínese shíp*. The availability of the RR is predicated on numerous factors, including whether there is a previous syllable that is able to bear stress or not. Another way to resolve SC is to partition the stressed syllables into separate phonological phrases, separated by a prosodic boundary, as in *Chinése # shíp*.

SC is strongly disfavored in English; listeners often repair SC violations perceptually (Kimball & Cole, 2014), and speakers routinely avoid it in spontaneous speech (Hammond, 2016; Breiss & Hayes, 2020). SC is also known to exact a processing cost, with clashing sequences resulting in increased cognitive load (Kentner, 2015).

In this paper, we use pupillometry to examine how stress shift motivated by avoidance of SC, as well as spurious stress shift which lacks phonological motivation, impacts the online auditory processing of sentences containing (possible) instances of SC. We find evidence that unrepaired SC induces a processing penalty, but that SC which has been repaired by RR facilitates processing above and beyond the simple lack of overt SC. We suggest that the exceptional facilitation for repaired SC may arise from participants using the non-citation form stress placement on a pre-nominal adjective (*Chínese*) as an early clue predicting the metrical structure of the upcoming noun (*shíp*), thereby facilitating word recognition. We close with a discussion of the implications of these results for different theoretical models of phonological processing.

Pupillometry

In many respects, pupillometry is ideal for exploring auditory manipulations, as it provides a noninvasive, naturalistic, largely unconscious window into on-line processing. Pupillometry measures changes in pupil diameter in response to changes in the environment or as a function of internal cognitive processes. While the pupil responds most dramatically to changes in luminance, it has also been found to index changes in mental effort and cognitive load. For example, Hess & Polt (1964) observed that pupil size increased when subjects attempted to solve difficult mental arithmetic problems compared to easy ones. Subsequent studies found an association between pupil size and other cognitive functions related to emotional valence, attention, memory, and language (e.g., Laeng et al., 2012, for a review). Linguistic variables reflected in the pupil have been found to include relative accessibility of word forms (Kuchinke et al., 2007; Schmidtke, 2017) and syntactic complexity (Engelhardt et al., 2010; Just & Carpenter, 1993), as well as prosodic (Zellin et al., 2011; Harris et al., 2019) and metrical (Scheepers et al., 2013) violations.

Scheepers et al. (2013) measured pupillary responses to highly expected metrical patterns imposed by limericks, forms of poetry that place strong metrical constraints on the verse (see also Breen & Clifton, 2011, 2013). Metrical violations were created by manipulating a single word before the final word of the poem. They found that pupil size increased in response to violations of metrical structure, but not to violations of syntactic, semantic, or rhyming conventions.

In keeping with the current pupillometry literature, we assume the following basic linking hypothesis between cognition and behavior: increased pupil size indirectly reflects increased cognitive load, including mental effort directed at managing language comprehension processes. The method provides an appealing online measurement of cognitive processing during passive listening. In addition, as pupil size is not under conscious control, the technique is resistant to individual strategies that may be employed by participants in tasks that may involve elements of strategic processing, such as self-paced reading or eye-tracking while reading.

We now turn to a pupillometry experiment that was designed to investigate the effects of stress clash and repair on sentence processing during passive listening.

Hypotheses and predictions

Given the prominent role that prosody is known to play in on-line processing, this study was conducted to explore the following questions:

1. *Research questions:*
   A. Does the pupil index processing costs due to SC?
   B. Does the RR help or hinder processing?

As discussed, violations of metrical expectations imposed by limerick structure have been found to induce processing penalties in studies with auditory stimulus-presentation (Scheepers et al., 2013), as well as during silent reading (Breen & Clifton, 2011, 2013). However, the strictness of the metrical constraints imposed by limericks are extreme and likely do not reflect the day-to-day auditory experience of most listeners. Therefore, it is not certain that the more subtle ill-formedness due to SC will be indexed by pupillary response. This motivates research question (1), since our study uses more naturalistic sequences of stress placement to assess how listeners respond to stress clash and its repair.

If the pupil does index SC violations, we explore three central possibilities in response to (1B).

2. *Possible hypotheses addressing (1B):*
   A. **Inhibition:** Non-citation form pronunciation of adjectives due to RR could hinder lexical access.
   B. **Facilitation:** Non-citation form adjective stress could constrain possibilities for upcoming metrical structure, allowing the early formation of lexical predictions.
   C. **Facilitation:** Earlier adjective stress could cue earlier recognition of a word boundary, facilitating word segmentation.

A scenario (2A) in which RR application hinders processing could arise from the non-citation form of the adjective undergoing stress retraction (*Chínese*) which could, in principle, inhibit lexical access and consequently lead to a more effortful parse of the sentence. On the other hand, encountering the non-citation form of a pre-nomimal adjective resulting from stress shift could facilitate processing in two ways.

First (2B), stress retraction in a pre-nominal adjective could act as an early cue to the stress-initial prosodic structure of the upcoming noun, thus facilitating word recognition and speeding up subsequent processing at the noun by both narrowing the space of possible nouns and possibly speeding word segmentation due to anticipating a primary stress (Cutler & Norris, 1988). Second (2C), early phonetic cues to primary stress in *Chínese* may enable earlier word recognition, due to easier word segmentation at that word's edge (Cutler & Norris, 1988) compared to canonical *Chinése*. In this case, we would expect that the RR would facilitate lexical processing of the adjective, but not necessarily the noun.

Experiment

Participants

18 undergraduates from the University of California, Los Angeles participated in the experiment. No participants reported a history of hearing loss or language disorders, and all reported being monolingual native listeners of North American English. Experimental sessions lasted approximately 45 minutes, and participants were compensated with course credit.

Materials and method

20 sentence quartets were constructed as shown in Table 1. The 2×2 design crossed a pre-nominal adjective and following noun's citation form stress placements (Clash, NoClash) with whether the RR applied (Shift, NoShift). The nouns in the critical adjective-noun pairs were identical within a quartet. Adjectives were all bisyllabic and monomorphemic. Thus, the stress of the citation form of an adjective was in either final (Clash; *Chinése*) or initial (NoClash; *Spánish*) position. The application of a stress shift to an adjective of the Clash type was motivated under the RR (*Chínese shíp*), whereas applying stress shift to a NoClash adjective (*Spánish*) is unmotivated by the RR (*Spanísh shíp*).

Adjectives with similar meanings were selected for the manipulation. Pairs were matched for adjective frequency and phonological neighborhood density using the English Lexicon Project database (Balota et al., 2007), and critical adjective-noun bigram frequency using the *English One Million* subset of the Google Books corpus (Lin et al., 2012). In addition, we conducted an offline ratings experiment to select the quartets of stimuli whose critical adjective-noun pairs were sufficiently similar in naturalness given the context of the carrier sentence. In other words, the rating study was designed to select quartets where there were minimal baseline differences in naturalness between the critical adjective-noun pairs, e.g., *Spanish ship* and *Chinese ship*, in the context of the carrier sentence, regardless of overt stress placement.

Forty-six subjects recruited from the same population as the main experiment completed an off-line ratings task, where they were asked to judge the naturalness of written Clash and No Clash sentences using a 7-point Likert scale (1 = completely unnatural, 7 = completely natural). From a starting set of 24 candidate quartets, we selected the 20 in which the critical adjective-noun pairs differed the least in rating. Ratings on the final set of stimuli were analyzed with a Bayesian mixed effects ordinal regression model fit in `brms` (Bürkner, 2017) with by-subject and by-item random intercepts, and a sum-coded fixed effect of Condition (Clash = -1, No Clash = 1), with a random slope of Condition by subject. The model also included a by-subject random intercept for the dispersion parameter of the ordinal model. The analysis revealed that the central 95% of values for Condition included zero, indicating no strong support for a difference in naturalness between Conditions. An example quartet is shown in Table 1.

Sentences were recorded in a sound-attenuated room by a female phonetically-trained native speaker of American English, who recorded the full sentences shown in Table 1. To ensure maximum acoustic similarity between different conditions within same quartet, we used the NoClash, NoShift condition (as in *Spánish shíp*) as a base frame, into which we spliced in the critical adjective-noun pairs from the other three conditions using Praat (Boersma & Weenink, n.d.) in order to yield the other three items in each stimulus quartet. Following the critical adjective-noun pair, we inserted 50ms of computer-generated silence to serve as a stable baseline for pupil size, and then normalized the audio intensity of the whole sentence to 70 dB. The materials were divided into four counter-balanced lists, interspersed with 58 filler items for a total of 78 trials per list. To ensure thorough processing of the sentences, participants answered a two-alternative forced-choice comprehension question after half of the trials.

The experiment was programmed and presented with Experiment Builder (SR Research). Subjects listened to sentences played over sound-isolating headphones in a dimly lit room dedicated to experimentation. Eye position and pupil area were recorded using an SR Research EyeLink 1000 Plus eye tracker sampling at 250Hz. Subjects' heads were stabilized with a chin rest attached to a tracker mounted to the table 55cm from a 27 inch LCD monitor with a light gray background to minimize light exposure from the screen. A 5-point calibration procedure was used before recording as necessary, and drift correction was conducted after every trial.

At the start of each trial, subjects fixated a cross at the center of the screen. They were instructed to avoid blinking for the duration of the sentence. They were encouraged to blink between trials, and to rest their eyes as needed. In order to minimize the possibility of eye blinks due to fatigue, participants initiated the trial when they were ready to begin. Calibration was also performed after breaks and as needed.

Experimental predictions

If SC induces a processing cost that is indexed by pupil dilation, we expect to observe increased pupil diameter in response to the Clash NoShift (*Chinése shíp*) and NoClash Shift (*Spanísh shíp*) conditions relative to the NoClash NoShift (*Spánish shíp*) condition. Further, if shift facilitates lexical processing, pupil size should decrease following the Clash Shift (*Chínese shíp*) condition as compared to a NoClash NoShift baseline. In contrast, if RR-induced stress shift hinders processing, pupil size should increase in this condition. The spurious application of a stress shift in the NoClash Shift (*Spanísh shíp*) condition was intended as a methodological check to determine whether the pupil would respond to metrical anomaly, regardless of the central manipulation.

Results

On average, subjects answered comprehension questions correctly in 89% of trials. No differences in responses between conditions were observed in a mixed effects logistic regression model with by-subject and by-item random intercepts.

Pupil size data were preprocessed using PupilPre (Kyröläinen et al., 2019) following the procedure outlined in Mathôt et al. (2018). Fixations outside of the fixation cross area were excluded. Blinks and other artefacts were removed from the data, along with 200ms on either side of the event. The missing data was interpolated with spine smoothing. Each event was examined individually and corrected by hand if needed. A baseline of 50ms was established for each event, corresponding to the interval of inserted computer-generated silence. Observations of pupil size following the baseline were calculated by subtracting the value of each observation from the mean pupil diameter during each trial's baseline. Data were then downsampled to 25Hz in order to reduce autocorrelation between points (van Rij et al., 2019).

The remaining analyses were conducted in R (R Core Team, 2020). Data were fit with a growth curve model (Mirman, 2016), which avoids assuming a linear form or an arbitrary time window for analysis. Growth curve models have been used to quantify continuous changes in pupillary response (e.g., Kuchinsky et al., 2013; McGarrigle et al., 2017; Harris et al., 2019). Experimental predictor variables were coded with treatment contrasts, with the NoClash and NoShift conditions specified as the reference levels for their respective factors. We used orthogonal polynomials (ot1–ot3) to further reduce multicollinearity between neighboring samples in the time series. Multiple models of the data were computed, starting with a third-order cubic model with fixed effects of Clash, Shift, and their interaction on polynomial terms, and by-subject and by-item random intercepts.

This model was compared with two simpler models, consisting of either quadratic and linear polynomials and the same fixed and random effect structures as before. The best fitting model included quadratic polynomial predictors and is reported in Table 2. The left panel of Figure 1 shows the model fits on the mean percent change for 2000ms from the

Citation form stress pattern	Stress realization *No Shift*	*Shift*
Clash	As the crowd cheered, the **Chinése** <u>shíp</u> // from across the ocean pulled into the harbor.	As the crowd cheered, the **Chínese** <u>shíp</u> // from across the ocean pulled into the harbor.
No Clash	As the crowd cheered, the **Spánish** <u>shíp</u> // from across the ocean pulled into the harbor.	As the crowd cheered, the **Spanísh** <u>shíp</u> // from across the ocean pulled into the harbor.

Table 1: Sample quartet item crossing Citation form stress pattern (Clash, No clash) and Stress realization (Shift, No shift). The <u>underlining</u> identifies the critical adjective-noun pair. Material following the // symbol (*from across the ocean pulled into the harbor*) was spliced into the recording from the NoClash NoShift condition following 50ms of computer generated silence, and was thus acoustically identical across the items within each quartet. All items are provided in the Appendix.

offset of the critical noun (*shíp*). The right panel illustrates the overall mean percent change in pupillary response during the same time window.

```
lmer(Pupil ~ (ot1+ot2+ot3) * # Orthogonal
                               polynomials
    Clash*Shift +            # Fixed effects
    (1|Subject) + (1|Item), # Random effects
    control = lmerControl(optimizer="bobyqa"),
    REML = F)
```

Parameters	Estimate	SE	t-value	
(Intercept)	-7.12	6.95	-1.02	
Linear	-23.23	14.22	-1.63	
Quadratic	7.58	14.22	0.53	
Clash	8.79	2.91	3.02	***
Linear: Clash	-4.99	19.73	-0.25	
Quad: Clash	-40.10	19.73	-2.03	*
Shift	12.99	2.91	4.47	***
Linear: Shift	3.86	20.11	0.19	
Quad: Shift	-38.78	20.11	-1.93	+
Clash × Shift	-22.63	4.12	-5.49	***
Linear: Clash × Shift	-33.50	28.08	-1.19	
Quad: Clash × Shift	64.83	28.08	2.31	*

Table 2: Linear mixed effects regression model of a growth curve analysis with two orthogonal polynomial terms (Linear; Quadratic). Significant terms demarcated with '***' indicating $p < .001$, '*' indicating $p < .05$, and '+' indicating $p = .05$.

We adopt Kuchinsky et al.'s (2013) interpretation of coefficient terms in a growth curve model of pupil size data. Larger positive intercepts reflect a greater overall response from the pupil, larger linear terms indicate a steeper deflection from a hypothetical flat line, more negative quadratic terms are associated with increased non-linearity, and more positive cubic terms reflect a more steep rise and fall from the peak.

Effect of citation form clash Sentences containing adjectives with a word final stress in their citation forms (*Chinése*) were associated with increased pupil size overall (Clash intercept: $t = 8.79, p < .001$), and exhibited a more linear curvature (Quadratic: Clash, $t = -2.03, p < .05$).

Effect of stress shift Sentences with a stress shift (*Chínese shíp, Spanísh shíp*) induced greater pupillary responses in listeners than sentences without stress shift. Shifting was found to elicit increased pupil size overall, as indicated by the significant coefficient on the Shift intercept, $t = 2.91, p < .01$, and marginally interacted with the Quadratic polynomial, producing a curve with a less linear form, $t = -1.93, p = .05$).

Clash × Shift interaction Crucially, these main effects were moderated by an interaction showing a differential benefit for stress shift on adjectives with word-final stress, reducing the overall pupil excursions (Clash×Shift interaction: $t = -5.49, p < .001$) and a sharper, more transient peak (Quadratic: Clash×Shift, $t = 2.31, p < .05$). The pattern suggests that RR application to avoid a clash (*Chínese shíp*) facilitated, rather than inhibited, speech processing.

Discussion

Our findings indicate that pupillary response is sensitive to violations of preferred metrical structure during auditory processing (1A). RR application that avoided a SC appeared to alleviate any cost of shifting stress. But this is not to say that deviations from citation forms are necessarily cost free. As indicated by the penalty in the NoClash Shift condition (*Spanísh shíp*), shifting stress facilitated processing only when it was motivated by general metrical well-formedness pressures.

Our results cast doubt on the hypothesis that a metrically-licensed application of the RR exacts a cognitive cost (2A). By extension, it is unlikely that the means for resolving metrical violations would be associated with a cognitive process penalizing deviations from citation forms. We instead take these results to be compatible with one of the hypotheses that stress shift may facilitate auditory processing (2B–C).

On one variant (2C), placing main stress earlier in the adjective allows word recognition to occur faster than for coun-

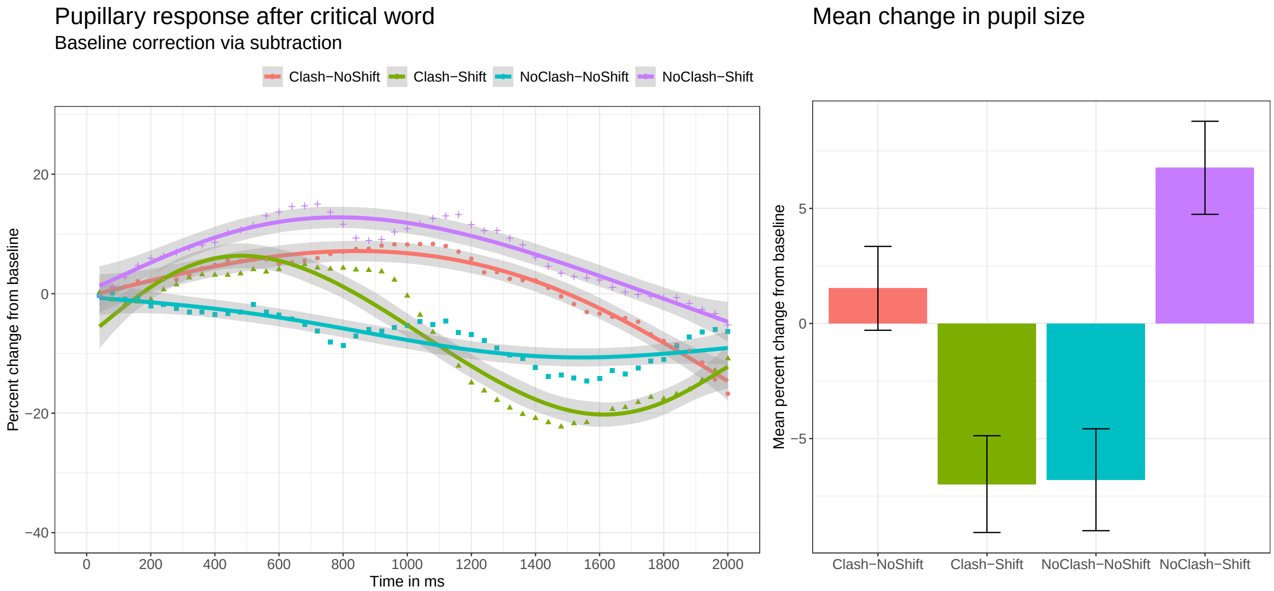

Figure 1: Percent change from baseline in pupillary response from critical word (left); mean and standard error of percent change in pupil size during 2000ms window (right).

terparts with later stress. This proposal was motivated by research finding that syllables bearing main stress facilitate speech segmentation processes (Cutler & Norris, 1988), because native English listeners use word-initial stress to identify the beginnings of words. However, this explanation for the present results seems insufficient on its own, as the NoClash NoShift (*Spánish shíp*) condition, in which adjectives always began with word-initial stress, was not apparently as easy to process as the Clash Shift condition.

The other variant (2B) proposed that stress shift facilitated word recognition by constraining a listener's expectations about the metrical structures of the following word. Under this account, if a native listener of American English encounters a non-citation form of a bisyllabic adjective like the ones used in our Clash condition, then their experience of their native language is highly consistent with a stressed syllable at the beginning of the next word that they will hear. Indeed, it is difficult to imagine an alternative reason for which a stress shift would have occurred in this case. We suggest that native American English listeners have learned to use stress retraction / non-citation forms as a cue to upcoming primary stress in an immediately following word. When a prediction based on this cue was fulfilled, they would have experienced the facilitation that we observed in the Clash Shift condition.

Although not conclusive, the present evidence supports the prediction variant (2B) over the other (2C). This account is compatible with Cutler's (2012) observation that while native English listeners do not need to use stress for *identifying* words, word-initial stress is nonetheless useful in locating word *boundaries* during segmentation. In English, the stressed versus unstressed distinction covaries perfectly with whether the vowels and consonants in a syllable are fully or hypo-articulated. It is possible that we observed a benefit of motivated stress shift in the Clash Shift condition relative to the baseline NoClash NoShift condition precisely because stress in English is not used for identifying a word itself, but only for finding a potential word boundary faster. But in other languages, vowels and consonants are realized in their full forms in unstressed syllables. In a language in which (i) stress is a useful cue for word identification *and* (ii) stress does not covary with segmental articulation, stress shift might be costly to process even when motivated by the RR, as shifting would disrupt a cue that is more fundamental to word identification.

In future studies on English, we hope to test the two hypotheses about word recognition more directly by comparing the effect of pre-nominal stress shift on bisyllabic nouns with initial (*káyak*) or final stress (*canóe*) placement. If stress shift on an adjective constrains expectations for the metrical structure of the following word, then metrically-motivated applications of the RR should facilitate word recognition of nouns with initial stress (*Chínese káyak*), but not of nouns with final stress (*Chínese canóe*).

The present results may also bear on current issues in phonological theory, which concern the trade-off between lexical listing and grammatical derivation in the representation of words in the lexicon. For illustration, it is useful to consider two kinds accounts of how word recognition might reference phonological knowledge: phonological inferencing and exemplar matching (see Farris-Trimble & Tessier, 2019, for an overview). *Phonological inferencing* accounts assume that lexical items are associated with one underlying word form, which must be recovered from a realized form in online word recognition. This is a process-oriented view, similar to rule-based models of phonology (e.g., Chomsky & Halle, 1968, i.a.), which posits a cost for undoing each successive phonological process that has been applied to a perceived word form. This model predicts that non-citation variants like *Chínese* would be costly to process in all contexts, since only citation forms *Chinése* are stored in the lexicon.

In contrast, *exemplar matching* accounts hold that context-specific, non-citation forms are stored side-by-side with canonical forms (*Chinése* $\sim$ *Chínese*). They predict that the cognitive effort of accessing lexical variants should be equivalent, modulo any differences in the frequency of each realization. Assuming that the stress-initial form *Chínese* is less frequent than the citation form *Chinése*, the non-citation form should induce a penalty compared to the latter, regardless of metrical context. However, we observed that a non-citation form was associated with a reduction in processing cost just in the case that it was motivated by stress clash avoidance.

Neither account in its unaugmented form seems, at present, to be immediately compatible with our findings. But we note that other psycholinguistic work on speech processing by Cutler and colleagues (see Cutler, 2012, for an overview) has demonstrated that not all differences in frequency of occurrence necessarily result in differences in listeners' behaviors. It could thus be the case that the non-citation forms in our Clash conditions are frequent enough that their use is not costly, even if they are not as frequent as their citation form counterparts. Thus, it seems that the exemplar matching account might be more easily augmented in light of our results. Alternatively, it could be the case that the processing cost predicted by both accounts is in fact exacted by the non-canonical form of the pre-nominal adjective, but that the penalty is too short lived to be detected in the relatively slow-moving pupillary response. In addition, any cost introduced by a non-citation form could have fleeting effects, rendering it undetectable in the presence of a larger word segmentation benefit associated with stress-initial adjectives.

Finally, we note that the stimulus construction method – splicing adjective-noun pairs from other conditions into the base recording of the NoClash NoShift (*Spánish shíp*) condition — may have created differences in auditory naturalness across the conditions of the present study. This seems unlikely to have confounded our results, because the Clash Shift condition (*Chínese shíp*) was facilitated even relative to the NoClash NoShift condition. While it seems unlikely that the splicing scheme caused the apparent facilitation, future studies may be required to address this possibility.

Conclusion

This study investigated the interplay between stress clash and its repair in online processing using pupillometry. We provided initial support for the claim that the pupil is sensitive to the prosodic ill-formedness of SC. As our findings were observed in speech unconstrained by strong metrical expectations, the results extend those observed in previous literature (Scheepers et al., 2013). We also found evidence that metrically-motivated application of the RR was associated with facilitated processing. We attributed the facilitation to an advantage of metrical information in the process of word recognition, and speculated that listeners used the non-citation form of an adjective to narrow the space of possible upcoming nouns to ones which would motivate application of the RR. Further research is needed, both with speakers of English to further support the claims made here, and with speakers of other languages, especially ones in which stress and hypo-articulation do not completely covary, to establish the validity of our interpretation of the role of stress and segmental reduction. We hope that this study can act as a starting point for such work, having demonstrated that pupillometry presents a promising method for exploring how metrical information guides auditory processing in real time.

Acknowledgments

Thanks to the undergraduate research assistants at the UCLA Language Processing Lab and lab manager Angelica Pan for helping run participants, and to Maura O'Leary for recording stimuli. This work was supported in part by NSF Graduate Research Fellowship DGE-1650604 to Canaan Breiss and by a UCLA Faculty Senate Grant to Jesse Harris.

References

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... Treiman, R. (2007). The english lexicon project. *Beh Res Meth*, *39*(3), 445–459.

Boersma, P., & Weenink, D. (n.d.). *Praat.* Retrieved from `https://praat.org`

Breen, M., & Clifton, C., Jr. (2011). Stress matters: Effects of anticipated lexical stress on silent reading. *JML*, *64*, 153–170.

Breen, M., & Clifton, C., Jr. (2013). Stress matters revisited: A boundary change experiment. *QJEP*, *66*, 1896–1909.

Breiss, C., & Hayes, B. (2020). Phonological markedness effects in sentence formation. *Language*, *96*(2), 338–370.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models. *J. of Stat. Software*, *80*(1), 1–28.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. MIT Press.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. MIT Press.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *JEP:HPP*, *14*, 113.

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *QJEP*, *63*(4), 639–645.

Farris-Trimble, A., & Tessier, A.-M. (2019). The effect of allophonic processes on word recognition. *Language*, *95*(1), e136–e160.

Hammond, M. (2016). Input optimisation. *Phonology*, *33*(3), 459–491.

Harris, J. A., Kaps, M., & Lawn, A. (2019). Investigating sound and structure in concert: A pupillometry study of relative clause attachment. *CogSci*, *41*, 1880–1886.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. University of Chicago Press.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Jrnl of Exp Psych*, *47*(2), 310–339.

Kentner, G. (2015). Stress clash hampers processing of noncanonical structures in reading. *Rhythm in Cognition and Grammar*, 111–135.

Kimball, A. E., & Cole, J. (2014). Avoidance of stress clash in perception of conversational american english. In *Proceedings of speech prosody* (Vol. 7, pp. 497–501).

Kiparsky, P. (1966). *Uber den deutschen akzent. studia grammatica 7: Untersuchungen über akzent und intonation im deutschen, 69–98.* New York: Akademie Verlag.

Kuchinke, L., Võ, M. L.-H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *Int Jrnl Psychophysiology*, *65*, 132–140.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden Jr, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*, 23–34.

Kyröläinen, A.-J., Porretta, V., van Rij, J., & Järvikivi, J. (2019). *PupilPre: Tools for preprocessing pupil size data.* (Version 0.6.2, updated 2020-03-08)

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Persp on Psych Sci*, *7*(1), 18–27.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *LI*, *8*(2), 249–336.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google books Ngram corpus. In *Assn for comp ling* (pp. 169–174).

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Beh. Res. Meth.*, *50*(1), 94–106.

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology*, *54*, 193–203.

Mirman, D. (2016). *Growth curve analysis and visualization using R*. CRC Press.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Scheepers, C., Mohr, S., Fischer, M. H., & Roberts, A. M. (2013). Listening to limericks. *PLoS One*, *8*(9), e74986.

Schmidtke. (2017). Pupillometry in linguistic research: An introduction and review for second language researchers. *Stud in Sec Lang Acq*, 1–21.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trnds in Hrng*, *23*, 1–22.

Zellin, M., Pannekamp, A., Toepel, U., & van der Meer, E. (2011). In the eye of the listener: Pupil dilation elucidates discourse processing. *Intl Jrnl of Psyphys*, *81*(3), 133–141.

Appendix: stimuli

Critical adjective-noun sequence underlined, with the clashing adjective and non-clashing adjective separated by a tilde ($\sim$); // indicates 50ms of computer-generated inserted silence.

1. On the way home, the <u>thirteen ∼ thrifty men</u> // from the construction-site stopped at the dollar store.
2. Just as the party ended, the <u>unkind ∼ spiteful guest</u> // from the next town over criticized the host's hair.
3. Despite the high cost, the <u>unkind ∼ stingy boss</u> // from the new management didn't cover the meal.
4. As the car passed, <u>fourteen ∼ countless dogs</u> // by the side of the road started howling loudly.
5. Despite investors' worries, <u>robust ∼ sturdy growth</u> // from the economy kept recession at bay.
6. Three years after the disaster, the <u>intact ∼ stable house</u> // below the old dam site finally went on sale.
7. As the crowd cheered, the <u>Chinese ∼ Spanish ship</u> // from across the ocean pulled into the harbor.
8. As the dorm door opened, the <u>unclean ∼ pungent smell</u> // from the old pizza box met the janitor's nose.
9. On a bright spring day, the <u>uncouth ∼ rustic lad</u> // from the small town arrived in the city.
10. As the dancers twirled, a(n) <u>insane ∼ crazy thought</u> // for a new, bold costume entered Angela's mind.
11. Late in the evening, the <u>complex ∼ easy task</u> // for his demanding boss seemed insurmountable.
12. Hearing the barking dog, the <u>innate ∼ native sense</u> // for protection in heights drove squirrels up the tree.
13. As the salespeople left the room, the <u>compact ∼ bundled notes</u> // from the boring meeting were thrown in the trash can.
14. Although much money was paid, the <u>unreal ∼ foolish dream</u> // to run a dog-daycare was never realized.
15. As the thief left the crime scene, a <u>covert ∼ furtive glance</u> // behind him up the road ensured him he was safe.
16. Hikers need to pay attention; a(n) <u>unsure ∼ hasty step</u> // through such dense underbrush could send you tumbling down.
17. Across the decades, the <u>convex ∼ arching roof</u> // of the old cathedral has been weathered by rain.
18. At the end-of-year dance, the <u>sixteen ∼ sixty girls</u> // by the water cooler discussed their summer plans.
19. At the end of the week, <u>eighteen ∼ extra shirts</u> // from the town's spare-clothes-drive were given to the poor.
20. Every morning after the cows are milked, the <u>fifteen ∼ fifty cats</u> // by the dairy farm drink the left-over milk.