Context speech rate and duration as cues to native and non-native perception of casually-spoken words in Russian

Elina Banzina[1], *Laura C. Dilley*[2]

[1] Department of Communication Sciences and Disorders, Bowling Green State University, Bowling Green, Ohio, USA

[2] Department of Communicative Sciences and Disorders, Michigan State University, East Lansing, Michigan, USA

EBanzi@bgsu.edu, LDilley@msu.edu

Abstract

How word duration and context speech rate affect lexical perception is unclear. We investigated the influence of these attributes on perception of casually-spoken Russian sentences. In Experiment 1, native Russian speakers performed a transcription task on sentences containing rate manipulations. Experiment 2 was a forced choice task using the same materials involving native Russian speakers and native English speakers with high or low proficiency in Russian. In both experiments, word duration and context speech rate influenced Russian lexical perception in all groups. The results suggest that relative timing cues are critical to accurate lexical perception in casual speech.

Index Terms: L2 speech perception, duration, speech rate, tempo, casual speech, spoken word recognition.

1. Introduction

Speech rate (i.e., the rate of articulation of speech segments) varies throughout the course of a connected discourse (Miller, Grosjean, & Lomanto, 1984). This variation in speech rate can be related to differences in information structure of utterances (e.g., Dahan, Tanenhaus, & Chambers, 2002) or to differences in phrase-level prosodic structure, e.g. slowing down at intonation boundaries or syntactic clauses (Turk & Shattuck-Hufnagel, 2000). Variation in speech timing also can cause changes to the perceived identity of a segment, thereby cueing differences in lexical content (e.g., Kidd, 1989). However, understanding how duration influences perception of spoken words has proven challenging (e.g., Davis et al. 2002).

Recently, we have been exploring whether variation in speech timing can cause changes to the perceived presence or absence of phonological units larger than a segment, i.e., whole words or syllables. For example, Dilley and Pitt (2008) presented listeners with English sentences which contained a function word spoken casually (e.g. the word *or* spoken as /ɚɹ/) in the context of a phonetically similar segment (e.g., /ɚ/ at the end of *minor*) in syntactic contexts where the function word was not obligatory. Either the duration of the function word plus immediately surrounding segments ("target") and/or the speech rate of the surrounding context ("context") were time-altered using computer software. Dilley and Pitt found that fewer function words were perceived when the relative timing pattern of the target and context regions was mismatched than when the rate matched across the entire sentence (either by speeding the whole utterance to the same extent or else applying no rate alteration). Moreover, a second experiment using similar sentences in which a function word was never spoken showed the opposite pattern: *more* function words were perceived when the relative timing pattern of the target and context regions was mismatched than when the rate matched. These results supported the hypothesis that whether a reduced syllable is heard (in this case, a monosyllabic function word) depends on segmental material having a certain minimum relative duration compared with the surrounding speech rate (see also Henry et al., 2009; Vinke et al., 2009; Niebuhr, 2008 for similar results).

The present studies extend this work in several ways. First, we investigated whether speech timing influences lexical perception of whole words or syllables in Russian, a language which shares prosodic and phonological properties with English (e.g., stress and vowel reduction) but which is nevertheless quite distinct (Avanesov, 1956). Assuming Russian speakers show influences of speech timing on lexical perception of whole words or syllables, then the same can be asked of non-native Russian speakers. The second issue addressed in this work was therefore whether native English-speaking individuals learning Russian would use duration and speech rate cues in a manner similar to how native Russians use them, and whether non-natives' ability to use these cues in a native-like manner would increase with second language proficiency level.

The third and final issue addressed by these studies concerned the precise manner in which word duration and context speech rate influence lexical perception. In the previous studies of Dilley and colleagues (Dilley & Pitt, 2008; Henry et al., 2009; Vinke et al., 2009), a mismatch between the rate of a critical word (or word sequence) and the rate of the speech context consistently led to less accurate perception. The precise pattern of responses in these experiments was most compatible with the hypothesis that an extra word or syllable was perceived only when the relative duration of critical, target speech material exceeded some minimum threshold relative to the context speech rate. However, an alternative hypothesis is that any mismatch in rate between a critical word or word sequence and the context speech rate will lead to a drop in lexical perception accuracy.

To address these three issues, we constructed sentences in Russian which contained a critical lexical sequence of one or more words which was phonologically (but not semantically) related to another lexical sequence with one less syllable. Thus the "Long" sequence /stərɑnɑ/ ("side") has one more syllable than the phonologically-related "Short" sequence /strɑnɑ/ ("country"). Critical Long-Short lexical sequence pairs were otherwise morphologically and phonologically heterogeneous, one to the next. Carrier sentences were semantically congruent with both the Long and the Short interpretations of each lexical sequence.

2. Experiment 1

The goals of Experiment 1 were (i) to investigate whether duration influences lexical recognition in perception of casual Russian by Russian native speakers; and (ii) to determine whether any mismatch in rate between a critical word or word sequence, on the one hand, and the context speech rate, on the other hand, is sufficient to yield a drop in lexical perception accuracy.

2.1. Method

Eighteen phonologically-related phrase pairs (e.g., "Short" *страна* /strɑnɑ/ vs. "Long" *сторона* /stərɑnɑ/) were identified; each member of a pair was embedded in semantically unbiased sentence contexts, e.g.:

"Это для меня {*сторона/ страна*} незнакомая". Translation: "This {*side (of town)/ country*} is unknown to me".

Sentences were recorded in Russian in a sound-attenuated booth by three native Russian speakers (2 male and 1 female), all graduate students from Bowling Green State University. Speakers were given a list of both Long and Short phrase experimental sentences and filler sentences, 244 in total; to ensure that speakers notice the one syllable difference in Long and Short words, initial contextual cues were added to the otherwise neutral sentences to give only one possible reading. Speakers were instructed to first read each sentence silently and then speak from memory twice. Instead of explicitly asking speakers to act naturally, casual speech productions were obtained by instructing talkers to speak from memory instead of reading, and placing experimental items strategically later in the long list, when speakers became fatigued and less careful in their speech articulation.

A single token of the Long phrase version of each sentence pair was selected as the basis for experimental items. Tokens were selected for which the critical Long phrase was judged to have been spoken casually and whose intonation patterns were deemed natural in both Long and Short phrase contexts.

Recorded sentences were then subjected to time manipulation using using Praat software (Boersma & Weenink, 2002) by altering the duration of either the Target (the unstressed vowel(s) that distinguish(es) the Long word/phrase from the Short word/phrase, plus one to two immediately surrounding phonemes: not more than 3 segments in total), or the Context (all sentence material before and after the Target). Target and Context portions were spliced out of original utterances, time-compressed by a factor of 0.6 or time-expanded by a factor of 1.9, and recombined. Special care was taken to prevent discontinuities at splicing points (i.e., zero crossings).

The single within-subjects independent variable was Time Manipulation with five levels (Unaltered, Target Compressed, Context Expanded, Target Expanded, and Context Compressed; see Fig. 1). The precise manner of manipulating Target and Context rates/durations enabled testing of two hypotheses about how speech rate influences lexical perception. (i) For the Unaltered condition, no change in rate was imposed on either the Target or Context portions relative to the original rate, i.e., rates of the Target and Context portions *matched*. Moreover, the duration of the Target relative to the Context was expected to be long enough for the extra syllable spoken in the Long phrase to be perceived; that is, the duration of the Target relative to Context was expected to be longer than the minimum relative duration necessary for the syllable to be perceived (i.e., it was *relatively long*). (ii) For the Target Compressed condition, the Target was time-compressed, while the Context rate was unaltered. Thus, the rates of Target and Context were *mismatched*; moreover, the duration of the Target relative to the Context was expected to be shorter than the minimum relative duration necessary for the syllable to be perceived (i.e., it was *relatively short*). (iii) For the Context Expanded condition, the Context was time-expanded in rate, while the Target rate was unaltered. Thus, rates of Target and Context were *mismatched;* moreover, the duration of the Target relative to the Context was expected to be shorter than some minimum relative duration necessary for the syllable to be perceived (i.e., it was *relatively short*). (iv) For the Target Expanded condition, the Target was time-expanded, while the Context was unaltered in rate. The rates of the Target and Context were thus again *mismatched*; however, this time the duration of the Target relative to Context was expected to be much longer than the minimum relative duration necessary for the syllable to be perceived (i.e., it was *relatively long*). (v) For the Context Compressed condition, the Context was time-compressed, while the Target was unaltered in rate. The rates of the Target and Context were thus once again *mismatched*; moreover, the duration of the Target relative to Context was expected to be much longer than the minimum relative duration necessary to perceive the extra syllable (i.e., it was *relatively long*).

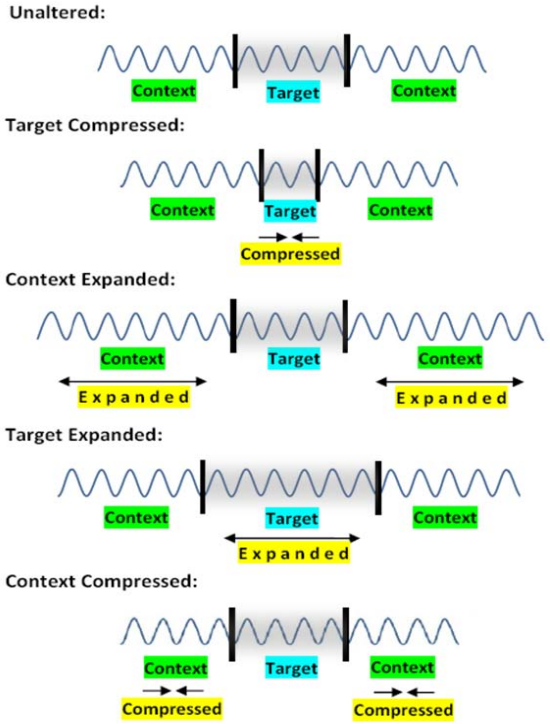

Figure 1: *Illustration of temporal manipulations to Target and/or Context material for each of the five Time Manipulation conditions. Arrows pointing outward indicate time-expansion, while arrows pointing inward indicate time-compression.*

Five lists were constructed from 18 experimental items and 22 filler items; the first five stimuli on the list were filler items, and the remaining items occurred in quasi-random order with the constraint that no more than three of one type of item (experimental or filler) occurred in a row. Each experimental item was presented only once on a list, with the pairing of experimental items and conditions counterbalanced across the five lists. Approximately one-third of the filler items was temporally modified by time-compressing by a rate of 0.6,

while another third was time-expanded by a rate of 1.9, respectively; the remaining items were not altered in rate.

The participants were twenty native Russian speakers residing in Latvia (13 male, 7 female), all at least 18 years of age and with self-reported normal hearing. The experiment was presented via Praat software. Participants were seated in front of a computer with headphones on. A paper answer sheet was provided to participants on which a series of sentences appeared, each with a blank space. Participants were instructed to click on a button on the computer screen, which would cause a sound file to play; they then wrote down the word they heard corresponding to the blank in each sentence. Participants could listen to each sentence twice, and could proceed through trials at their own pace.

2.2. Results and Discussion

The rate of a "Long" lexical sequence response was coded; participants gave a "Long" or "Short" response on all trials. Figure 2 shows the proportion of "Long" responses for each Time Manipulation condition. A one-way repeated measures ANOVA revealed a significant effect of Time Manipulation on proportion of "Long" responses [$F(4,76) = 32.135$, $p < 0.01$]. To further assess differences across conditions, a series of planned comparisons were conducted using two-tailed, paired-samples *t*-tests with Bonferroni corrections. Significantly more "Long" responses were given in the Unaltered condition than in either the Target Compressed condition ($p < .001$) or the Context Expanded condition ($p < .001$). Moreover, there was no difference in the proportion of "Long" responses comparing the Target Compressed and Context Expanded conditions ($p = .10$), suggesting that slowing the Context rate was just as effective at eliciting "Short" responses as speeding the Target. Moreover, there was no difference in the proportion of "Long" responses comparing the Unaltered condition and the average of the Target Expanded and Context Compressed conditions ($p = .57$).

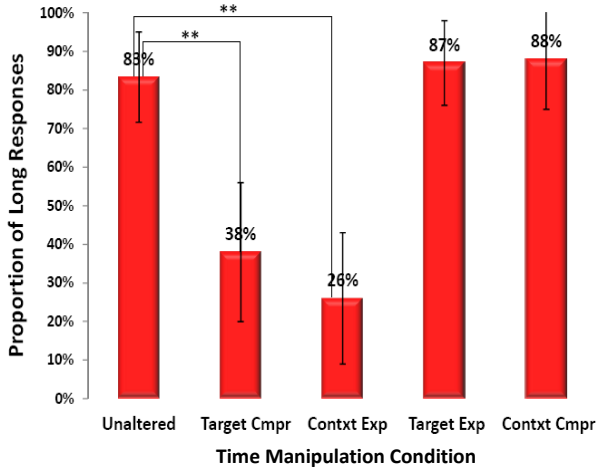

Figure 2: *Rate of "Long" responses to ambiguous portions of each experimental stimulus in Experiment 1. Conditions which were significantly different are shown with asterisks (**). See text.*

These results show that native Russian listeners relied on temporal information in lexical perception of Russian speech. Moreover, the present experiment clarifies the manner in which timing information may influence duration on lexical perception. In particular, not every mismatch in rate between the Target and Context portions of the speech resulted in a drop in accuracy (i.e., a drop in "Long" responses, given that sentences containing Long phrases were actually spoken). Instead, the only rate conditions which altered lexical perception to be something other than the veridical phrase were those in which the Target duration was *relatively short* compared with the Context duration (through either shortening the Target or lengthening the Context). These findings suggest therefore that in order for casually spoken, reduced syllables in Russian to be heard, they must exceed a certain minimum duration as defined relative to the context speech rate.

3. Experiment 2

3.1. Method

The goal of Experiment 2 was to determine whether non-native learners of Russian would similarly rely on temporal information in word recognition in Russian, and whether this reliance grows with experience in a second language (L2).

The experiment was a 3 x 5 mixed factorial design, with Proficiency level (Native, High-proficiency and Low-proficiency) as a between-subjects variable, and Time Manipulation (Unaltered, Target Compressed, Context Expanded, Target Expanded, and Context Compressed) as a within-subjects variable.

There were 28 participants in the experiment, all of whom were at least 18 years of age. The Native Russian-speaking group consisted of ten participants, all graduate students from Bowling Green State University (6 male, 4 female). The Low-proficiency, non-native Russian-speaking group consisted of ten native English speakers from Bowling Green State University and Michigan State University (3 male, 7 female). Low-proficiency participants had either (i) 1-2 years of formal instruction in Russian, and/or (ii) 1-2 years experience living in a Russian speaking country, and/or (iii) active daily communication in Russian with a native speaker of Russian for 2 years. The High-proficiency, non-native Russian-speaking group consisted of eight native English speakers from Bowling Green State University and Michigan State University (5 male, 3 female). These participants had either (i) formal instruction in Russian not less than 4 years, and/or (ii) a minimum of 4 years living in a Russian-speaking country, and/or (iii) a minimum of 5 years active daily communication in Russian with a Russian native speaker. Proficiency level was based on demographic data and self-reports.

The stimuli for Experiment 2 were the same as in Experiment 1. A two-alternative forced choice task was used due to the limited vocabulary knowledge of the lower-level Russian learners. Each participant saw a list of 18 experimental sentences and 22 filler items, each with a choice of Long and Short phrases from phonetically-related pairs. Participants were instructed to listen to sound files over headphones by clicking on buttons on the computer screen, which would cause a sound file to play, and then to circle one out of the two options provided for each sentence in their answer sheets. Again, participants could listen to each sentence twice, and could proceed through trials at their own pace.

3.2. Results and Discussion

Figure 3 shows the proportion of "Long" responses according to Time Manipulation condition for the three groups differing in Russian proficiency level. A two-way, mixed measures ANOVA showed a significant main effect of Time Manipulation [$F(4,100) = 24.677$, $p < .001$] and a marginally significant effect of Proficiency level [$F(2,25) = 2.787$, $p = .081$]; there was no interaction. Post-hoc, Tukey's Honestly Significant Difference tests revealed that Native speakers were

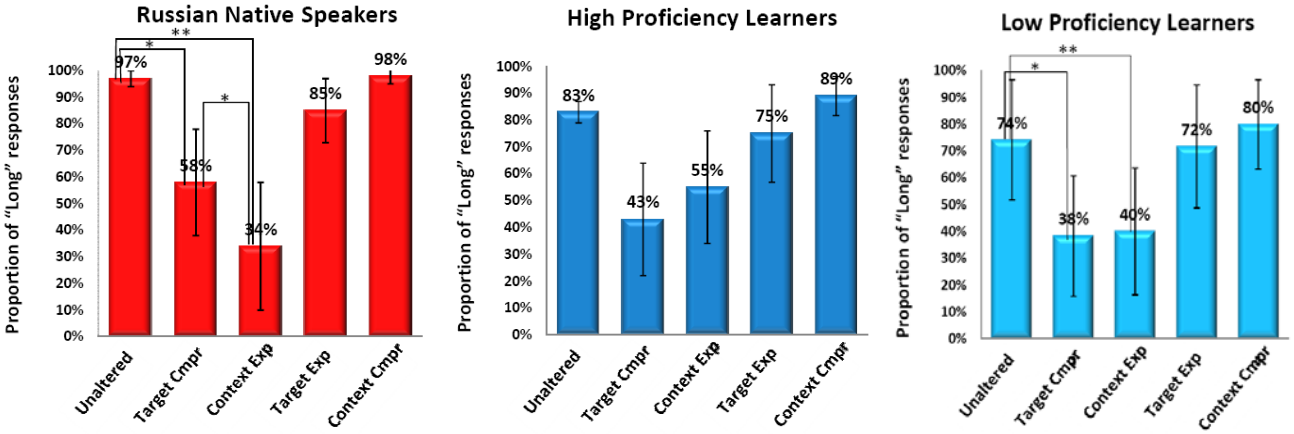

Figure 3: *Rate of "Long" responses to ambiguous portions of each experimental stimulus in the forced-choice task of Experiment 2. Conditions which were significantly different using planned comparisons (paired samples t-tests) with Bonferroni correction are shown (**p < .001, *p < .05).*

marginally significantly different in performance compared with Low proficiency learners ($p = 0.067$).

Separate one-way, repeated measures ANOVAs for each group with Time Manipulation as the factor showed a significant effect for Russian Native speakers [$F(4,36) = 20.946$, $p < .001$], High Proficiency learners [$F(4,28) = 3.645$, $p < .001$] and Low Proficiency learners [$F(4,36) = 8.749$, $p < .001$]. To further assess differences across conditions, a series of planned comparisons were conducted using two-tailed, paired-samples *t*-tests with Bonferroni corrections. Significant differences are indicated in Figure 3.

These results suggest that non-native Russian speakers use duration cues during word recognition, even if those speakers are not proficient in their L2. Moreover, preliminary evidence suggests that ability to use duration cues in word recognition in a native-like manner increases with learners' proficiency in a second language, as evidenced by the trend in increases in "Long" responses with greater proficiency level in Russian. Finally, the finding that temporal cues are used in Russian in perception of lexical items containing reduced syllables was confirmed using a different task than that of Experiment 1.

4. General Discussion and Conclusions

In this paper, we investigated the role of duration and context speech rate in perceiving casual speech in Russian. Previous research (Dilley & Pitt, 2008; Henry et al., 2008; Vinke et al., 2009) has shown that native English listeners can use temporal information to perceive lexical items in casual speech in English. In two experiments, both native and non-native speakers of Russian were demonstrated to show reliance on duration and context speech rate in Russian word recognition in the presence of low-quality spectral information. In Experiment 1, native Russian speakers gave a free response about the words they heard. In Experiment 2, native Russian speakers, as well as high- and low-proficiency native English-speaking learners of Russian, performed a two-alternative, forced choice task about what words they heard.

Across both experiments, all groups showed reliance on lexical duration and context speech rate in order to determine whether they heard phonologically shorter or longer lexical sequences. Moreover, speech timing information was used in a relative manner; participants gave lexical responses which were phonologically longer (i.e., contained more syllables, relative to an alternative shorter lexical interpretation) only when the duration of the target material exceeded a minimum relative threshold compared with the duration of the context, not merely when the speech rates of the target and context material mismatched. In addition, preliminary evidence was obtained that the ability to use duration and context speech rate to perceive words in a non-native language increases with proficiency level in one's L2. These results collectively have implications for understanding how timing information is used in perceiving lexical and prosodic information in spoken language by L1 and L2 speakers. These findings help to explain how spoken word recognition can be so robust in spite of often impoverished spectral information to phonemic and lexical content.

5. Acknowledgements

We are grateful to Amanda Millett, Claire Carpenter and Shaina Selbig for their research assistance. This research is supported by NSF Award BCS0874653 to L. Dilley.

6. References

Avanesov, R. I. (1956). *Fonetika Sovremennogo Russkogo Literaturnogo Jazyka (Phonetics of Modern Russian Language).* Moscow: Izdatelstvo Moskovskogo Universiteta.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47, 292-314.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, G. (2002). Leading up the lexical garden path: segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 218-244.

Dilley, L. C. & Pitt, M. A. (2008). Now you hear it, now you don't: Effects of speech rate on function word perception. Paper presented at the 49th Annual Meeting of the Psychonomic Society, Chicago.

Henry, M. J. , Dilley, L. C., Vinke, L. N. & Weinland, C. J. (2009). Duration and context speech rate as cues to lexical perception and word segmentation. *Journal of the Acoustical Society of America*, 125, 2655.

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 736-748.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41, 215-225.

Niebuhr, O. (2008). Identification of highly reduced words by differential segmental lengthening. Talk presented at the First Nijmegen Speech Reduction Workshop, Max Planck Institute, Nijmegen, The Netherlands.

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, 28, 397-440.

Vinke, L. N., Dilley, L. C., Banzina, E., & Henry, M. J. (2009). Lexical perception and segmentation of words beginning with reduced vowels: A role for timing cues. Paper presented at the 50th Annual Meeting of the Psychonomic Society, Boston.