# “Russian Pear Chats and Stories”: Vocal annotation guide

*(Version 10.01.2019)*

**Nikolay A. Korotaev**[1]

Vocal annotation embraces two communicative channels, the verbal and the prosodic (see Kibrik 2018), which implies two large groups of tasks for the annotators:

(i) The speech flow should be segmented into significant units of various levels (elementary discourse units, words, filled pauses, etc.); for each unit, its precise time boundaries should be indicated. Various non-speech vocals acts, laugh, whistles, clicks and such should undergo similar analysis.

(ii) Also, we annotate specific features of speech units such as prosodic characteristics, illocutionary and phase values, speech repairs, etc. that contribute most to the local discourse structure.

The main principles of vocal annotation were developed and described by Kibrik and Podlesskaya (2009); see also http://www.spokencorpora.ru/, Kibrik et al. (2019) in print. For this project, however, we have introduced several new solutions that adjust the annotation procedure taking into account the interactive and multichannel properties of the sessions. In this text, we provide a brief overview of the general annotation principles, describe the stages of the annotation procedure and explain transcription symbols. A more detailed analysis of the phenomena in question can be found in the above-mentioned works published by our research group.

## 1. Annotation procedure and individual annotation format

### 1.1. General observations

For each session, the vocal behavior of the three main participants (the Narrator, the Commentator, and the Reteller) is annotated *separately*. (Below, in Section 7, we describe a scores transcript format that gives a unified vocal representation of all the three participants. However, such transcripts are not created within the annotation process as such but are automatically compiled once all individual annotations have been completed.)

The main source for annotation is found in the individual files that contain recordings from the individual microphones of the participants (see http://multidiscourse.ru/solutions/). The common audio files and the video files with superimposed sound are also used as supplementary materials. The common audio files might help with understanding the overall context of the communicative exchange. In the individual audio files, the voice of the speaker who is being recorded dampens the vocal acts of the other participants; in the common audio files, the volume of all participants is brought to an equal level. It is particularly important to compare individual audio files with the common one when analyzing dialogical fragments. The common audio file also contains useful information about integral characteristics of the session (see Section 1.2 below). Individual video files provide the necessary information for a more precise interpretation of the vocal acts: unclear fragments may become easier to understand when one analyzes lip movements and overall kinetic behavior. The video signal also helps to identify the source of sounds which are not directly connected with one’s vocal behavior (such as when speakers are scratching, accidentally touch the microphone, clap their hands, etc.). Nevertheless, the most exhaustive and thus the priority annotation method is the one provided by the perceptive and instrumental analysis of individual audio files.

[1] To cite this version:

Korotaev, N.A. “Russian Pear Chats and Stories”: Vocal annotation guide. Version 10.01.2019. http://multidiscourse.ru

We recommended using Praat software to work with the audio files and to analyze their acoustic characteristics. This software is available for free download, and it is equipped with the detailed reference system. Readers who are not familiar with the general principles of using Praat or with the specifics of creating and editing textgrids (see Sections 2 and 3 below) can find respective information on the subjects on the software webpage http://www.fon.hum.uva.nl/praat/.

## 1.2. Stages of annotation

The procedure for vocal annotation is split into several stages. The number of annotators as well as the level of their expertise influences the way the overall annotation work is distributed across stages. Below you will find a scheme where the first stages are delegated to the less experienced annotators, and then the results of their work are further processed by the annotators with more experience in the field.

### *Preliminary stage: Getting acquainted with the session*

Before starting the annotation procedure, we recommend listening to the file from the beginning until the end. This is the most convenient way to familiarize oneself with the contents of the recording, the participants' manner of speaking, fragments of overlapping, etc. The manner of speaking can be connected with the overall communicative purposes of the participants, which is exactly why it is also recommended to watch the cover shot video files, either partly or from the beginning until the end. Observations made during this preliminary stage play a very important role for the subsequent annotation process. Since the main annotation is carried out on relatively short speech fragments, it is useful to understand integral characteristics of the entire session and its larger parts. This allows you to consider specific phenomena as driven by the overall strategies of particulars speakers and/or by the interaction context at a particular stage of the session, rather than treating them as isolated instances.

### *Stage one: Basic transcribing, setting time boundaries and preliminary segmentation into EDUs*

At the first stage of annotation, separate words, filled pauses and several kinds of non-speech vocal events are identified for each participant. For all of these units, which will later be referred to as *lower-level segmentation units*, we define the time boundaries, i.e. the point where the sound begins and where it ends in reference to the beginning of the file. One should pay special attention so that each unit is ascribed to the right speaker and no unit is omitted in the transcript. To achieve this, annotators should listen to the audio files very attentively and always do so from the beginning until the end of the file, including those fragments where the corresponding speaker remains silent due to their specific role in the session. For instance, Retellers, who were asked not to interrupt Narrators at the stage of their first telling, nevertheless regularly provide feedback signals which should be accordingly annotated. In difficult cases, we recommend turning to the common audio file and/or to the cover shot and individual video files. The specifics of lower-level annotation are discussed later in Section 2.

The first stage of annotation also implies preliminary segmentation of the speech flow into *elementary discourse units* (further referred to as EDUs). The EDU is a central notion of the discourse transcription system used in our project. Each EDU is a minimal step in the development of discourse; for more detailed information see Section 3 and the works referenced there. In the overall annotation scheme (see "Multichannel annotation in ELAN" document, http://multidiscourse.ru/data/ann/pears_multichannel_annotation_en.pdf), EDUs are regarded as *higher-level segmentation units*.

*Stage two: Revision of the segmentation into EDUs, detailed prosodic and discourse annotation*

At the second stage, the boundaries of the EDUs and other higher-level segmentation units are revised (see Section 3), and, if necessary, corrections are introduced to the set of the lower-level segmentation units as well as their time boundaries. After that, a detailed prosodic and discourse annotation of the EDUs and of the words included in these units is carried out (see Section 4). Collateral vocal acts (such as laughter, smiling or a creaky voice) that overlap with the speech are also detected (see Section 5).

As has been noted earlier, the scheme in question is relevant for situations when the annotation work is conducted by annotators of various levels of expertise. However, if this process takes place under different circumstances, then the first and second stages can be realized simultaneously. Meanwhile, the result of the first stage (annotation of all lower-level segmentation units, their time boundaries and preliminary EDU segmentation) can already be viewed as an incomplete but inherently consistent annotation.

## 1.3. Individual annotations: Final format

Final annotation that reflects the vocal behavior of all three session participants is saved in two formats.

(i) Information concerning the time boundaries of each unit is stored in *textgrid* files that can be created and edited using Praat software.

(ii) Detailed prosodic and discourse annotation is created and edited in .doc format using MS Word. It is also possible to use any other freely available word processor. However, it is specifically MS Word where the best results can be achieved.

Below you will find a screenshot of a textgrid that corresponds to an annotated fragment of the Reteller's vocal behavior in Session #22.

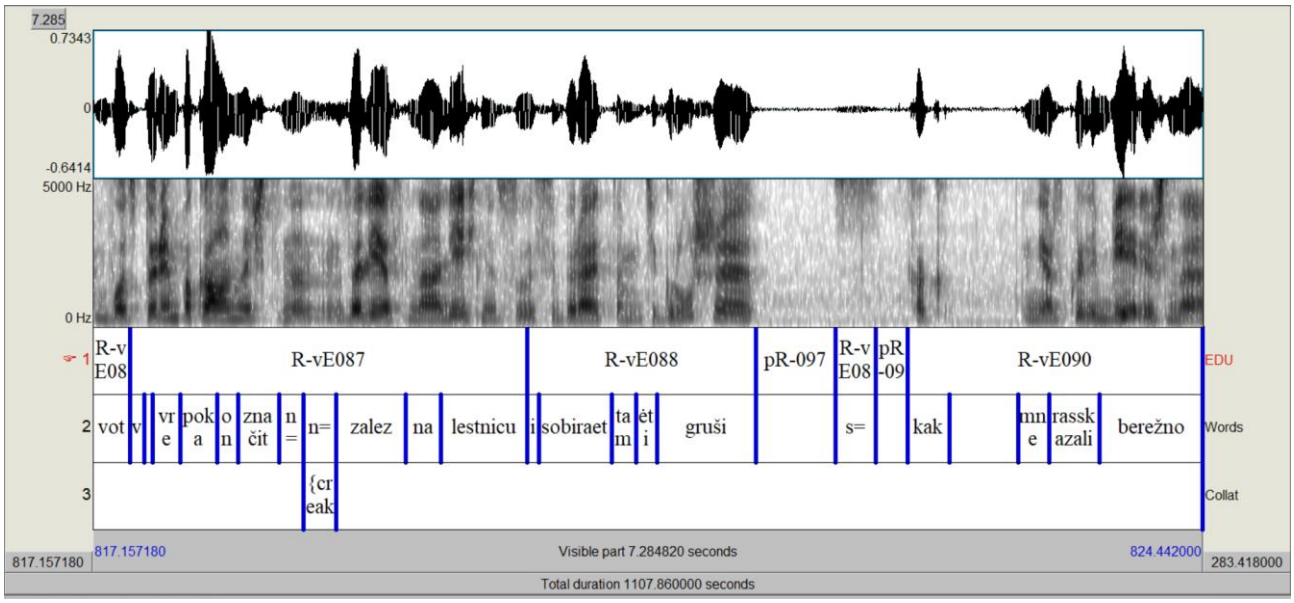


*Figure 1.* Sample of a textgrid (fragment of Session #22, Reteller)

Information in the textgrid is divided into three tiers:

(i) The "EDU" tier includes information about the boundaries of the EDUs and other higher-level segmentation units. Silent pauses between the units are also annotated here using special numbering. You will find more detailed information concerning the annotation of these units in Section 3.

(ii) The “Words” tier includes information about the boundaries of words, filled pauses and other lower-level segmentation units (see Section 2).

(iii) The “Collat” tier provides information concerning non-verbal acts which accompany the speech (see Section 5).

In the vocal transcripts, information about the units marked in the “EDU” and “Words” tiers is represented in tables. Appendix 2 provides a fragment of the individual transcript of the Reteller in Session #22; the start of the fragment corresponds to the leftmost point of Figure 1. Each line of the table corresponds to a higher-level segmentation unit, i.e. to an interval of the “EDU” tier in the textgrid. For each line, the following information is indicated in separate columns:

(i) The start time of the corresponding interval with a precision of 10 ms (“Time” column).

(ii) The higher-level segmentation unit code number (“EDU” column).

(iii) Transcription of the lower-level segmentation units which are a part of the given higher-level unit (“Transcription” column; for more detailed information see Section 4 and Appendix 1).

(iv) Comments written in free form (“Comments” column).

Symbols used in the transcripts are discussed in Sections 2 to 4 and then summarized in Appendix 1.

### 1.4. Technical specifics

As one can notice in Figure 1 and Appendix 2, specific numbering is implemented for the higher-level units both in textgrids and in transcripts (for more detail see Section 3.3). Moreover, for the sake of consistency, it is highly important to make sure that information in the textgrids completely corresponds to that in the transcripts. In order to provide a quick and reliable solution to this task, as well as to several other tasks within the frame of the given project, supplementary software products have been developed, namely (see also Sections 3.3 and 3.4):

(i) a Praat script and a VBA macro which perform semiautomatic numbering of the higher-level units in the textgrids and in the vocal transcripts correspondingly;

(ii) a Praat script that checks if the interval boundaries in the “EDU” tier corresponds to the interval boundaries in the “Words” tier; and

(iii) a VBA macro which checks the consistency of the data in the textgrids and in the vocal transcripts; it also exports data on unit time boundaries to the vocal transcript.

Moreover, automation tools are used for incorporation of the vocal annotation into a multichannel annotation in ELAN (see Section 6) and for creating scores transcripts (see Section 7).

## 2. Lower level of segmentation: Basic annotation

In this section, we will describe the principles of the basic annotation for words and other lower-level segmentation units. We single out the following types of these units:

- words and their analogues;

- filled pauses;

- laughter and other non-verbal acts;

- pauses filled with loud inhalation; and

- silent pauses.

The rules of transcription of these units as well as principles of setting their time boundaries in Praat are given below.

## 2.1. Transcription conventions

The transcription conventions discussed below apply both to intervals of the "Word' tier in textgrids and to the content of the "Transcription" column in text transcripts. Caution: in the subsequent annotation stages, transcription of the units in text transcripts may vary; see Section 4.

### *2.1.1. Transcription of words*

**1.** Words are written down using standard spelling (but see also points 6 and 7). Rare exceptions like ščas ('now', instead of sejčas) are allowed, but in general they are not recommended. If the transcriber notices that a word is pronounced in a non-standard way, this fact should be noted in a separate comment. The same holds for non-standard pronunciation, as in the example below:

on pytaetsja ix peresčitat' //pronounced as "peresčiDat' "

'he tries to count them'

For confirmation signals like ugu 'yep' or aga 'uh-huh', a conventionalized method is used. If the speaker pronounces these signals not in the way they are typically pronounced but somewhat 'more vocalized' (so that they seem to consist of the sounds they are usually written down with), then it seems reasonable to comment on this:

aga //non-reduced pronunciation: [aga]

**2.** The letter Ë should by all means be used in all cases which are allowed by the standard spelling:

berët '(s)he takes'; svoëm '[in] his / her'; neë '[at] her'; etc.

**3.** All word forms are written down the way they are pronounced, even when the 'wrong' cases/numbers/tenses, etc. are used. No adjustments toward 'the way it should be' are allowed for annotators. For example (here a red-colored font is used for clarification purposes only, and is NOT a transcription convention!):

i on (ə) vtoruju točnee dopolnjaet,

d= (ə) ot= ostavšixsja grušami,

'and he fills the second one [basket] with the rest of the pears'

**4.** All repetitions, including partial ones, should be written down:

na na svoj ba= <kakoj-to> bagažnik?

'on... on his bi... a bike rack?'

to est' pere= peresčityvaet snačala

'that is he re... recounts them from the start'

**5.** All truncated fragments of words should be written down. Such fragments are finished with an equal sign without a space:

on on on kakoj'to pož= on požiloj dopustim?

'he ... he ...he is somewhat ol... say, is he old?'

If the annotator is quite sure which word the speaker was planning to say, then the fragment that got cut off should be written down using traditional spelling:

poètomu kažetsja snačala,

č= čt= čto eë tam net,

'that is why at first it seems th... th... that she is not there'

However, if the annotator is not exactly sure what got truncated, then the fragment should be written the way they hear it:

Jura pravil'no skazal š= pro svist mal'čikov,

'Yura was right about the boys whistling'

See also the transcription of truncated words in the lines R-vE088, R-vE090, and R-vE098 of the transcript in Appendix 2.

**6.** If a word is hyphenated in its standard spelling but each part of the word is stressed separately, these parts are considered *separate words* for the annotation, and they also occupy separate intervals in the "Words" tier of a textgrid. In the transcripts, a space is inserted between these parts, while hyphens are preserved:

Mužčina

gde-to let soroka- pjatidesjati

'A man, about forty or fifty years old'

**7.** If, on the contrary, a sequence of orthographic words is pronounced in such a contracted manner that it makes no sense to set a boundary between them, then the whole sequence is *considered one word* and is spelled with an underscore.

v_smysle da,

'I mean yes'

**8.** If a word is pronounced so unclearly that the annotator is not sure whether it is in fact that word, then that unit is enclosed in <angle brackets>:

čto vot Jura <vot> kak raz rasskazal,

'what Yura has just said'

If the annotator is not sure about a sequence of words, then angle brackets are used for each of the words of the sequence:

<vsë> <tam>.

**9.** If the annotator can clearly hear the segmental structure of a word but cannot interpret it, then this unit is enclosed in the >inverted angle brackets<:

a potom >pst< uvidela košku,

'and then >pst< I saw a cat'

The same symbols are used for those cases where non-standard pronunciation can be interpreted, but it is clearly random:

a èti troe rebjat pošli v_storonu >fervela<. //Apparently, it means "fermera", 'farmer'.

'and these three guys went towards a >farver<'

**10.** For completely indiscernible fragments, the label <UNCLEAR> is used:

potomu čto on kak-to pytaetsja za= sest' na ètot <UNCLEAR>

'because he is trying to mount... to get on this <unclear>'

### *2.1.2. Transcription of onomatopoeias*

We consider onomatopoeias as a subclass of words. They are written down between hash signs the way one hears them:

mal'čik edet na velosipede,

#ty-dyš# # ty-dyš # # ty-dyš # # ty-dyš #!

'the boy is riding his bike like "ty-dyš" '

If an onomatopoeia is difficult to transcribe as it is heard, then a verbal description appears between hash signs:

#roars#

#bleats#

Hash signs are also used for transcribing non-vocal acts that have a function similar to that of the lexical unit but are not typically expressed orthographically:

i on takoj #whistle#

'and he's like...'

### *2.1.3. Transcription of filled pauses*

Filled pauses are vocalized fragments a speaker uses to fill a period of hesitation. There is no clear boundary between these units and regular words; see, for instance, Clark & Fox Tree (2002), where similar elements are treated as lexical items. In any case, filled pauses constitute a separate class; however small it may be, it is quite frequent in natural discourse. Each speaker fills their hesitation time in their own way, although four rough types can be singled out:

| **Pause type** | **Symbols used in transcripts** | **Unicode** |
|---|---|---|
| *um*-like filled pause | (ɯ) | 026F |
| *uh*-like filled pause | (ə) | 0259 |
| *ah*-like filled pause | (ɐ) | 0250 |
| pause filled with glottal creak | (ˀ) | 02C0 |

*Table 1.* Symbols used for filled pauses

In addition, various combinations are possible: (ɯə), (əɯ), etc. In these cases, one sound merges with another, and there is no gap and / or interruption between different parts.

*Note*. In the final transcripts, symbols for filled pauses are accompanied by an indication of their duration: (ɯ 0.23), (ɐɯ 0.40), etc. This is a result of the automatic export of data from the textgrids. The same applies to laughter and other non-speech vocal acts, silent pauses and pauses filled with loud inhalations.

### *2.1.4. Annotation of laughter and other non-verbal acts*

In addition to words and filled pauses, participants can perform a variety of non-speech vocal actions: clicking the tongue, smacking lips, etc. All these phenomena are recorded in {curly brackets} and are encoded as follows:

| **Non-verbal act** | **Transcribing convention** |
|---|---|
| expectoration | {exp} |
| clicking of the tongue | {cl} |
| smacking lips | {sm} |
| sigh | {sg} |
| gulp | {gp} |

| | |
|---|---|
| whistle | {wh} |
| snorting | {st} |
| coughing | {cg} |
| sniffing | {sf} |
| hemming | {hm} |
| laughter | {laugh} |

*Table 2.* Transcribing conventions for non-verbal acts

Laughter stands out among other the non-verbal acts, since it can be present as a separate segment or overlap with the speech. In the former case, the laughter is recorded as indicated in Table 2. In the latter, it is annotated in the "Collat" tier of the textgrid and is not marked in the transcript (see Section 5).

### *2.1.5. Annotation of silent pauses and pauses filled with a loud inhalation*

Silent pauses, i.e. the absence of any vocalization of a given speaker, are detected as gaps between vocalized units (but see Section 2.2 below for more detail). In textgrids, silent pauses correspond to empty intervals in the "Words" tier (for example, see the empty interval under the unit pR-021 of the "EDU" tier in Figure 1), while in transcripts they are designated with empty parentheses: ().

One should bear in mind that when annotating the vocal behavior of a *particular speaker*, a silent pause is understood as a segment of silence for this speaker; the other participants can still be pronouncing their utterances at this moment. Periods of *shared silent pauses* are further detected when compiling a scores transcript, as described later in Section 7.

Formally and functionally similar to silent pauses are pauses filled with loud inhalations. Inhalations are to be distinguished from sighs: when *sighing*, the participant takes in the air loudly and then immediately exhales loudly as well; a loud *inhalation* is typically followed by speech or a simple exhalation. An inhalation often indicates the participant's intention to take a turn in conversation. Pauses that contain loud inhalations are marked as (ɥ; Unicode number 0265).

## 2.2. Setting time boundaries

For every interval in the "Words tier", the exact times of its beginning and ending should be established. When setting boundaries, annotators must rely on auditory perception and on acoustic representations of the audio flow visualized in the Praat program via:

- waveform;

- spectrogram;

- pitch graph; and

- intensity graph.

Auditory perception and spectrograms are the most reliable instruments, but none of the above-mentioned means can serve as an absolutely accurate indicator of boundaries. Nevertheless, a combination of several parameters usually allows the annotator to determine the place of a boundary with sufficient accuracy.

In Figure 2 below, you will find a fragment of the textgrid for the Narrator's individual audio file of Session #22. Among other acoustic properties, you may notice the following.

(i) Virtually flat sections of the waveform, the absence of dense areas on the spectrogram, and interruption of the F0 curve correspond to silent pauses (empty intervals in the “Words” tier).

(ii) Minimum waveform oscillations, interruption of the F0 curve, and a characteristic spectrum correspond to a pause filled with a loud inhalation (interval (ɥ) in the “Words” tier).

(iii) A characteristic intermittent waveform and traces of stops on the spectrogram correspond to a pause filled with a glottal creak (ʔ).

(iv) Vowel-like waveforms and spectrograms correspond to an *uh*-like filled pause (ə), as well as a level pitch movement at modal F0.

(v) For most words, the spectrogram shows a clear formant structure of vowels and sonorant consonants and characteristic symptoms of plosive and fricative consonants.

(vi) For vowels and sonorant segments, an F0 curve is created.

In a standard case, the formal characteristics observed in the graphs are confirmed perceptually.

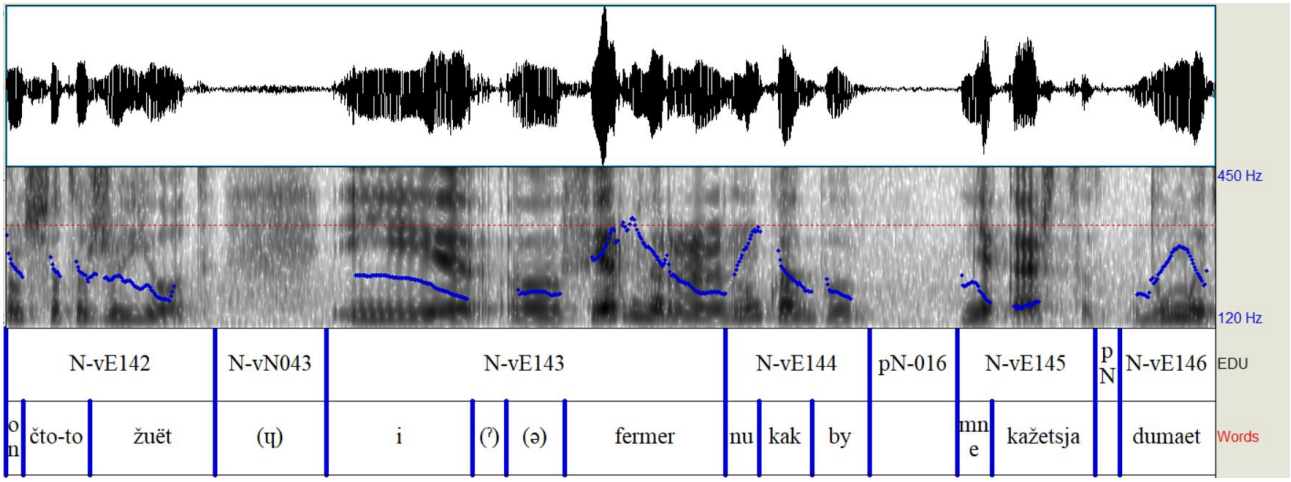


*Figure 2.* Fragment of the textgrid for Session # 22 (Narrator); time boundaries of lower-level units are set in the “Words” tier.

### *2.2.1. Important remarks*

**1.** Since one of the basic tasks of the project is to describe the coordination of different communication channels, time boundaries must be established with maximum accuracy. The best-case scenario is when two or more annotators are involved in marking the boundaries for one audio file, and the negligible difference between their annotation variants should not exceed 5 milliseconds.

**2.** When checking boundaries by ear, it is practically of no use to listen to very short intervals. Even pauses contain some noise, and microfragments of sounds shorter than a certain duration can no longer be perceived as phonemes by our hearing. Therefore, one should listen to longer fragments before and after the boundary in question, and then turn to the fragment where this boundary appears more or less in the middle, looking for a place where the quality of the sound changes considerably.

**3.** When working with a spectrogram, one should pay attention not only to the formant structure but also to such signs as, for example, the opening of the vocal cords at the beginning of speaking. Quite often, this allows annotators not to miss the beginning of the pronunciation of a unit.

**4.** At the beginning and end of words, various stops, vowel glide sounds and aspirations can occur. These phenomena are additionally annotated in the text transcripts (see Section 4.7.6), but are not marked in the textgrid. However, they must be included within the boundaries of the corresponding intervals of the “Words” tier.

**5.** Silent pauses between words and other lower-level segmentation units can theoretically be of any duration. However, the shorter the interval, the more difficult it becomes to distinguish between a pause and a voiceless stop or other similar phenomena. Therefore, we propose the following rules when setting the time boundaries of silent pauses.

(i) Pauses with a duration exceeding 0.1 second should be marked at all times.

(ii) Pauses with a duration below 0.1 second are marked only under the following circumstances: (a) the pause is detected as such by ear; and (b) to the left and to the right of the pause there are sounds the boundaries of which are easily and unmistakably defined.

(iii) If a potential pause is shorter that 0.1 second, it is difficult to detect it by ear and it is surrounded by voiceless consonants on the left and on the right, or if there are other phenomena hindering the definition of its boundaries (for instance, other speakers' activity, background noise, etc.), then the pause is not marked. In this case, a word boundary is placed at the extreme left point defined using a spectrogram or by ear.

#### *2.2.2. Automation of setting time boundaries*

There are ways to automatically detect the time boundaries of words and other lower-level units. In our project, we rely on routines developed by our partners from STEL (http://speech.stel.ru/). Basic transcripts are used as input, and the output is a text file with automatically detected time codes for each transcribed item. Then, with the help of an additional script, this text file is further transformed into a textgrid of the required format.

Partial automation of establishing time boundaries significantly increases productivity, yet its results should be checked manually. Various factors can cause inaccuracies in automatic detection, among them:

- non-trivial ways of filling a pause;

- reduced and / or quiet pronunciation; or

- interference from other speakers and background noise.

When checking the results of the automatic procedure, one should follow the principles described above in this section.

## 3. Higher level of segmentation: Basic annotation

### 3.1. Elementary discourse units

The *elementary discourse unit* (EDU) is the central concept of our approach to analyzing spoken discourse. An EDU is a basic unit of the higher segmentation level; each EDU constitutes a minimal step in discourse production. In most canonical cases, the elementary nature of EDUs is consistently manifested at different levels of speech production such as the physiological, prosodic, syntactic, semantic and cognitive levels (see Chafe, 1994; Kibrik & Podlesskaya, 2009; Kibrik et al. in print).

When segmenting the speech flow into EDUs, annotators should primarily rely on a set of prosodic criteria:

- loudness pattern (the beginning of an EDU is usually pronounced louder than the end);

- tempo pattern (the beginning of an EDU is often pronounced faster than the end); and

- holistic intonation contour and accents.

The last criterion is the most significant. In most cases, EDU boundaries can also be related to communicative prosodic constituents (see Korotaev, 2015).

Each EDU is marked as a separate interval in the "EDU" tier in textgrids and as a separate line in the vocal text transcripts (see Appendix 2). Although EDUs are singled out on prosodic grounds,

they tend to correlate with simple clauses. Thus, in Appendix 2, out of fifteen EDUs, nine correspond to this syntactic format, and only six (R-vE087, R-vE090, R-vE092, R-vE096, R-vE099, R-vE101) are not clauses (see Section 3.3 for the principles of numbering).

## 3.2. EDUs and other higher-level units

An EDU can include any lower-level units: words, silent and filled pauses (see lines R-vE093, R-vE094 in Appendix 2), and non-vocal acts (see R-vE095). Apart from EDUs, we also distinguish other higher-level segmentation units that have a much simpler internal structure. Such units may consist of isolated (i.e. not included in any EDU) filled pauses or their clusters; pauses filled with loud inhalations; isolated laughter; or other non-speech vocal acts. All such units are also formatted as separate lines in the text transcripts and as separate intervals in the "EDU" tier of textgrids.

The following conventions are used when drawing boundaries between EDUs and other higher-level segmentation units.

**1.** Laughter that is not located inside any EDU (i.e. not surrounded *on both sides* by elements belonging to the same EDU) constitutes a separate higher-level segmentation unit (see lines R-vL024, R-vL025 in Appendix 2).

**2.** Each continuous sequence of other non-speech vocal acts and / or pauses filled with loud inhalations that are not *included* in any EDU also constitutes a separate higher-level segmentation unit (see R-vN020, R-vN021, R-vN022, R-vN023).

**3.** An EDU may begin with a filled pause or a group of filled pauses if the first "real" word of this EDU immediately follows the last pause (see R-vE099). If a filled pause is separated from the first word of an EDU by a silent pause or some other element, then this filled pause constitutes a separate higher-level segmentation unit (see R-vF005, R-vF006).

**4.** All silent pauses located between higher-level units are also recorded in separate lines in the transcripts (see pR-098 — pR-104) and as non-empty intervals of the "EDU" tier in the textgrids. See also Section 7 on scores transcripts, where further evidence is provided on why it is important to mark silent pauses in separate lines.

## 3.3. Boundaries and numbering

"EDU" tiers in the textgrids should be divided into non-empty intervals *without remainder*. It is necessary to make sure that the boundaries of each higher-level unit fully coincide with the boundaries of some lower-level units. In other words, it is not allowed for an interval of the "Words" tier to have a non-empty intersection with more than one interval of the "EDU" tier.

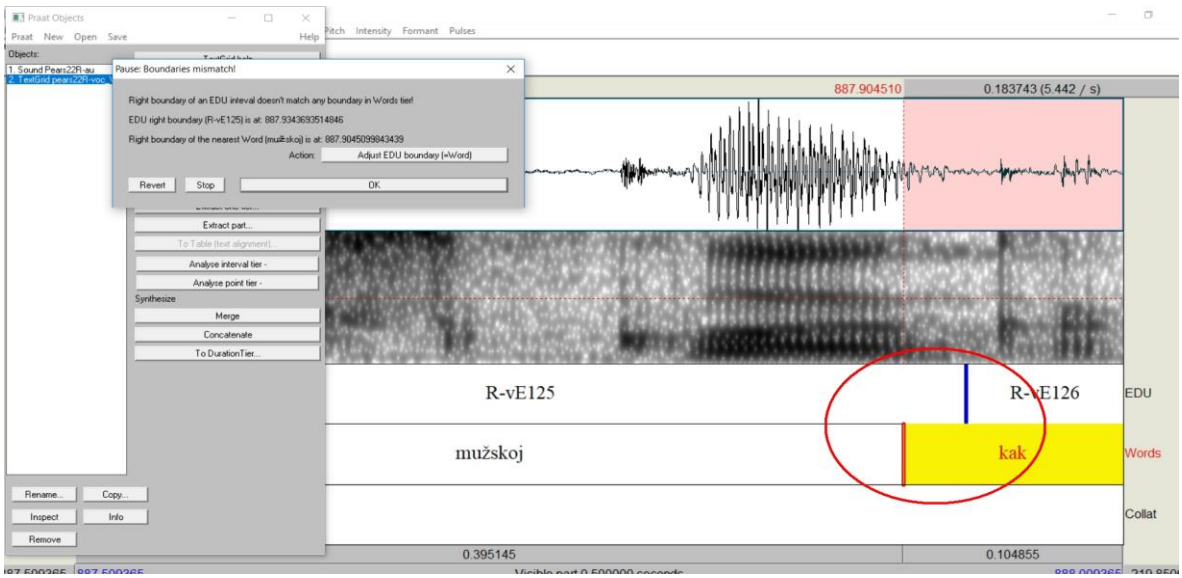


*Figure 3.* Example of an error detected by a script that checks interval boundaries in a textgrid.

Figure 3 above shows a screenshot where one can see that the right boundary of EDU R-vE125 does not coincide with any of the boundaries of the words mužskoj (‘masculine’) and kak (‘like’). This situation is unacceptable and must be corrected when the annotation is checked against the technical requirements (see Section 1.4).

All higher-level segmentation units are uniformly numbered in textgrids and in transcripts. Each code number consists of:

- the speaker code (N for the Narrator, C for the Commentator, R for the Reteller);

- the vocal channel code (-v);

- a symbol denoting the unit type; and

- a three-digit number.

The numbering is continuous and uninterrupted within each type of unit. Silent pauses located between higher-level units also undergo continuous uninterrupted numbering. In the transcripts, code numbers of different unit types are additionally distinguished by formatting. The specific conventions used for numbering are given in Table 3.

<table>
<tr><th>Unit type</th><th>Prefix</th><th>Type symbol</th><th>Number format</th><th>Example</th><th>Additional formatting in the transcript</th></tr>
<tr><td>EDU</td><td rowspan="4">N/C/R-v</td><td>E</td><td rowspan="5">000</td><td>N-vE123</td><td></td></tr>
<tr><td>Filled pause / a cluster thereof</td><td>F</td><td>C-vF056</td><td>Reduced font size</td></tr>
<tr><td>Laughter</td><td>L</td><td>R-vL002</td><td>Reduced font size, italics</td></tr>
<tr><td>A cluster of other non-speech vocal phenomena and / or pauses filled with loud inhalations</td><td>N</td><td>R-vN096</td><td rowspan="2">Reduced font size, grey color</td></tr>
<tr><td>Silent pause between higher-level segmentation units</td><td colspan="2">pN/C/R-</td><td>pC-201</td></tr>
</table>

*Table 3.* Numbering of the higher-level segmentation units and silent pauses between them.

## 4. Annotating properties of EDUs and words

Apart from segmenting the participants’ vocal behavior into units of higher and lower levels, vocal annotators should register additional properties of EDUs and words. These are properties that are closely connected to the way a local discourse structure is produced and perceived in real communication. This extended annotation is only marked in the text transcripts and is carried out using additional symbols; in some cases, the basic mark-up discussed above in Sections 2 and 3 gets modified in the course of this process.

A full list of transcribing conventions used in the text transcripts is given in Appendix 1. In what follows, a brief explanation of the most important phenomena is provided. More elaborated treatment can be found in Kibrik & Podlesskaya (2009).

## 4.1. Accents and pitch movements

Words that bear discourse *accents* are prosodically prominent. Functionally, accents serve primarily (i) to identify the central element of a communicative constituent and (ii) to establish the way this constituent relates to the external context. *Pitch movements* associated with accentuated words are mostly responsible for the second function (see Kodzasov, 2009; Yanko, 2008). In the transcripts, accents and the nature of pitch movements on stressed syllables are denoted jointly by means of iconic slashes: / before a word stands for a rising pitch accent, \ stands for a falling pitch accent, and − stands for a level pitch accent. For example, the word /kozoj '(with) a goat' in line R-vE093 of Appendix 2 is pronounced with a rising accent; the word \ljubov′ju '(with) love' in line R-vE092 is pronounced with a falling accent; the word –lestnicu 'ladder' in R-vE088 is pronounced with a level accent. If a complex pitch movement is realized on the stressed syllable, combinations of slash symbols are used (see examples in the scores excerpt from Appendix 3: rising-falling accents in /\net 'no' in EDU N-vE262 and /\devočki 'girls' in C-vE167; rising-level accent on the word mal′čika 'boy's' in R-vE029).

In some cases, pitch movements outside the stressed syllable are also important; such movements are indicated by arrows (↑, ↓, →) located before or after the slash. Thus, to interpret the accent on the word /↑mimo 'past' (line R-vE097 in Appendix 2), not only the rising pitch movement on the stressed syllable (mi-) should be taken into account, but also the continuation of this movement in the post-stressed syllable (-mo).

In each EDU, there is usually one *primary accent* that takes most responsibility for incorporating the EDU into the local discourse structure (see also below, Section 4.2). In the transcripts, primary accents are additionally indicated by underlining the stressed vowels of the words that bear such accents. EDUs may also contain *secondary accents*; see line R-vE098, where the falling primary accent on the word \lestnice 'ladder' is preceded by two secondary rising accents on the words /vremja 'time' and /sadovnik 'gardener'. In this case, the opposition of the primary and secondary accents is rooted in their communicative functions: the primary accent marks the rheme (or, focus) of the utterance, while the secondary accents are placed on thematic (or, topical) elements.

In some cases, more than one primary accent can be found in an EDU (see line R-vE095). This typically occurs when one of the two primary accents mark the the rhematic component, and the other indicates the phase value of incompleteness (see Section 4.2).

## 4.2. Illocutionary / phase value of an EDU

Each successful (completed) EDU expresses a certain set of values associated with:

(i) the illocutionary force of the utterance;

(ii) the phase opposition of completeness vs. incompleteness; and

(iii) additional meanings modifying (i) and / or (ii).

To indicate the *illocutionary / phase EDU values* in transcripts, punctuation marks are used at the end of lines. We distinguish between marks that (a) indicate the illocutionary force, (b) express the type of discourse incompleteness, and (c) indicate the type of modifying meaning.

**A.** Illocutionary marks

The completion of a *statement* illocution is marked with a period (.): see lines R-vE091, R-vE092 and R-vE098 in Appendix 2. In statement-final EDUs, the primary accent usually bears a deep falling pitch movement.

General and special *questions* are coded with a final question mark (?). A general question is typically characterized by a rising primary accent. An inverse question mark (¿) denotes a *semi-statement* — an illocution that demonstrates formal features of a statement but functions as a

request to confirm an expressed assumption (see Korotaev, 2018). Semi-statements are represented in EDUs R-vE025 and R-vE030 of the scores fragment in Appendix 3.

*Directives* (illocutions that contain a certain call for action) are denoted by an inverted exclamation mark (¡); *vocatives* are designated with a @ sign.

**B.** Incompleteness

A default way of expressing incompleteness is encoded with a comma (,). Incompleteness is quite often expressed by a rising primary accent (see lines R-vE093, R-vE094, R-vE095 (second accent), R-vE097 and R-vE100 in Appendix 2). However, this is not the only way; e.g., incompleteness is regularly accompanied by a non-final falling primary accent (see R-vE087).

Incompleteness can be combined with *inexhaustiveness* — a value associated with the speaker's inability to "perform an unambiguous and discrete illocutionary act of a statement" (Kibrik & Podlesskaya, 2009, p. 168). In this case, the "triple commas" sign (,,,) is used in the transcripts (see lines R-vE089, R-vE101). Incompleteness can also be complicated by *elucidation*, for example, before direct quotations (see Section 4.6). In such cases, a colon (:) is used in the transcripts.

**C.** Modifying meanings

Inexhaustiveness, as mentioned above, is one of the meanings that can modify illocutionary / phase values. If it is not combined with incompleteness, dots (…) are used to denote it. *Exclamation* is another important modifying value; it is denoted by an exclamation mark (!), as in N-vE262 in the scores transcript of Appendix 3.

The phenomena discussed above can co-occur in one EDU: for example, a directive illocution can be combined with incompleteness (¡,), a question with an exclamation (?!), etc. Several technical rules apply in these cases:

(i) When an incompleteness mark is combined with illocutionary or modifying marks, the former should be placed at the end (!,; ¿,,, etc.).

(ii) When combined with other marks, the period is omitted. For example: !, but not!..

If an EDU ends with an illocutionary punctuation mark or a combination of an illocutionary mark with a modifier sign, this EDU is said to be *sentence-final*. As in standard written punctuation, the beginning of each new sentence is marked with a capital letter of the first word (see, for instance, EDU R-vE099 in Appendix 2).

## 4.3. Speech disfluencies

In the transcripts, we annotate disfluencies that have an obvious *interruption point*, i.e. a signal of discourse disruption. Word truncation is one such signal (see Section 2.1.1, point 5).

An interruption point is typically followed by the speaker's attempt to revise the failed wording. Most frequently, this results in a *self-repair*. Two types of self-repairs are distinguished in the transcripts:

(i) A mild repair is a self-repair that does not lead to abandoning the current EDU. The || symbol is used for mild self-repairs (see multiple mild repairs in lines R-vE088 and R-vE098 in Appendix 2). By definition, the sign of a mild repair can only be placed inside an EDU.

(ii) A severe repair is a self-repair that results in the speaker abandoning the current EDU and starting a new one. Severe self-repairs are indicated by a double equal sign (==) at the end of the line (see R-vE090).

For more information on the criteria for distinguishing between mild and severe repairs, see Kibrik & Podlesskaya (2009, pp. 187-208); a classification of self-repairs can be found in Podlesskaya (2015).

In conversation, repairs can be *induced externally*: interlocutors can interrupt each other, provoke reactions by non-verbal means, etc. When there are reasonable grounds to believe that a disfluency was caused by external factors, the ⁞⁞ symbol (two Unicode 2E3E characters) is used for a mild repair and the ≈≈ symbol (two Unicode 2248 characters) for a severe repair; see EDUs R-vE024, R-vE026, C-vE164 and C-vE168 in the scores fragment of Appendix 3).

EDUs that are not completed because of a severe self-repair (i.e. considered unsatisfactory by the speaker) should be distinguished from those that are intentionally incomplete. In Example (1) below, the speaker deliberately leaves the line C-vE139 incomplete, believing that the already produced words suffice to express her idea ('it is unlikely that the film's characters could wear jeans had they lived in the Crimea'). Examples of this type are interpreted as instances of *aposiopesis* and are marked by a tilde (~).

(1) pears23: C-vE139

| | | |
|---|---|---|
| 441.66 | C-vE137 | Ja eščë /podumala,<br>I also thought |
| 442.12 | C-vE138 | čto točno ne /\Krym,<br>that this is definitely not the Crimea |
| 442.95 | C-vE139 | potomu čto /–džinsov (0.09) –nu ~<br>because the jeans well ... |

## 4.4. Co-construction

Participants of a conversation regularly resort to *co-construction*, when the first part of an utterance is produced by one speaker and the final part by another (see Helsavuo, 2004; Grenoble, 2008). In the transcripts, co-construction is designated with the percent sign (%) placed at transition places, i.e. at the end and at the beginning of the involved EDUs. For example, in the fragment presented in Appendix 3, the construction started by the Reteller in EDU R-vE027 is completed by the Narrator in N-vE264. In this excerpt, the first participant (Reteller) immediately stopped speaking when her interlocutor intervened to complete the utterance. However, this is not always the case: the first speaker may decide to continue despite the intervention. See Example (2) below, where the Narrator chooses to complete the 'basket' construction begun by the Commentator, but the latter does not give up his turn, which results in an overlap. In such cases, the co-construction mark (%) is only put at the beginning of the contribution by the second speaker. See Section 7 for details on scores transcripts.

(2) pears22: C-vE146 – N-vE204

| | | | | | |
|---|---|---|---|---|---|
| 446.30 | 447.84 | | | C-vE145 | On /–vzjal samuju (0.07) \pervuju /korzinu-u,<br>He took the closest basket |
| 447.84 | | | | C-vE146 | kotoraja byla s= ‖ (0.18) \polnaja.<br>which was the mo... fullest |
| 448.49 | | N-vE204 | % bliže \vsego k nim-m.<br>the closest to them | | |
| | 449.06 | | | | |
| | 449.61 | | | | |

## 4.5. Insets

An *inset* is an EDU or a group of EDUs wherein the speaker temporarily deviates from the mainline. Usually, insets are wedged in between units that are closely interrelated both formally and semantically. Sometimes, an inset can even be mentally removed without spoiling the local coherence. Insets often contain clarifications to the content of the mainline. The beginning and end of the inset are marked with brackets. Consider lines R-vE091 and R-vE092 in Appendix 2; information about how exactly one of the film's characters collects pears appears as an inset between the subordinate (lines R-vE088 to R -vE089) and the main (starting with the line R-vE093) clauses of a complex construction.

If an inset appears inside an EDU, then this is additionally marked with an em-dash (—) at the end of the first and beginning of the second part of the interrupted EDU. We call this phenomenon a *split*; see the following example:

(3) pears04: C-vE014 — C-vE016

| | | |
|---|---|---|
| 359.13 | C-vE014 | Kstati na vot ètom /mal'čike —<br>By the way, this boy ... |
| 360.63 | C-vE015 | (kotoryj na \velike ezdil',)<br>who was riding the bike |
| 362.05 | C-vE016 | — 'i na-a /djad'ke /fermere byli platki \↑odinakovye!<br>... and the farmer dude wore identical neckscarves |

Note that parts of the truncated EDU are annotated as separate lines, and each of them has its own standard code number.

Sometimes, the speaker begins a discourse stretch as an inset, but then gradually transforms it in such a way that it becomes mainline. In the transcripts, such situations are indicated by a combination of characters (* at the beginning of the line that opens a "one-sided" inset.

## 4.6. Quotation

Since the characters of the Pear film that served as stimulus material for the corpus do not speak to one another, instances of *reported speech* are considerably less frequent here than in some other oral texts. Still, in our transcription system we use special conventions for quotations; see Kibrik & Podlesskaya (2009, pp. 288-309). Direct and semi-direct quotations are enclosed in the quotation marks; see Example (4) taken from the corpus "Funny Life Stories". The speaker in (4) uses a semi-direct quotation technique, as deictic and intonational characteristics of direct speech are combined with the presence of a conjunction čto 'that', which usually introduces an indirect quotation (see Podleskaya, 2018 for details).

(4) FS_28: 12-15

| | | |
|---|---|---|
| 18.74 | 10. | i /stojali,<br>and they stood |
| 19.29 | 11. | i \nyli:<br>and moaned |
| 19.84 | 12. | ( ) čto "Kak ne /–xo-očetsja,,,<br>like "We don't want to |
| 21.92 | 13. | kak /–zdo-orovo,,,<br>that would be great |

| | | |
|---|---|---|
| 23.05 | 14. | vot by /opozdat’,<br>if only we could be late |
| 24.12 | 15. | vot by zdes’ \ostat’sja.”<br>if only we could stay here” |

## 4.7. Other phenomena

### *4.7.1. Tempo*

When assessing speech tempo, we rely on the idea of a normal tempo for a given speaker, and then note cases of accelerated and decelerated pronunciation. Words are taken as the minimum units of tempo alternation. Accelerated pronunciation is marked with *italics*, and slowing down is marked with s p a c i n g. An example of accelerated tempo is presented in line R-vE091 of Appendix 2 (the words *mne rasskazali* ‘I was told’). It is no accident that this stretch is found inside an inset, as accelerated pronunciation is a typical prosodic property of such instances.

### *4.7.2. Reduction*

A significant phonetic *reduction* takes place when several phonemes are lost in one’s pronunciation although they are typically present when the word is pronounced in a neutral manner (see Kibrik & Podlesskaya, 2009, p. 350). We distinguish between a complete reduction affecting the whole word and a partial reduction affecting only a part of the word. In the transcripts, the reduction is marked with a grey font. See proxodit ‘passes by’ in line R-vE093 of Appendix 2.

### *4.7.3. Lengthening*

When a phoneme is *pronounced in a prolonged way*, the orthographic form of the word gets altered. The letters corresponding to the lengthened phonemes are duplicated using the symbol ‑ (Unicode 02D7). For iotified vowels, the following conventions are used: ja‑a, ë‑o (see eščë‑o ‘another’ in line R-vE093 of Appendix 2), e‑e, j‑ja, j‑ju, etc. Lengthening may be due to hesitation (see instances in lines R-vE090, R-vE095, R-vE098). For other functions of lengthening, see Kibrik & Podlesskaya (2009, pp. 344-349).

### *4.7.4. Emphasis*

When speakers express emotional attitudes toward the content of their utterances, they may resort to an *emphatic* pronunciation, i.e. to pronouncing (a group of) words with additional stress; see Kibrik & Podlesskaya (2009, pp. 353-354). In the transcripts, emphasis is marked with bold, see **berežno** ‘carefully’ in R-vE091.

### *4.7.5. Register*

Sometimes, speakers resort to pronouncing (groups of) words in a shifted F0 register. In these cases, the F0 curve temporarily shifts to the lower or higher level of the range that is standard for a given speaker or might even go beyond the standard range. In the transcripts, this phenomenon is noted with the help of a reduced point size, and for a lowered register (this modification occurs much more often than a heightened register), placing it below the baseline is used. Inset (see Section 4.5) is a typical context for a lowered F0 register; see the following example:

(5) pears04: C-vE230

| | | |
|---|---|---|
| 1050.95 | C-vE229 | A /potom uže —<br>and then ... |
| 1051.58 | C-vE230 | (v /sl̲edujuščij kadr,)<br>into the next shot |
| 1052.41 | C-vE231 | — on \ležit̲,<br>... he is lying on the ground |

### *4.7.6. Stops and other phenomena at the beginnings and ends of words*

Additional vocal phenomena such as aspirations, glide vowels and stops may accompany the beginning and / or the end of a word. Some of them are functionally loaded, while others instead reflect individual characteristics of the speakers; see Kibrik & Podlesskaya (2009). In the transcripts, the following are marked as superscripts:

- schwa glide vowel at the beginning or the end of a word (symbol ə; see ещё-оᵊ ‘another’ in EDU R-v093 of Appendix 2);

- glottal stop at the beginning or the end of a word (symbol ʔ; for example, ʔoni ’they’);

- labial stop at the end of a word (symbol ɯ; for example, idtiᵚ ‘to go’); and

- aspiration at the end of a word (symbol h; for example, gruši$^{h}$ ‘pears’).

### *4.7.7. Non-standard / variable stress*

In words with variable lexical stress (твороґ vs. твóрог ‘cottage cheese’), as well as in cases of non-standard stress (ранчó ‘ranch’) or homography resolved through stress (зáмок ‘castle’ vs. замóк ‘lock’), the stressed vowel is additionally marked with the Unicode symbol 0301.

### *4.7.8. Comments*

In the “Comments” column of the transcript, one can indicate additional properties of EDUs and / or their parts. These may relate to the way a unit is pronounced (for example, “quiet”, “whisper”, “loud”, etc.), provide an interpretation (see a motivation for having two primary accents in R-vE095, Appendix 2), refer to the general communicative context not directly derived from the vocal signal (for example, “addressed to the Commentator”), etc. The “Comments” column is optional and not subject to strict rules.

## 5. Collateral vocal phenomena

Non-speech vocal acts can overlap the speakers’ main vocal activity. Such phenomena are annotated *only in textgrids* in a separate “Collat” tier. We annotate three types of collateral phenomena:

- {laugh};

- {smile}; and

- {creaky} voice.

In general, the boundaries of collateral phenomena should be marked regardless of the boundaries in the “Words” tier. For example, if one can hear that only a part of a word is pronounced with a creaky voice, this must be properly marked in the “Collat” tier. See Figure 4, where it is clearly seen on the spectrogram that the creak is found at the end of the filled pause (ə) and at the beginning of the conjunction i ‘and’.

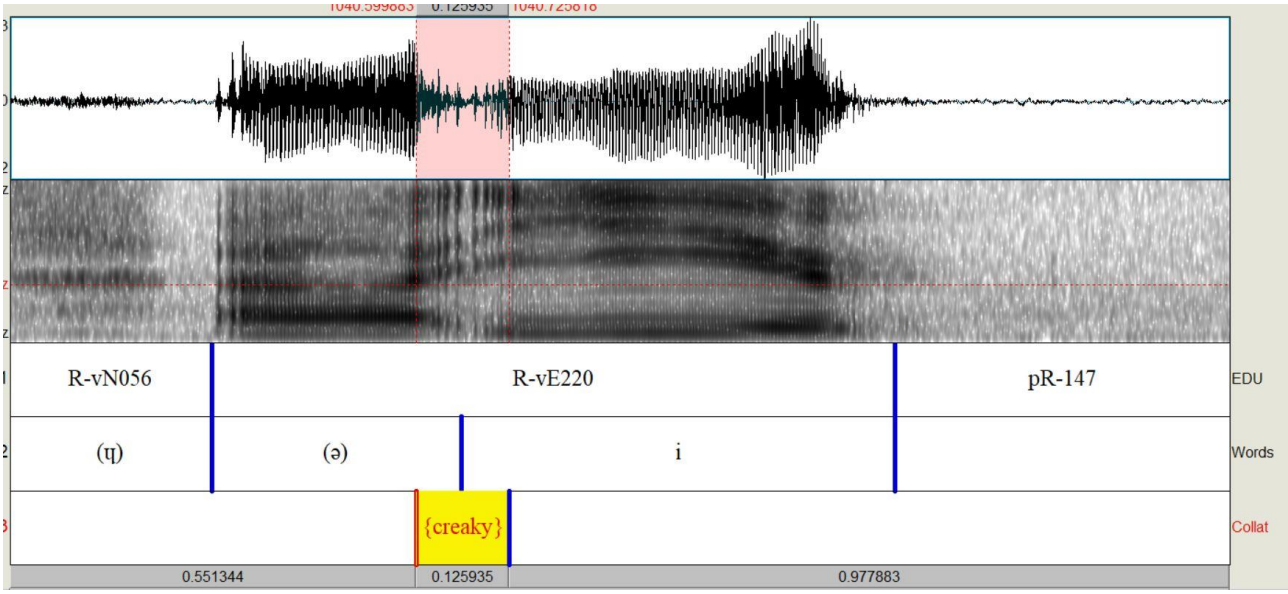


*Figure 4.* Independent interval boundaries in the "Words" and "Collat" tiers.

However, it is sometimes impossible to establish such independent boundaries. For example, it is not always easy to understand where a laugh overlaps with a word. In such cases, one can set the boundary of the laughter where the boundary of the word is.

## 6. Converting vocal annotation into ELAN

To empower an analysis of multichannel communication, the vocal annotation performed in Praat and Word software should be automatically converted to the ELAN format. The basic principles for converting data contained in textgrids and vocal transcripts into an ELAN structure are described below. For more information about the multichannel annotation scheme, see the "Multichannel annotation in ELAN" document (http://multidiscourse.ru/data/ann/pears_multichannel_annotation_en.pdf), which also contains a list of all tiers and sets of possible values for the intervals of each tier.

**1.** Higher-level segmentation units, with their time boundaries and code numbers, are transferred to the *-vLine tier (where * takes one of the values N, C and R depending on the role of the participant in the session); lower-level segmentation units (except for silent pauses) are transferred to the *-vSegm tier; and intervals corresponding to the collateral vocal phenomena are transferred to the *-vCollat tier.

**2.** Silent pauses are marked and additionally numbered in a separate tier *-vPause; for each pause, in the dependent tier *-vPauseInOutEDU, there is an indication whether it belongs to an EDU or is located between higher-level units.

**3.** In the *-vLineType and *-vSType tiers, the type is indicated for the higher- and lower-level segmentation units, respectively: EDU, word, filled pause, etc. *-vSForm and *-vCollatForm tiers contain "lexical" forms of lower-level units and collateral phenomena, respectively. The *-vLineVerbatim tier contains a word-by-word record of the lower-level units that compose the given higher-level unit.

**4.** The properties of words and EDUs, as discussed in Section 4, are converted from transcript marks into interval values on the dependent tiers. For each property, a separate tier is used, among them:

- *-vIllocPhase and *-vCombIllocPhase tiers for indication of the illocutionary / phase values of EDUs (see Section 4.2);

- *-vCoConstr tier for indication of the EDU's role in a co-construction (4.4);

- *-vParenth and *-vInSplit tiers for indication of the EDU's role in an inset and / or split constructions (4.5);

- *-vCitation tier for indication of the EDU's role in a construction with a (semi) direct quotation (4.6);

- *-vAccents and *-vMainAccent tiers for indication of pitch movements in accented words and the status of accents as primary or secondary (4.1);

- *-vInterrupt tier for indication of the type of disfluency associated with the given interruption point (4.3);

- *-vTempo and *-vRegister tiers for indication of deviations from the standard tempo (4.7.1) and standard register (4.7.5);

- *-vReduction and *-vEmph tiers for indication of reduction (4.7.2) and emphatic pronunciation (4.7.4);

- *-vLength and *-vStress tiers for indication of the prolonged realization of phonemes (4.7.3) and non-standard / atypical stress (4.7.7); and

- *-vStops tier for indication of the stop type / other phenomenon at the word boundary (4.7.6).

**5.** For the convenience of further calculations, more tiers are created in addition to the above-mentioned: in the intervals of these tiers, one finds values that are specially computed during the course of conversion. Additional tiers related to EDUs include:

- *-vWordsCount: this tier indicates the number of words in an EDU;

- *-vPausesCount: the number of silent pauses within an EDU;

- *-vFilledCount: the number of filled pauses within an EDU;

- *-vStartFilled: this tier indicates whether an EDU begins with a filled pause (see Section 3.2);

- *-vAccentsCount: the number of accented words in an EDU;

- *-vMainAccentsCount: the number of words produced with a primary accent;

- *-vMainAccents: this tier lists pitch movements realized in the primary accents of an EDU;

- *-vAccentsAfterMainCount: the number of secondary accents after the (last) primary accent in an EDU; and

- *-vInterruptCount: the number of interruption points in an EDU.

Computed properties of the lower-level segmentation units are reflected in the following tiers:

- *-vInOutEDU indicates whether a unit belongs to any EDU;

- *-vNearPause indicates whether a unit is adjacent to a silent pause;

- *-vWordNum indicates the position of a word from the beginning of its EDU; and

- *-vWordNumReversed indicates the position of a word from the end of its EDU.

## 7. Scores transcripts

The principles of vocal annotation described above are based on a system of discourse transcription which was developed on the basis of monologic spoken discourse. As indicated in Section 1.1, when annotating the "Pear Chats and Stories" corpus, we continue to transcribe the vocal behavior of each participant separately — even in those parts of the recording where full-fledged speech interaction among several participants is taking place. This approach is partly justified by the fact that such basic concepts as EDU, accents, illocutionary / phase values, disfluencies, etc., are relevant not only for monologues but also for dialogical situations. At the same time, one can single

out characteristics that significantly distinguish a genuine conversation from a (mostly) monologic production. At least two of them directly relate to the process of vocal annotation and to the ways of representing annotated data.

First, the status of *boundary silent pauses* varies significantly in monologues and in conversation. When transcribing monologic discourse, it is possible to attribute boundary pauses to subsequent EDUs (see Chafe, 1994; Kibrik & Podleskaya, 2009). In dialogues, however, intervals of silence cannot be attributed to particular EDUs (see Section 3.2), and, sometimes, even to particular speakers. In fact, it is important to distinguish situations when only one speaker is silent (such segments are marked in individual transcripts and textgrids) from *shared pauses*, when all active participants keep silence simultaneously. Shared pauses can play a role in turn-taking, preference organization in adjacent pairs, etc. (see Sacks et al., 1974 and other works on the Conversation Analysis).

Second, conversations are filled with *overlap*, i.e. simultaneous speaking of more than one participant. According to our data, at the conversation stage, overlaps take up about 15% of the time; moreover, less than half of all EDUs at this stage are pronounced without any overlaps. In general, the question of how participants coordinate their vocal actions on a shared timeline is one of the central issues for the analysis of dialogical data.

In order to visually display the properties and phenomena described so far, along with individual vocal transcripts, a single *scores transcript* is used for every annotated session. The scores transcript is automatically generated in Excel from individual vocal transcripts and textgrids (a special script was created to perform this procedure). A screenshot with a fragment of the scores transcript of Session #22 is presented in Appendix 3. Compilation of scores transcripts is based on the following principles:

**1.** The segmentation of the speech flow into lines corresponding to EDUs and other higher-level units is preserved along with their numbering (see Section 3). At the same time, the vocal contribution of each participant is recorded in separate columns marked with different colors: green for the Narrator, blue for the Commentator and dark red for the Reteller.

**2.** For each higher-level segmentation unit, the time of the start and the end of vocalization is indicated in the columns “TimeS” and “TimeE”, with a precision of 10 ms. The lines are sorted by the start time of vocalization, i.e. by the values in the “TimeS” column.

**3.** Shared pauses are written down in separate columns under the “Pauses” heading, to the left of the “TimeS” and “TimeE” columns. For each shared pause, its code number (vp001, etc.) and duration (with a precision of 10 ms) are provided. The arrangement of lines with shared pauses also follows the basic sorting rule.

**4.** Each graphic line of the scores transcript is numbered for the convenience of further reference (“Line #” column).

**5.** If a higher-level segmentation unit of one participant ends later than the nearest unit of the other participant begins (i.e. there is an overlap), then the end time n of the first unit is indicated in an additional graphic line. Thus, the values in the “TimeE” column are also sorted in ascending order; the cells in the transcript column are visually merged: see, for example, the graphic lines 0810-0813 (EDU C-vE163 of the Commentator) and 0838-0842 (EDU R-vE030 of the Reteller) in the fragment presented in Appendix 3.

**6.** Continuous periods of vocalization of one participant are additionally highlighted with color according to the principle described above in Paragraph 1). The same colors fill cells with the indication of the start / end times of vocalization in the “TimeS” and “TimeE” columns. This helps visually identify instances of overlaps. See, for example, graphic lines 0810-0811, where there is an overlap of the Commentator’s and the Narrator’s contributions; graphic lines 0813-0815 and 0817-0818, with overlaps of the Reteller’s and the Commentator’s contributions; and graphic lines

0819-0821, where all three participants speak simultaneously. In graphic lines 0830-0834, it can be seen that an overlap accompanies a co-construction: the Narrator intervenes to finish the construction started by the Reteller before the latter stops speaking.

**7.** Also filled with colors are empty areas in the "TimeS" and "TimeE" columns that correspond to the middle sections of one speaker's contribution. See, for example, the color filling in graphic lines 0809 and 0816. In 0809, the Narrator continues to pronounce EDU N-vE261, after the Reteller completed EDU R-vE024, while the Commentator has not yet started producing EDU C-vE163. In 0816, the Reteller continues speaking, while the Commentator takes a pause between two EDUs.

## References

Chafe, W. (1994) Discourse, consciousness, and time. Chicago: University of Chicago Press.

Clark, H. H., Fox Tree, J. E. (2002) Using *uh* and *um* in spontaneous speaking // Cognition, 84. - 73-111.

Grenoble, L. A. (2008). Sintaksis i sovmestnoe postroenie repliki v russkom dialoge [Syntax and co-construction in Russian dialogues] // Voprosy jazykoznanija, 1. 25-36.

Helsavuo, M.-L. (2004) Shared syntax: the grammar of co-construction // Journal of pragmatics. 36.

Kibrik, A. A. (2018). Russkij mul'tikanal'nyj diskurs. Čast' I. Postanovka problemy [Russian multichannel discourse. Part I. Setting up the problem] // Psixologičeskij žurnal 39 (1). 70–80.

Kibrik, A., Korotaev, N., Podlesskaya, V. (2019). Russian spoken discourse: Local structure and prosody. In print.

Kibik, A. A, Podlesskaya, V. I. (eds.) (2009). Rasskazy o snovidenijax: korpusnoe issledovanie ustnogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki slavjanskix kul'tur.

Kodzasov, S. V. (2009). Issledovanija v oblasti russkoj prosodii [Studies in the field of Russian prosody]. Moscow: Jazyki slavjanskix kul'tur.

Korotaev, N. A. (2015). Kommunikativno-prosodičeskij podxod k vyjavleniju èlementarnyx diskursivnyx edinic v ustnom monologičeskom tekste [Elementary discourse units in spoken monologues: Evidence from communicative prosody]. Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialog", 14 (21), 294–307.

Korotaev, N. A. (2018). Vopros i poluutverždenie v strukture mul'tikanal'nogo diskursa [Questions and semi-statements in multichannel discourse] // Vos'maja Meždunarodnaja konferencija po kognitivnoj nauke.

Podlesskaya, V. I. (2015). A corpus-based study of self-repairs in Russian spoken monologues. Russian Linguistics, 39 (1), 63–79.

Poddleskaya, V. I. (2018). Čužaja reč' v svete korpusnyx dannyx [Reported speech: a corpus-based approach] // Voprosy jazykoznanija, 4. 47-73.

Sacks, H., Schegloff, E., Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation // Language, 50. – 696-735.

Yanko, T. E. (2008). Intonacionnye strategii russkoj reči v tipologičeskom aspekte [Intonational strategies in spoken Russian from a comparative perspective]. Moscow: Jazyki slavjanskix kul'tur.

## Appendix 1: Conventions used in vocal text transcripts

The table below presents a list of conventions used in vocal text transcripts. For every convention, the denoted phenomenon, the corresponding section in the text, and correspondence in ELAN are provided.

| Convention | Phenomenon | See Section | Correspondence in ELAN |
|---|---|---|---|
| Lines in transcripts | Segmentation of the speech flow into higher-level units (EDUs and other). | 3.1, 3.2, 3.3 | Numbered intervals in *-vLine tiers |
| | Silent pauses between the higher-level segmentation units | 3.2, 3.3 | Numbered intervals in *-vPause tiers, with Out value in the dependent *-vPauseInOutEDU tier |
| (0.23) | Silent pause and its duration, s | 2.1.5 | Numbered intervals in *-vPause tiers |
| (ɥ 0.73) | Silent pause filled with a loud inhalation sound and its duration, s | | Numbered intervals in *-vSegm tiers, with HPause value in the dependent *-vSType tier |
| (ə 0.20), (ɐ 0.33), (ɯ 0.48), (ˀ 0.34), (əɯ 0.62) etc. | Filled pauses and their durations, sec | 2.1.3 | Numbered intervals in *-vSegm tiers, with Filled value in the dependent *-vSType tier |
| {laugh 1.02} | Laughter and its duration, sec | 2.1.4 | Numbered intervals in *-vSegm tiers, with Laugh value in the dependent *-vSType tier |
| {cl 0.12}, {st 0.23}, {gp 0.18} etc. | Other non-speech vocal phenomena | | Numbered intervals in *-vSegm tiers, with Other value in the dependent *-vSType tier |
| / \ − /\ etc. (placed before a word) | Pitch movements on stressed syllables of accented words | 4.1 | Non-empty intervals in *-vAccents tiers |
| ↑ ↓ → | Significant pitch movements on other syllables | | |
| Underlining of a word's stressed vowel | The given word bears the EDU's primary accent | 4.1 | Main value in the *-vMainAccent tier |

| Convention | Phenomenon | See Section | Correspondence in ELAN |
|---|---|---|---|
| Capitalization at the beginning of a line | Beginning of a new spoken sentence | 4.2 | |
| . | Statement | | Period value in the *-vIllocPhase tier |
| ? | Question | | Quest value in the *-vIllocPhase tier |
| ¿ | Semi-statement | | Semi-St value in the *-vIllocPhase tier |
| ¡ | Directive | | Dir value in the *-vIllocPhase tier |
| @ | Vocative | | Addr value in the *-vIllocPhase tier |
| , | Default incompleteness | | Comma value in the *-vIllocPhase tier |
| : | Incompleteness with further elucidation | | Colon value in the *-vIllocPhase tier |
| … | Inexhaustiveness combined with an illocutionary completion | | Dots-f value in the *-vIllocPhase tier |
| ,,, | Inexhaustiveness combined with an incompleteness | | Dots-nf value in the *-vIllocPhase tier |
| ! | Exclamation | | Exclam value in the *-vIllocPhase tier |
| — (beginning of line) | Entering a split-related inset | 4.5 | Split value in the *-vIllocPhase tier,<br>Enter value in the *- vParenth tier + InSplit value in the *- vInSplit tier |
| — (end of line) | Returning from a split-related inset | 4.5 | Return value in the *- vParenth tier + InSplit value in the *- vInSplit tier |
| ( ) | Inset | | Start, Final, Full values in the *- vParenth tier |
| (* | “One-sided” inset | | *Start value in the *- vParenth tier |
| “ ” | Direct or semi-direct quotation | 4.6 | Begin, End, Whole values in the *-vCitation tier |

| Convention | Phenomenon | See Section | Correspondence in ELAN |
|---|---|---|---|
| % | Co-construction in conversation | 4.4 | Non-empty intervals in *-vCoConstr tiers |
| \|\| | Mild internally-induced false start (the current EDU is not abandoned) | 4.3 | Mild value in the *-vInterrupt tier |
| == | Severe internally-induced false start (the current EDU is abandoned) | | Fst value in the *-vIllocPhase tier; Severe value in the *-vInterrupt tier |
| ⁑ | Mild externally-induced false start | | Mild-other value in the *-vInterrupt tier |
| ≈≈ | Severe externally-induced false start | | Interrupt value in the *-vIllocPhase tier; Severe-other value in the *-vInterrupt tier |
| ~ | Aposiopesis | | Tilde value in the *-vIllocPhase tier |
| = | Word truncation | 2.1.1 (pt. 5) | Truncated value in the *-vTruncated tier |
| written with the_underscore | Super-contracted pronunciation which corresponds to separate writing in standard spelling | 2.1.1 (pt. 7) | |
| ˀwordˀ | Glottal stop at the beginning / end of a word | 4.7.6 | Gl-st, Gl-en values in the *-vStops tier |
| ᵊwordᵊ | *schwa*-sound at the beginning / end of a word | | Schw-st, Schw-en values in the *-vStops tier |
| wordʷ | Labial stop at the end of a word | | Lab-en value in the *-vStops tier |
| wordʰ | Aspiration at the end of a word | | Asp-en value the *-vStops tier |
| a-a s-s ja-a j-ja | Phoneme lengthening | 4.7.3 | Len value in the *-vLength tier |
| rančó | Non-standard lexical stress | 4.7.7 | Stress value in the *-vStress tier |
| *Italics* | Accelerated tempo | 4.7.1 | Fast value in the *-vTempo tier |

| Convention | Phenomenon | See Section | Correspondence in ELAN |
|---|---|---|---|
| Increased letter-spacing | Decelerated tempo | | Slow value in the *-vTempo tier |
| Grey | Perceptible phonetic reduction | 4.7.2 | Non-empty intervals in th *-vReduction tiers |
| **Bold** | Emphasis | 4.7.4 | Emph value in the *-vEmph tier |
| Reduced font size | Heightened F0 register | 4.7.5 | Hi value in the *-vRegister tier |
| Reduced font size below the baseline | Lowered F0 register | | Lo value in the *-vRegister tier |
| #ty-dyš#<br>#whistle# | Onomatopoeias / non-verbal acts functionally identical to words | 2.1.2 | Onom value in the *-vOnom tier |
| <vot> | Presumable transcription of an uncertain fragment | 2.1.1 (pt. 8) | |
| >fervela< | Discernible but unidentified fragment | 2.1.1 (pt. 9) | |
| <UNCLEAR> | Unintelligible fragment | 2.1.1 (pt. 10) | |

# Appendix 2: a fragment of the individual vocal transcript (Session #22, Reteller)

| Time | EDU | Transcription | Comments |
|---|---|---|---|
| 817.15 | R-vE087 | \Vot, | |
| | | Well, | |
| 817.40 | R-vE088 | v to vremja poka on značit n= \|\| n-n= \|\| zalez na –lestnicu, | |
| | | while he well climbed the ladder | |
| 820.01 | R-vE089 | i sobiraet tam ėti /↓gruši-i,,, | |
| | | and is picking up these pears, | |
| 821.51 | pR-098 | (0.52) | |
| 822.03 | R-vE090 | s-s= == | |
| | | with… | |
| 822.29 | pR-099 | (0.21) | |
| 822.50 | R-vE091 | (Kak (0.45) *mne rasskazali* **berežno*. | |
| | | As I was told **carefully**. | |
| 824.44 | *R-vL024* | {laugh 0.71} | |
| 825.15 | R-vE092 | I s \ljubov'ju.) | |
| | | And with love. | |
| 825.94 | *R-vL025* | {laugh 0.77} | |
| 826.72 | R-vF005 | (ʁ 0.41) | |
| 827.13 | pR-100 | (1.64) | |

| Time | EDU | Transcription | Comments |
|---|---|---|---|
| 828.77 | R-vE093 | mimo proxodit (ə 0.39) kakoj-to eščë-oᵊ (0.29) drugoj mužčina s /kozoj, | |
| | | Past him another man passes by with a she-goat, | |
| 832.49 | pR-101 | (0.17) | |
| 832.66 | R-vN020 | (ɥ 0.32) | |
| 832.98 | R-vE094 | i koza-a (1.63) xočet vidimo-o (0.19) (ʔ 0.08) (0.20) polakomit'sja /grušami, | |
| | | and the she-goat wants apparently to regale itself with pears, | |
| 837.89 | R-vE095 | no-o {sf 0.48} (ɯ 0.57) (0.62) eë xozjain ej ne \daët ėtogo /sdelat', | The first main accent is on the verifying rheme, and the second one marks incompleteness. |
| | | but its owner doesn't let it do this, | |
| 841.67 | pR-102 | (0.26) | |
| 841.93 | R-vN021 | (ɥ 0.11) | |
| 842.04 | R-vE096 | (–vot,) | |
| | | well, | |
| 842.20 | R-vE097 | oni proxodjat /↑mimo, | |
| | | they pass by, | |
| 843.16 | pR-103 | (0.25) | |
| 843.40 | R-vN022 | (ɥ 0.28) | |
| 843.68 | R-vF006 | (ə 0.34) | |
| 844.02 | pR-104 | (0.72) | |
| 844.74 | R-vE098 | vsë ėto /vremja značit /sadovnik na de= \|\| n-na= \|\| na-a = \|\| na= \|\| na \lestnice. | |

| Time | EDU | Transcription | Comments |
|---|---|---|---|
| | | all this time the gardener is in the tr…, on on on… on the ladder. | |
| 847.96 | R-vN023 | (ɥ 0.48) | |
| 848.45 | R-vE099 | (ə 0.18) –Vot, | |
| | | Well, | |
| 848.82 | R-vE100 | potom priezžaet /mal'čik, | |
| | | then comes a boy, | |
| 850.15 | R-vE101 | po-moemu na \krasnom /↓velosipede,,, | |
| | | I think on a red bike, | |

# Appendix 3: a fragment of the scores vocal transcript (Session # 22)

| Line # | Pauses | | TimeS | TimeE | Narrator | | Commentator | | Reteller | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0806 | vp084 | (0.04) | 538.63 | 538.67 | | | | | | |
| 0807 | | | 538.67 | | N-vE261 | U nix u \vsex šljapy. | | | | |
| 0808 | | | 538.89 | 539.23 | | | | | R-vE024 | Li ≈≈ |
| 0809 | | | | | | | | | | |
| 0810 | | | 539.46 | | | | C-vE163 | U nix u /vsex, | | |
| 0811 | | | | 539.68 | | | | | | |
| 0812 | | | | | | | | | | |
| 0813 | | | 539.86 | | | | | | R-vE025 | I u /maľčikov troix \tože šljapy-y¿ |
| 0814 | | | | 540.12 | | | | | | |
| 0815 | | | 540.12 | 540.30 | | | C-vE164 | k°= ≈≈ | | |
| 0816 | | | | | | | | | | |
| 0817 | | | 540.90 | 541.23 | | | C-vE165 | \Net. | | |
| 0818 | | | 541.23 | | | | C-vE166 | Vot u troix maľčikov /net, | | |
| 0819 | | | 541.26 | 541.68 | N-vE262 | /\Ne-et! | | | | |
| 0820 | | | 541.68 | | N-vE263 | U detej \netu. | | | | |
| 0821 | | | | 541.74 | | | | | | |
| 0822 | | | | 542.27 | | | | | | |
| 0823 | | | 542.27 | | | | C-vE167 | i u /\devočki netu. | | |
| 0824 | | | 542.48 | | | | | | R-vE026 | To est' ≈≈ |
| 0825 | | | | 542.58 | | | | | | |
| 0826 | | | | 542.69 | | | | | | |
| 0827 | | | | 543.19 | | | | | | |
| 0828 | vp085 | (0.42) | 543.19 | 543.61 | | | | | | |
| 0829 | | | 543.61 | | | | C-vE168 | /Šljapy ≈≈ | | |
| 0830 | | | 543.62 | | | | | | R-vE027 | /Šljapy toľko-o % |
| 0831 | | | | 544.09 | | | | | | |
| 0832 | | | | | | | | | | |
| 0833 | | | 544.32 | | N-vE264 | % u \vzroslyx. | | | | |
| 0834 | | | | 544.66 | | | | | | |
| 0834 | | | | 545.10 | | | | | | |
| 0835 | vp086 | (0.01) | 545.10 | 545.11 | | | | | | |
| 0836 | | | 545.11 | 545.65 | | | | | R-vE028 | U /vzroslyx. |
| 0837 | | | 545.65 | 546.31 | | | | | R-vE029 | i u /–maľčika. |
| 0838 | | | 546.31 | | | | | | R-vE030 | kotoryj na \velosipede sootvetstvenno¿ |
| 0839 | | | 547.13 | 547.43 | | | C-vE169 | (u 0.11) \Da-a. | | |
| 0840 | | | | | | | | | | |
| 0841 | | | 547.80 | | | | C-vE170 | –Da. | | |
| 0842 | | | | 547.83 | | | | | | |
| 0843 | | | | 548.03 | | | | | | |
| 0844 | vp087 | (0.08) | 548.03 | 548.10 | | | | | | |