RUNNING HEAD: Lexical stress and phrasal prosody in segmentation

Transition Probabilities and Different Levels of Prominence in Segmentation

Mikhail Ordin[a,b] & Marina Nespor[b]

[a]Bielefeld University & [b]International School for Advanced Studies of Trieste

Abstract

A large body of empirical research demonstrates that people exploit a wide variety of cues for the segmentation of continuous speech in artificial languages, including rhythmic properties, phrase boundary cues, and statistical regularities. However, less is known regarding how the different cues interact. In this study we addressed the question of the relative importance of lexical stress, phrasal prominence, and transitional probabilities (TP) between adjacent syllables for the segmentation of an artificial language. We explored how duration increase, pitch rise, and the combination of duration and pitch on the antepenultimate, the penultimate, and the final syllable of a three-syllabic word affect segmentation by native speakers of Italian. Our results indicate that, if the most frequent location of stress in the participants' native language and a lengthened syllable in the artificial language do not coincide, segmentation is disrupted. If there is no conflict between the location of stress in the native language of the participant and the lengthened syllable in the artificial language, segmentation is neither impeded nor facilitated. Pitch marked the edges of the TP-defined words in a continuous speech stream. When TPs and pitch cues are in conflict, segmentation fails; if pitch rise coincides with the edges of TP words, segmentation succeeds, but is not facilitated. Phrasal prominence comprising both pitch and duration facilitates segmentation when aligned with the word edges. Our findings show that language-specific peculiarities of how nuclear pitch accents are realized in the native language of the listener might interact with statistical cues in the segmentation of an unfamiliar language.

Keywords: *speech segmentation; lexical stress; phrasal prosody; transitional probabilities; F0; pitch; duration*

Author Note

We are thankful to Alan Langus and to two anonymous reviewers for their valuable comments and advice. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (*FP7/2007-2013)* / ERC *grant agreement* n° 269502 (*PASCAL*). Correspondence concerning this article should be addressed to Mikhail Ordin, Bielefeld University, Fakultät für Linguistik und Literaturwissenschaft, Universitätstrasse 25, Bielefeld 33615, Germany. E-mail: mikhail.ordin@gmail.com

Introduction

One of the central problems in language acquisition research is the identification of the mechanisms that enable learners to extract discrete units from continuous speech. Research focusing on the role of statistical learning has provided evidence that infants, children, and adults may track simple statistical regularities for the purposes of speech segmentation in an unknown artificial language (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). More specifically, researchers have shown that troughs in transitional probabilities (TPs)[1] between syllables or diphones are exploited to segment speech and extract words. TPs are generally higher within words than between words. That is, syllables within words have higher TPs than syllables that straddle word boundaries. Tracking these statistical regularities is one of the mechanisms humans may exploit to detect word boundaries in the language they are acquiring (Hayes & Clark, 1970). Lexical stress at word onset or offset might also play a role in the segmentation of an unknown language, especially if its stress pattern matches that of the participants' native language (Tyler & Cutler, 2009). In particular, language-specific biases for stress placement might provide reliable cues to the word onset (in languages like English, Hungarian, Dutch, Finnish, etc.) or offset (in languages like Turkish, French, etc.), when the stress location coincides with the word's edge. However, it is less clear what role lexical stress plays in the segmentation of languages where stress is not aligned with one of the edges of words (e.g., in Italian or Spanish) and how speakers of these languages exploit lexical stress for segmentation of an unknown language. Duration, pitch and intensity as key correlates of lexical stress all contribute to the differentiation of stressed and unstressed syllables, but not equally, and the weight of these acoustic correlates in

stress perception also varies cross-linguistically. Finally, the relative importance of stress cues and TPs in segmentation for speakers of such languages is also an open issue.

In this article, we report on three different experiments which we designed with the goal to understand the way in which different cues to word segmentation are exploited in language acquisition. The participants were adult native speakers of Italian, a language with complex stress patterns where stress is not aligned with one of the edges of words. We pitted TPs against duration (i.e., the most important correlate of lexical stress in Italian) in the first experiment, against pitch (i.e., the most salient prosodic cue to prominence cross-linguistically) in the second experiment, and against duration and pitch combined (i.e., accent) in the third experiment. We were interested in the mechanisms that allow people to segment continuous speech regardless of language-specific phonetic realizations of linguistic events.

Background to the Study

While it is a well established finding that TPs between syllables and phonemes are exploited to segment speech and extract words (Hayes & Clark, 1970; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), they are not the only statistical cues available to language learners for segmentation. Among other statistical cues that are successfully exploited are phonotactic regularities (Finn & Hudson Kam, 2008; Onishi, Chambers, & Fisher, 2002), non-adjacent TPs (Peña, Bonatti, Nespor, & Mehler, 2002), the relative frequency of functors and lexical items (Gervain, Nespor, Mazuka, Horie, & Mehler, 2008), as well as TPs only between the consonants (Bonatti, Peña, Nespor, & Mehler, 2005), in addition to distributional properties of phonemes and allophones (Brent & Cartwright, 1996; Batchelder, 2002; Maye, Werker, & Gerken, 2002). However, TPs

between adjacent syllables, in the absence of different statistical or other types of cues, are sufficient for the segmentation of artificial streams (Aslin, Saffran, & Newport, 1998).

The use of prosodic information in speech segmentation has also been well documented (Selkirk, 1984; see also Nespor & Vogel, 1986, 2007). The contours of the fundamental frequency (F0), which represent the acoustic correlates of pitch, mark the boundaries of Intonational Phrases (IPs). IP boundaries coincide with word boundaries and are therefore also relevant for the segmentation of speech into words. Langus, Marchetto, Bion, and Nespor (2012) showed that final lengthening in phrases and pitch declination in sentences are also successfully exploited in speech segmentation. In addition, Shukla, Nespor, and Mehler (2007) showed that adults are able to exploit the prosodic markers of IP edges for word segmentation. They found that trisyllabic statistically defined words of an artificial language were segmented better when aligned with IP boundaries than when they occurred in the middle of IP contours. They also found that if a statistical word straddles two F0 contours, it is not recognized. Shukla et al. (2007) have also shown that participants can exploit F0 contours extracted from a foreign language and imposed on an artificial speech stream for the purposes of segmentation. This indicates that prosody at the higher linguistic levels, for example, the IP level, offers universal cues. This conclusion is to some extent confirmed in Toro, Sebastián-Gallés, and Mattys (2009): After a 7-minute exposure, both Spanish and English listeners were able to segment an artificial stream into trisyllabic words in the TP-only condition, as well as in pitch-initial and pitch-final conditions, but segmentation failed in the pitch-middle condition[2].

The use of word level prosody (e.g., lexical stress) for word segmentation of an unknown language by adults (Cutler, Dahan, & van Donselaar, 1997; Cutler & Norris, 1998; Cutler, Norris, Mehler, & Segui, 1992), or first language by infants (Johnson & Jusczyk, 2001) is also

attested. However, all studies on the role of lexical stress for segmentation concern languages in which lexical stress is either exclusively or predominantly at one of the word's edges, as in French and English, respectively. In addition, only a few studies compared directly the relative importance of word stress and TPs in segmentation, and the results of these studies are not coherent. McQueen (1998) and Cairns, Shillcock, Chater, and Levy (1997) concluded that stress is a cue of minor importance in Dutch and English when alternative cues are available. Mattys, White and Melhorn (2005) showed that in English low-probability diphones are interpreted as word boundaries regardless of stress pattern, and higher probability diphones suppress the perception of word onsets signaled by stress cues. However, Mattys et al. (2005) and McQueen (1998) found a substantial effect of stress cues in acoustically degraded signals, for example, in the presence of environmental noise. This conclusion is in line with Smith, Cutler, Butterfield, and Nimmo-Smith (1989) and Liss, Spitzer, Caviness, Adler, and Edwards (1998), who found that in English, stress outweighs TPs in acoustically impoverished conditions caused either by background noise or pathological speech due to motor speech disorders. Yet, in some studies, stress has been claimed to be as important or even more dominant than TPs in the segmentation of acoustically clear speech signals, for example, of nonsense words by native speakers of English (Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997) and Finnish (Vroomen, Tuomainen, & de Gelder, 1998).

During language acquisition, developmental changes in the relative importance of lexical stress and statistical cues for speech segmentation have been detected. Thiessen and Saffran (2003) found that 6-month-olds from an English-speaking environment relied more on statistical cues than on lexical stress in segmenting new words, while 9-month-olds shifted their attention to prosody. These authors suggested that infants first acquire prosodic knowledge of their

language of exposure with the help of statistical learning mechanisms. Having acquired the prosodic regularities of their language of exposure, they develop a new segmentation strategy based on stress cues. By the end of the first year of life, infants change their strategy again (Thiessen & Saffran, 2007): They then learn to integrate multiple cues for word segmentation, and once TPs and stress cues are in conflict, English infants favor statistical cues (Johnson & Jusczyk, 2001).

From previous research we know that lexical stress at word onset or offset might play a role in the segmentation of an unknown language, especially if its stress pattern matches that of the participants' native language (Tyler & Cutler, 2009). The basic principle is that language-specific biases for stress placement might provide reliable cues to the word onset (in languages like English, Hungarian, Dutch, Finnish, etc.) or offset (in languages like Turkish, French, etc.), when the stress location coincides with the word's edge. The role of lexical stress in the segmentation of languages where stress is not aligned with one of the edges of words (e.g., in Italian or Spanish) and how speakers of these languages exploit lexical stress for segmentation of an unknown language is not clear. The relative importance of stress cues and TPs in segmentation for speakers of such languages is also an open issue.

Lexical stress is a complex interplay of several characteristics that make one syllable more salient perceptually than its neighboring syllables. The prosodic correlates of stress are: (1) duration, as stressed vowels are longer than unstressed ones (Gussenhoven, 2003, pp. 12-19); (2) overall intensity, as stressed vowels have higher intensity level (Cutler, 2005); and (3) F0, as stressed vowels have a higher fundamental frequency (Cutler, 2005).

Duration, pitch, and intensity all contribute to the differentiation of stressed and unstressed syllables, but not equally, and the differentiating power of each acoustic correlate

varies across languages. Thus, the weight of these acoustic correlates in stress perception also varies cross-linguistically. In what follows we briefly discuss the language-specific complex interplay of pitch, duration, and intensity in the manifestation of prominence.

In Italian, duration is the major correlate of lexical stress both in production and in perception (Bertinetto, 1980), while pitch plays a major role in prominence at the IP level. Overall, stressed vowels in open syllables are longer than unstressed vowels or stressed vowels in closed syllables. However, the increase in duration is not equal in all syllables. Stressed open penultimate syllables are much longer than open antepenultimate, and the stressed open final syllables are the shortest (D'Imperio & Rosenthall, 1999). D'Imperio and Rosenthall presented a phonological analysis of lexical stress in Italian and argued for two different lengthening phenomena of stressed vowels: phonological and phonetic. Increase in duration of stressed vowels is accounted for by phonetic lengthening, while phonological lengthening accounts for vowel duration exclusively in stressed open penultimate syllables. Over 70% of trisyllabic words in the Italian lexicon bear stress on the penultimate syllable, less than 30% bear stress on antepenultimate syllable, while final stress is much less frequent (Krämer, 2009). Besides being unmarked, penultimate stress is obligatory if the penultimate syllable is heavy. In other positions heavy syllables do not attract stress.

A bulk of research has shown that duration is a cross-linguistic and most universal correlate of lexical stress in unaccented positions (lexical stress without phrasal prominence). Stressed vowels are significantly longer than unstressed vowels in Dutch (Sluijter & van Heuven, 1996), Welsh (Williams, 1985), Italian (Bertinetto, 1980), Spanish and Catalan (Ortega-Llebaria & Prieto, 2011), Thai (Potisuk, Gandour, & Harper, 1996), Romanian (Manolescu, Olson, & Ortega-Llebaria, 2009), Estonian and Russian (Eek, 1987), German (Dogil & Williams, 1999;

Kohler, 2012), English (Crystal & House, 1987), and Greek (Arvaniti, 2000; Kastrikani, 2003), among other languages.

Pitch can only be a prosodic correlate of stress in accented syllables. Even in those languages in which pitch has been claimed to be the strongest correlate of stress (e.g., English), it can be a reliable correlate of stress in some contexts and almost irrelevant in others (Ladd, 2008, pp. 50-52). The misconception that stress is realized by F0 fluctuations initially originated from Fry's (1958) perception experiments. F0 changes do provide powerful cues to the location of stress because the presence of F0 movement is aligned with stressed syllables. However, the position of the pitch accent and the shape of F0 contour (phonological tone) is part of the intonational grammar of language. As pitch accent is aligned with a stressed syllable, it can cue lexical stress, but not every stressed syllable is pitch-accented.

The distinction between accent and lexical stress was first clearly formulated by Bolinger (1958), who defined a stressed syllable as a syllable that can potentially bear a pitch accent. This distinction between stress and accent was further developed within Autosegmental Theory (e.g., Goldsmith, 1976) and Metrical Theory (e.g. Liberman & Prince, 1977): Representations for phonological tones were created and lexical stress was represented separately from F0 patterns (see Ladd, 2008, for an overview).

Taking into consideration the language-specific complex interplay of pitch, duration, and intensity in the manifestation of prominence, we decided to tackle duration and pitch separately in order to investigate their separate contribution in the segmentation of continuous speech. We did not include intensity as a stress correlate in our study for several reasons. First, although intensity is a well-determined acoustic correlate of stress (e.g., Cutler, 2005), its role as a perceptual correlate of stress is less clear. Sluijter and van Heuven (1996) showed that spectral

tilt (i.e., a downward slope towards the higher end of the spectrum) is a much more reliable perceptual correlate of stress than overall intensity. Second, the overall intensity level is highly correlated with vowel duration: Averaged intensity level on a longer vowel is by default higher than that on a shorter vowel, all other factors being equal. Third, intensity alone can never mark prominence, and it always works in a bundle with other prosodic features, while duration and pitch can mark prominence to the listener on their own (Turk & Sawusch, 1996). Finally, differences in overall intensity between stressed and unstressed vowels are very small, in the vicinity of 3-4 dB (e.g., see Ortega-Llebaria & Prieto, 2011; Ortega-Llebaria, Vanrell, & Prieto, 2010), while the minimum perceptual threshold for the differences in intensity varies between 1-2 dB. Thus, the increase in intensity caused by stress is perceptually very small. Much larger differences in overall intensity, up to 5-7 dB, are caused by adjacent consonants (House & Fairbanks, 1953), as well as by other factors such as syllable complexity (Parker, 2008). Even differences in intrinsic intensity between different vowels can be larger (up to 5 dB) than those between the same vowel in stressed and un-stressed positions (Ordin, 2011). Consequently, the overall intensity level is prone to modifications caused by multiple factors to a much greater degree than by stress. In addition, infants are much less sensitive to differences in intensity (Saffran, Werker, & Werner, 2006, for an overview) and consequently less likely to exploit intensity for stress perception. Since the ultimate goal of our study is to understand the way in which different cues to word segmentation are exploited in language acquisition, we decided not to include this parameter in our study.

People exposed to an unfamiliar language employ the segmentation strategies they have developed in their native language (Cutler, Mehler, Norris, & Segui, 1986; Finn & Hudson Kam, 2008; Toro, Pons, Bion, & Sebastián-Gallés, 2011; Vroomen et al., 1998). If a novel language

and the native language of the listener have the same cues for segmentation — for example, have the same type of vowel harmony (Vroomen et al., 1998), vocalic structure (Toro et al., 2011), or stress location (Tyler & Cutler, 2009) — segmentation is facilitated. However, the evidence for the facilitation effect of lexical stress was obtained in these studies with native speakers of languages in which lexical stress coincides with the word edges (Dutch, English, Finnish, and French). It is much less known if more complex stress patterns of the first language (e.g., Italian) are exploited in word segmentation in a novel language. We wanted to investigate whether the location of lexical stress is an aid to segmentation for participants of a native language in which the unmarked location of lexical stress is not aligned with the word edges. We also investigated whether lexical stress at the word edge — allowed but marked in the participants' native language — can help participants segment an artificial language. In addition, we set out to determine the different roles for segmentation of lexical stress and phrasal prominence manifested by pitch accent, as well as to evaluate the relative importance of prosodic cues and TPs in segmentation.

To address these questions, we decided to pit TPs of adjacent syllables against prosodic cues in an artificial language and test the specific cues native Italian speakers will exploit in segmentation. We pitted TPs against duration (i.e., the most important correlate of lexical stress in Italian) in the first experiment, against pitch (i.e., the most salient prosodic cue to prominence cross-linguistically) in the second experiment, and against duration and pitch combined (i.e., accent) in the third experiment.

We did not introduce Italian-specific phonetic realizations in our stimuli. All languages use F0 and duration to mark prominence for lexical stress and phrasal accent. Although the phonetic realizations of these phenomena are language-specific, adults can learn to segment

speech in a foreign language, even if the phonetic realization of prominence in the target language differs from that in the native language of the learner. We were interested in the mechanisms that allow people to segment continuous speech regardless of language-specific phonetic realizations of linguistic events. When people learn to segment speech in an unknown language, they apply the phonology and phonetic peculiarities of their native language to the incoming speech stream. However, a new unknown language is not necessarily similar to the native language of the learner, and this might present segmentation difficulties. We avoided implementing Italian-specific phonetic patterns into the artificial language of our experiments in order to investigate whether and how Italians will use their native phonology and phonetic regularities when segmenting an unknown language that might have different patterns of stress and phrasal accent.

Experiment 1

Participants

The participants for Experiment 1 were 24 (16 females and 8 males) monolingual Italian speakers who received monetary contribution for participation in the experiment. All participants came from monolingual parents, learned Italian from birth, and none were regularly exposed to any foreign language on a regular basis. Although English is a compulsory school subject in Italy, care was taken to select the participants with as little experience with English or any other foreign language as possible. Participants were first- and second-year students from the University of Trieste at the time of the experiment (approximate age: 19-20 years). None reported nor showed any speech and hearing disorders.

Stimuli

For the three experiments, we designed an artificial language and created a series of audio files representing different experimental conditions. Readers will find the complete audio files for all materials used in the three experiments in Appendix S1 of the Supporting Information online.

We constructed an artificial language using five vowels and 11 consonants: /k/, /m/, /p/, /b/, /l/, /t/, /g/, /v/, /n/, /f/, /d/, /i/, /e/, /a/, /o/, /u/. We selected these phonemes because they occur more frequently in the world languages (Ladefoged & Maddieson, 1995; Maddieson, 1984). Concatenations of these phonemes produced twelve three-syllabic words (*komipa, bolatu, kupige, vunelu, bamofe, defida, bukite, vifole, dubipo, vaputa, donume, ginefa*) and a set of 36 unique syllables. Six words were used to create stream 1 (the first artificial language) and 6 words were attributed to stream 2 (the second artificial language). Thus, the TPs between syllables within words were 1.0 throughout, and the TPs between syllables at word boundaries were .15 in both streams. We made sure that neither the words themselves nor any concatenation of these words within a stream produced real Italian polysyllabic words. We did so by presenting them auditorily to Italian speakers and asking them to listen for any part of the stream that sounded like an Italian word. During this pretest Italian speakers did not hear any Italian word in the streams. The words were concatenated so as to avoid adjacent word repetitions.

The acoustic streams were generated using the MBROLA speech synthesizer (Dutoit, Pagel, Pierret, Bataille, & van der Vrecken, 1996). We adopted the paradigm used by Peña et al. (2002) and by Tyler and Cutler (2009) and assigned an equal base length to each vowel and consonant. The duration of each phoneme was set to 100 ms, which produced 200-ms syllables. Each word was repeated 166 times in a stream. Stream duration was 597.6 sec (9.96 min). We

used the *it4* MBROLA diphone database (female voice). Thus, we set the fundamental frequency (F0) to 200 Hz throughout, since this is the average female F0. The acoustic signal was generated at 16 kHz sampling frequency, and a 5-second fade-in and fade-out was applied to the stream edges so that participants did not have access to the word-boundaries at the beginning and at the end of the streams. As each phoneme was coarticulated with the following phoneme regardless of its position in the stream, coarticulatory cues were not available to the listeners for the purposes of segmentation. Thus, the only cues for segmentation were TPs between syllables. Further in the text, across all three experiments, we will refer to this condition as TP-only.

Each stream was then modified to implement prosody at the word level. Either the first, the second, or the third vowel in each word was lengthened by 80 ms (consonantal durations were left intact). Thus, the stressed syllable in each word was 280 ms, compared to the 200 ms duration of unstressed syllables. This lengthening increased the total duration of the stream to 677.28 sec (11.29 min). As a result, each stream was prepared in four different conditions: TP-only, initial-lengthening, middle-lengthening, and final-lengthening (see Appendix S1 in the Supporting Information online).

Duration values for the stimuli were chosen irrespectively of Italian-specific durations because we were interested in the general mechanisms exploited by Italians to process durational cues to segment an unknown language. We were rather more concerned about obtaining results that would be comparable to other studies on segmentation. In one of the initial studies using artificial languages, Saffran, Newport and Aslin (1996) set the syllable duration to 277 ms, and increased the duration by 100 ms on the cued syllables. Peña et al. (2002), Toro et al. (2009), and Tyler and Cutler (2009) set the syllabic duration to 232 ms with 60 ms increase on cued syllables. Vroomen et al. (1998) used 220 ms as syllable duration. Shukla et al. (2007) used a

syllable duration of 200-280 ms. Kim, Broersma and Cho (2012) varied the syllable duration between 252 and 446 ms. Considering this variability, we decided that 200 ms baseline and 280 ms for increased duration values would provide a good comparison across studies.

Procedure

Each participant came for the experiment twice, with an interval of 1 to 2 weeks between the two sessions. She or he was exposed to stream 1 and stream 2 in two different conditions in the first session, and to stream 1 and stream 2 in the other two conditions in the second session. The combination of stream × condition × order of presentation was randomized (24 unique combinations), and each participant was assigned to one unique combination.

In the familiarization phase, participants were instructed to listen carefully to an imaginary language that contains its own words that do not have any meaning in any attested language. Participants were aware that there was going to be a test phase after the familiarization phase. This awareness supposedly helped them to keep focused on the task (listening to the artificial language carefully). Immediately after exposure to one stream, in a dual forced-choice task, we asked participants to listen to pairs of imaginary words and decide which of the two they thought had been presented in the familiarization language. Three partwords were formed from the third syllable of one statistically-defined word and the first and second syllables of the following word, and three partwords were formed from second and third syllables of one word and the first syllable of the next word. Pitting all possible words against all possible partwords gave 36 pairs, each containing one word and one partword. The order of words and partwords in the pairs was counterbalanced. The order of the pairs was randomized for each participant. The items in the pair were separated by a 500-ms pause. Participants were instructed to listen to the pair and to click either button 1 or button 2, depending on whether they considered the first or

the second item in the pair a word in the language they had just listened to. Participants were instructed to give the first answer that comes to mind and not to spend much time to decide the correct answer. The stream and the test items were presented via headphones in a sound-attuned booth, and each participant was instructed and tested individually. After the test was over, participants had a 5-minute pause, and the second stream was presented, followed by a new test. During the second session, one or two weeks later, the participants were exposed to the streams in the other two conditions. In all cases, familiarization was followed by a dual forced-choice test of words versus partwords.

Franco, Cleermans, and Destrebecqz (2011) showed that people are able to learn two artificial languages sequentially and to easily differentiate between them, while Gebhart, Aslin and Newport (2009) found interference between statistically-coherent languages when they are presented sequentially. The latter authors showed that successful extraction of the statistical structure of the first language reduces the performance in processing the subsequent artificial structure. However, the participants in experiments by Gebhart et al. (2009) were exposed to two languages sequentially within one familiarization phase, and they had to perform the test that included items from both artificial languages. The same approach was used by Weiss, Gerfen, and Mitchell (2009) who showed that participants can track statistics from several languages provided that they have sufficient indexical cues (e.g., different voices) for each language. Unlike in these studies, in our experiments the familiarization phase included only one language at a time, and also the test items were taken from one language. In other words, we did not mix the items from different languages within the same testing session. Only after the test was completed did the familiarization with the second set of words start. In addition, Gebhart et al. (2009) did not find an interference effect if the exposure to the second language was lengthy

enough, or the presence of two different structures was marked explicitly (e.g., in instructions), or when the two subsequent languages were separated by a pause. All three conditions are fulfilled in our experiments. We thus assume that that one stream did not influence the other during either familiarization or testing. To confirm the assumption that there was no bias for either presentation order or language (i.e., the specific familiarization stream), an additional series of statistical tests was performed.

Results

We performed data screening to monitor for outliers and to ensure that the requirements for normality and linearity were not violated. Parametric tests were subsequently run, and the significance of the individual *t*-tests was evaluated after applying the Bonferroni correction, so alpha value for the whole set of tests was set at $p < .005$. Finally, the effect size was measured by calculating the correlation coefficient from *t*-statistics and *df* to evaluate whether the difference between the chance level and the mean number of correct responses in a certain condition, or between mean numbers of correct responses in two different conditions is large enough to be practically meaningful. We used Cohen's (1988: 284-287) suggestions as guidelines to interpret the effect size values. Significant test and big ($r > .5$, experimental manipulation accounts for at least 25% of variance in the responses) or medium effect size ($r > .3$, experimental manipulation accounts for at least 9% of variance) represents an important result that confirms the hypothesis and has practical value. Non significant test and big effect size represent lack of power of the test. Non significant test and small effect ($r > .1$) represents the lack of difference between the chance level and the mean or between two means.

Preliminary tests revealed no apparent bias for either stream (one of the streams was not better segmented than the other) or session (people did not perform better during one of the two

sessions). Readers will find the results of these preliminary tests for Experiment 1 (as well as the other two experiments) in Appendix S2 in the Supporting Information online.

One-sample *t*-tests were performed to compare the number of correct answers in each condition with the chance level (50% or 18 correct answers). We assumed that if participants successfully segmented the words in the continuous acoustic stream, the number of correct responses should be significantly above chance. The tests showed that in the TP-only condition, $t(23) = 3.93$, $p = .001$, $r = .63$ and in the middle-lengthening condition, $t(23) = 2.99$, $p = .007$, $r = .53$, the number of correct responses was significantly and substantially above chance, as indicated by the large effect size, while in initial-lengthening, $t(23) = .82$, $p = .422$, $r = .17$, and final-lengthening conditions, $t(23) = .1$, $p = .333$, $r = .02$, the number of correct answers was at chance. The mean number of correct answers for each condition ($n = 24$) and the bars showing ±2 standard errors are presented in Figure 1.

Figure 1. Mean number (±2 standard errors) of correct answers in the test phase for each condition in Experiment 1 (only durational cues were used for the lexical stress)

We did not find any facilitation effect of durational cues on segmentation. Segmentation performance was the same in the TP-only and the middle-lengthening conditions, $t(23) = .72$, $p = .481$, $r = .15$. Participants in our experiment apparently paid attention to the phonological lengthening typical of their native language, which in Italian occurs exclusively in open stressed penultimate syllables, and not to phonetic lengthening. There was no difference in performance between the final-lengthening and initial-lengthening conditions, $t = -.09$, $p = .928$, $r = .02$, both being at chance. If participants had paid attention to phonetic lengthening, performance would have been better in the antepenultimate condition than in the final condition, because in Italian vowels in stressed open antepenultimate syllables are longer than in final syllables, all other factors being equal.

The partwords that we used in the test fall into two categories: (a) those which consist of the third syllable of one word and the first and second syllables of the following word, or type A, and (b) those which consist of the second and third syllables of one word and first syllable of the following word, or type B. When contrasted with words, partwords of type A and of type B bear different prosodic cues for word identification, as shown in Table 1. Therefore, we performed the analysis for the two different types of partwords separately. The results are presented in Figure 2.

TABLE 1 Contrasting words and partwords of two different types. Circles stand for the syllables. Prosodically marked syllables are represented by larger circles. The brackets represent the edges of the words and partwords.

	Initial prominence	Middle prominence	Final prominence
WORD	... [○ ∘ ∘][○ ∘ ∘] ...	... [∘ ○ ∘][∘ ○ ∘] ...	... [∘ ∘ ○][∘ ∘ ○] ...
PARTWORD A	... ○ ∘][∘ ○ ∘][∘ ...	... ∘ ○][∘ ∘ ○][∘ ...	... ∘ ∘][○ ∘ ∘][○ ...
PARTWORD B	... ○][∘ ∘ ○][∘ ∘...	...∘][○ ∘ ∘][○ ∘ ...	...∘][∘ ○ ∘][∘ ○ ...

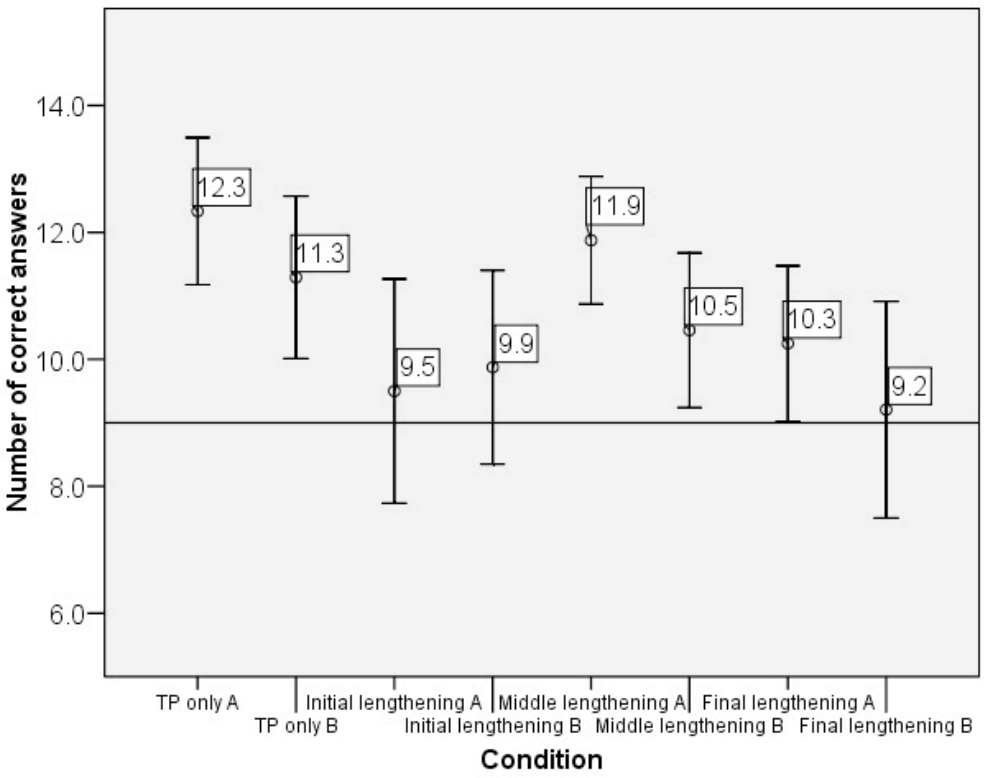

Figure 2. Mean number (±2 standard errors) of correct answers in the test phase for different types of partwords (A and B partwords) in each condition in Experiment 1 (durational cues for lexical stress)

We performed paired *t*-tests to compare the number of correct answers for partwords A and partwords B in each condition. We did not find any significant differences in number of correct responses for partwords A and partwords B. The difference in performance in the middle-lengthening condition, $t(23) = 2.37$, $p = .027$ was no longer significant after applying the Bonferroni correction. However, the substantial effect size $r = .44$ indicates that the correction is over-stringent here and most probably leads to rejecting the factual and important difference in performance as nonsignificant. Therefore, we carried out one-sample *t*-tests comparing the number of correct answers with chance (50%) in each condition for partwords A and partwords B. The number of correct responses was significantly above chance in the TP-only condition for

both types of partwords, and in the middle-lengthening condition for partwords A (at the set alpha level of $p < .005$). Segmentation fails in the middle-lengthening condition for partwords B, $p = .083$, $r = .35$ (a moderate effect size). This shows that Italian participants disfavored final lengthening. Participants made significantly fewer mistakes when they chose between a word with medial lengthening and a partword with final lengthening, than when they chose between a word with medial lengthening and a partword with initial lengthening.

Discussion

As it has been shown in previous studies on statistical cues to segmentation (e.g., Aslin et al., 1998), participants can reliably segment continuous streams of speech into words on the basis of distributional cues. When prosodic cues and statistical cues collide, segmentation is at chance. In Experiment 1, participants segmented reliably when the lengthened syllable was in penultimate position, which is the unmarked (i.e., most common) stress in their native language. However, participants did not benefit from lengthening. In other words, lengthening did not have a facilitation effect on segmentation.

Alternatively, we could expect final lengthening to enhance segmentation, as it has been reported in a number of previous studies (Saffran, Kim, Broersma, & Cho, 2012; Saffran, Newport, & Aslin, 1996; Tyler & Cutler, 2009). Italian participants in our study, however, did not benefit from lengthening in word-final syllables. Final lengthening is a phenomenon that has been attested in a number of languages. However, final lengthening is primarily a cue to a phrasal prosodic boundary, and Italians in Experiment 1 appeared to extract words independently of their position in larger prosodic constituents. In addition, since Italian exhibits phonological lengthening on the stressed penultimate open syllables, vowel lengthening in this position might possibly outweigh the cross-linguistic cue of final lengthening. We suggest that adult Italian

participants are likely to have transferred their native linguistic competence while processing a phonotactically similar unknown language, and perceived lengthening as the cue to penultimate stress.

Our findings suggest that adults are not able to ignore statistical cues and to segment words on the basis of durational cues only. Partwords A in the initial-lengthening condition and partwords B in final-lengthening condition have penultimate lengthening, and the relevant TP words in the test pairs are prosodically ill-formed. If participants had preferred lengthening cues to statistical cues, the number of correct responses in the initial-lengthening condition for partwords A and in the final-lengthening condition for partwords B would have been significantly below chance, not at chance.

To summarize, the results of Experiment 1 show that (a) lexical stress does not facilitate segmentation if the location of stress in the native language of the listener and the novel language coincide, but it can disrupt segmentation if TPs and lengthening cues for segmentation clash; (b) Italians are sensitive mostly to phonological lengthening; and (c) Italians disfavor lengthening of stressed vowels in word final syllables, while phonetic lengthening of stressed word initial vowels is not disfavored.

Experiment 2

Participants

Twenty-four students (14 females, 10 males, approximate age: 19-20 years) who did not participate in the first experiment were recruited also in Trieste. None reported any speech or hearing disorders; all came from monolingual Italian families and were not exposed to foreign languages on a regular basis. Care was taken to select people with little or no exposure to foreign languages.

Stimuli

The same stimuli were used as in Experiment 1, but this time we used pitch-induced prosodic cues instead of durational cues to mark stress. The streams were generated in four different conditions: TP-only, pitch-initial, pitch-middle and pitch-final (to listen to each condition, see Appendix S1 in the Supporting Information online). All generated streams were equal in durations (9.96 min). Following Thiessen and Saffran (2003) and Tyler and Cutler (2009), we created parabolic F0 contour on the cue-bearing syllable. We increased F0 from 180 Hz to 240 Hz on the prominent syllable of the statistically-defined word.

As we were interested in how Italians will use pitch in an unknown language for segmentation purposes, our pitch boundary cues are realized differently from the native language of the participant. The values for the F0 increase were chosen to insure comparability across studies. Increase from 180 to 240 Hz in our study corresponds to 5 semitones (ST). Kim, Broersma and Cho (2012) used similar increase (4.8 ST) on cued syllables. Tyler and Cutler (2009) and Vroomen et al. (1998) increased pitch by 6 ST. Toro et al. (2009) used a much smaller F0 range corresponding to 1.7 ST.

Procedure

The procedure was equivalent to that used in Experiment 1.

Results

We screened the data for outliers and confirmed normality and linearity were not violated. Parametric tests were then run and significance of the individual *t*-tests was evaluated after applying the Bonferroni correction ($p<.005$), and effect sizes were inspected by calculating *r*.

Preliminary analysis of the data revealed no influence of the stream or the session on the number of correct responses (Appendix S1 in online supplementary information). We performed one-sample *t*-tests to compare the number of correct answers in each condition with the chance level (50% or 18 correct responses). The tests showed that in TP-only, $t(23) = 3.55$, $p = .002$, $r = .59$), pitch-initial, $t(23) = 4.12$, $p < .0005$, $r = .65$, and pitch-final, $t(23) = 6.4$, $p < .0005$, $r = .8$ conditions the number of correct responses was significantly and substantially above chance, while in pitch-middle condition the performance was at chance, $t(23) = 1.64$, $p = .115$, $r = .32$. Thus, segmentation failed in the pitch-medial condition and was successful in the two other conditions, when prominence is manifested by F0 increase. The mean number of correct answers for each condition ($n = 24$) is presented in Figure 3.

We also performed paired *t*-tests to compare the number of correct answers for partwords A and partwords B in each condition (see Table 2 for illustration of difference between partword of type A and type B). As shown in Figure 4, we did not find significant differences in number of correct responses for partwords A and partwords B.

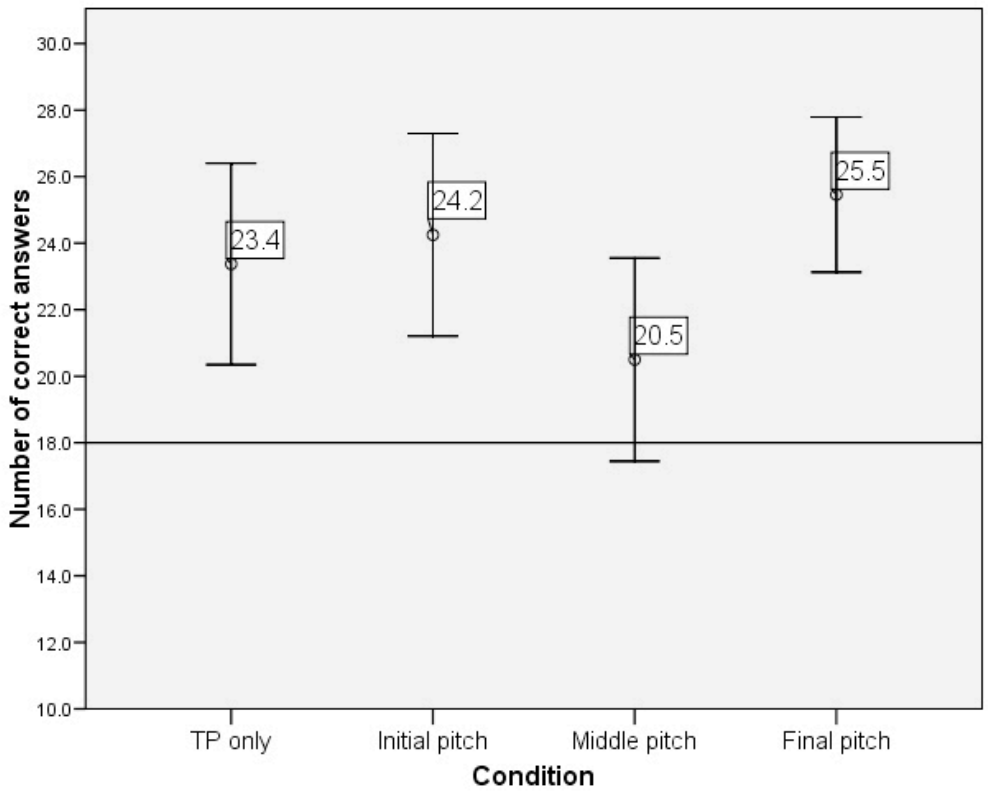

Figure 3. Mean number (±2 standard errors) of correct answers in the test phase for each condition in Experiment 2 (pitch cues for the lexical stress).

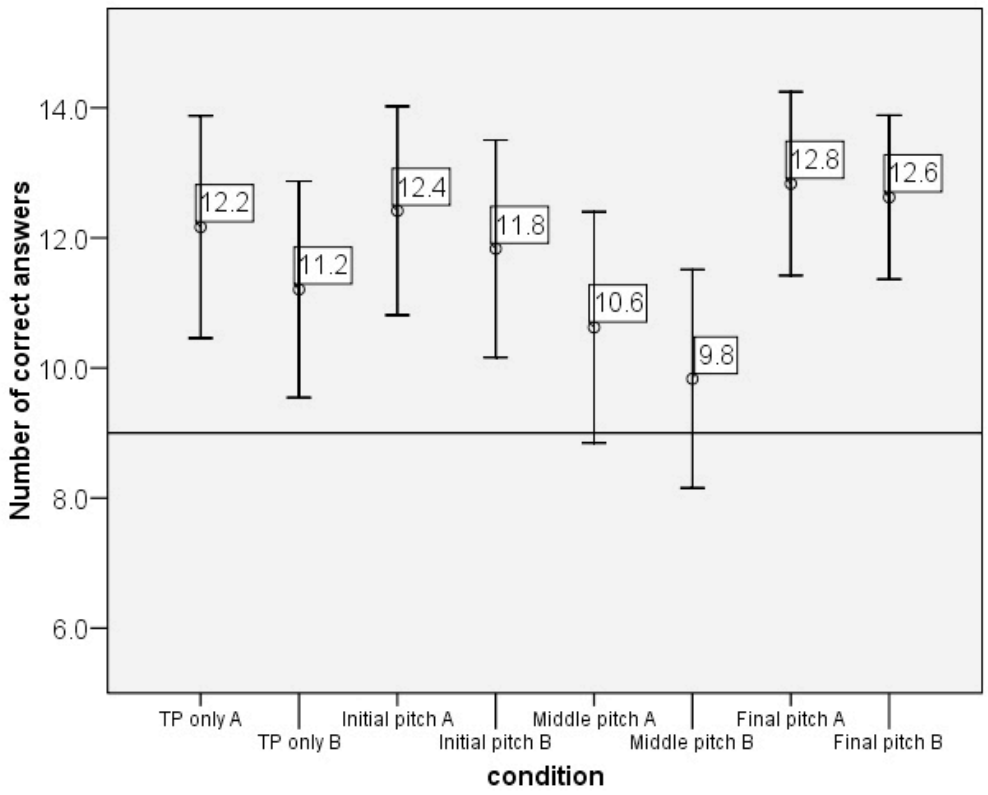

Figure 4. Mean number (±2 standard errors) of correct answers in the test phase for different types of partwords (A and B partwords) in each condition in Experiment 2 (pitch cues for the lexical stress).

Discussion

The results of this experiment are completely different from those of Experiment 1. Italians failed to segment speech when the penultimate syllable was marked by pitch, that is, in the pitch-medial condition. In the other conditions, participants were able to segment continuous speech. Although Figure 3 shows a slight tendency for the number of correct answers to be higher in pitch-final position compared to TP-only and pitch-initial conditions, the only significant differences are between the pitch-middle condition, on the one hand, and the pitch-initial and the pitch-final conditions, on the other hand. Thus, we do not see any significant facilitation effect of pitch for word segmentation based on TPs.

Our results agree with those obtained by Toro et al. (2009). They found that Spanish and English listeners segment artificial streams in both pitch-initial and pitch-final condition, but segmentation fails in pitch-middle condition. A slight yet not-significant tendency for the percent of correct answers in pitch-final position was detected, similar to the one we observed in our experiment. No difference in performance was found between English and Spanish listeners, who behaved in exactly the same way as the Italian participants did in our study. Toro et al. (2009) raised pitch on the cued syllable only by 1.7 ST. We obtained the same results despite a much larger F0 rise on cued syllables (from 180 to 240 Hz, i.e., 5 ST).

Kim, Broersma and Cho (2012) also found no facilitation effect of either initial or final F0 rise for the segmentation of trisyllabic words in an artificial language by Dutch listeners. However, Korean listeners segmented better in pitch-final condition compared to TP-only condition. The authors did not test the effect of medial pitch in their study. Vroomen et al. (1998) tested the effect of vowel harmony and initial pitch on segmentation in an unknown artificial language by Dutch, French, and Finnish participants. They detected that Dutch listeners segmented both harmonious and disharmonious statistically-defined constituents better in pitch-initial condition compared to no-prosody condition, Finnish listeners segmented better only disharmonious speech in pitch-initial condition, and the performance of the French participants did not reveal any effect of pitch on segmentation. Consequently, the authors concluded that initial pitch might facilitate segmentation by people from specific language groups.

The discrepancy between our results and Vroomen et al. (1998) can be explained by the differences in the stimuli. In our study, we used parabolic F0 contour on the cued syllable, and the baseline F0 pitch across the other syllables. In the stimuli used by Vroomen et al. (1998), F0 linearly increased on one syllable from 120 Hz to 170 Hz and then gradually decreased over the

next two syllables back to the baseline of 120 Hz. Their pitch contour thus looks more like an intonation contour, hence participants in their study might have detected the beginning of a higher hierarchical constituent (e.g., the intonational phrase) and associated the F0 rise with the beginning of an intonational phrase, since this is always aligned with the beginning of a word. Shukla et al. (2007) also found a facilitative effect of F0 intonation contour on segmentation.

The comparison of results across studies clearly shows that pitch draws listeners' attention to the edges of words. If the F0 peak is not aligned with the boundary syllable of the statistically-defined word, segmentation fails. We might assume that participants extract the structural units from the continuous acoustic stream based on TPs, and then they align pitch peaks with the edges of these units. Pitch peaks in medial position are not expected and lead to segmentation failure.

English has a strong tendency for initial stress, while in Spanish and Italian stress on the penultimate syllable is most common. Yet regardless of differences in the preferred stress placement, speakers of all these languages segment continuous speech better in pitch-initial or pitch-final positions than in middle position. On the basis of the similar result patterns in experiments with English, Italian, and Spanish participants, we conclude that (a) pitch is not associated with lexical prominence in the absence of durational cues and (b) the detected tendency of pitch to mark edges of linguistic constituents is more general and less language-specific than lexical stress. There is, however, evidence against the universality of this mechanism in Kim, Broersma and Cho (2012), who showed that Korean listeners act in a language-specific manner and use their native linguistic competence in the segmentation of an unknown language. Usually pitch and duration interact and contribute together to the perception of linguistically relevant prominence (e.g., in accented syllables). Therefore, in the third

experiment we explored the combined effect of these two cues on the segmentation of an unknown language by native speakers of Italians.

Experiment 3

Participants

Twenty-four students (17 females, 7 males, approximate age: 19-20 years) who did not participate in the first or second experiments were recruited in Trieste. None reported any speech or hearing disorders; all came from monolingual Italian families and were not exposed to foreign languages on a regular basis. Care was taken to select people with little or no exposure to foreign languages.

Stimuli

The same TP streams of Experiments 1 and 2 were used also in Experiment 3. The prosodically cued versions were created by raising the F0 on a prominent syllable from a lower preceding value, thus creating an inverted parabola shape of F0 contour. The same contour was used in other experiments, for example, Tyler and Cutler (2009) or Thiessen and Saffran (2003). Duration was increased on the prominent syllable. The F0 values ranged between 180 and 240 Hz, and duration was increased from 100 to 180 ms either on the first, on the second, or on the third syllable. By combining durational and pitch cues, we therefore created stressed syllables with pitch accent. The streams were generated in 4 different conditions: TP-only, accent-initial, accent-medial and accent-final (the audio files are available in Appendix S1 in the Supporting Information online).

Procedure

The procedure was identical to that used in Experiments 1 and 2.

Results

As in Experiments 1 and 2, we monitored for outliers and confirmed the data were normal and linear. We then ran individual t-tests after applying the Bonferroni correction at $p<.005$ and calculated r in order to gauge the magnitude of the results. Preliminary analysis of the data revealed no influence of the stream or the session on the number of correct responses (Appendix S2 in Supporting Information online). One-sample t-tests were performed to compare the number of correct answers in each condition with the chance level (50% = 18 correct responses). The tests showed that in TP-only, $t(23) = 4.26$, $p < .0005$, $r = .66$, accent-initial, $t(23) = 2.23$, $p = .036$, $r = .42$, and accent-final $t(23) = 7.36$, $p < .0005$, $r = .84$ conditions, the number of correct responses was significantly and substantially above chance, while in accent-medial condition the performance was at chance, $t(23) = .27$, $p = .791$, $r = .06$. The mean number of correct answers for each condition ($n = 24$) and the bars showing ±2 standard errors are presented in Figure 5.

Comparisons revealed that the number of correct responses was significantly higher in the accent-final condition than in the TP-only condition, $t(23) = -2.72$, $p = .012$, $r = .49$, and the number of correct responses in the accent-medial condition was significantly and substantially lower than in the TP-only condition, $t(23) = 2.54$, $p = .018$, $r = .47$. This shows that segmentation is facilitated by final accent that is manifested by a combination of durational and pitch cues. Medial accent results in segmentation failure. Initial accent neither facilitates nor disrupts segmentation.

Paired t-tests were performed to compare the number of correct answers for partwords A and partwords B in each condition (see Table 2 for illustration of the difference between partword of type A and type B). The only significant difference was found in the accent-medial

condition, $t(23) = -2.78$, $p = .011$. The effect size $r = .50$ shows that the difference is large. Participants made more mistakes when they had to choose between a word with medial accent and a partword with final accent, than when they had to choose between a word with medial accent and a partword with initial accent. This shows that the Italian participants preferred final accent (Figure 6). This confirms the hypothesized facilitation effect of final accent in segmentation of continuous speech stream.

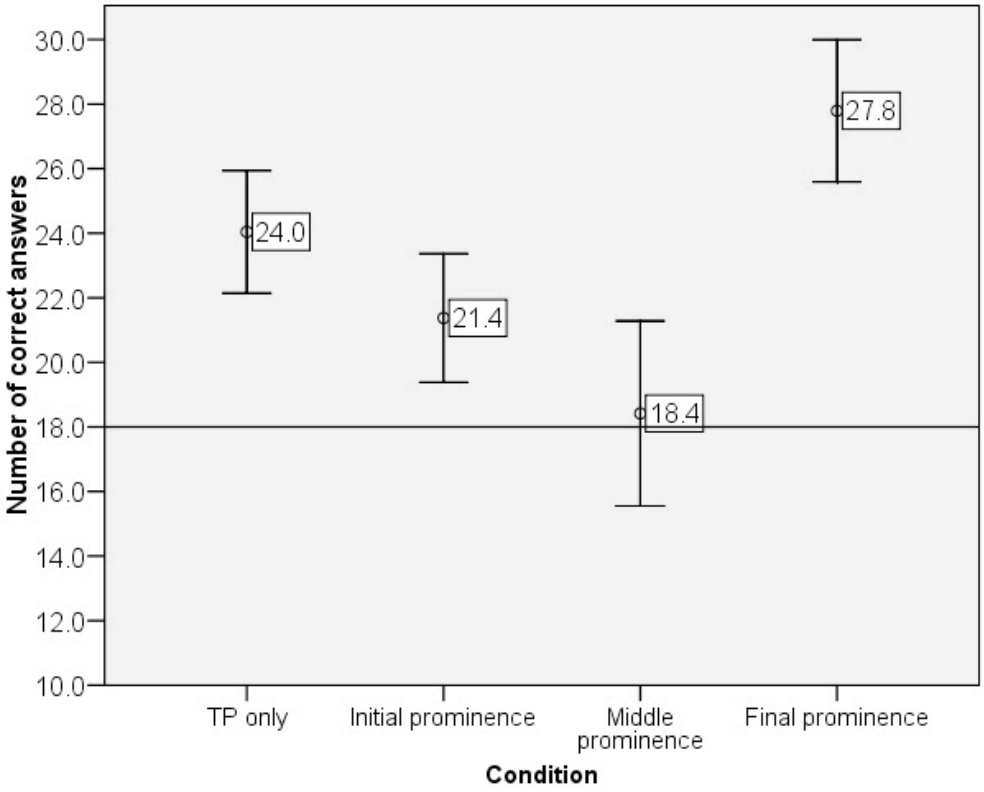

Figure 5. Mean number (±2 standard errors) of correct answers in the test phase for each condition in Experiment 3. Prominence is realized by simultaneous increase of pitch and duration.

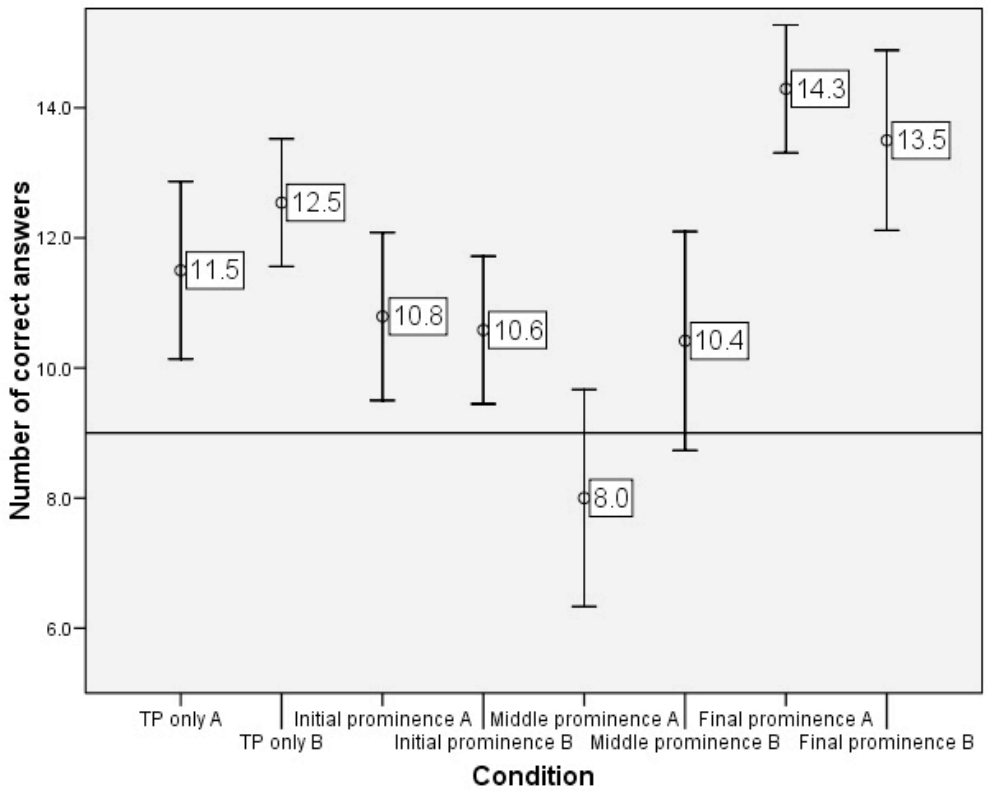

Figure 6. Mean number (±2 standard errors) of correct answers in the test phase for different types of partwords (A and B partwords) in each condition in Experiment 3. Prominence is realized by simultaneous increase of pitch and duration.

Discussion

The results of the third experiment are similar to those of Experiment 2, except for the fact that when prominence is manifested by both pitch and duration, the tendency for a higher number of correct responses in the prominence-final condition becomes significant with respect to those in the TP-only condition. We suggest that this type of prominence is perceived as phrasal prominence, or nuclear accent.[3] And we clearly observed its facilitation effect in final position. The comparison of the number of correct responses in accent-medial condition for

different types of partwords also shows that participants preferred final accent to initial accent: they were more prone to take the partword with final accent for a word.

The fact that in this experiment, where prominence is expressed by a combination of duration and pitch, segmentation fails in prominence-medial condition suggests that pitch overrides duration. In the first experiment, where we created the artificial language with prominence manifested only by duration, Italian participants were driven by the phonological knowledge of their native language. We concluded that they perceived increase in duration as a phonological lengthening which takes place in Italian exclusively in open penultimate syllables bearing lexical stress. However, when duration was combined with pitch in Experiment 3, participants' attention was diverted towards the edges of the words by pitch cues, just like in the second experiment. Why does pitch override duration? It might be suggested that this happens because pitch marks the edges of the structural units of speech cross-linguistically, while the perception of durational cues is driven by the native phonology of the participants. When adults attend to an unfamiliar language, universal phonetically rich cues can override phonological knowledge specific to their native language (e.g., see Kim, Cho, & McQueen, 2012).

The facilitation effect of final accent can be explained by two factors. First, we can assume that the speech stream in the accent-final condition creates the perceptual effect of a list intonation. If so, then each statistically defined word is actually perceived as an IP in a list. Increased F0 marks the right edge and imprints into memory the right edge syllables. As the syllables we used to construct the streams are unique, bringing the edge syllables into memory provides a new and reliable cue for segmentation. In this case, the longer vowels are not perceived as lengthened due to lexical stress, but to phrase-final position. Many studies in speech acoustics have revealed that at utterance boundaries, segments or syllables increase in duration

and this tendency has been suggested to be universal (Nakai, Kunnari, Turk, Suomi, & Ylitalo, 2009; Turk & Shattuck-Hufnagel, 2007; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992), although implemented in a language-specific way. Analyses of the two types of partwords in the third experiment shows that participants preferred final prominence and disfavored initial prominence manifested by a combination of duration and pitch. Therefore, we suggest that participants disfavored initial lengthening in a phrase and went for final lengthening, which is a cross-linguistic boundary signal.

The second factor that accounts for the facilitation effect of final accent is the discrepancy in phonetic realization of pitch accents in Italian and in the synthesized speech in the artificial language used in our experiments. In the synthesized speech, parabolic F0 contours on accented syllables with the peak at 240 Hz, rising from the preceding value of 180 Hz are best labeled as H* or L+H* (see Appendix S3 in the online Supporting Information).

In standard Italian the nuclear pitch accent is L*, H+L* or even L*+H (see D'Imperio, 2002, for an overview), and the choice of the pitch accent is dependent on information structure, or focus, to be more precise (D'Imperio, 2001). The choice of the accent which falls steeply (H+L*) or moderately (L*) from a much higher preceding F0 value is a very common type of accent in declarative sentences in standard Italian, northern dialects, Palermo Italian, Bari Italian, and Neapolitan Italian (D'Imperio, 2002; Grice, 1995; Prieto, D'Imperio, & Fivela, 2005; Rossi, 1997).

However, in the extant literature we did not find any discussion of whether the realization of the nuclear pitch accent differs depending on the position of the accented syllable in the word. We thus investigated whether pitch accents associated with the antepenultimate, penultimate, or final syllables differ. For this purpose we asked 10 native Italians from the University of Trieste

to read six sentences. The sentences were read three times. Three of these sentences were constructed so that the same syllable "-ta" was stressed in either antepenultimate (*tavola*), penultimate (*gitano*), or final (*novità*) syllables, and the pitch accent is associated with these syllables. All words were trisyllabic and consisted only of open syllables. The analysis showed that: (1) 80% of accents associated with antepenultimate syllables are falling from a higher F0 value (H+L* or L*); (2) 66% of accents associated with penultimate syllables are falling from a higher F0 value (H+L* or L*); and (3) 60% of accents associated with ultimate syllables are rising from a lower F0 value (H* or L+H*). Further details on the acoustic analysis, methodology, and supplementary production data can be found in Appendix S3 in the Supporting Information online.

The H* or L+H* accents implemented in the synthesized speech in our artificial language streams are most frequently associated with final stressed syllables and are rarely associated with antepenultimate and penultimate stressed syllables in Italian. This might be the factor explaining why we detected a substantial facilitation effect of prosody on the right edge, that is, where the final syllables in statistically defined words are simultaneously cued with duration and pitch. Figure 7 illustrates the typical realization of pitch accent associated with either antepenultimate, penultimate, or final syllables in trisyllabic Italian words and the realization of pitch accent in the synthesized speech.

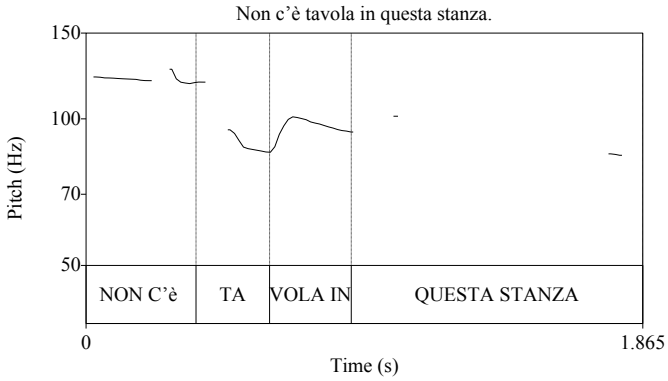

Figure 7a. H+L* accent associated with antepenultimate syllable in the word bearing the nuclear accent

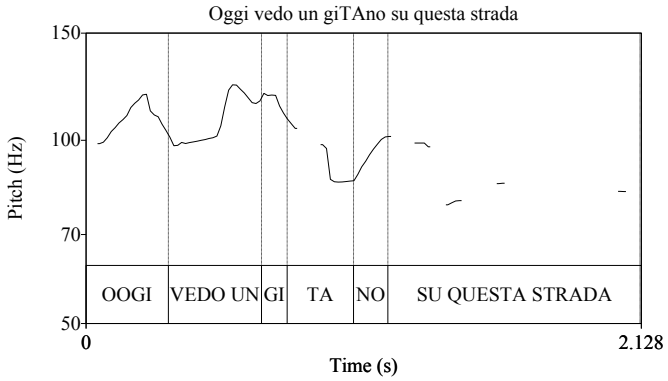

Figure 7b. H+L* accent associated with penultimate syllable in the word bearing the nuclear accent

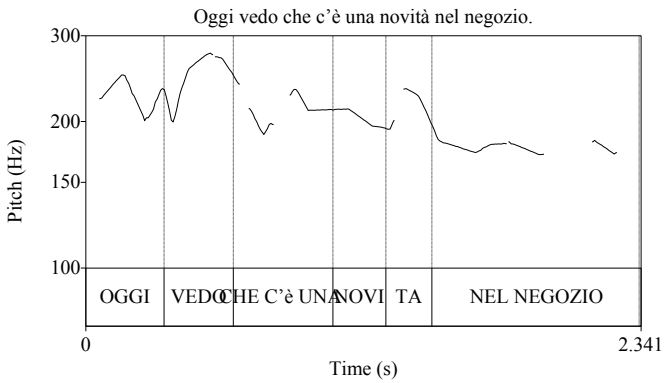

Figure 7c. H* accent associated with ultimate syllable in the word bearing the nuclear accent.

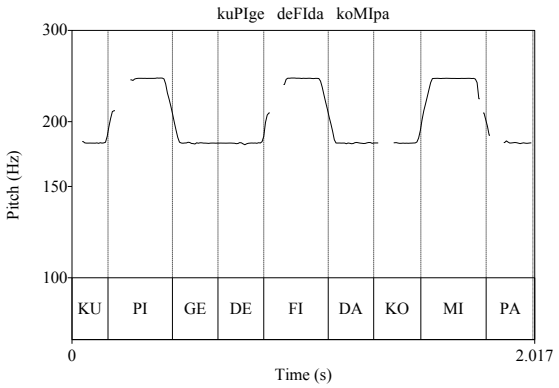

Figure 7d. L+H* accents in the synthesized speech stream.

Figure 7. Typical realization of pitch accent in trisyllabic Italian words and realization of pitch accent in the synthesized speech.

General Discussion

Our results indicate a poor contribution of lexical stress to the segmentation of an unknown language by speakers whose native language does not have default lexical stress

aligned with a word edge. In nonsense words consisting of three open syllables and in the absence of other prosodic cues, duration is perceived by native speakers of Italian as *phonological* lengthening triggered by lexical stress. Phonological lengthening in Italian words takes place only in stressed open penultimate syllables. If this constraint is violated, that is, if in the experimental material lengthening takes place either on the antepenultimate or on the final syllables of the words, thus conflicting with TPs, segmentation fails. However, if the constraint is respected, lengthening cues, which act as correlates of lexical stress, do not facilitate segmentation based on TPs.

Our results also indicate a facilitation effect on segmentation when pitch accent is on the right edge of a word: The combination of F0 increase and vowel lengthening is perceived as pitch accent on the stressed syllable. In Italian, pitch accent is manifested by a F0 local valley on a prominent syllable in antepenultimate and penultimate positions. The final position is the only position in which we have attested prominent syllables with pitch accent manifested by a local F0 peak in more than 50% of the cases. In our synthesized speech, we increased the F0 to mark prominence on the longer syllable, regardless of whether the syllable is antepenultimate, penultimate, or final. The facilitation effect on segmentation detected when pitch accent is on the final syllable is likely due to the fact that the phonetic realization of pitch accent in the experimental language is similar to that of the native language of the participant.

The effect of language-specific phonological knowledge on the processing and on the segmentation of words in an unfamiliar language has been widely attested (e.g., Finn & Hudson Kam, 2008; Toro et al., 2011; Tyler & Cutler, 2009; Vroomen et al., 1998). It is also known that universal prosodic boundary markers facilitate segmentation (Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; Endress & Hauser, 2010; Shukla et al., 2007). However, to the best of

our knowledge, our study is the first to show that the language-specific phonological competence of phrasal prosody influences segmentation. Our study adds to the existing body of knowledge the fact that language-specific peculiarities of how nuclear pitch accents are realized in the native language of the listener might interact with statistical cues in the segmentation of an unfamiliar language.

Comparing the results of the first and the third experiments, we can conclude that for participants whose native language does not have stress at one of the word's edges, lexical stress does not facilitate segmentation in an unknown language; instead, lexical stress is only able to disrupt segmentation if the stress patterns differ from those of the native language of the listener. Phrasal prosody, on the other hand, can facilitate segmentation, if the prosodic characteristics of the presented language conform either to the native language or to universal prosodic phenomena (e.g., final lengthening, as shown, for example, in Langus et al., 2012). This agrees with the body of experimental evidence suggesting that lexical stress is a much less reliable correlate for segmentation than prosodic boundary cues (Endress & Hauser, 2010) or statistical regularities (Mattys et al., 2005).

Previous studies that have shown a significant facilitation role of stress in segmentation were performed with participants whose native language aligns stress with word edges. If the native language of the listener does not align word stress with the word edge, as in Italian, stress in a novel language does not induce a facilitation effect for segmentation. In addition, a facilitation effect of stress was only found when the stressed syllables were cued by F0 peaks. We suggest that when lexical stress coincides with the word edges and is realized by an F0 increase, it is interpreted as a stressed syllable that bears pitch accent, and it thus acts as a prosodic cue at the phrasal level. Our research suggests that background phonological

competence modulates segmentation and recognition of statistically defined words in an unknown language. Word-level prosody can restrain or disrupt segmentation, while phrasal prosody can facilitate segmentation. We propose that the interaction of TPs and prosody in the process of word segmentation is accounted for by extending the model proposed by Shukla et al. (2007) to other levels of the prosodic hierarchy, from the IP to the phonological phrase and to the word. Shukla et al. (2007) showed that IP boundaries act as a filter disallowing TP defined words that straddle IP boundaries. They also showed that TPs are computed online and independently from prosodic regularities, though they are extracted simultaneously. This conclusion was later supported by Toro et al. (2011), who showed that TP computations are performed online, irrespective of either prosodic regularities or language-specific phonological constraints.

Indirect evidence in support of the suggestion that TPs are calculated independently of phonological constraints on segmental structures and prosody comes from the field of neuroscience brain imaging. Prosody and phonemes are processed in different hemispheres. For example, Telkemeyer et al. (2009) revealed that, already in infancy, the left and the right hemispheres are sensitive to different time resolutions. The left hemisphere, which is traditionally considered dominant for speech processing, is sensitive to fast temporal modulations and, consequently, is tuned to phonemic perception and discrimination. The right hemisphere, which is traditionally considered engaged in music perception, is responsive to slow temporal modulations and is thus tuned to prosodic information. If prosody and phonemes are processed in different brain areas from infancy, this can explain why TP computations and extraction of prosodic structures are processes that develop in parallel, but independently from one another, and why online TP computation is not disrupted by prosodic regularities. Prosody can only filter word candidates, by selecting which sequences of syllables with high internal TPs

are coherent with prosody. This filtering effect is noticeable only at the retrieval phase, during test, when a word also retrieves its associated prosody; if the retrieved prosody is illegal, the word is suppressed (Shukla et al., 2007).

We suggest that prosody is used to construct the frames for the sequential constituents of the acoustic speech signal, and then the discrete units, segmented on the basis of TPs, are used to fill in the frames. If the content does not fit the frame, it is suppressed and not recognized as a constituent of the speech signal, although it was segmented online using statistical cues. Making use exclusively of duration cues, participants construct the frames for the lexical words on the basis of phonological knowledge specific to their native language. These frames contain the slots for consecutive syllables, with a longer penultimate slot intended for a stressed syllable. On the basis of pitch cues, two types of frames are constructed, one with the slot for the right edge syllable and another one for the left edge syllable bearing the F0 peak. Pitch and duration together are used to construct the frames for the phrases, with the final slot for the longer syllable due to final lengthening. TPs are used to extract the words from the speech stream. During recognition, sequences of syllables with high internal TPs are used to fill in the slots in the frames. Those sequences that fit the frames are recognized as words, and those that do not fit the frames are suppressed.

Conclusion

In the present study we addressed the issue of how prosodic cues and TPs between adjacent syllables interact in word segmentation in an artificial language. We also investigated the different roles of lexical stress and phrasal prominence in segmentation. In addition, we investigated whether the location of lexical stress is relevant to word segmentation if stress does not coincide with word boundaries. Based on the performed experiments, we are able to draw the

following five conclusions. First, if not regularly at word edges, lexical stress does not facilitate segmentation but is able to disrupt segmentation and word recognition. Second, in the absence of durational cues, pitch peaks mark the edges of the segmented discrete units. Association of the pitch peaks with the boundaries of the sequential constituents in the continuous acoustic stream might be a universal perceptual primitive. Third, prominence manifested with pitch and durational cues simultaneously is interpreted as pitch accent and thus belongs to phrasal prosody that can facilitate segmentation, when aligned with word edges. Fourth, the mechanism of interaction of prosodic cues and TPs can be accounted for by extending Shukla's et al. (2007) model to the word level and the level of phonological phrase. Finally, phonological competence in the native language and cross-linguistic prosodic phenomena are both able to modulate segmentation and recognition of discrete units in an artificial language.

Multiple cues interact in the segmentation of continuous speech streams at different levels of the prosodic hierarchy. Our study shows that linguistic competence restricts statistically-defined word candidates at a lower level of the prosodic hierarchy,, that is, at the level of the phonological word, and facilitates segmentation at higher levels, that is, at the levels of phonological phrase and IP. This suggests that in order to have a clearer account of how listeners deal with multiple segmentation cues, one needs to take into account the hierarchical level on which these cues operate.

Final revised version accepted 13 May 2013

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Audio Files with all Experimental Conditions across the three Experiments

Appendix S2: Preliminary Tests for Potential Stream or Session Bias in the three Experiments

Appendix S3: Acoustic Analysis of the Realization of Pitch Nuclear Accents by 10 Native Italians

REFERENCES

Arvaniti, A. (2000). The phonetics of stress in Greek. *Journal of Greek Linguistics, 1(1)*, 9–39.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Batchelder, E. (2002). Bootstrapping the lexicon: a computational model of infant speech segmentation. *Cognition, 83*, 167–206.

Bertinetto, P.M. (1980). The perception of stress by Italian listeners. *Journal of Phonetics, 8*, 385–395.

Bolinger, D. (1958). A theory of pitch accent in English. *Word, 7*, 109–149.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on Statistical Computations. *Psychological Science,* 16, 451-–9.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93–125.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. P. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33*, 111–153.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language, 51*, 523–547.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates.

Crystal, T. H., & House, A. S. (1987). Segmental duration in connected speech syllables: Syllabic stress. *Journal of the Acoustical Society of America, 83*, 1574–1585.

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

Cutler, A., Dahan, D., & van Donselaar, W. A. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*, 141–202.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25*, 385–400.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology, 24*, 381–410.

Cutler, A., & Norris, D. G. (1988). The role of stressed syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 113–121.

D'Imperio, M. (2001). Focus and tonal structure in Neapolitan Italian. *Speech Communication, 33*, 339–356.

D'Imperio, M. (2002). Italian intonation: An overview and some questions. *Probus, 14(1)*, 37–69.

D'Imperio, M., & Rosenthall, S. (1999). Phonetics and phonology of main stress in Italian. *Phonology, 16*, 1–28.

Dogil, G., & Williams, B. (1999). The phonetic manifestation of word stress. In H. van der Hulst (Ed.), *Word Prosodic Systems in the Languages of Europe* (pp. 273–311). Berlin: Mouton de Gruyter.

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress ''deafness''. *Journal of the Acoustical Society of America, 110*, 1606–1618.

Dutoit, T., Pagel, N., Pierret, F., Bataille, O., & van der Vrecken, O. (1996). The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proceedings of the Fourth International Conference on Spoken Language Processing*, 1393–1396. Accessed January 2 2013 from: http://www.asel.udel.edu/icslp/cdrom/icslp96.htm

Eek, A. (1987). The perception of word stress: A comparison of Estonian and Russian. In Channon, R. & Shockey, L. (Eds.) *In honour of Ilse Lehiste* (pp. 19–32). Providence, RI: Foris.

Endress, A., & Hauser, M. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology, 61*, 177–199.

Finn, A., Hudson Kam, C. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition, 108*, 477-499.

Franco, A., Cleeremans, A., & Destrebecqz, A. (2011). Statistical learning of two artificial languages presented successively: how conscious? *Frontiers in Psychology*, *229*, 1–12.

Fry, D. (1958). Experiments in the perception of stress. *Language and Speech 1*, 205–213.

Gebhart, A., Aslin, R., & Newport, E. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science, 33*, 1087–1116.

Gervain, J. Nespor, M., Mazuka, R., Horie, R., & Mehler J. (2008) Bootstrapping word order in prelexical infants: a Japanese-Italian cross-linguistic study. *Cognitive Psychology*, *57,* 56–74

Goldsmith, J. (1976). *Autosegmental phonology*. . Doctoral dissertation, Massachusetts Institute of Technology. (published by Garland Press, New York, 1979)

Grice, M. (1995). *The intonation of Palermo Italian: Implications for intonation theory.* Tübingen: Niemeyer.

Gussenhoven, C. (2003). *The phonology of tone and intonation*. Cambridge, UK: Cambridge University Press.

Hayes, J.R. & Clark, H.H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley.

House, A., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of Acoustical Society of America, 25*, 105–113.

Johnson, E. K., Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*, 548–567.

Kastrikani, A. (2003). *The temporal correlates of lexical and phrasal stress in Greek*, *exploring rhythmic stress: Durational patterns for the case of Greek words*. Unpublished master's thesis, University of Edinburgh.

Kim, S., Broersma, M., & Cho, T. (2012). The use of prosodic cues in learning new words in an unfamiliar language. *Studies in Second Language Acquisition, 34*, 415–444.

Kim, S., Cho, T., & McQueen, J. (2012). Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language. *Journal of Phonetics, 40*, 443–452.

Kohler, K. (2012). The perception of lexical stress in German: Effects of segmental duration and vowel quality in different prosodic patterns. *Phonetica, 69*, 68–93.

Krämer, M. (2009). *The phonology of italian*. Oxford, UK: Oxford University Press.

Ladd, R. (2008). *Intonational phonology*. Cambridge, UK: Cambridge University Press.

Ladefoged, P., & Maddieson, I. (1995). *The sounds of the world's languages*. Cambridge, MA: Wiley-Blackwell.

Langus, A., Marchetto, E., Bion, R., & Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language, 66*, 285–306.

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

Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., & Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *Journal of the Acoustical Society of America, 104,* 2457–2466.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge, UK: Cambridge University Press.

Manolescu, A., Olson, D., & Ortega-Llebaria, M. (2009). Cues to contrastive focus in Romanian. In M. Vigário, S. Frota, and M. J. Freitas (Eds.), *Phonetics and phonology: Interactions and interrelations*. *Current issues in linguistic theory 306* (pp. 71–90). Amsterdam: John Benjamins.

Mattys, S., White, L., & Melhorn, J. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134*, 477–500.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can effect phonetic discrimination. *Cognition, 82*, B101–B111.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39*, 21–46.

Nakai, S., Kunnari, S., Turk, A., Suomi, K., & Ylitalo, R. (2009). Utterance-final lengthening

and quantity in Northern Finnish. *Journal of Phonetics, 37*, 29–45.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology.* Dordrecht: Foris.

Nespor, M., & Vogel, I. (2007). *Prosodic phonology.* Berlin: Walter de Gruyter.

Onishi, K., Chambers, K., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition, 83*, B13–B23.

Ordin, M. (2011). Palatalization and intrinsic prosodic vowel features in Russian. *Language and Speech, 54*, 547–568.

Ortega-Llebaria, M., & Prieto, P. (2011). Acoustic correlates of stress in central Catalan and Castilian Spanish. *Language and Speech, 54*, 73–97.

Ortega-Llebaria, M., Vanrell, M. M., & Prieto, P. (2010). Catalan speakers' perception of word stress in unaccented contexts. *Journal of the Acoustical Society of America, 127*, 462–471.

Parker, S. (2008). Sound level protrusions as physical correlates of *sonority*. *Journal of Phonetics, 36*, 55–90.

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*, 604–607.

Peperkamp, S., Vendelin, I., & Dupoux, E. (2010). Perception of predictable stress: Across-linguistic investigation. *Journal of Phonetics, 38*, 422–430.

Potisuk, S., Gandour, J., & Harper, M. P. (1996). Acoustic correlates of stress in Thai. *Phonetica, 53*, 200–220.

Prieto, P., D'Imperio, M., & Fivela, B. (2005). Pitch accent alignment in Romance: Primary and secondary associations with metrical structure. *Language and Speech, 48*, 359–396.

Rossi, M. (1997). Italian Intonation. In D. Hirst & A. Cristo (Eds.) *Intonation systems: A survey of twenty languages.* (pp. 219–238). Cambridge, UK: Cambridge University Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101–105.

Saffran, J., Werker, J., & Werner, L. (2006). The infants' auditory world: Hearing, Speech and the beginning of language. In D. Kuhn and R. Siegler (Eds.), *The handbook of child psychology* (Vol. 2, pp. 58–108). Malden, MA: Wiley.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure.* Cambridge, MA: The MIT Press.

Shukla M., Nespor M., & Mehler J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology, 54*, 1–32.

Sluijter, A.M., & Van Heuven, V. (1996). Spectral tilt as an acoustic correlate of linguistic stress. *Journal of Acoustical Society of America, 100*, 2471–2485.

Smith, M. R., Cutler, A., Butterfield, S., & Nimmo-Smith, I. (1989). The perception of rhythm and word boundaries in noise-masked speech. *Journal of Speech and Hearing Research, 32*, 912–920.

Telkemeyer, S., Rossi, S., Koch, S., Nierhaus, T., Steinbrink, J., Poeppel, D., Obrig, H., & Wartenburger, I. (2009). Sensitivity of newborn auditory cortex to the temporal structure of sounds. *The Journal of Neuroscience, 29*, 14726–14733.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Statistical and stress cues in infant word segmentation. *Developmental Psychology, 39*, 706–716.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn*:* Infants′ acquisition of stress-based strategies for word segmentation*. Language Learning and Development, 3*, 73–100.

Toro, H.M., Sebastián-Gallés, N., & Mattys, S. (2009). The role of perceptual salience during the segmentation of connected speech. *European Journal of Cognitive Psychology, 21*, 786–800.

Toro, J., Pons, F., Bion, R., & Sebastián-Gallés, N. (2011). The contribution of language-specific knowledge in the selection of statistically-coherent word candidates. *Journal of Memory and Language, 64*, 171–180.

Turk, A., & Sawusch, J. (1996). The processing of duration and intensity cues to prominence. *Journal of Acoustical Society of America, 99*, 3782–3790.

Turk, A., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35*, 445–472.

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

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactic and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech, 40*, 47–62.

Vroomen, J., Tuomainen, J., & de Gelder, B. (1998). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language, 38*, 133–149.

Weiss, D., Gerfen, C., & Mitchel, A. (2009). Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? *Language Learning and Development, 5*, 30–49.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*, 1707–1717.

Williams, B. (1985). Pitch and duration in Welsh stress perception: The implications for intonation. *Journal of Phonetics, 13*, 381–406.

---

[1] The transitional probability (TP) from syllable A to syllable B is computed by dividing the frequency of an AB syllable sequence by the frequency of A syllable in the speech stream, that is, TP (A→B) = frequency (AB)/frequency (A).

[2] French participants in the study by Toro et al. produced a different pattern of results. They did not show a facilitation effect of final or initial pitch, neither did they show inhibitory effect of pitch in medial position. The French participants succeeded to segment in all four conditions. Their performance was never different from that in the TP-only condition. The authors accounted for the different results with French participants with what they called the stress-deafness effect found in French speakers (Dupoux, Pepperkamp, & Sebastián-Gallés, 2001; Peperkamp, Vendelin, & Dupoux, 2010).

[3] By the term nuclear accent we mean the most prominent and usually the last accent in a phrase.