Developing a Transcription of Russian Intonation (ToRI)

Two experiments for an evaluation of Russian pitch accents

Cecilia Odé
University of Amsterdam
Institute of Phonetic Sciences/ACLC, Herengracht 338, 1016 CG Amsterdam, The Netherlands
E-mail: cecilia.ode@hum.uva.nl

ABSTRACT

A manual transcription system using unambiguous symbols that express experimentally verified forms and contextual functions of Russian intonation is under development: ToRI (Transcription of Russian Intonation). ToRI, inspired by ToDI [1], will be implemented on the Internet as an interactive research tool and learning module. For theoretical and language-specific reasons ToRI will differ considerably from ToDI, also because contextual functions will be described. A first step toward ToRI symbols is the evaluation of the classification in my 1989 thesis [2] in eleven experimentally verified types of pitch accent occurring in spontaneous and prepared speech. The question was whether the types would be directly usable for ToRI. Two experiments were conducted to verify the perceptual equivalence of the types with realizations of types in spontaneous and prepared speech by other speakers. Results show that my classification can be used, if taking into account the limits of perceptual tolerance of the types of pitch accent. Intonation change may also be an issue.

1. INTRODUCTION

The aim of the two experiments was to check whether my classification of Russian pitch accents described in [2] covers all existing types, and thus would directly be usable for ToRI, or needs to be adjusted. In other words, is the then used corpus of 15 minutes of spontaneous and prepared speech (henceforth: corpus A) an exhaustive representation of types of Russian intonation? If the answer to the question is positive, the classification can serve as a basis for the development of ToRI symbols for the form of Russian pitch accents. The classification was tested in the two experiments as follows. I compared realizations of one type of pitch accent from corpus A to melodically similar realizations in a new corpus in order to verify their perceptual equivalence, that is, one realization being a good imitation of a melodically similar realization. The new corpus (henceforth: corpus B) consists of not yet analysed spontaneous and prepared speech from other speakers than corpus A. The digital recordings used for Corpus B have recently been made in St Petersburg; speakers are from 18 to 50 years old. On the basis of perceptual equivalence I tentatively classified realizations of pitch accents in corpus B into one of the types of pitch accent for verification in a paired-comparison experiment. I expected that listeners could have problems with corpus B realizations of types of pitch accent with phonetic specifications at the extreme of the limits of perceptual tolerance of that type. These extremes are the minimum and maximum values of the phonetic specifications of pitch accents: excursion size, posttonic part, timing (see section 2). According to the results of perception experiments and the phonetic specifications described in [2], stimuli were still within those limits. I expected stimulus pairs that are *not* realized at the extreme of those limits to score high as to perceptual equivalence. Realizations in corpus B of which the type of accent could not be identified were not included. In order to find out to which type those realizations belong, or to see if any new types of accent would be revealed, I selected them for a classification experiment. The experiments were carried out in St Petersburg and Moscow in 2002.

2. STIMULI FOR THE EXPERIMENTS

Stimuli for the *paired-comparison experiment* were utterances pronounced by male and female speakers of standard Russian and were segmented by means of PRAAT [3] from original recordings. Stimuli were presented in their original realization without any manipulation, except for levelling out the volume of stimuli that came from different recordings. For each of the eleven types of pitch accent (see below), ten realizations in short utterances were selected from corpus A and ten from corpus B that according to my tentative classification (see section 1) belong to the same type. The ten realizations for each of the eleven types selected from corpus A were considered representative. That is, they were good examples of these types according to their phonetic specifications described in chapter 6 of [2] and not situated at the extremes of the limits of perceptual tolerance. Since the question was whether my classification covers all possible realizations of the eleven pitch accents, this was not a criterion for the realizations from corpus B. The paired comparison thus consisted of 110 pairs of utterances: ten realizations of eleven types of pitch accent from the two corpora. Stimuli were selected from utterances in such a way that, though isolated from their semantic context, they did not sound odd. In order to avoid problems in comparing speakers with different registers, two stimuli of a pair were always pronounced by two male or two female speakers having more or less the same register. There were 26 pairs pronounced by four female, 84

 ISBN 1-876346-48-5

pairs by eleven male speakers. The eleven pitch accents described with names after their phonetic specifications are the following: Rh-, Rl-, Rm-/+, rl-/+, rm-/+, Fl-, Fl+, Fnl-, Fnl+, Fh-, f-/+ [2]. The five rising pitch accents have a large excursion (R), a small excursion (r), three different posttonic parts, high (h), mid (m) and low (l), and different timing (the position in the accented syllable where the end frequency of a pitch movement is reached), namely early (-) or late (+). The six falling accents have a large excursion (F), a small excursion (f), three different posttonic parts, low (l), non-low (nl) and high (h), and different timing, early (-) or late (+). For type rl-/+ there were only seven realizations in corpus A, therefore three of them occurred twice. Two types of pitch accent described in [2], viz. Rø and F^n+, were not considered necessary to be included in the experiment. The former is a neutralization of accents Rh- and Rl- (the accent is situated in the final syllable, so there is no posttonic part h or l), and the latter is a pitch contour with repeated realizations of accent Fnl+/Fl+.

The *classification experiment* consisted of fifty utterances from corpus B, 25 by two female and 25 by two male speakers, with unidentified pitch accents (see section 1). As reference accents, two series of eleven short utterances with a representative (see above) realization of each type were selected from corpus A and stored under two rows of eleven buttons: one row with realizations by two female and one row with realizations by five male speakers. Realizations were selected that according to [2] have the most frequently occurring phonetic specifications. Reasons for a male and a female series are described above; two series also enabled subjects to compare stimuli to two reference accents.

3. TASKS AND SUBJECTS

Subjects that participated in the *paired-comparison experiment* were allowed to repeatedly listen to a pair of realizations of pitch accent, one from each corpus. Their task was to indicate the perceptual equivalence of the members of each pair. After an instruction in Russian on screen, pairs were presented to them without text (problems in understanding the stimuli were not expected, but see section 6) and in random order. Subjects did not know that the experiment started with five training stimuli that were randomly chosen from the list of stimuli and that occurred in the experiment once more. Subjects accepted or rejected the perceptual equivalence of a pair by pressing the button "same" or "different", same implying that a pair is perceptually equivalent, different that it is not. By pressing the button "next" a following pair appeared.

For the *classification experiment* in two runs, subjects repeatedly listened to stimuli in random order that like the paired comparison appeared without text. Subjects could press the button "next" to proceed. They compared the test stimuli to the realizations of the eleven reference types by pressing the unnamed buttons numbered 1 to 11 under which the types were stored. In the first run subjects matched accents by pressing the button with the number of the type they considered the realization perceptually equivalent to (forced choice), and in the second run they were allowed to choose an extra, twelfth button "no match" whenever they found that none of the eleven reference accents was perceptually equivalent to the given stimulus. In this type of experiment it was not considered necessary to include training stimuli.

Twelve subjects, two male and ten female, participated in the paired comparison. Six subjects, five female and one male, took part in the classification experiment. Three subjects participated in both experiments with an interval between them of a few days (one subject) or weeks (the other two). Subjects did not need further explanations after the instruction. Their comments after the experiments will be discussed in the sections below. Subjects, phoneticians with a linguistic background or linguists, were experienced native listeners. The experiments were controlled by a computer and subjects used headphones. The tasks were executed individually without limitation in time.

4. RESULTS PAIRED COMPARISON

The results for the paired comparison are presented in Figures 1 and 2.

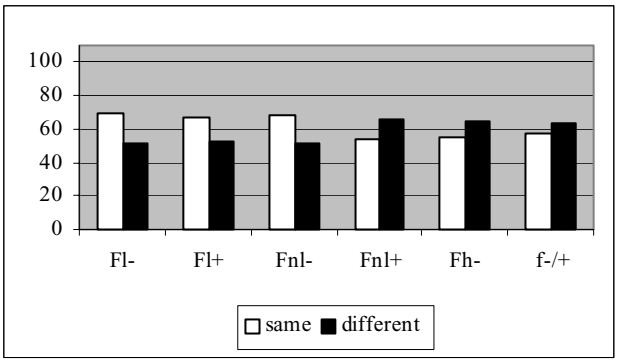

Figure 1: Results for the falling pitch accents. The histograms show real numbers of same/perceptually equivalent (white) and different/not perceptually equivalent (black) realizations. On the y-axis the number of stimulus pairs (maximum 110); on the x-axis the types of pitch accent.

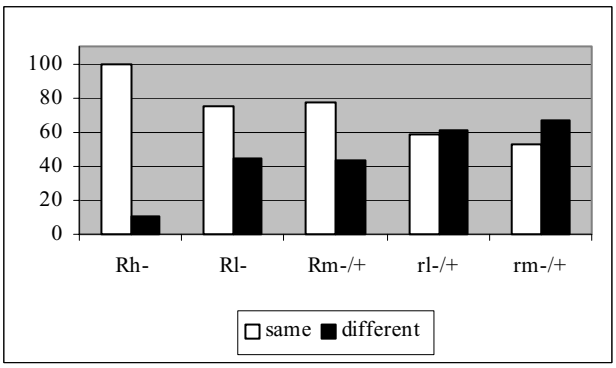

Figure 2: Results for the rising pitch accents. The histograms show real numbers of same/perceptually equivalent (white) and different/not perceptually equivalent (black) realizations. On the y-axis the number of stimulus pairs (maximum 110); on the x-axis the types of pitch accent.

The histograms show real numbers of same/perceptually equivalent (white) or different/not perceptually equivalent (black) realizations, with on the y-axis the number of stimulus pairs (maximum 110), on the x-axis the types of pitch accent. Figure 1 shows that falling accents Fl-, Fl+, Fnl-, and Figure 2 that rising accents Rh-, Rl- and Rm-/+ give the highest score for perceptual equivalence; falling accents Fnl+, Fh-, f-/+, and small rising accents rl-/+ and rm-/+, score almost chance. The time needed to fulfil the task was one hour (1 subject), 45 minutes (1 subject), 30 minutes (5 subjects) and about 15 minutes (5 subjects).

After completing the experiment, subjects reported their problems with the task. Most of them did not know what to concentrate on if the duration of stimuli or the number of accents in the stimuli were not equal. The stimulus pairs with supposedly perceptually equivalent pitch accents were indeed not always equal as to their duration and number of pitch accents occurring in them. In this respect there were three types of stimulus pairs: 1) pairs with one pitch accent in each stimulus, 2) pairs with two different pitch accents in each stimulus, and 3) pairs with an unequal number of pitch accents. Note that if two accents occurred in one stimulus, there was in my perception only one type in each stimulus that could possibly be compared, the others being completely different. In Figures 3 and 4 results are now presented separately, taking into account the inequality of stimuli: same/perceptually equivalent pairs in "equal" pairs (white), same/perceptually equivalent pairs in "unequal" pairs (diagonal stripes), and different/not perceptually equivalent pairs in "equal" pairs (black) and different/not perceptually equivalent pairs in "unequal" pairs (horizontal stripes).

Given the expectation that most pairs would be perceptually equivalent, subjects had less problems with "equal" than with "unequal" stimulus pairs, especially for rising pitch accents. According to the Chi-square test for equality of distributions, the difference between "equal" and "unequal" pairs is highly significant: $p<0.007$, except for Fl+ and Fnl+ with $p<0.02$ and $p<0.03$, respectively.

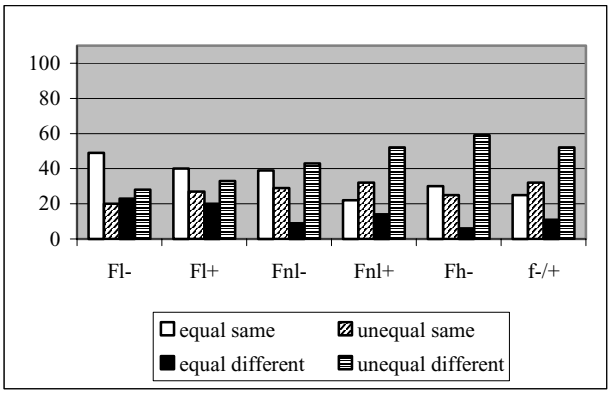

Figure 3: Results of "equal" vs. "unequal" pairs for falling pitch accents. The histograms show real numbers of same/perceptually equivalent realizations in "equal" (white) and "unequal" (diagonal stripes) pairs, and different/not perceptually equivalent in "equal" (black) and in "unequal" pairs (horizontal stripes). On the y-axis the number of stimulus pairs (maximum 110); on the x-axis the types of pitch accent.

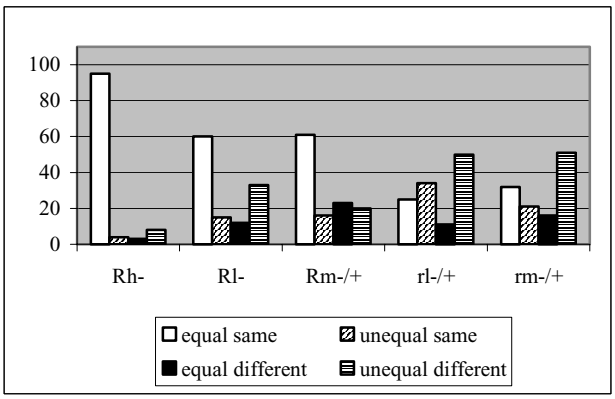

Figure 4: Results of "equal" vs. "unequal" pairs for rising pitch accents. The histograms show real numbers of same/perceptually equivalent realizations in "equal" (white) and "unequal" (diagonal stripes) pairs, and different/not perceptually equivalent in "equal" (black) and in "unequal" pairs (horizontal stripes). On the y-axis the number of stimulus pairs (maximum 110); on the x-axis the types of pitch accent.

By looking at the scores for subjects individually, I found that according to the Chi-square test for equality of distributions, there is a significant difference between them of $p<0.000000$. Their scores are shown in Figure 5. Only subjects 2, 5 and 8 had less problems with "unequal" pairs.

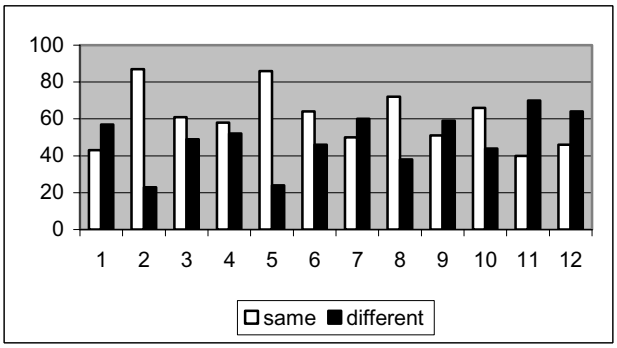

Figure 5: Results of the paired comparison for the individual subjects. The histograms show real numbers of same/perceptually equivalent (white) and different/not perceptually equivalent (black) realizations. On the y-axis the number of stimulus pairs (maximum 110); on the x-axis the twelve subjects.

Finally, I expected problems with realizations of types of pitch accent with phonetic specifications at the extreme of the limits of perceptual tolerance. Yet stimuli at those extremes were also selected for verification, because, ideally, my classification should cover *all* realizations of a given type. This issue is discussed in section 6.

5. RESULTS CLASSIFICATION

In the first run the six subjects that participated in the classification experiment needed for their matching task from 19 minutes to more than one hour: 19', 25', 32', 27', 53', 1:08'. In the second run they needed less time, except subject 3. In the same order of subjects the task took them 10', 16', 48', 43', 23', 33', respectively. As said in section 1, it was expected that it would be easier to compare the 50 test stimuli to the eleven reference accents if pronounced in

ISBN 1-876346-48-5

more or less the same register. Indeed, in only 6% of the responses subjects matched a male stimulus with a female reference accent or the reverse. The aim of the first run was to find out which of the eleven reference accents from corpus A, the 50 realizations from corpus B were perceptually equivalent to (forced choice); in the second run subjects were allowed to indicate that a given realization was not perceptually equivalent to any of the eleven reference accents ("no match"). The six subjects did not much agree among each other and among themselves. For example, female stimulus 18 was for two subjects perceptually equivalent to accent Rl-, for two to Rh-, for one to Rm- and for one to Fh-, all different and rather salient types of accent. There are many such examples. The number of stimuli that were matched with the same reference accent in both runs varies per subject from 10 to 22 stimuli with a total of 96 stimuli out of 300 (50 stimuli x 6 subjects) or 32%. The number of stimuli that in the first run were matched with another reference accent than in the second run varies per subject from 15 to 29 stimuli with a total of 134 out of 300 or 44%. The number of "no match" stimuli in the second run varies per subject from 4 to 19 with a total of 70 out of 300 or 23%. See Table 1.

	chosen accents in percentages
same reference accent	32 (20-44)
other reference accent	44 (30-58)
"no match"	23 (8-38)

Table 1: Results in percentages for the second run as compared to the first run: same, other reference accent or "no match"; between brackets the minimum and maximum percentages for the subjects.

The reference accents that were chosen in the first run for the 70 stimuli that were "no match" in the second run are the following: 6x Rh-, 5x Rl-, 13x Rm-/+, 11x rl-/+, 2x rm-/+, 6x Fl-, 4x Fl+, 0x Fnl-, 5x Fnl+, 5x Fh-, 13x f-/+. But from whatever point of view one arranges the results, stimuli have been scattered seemingly random over the eleven reference accents.

6. DISCUSSION

The results for the *paired-comparison experiment* in "equal" pairs show that the classification in eleven pitch accents as described in [2] needs not to be adjusted as long as pitch accents are realized away from the extremes of the limits of perceptual tolerance. Realizations close to or at the extreme of these limits cause confusion among subjects. It was probably those realizations that were responsible for the fact that the score for "equal" pairs was not higher. There may also be some overlap at the formal borderline between types of accents. This means that the distance between the minimum and maximum values of the phonetic specifications of the accents that will be translated into symbols for ToRI must be much smaller. For example, if the experimentally verified excursion size of an accent lies between 13 and 21 semitones, the corresponding symbol for this accent must be defined as an accent with an excursion size of 17 semitones. By analysing some of the "worst scoring" stimuli from corpus B it was observed that those stimuli are indeed situated at the extremes. This must be further analysed. Next, in future experiments, stimuli of different duration with an "unequal" number of pitch accents must be avoided. In the present experiment a solution for "unequal" pairs would have been to present the text of the stimuli and to underline the words in which the pitch accents to be compared occur. However, texts may introduce other problems. Finally, pairs with different contextual functions should be excluded if the perceptual equivalence of forms must be compared. The problem is whether pitch accents that are *in melody perceptually equivalent* can have *different contextual functions* and pitch accents that are *in melody not perceptually equivalent* can have *same contextual functions*.

The confusing results from the *classification experiment* can partly be explained. I could not identify the stimuli. Obviously, with a few exceptions, neither can the subjects. Stimuli that did not match in the second run were in the first run considered to be perceptually equivalent to types of accents that I would call more or less "neutral" with respect to their contextual function. These types, frequently chosen by subjects, are rm-/+, rl-/+, f-/+, Rm-/+. The former three occur sentence internal, not at boundaries, and Rm-/+ frequently occurs at a boundary as a continuation (see section 5). Furthermore, I suspect that also for this experiment "unequal" stimuli and realizations of stimuli at the extreme of limits of perceptual tolerance are responsible for this outcome. Another factor that one cannot rule out is the not yet analysed effect of observed changes in intonation during one generation. Many stimuli were pronounced by young speakers from 18 to 24 years old; two subjects who are also from that generation, carried out tasks quicker than the generation above 40. In developing ToRI, the factors mentioned must be taken into account.

ACKNOWLEDGMENTS

This research is financially supported by the Netherlands Organisation for Scientific Research, NWO. The author expresses her gratitude to the Phonetic Institute of St Petersburg University for supplying recordings, to the subjects involved in the experiments, and to Rob van Son and Ton Wempe for their technical assistance.

REFERENCES

[1] Gussenhoven, C., T. Rietveld, J. Terken. *ToDI, Transcription of Dutch intonation, Courseware.* http://lands.let.kun.nl/todi, Nijmegen 1999.

[2] Odé, Cecilia. *Russian Intonation: A Perceptual Description.* Amsterdam, Rodopi 1989.

[3] Boersma, P. PRAAT version 4.0.34, http:// www.fon.hum.nl/praat.