Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Synchronous speech and semantic incongruity: what do outliers tell us about it?

Verônica Penteado Siqueira[1], *Beatriz Raposo de Medeiros*[2]

[1]University of São Paulo, Brazil
[2]University of São Paulo, Brazil
veronica.siqueira@usp.br, biarm@usp.br

Abstract

Synchronous speech, considered to be an easily performed task, is investigated in two experimental conditions designated as original (OC) and altered (AC), with focus on the outliers' behavior. The hypothesis raised is that AC, which offers semantic incongruities, would lead individuals to a poorer synchronization performance, i.e, producing greater lag duration. Divided equally in two groups (A and B), 24 dyads were recorded reading two fables in Brazilian Portuguese, in both original and altered conditions. Asynchrony duration was obtained by extracting the lag between vowel onsets, after aligning speakers' waveforms in each dyad. Considering results related to the entire dataset, speakers are able to synchronize in both conditions (OC and AC). However, a great number of outliers was observed throughout the dataset. Its distribution in AC is significantly different from the distribution in OC, the former showing greater values for both variance and mean. In this exploratory study, one promising explanation for these results will be discussed taking into account aspects such as the outliers' location throughout the text. These initial results prompt further investigation, in order to verify a more accurate relation between the outliers' duration and the semantic incongruities' place of occurrence.

Index Terms: synchronous speech, semantic incongruity, outliers

1. Introduction

Synchronization is a phenomenon observed throughout nature, animals, and humans alike, seen in the rhythmic flashing of fireflies as well as the synchronous matching of human dancing and musical beat [1], [2]. Synchronization normally requires a level of coordination so as to lead moving parts to a state of interaction and mutual influence [3].

Synchronized speech happens when two or more individuals speak the same text at the same time, i.e., in unison, and the same can be said about joint speech, which was first characterized as such and investigated in the works of Cummins [4]–[6]. It is worth highlighting that in an experimental environment, where two people are asked to read in unison, we call it "synchronous speech" [6]. Although exposed to interferences such as noise in the acoustic signal, different types of texts and different musical genres, speakers synchronize easily and temporal delay is, on average, 40 milliseconds [7]–[9].

This paper presents an experiment aiming to understand if the ability of speaking in synchrony can be influenced by the pragmatic-semantic level of language. More specifically, we observed the interplay between synchronized speech production, and semantic incongruities artificially introduced in the text.

For this study, we considered that a semantic violation in a given discourse would cause a disturbance in the expectation an individual has in relation to its meaning. The idea that we create expectancies based on previous information or stimuli is known in perceptual studies as "the oddball paradigm" [10]–[12]. Kutas and Hillyard [13] started applying this idea to language studies, with the aid of electroencephalogram (EEG) to record brain activity. It was noticed that individuals will create an expectancy of what is about to come in a text based on the context and the words used before. Likewise, if that expectancy is violated, that is, if we encounter an unexpected word that might seem nonsensical in the context, it is registered as a change in our brain activity. More specifically, these works, e.g., [13]–[15] observed the effect of semantic incongruities in brain activity. In other words, when reading a sentence such as "For breakfast, I have bread and …", you would expect a word like "butter"; however, if the sentence finishes with a word incongruous to the semantic context, like "shoes", there will be a change in the processing of that sentence.

Our question is: Can the task of synchronous speech be affected by the reading of a text with semantic incongruities? If the answer is positive, we expect to find greater differences between two text conditions related to synchronization, which would be a result of this "disruptive task".

The contribution that this paper aims to make in the language prosody domain is that synchronicity in speech reveals our rhythmic ability to align, for example, vowel onsets and phrases endings, when speaking simultaneously the same text with another individual. Any element that would cause disruption to this entrained interaction would influence prosodic performance.

The study rationale is to argue in favor of a dynamicist approach of speech production [16]. In a traditional linguistic view, speech production has been seen as a physical phenomenon that needs a translation to mental units [17], thus underpinning a theoretical dualism to explain all linguistic facts, including the prosodic ones. We then assume that, for the dualistic view, speech components should be decoded only through low level cognitive processes. This being true, there would not be any influence of text semantic content to the act of speaking. We propose that the interplay between the semantic-pragmatic level and the speech level could be observed in a task such as the synchronous speech task. Thus, we expect to see the interference of text meaning on speech synchronization as a result of the interaction between these two cognitive levels.

10.21437/SpeechProsody.2022-42

2. Experimental procedures

2.1. Participants, recordings and selected texts

Fifty-two Brazilian Portuguese native speakers (26 women and 26 men), forming 26 pairs (mixed-sex dyads), and divided into two groups of 13 pairs were recorded at Lafalin, the laboratory of phonetics at the University of São Paulo, Brazil. Recordings were done in an acoustic isolated booth, using a BR-800 Digital Recorder and two SM10A headset microphones, in two different channels, with a sample rate of 48 kHz and a 16-bit resolution. All participants were undergraduate and graduate students, ages between 18 and 31 (M = 21.7, SD = 2.7). They read the texts in synchrony, facing each other, at about 1.5 meter of distance. Dataset is available at https://lafalin.fflch.usp.br/.

Two Aesop fables in Brazilian Portuguese were chosen to be used in the reading task, plus a third one used to divert participants' attention. All texts have similar length (about 100 words), syntactic structure and narrative elements. Semantic incongruities were inserted in each of the two chosen fables, *O vento sul e o sol* (Text 1: "The North Wind and the Sun") and *A reunião geral dos ratos* (Text 2: "Belling the Cat"), so that an altered version was obtained. This was done based on previous experiments [13] and taking into account phonological, morphological and syntactic properties, as well as features such as concreteness or abstractness.

The experiment dependent variable is the asynchrony duration between speakers, i.e., the lag. The independent variable we are interested in is the condition, a fixed factor with two levels: the original condition (OC), in which the participants read a fable in its original version; and the altered condition (AC), with semantic incongruities.

In order to avoid a pair reading the same text in both conditions, and thus learning the test purpose, we distributed them to two different groups of participants. Group A read the original version of Text 1 and the altered version of Text 2; group B read the original version of Text 2 and the altered version of Text 1. No repetitions were made and the order in which the texts were read was random.

2.2. Obtaining lag duration

In order to obtain lag durations, data preparation involved semi-automatic procedures that will be briefly described here and whose details are found in [18]. Firstly, recordings were divided into files of about 10 seconds and the channels were split into two. The script BeatExtractor [19] was run to segment sentences into VV units (vowel-to-vowel syllables). This script bounds units from the onset of one vowel until the onset of the following vowel, based on the notion of perceptual center [20]. After this segmentation, the files were joined back together into two channels. The VV units' boundaries were then checked, manually corrected, when necessary, and labeled. This was done in software Praat [21].

However, the VV unit is not the measure we are interested in; rather, the measure is the difference between the onsets of the same units produced by both speakers. After aligning the sound waves, the function List (in Tabulate), in Praat, provides the boundary times from both channels. The lag is then obtained by subtracting the time of each boundary marking a vowel onset in both channels.

The statistical analysis was done in R [22]. The lag distribution in the two experimental conditions was analyzed through descriptive statistics and normality tests, followed by a Fligner-Killeen test to compare variances for non-normal distributions, and a non-parametric Wilcoxon Rank Sum test for independent samples.

3. Results

3.1. Full dataset and variances difference

Firstly, we observed the entire dataset distribution, comparing lag durations obtained at each speech condition. Groups A and B were analyzed together as experimental conditions were similar for both. Pairs were treated as a random factor for this analysis. From the 26 pairs recorded, one pair from group A (pair A7) and one pair from group B (pair B13) were discarded, as they showed difficulties performing the task.

Also taken out from the analysis were the durations equal to 0 milliseconds, as they represent perfect synchrony between the speakers. This was done for three reasons: we were interested in durations that would represent a delay between the speakers; secondly, the logarithmic transformation applied to the lag duration required values to be greater than 0; and thirdly, zeros represented less than 1% of the data (74 occurrences). It is worth mentioning that the same analysis was done to the dataset including zeros, and the results were the same as those reported here: a mean of 43.4 ms in the OC (SD = 50.3) and a mean of 48.1 ms in the AC (SD = 67.4).

A total of 8012 lags was analyzed. We found higher numbers for mean and standard deviation in AC: a mean of 43.7 milliseconds (hence, ms) in OC and 48.5 ms in AC; and a standard deviation of 50.4 ms for OC and 67.7 ms for AC. Although the difference between mean lag duration is small, there is a considerable difference between the values for standard deviation. The distribution is shown in Figure 1.

A Shapiro-Wilk test indicated that the samples were not normally distributed, so a non-parametric test was used. A Fligner-Killeen test showed that OC and AC variances are different (med chi-squared = 6.3, $p = .01 < .05$). It seems that the pairs were able to synchronize in a similar way in both conditions, which may account for similar mean durations. However, the variances are significantly different, suggesting that the pairs might have produced more and greater lags in AC. In addition to the difference between variances, we have to take into account that outliers were produced in both conditions. These observations led us to the main analysis in this paper: the outliers in synchronous speech.

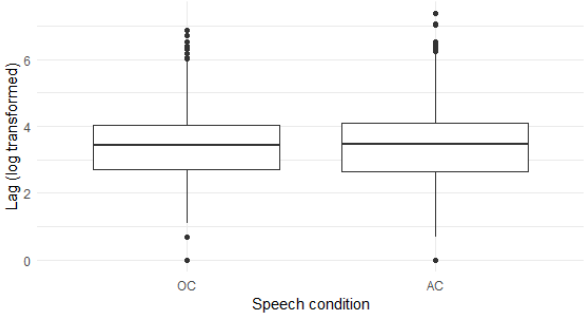

Figure 1. *Boxplot of full dataset distribution with lag log-transformed, in original condition (OC) and altered condition (AC).*

3.2. Outliers: do they differ between speech conditions?

Initially, outliers were classified using the interquartile range criterion (IQR), which allows detecting observations higher than 1.5 times the third quartile (representing 75% of the data) or lower than 1.5 times the first quartile (representing 25% of the data). There are 245 outliers in the OC subset and 263 outliers in the AC subset. Figure 2 shows a positively skewed distribution for both conditions, however, there is a longer and heavier tail in AC, while the OC data are more concentrated on the left.

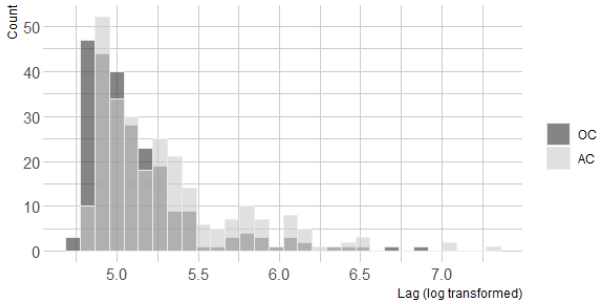

Figure 2. *Histogram of outliers' distribution with lag log-transformed, in original condition (OC) and altered condition (AC).*

A Q-Q plot, in Figure 3, shows the heavier tail to the right. There is a clear departure from the 45-degree reference line, with a longer till to its right, which confirms the positively skewed distribution and suggests that both OC and AC set of quantile values do not come from the same distribution.

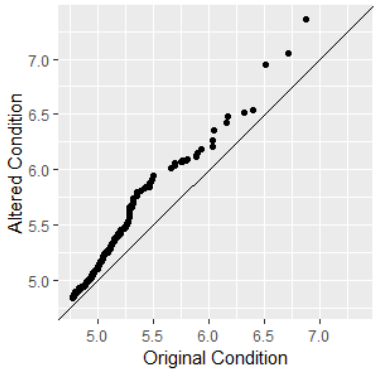

Figure 3. *Q-Q plot (quantile-quantile plot) of the distributions for OC (original condition), in the x axis, and AC (altered condition), in the y axis. Lag values are log-transformed.*

In Table 1, descriptive values for the outliers' subsets are presented. Even though the amplitudes of those samples are similar, there is a noticeable difference between mean and standard deviation.

Table 1: *Minimum, maximum, mean and standard deviation values, in milliseconds, in the original condition (OC) and the altered condition (AC).*

	Outliers	
	OC	**AC**
Min.	118	127
Max.	976	1569
Mean	180.9	224.9
St. dev.	103	156.9

A Shapiro-Wilk test also indicated that these subsets are not normally distributed. A Fligner-Killeen test shows that the variances of the outliers' subsets are significantly different (med chi-squared = 17.74, df = 1, p = .00 < .05). A Wilcoxon Rank-Sum test for independent samples shows a significant difference between the median values from the two outliers' samples (W = 22,624, p = .00 < .05).

3.3. Outliers: when do they occur?

We now question whether the statistical difference found between speech conditions is due to the semantic incongruities in AC. One promising way of answering this question is identifying the moments throughout the reading when these outliers occur. If they are produced close to - before or after - the semantic incongruities, it might be an indication of its effect on speech synchronization. To answer this question, we made a qualitative description of lag distribution, by pair, generating graphs as the one seen in Figure 4, which shows the reading done by pair A13. This pair has a lag mean of 28.9 ms in the OC (SD = 23.6) and a mean of 58.3 ms in the AC (SD = 84.6).

Figure 4 depicts lag occurrences throughout the text, with highest durations produced at the second semantic incongruity, while other outliers are located at the text final portion. Most pairs show this tendency, i.e., to produce outliers both close to semantic incongruities or in the text second half.

Our initial expectation related to a greater asynchrony was that long lags would always appear next to the incongruent words. However, this was not verified. On the other hand, after observing the distribution tendency among pairs, we assumed that a greater asynchrony would be triggered after more than one incongruity and would be manifested in long lags located at various places after that.

Outliers can also occur at the beginning of sentences, after the production of a longer pause, which is expected [5], although according to our assumptions they are not necessarily related to the incongruities. We refer here to pauses produced by both speakers between phonological phrases.

4. Discussion

In this experiment we sought to observe if text meaning could interfere with the task of speaking in synchrony, known to be an easy one. An experimental condition was created (AC) in which a dyad read texts with an incoherent or even whimsical meaning. It was, then, expected that speakers would synchronize less easily in AC.

The phenomenon seems to be more complex than our initial hypothesis assumed. On the one hand, the speakers were able to synchronize well in both conditions. Even when there was a

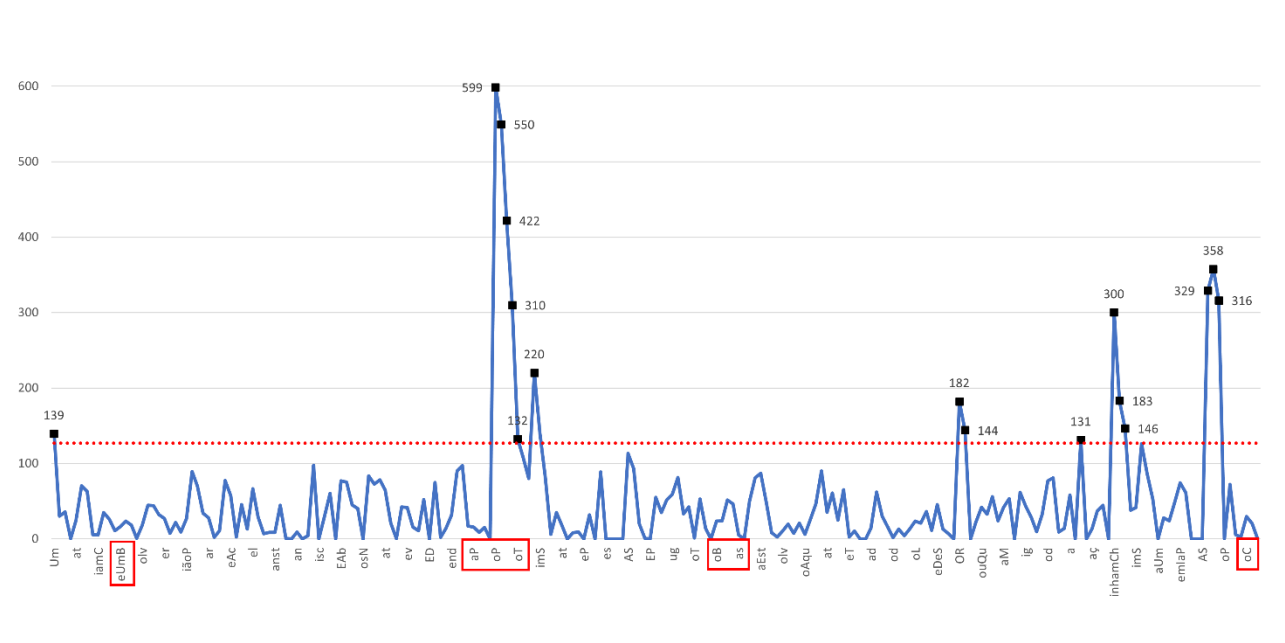

Figure 4. *Lag distribution and lag duration (ms) in the altered condition (AC), for pair A13 (straight line). Horizontal axis shows VV-units. Dotted line indicates outlier (squares) minimum value (127ms). Rectangles indicate the words' syllables conveying semantic incongruities.*

great delay between participants, represented by a long lag, they were able to get back to synchronicity. This corroborates what has been found in previous studies about synchronous speech [5], [7], [23]. On the other hand, very long lags might be an effect caused by the semantic incongruities in the text. They are not enough to affect the lag mean values, but enough to cause greater variation in AC, in comparison to OC.

The outlier variation per se would support our hypothesis, and yet it would not be enough, as it could be a random effect. Therefore, we looked for the moment in the text when outliers would occur. For this search we relied on graph analysis such as the one depicted in Figure 3. The direct relation between the incongruity adjacencies and the presence of longer lags was not verified. Instead, outliers were mostly found at the text second half, not necessarily close to incongruent words, but between them.

It is important to reason about two aspects that might have influenced the results. Studies show that individuals often anticipate their reading by moving their eyes to the following words, before actually reading them out loud [24]–[26]. This might explain why greater lags can occur before the reading of semantic incongruity. Another aspect is the one caused by the occurrence of five semantic incongruities throughout the reading that may affect the whole discourse and not exact places in the text. This effect can be called a cumulative effect.

The greatest lags that correspond to outliers as focused on this study indicated that our hypothesis is, to some extent, tenable. However, as a preliminary study, it needs further inquiries that account for the lag behavior found in our analyses. Studies of priming effect can be useful as well to investigate a possible cumulative effect throughout the reading, which could account for great lags occurring more predominantly at the end of the text, rather than at the beginning [27], [28]. An investigation of head movements, for example, is in progress, as all dyads were video and audio recorded simultaneously. A video images dataset, then, is to be analyzed related to the present paper results. We are aware of the necessity of quantifying data for our question "Outliers: when do they occur?", as we acknowledge the exploratory nature of this qualitative analysis.

In this experiment we proposed a perturbation effect to the synchronized speech task. Asynchrony was greater at AC, as we observed outliers' behavior in this condition. These extreme values became a relevant part of the data, which can pinpoint important information about the phenomenon in hand.

Our prosodic skills allow us to produce speech in a temporal organized manner that can be more or less varying. We can slow down or speed up speech, insert longer pauses, or we can, in the case of synchronized speech, restrict these variations. The task assigned to subjects in this study aimed at disturbing the speech that attempted to adjust the timing between two individuals, this disturbance being of a more abstract order: the disruptive semantic. In the face of this disturbance, we found that a greater speech asynchrony was generated, thus revealing that the prosodic timing adjustments were affected.

Relying on the present results, we hope that our findings will be a contribution to a better understanding of speech production, taking into consideration that it is not a phonetic physical phenomenon devoid of the abstract nature of planning [29]; and synchronized speech can be an experimental means to support this idea, besides, of course, being an important theme to study other speech production aspects.

5. Acknowledgements

This research (Project number 130269/2018-2) was funded by CNPq, Brazil, to the first author. We thank André Baceti, from Murabei Data Science, for his contributions to statistical analysis.

6. References

[1] S. H. Strogatz and I. Stewart, “Coupled Oscillators and Biological Synchronization,” *Sci Am*, vol. 269, no. 6, pp. 102–109, Dec. 1993, doi: 10.1038/scientificamerican1293-102.

[2] M. Wilson and P. F. Cook, “Rhythmic entrainment: Why humans want to, fireflies can’t help it, pet birds try, and sea lions have to be bribed,” *Psychon Bull Rev*, vol. 23, no. 6, pp. 1647–1659, Dec. 2016, doi: 10.3758/s13423-016-1013-x.

[3] F. Cummins, “Periodic and Aperiodic Synchronization in Skilled Action,” *Front. Hum. Neurosci.*, vol. 5, 2011, doi: 10.3389/fnhum.2011.00170.

[4] F. Cummins, “On synchronous speech,” *Acoustics Research Letters Online*, vol. 3, no. 1, pp. 7–11, Jan. 2002, doi: 10.1121/1.1416672.

[5] F. Cummins, “Practice and performance in speech produced synchronously,” *Journal of Phonetics*, vol. 31, no. 2, pp. 139–148, Apr. 2003, doi: 10.1016/S0095-4470(02)00082-7.

[6] F. Cummins, *The ground from which we speak: joint speech and the collective subject*. Cambridge: Cambridge Scholars Publishing, 2018.

[7] F. Cummins, “Rhythm as entrainment: The case of synchronous speech,” *Journal of Phonetics*, vol. 37, no. 1, pp. 16–28, Jan. 2009, doi: 10.1016/j.wocn.2008.08.003.

[8] B. R. de Medeiros and F. Cummins, “Speech and song synchronization: A comparative study,” in *Speech Prosody 2014*, May 2014, pp. 748–751. doi: 10.21437/SpeechProsody.2014-138.

[9] C. A. A. de A. Santos, “A sincronização em dois ritmos da canção: uma observação experimental acerca da fala cantada,” text, Universidade de São Paulo, 2012. doi: 10.11606/D.8.2012.tde-27092012-120540.

[10] N. K. Squires, K. C. Squires, and S. A. Hillyard, “Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man,” *Electroencephalography and Clinical Neurophysiology*, vol. 38, no. 4, pp. 387–401, Apr. 1975, doi: 10.1016/0013-4694(75)90263-1.

[11] C. C. Duncan-Johnson and E. Donchin, “On Quantifying Surprise: The Variation of Event-Related Potentials With Subjective Probability,” *Psychophysiology*, vol. 14, no. 5, pp. 456–467, 1977, doi: 10.1111/j.1469-8986.1977.tb01312.x.

[12] T. W. Picton, “The P300 Wave of the Human Event-Related Potential:,” *Journal of Clinical Neurophysiology*, vol. 9, no. 4, pp. 456–479, Oct. 1992, doi: 10.1097/00004691-199210000-00002.

[13] M. Kutas and S. A. Hillyard, “Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity,” *Science*, vol. 207, no. 4427, pp. 203–205, Jan. 1980, doi: 10.1126/science.7350657.

[14] M. Kutas and S. A. Hillyard, “Event-related brain potentials to grammatical errors and semantic anomalies,” *Memory & Cognition*, vol. 11, no. 5, pp. 539–550, Sep. 1983, doi: 10.3758/BF03196991.

[15] M. Kutas and S. A. Hillyard, “Brain potentials during reading reflect word expectancy and semantic association,” *Nature*, vol. 307, no. 5947, pp. 161–163, Jan. 1984, doi: 10.1038/307161a0.

[16] L. Goldstein and C. Fowler, “Articulatory Phonology: A phonology for public language use,” in *Phonetics and phonology in language comprehension and production: differences and similarities*, 2003, pp. 159–207. doi: 10.1515/9783110895094.159.

[17] C. Fowler, P. Rubin, R. E. Remez, and M. E. Turvey, “Implications for Speech Production: A General Theory of Action,” in *Language Production, Vol. I: Speech and Talk*, New York: Academic Press, 1980, pp. 373–420.

[18] V. P. Siqueira and B. R. de Medeiros, “Uma proposta metodológica para o estudo da sincronização da fala em interação com a semântica,” *Gradus - Revista Brasileira de Fonologia de Laboratório*, vol. 5, no. 1, pp. 99–124, Aug. 2020, doi: 10.47627/gradus.v5i1.151.

[19] P. A. Barbosa, *Incursões em torno do ritmo da fala*. São Paulo: Pontes Editora, 2006. Accessed: Nov. 15, 2021. [Online]. Available: http://ponteseditores.com.br/loja/index.php?route=product/product&product_id=301

[20] J. Morton, S. Marcus, and C. Frankish, “Perceptual centers (P-Centers),” *Psychological Review*, vol. 83, pp. 405–408, Sep. 1976, doi: 10.1037/0033-295X.83.5.405.

[21] “Praat: doing Phonetics by Computer.” https://www.fon.hum.uva.nl/praat/ (accessed Nov. 14, 2021).

[22] “R: The R Project for Statistical Computing.” https://www.r-project.org/ (accessed Nov. 14, 2021).

[23] K. Cerda-Oñate, G. T. Vega, and M. Ordin, “Speech rhythm convergence in a dyadic reading task,” *Speech Communication*, vol. 131, pp. 1–12, Jul. 2021, doi: 10.1016/j.specom.2021.04.003.

[24] D. A. Balota, A. Pollatsek, and K. Rayner, “The interaction of contextual constraints and parafoveal visual information in reading,” *Cognitive Psychology*, vol. 17, no. 3, pp. 364–390, Jul. 1985, doi: 10.1016/0010-0285(85)90013-1.

[25] K. Rayner and A. D. Well, “Effects of contextual constraint on eye movements in reading: A further examination,” *Psychonomic Bulletin & Review*, vol. 3, no. 4, pp. 504–509, Dec. 1996, doi: 10.3758/BF03214555.

[26] S. McDonald and R. Shillcock, “Eye Movements Reveal the On-Line Computation of Lexical Probabilities During Reading,” *Psychological science*, 2003, doi: 10.1046/j.0956-7976.2003.psci_1480.x.

[27] M. A. Moreno and G. C. van Orden, “Word Recognition, Cognitive Psychology of,” in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 16556–16561. doi: 10.1016/B0-08-043076-7/01553-9.

[28] R. S. Hoedemaker and P. C. Gordon, “It takes time to prime: Semantic priming in the ocular lexical decision task.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 6, pp. 2179–2197, 2014, doi: 10.1037/a0037677.

[29] B. Hommel, J. Müsseler, G. Aschersleben, and W. Prinz, “The Theory of Event Coding (TEC): A framework for perception and action planning,” *Behav Brain Sci*, vol. 24, no. 5, pp. 849–878, Oct. 2001, doi: 10.1017/S0140525X01000103.