11-16-2018

Explaining Cross-Language Asymmetries in Prosodic Processing: The Cue-Driven Window Length Hypothesis

Marta Ortega-Llebaria
University of Pittsburgh

Daniel J. Olson
Purdue University, danielolson@purdue.edu

Alba Tuninetti
Western Sydney University

Recommended Citation

Ortega-Llebaria, M., Olson, D. J., & Tuninetti, A. (2018). Explaining cross-language asymmetries in prosodic processing: The Cue-Driven Window Length Hypothesis. *Language and Speech*. https://doi.org/10.1177/0023830918808823

Explaining Cross-Language Asymmetries in Prosodic Processing: The Cue-Driven Window Length Hypothesis

Marta Ortega-Llebaria
University of Pittsburgh, USA

Daniel J. Olson
Purdue University, USA

Alba Tuninetti
Western Sydney University, Australia; ARC Centre of Excellence for the Dynamics of Language, Australia

Abstract
Cross-language studies have shown that English speakers use suprasegmental cues to lexical stress less consistently than speakers of Spanish and other Germanic languages ; accordingly, these studies have attributed this asymmetry to a possible trade-off between the use of vowel reduction and suprasegmental cues in lexical access. We put forward the hypothesis that this “cue trade-off” modulates intonation processing as well, so that English speakers make less use of suprasegmental cues in comparison to Spanish speakers when processing intonation in utterances causing processing asymmetries between these two languages. In three cross-language experiments comparing English and Spanish speakers’ prediction of hypo-articulated utterances in focal sentences and reporting speech, we have provided evidence for our hypothesis and proposed a mechanism, the Cue-Driven Window Length model, which accounts for the observed cross-language processing asymmetries between English and Spanish at both lexical and utterance levels. Altogether, results from these experiments illustrated in detail how different types of low-level acoustic information (e.g., vowel reduction versus duration) interacted with higher-level expectations based on the speakers’ knowledge of intonation providing support for our hypothesis. These interactions were coherent with an active model of speech perception that entailed real-time adjusting to feedback and to information from the context, challenging more traditional models that consider speech perception as a passive, bottom-up pattern-matching process.

Keywords: Speech perception, intonation, vowel reduction, English, Spanish

Corresponding author: Marta Ortega-Llebaria, University of Pittsburgh, 2816 Cathedral of Learning, Pittsburgh, PA 15261, USA. Email: m.ortega.llebaria@gmail.com

1 Introduction

Relevant speech events are expressed in the speech signal across different time scales ranging from short stop bursts (e.g., 4–10 ms) to progressively longer events such as formant transitions, syllables, and prosodic units like trochees and utterances (e.g., Pisoni, 1973; Rosen, 1992; Kubanek, Brunner, Gunduz, Poeppel & Schalk, 2013) making the temporal organization of speech a complex matter. Psychophysical and physiological research showed that information unfolded over time is chunked into temporal windows and provided evidence for at least two—namely, a shorter window of approximately 25–40 ms (Joliot, Rubary & Llinas, 1994; Singer, 1993) and a longer one of 150–250 ms (Nätäänen, 1992; Yabe, Tervaniemi, Reinikainen & Nätäänen, 1997). The integration of these windows provides a framework to organize temporally developing information, a concept used in models such as the Adaptive Window of Analysis (Nusbaum & Henly, 1992) and the Multiple Look Model (Veimester & Wakefield, 1991). Proposals offered by these models differed, for example, in their number of windows, that is, several windows of different lengths versus one temporal window but multiple looks in subsequent analysis. Nevertheless, all models accounted for the observation that not a single window always took precedence, but it changed according to the situation. It is in this light that we propose the Cue-Driven Window Length (CDWL) hypothesis, a mechanism of temporal organization that accounts for the processing of prosody in utterances by combining bottom-up information, such as available cues in the speech signal, with top-down information, such as task goals. Like the Adaptive Window of Analysis (Nusbaum & Henly, 1992), the CDWL hypothesis proposes that the listener adjusts the time of the window length according to the cues available in the speech signal in order to accomplish a specific task in a particular language. Once this window length is adjusted, it conditions the interpretation of additional information. The novelty and value of the CDWL hypothesis is that it provides a single mechanism to account for cross-language differences in prosodic processing at both the lexical and utterance levels. We define "lexical level" as single word processing as it is described in most current word recognition models (e.g., Cutler, 2012; Mirman, 2016), and "utterance level" refers to the words constituting one intonation unit, such as an Intonation Phrase (Ladd, 2008).

The remainder of the Introduction is organized as follows. In CDWL Hypothesis: A Mechanism to Account for Cross-Language Differences in Prosodic Processing (section 1.1), we present our hypothesis by first addressing how languages differ in their acoustic cues to stress. Then, we describe how Cutler and colleagues addressed these cross-language differences. Finally, we show how our CDWL hypothesis provides a processing mechanism to Cutler's explanation of cross-language differences in stress processing and also show that the CDWL hypothesis predicts similar cross-linguistic differences in processing other prosodic units such as utterance intonation. In Testing CDWL, Comparing English and Spanish Speakers' Processing of Utterance Intonation, we explain why Spanish and English provide an excellent testing ground to the CDWL hypothesis and describe in detail how Spanish and English speakers predict upcoming hypo-articulated utterances based in their knowledge of sentence prosody. In Experiments and Research Questions, research questions are then related to each of the three experiments.

1.1 CDWL hypothesis: A mechanism to account for cross-language differences in prosodic processing

Cross-linguistically, languages make use of different sets of cues, both segmental and suprasegmental, to discriminate stressed and unstressed syllables during the course of lexical access. Exemplifying this cross-linguistic difference, consider the Spanish–English cognate *banana.* In both languages, stress falls in the second syllable making this syllable longer and louder (i.e., for Spanish, see Navarro-Tomás, 1974a, 1974b; for English, see Fry, 1955, 1958). However, English speakers mark stress with an additional cue, namely, vowel reduction. Whereas in English, stressed syllables consistently have full vowels and unstressed syllables become reduced (for example, [bə'naːnə]), in Spanish, vowel reduction is at best small in range and it is not phonological (for example, [ba'naːna]) (Nadeu, 2013; Torreira, Simonet & Hualde, 2014). As a result, for Spanish listeners, suprasegmental cues are crucial for word recognition (Soto-Faraco, Sebastián-Gallés & Cutler, 2001), whereas such cues are largely rendered redundant for English listeners (Cutler, 1986, 2005), owing to the consistency in vowel reduction.

Cutler and colleagues observed that counterintuitively, other Germanic languages like Dutch and German behaved like Spanish rather than English in that speakers of Dutch and German used suprasegmental cues to stress more effectively than English speakers did (Cooper, Cutler & Wales, 2002; Tyler & Cutler, 2009; van Donselaar, Koster & Cutler, 2005). For example, Cooper et al. (2002) showed that in a primed lexical decision task, Dutch speakers made use of pitch, duration, and intensity cues in a prime (e.g., a longer and louder MUSversus a shorter and softer mus-) to selectively activate either *MUS*ic or *musEUM*, but not both. English speakers, in contrast, failed to make use of such suprasegmental differences in the primes, activating both *music* and *museum*. In Cutler and colleagues' own words,

> [In Dutch] there may be on-line directive use of [suprasegmental cues to] stress information in lexical access [...]. This result was also observed with similar fragments of Spanish words [...]. [In English] stress information can nearly always be derived from segmental structure, and words can virtually always be distinguished by segmental analysis without recourse to stress (Cutler, Dahan & van Donselaar, 1997, p. 154).

Providing a framework for understanding the above relationship between segmental and suprasegmental cues at the lexical level, Cutler and colleagues (Cooper et al., 2002; Cutler et al., 1997; Cutler, 2005) have proposed a cue-tradeoff, such that segmental cues, like vowel reduction, render suprasegmental cues unnecessary during word recognition when highly correlated with stress as it is in English. Consequently, speakers of languages with either no phonological vowel reduction like Spanish or with a weaker correlation between vowel reduction and stress like Dutch (Quené & Koster, 1998) would make more efficient use of suprasegmental cues to stress than English speakers, a language with consistently reduced vowels in unstressed positions.

The CDWL hypothesis provides a mechanism that accounts for Cutler and colleagues' "trade-off" explanation of cross-language differences in lexical stress processing. The CDWL hypothesis assumes that in order to maximize processing efficiency, speakers adjust the length of the processing window to the minimal duration necessary to interpret acoustic input with regard to the task at hand and the available cues in the speech signal. Accordingly, when processing lexical stress, Dutch and Spanish speakers adjust their processing windows to a length that is efficient enough to identify stressed and unstressed syllables based on the suprasegmental cues of duration and intensity. These cues are relative measures, that is, a syllable of certain duration is perceived as longer if adjacent syllables are shorter (e.g., Massaro, 1984). The same syllable, however, is perceived as shorter if the contiguous syllables are longer. As a result, the temporal window to process stress in Spanish and Dutch has to be at least bi-syllabic. In contrast, a one-syllable window is long enough to process stress in English because English speakers relate stress to vowel reduction. They can perceive the first syllable of [bə'naːnə] as unstressed because it has a schwa and the second as stressed because it has a full vowel. Consequently, Cutler's trade-off explanation between segmental and suprasegmental cues can be understood as the effect of adjusting the processing window to the minimal length required to perform stress detection tasks with the relevant cues available in each language. That is, duration and intensity are the relevant cues to stress in Dutch and Spanish requiring at least two-syllable windows. Because vowel reduction is the relevant cue to stress in English, a shorter one-syllable window is required. Based on the second tenet of the CDWL hypothesis—namely, in setting the window length, speakers regulate the type of acoustic information that is amenable to interpretation—the trade-off between segmental and suprasegmental cues is motivated. The one-syllable window used by English speakers in stress perception tasks is less optimal than two-syllable windows to perceive syllabic differences in duration and intensity, making English speakers use suprasegmental cues to stress less efficiently than speakers of two-syllable window languages like Dutch and Spanish.

As a corollary to the CDWL hypothesis, we make the following prediction. In comparison to English speakers, speakers of languages like Dutch and Spanish will make a more efficient use of suprasegmental cues to interpret events relevant to utterance-level intonation such as pitch accents, pitch range, duration compression in post-focal utterances, and duration expansion in syllables with contrastive pitch accent. For example, interpreting an ascending F0 trajectory into a given pitch accent contour requires two syllables to discern whether the ascending F0 ends at the stressed syllable (e.g., LH* in ToBI notation; Pierrehumbert, 1980) or continues into the post-tonic (e.g., L*H). To test our prediction, we designed three cross-language lexical identification experiments comparing English and Spanish speakers' perception.

1.2 Testing the CDWL hypothesis: Comparing English and Spanish speakers' processing of utterance intonation

English and Spanish provide an ideal comparison to test the CDWL hypothesis because of their lexical stress and intonation patterns. As for lexical stress, English and Spanish are similar insomuch as stressed syllables in both languages have longer durations, louder intensities, and higher pitch (i.e., F0) than their unstressed counterparts (for

Spanish, see Navarro-Tomás, 1974a, 1974b; for English, see Fry, 1955, 1958). The stressed syllable may fall on one of the last four syllables of a word, and despite that its exact position is largely unpredictable, both languages have a trochee bias: the stressed-unstressed pattern of “table” (or *mesa*) is more common than the unstressedstressed pattern of “saloon” (or *mesón*).

English and Spanish, however, differ with respect to the phonetic expression of lexical stress in two key parameters: duration ratios and vowel reduction (see also Ortega-Llebaria, Gu & Fan, 2013 for the effects of pitch accent frequencies). First, the duration differences between stressed and unstressed syllables are consistently larger in English than in Spanish (e.g., Borzone de Manrique & Signorini, 1983; Delattre, 1966; Ramus, Nespor & Mehler, 1999; White & Mattys, 2007). For instance, Delattre (1966) reported that stressed to unstressed duration ratios were 6:1 in English and 3:1 in Spanish. Second, in English, vowel reduction patterns consistently correlate with stress: unstressed vowels are produced with significant vowel reduction, usually expressed as a schwa, whereas stressed vowels are produced as a full, unreduced vowel. In contrast, there is no phonological vowel reduction in Spanish. Since vowel reduction is a consequence of hypo-articulation (Lindblom, 1990) and as such, it makes reduced vowels shorter, it motivates that the duration differences between reduced unstressed vowels and fully realized stressed vowels in English are larger than between stressed and unstressed vowels in Spanish which, unlike English, have no phonological reduction and express duration differences only in relation to stress but not in relation to vowel reduction. Together, these duration and vowel reduction patterns make the acoustic expression of stress in English more salient than that of Spanish, providing an ideal test case to the CDWL hypothesis. As explained in CDWL Hypothesis: A Mechanism to Account for Cross-Language Differences in Prosodic Processing, the presence of phonological vowel reduction as a cue to stress in English and its absence in Spanish causes a trade-off where Spanish speakers rely more heavily on suprasegmental cues to stress than English speakers do during word recognition tasks. Thus, if Spanish speakers rely more on suprasegmental cues than English speakers despite the fact that suprasegmental cues to stress in Spanish are acoustically less salient than in English, this cross-language difference will provide compelling evidence in support of the CDWL hypothesis.

With respect to intonation, English and Spanish are stress-accent languages that use the suprasegmental cues of duration, pitch, and intensity to express sentence-level prominence (Beckman, 1986; Hualde, 2005). In both languages, a well-formed intonation contour requires a minimum of a nuclear pitch accent and a boundary tone, with optional pre-nuclear pitch accents, to express a discourse-level meaning (see Beckman, 1986 for a detailed description of the intonation system in stress-accent languages). Yet, it is possible to find flat-F0 utterances, like reporting clauses—see Figure 1(c) and (d)—and post-focal contexts—see Figure 1(a) and (b). (For similar examples in Spanish and additional examples in English, see Appendix 3.) Reporting clauses refer to the portion of speech that identifies the speaker in direct speech—Figure 1(c) and (d)—and they can precede, follow, or occur in the middle of the reported speech, making their position unpredictable within the directed speech utterance (e.g., Navarro-Tomás, 1974a, 1974b). In contrast, the flat-F0 utterances in post-focal contexts are always preceded by a contrastive pitch accent—for example, “LOVED” in Figure 1(a) and “ROSIE” in Figure 1(b)—and express a contrastive meaning. For example, in Figure 1(a), the accented word “LOVED” precedes the flat-F0 utterance “nana’s present” and the whole sentence corrects the assumption that “my grandparents hated her present.” Despite that sentences with contrastive pitch accents are more common in English than in Spanish (see Zubizarreta, 1998, for an alternative syntactic mechanism for the expression of focus in Spanish, and see Vanrell & Fernandez, 2013, for the dialectal variation to express focus), they are possible in both languages; and more importantly for these experiments, when they happen, contrastive pitch accents always precede a flat-F0 post-focal utterances. Consequently, the presence of a contrastive pitch accent predicts an upcoming flat-F0 utterance in both languages. In contrast, the flat-F0 reporting clauses embedded in directed speech are not predictable because there is no single pitch accent type or any other consistent prosodic unit in the direct speech utterance that always precedes the flat-F0 reporting clause.

It is worth noting in Figures 1(a) and (b) that, as a prosodic landmark, the focal pitch accent always precedes the flat-F0 clause. There is no such landmark in reporting sentences shown in Figures 1(c) and (d). The flat-F0 stretches contain the target word *Nana* either at the beginning of the clause in Figures 1(a) and (c) or in the middle in Figures 1(b) and (d).

Important to this experiment is the fact that the flat-F0 utterances of reporting clauses and postfocal contexts are produced with a compressed pitch range and softer intensity that we will call “hypo-articulated utterances” (Lindblom, 1990; and listen to speech files in Appendix 3). These hypo-articulations makes the cues to stress less perceptible (for English, see Beckman, 1986; for Spanish, see Navarro-Tomás, 1974b; Hualde, 2005). As a result, the requirements of sentence intonation shape the acoustic expression of lexical stress: in the hypo-articulated clauses, lexical stress is expressed mainly by vowel quality and duration cues (e.g., Beckman & Edwards, 1994 for English; Ortega-Llebaria, 2006 for Spanish; Ortega-Llebaria et al, 2011) because pitch compression makes pitch contrasts less perceptible, making it less relevant as a cue to stress. In contrast, the stressed syllables of the pre-focal clause and reported sentences will preserve all the cues to stress. Finally, the stressed syllable of the word in focus will preserve all the cues to stress (i.e., pitch, duration, intensity, and vowel quality), expressed with enlarged dimensions.

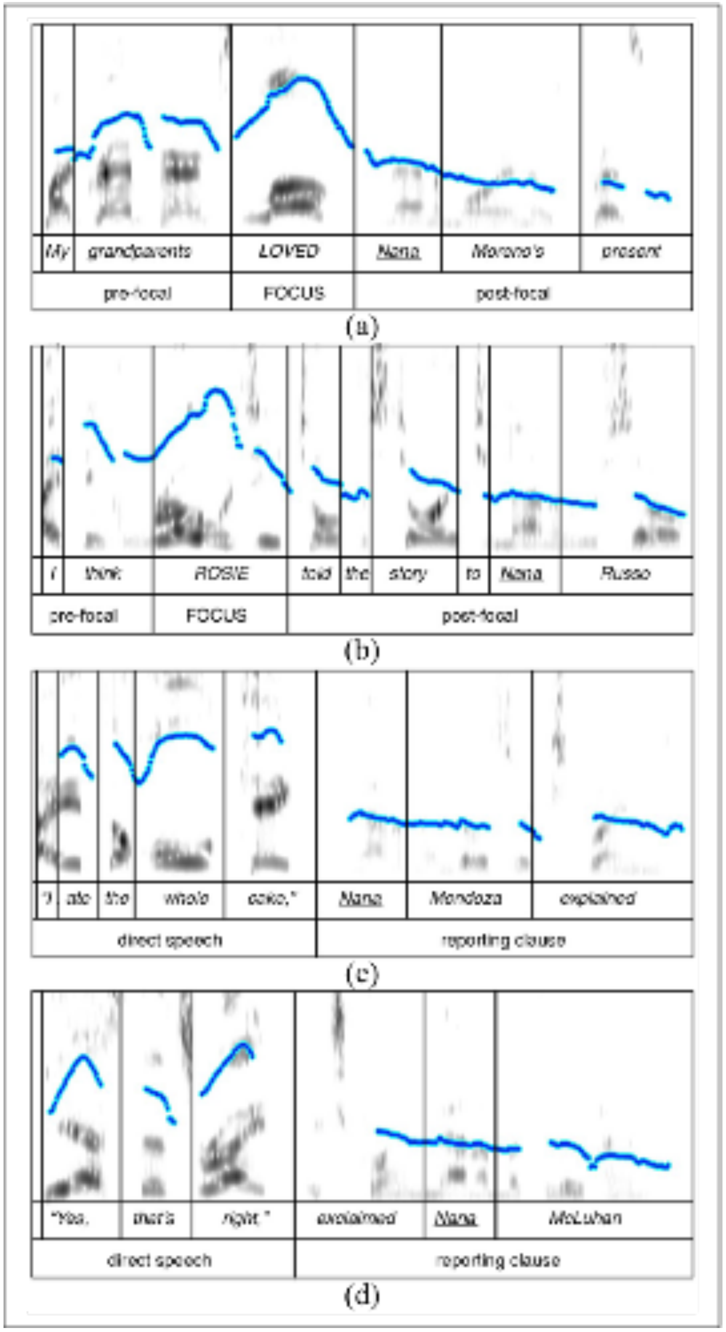


Figure 1. Flat-F0 clauses embedded in sentences spoken with focal and reporting intonations. (a) Focal sentence with target word "Nana" at the beginning of the post-focal flat-F0 stretch (i.e., post-focal, initial). (b) Focal sentence with target word "Nana" in the middle of the post-focal flat-F0 stretch (i.e., post-focal, medial). (c) Reporting utterance containing the target "Nana" at the beginning of the flat-F0 stretch (i.e., reporting, initial). (d) Reporting utterance containing the target "Nana" in the middle of the flat-F0 stretch (reporting, medial).

When processing speech, predicting an upcoming hypo-articulated utterance may facilitate lexical access by readjusting the weights to pitch and duration stress cues accordingly, and consequently, reducing the number of activated candidates to those that configure a contextually appropriate stress pattern. For example, detecting iambic "Nana" instead of "Naomi," "Nancy" or trochee "Nana" in the sentences of Figure 1, will be easier in contexts where the upcoming hypoarticulated utterance is predicted making reaction times to word detection faster in these contexts. While attending to the focal pitch accent with expanded pitch range and duration—that is, "LOVED" in Figure 1(a)—drawing on their (implicit) knowledge of the prosodic structure of contrastive focus sentences, speakers may anticipate the upcoming hypo-articulated utterance. In this case, expectations of a flat-F0 clause may be generated and weights to stress cues adjusted accordingly facilitating detection of iambic "Nana." This example can be contrasted with reporting clauses. Reporting clauses are also hypo-articulated utterances but unlike post-focal utterances, they lack a specific pitch accent type (or any other prosodic unit) that consistently precedes it, and consequently, that could be associated with the reporting clause and predict it. As such, although both post-focal and reporting clauses contain a flat F0 contour, only contrastive focus utterances contain a prosodic cue, the contrastive pitch accent, that could allow speakers to anticipate the hypo-articulated utterance.

In addition to the use of a contrastive pitch accent, listeners may also anticipate a flat F0 contour on a target token if the preceding string is also part of a hypo-articulated utterance. Consider the case of the reporting clause, in which there are no prosodic landmarks to signal the upcoming deaccented F0 contour. In this case, listeners may only begin to anticipate a flat F0 contour as they begin processing the first word of the reporting clause—for example, "Nana" in Figure 1(c). As such, although speakers may not be able to anticipate a deaccented F0 contour for the first word of the reporting clause—"Nana" in Figure 1(c)—they could reasonably anticipate a flat contour for a clause-medial token—"Nana" in Figure 1(d).

Thus, when predicting an upcoming hypo-articulated utterance, lexical access may be facilitated by either a preceding contrastive pitch accent, as in contrastive focus utterances, or a preceding flat F0 contour (i.e., clause medial position of the target word). As such, lexical access, or reaction times in a lexical identification task, may be expected to be longer for target words in a reporting clause relative to a post-focal clause, specifically for clause-initial tokens. Moreover, it would be expected that clause-initial tokens would evidence longer reaction times than clausemedial tokens. Finally, the reporting sentence with a target in initial position—Figure 1(c)—is the only context containing no prosodic cues on which speakers may predict the upcoming hypoarticulated utterance, and therefore, this context would be expected to demonstrate the longest reaction times.

1.3 Experiments and research questions

The current study explores the CDWL hypothesis as a mechanism to explain prosodic processing. More specifically, based on Cutler and colleagues' trade-off between segmental (i.e., vowel reduction) and suprasegmental cues at the lexical level (i.e., duration and pitch, Cooper et al., 2002; Tyler & Cutler, 2009; van Donselaar et al., 2005), the three proposed experiments examine the prediction explained at the end of CWDL Hypothesis: A Mechanism to Account for Cross-Language Differences in Prosodic Processing (section 1.1), namely, whether the trade-off at the lexical level modulates as well listeners' ability to generate expectations of upcoming hypo-articulated utterances. These expectations are based in the fact that that post-focal utterances are always preceded by a contrastive pitch accent that makes them predictable, whereas reporting sentences have no consistent cue, which makes them unpredictable. Three specific research questions are addressed:

Research Question 1: Can we observe cross-language asymmetries between English and Spanish speakers in their processing of intonation in naturally spoken focal sentences and reporting clauses, which contain phonological vowel reduction only in English? To answer this question, a lexical identification task was conducted in Experiment 1 using English and Spanish materials distributed across the four intonation contexts described above (Figure 1). These naturalistic materials preserved cross-linguistic differences, such that iambic target "Nana" was expressed via vowel reduction and duration cues in English (i.e., [nənaː]) and only by duration cues in Spanish (i.e., [nanaː]). If the presence of vowel reduction in the English materials mitigates the use of suprasegmental cues while the absence of vowel reduction does not, then English speakers will rely less than Spanish speakers on their predictions of the upcoming hypo-articulated utterances. Consequently, it is expected that significant reaction time differences between the four intonation contexts illustrated in Figure 1 will be produced by Spanish speakers but not by English speakers.

If results from Experiment 1 show evidence of cross-language differences, we will have obtained evidence that Cutler and colleagues' trade-off (Cooper et al., 2002; Tyler & Cutler, 2009; van Donselaar et al., 2005) between segmental and suprasegmental cues at the lexical level also modulates prosodic processing at the utterance level. Then, the next step is to explore the two tenets of the CDWL hypothesis with utterance materials that control for the cues of duration and vowel reduction across languages. Recall from CDWL Hypothesis: A Mechanism to Account for CrossLanguage Differences in Prosodic Processing that the first CDWL hypothesis assumption posits that listeners adjust the length of the processing window (i.e., amount of acoustic information processed from the incoming signal) to perform a given task. This adjustment is driven by both the type of acoustic cue available in the speech signal and the task to be performed. For example, if the task consists of perceiving syllabic prominence based on duration cues, listeners are expected to use a window length of at least two syllables in order to discern which syllable is shorter and which is longer. In our stimuli, listeners are expected to use two syllable windows to differentiate [nanaː] from [naːna]. In contrast, if the task consists of perceiving syllable prominence based on patterns of vowel reduction, then a window length of one syllable is sufficient to determine whether the vowel is reduced and the syllable is non-prominent, or whether the vowel is fully realized and the syllable is prominent. In our stimuli, English listeners are expected to make reliable decisions on syllabic prominence by detecting the schwa in [nəna] and the full vowel in [nanə]. This hypothesis leads to Research Question 2.

Research Question 2: Do listeners adjust the length of a processing window based on the nature of the available acoustic information? In other words, does a lexical-level segmental cue (i.e., vowel reduction) favor a shorter processing window than a suprasegmental cue (i.e., duration)? A gating task (Experiment 2) was designed to answer this question. Native English and native Spanish listeners were asked to perform a lexical identification task to differentiate iambic and trochee renditions of "Nana." Crucially, vowel quality and duration cues are manipulated in these words so that they contain either duration cues (i.e., [nanaː] and [naːna]) or vowel reduction cues (i.e., [nəna] and [nanə] in English; [nena] and [nane] in Spanish; note we did not use "schwa" in Spanish because it is not part of the Spanish vowel inventory). Support for the CDWL hypothesis will come from both English and Spanish speakers successfully identifying target words with vowel quality cues at an earlier gate than target words with durational cues.

The second CDWL hypothesis assumption establishes that the window length, adjusted according to the task goals and the available acoustic information in the speech signal, subsequently determines which top-down information is integrated into the perceptual process. Longer windows are needed to perceive suprasegmental features, such as pitch and duration, which are perceived as relative rather than absolute values (i.e., pitch in a syllable is perceived as high if the pitch of the next syllable is lower). This longer window length facilitates the interpretation of suprasegmental cues, which in turn, favors the generation of expectations based on sentence intonation such as predicting upcoming hypo-articulated utterances immediately after listening to a contrastive pitch accent. Shorter windows are needed to perceive vowel quality, which then impedes the interpretation of suprasegmental cues and hinders the generation of expectations driven by sentence intonation. This CDWL hypothesis prediction leads us to the Research Question 3.

Research Question 3: Do longer processing windows favor a more efficient use of speakers' prosodic knowledge such as the association between a contrastive pitch accent and its following F0-flat hypo-articulated post-focal utterance? A lexical identification task (Experiment 3) was conducted with native English and Spanish listeners, parallel to Experiment 1, with one key difference. Whereas Experiment 1 employed naturalistic speech, target tokens in Experiment 3 were the manipulated tokens of Experiment 2 (i.e., the iambic and trochee renditions of "Nana" contained *either* vowel quality or duration cues). Support from the second CWDL hypothesis will come from a reduction of the cross-language differences of Experiment 1 in Experiment 3. In Experiment 3, it is expected that both English and Spanish speakers will rely on sentence prosody when target words contain duration cues because duration cues require it. A two-syllable processing window (which, in turn, facilitates processing duration and pitch cues) is relevant to intonation. However, participants are expected not to rely on sentence prosody when the target words contain vowel reduction cues because one-syllable windows to process vowel reduction are too short to interpret the duration and pitch cues relevant to intonation.

Altogether, results from the above experiments will illustrate in detail how different types of low-level acoustic information (e.g., vowel reduction vs. duration) interact with higher-level expectations based on the speakers' knowledge of sentence intonation to predict upcoming hypoarticulated utterances. In doing so, we will discuss the CDWL hypothesis and its implications for models of speech perception. In general, models of speech perception, in contrast to more interactive models of word recognition (e.g., McClelland & Rumelhart, 1985; Seidenberg & McClelland, 1989) and sentence comprehension (e.g., Norris, 1994), have a long tradition of favoring a linear, bottom-up approach with limited top-down interaction (e.g., Fant, 1967; Jusczyk, 2000; Liberman & Mattingly, 1985).

2 Experiment 1: Word detection task in natural speech

2.1 Methodology

Participants. Eighteen native English speakers (10 females) and 17 native Spanish speakers (13 females) participated in the word detection task. All participants, ranging in age from 23 to 41 (*M* = 26), were students at the University of Texas, Austin. Participants were considered native English speakers if they had learned English from birth to the exclusion of any other language, rated themselves as more dominant in English than any other language, and acquired any other language after the age of 5. Likewise, native Spanish speakers who had learned Mexican Spanish from birth rated themselves as more dominant in Spanish than any other language, and had acquired any other language after the age of 5. Worth noting, although the native Spanish speakers resided in the USA, they all reported speaking Spanish in their daily lives, both at work and socially, and using Spanish more frequently than English. They were fluent speaking and writing non-colloquial, educated registers of Spanish. All subjects reported normal speech and hearing.

Stimuli. Auditory stimuli consisted of recordings in English and Spanish containing target words embedded within the four sentence intonation contexts described in Testing CDWL, Comparing English and Spanish Speakers' Processing of Utterance Intonation. Specifically, 180 utterances were recorded in English, by a native English speaker (Midwestern American variety, see Appendix 3 for examples with their sound files), and 180 utterances were recorded in Spanish, by a native Spanish speaker (Peninsular variety; see Appendix 3). Utterances were recorded in a sound-proofed booth, with a 44.1 kHz sampling rate. Of the 180 utterances recorded in each language (120 target words, 60 fillers), 90 consisted of utterances containing contrastive narrow focus, correcting a previous utterance, and 90 contained reporting clauses. As shown in Figure 1, both post-focal and reporting utterances contained a flat-F0 contour clause (e.g., Huss, 1978 for English; NavarroTomás, 1974b for Spanish). Within this flat-F0 clause, we embedded tokens of the target tokens of the personal names *NAna* (*n* = 60) and *naNA* (*n* = 60), with stress on either the first syllable (i.e., iambic) or second syllable (i.e., trochee). An additional 60 filler tokens were included containing personal names starting with [na] and [nə] (e.g., Naomi). Tokens were positioned in either clause initial or clause medial position. Tokens in initial position were the first constituent of the deaccented section, while tokens in medial position were a minimum of three syllables, one of them with primary stress, from the start of the de-accented portions (Spanish: *M* = 4.3, *SD* = 1.01; English: *M* = 3.25, *SD* = 0.86; see examples in Figure 1). Thus, stimuli consisted of four intonation contexts (post-focal initial, post-focal medial, reporting initial, reporting medial), three types of tokens (iambic target, trochee foil, names with initial *Na-*), 15 sentences, and two languages.

Two Spanish–English bilingual speakers and trained phoneticians listened to the utterances and independently marked the contrastive pitch accents they heard. They selected the focal utterances where both listeners clearly heard a contrastive pitch accent before the flat-F0 contour clause, and the reporting sentences where no focal accent was heard before them. From these utterances, the 10 sentences (out of the 15) per condition that sounded the clearest were selected, yielding 120 utterances per language (80 target words, 40 fillers).

Target tokens. As explained in the Introduction, the phonetic expression of stress is more salient in English than it is in Spanish because duration differences in English are larger than those in Spanish, and English has an additional cue to stress, namely vowel quality (e.g., Delattre, 1966; Borzone de Manrique & Signorini, 1983). To confirm these expected differences, the duration, intensity, and vowel quality were analyzed for each of the 80 target tokens in English and Spanish. For each factor, a ratio was calculated by dividing the duration, intensity, and vowel quality measurements of the second syllable by those of the first syllable. As such, a value close to 1.0 corresponds to similar duration, intensity, and vowel quality in the stressed and unstressed syllable. Values greater than 1.0

correspond to a greater duration, intensity, and vowel quality (F1) associated with the second syllable, while values less than 1.0 correspond to greater duration, intensity, and vowel quality in the first syllable.

Figure 2 shows that the iamb and trochee realizations were clearly differentiated in both languages. One-way analyses of variance (ANOVAs) showed that trochee and iamb realizations differed significantly in all three dimensions (duration, intensity, and vowel quality) in both languages (in English: duration, $F(1,78) = 435.418$, $p < 0.0001$; intensity, $F(1,78) = 130.458$, $p < 0.0001$; vowel quality, $F(1,78) = 273.939$, $p < 0.0001$; and in Spanish: duration, $F(1,78) = 251.871$, $p < 0.0001$; intensity, $F(1,78) = 25.274$, $p < 0.0001$; vowel quality, $F(1,78) = 10.953$, $p = 0.086$).

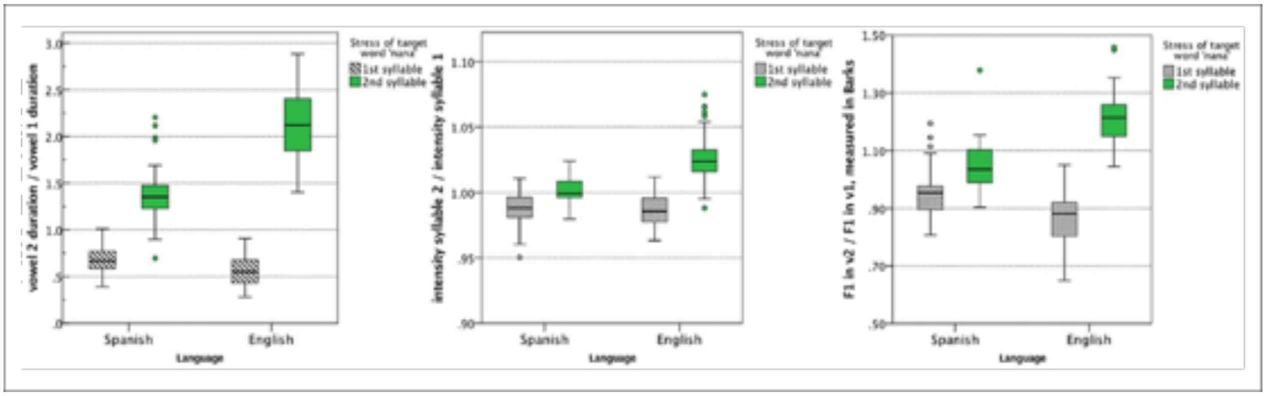


Figure 2. Duration, intensity, and vowel quality ratios of trochee and iamb pronunciations of "Nana" in English and Spanish. The first graph depicts duration, the second depicts intensity, and the third depicts vowel quality. Bars with lines depict ratios of trochees, and bars with solid filling depict ratios of iambs.. v: vowel.

However, the larger mean differences in English indicated that the contrast was more salient in English than in Spanish. For example, mean duration differences between the syllables of trochee "Nana" were 67 ms in English and 60 ms in Spanish, and between the syllables of iambic "Nana" were 57 ms in English and 42 ms in Spanish. Considering intensity, stressed "na" was louder than unstressed "na" in both languages. On average, stressed syllables were 1.8 dB louder in English and 1.1 dBs in Spanish. Similarly, vowel reduction contrasts were larger in English. F1 in the unstressed schwa in English was on average 1.2 Barks higher than the F1 of the corresponding stressed [a]. However, an analogue comparison in Spanish scored 0.4 Barks, indicating that vowel quality differences consistently cued a contrast in English but not in Spanish.

Procedure. In order to capture the effects of the four sentence intonation contexts described in Testing CDWL, Comparing English and Spanish Speakers' Processing of Utterance Intonation on lexical access, participants performed a lexical identification task (Kilborn & Moss, 1996; Marinis, 2010; Marslen-Wilson & Tyler, 1980) in their native language in the Spanish Phonetics Lab at the University of Texas at Austin. With headphones adjusted to a comfortable loudness, participants self-started instructions and stimulus presentation using SuperLab Pro 4.1.2 (Cedrus, 2010). Instructions directed the participants to indicate via response pad (Cedrus RB-800), as soon as they identified the iambic target word [nə'naː] in English and [na'naː] in Spanish. The explicit instructions written in the computer screen were "push this key as soon as you hear the name 'Nana' with stress in the last syllable like in 'payee' (or 'mesón' for Spanish speakers)." Thus, this task required participants to focus their attention on iambic target words while eliminating task-irrelevant lexical items (e.g., trochee ['naːna] in Spanish and ['naː nə] in English) and fillers beginning with [n], like Naomi, Natasha, Nora, Nanni (for a similar task, see Marinis, 2010, p. 142). Following a set of practice utterances, stimuli were presented to participants in blocks of 20, with an inter-stimulus interval of 1500 ms. To limit fatigue, participants were permitted a brief break between blocks. Stimulus order was randomized, and each subject received a different randomized order. The effect of sentence intonation was assessed by comparing reaction times to the iambic target word between the four sentence intonation contexts.

2.2 Results

In order to assess participants' sensitivity towards the iamb target word, that is, [nə'naː] in English and [na'naː] in Spanish, *d′* scores were calculated as the difference of *z* scores between hits (key pressed on hearing the iambic target word) and false alarms (key pressed on hearing the corresponding trochee). Out of 40 possible answers, English participants obtained a mean of 39.3 hits (*SD* = 0.6) and Spanish 37.09 (*SD* = 1.8). As for false alarms, English speakers scored a mean of 2.9 (*SD* = 3.5) and Spanish speakers 3.4 (*SD* = 1.1). Two participants (one English and one Spanish) obtained *d′* scores lower than 2 and they were eliminated from further analysis. For the remaining participants, *d′* scores averaged 2.83 for English speakers and 2.7 for Spanish speakers, confirming they were all highly sensitive to the target word.

In order to assess the effect of intonation contexts in the perception of the target word, reaction times to correctly identified target words were further analyzed. Reaction times, measured in milliseconds, were defined as the temporal delay from the offset of the target word to the participant response. Responses with a negative reaction time or a reaction time over 2000 ms were eliminated, representing 1.17% of the data.

Figure 3 illustrates the mean reaction times to target word detection in each of the intonation contexts (post-focal initial, post-focal medial, reporting initial, and reporting medial) and language. A visual inspection of the graphs reveals clear differences across languages. In general, reaction times are longer in Spanish (*n* = 533, *M* = 504.3, *SD* = 316.2) than in English (*n* = 642, *M* = 378, *SD* = 184.2). In particular, reaction times for Spanish speakers were visibly longer in the reporting initial condition than in any other context. In contrast, no clearly discerning pattern appeared in the English data. To assess the significance of these patterns, hierarchical linear models were conducted in R 3.1.3 (R Core Team, 2013) using the LMER package (Bates, Maechler, Bolker & Walker, 2014) for each language group with the four intonation contexts (intonation) as the fixed effect (i.e., post-focal initial, post-focal medial, reporting initial, and reporting medial) and Subject and Item as random effects. Results showed a significant effect of intonation context for Spanish, *F*(3,34.92) = 3.98, *p* = 0.0153 but not for English speakers, *F*(3,32.23) = 2.419, *p* = 0.0841, indicating that expectations based on sentence prosody affected word detection in Spanish. Multiple comparisons with the Bonferroni adjustment on the Spanish data showed that reaction times in the reporting initial context (*n* = 141, *M* = 627.63, *SD* = 392.21) were significantly longer than those in post-focal medial (*n* = 124, *M* = 472.64, *SD* = 268.8) and post-focal initial (*n* = 130, *M* = 435.77, *SD* = 300.4) and reporting medial contexts (*n* = 138, *M* = 535.8, *SD* = 387.78) at the $p < 0.05$ level, confirming that it took longer for Spanish speakers to detect the target word in the reporting-initial sentences. Again, it is important to note that the reporting-initial context contained no prosodic cues to the upcoming hypo-articulated flat-F0 clause. Moreover, there were no significant differences between the three contexts that contained prosodic cues, that is, post-focal initial, post-focal medial, and reporting medial.

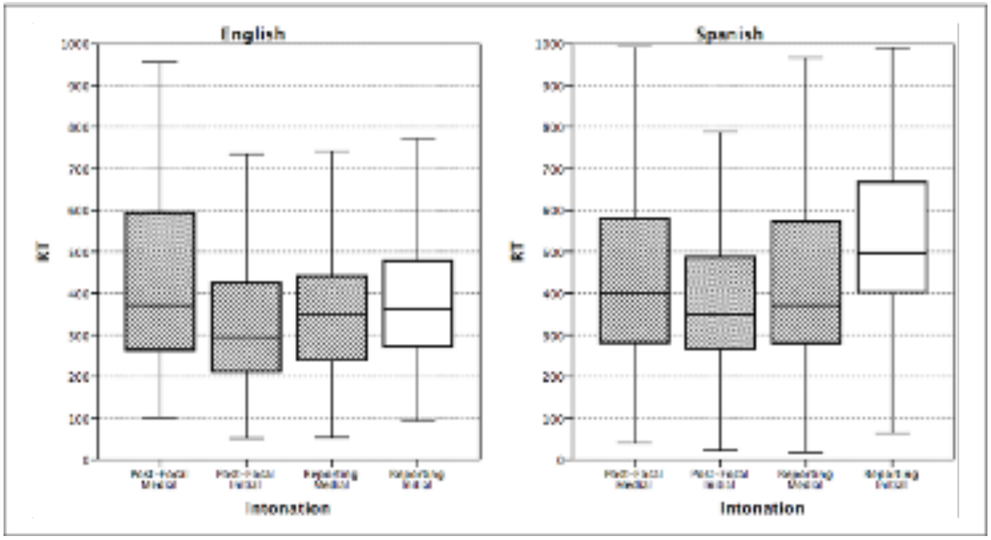


Figure 3. English (left panel) and Spanish (right panel) participants' reaction times to the target word embedded in different intonation contexts. The boxplots with line patterns refer to contexts that contain one or more prosodic cues to the upcoming hypo-articulated utterance (the focal pitch accent or the target is in medial position of hypo-articulated utterances). The boxplot in white refers to the only context that contains no cues to the upcoming hypo-articulated utterance. RT: reaction time

2.3 Discussion

Two main findings can be drawn from Experiment 1. First, for Spanish speakers, the above results showed variation in reaction times based on the preceding intonational context. Specifically, significantly longer reaction times were evidenced when Spanish speakers had to detect the target word in the absence of cues that allowed them anticipate the hypo-articulated utterance (i.e., reporting initial context). These findings confirm that, when available, Spanish speakers used the prosodic cues in order to predict up-coming hypo-articulated utterances facilitating the detection of the target word. In contrast, English speakers' reaction times reflected no significant differences across the four intonation contexts. These results demonstrate that English speakers did not use the available prosodic cues to create expectations that facilitate word detection. Thus, reaction times revealed a cross-language difference in the extent to which speakers are able to effectively predict upcoming hypo-articulated utterances based on the preceding sentence-level prosody. In short, Spanish speakers made more effective use of preceding sentence-level intonation than English speakers.

Second, the acoustic analysis of the target words confirmed that vowel reduction and suprasegmental cues to stress were more salient in English because duration, intensity, and vowel quality differences between stressed and unstressed syllables were larger in English than in Spanish. Together, these two results show that despite the fact that suprasegmental cues were more salient in the English materials, English speakers used them to a lesser extent than Spanish speakers, making it plausible that Cutler's trade-off hypothesis between vowel reduction and the suprasegmental cues to stress modulated speakers' use of sentence prosody, answering Research Question 1 affirmatively. Vowel reduction prevented English speakers from using suprasegmental cues to sentence prosody whereas Spanish speakers, in the absence of vowel reduction, used their less salient suprasegmental cues more effectively to predict upcoming hypo-articulated utterances.

3 Experiment 2: Gating task

The next two experiments explored the underlying mechanism responsible for Experiment 1's crosslanguage difference in the use of suprasegmental cues by testing the CDWL hypothesis described in Experiments and Research Questions. Experiment 2 tests the first assumption of the CDWL hypothesis by addressing Research Question 2; namely, do listeners adjust the length of a processing window based on the nature of the available acoustic information? A gating experiment was designed where the sentences of Experiment 1 are truncated in a series of progressively longer windows or gates: after the first syllable of the target word, after the second syllable of the target word, and one syllable after the target word (see Figure 4). Moreover, cues to target words are manipulated so that target words in both languages contained either duration cues (e.g., [naˈnaː] vs. [ˈnaːna]) or vowel quality cues (e.g., [nəˈna] vs. [ˈnanə]), but not both. The participants' task consists of deciding whether they heard an iambic or a trochaic rendition of "Nana" by pressing the appropriate button. The CDWL hypothesis will be supported if English and Spanish speakers obtain reliable answers after listening to the first syllable only in the stimuli containing vowel quality cues (e.g., [nəˈna] vs. [ˈnanə] in English and [neˈna] vs. [ˈnane] in Spanish), while needing to listen to more than one syllable in the stimuli with duration cues (e.g., [naˈnaː] vs. [ˈnaːna]). In Spanish, in order to use Spanish vowels, vowel quality was cued by using [e] instead of a schwa in unstressed positions.

3.1 Methodology

Participants. For Experiment 2, a new set of participants was recruited, consisting of 12 native speakers of American English and 10 native speakers of Mexican Spanish, as determined by the same criteria employed in Experiment 1. That is, participants were considered native English (or Spanish) speakers if they were exposed to English (or Spanish) from birth to the exclusion of any other language, rated themselves more dominant in English (or Spanish), and did not learn any other language until after age 5. Participants were students at the University of Texas at Austin.

Stimuli. Utterances recorded for Experiment 1 were employed again in Experiment 2, with crucial manipulations made to the target tokens. Target tokens were manipulated, via Praat (Boersma & Weenink, 2012) to create two separate sets of 80 utterances; those containing target words contrasting in duration (i.e., [na'naː] vs. ['naːna], duration utterances), and those containing target words contrasting in vowel quality (i.e., [nə'na] vs. ['nanə] vowel quality utterances). In the duration utterances, the vowel quality in both the first and second syllable of [nana] were kept identical by copying the first vowel into the second one while maintaining the original (i.e., naturalistic) duration differences, so that the duration contrast served as the main cue to stress placement. Similarly, in the vowel quality utterances, the duration of both the first and second syllables were made identical, with vowel quality serving as the main cue to stress placement. Given the inherent lack of vowel quality contrast in Spanish, and lack of vowel reduction, target tokens in the Spanish vowel quality condition consisted of [ne’na] and ['nane]. Intensity (R^2) was manipulated such that both vowels of the target token had similar intensity levels (difference < 2 dB). To ensure that manipulations of the target tokens were done appropriately, one-way ANOVAS with stress as the grouping factor were performed on duration and vowel quality ratios. Results, summarized in Appendix 2, confirmed that duration differences served as the sole cue to stress in the duration utterances, and vowel quality served as the sole cue to stress in the vowel quality utterances.

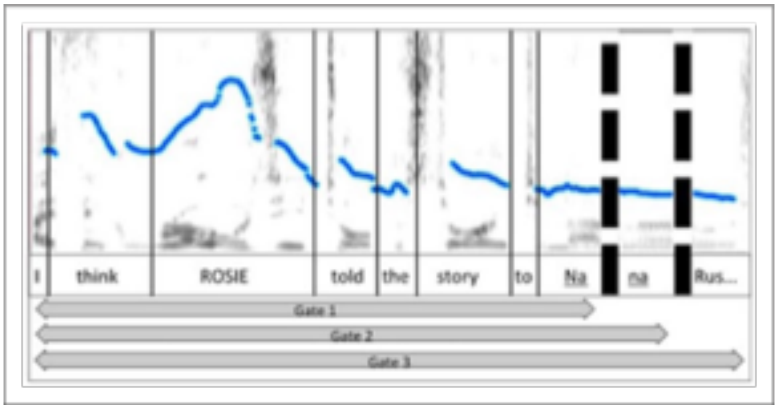


Figure 4. Gates played to the participants.

The gating paradigm employs the manipulated duration utterances and vowel quality utterances. Creating three progressively longer versions of the stimuli, each utterance was truncated at a different point relative to the target token (Figure 4 above). The first iteration consisted of the utterance and the first syllable of the target token. The second iteration consisted of the utterance and the first and second syllables of the target token. Finally, the third iteration consisted of the utterance, the target token, and one additional syllable from the following word. Sentences were balanced for intonation pattern (post-focal vs. reporting) and clause position (initial vs. medial).

Procedure. Though a brief, automated training, subjects were instructed to respond to each auditory presented stimulus, via a response box. Subjects were asked to indicate whether they perceived the target word either as a trochee or as an iambic rendition of “Nana” (i.e., [naːna] or [nanaː] in the duration condition, [nanə] or [nəna] in English, and [nane] or [nena] in Spanish in the vowel quality condition). The three iterations of each stimulus were presented in sequence, with each progressively longer than the previous. Response was required after each iteration of the stimulus, such that participants responded three times per stimulus. Stimuli were blocked by condition, and the order of the conditions was counterbalanced. Within each block, stimuli were randomized, and each participant received a different randomized order. Again, a brief break was given after 20 stimuli to limit fatigue.

3.2 Results and discussion

Figure 5 illustrates the mean percentage of correctly identified targets as iambic or trochee across each of the three gates. Decisions made on Gate 1 were close to chance when based on duration cues (51.3% correct in English and 53.8% in Spanish) and above chance when based on vowel quality cues, (69.2% correct in English and 98.8% correct in Spanish). Exact Binomial Tests confirmed the statistical significance of these patterns. When stimuli contained only durational cues to stress, probabilities of correctly identifying the syllable in Gate 1 as long or short were not significantly different from chance in either language (English, *p* = 1, 95% CI: 0.35–0.68; Spanish, *p* = 0.75, 95% CI: 0.37–0.70). In contrast, in stimuli containing vowel quality differences, probabilities

of success were significantly greater than chance (English, *p* = 0.023, 95% CI: 0.52–0.83; Spanish, *p* < 0.0001, 95% CI: 0.91–1).[1] When considering performance in Gates 2 and 3, decisions made by speakers of both languages were close to ceiling, with values ranging from 77% to 100%.

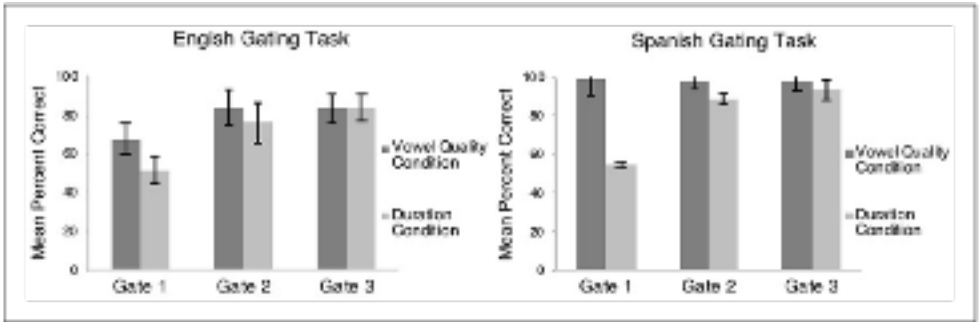


Figure 5. Mean percentage correct answers for the gating task for English speakers (left panel) and Spanish speakers (right panel). Gate 1 occurred at the end of the first syllable of the target word “nana”; Gate 2 occurred at the end of the second syllable; Gate 3 occurred at the end of the syllable after the target word. Error bars represent ±1 SD.

Thus, overall these results confirm the first part of the CDWL hypothesis, namely, that the length of the processing window is adjusted in relation to the type of acoustic cue available in the speech signal. Duration is a relative cue (e.g., Massaro, 1984), and therefore, a syllable is perceived as long (or short) in relation to other syllables. When duration cues were present in the speech signal, both English and Spanish speakers needed at least two syllables to make reliable decisions on whether they were hearing an iambic [nana:] or a trochee [na:na] word. In contrast, vowel reduction is not a relative cue, and therefore, vowel quality cues perceived in isolated vowels constitute sufficient information for the speakers to identify a vowel. Thus, when vowel quality cues were present in the target words [nəˈna] and [’nanə] ([ne’na] and [’nane] in Spanish), the first syllable was enough to differentiate the schwa in the iamb [nəna] (or [ne] in Spanish), from the full vowel in the trochee [nanə] (or [na] in Spanish).

4 Experiment 3: Word detection task when controlling for duration and vowel quality cues

This experiment was designed to test the second assumption of the CDWL hypothesis by addressing Research Question 3, namely, do longer processing windows favor the generation of prosodic expectations because these are long enough to interpret the suprasegmental cues than configure sentence intonation? As explained in Experiments and Research Questions, linguistic constructs such as sentence intonation (which is based on the suprasegmental cues of duration), pitch, and intensity, will require longer window lengths to be interpreted than, for instance, segmental differences in vowel quality. To test the second CDWL hypothesis assumption, we replicated Experiment 1 design in Experiment 3 but inserted the cue-manipulated target words of Experiment 2. Thus, speakers had to detect the iambic tokens of “Nana” embedded in the same intonation contexts as in Experiment 1 based either on duration cues (e.g., [na’na:] and [’na:na]) or vowel quality cues (e.g., [nəˈna] and [’nanə]). Because in Experiment 2 both English and Spanish speakers used a larger processing window when perceiving target words with duration cues than with vowel quality cues, we expected that the cross-language difference obtained in Experiment 1—namely, only Spanish speakers used expectations based on sentence intonation to detect the target word—will become reduced in Experiment 3. When target words contain duration cues, both Spanish and English speakers are expected to predict upcoming hypo-articulated utterances. Similarly, when target words contain vowel quality cues, both Spanish and English speakers are expected not to use expectations based on sentence intonation.

4.1 Methodology

Participants. For Experiment 3, a new set of participants was recruited, consisting of 57 native speakers of American English (44 female) and 27 (18 female) native speakers of Mexican Spanish, as determined by the same criteria employed in Experiment 1. That is, participants were considered native English speakers if they learned English from birth, to the exclusion of any other language, rated themselves as more dominant in English than Spanish, and did not begin to acquire any other language before the age of 5. Similarly, native Spanish speakers learned Spanish from birth, rated themselves as more dominant in Spanish, and did not learn any other language until after the age of 5. All participants were students at the University of Texas at Austin.

Stimuli. The whole utterances used in Experiment 2 were employed again in Experiment 3. Recall that in Experiment 2, iambic and trochee "Nana" tokens were manipulated, via Praat (Boersma & Weenink, 2012), to create two separate sets of 80 utterances: those containing target words contrasting in duration (i.e., [nana:] vs. [na:na], duration utterance), and those containing target words contrasting in vowel quality (i.e., [nəna] vs. [nanə], vowel quality utterances).

To further ensure that any differences observed in the duration utterance and vowel quality utterance stimuli were due to the surrounding prosodic cues, the manipulated target tokens were excised from the sentences and also presented in isolation to participants (i.e., duration words; vowel quality words). As such, differences present in the duration utterances and vowel quality utterances, which are embedded in the four sentential contexts, but not in isolation (i.e., duration words and vowel quality words), can be attributed to the sentential context. Despite the fact that sentential contexts carry information beyond intonation, such as semantic and syntactic information, a comparison between the same words presented in isolation versus embedded sentences with focal and reporting intonations may constitute a reasonable control for prosodic effects. Results are presented below first for the targets embedded in sentential contexts and then for targets in isolation.

Procedure. In one session of 50 min, participants were given two separate lexical identification tasks: one containing the targets with duration cues, and a second containing the targets with vowel quality cues. The order of presentation was counter-balanced, such that half of the participants received duration targets first, and the other half received vowel quality targets. Within each task, participants listened first to the 80 (40 trochee and 40 iambs) word targets presented in isolation and then listened to the targets embedded into the sentences. Stimuli were randomized and presented in blocks of 20 utterances, with a brief break between blocks. For both tasks, the procedure for Experiment 3 paralleled that of Experiment 1. Participants were instructed to respond as quickly as possible via response pad when hearing the target word with trochaic stress, while ignoring those with iambic stress. Again, participants were only given the stimuli corresponding to their native language.

4.2 Results

As in Experiment 1, *d′* scores were calculated as the difference of *z* scores between hits (key pressed on hearing the iambic target word) and false alarms (key pressed on hearing the corresponding trochee) in the sentence stimuli. Out of 40 possible answers based on targets with duration cues in sentences, English participants obtained a mean of 35.3 hits (*SD* = 4.3) and Spanish, 32.09 (*SD* = 7.2). As for false alarms, English speakers scored a mean of 10.2 (*SD* = 7.1) and Spanish speakers, 9.2 (*SD* = 9.1). With regard to the 40 possible targets with vowel quality cues in sentences, English participants scored a mean of 35.3 hits (*SD* = 5.03) and Spanish, 31.08 (*SD* = 7.1). As for false alarms, English speakers scored a mean of 9.08 (*SD* = 6.5) and Spanish speakers, 9.5 (*SD* = 9.10). The one English and three Spanish participants who pressed the key for both the iambic and trochee "Nana" across tasks were eliminated from further analysis. For the remaining participants, *d′* scores for sentence stimuli averaged 1.8 and 2 in the duration and vowel quality targets, respectively, for English speakers and 2 and 1.6, respectively, for Spanish speakers. As expected, participants' sensitivity to target words decreased in Experiment 3 in comparison to Experiment 1. This difference is not unexpected, as targets in Experiment 3 were manipulated to contain fewer cues than in Experiment 1.

As for reaction times in the sentence stimuli, responses greater than 2000 ms (*n* = 16) or less than 0 ms (*n* = 47) were eliminated from analysis, yielding a total of 5270 responses. In word stimuli, responses with reaction times over 1200 ms were eliminated from further analysis, leaving a total of 5127 tokens.

A visual inspection of the graphs in Figure 6 shows that when stimuli contained duration cues (duration utterances; see the first four box-plots in two graphs at the top), participants from both language groups evidenced longer latencies when detecting target words at the beginning of reporting sentences (reporting initial, see box-plot in white) relative to the other intonation contexts (see box-plots with line patterns). This pattern could not be observed in the corresponding duration utterance tokens presented in isolation (see the two graphs at the bottom of Figure 6). In contrast, for stimuli containing vowel quality cues to stress (vowel quality utterances), there were no visible effects of sentence intonation in English. Only Spanish speakers obtained longer reaction times in the reporting initial context. Again, no similar patterns were visible on the corresponding isolated word targets.

Hierarchical linear models with the fixed effects of cue (duration and vowel quality), intonation (post-focal medial, post-focal initial, reporting medial, and reporting initial), and order (duration first and vowel quality first) and the random effects of Subject and Item were run separately for each language on sentence stimuli. Results, which are displayed in Table 1, showed that cue and intonation were strongly significant main factors in both languages, suggesting that both the type of acoustic cue available in the signal and the expectations based on prosody had an effect on lexical stress perception in both English and Spanish speakers. However, the two-way interaction of cue × intonation was significant only in English speakers, indicating some cross-language differences in these effects. “No significant order effects” indicated that the order of task presentation did not affect participants’ reaction times.

To further examine the significant main effects and interaction depicted in Table 1, a hierarchical linear model with the fixed factor of intonation and the random effects of subject and item were run separately on each cue and language. Results showed a significant effect of intonation in duration cues for both languages, in English, *F*(3, 1933) = 19.557, $p < 0.0001$ and in Spanish, *F*(3, 690.8) = 16.453, $p < 0.0001$. However, in the vowel quality stimuli, intonation was strongly significant only for Spanish speakers, *F*(3, 565.82) = 16.47, $p < 0.0001$, but not for English speakers, *F*(3, 1910.7) = 2.428, $p = 0.06371$. Thus, intonation had a significant effect on word detection in English and Spanish speakers when stimuli contained duration cues. However, this effect was not consistent across languages when stimuli contained vowel quality cues. Multiple comparisons with the Bonferroni adjustment on the effects of intonation context in duration stimuli (Table 2) showed that for both English and Spanish speakers, reaction times on reporting initial sentences were different from those in reporting medial and post-focal contexts (all *p*’s < 0.0001). However, no differences were observed between reaction times in post-focal contexts and reporting-medial sentences. Altogether, these results showed that both English and Spanish speakers integrated their expectations based on sentence prosody into word detection tasks when duration was the cue present in the target word. In contrast with duration, results from vowel quality cues are not as consistent since intonation effects were significant only for Spanish speakers.

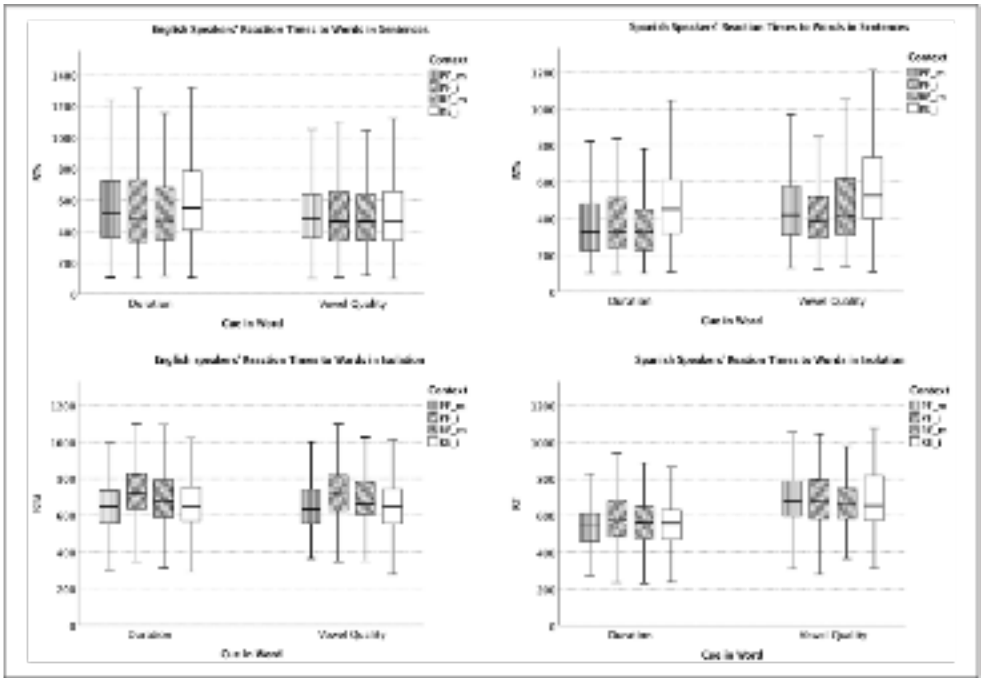
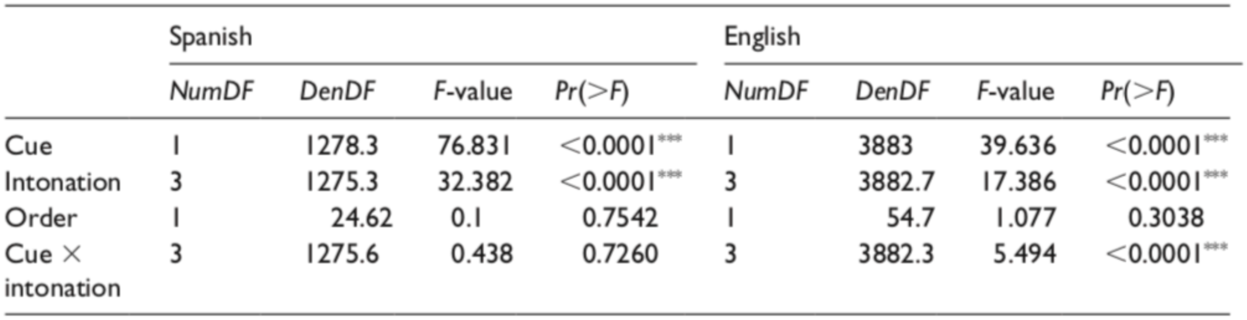


Figure 6. Participants' reaction times to the target word in different intonation contexts while controlling for duration and vowel quality cues. English speakers (two graphs on the left) and Spanish speakers (two graphs on the right) listened to target words containing duration cues and vowel quality cues embedded in sentences (top graphs). These words were excised from sentences and presented in isolation (bottom graphs). Sentences were spoken in four intonation contexts (PF_m, PF_i, RE_m, and RE_i). RTs: reaction times; PF_m: post-focal medial; PF_i: post-focal initial; RE_m: reporting medial; RE_i: reporting initial.

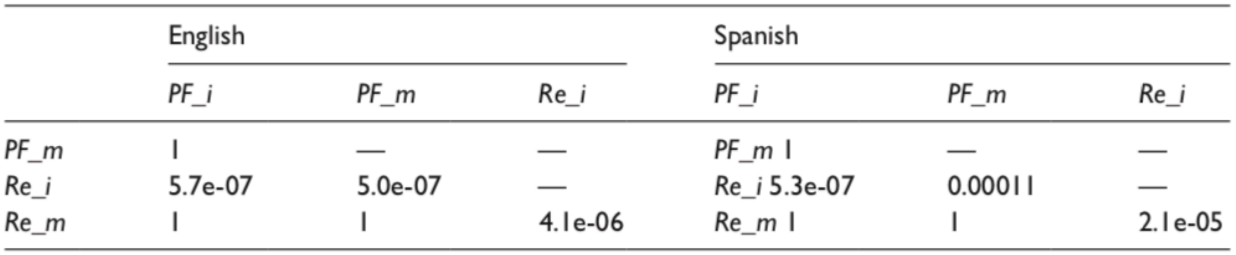
Table 1. Hierarchical linear models on sentence stimuli for English and Spanish.



	Spanish				English			
	NumDF	*DenDF*	*F*-value	*Pr*(>*F*)	*NumDF*	*DenDF*	*F*-value	*Pr*(>*F*)
Cue	1	1278.3	76.831	<0.0001***	1	3883	39.636	<0.0001***
Intonation	3	1275.3	32.382	<0.0001***	3	3882.7	17.386	<0.0001***
Order	1	24.62	0.1	0.7542	1	54.7	1.077	0.3038
Cue × intonation	3	1275.6	0.438	0.7260	3	3882.3	5.494	<0.0001***

NumDF: numerator degrees of freedom; *DenDF*: denominator degrees of freedom; *F*-value: test statistic; *Pr*(>*F*): significance of probability. ****p* = at an alfa level of .05.

Table 2. Comparison across intonation contexts of reaction times to duration targets embedded in sentences.



	English			Spanish		
	PF_i	*PF_m*	*Re_i*	*PF_i*	*PF_m*	*Re_i*
PF_m	1	—	—	*PF_m* 1	—	—
Re_i	5.7e-07	5.0e-07	—	*Re_i* 5.3e-07	0.00011	—
Re_m	1	1	4.1e-06	*Re_m* 1	1	2.1e-05

PF_i: post-focal initial; *PF_m*: post-focal medial; *RE_m*: reporting medial; and *RE_i*: reporting initial.

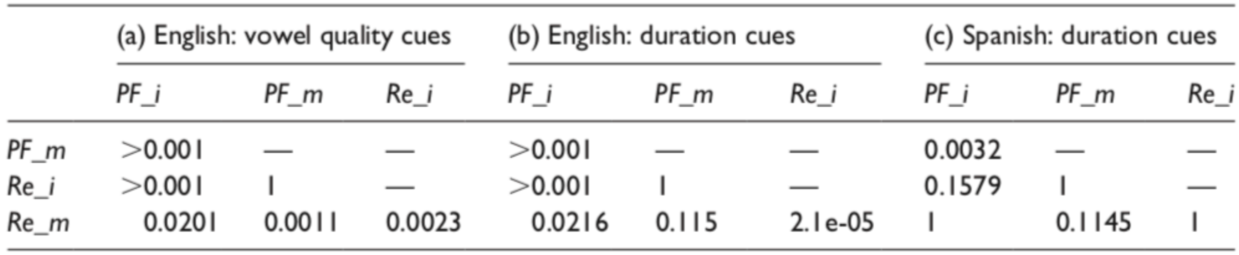
In order to ensure that the above results were indeed related to the effects of expectations based on sentence prosody and not to any possible idiosyncrasies present in the target words, a second hierarchical linear model was performed as well for the reaction times obtained from the same target words presented in isolation. This model, which included the fixed factor of intonation and the random effects of Subject and Item, was performed separately for cue and language. Results showed a significant effect of intonation for English words with both vowel quality cues, $F(3, 1752.8) = 33.78$, $p < 0.0001$, and duration cues, $F(3, 1485) = 2078$, $p < 0.0001$. A significant effect of intonation was shown for Spanish words with duration cues, $F(3, 972.9) = 5.154$, $p = 0.0015$. Multiple comparisons with the Bonferroni adjustment showed that when words with duration cues were presented in isolation, reaction times for the reporting-initial context did not differ from those in the other three contexts in Spanish—see Table 3(c)—and only from the post-focal medial context in English—see Table 3(b). These results contrast sharply with those obtained with the same words embedded in sentences where reaction times for the reporting initial context differed significantly from those in the other contexts in both English and Spanish. Similarly, the results with words containing vowel quality cues differed sharply when presented in isolation or embedded within sentences. While in the sentence context the effect of intonation was non-significant, the same words presented in isolation obtained significant results for all paired comparisons—see Table 3(a). These consistent reaction time differences between words presented in sentences and the same words presented in isolation confirm that results obtained in the sentence condition were related to the effects of expectation-based sentence prosody rather than to any possible idiosyncrasies of the target words.

4.3 Discussion

Two main results emerge from Experiment 3. First, the significant effects of Intonation obtained in target words embedded in sentences differed from those obtained in target word presented in isolation, confirming that the former effects were related to the expectations based on speakers' knowledge of intonation.

Second, when listening to the target words embedded in sentences, cue manipulations had a clear effect on participants' responses, especially in English speakers. When detecting [nana:] based on duration differences, they obtained slower reaction times in the reporting initial context, the only context containing no cues on which to build prosodic expectations in comparison to the other three intonation contexts (which all contained cues on which to build prosodic expectations), showing that they used suprasegmental cues and built expectations based on their knowledge of sentence intonation. However, when detecting [nəna] based on vowel quality differences, these speakers did not build expectations based on sentence prosody. Similar results were obtained for Spanish speakers in targets containing duration cues, where they clearly used prosodic expectations. These results support the second part of the CDWL hypothesis in that duration cues, which are processed in longer windows, promoted the generation of expectations based on sentence intonation in both English and Spanish speakers. However, vowel quality cues, which require shorter windows, did not, at least in English speakers.

Table 3. Comparison across intonation contexts of reaction times to vowel quality and duration targets.



	(a) English: vowel quality cues			(b) English: duration cues			(c) Spanish: duration cues		
	PF_i	*PF_m*	*Re_i*	*PF_i*	*PF_m*	*Re_i*	*PF_i*	*PF_m*	*Re_i*
PF_m	>0.001	—	—	>0.001	—	—	0.0032	—	—
Re_i	>0.001	1	—	>0.001	1	—	0.1579	1	—
Re_m	0.0201	0.0011	0.0023	0.0216	0.115	2.1e-05	1	0.1145	1

PF_i: post-focal initial; *PF_m*: post-focal medial; *RE_m*: reporting medial; and *RE_i*: reporting initial.

There is one caveat that needs to be addressed in further research, namely why Spanish speakers failed to associate differences in vowel quality with a contrast between trochee and iamb. In the final debriefing Spanish speakers reported expecting to hear "['nena], [ne'na], ['nane], or [na'ne]," when the only possible options were ['nena] and [na'ne]. This lack of association between vowel quality and the trochee-iamb contrast, which has been found in previous literature as well (e.g., Flege & Bohn, 1989), casts doubt that Spanish speakers used vowel quality cues to perform this task. One might speculate that Spanish speakers maintained two-syllablelong processing windows and turned to any residual intensity cues (e.g., recall that intensity differences were reduced to 2 dBs or less).

5 General discussion

In summary, results from Experiment 1 showed that indeed, there is a cross-language difference in the use of sentence intonation to predict upcoming flat-F0 hypo-articulated utterances; that is, Spanish speakers more effectively anticipated these utterances than English speakers. Then, results from Experiments 2 and 3 showed that this cross-language difference was accounted for by the CDWL hypothesis. Participants in Experiment 1, based on their language knowledge, scanned the unfolding speech signal in search of the acoustic information that was relevant to the task at hand. Since the goal was to detect iambic "Nana" as fast and as reliably as possible, speakers choose the acoustic cues that, in the shortest possible time, gave sufficient information to make a reliable decision. For example, in the naturally spoken sentences of Experiment 1, Spanish speakers were relying on duration cues (e.g., [nana:]) and English speakers on vowel reduction cues (e.g., [nəna]) to perceive the iambic target. In doing so, the length of the processing window was set to ensure maximum efficiency. In order to detect the target word [nəna] in natural speech, English speakers scanned the speech signal in onesyllable windows to detect schwas and full vowels. In contrast, Spanish speakers had to rely on the available duration cues (i.e., [nana:]) to detect the iambic word. Because duration is a relative measure and a syllable is long only in relation to its adjacent syllable (e.g., Massaro, 1984), a two-syllable window becomes the appropriate minimal length to scan the unfolding speech signal in search of the short–long pattern of iambs. Thus, English and Spanish speakers adjusted the length of the processing window according to the acoustic information that, in their respective native language, allowed the fastest and most reliable detection of the target word.

However, there is an alternative interpretation to the results of Experiment 1, namely, speakers adjusted the length of the processing window according to their L1 processing routines rather than the cues present in the signal; that is, English speakers tend to use shorter one-syllable windows and Spanish speakers longer two-syllable windows regardless of the cues present in the speech signal. Results from the gating task of Experiment 2 ruled out this possible interpretation. While speakers from both languages consistently used two-syllable-long windows when processing target words containing only duration cues to stress, speakers switched to one-syllable windows when target words contained only vowel quality cues to stress. Thus, results from Experiment 2 showed that adjusting the window length was contingent to the cues present in the speech signal rather than to the speakers' L1 processing routines, providing evidence in support of the first tenet of the CDWL hypothesis.

The second tenet of CDWL hypothesis says that, in setting the window length according to the cues present in the speech signal, speakers regulate the type of top-down information that is amenable to interpretation. For example, when scanning a signal in one-syllable windows to detect schwas versus full vowels, the interpretation of suprasegmental cues, such as duration, intensity, or pitch becomes impossible because these are relative cues that require at least two syllables to be interpreted. That is, a syllable is perceived as long or short in contrast to the length of its adjacent syllables. In contrast, the strategy of scanning the speech signal in two-syllable windows to capture duration contrasts of iambic “Nana” makes these windows long enough to interpret not only iambic “Nana” but also a variety of prosodic cues relevant to sentence intonation such as contrastive pitch accents, flat-F0 utterances, and pitch compression. Thus, two-syllable windows are long enough to interpret suprasegmental cues into units relevant to sentence intonation, making it possible to predict upcoming F0-flat hypo-articulated utterances, whereas one-syllable processing windows do not. Experiment 3 provided supporting evidence for this second tenet. In Experiment 3, participants listened to the same sentences of Experiment 1, but the cues of target words were manipulated so that they contained either only vowel reduction or only duration cues to stress. Results showed that when listening to target words with duration cues, both Spanish and English speakers detected iambic “Nana” faster in hypo-articulated utterances that could not be anticipated by intonation cues reducing the cross-language differences obtained with the same unmodified sentences of Experiment 1. Since intonation modulates cues to stress, predicting incoming hypoarticulated utterances allows re-weighting cues to stress accordingly, facilitating in this way the detection of iambic “Nana.” Partial evidence was obtained from sentences containing vowel reduction cues to stress, since only English speakers showed no effect of intonation. Spanish speakers showed these effects by detecting “Nana” more slowly in the reporting sentences with initial “Nana.” Since there were no duration cues, we speculate that Spanish speakers may have based their answers on residual intensity cues.

Altogether, the above results showed that there is a fine interplay among the task at hand, the acoustic cues present in the speech signal, the generation and access to expectations based in prosodic knowledge, and the length of the processing window. Bottom-up acoustic information interacts with top-down information to perform a particular task such that once the length of the processing window is adjusted to efficiently process the cues in the speech signal relevant to the task, this length modulates which additional high-level information is interpretable. Thus, results from the three experiments showed that this mechanism, which constitutes the core of CDWL hypothesis, explains prosodic processing at the utterance level. Furthermore, this same mechanism accounts for prosodic processing at lexical level. As explained in CDWL Hypothesis: A Mechanism to Account for CrossLanguage Differences in Prosodic Processing, the CDWL hypothesis motivates Cutler and colleagues' trade-off between vowel reduction and suprasegmental cues to stress during word recognition processes. Because vowel reduction is the relevant cue to stress in English, and duration and pitch in Dutch and Spanish, the CDWL hypothesis predicts that English speakers will process words in onesyllable windows and Dutch and Spanish speakers in two-syllable windows. In setting the window length, the CDWL hypothesis predicts the cross-language trade-off between segmental and suprasegmental cues, namely, using one-syllable windows to detect stress in words, English speakers are set to use suprasegmental cues to stress less efficiently than Dutch and Spanish speakers. Thus, the CDWL hypothesis explains prosodic processing in both words and utterances by showing how bottom-up and top-down information interact in a complex bi-directional flow.

5.1 Theoretical implications and limitations of the study

The above interactions have important implications for models of speech perception. These interactions illustrate a constant feedforward/feedback among task goals, acoustic information in the speech signal, and the generation and use of prosodic expectations, which in turn determine the optimal length of the processing window for maximum efficiency. These bi-directional interactions between higher-level and lower-level processes challenge the traditional view of speech perception as an automatic process of pattern matching between the incoming speech signal and a stored phonological representation, where activated candidates passively percolate onto higher order operations. Instead, these interactions indicate that speech perception is an active process that entails real-time adjusting to feedback and to information from the context. A growing body of research supporting this view of speech perception comes from computational modeling and neuroscience. For example, the C-Cure model proposed by Murray and Jongman (2011), where cues are interpreted relative to expectations, obtained better results than two models which excluded expectations and contextual compensations. Similarly, the analysis by synthesis models (AxS)—which were proposed by Stevens and Halle in the 1960s and are now revisited by Poeppel and Monahan (2011)—include a hypothesis-and-test circuit that, for example, could account for the on-line readjusting of the processing window length to the type of cues present in the signal. Moreover, neuroanatomy research showed that expectations based on higher-level knowledge altered low-level processing in the auditory brainstem (e.g., Galbraith & Arroyo, 1993) or even in the cochlea (e.g., Giard et al., 1994). Recent research also showed that regularity encoding and deviance detection are modeled by stimuli complexity along the auditory pathway, starting from the brainstem with less complex stimuli ascending to the auditory cortex (Escera, Leung & Grimm, 2014). Altogether, this evidence supports the plausibility of a speech perception model that, in addition to a passive bottom-up path, also includes a top-down path where feedback signals from the cortical level change processing in real-time at lower levels according to context and expectations (e.g., Heald & Nusbaum, 2015). Although models of word recognition (e.g., McClelland & Rumelhart, 1985; Seidenberg & McClelland, 1989) and sentence comprehension (e.g., Norris, 1994) have included a dynamic, feedforward approach in which context restricts the possible set of targets, speech perception has been considered a passive, bottom-up pattern-matching process (e.g., Heald & Nusbaum, 2014). In contrast, the current study proposes a dynamic, highly interconnected framework for speech perception.

There is a final caveat that would deserve further exploration. English speakers showed a high degree of flexibility in the bi-directional communication between higher-level knowledge and the acoustic cues to stress present in the speech signal. To illustrate, in Experiment 3 English speakers were able to successfully use either duration or vowel quality cues in the speech signal. In doing so, they adjusted the length of the processing window accordingly, which led to either the integration in two-syllable windows, or to the exclusion in one-syllable windows, of prosodic expectations into stress detection. In contrast, Spanish speakers did not show that flexibility. Although they were successful in detecting stress based on duration cues, Spanish participants had trouble associating vowel quality to stress, that is, [e] with a stressed vowel and [a] with an unstressed vowel.

This trouble was made clear by the participants' comments during the final debriefing. For example, one subject stated "I was not sure if I was hearing iambic or trochaic [nena] or [nane]", when the only possible options were trochee [nena] and iambic [nane]. Similarly, the instructions in Experiment 2 had to be simplified to a vowel quality contrast instead of a vowel quality in relation to stress, that is, "press the key when you hear [nane] not [nena]," instead of "press the key when you hear the iambic realization of the word minimal pair trochee [nena] versus iambic [nane]." Although this difficulty in associating vowel quality to stress appears to be reasonable in our test materials due to its unnaturalness—that is, within a sentence, vowel quality was independent from stress in all words except for the target word—previous research has also encountered similar difficulties for Spanish learners of English (e.g., Flege & Bohn, 1989). In English stimuli, however, the relation between vowel quality and stress is pervasive and reliable, showing that Spanish speakers' lack of flexibility to use vowel quality as a cue to stress is not related only to the unnaturalness of our stimuli. More research is needed to understand, in depth, this lack of cue-flexibility in Spanish speakers. Although it is true that vowel quality as a cue to stress is pervasive in English, whereas at best, it is marginal in Spanish (there is residual vowel reduction in unstressed vowels in some dialects; see Canellada & Kuhlman-Madsen, 1987; Delforge, 2008; Lipski, 1990), the reasons that make the use of vowel quality as a cue to lexical stress particularly difficult to Spanish speakers are not yet fully understood.

6 Conclusion

The CDWL hypothesis assumes that speech is processed within windows whose lengths are adjusted according to the acoustic cues in the speech signal that are relevant to the task at hand. This mechanism, which underlies the processing of suprasegmental and segmental features of speech, offers an explanation to observed cross-language asymmetries between English and Spanish in the processing of prosody at the lexical and utterance levels. It was observed by Cutler and colleagues (Cooper et al., 2002; Tyler & Cutler, 2009; van Donselaar et al., 2005) that in word recognition, English speakers processed suprasegmental cues to stress less efficiently than speakers of Dutch and Spanish. Furthermore, results from Experiment 1 showed that English speakers were less efficient than Spanish speakers in processing suprasegmental cues in relation to utterance intonation. The CDWL hypothesis explains these asymmetries by showing that adjusting the length of the processing window to the relevant cue determined how efficiently suprasegmental cues to lexical and utterance prosody were processed. If the task required two-syllable windows to process suprasegmental cues to stress, suprasegmental cues to utterance intonation were also amenable to interpretation. However, if speakers adjusted the processing window length to one-syllable in order to detect schwas and full vowels, suprasegmental cues of duration and pitch were not interpretable either in relation to stress or utterance intonation. English speakers were able to adjust the processing window length to both vowel quality and duration cues according to the task requirements providing strong evidence for the CDWL hypothesis. Spanish speakers, however, were better at adjusting the window length to two syllables in order to process suprasegmental cues to stress and utterance intonation. They were not capable of adjusting window length to one syllable in order to process vowel reduction cues to stress, showing that the CDWL hypothesis has some limitations that need to be addressed in the future. Overall, these results have important consequences for models of speech processing. By showing that highly interactive speech processing models with bi-directional flows of information, like the CDWL hypothesis, are able to account for cross-language asymmetries in prosodic processing, this study adds to the cumulative evidence that advocates for highly dynamic interconnected models of speech perception instead of more traditional feedforward passive pattern-matching models.

Acknowledgements

We would like to thank the participants who collaborated in the perception task and Professor Scott Fraundorf for his advice on statistics.

Funding

This work was supported by the Central Research Development Fund Grant, University of Pittsburgh and the Research Grant Competition, College of Liberal Arts, University of Texas awarded to the first author.

Note

1. The difference in correct identification probabilities in English and Spanish for vowel quality tokens likely owes to the degree of salience of the acoustic cues in the two languages. Specifically, Spanish stimuli, comparing [a] and [e] vowels, had a greater acoustic difference than the English stimuli, comparing [a] and [ə], a difference of approximately 100 Hz in F1 height. Another alternative explanation is that [e] and [a] have different graphemes in Spanish whereas full [a] and "a schwa" share the same grapheme in English.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package Version, 1*, 1–23.

Beckman, M. E. (1986). *Stress and non-stress accent*. Dordrecht, Netherlands: Foris.

Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 7– 33). Cambridge, UK: Cambridge University Press.

Boersma, P., & Weenink, D. (2012). Doing phonetics by computer [Computer program]. *Version, 5*, 52.

Borzone de Manrique, A. M., & Signorini, A. (1983). Segmental durations and the rhythm in Spanish. *Journal of Phonetics*, *11*, 117–128.

Canellada, M. J., & Kuhlman-Madsen, J. (1987). *Pronunciación del Español. Lengua hablada y literaria*. Madrid, Spain: Editorial Castalia.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech*, *45*, 207–228.

Cutler, A. (1986). Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech*, *29*, 201–220.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*, 141–201.

Cutler, A. (2005). Lexical stress. In D. Pisoni, & R. Remez (Eds.), *The handbook of speech perception* (pp. 264–289). Oxford, UK: Blackwell.

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

Delattre, P. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *4*, 183–198.

Delforge, A. M. (2008). Unstressed vowel reduction in Andean Spanish. In L. Colantoni & J. Steele (Eds.), *Selected proceedings of the 3rd conference on laboratory approaches to Spanish phonology* (pp. 107– 124). Somerville, MA: Cascadilla Proceedings Project.

Escera, C., Leung, S., & Grimm, S. (2014). Deviance detection based on regularity encoding along the auditory hierarchy: Electrophysiological evidence in humans. *Brain Topography*, *27*, 527–538.

Fant, G. (1967). Sound, features and perception. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QSPR)*, *8*, 1–14.

Flege, J. E., & Bohn, O. S. (1989). An instrumental study of vowel reduction and stress placement in Spanish accented English. *Studies in Second Language Acquisition*, *11*, 35–62.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America*, *27*, 765–768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, *1*, 126–152.

Galbraith, G. C., & Arroyo, C. (1993). Selective attention and brainstem frequency-following responses. *Biological Psychology*, *37*, 3–22.

Giard, M. H., Collet, L., Bouchet, P., & Pernier, J. (1994). Auditory selective attention in the human cochlea. *Brain Research*, *633*, 353–356.

Heald, S. L., & Nusbaum, H. C. (2015). Speech perception as an active cognitive process. The effect of hearing loss on neural processing. *Frontiers in Systems Neuroscience*. DOI: 10.3389/fnsys.2014.00035.

Hualde, J. I. (2005). *The sounds of Spanish*. New York: Cambridge University Press.

Huss, V. (1978). English word stress in the post-nuclear position. *Phonetica, 35*, 86–105.

Joliot, M., Ribary, U., & Llinas, R. (1994). Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. *Proceedings of the National Academy of Sciences*, *91*, 11748–11751.

Jusczyk, P. W. (2000). Wrapping things up. In *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kilborn, K., & Moss, H. (1996). Word monitoring. *Language and Cognitive Processes*, *11*, 689–694.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The tracking of speech envelope in the human cortex. *PloS ONE*, *8*, e53398.

Ladd, D. R. (2008). *Intonational phonology, 2nd Ed*. Cambridge, MA: Cambridge University Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Dordrecht, Netherlands: Springer.

Lipski, J. (1990). Aspects of Ecuadorian vowel reduction. *Hispanic Linguistics, 4*, 1–19.

Marinis, T. (2010). Using on-line processing methods in language acquisition research. In S. Unsworth & E. Blom (Eds.), *Experimental methods in second language research* (pp. 139–162). Philadelphia, PA: John Benjamins.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1–71.

Massaro, D. W. (1984). Time's role for information, processing, and normalization. *Annals of the New York Academy of Sciences*, *423*, 372–384.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*, 159–188.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*, 219.

Mirman, D. (2017). Zones of proximal development for models of spoken word recognition. In M. G. Gaskell & J. Mirkovic (Eds.), *Speech perception and spoken word recognition*. Psychology Press.

Näätänen, R. (1992). *Attention and brain function*. Hillsdale, NJ/London: Erlbaum/Psychology Press.

Nadeu, M. (2013). *The effects of lexical stress, intonational pitch accent and speech rate on vowel quality in Catalan and Spanish*. PhD Thesis, University of Illinois at Urbana-Champaign.

Navarro-Tomás, T. (1974a). *Manual de pronunciación Española*. Madrid, Spain: Publicaciones de la Revista de Filología Espanola.

Navarro-Tomás, T. (1974b). *Manual de entonación Española* (Vol. 175). Madrid, Spain: Ediciones Guadarrama.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

Nusbaum, N. S., & Henly, A. S. (1992). Listening to speech through an adaptive window of analysis. In M. E. H. Schouten (Ed.), *The auditory processing of speech: From sounds to words* (p. 339). Berlin, Germany: Walter de Gruyter.

Ortega-Llebaria, M. (2006). Phonetic cues to stress and accent in Spanish. In M. Díaz-Campos (Ed.), *Selected proceedings of the 2nd conference on laboratory approaches to Spanish phonetics and phonology* (pp. 104–118). Somerville, MA: Cascadilla Proceedings Project.

Ortega-Llebaria, M., Gu, H., & Fan, J. (2013). English speakers' perception of Spanish lexical stress: Context-driven L2 stress perception. *Journal of Phonetics*, *41*, 186–197.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. PhD Thesis, MIT, Cambridge, MA.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*, 253–260.

Poeppel, D., & Monahan, P. J. (2011). Feedforward and feedback in speech perception: Revisiting analysis by synthesis. *Language and Cognitive Processes, 26*, 935–951.

Quené, H., & Koster, M. L. (1998). Metrical segmentation in Dutch: Vowel quality or stress? *Language and Speech*, *41*, 185–202.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*, 265–292.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London B*, *336*, 367–373.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, *55*, 349–374.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language*, *45*, 412–432.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. WathenDunn (Ed.), *Models for the perception of speech and visual form* (pp. 88–102). Cambridge, MA: MIT Press.

Torreira, F., Simonet, M., & Hualde, J. I. (2014). Quasi-neutralization of stress contrasts in Spanish. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Proceedings of 7th international conference on speech prosody* (pp. 197–201). Dublin: Trinity College Dublin.

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America*, *126*, 367–376.

van Donselaar, W., Koster, M., & Cutler, A. (2005). Exploring the role of lexical stress in lexical recognition. *The Quarterly Journal of Experimental Psychology Section A*, *58*, 251–273.

Vanrell Bosch, M. D. M., & Fernández Soriano, O. (2013). Variation at the interfaces in Ibero-Romance. Catalan and Spanish prosody and word order. *Catalan Journal of Linguistics*, *12*, 253–282.

Viemeister, N. F., & Wakefield, G. H. (1991). Temporal integration and multiple looks. *The Journal of the Acoustical Society of America*, *90*, 858–865.

White, L., & Mattys, S. L. (2007). Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), *Segmental and prosodic issues in romance phonology, Current issues in linguistic theory*, Vol. 306 (pp. 237–257). Amsterdam, Netherlands and Philadelphia, PA: John Benjamins.

Yabe, H., Tervaniemi, M., Reinikainen, K., & Näätänen, R. (1997). Temporal window of integration revealed by MMN to sound omission. *Neuroreport*, *8*, 1971–1974.

Zubizarreta, M. L. (1998). In S. J. Keyser (Ed.), *Prosody, focus, and word order*. Cambridge, MA: MIT Press.

Appendix 1. Stimuli sentences in Spanish and English

Spanish

['nana], post-focal medial
¿A qué no sabes qué pasó? ROSA le regaló a ['nana] un collar.
Es cierto. ANTONIO llevo a ['nana] al colegio.
¿Ana María? No, ANDREA conoce a ['nana] Morera.
La hermana de Isabel, no. La hermana de MARIA no se habla con ['nana] desde hace años. No le cantó, sino que BAILÓ con la hermana de ['nana] toda la noche.

¿Tristes? No, saludaban SONRIENTES a la madre de ['nana] y Miguel.
El cuaderno viejo no. Mi amiga le entregó el LIBRO viejo a ['nana] Marti.
A sus padres, les gusto MUCHO el regalo de ['nana] y Miguel.
No se si sera cierto, pero me dijeron que Lola le PEGO al hermano de ['nana] Moron. Mañana no, HOY le damos a ['nana] su regalo de graduacion.

[na'na], post-focal medial
No te imaginas quien fue? JOSE le pidio a [na'na] doscientos dolares.
Es verdad. GUSTAVO fue a visitar a [na'na] a Madrid.
¿Carmen? LAURA es la prima de [na'na] Pereda.
Creo que mi amiga le PRESTO una falda a [na'na] Soler, no se la pidio.
Si, es asi. La madre de CARMEN es amiga de [na'na] Rosales.
Este fin de semana, Berta ira con sus PRIMOS a casa de [na'na] Vidal, no con sus tios.
No rompieron. Roberto SIGUE de novio con [na'na] Salinas.
Si, a Roberto SE LE ROMPIO el collar de [na'na] Martinez.
Elena Morera no, Elena PARERA es la prima de [na'na] y Lili.
Esta semana no podemos ir a su casa, pero el JUEVES le damos a [na'na] el regalo de cumpleaños.

['nana], post-focal initial
Sí, estoy segura. Mi amiga le entregó las LLAVES a ['nana] Vidal, no el auto. ¿Tristes? No, saludaban SONRIENTES a ['nana] Vidal y a su madre.
La madre de gustavo conoce MUY BIEN a ['nana] Dominguez.
¿Qué dices? La VECINA de ['nana] estaba con nosotras, no la sobrina.
No, la amiga no. La TIA de ['nana] llegó ayer.
La SOBRINA de ['nana] está casada con Juan, no la prima.
No te lo vas a creer. Me dijeron que Lola le GRITO a ['nana] Vidal.
Si, es cierto. Hoy, Andres SE FUGO con ['nana] Solinas.

Se que Miguel BAILO con ['nana] toda la noche.
Mañana, SE GRADUA ['nana] de la universidad.

[na'na], post-focal initial
Sí, estoy segura. Mi amiga le pidió la BOLSA a [na'na] Vidal, no el libro.
No parecian enfadados, sino CONTENTOS con [na'na] Pereda.
El tio no. La HERMANA de [na'na] conoce a mi padre.
¿Qué dices? La AMIGA de [na'na] estaba con nosotras, no su vecina.
Creo que no conoce al hermano, sino a la HERMANA de [na'na] Dominguez. A que no sabes quien llego. La TIA de [na'na] Morales.

La HERMANA de [na'na] comparte piso conmigo, no su prima.
Este verano, los PADRES de [na'na] vendran de vacaciones con nosotros, no sus hermanos. No es cierto que esten enfadadas. Mercedes HABLO con [na'na] toda la noche.
Ella faltar a clase? Al contrario, mi hija SE PRESENTO con [na'na] Toledo al examen.

['nana], reporting medial
¿Mañana?—le pregunta ['nana] a su amiga.
Hoy,—le digo a ['nana] Martinez—vi a tu mamá en la tienda.
¿Alegres?—me pregunta ['nana] con sorpresa—¿Cómo pueden estar alegres?
No le cantó—nos cuenta ['nana] Sabater—sino que BAILÓ con Elena Martí toda la noche.

¿A qué no sabes qué pasó?—nos pregunta ['nana] Sampere—Rosa me regaló la mermelada de fresa.
Ayer,—me explica ['nana] Solina—la madre de Berta conoció a sus primos.
No te lo vas a creer—nos dijo ['nana] Vidal—Me contaron que Lola le GRITO a Nuria en la plaza. A sus tíos, les gusto MUCHO el regalo de Maria a Miguel—exclama complacida ['nana] Vidal. Ayer, Laura conoció a los PRIMOS catalanes, no a los abuelos—nos informa ['nana] Rosales. Maria GARCIA es la tía de Juana, no Maria Masse—les aclara ['nana] a sus padres.

[na'na], reporting medial
Mañana vendre a comer—les avisa [na'na] Perales a sus padres.
¿Laura?—le pregunta [na'na] a su amiga.
Su tía no—me aclara [na'na] Martinez—la MADRE de Maria se casó con Gustavo.
No es cierto—exclama [na'na] Soler—La vecina de CARMEN estaba con nosotras, no la de Paco. ¿Sabes quien fue?—le dijo [na'na] contenta—ROSA me regaló el collar de la abuela.
Si.—afirma convencida [na'na] Vidal—Ella es la TIA de Juana, no la prima.
Es increible—exclama enfadada [na'na] Segura—Nuria ROMPIÓ con Jorge.
Mi madre agradeció MUCHISIMO el regalo de Nuria Soler—me cuenta [na'na] complacida. Hoy, Marta visito a sus PRIMOS, no a sus abuelos—nos explica [na'na] Perales.
La esposa de Antonio se llama Maria SOLANA, no Maria Solis—nos aclara [na'na] Martinez.

['nana], reporting initial
¿María?—['nana] me pregunta sorprendida.
Mañana,— ['nana] me advierte—iremos a visitar a tu tía.
¿Qué dices?—['nana] Pineda exclamo sorprendida—Jaime bailó conmigo toda la noche.
A su padre, no.— ['nana] me aclara—Se parece a su madre.
Estás en lo cierto.—['nana] Moreno nos cuenta—María no se habla con Marta desde hace años. No, la amiga no—['nana] Salinas me explica pacientemente—La TIA de mi madre llegó ayer de Santander.
¿A qué no sabes qué pasó?—['nana] Perales exclamo asustada—Ayer, María se escapó del colegio.
Esta mañana,—['nana] nos contaba—llegó mi hermana de Madrid.
No cenó—['nana] Soler me cuenta—sino que BAILÓ con Elena toda la noche.
Ana VIDAL es mi amiga, no Ana Nadal—['nana] me explica pacientemente.

[na'na], reporting initial
Es cierto—[na'na] Pereda afirma segura.
¿Elvira?—[na'na] me pregunta sorprendida.
Este verano,—[na'na] me advierte—te quedaras estudiando matematicas.
No es cierto.—[na'na] Solis me dijo—Ayer, Jaime fue a visitar a su prima.
Estás en lo cierto.—[na'na] nos confirmo—ELENA no visita a Nuria desde hace años.
No, la amiga no—[na'na] Moreno me explica pacientemente—La TIA de mi madre llegó ayer de Santander.
¿A qué no sabes qué pasó?—[na'na] Martinez exclamo preocupada—Ayer, Maria ingresó en el hospital.
Esta mañana,—[na'na] Rosales nos informo—llegó mi padre de Santander.
A mis abuelos, les gusto mucho el libro de Nuria—[na'na] me conto complacida.
Lola Marti no es mi amiga—[na'na] me explica pacientemente.

English

['nanə], post-focal medial
Do you know who did it? MARY offered ['nanə] that great job.
Today I had more time because TONY took little ['nanə] to the school.
It wasn't Molly. I think ROSIE told the story to ['nanə] Rousseau.
No, not Isabel's sister. MARY's sister agrees with ['nanə]'s position.
He did not talk to her, but he WROTE a long letter to ['nanə] Bartholomew.
Today is not a good day. But TOMORROW I will talk to ['nanə] Vidal.
You have the old book, so they gave the NEW book to ['nanə]'s sister.
My grandparents LOVED your present to ['nanə] Martinez.
That's bad. Mary FORGOT to invite ['nanə] Delgado.
No, you did not hear me. She is sending a LETTER to cousin ['nanə] for her birthday.

[nə'na], post-focal medial
Do you know who it was? JOHN offered [nə'na] the money.
They agree with Mary, but they ABSOLUTELY hate [nə'na]'s opinions.
No, they were not angry. They were HAPPY to meet with [nə'na]'s fiancée. She did not give that shirt to her. I think she LENT that skirt to [nə'na] Moreno. I'll go FISHING with my friend [nə'na] McMillan.
Imagine! Robert managed to ENGAGE cousin [nə'na] in a fun conversation. Yes, he is great. He MANAGED to give [nə'na] her present.
Sam cannot keep quiet. He TOLD Mary and [nə'na] the news.
Helen Smith no, Helen JOHNSON talks with [nə'na] McLuhan.
We'll go on THURSDAY to [nə'na] Delgado's house.

['nanə], post-focal initial
She knows VERY WELL ['nanə] Bartholomew.
This morning, Tony took TWO of ['nanə]'s suits to dry cleaners.
You are right: he does not talk much, but he WROTE ['nanə] the most beautiful letters. She did not meet her brother. She met her SISTER at ['nanə] García's party.
The book is still on my table, so they gave the LETTERS to ['nanə]'s apprentice.
Not Mary but LAURA's ['nanə]'s best friend.
Laura? No, I think she told SUZANNE of ['nanə]'s accident.
My grandparents LOVED ['nanə] Moreno's present.
How funny. Let's TELL ['nanə] what happened.
Yes, Bob is sending a PRESENT to ['nanə]'s sister.

[nə'na], post-focal initial
They were HAPPY with [[nə'na]'s results.
My sister did not come yesterday, but my FRIEND [nə'na] Rousseau.
My parents were amazed at her language abilities. They ENJOYED [nə'na]'s Chinese. He does not talk much. But he WROTE [nə'na] the most beautiful letters.

I do not know her brother, but I met her SISTER at [nəˈna] Garcia’s.
Not Mary but LAURA’s [nəˈna]’s fiancée.
Maya? No, I think she told LINDA of [nəˈna]’s apartment.
My friends LOVED [nəˈna] Moreno’s recital.
How funny. Let’s TELL [nəˈna] the story.
I am not sure how important it is, but Mary FORGOT [nəˈna]’s computer at home.

[ˈnanə], reporting medial
Yes, that’s right—exclaimed [ˈnanə] McLuhan.
Tomorrow?—asked [ˈnanə] excitedly.
Today—I said to [ˈnanə] Fernandez—I saw your mother in Whole Foods. Happy?—asked [ˈnanə] McCormick—They can’t be!
My friends—I explained [ˈnanə] McKenzie—are coming tomorrow. What’s up?—asked [ˈnanə] Bennett.
Stop it!—demanded [ˈnanə] to her friend.
Unbelievable!—exclaimed [ˈnanə] Mc Luhan.
I met your COUSINS, not your uncles—explained [ˈnanə] Mendoza. Laura McKenzie!—called [ˈnanə] Morales.

[nəˈna], reporting medial
Yesterday—I said to [nəˈna] McLuhan—I saw Molly in HEB. Angry?—asked [nəˈna] Moreno—They can’t be!
Are you sure?—asked [nəˈna] Solis—I was here this morning.
How are you doing?—asked [nəˈna] Burnett.
That’s not right—disagreed [nəˈna] Mendoza—Maya lives with her cousin. Awesome!—exclaimed [nəˈna] McLuhan.
They LOVED your speech—clarified [nəˈna] Molina.
Not Rosie, but MARY is here—explained [nəˈna] Mahone.
I met your BROTHER in the store—explained [nəˈna] Delgado.
Maya! What a surprise!—exclaimed [nəˈna] McMillan.

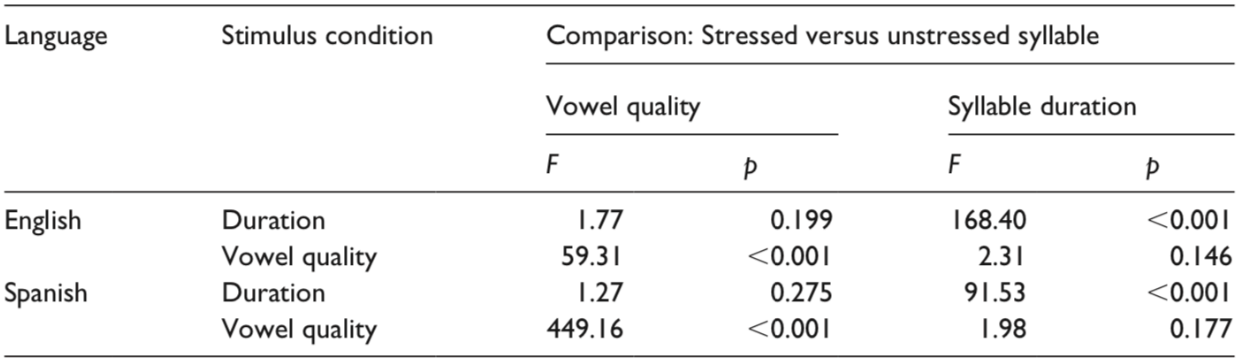
[ˈnanə], reporting initial
Next Monday?—[ˈnanə] complained to her sister.
On Wednesdays—[ˈnanə] Fernandez clarifies—I go JOGGING, not swimming. Go away!—[ˈnanə] Moreno cried.
They HATED my story—[ˈnanə] complained.
Your friends—[ˈnanə] McKenzie told me—are leaving tonight.
What’s up?—[ˈnanə] Fernandez asked.
I don’t think so—[ˈnanə] disagreed—She’s not coming.
I ate the WHOLE cake—[ˈnanə] Mendoza explained.
Look at the roses!—[ˈnanə] exclaimed happily—they are beautiful.
I saw TONI in the store—[ˈnanə] McKenzie exclaimed.

[nə’na], reporting initial
Here is your wallet—[nə’na] McLuhan said with relief.
I’ll be home at ten—[nə’na] García said to her husband.
This Tuesday—[nə’na] Fernandez explained—there is a concert in the park. Now?—[nə’na] complained—Now I am sleeping.
Come home soon!—[nə’na] Gonzales told to her sister.
My colleagues—[nə’na] McLuhan told me—are leaving tonight.
What’s wrong?—[nə’na] McDowell asked.
That’s absolutely right!—[nə’na] Delgado exclaimed—She’s coming tomorrow. We are traveling to CANADA—[nə’na] Mendoza explained.
I called YOU yesterday—[nə’na] McKenzie complained.

Appendix 2. Comparing stressed and unstressed syllables in manipulated tokens

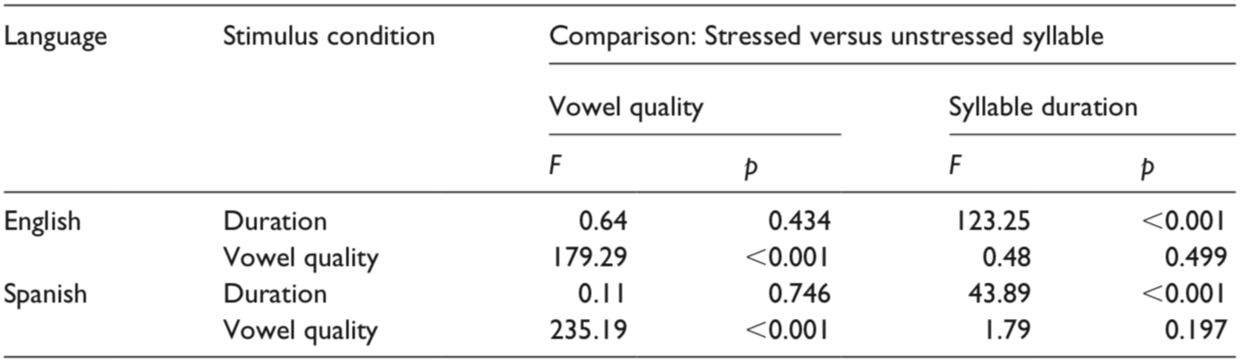
To ensure that manipulations of the target tokens were done appropriately, one-way ANOVAS with stress as the grouping factor were performed comparing syllable duration and vowel quality in the stressed and unstressed syllables of the target tokens. Results, summarized for each context in Tables 4–7, confirmed that for tokens in the duration utterances, only duration, and not vowel quality, yielded statistically significant differences between the stressed and unstressed syllables. Correspondingly, in the vowel quality utterances, vowel quality, and not duration, served to statistically differentiate the stressed and unstressed syllables.

Table 4. Post-focal initial context.



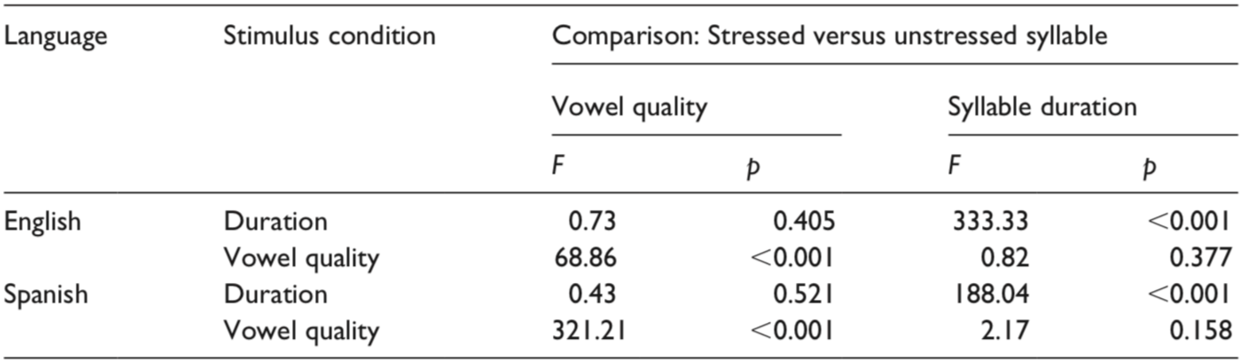
Language	Stimulus condition	Comparison: Stressed versus unstressed syllable			
		Vowel quality		Syllable duration	
		F	*p*	*F*	*p*
English	Duration	1.77	0.199	168.40	<0.001
	Vowel quality	59.31	<0.001	2.31	0.146
Spanish	Duration	1.27	0.275	91.53	<0.001
	Vowel quality	449.16	<0.001	1.98	0.177

Table 5. Post-focal medial context.



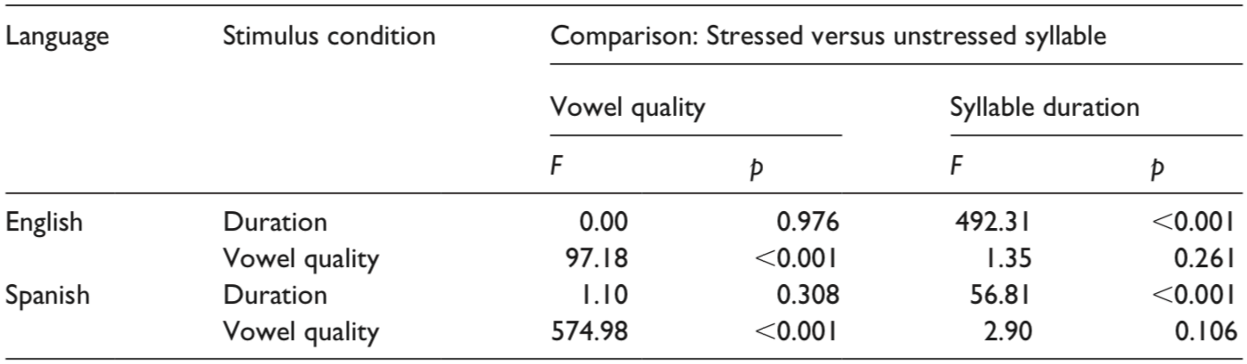
Language	Stimulus condition	Comparison: Stressed versus unstressed syllable			
		Vowel quality		Syllable duration	
		F	*p*	*F*	*p*
English	Duration	0.64	0.434	123.25	<0.001
	Vowel quality	179.29	<0.001	0.48	0.499
Spanish	Duration	0.11	0.746	43.89	<0.001
	Vowel quality	235.19	<0.001	1.79	0.197

Table 6. Reporting initial context.



Language	Stimulus condition	Comparison: Stressed versus unstressed syllable			
		Vowel quality		Syllable duration	
		F	*p*	*F*	*p*
English	Duration	0.73	0.405	333.33	<0.001
	Vowel quality	68.86	<0.001	0.82	0.377
Spanish	Duration	0.43	0.521	188.04	<0.001
	Vowel quality	321.21	<0.001	2.17	0.158

Table 7. Reporting medial context.



Language	Stimulus condition	Comparison: Stressed versus unstressed syllable			
		Vowel quality		Syllable duration	
		F	*p*	*F*	*p*
English	Duration	0.00	0.976	492.31	<0.001
	Vowel quality	97.18	<0.001	1.35	0.261
Spanish	Duration	1.10	0.308	56.81	<0.001
	Vowel quality	574.98	<0.001	2.90	0.106

Appendix 3. F0 pitch tracks in English and Spanish and corresponding sound files

English

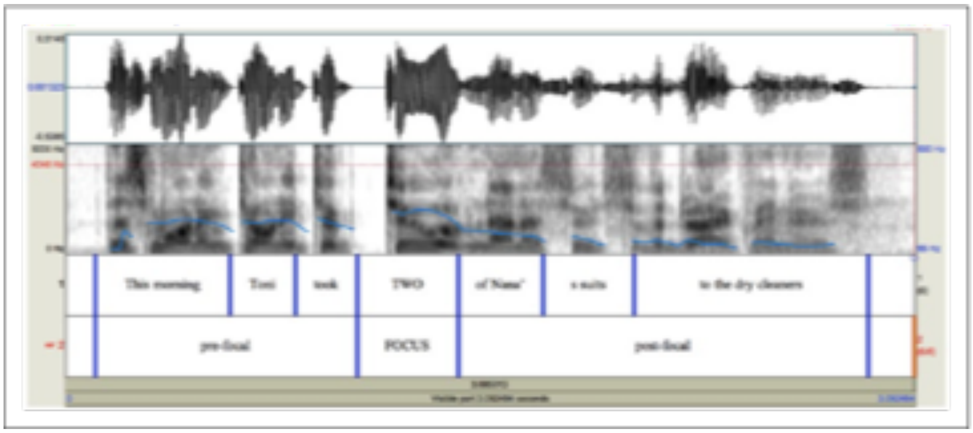


Figure 7. Focus initial: This morning Toni took two of Nana’s suits to the dry cleaners.

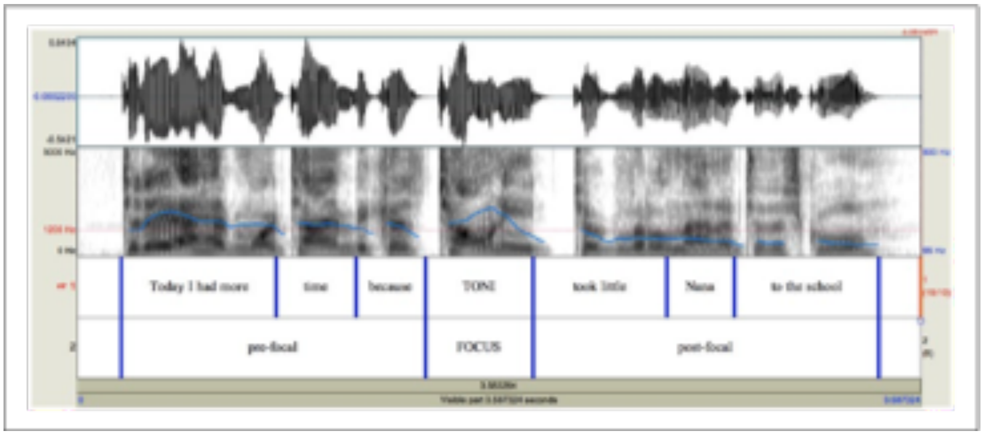


Figure 8. Focus medial: Today I had more time because Toni took little Nana to the school.

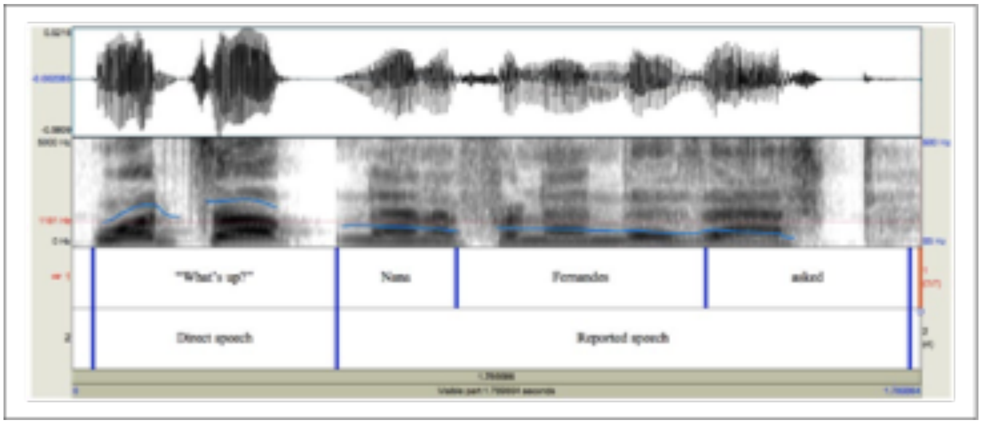


Figure 9. Reporting initial: “What’s up?” Nana Fernandes asked.

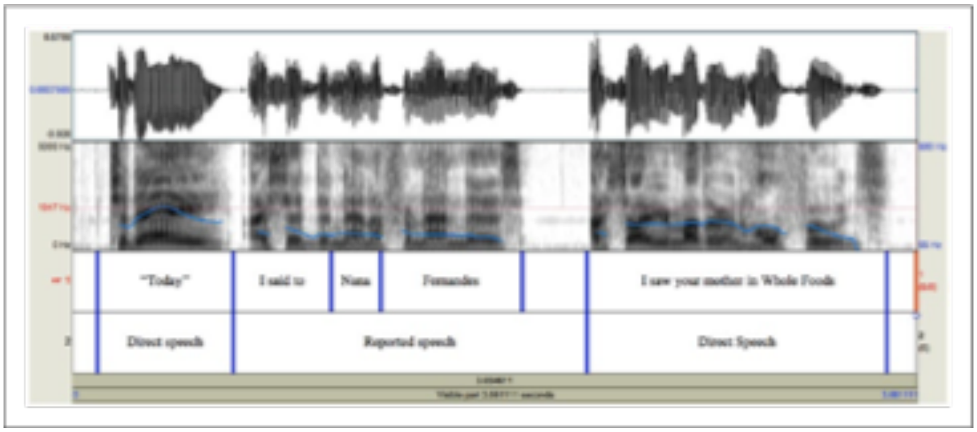


Figure 10. Reporting medial: "Today," I said to Nana Fernandes. "I saw your mother in Whole Foods."

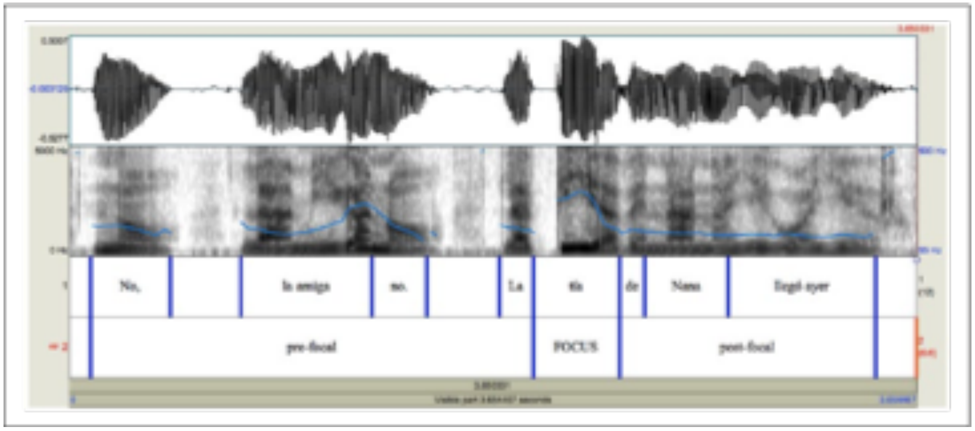


Figure 11. Focus initial: *No, la amiga no. La tía de Nana llegó ayer*.
Not your friend but Nana's auntie arrived yesterday.

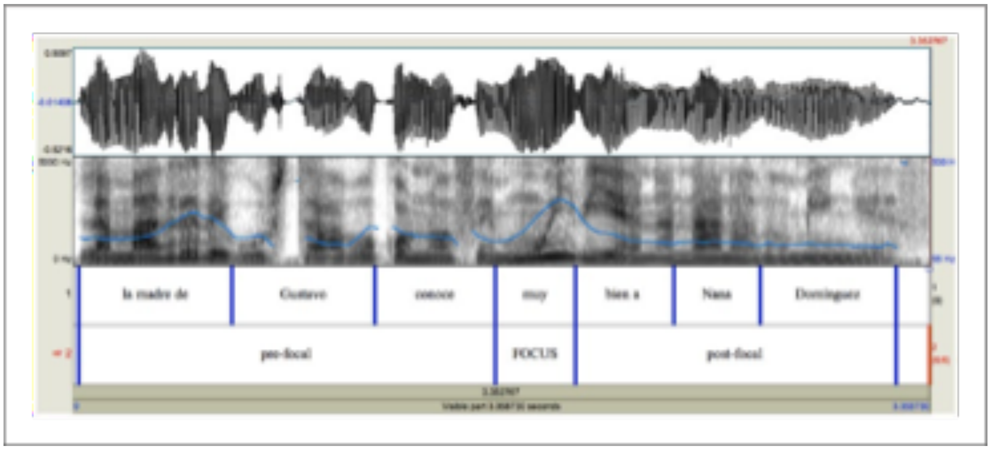


Figure 12. Focus medial: *La madre de Gustavo conoce muy bien a Nana Domínguez*.
Gustavo's mom knows Nana Dominguez very well.

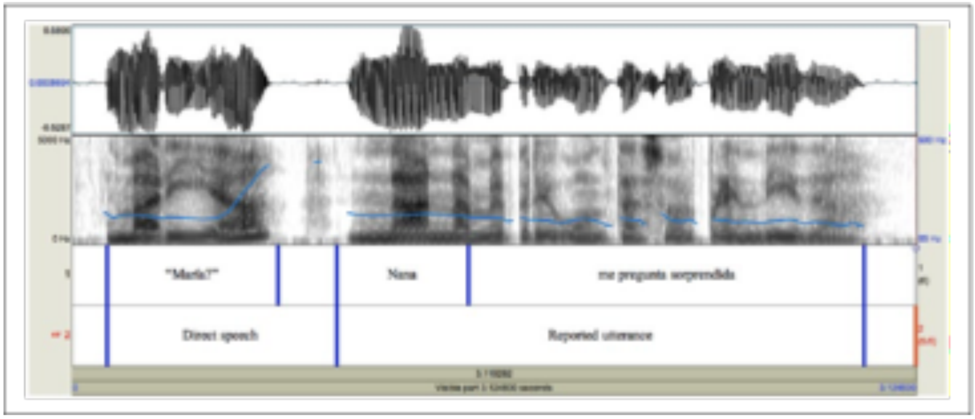


Figure 13. Reporting initial: *"Maria?" Nana pregunta sorprendida.* "Maria?" Nana asks with surprise.

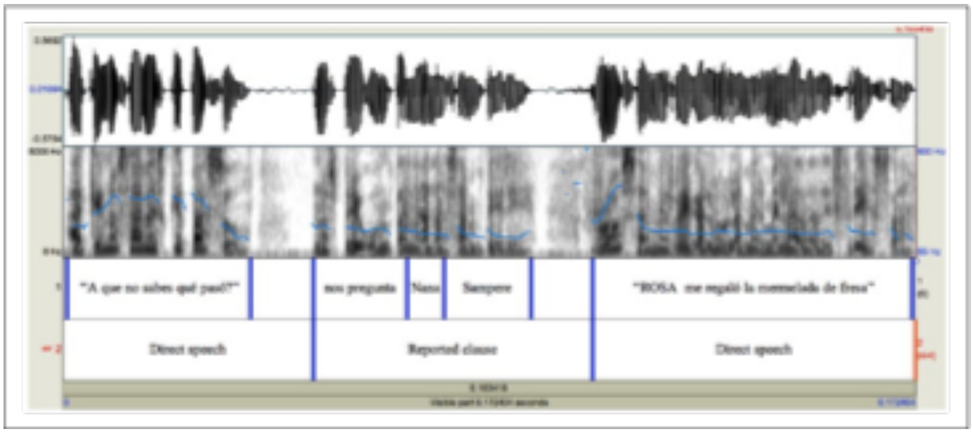


Figure 14. Reporting medial: *"A qué no sabes qué pasó?" nos pregunta Nana Sampere. "Rosa me regaló la mermelada de fresa."*
"Do you know what happened?" Asks Nana Sampere to us. "Rose gave us strawberry marmalade."