Conventions for transcribing speech.

0. Segmentation – when segmenting the audio for transcription, try not to make each segment overlong. For instance, two main clauses should be in different segments:

*S1: Давай, одну куколку я возьму. Все.* S2: Не, вторую куколку мне тяжело.

or

*S1: Давай, одну куколку я возьму. S2: Все. S3: Не, вторую куколку мне тяжело.

1. **Numerals** are spelled out in full, without abbreviations.

2. **Abbreviations**, which are written as one word but pronounced as separate letters, need to be written in capital letters (for instance, МГУ). Don’t put periods after each letter or after the whole abbreviation. Words that are often abbreviated in writing, but pronounced in full should be written in full, for instance, terms of address мистер, госпожа, expressions и так далее.

3. To make the reading process easier, it is necessary to use the **standard punctuation rules**; however, please limit your punctuation to using periods, commas, question marks, and exclamation marks. Quotation marks should be used exclusively to indicate direct speech or the thoughts of a speaker; quotation marks should not be used for any other purposes.

Please remember, that you DO NOT need to FULLY COMPLY with the Russian punctuation rules. We expect that your decisions about the use commas and exclamation marks will be the most difficult ones.

**Commas** should be used to indicate a small pause; it usually occurs between the main and the subordinate clauses inside sentences or after a person is addressed. DO NOT use commas to separate interjections or introductory phrases from the main clause. *(Ну конечно мы можем!* This sentence should not have any commas in it).

**Exclamation marks** should be used in a sentence when you hear or feel an emphatic stress or intonation. Compare the following: Пойдем! Выберите какую-нибудь книгу, мы её почитаем. and Пойдем, какая кукла красивая. Two instances of the same word “пойдем” are used with different intonation in these two sentences. In the first case, it is emphatic; however, in the second sentence we do not even observe a pause between the imperative пойдем and the following word какая.

You will have to make some relatively subjective decisions on the use of the exclamation mark after imperatives. An exclamation mark should be used if you hear an emphasis.

Please pay particular attention to phrases like Ой, какая хорошая. Most likely you would use an exclamation mark due to the inherently emphatic nature of this type of expression.
Hyphens. When using a hyphen, please abide by the Russian grammar rules. It is used before the particles -то, -нибудь, -либо and after кое- (кто-то, как-то, а как ты-то? кто-нибудь, где-нибудь, кто-либо, о чем-либо, кое-как, кое-кто).

Use hyphens to mark interjections, which are uttered without pauses between them. (This is what usually happens.) Ой-ои-ои, да-да, да-да-да, нет-нет. If, however, interjections are articulated as separate words please use periods to separate them. Ой. Ой. Ой.

CONSISTENCY. Being consistent in the use of commas, periods, and other punctuation marks is a very important component of your work. The end of a sentence should always be marked with a period even if this sentence consists of one word. Occasional, inconsistent use of a period to mark the end of a sentence is unacceptable.

4. Disfluent speech.

Disfluent speech is the hardest speech to transcribe. Speakers may repeat themselves, stumble, utter only parts of words, repeat beginning of a phrase or of a sentence, and they may also use multiple hesitation sounds in order to fill pauses in speech. You should pay particular attention to these aspects of speech, reflecting exactly what has been said including all redundancies, hesitations, self-corrections, partially uttered words, etc.

Unfinished phrases, where speakers interrupt themselves, or when they are interrupted by others or get distracted, should be marked with the tag <BREAK> (always leave spaces before and after all tags). Он пришел домой и сел <BREAK> Да, положи это обратно в холодильник.

5. Fillers

Filler words are non-lexemes, or non-words, which are used by a speaker to fill the pause in his speech indicating his hesitation yet intent to continue his thought while he is pondering over the form of expressing it.

Each language has its own set of filler words. Russian speaker may use such sounds, words and phrases as э-э-э, а-а-а, ну, вот, так сказать, это.

Please keep track of all filler words that you use in your transcription. Such practice will help standardize the orthography and improve these instructions. Please, periodically send your list to Sophia by email or save it in the folder titled Transcription Materials.

6. Partially produced words and stuttering.
You should transcribe everything as you hear it, even if a speaker aborts the production of a word after a few initial sounds. Please mark the word with a dash with no space between the dash and the word. Please write down the entire word immediately after the partially produced word with a plus sign in front of it, if you can guess which word it was.

И она пош- домой.
И она пош- +пошла домой.

A dash is used without a space immediately following the last letter; then a space is used, followed by a plus sign with a word following it without a space.

Please limit your guess to one word, and do not add multiple-word phrases even if the context clearly indicates which phrase was implied.

Please ignore stuttering, such as repetition of a syllable in the beginning or middle of a word.

7. Repetition of the beginning of a phrase, “false starts” or restarts.

When speakers repeat the beginning of a phrase or a sentence, changing their train of thought, mark this false-start with the tag <FS before the beginning of the repeated segment (always leave a space on both sides of the tag) and <$FS> after the end (again, with spaces).

Поскольку мы <FS> с Дальнего <$$FS> ну с Дальнего севера, мы всегда получали льготы.

8. Insertion of English or German words, newly coined words / non-words.

Use Latin alphabet for inserted English or German words:

Nein!

If the word is morphologically or phonetically adapted into Russian, please spell it the way you hear it, using Russian letters, and indicate that this is a borrowing as follows. Leave spaces before and after the tag.

Катя, пойдешь к <BOR> аме <$$BOR> с <BOR> опой <$$BOR> ?
Катя, давай <BOR> пошерим <$$BOR> с тобой бутерброд?

Neologisms invented by the speaker during the conversation or non-words should be marked as {neo} and their pronunciation should be tagged using <PR> <$PR> (see instructions on transcribing children’s pronunciation below).
9. Incorrectly pronounced words and non-standard words.

It is necessary to make a distinction between incorrectly pronounced words and regional or other dialects. Черное кофе vs черный кофе, and зво'нит vs звони'т are examples of the different regional dialects. You do not need to highlight them.

Каська is an example of an incorrect pronunciation of the word кашка; same goes for миционер vs милиционер. Although this is not obligatory, you may mark such words with an asterisk in front. However, please spell the word correctly!

Я люблю *кашку.

Every word may be marked with an asterisk if all the words in the phrase are pronounced incorrectly *будешь *мультик *смотреть?

Child’s pronunciation should be marked only in cases when it is drastically different from that of an average adult (for example, if part of a word is missing). In such cases please mark the child’s pronunciation as non-standard by using the tag <PR>. Include the actual pronunciation in the tag, and write the correct spelling (if you know what word was meant) in between the tags. If you cannot guess what word was meant, just write this as a non-word {neo}. Please see the examples below.

- The child said мат'я, it’s impossible to figure out what this means.
  - <PR матъ> {neo} <$SPR>
- The child said баки, and from the context it’s clear that бабочки were meant.
  - <PR баки> бабочки <$SPR>
- The child said заку, which could be за руку, or alternatively could be зайку.
  - <PR заку> [за руку] | зайку <$SPR>
- The child said something unclear, which sounded close to васьк, and from the context it’s clear that паровозик is meant.
  - <UNCLEAR> <PR васьк> паровозик <$SPR> <$UNCLEAR>
- The child said something unclear, which sounded either like асел (a non-word) or as он сел.
  - <UNCLEAR> <PR асел> {neo} <$SPR> | он сел <$UNCLEAR>

Катя: Там <PR хонана> холодно <$SPR>.
Мама: Холодно. Холодно.
Папа: «Хонана».  Here the transcriber has a choice: you can mark it as an incorrectly pronounced word холодно, as with Катя, <PR хонана> холодно <$SPR> or *холодно, or as a quotation «хонана».

In all other cases please spell the word using standard orthography:
  *енот or енот, and not анот or ёнот
  *обезьяна
  *отдай
  *будешь *мультик *смотреть
What is the extent of “incorrectness” that prompts us to use the mark-up with <PR> or an asterisk?
We do not yet know the full answer to this question. We intend these tags to be used sparingly,
mostly where syllables are missing.

Thus, if a child says ласатка, we will write лошадка.

10. Unintelligible speech.

Sometimes you will come across words or phrases which are difficult or impossible to understand. If you
cannot hear the phrase clearly but you can guess what it is with a high level of certainty, write your
guess between the tags <UNCLEAR> <$$UNCLEAR> (leave spaces before and after all tags).

И она мне сказала <UNCLEAR> что я должна уйти <$$UNCLEAR>.

If you cannot hear anything at all or if you cannot make a guess, mark this segment with the tag
{inaudible}.

If there are two options for what the phrase might be, but you are not sure which one is right, you can
write down both, separating them with a vertical line | (leave spaces before and after the line).

Я хочу <UNCLEAR> лошадку | лисичку <$$UNCLEAR>.
Дай мне, пожалуйста, <UNCLEAR> эту скалку | ту скалку <$$UNCLEAR>.

11. Dots over è and stress marks.
   DO NOT put dots over e.
   Do not indicate stress. This is a different task altogether.

12. Other tags.

The following symbols are used to indicate non-speech sounds:

   {BR} breath (Speaker takes a deep breath)
   {CG} cough (Speaker coughs or clears their throat)
   {LS} lip smack (Speaker smacks their lips)
   {LG} laughter (Speaker laughs)
   {NS} noise (Loud noise in the background: for example, a door slams shut, cars are honking, etc.)
   {SN} sneeze, sniff
   {CR} crying


Annotators should not try to correct non-standard grammatical structures, like я скучаю за ним. Such
phrases should be written the way you hear them. The same rule applies to transcribing words that are
used incorrectly. Annotators should transcribe what was said and not what should have been said or what they expected to hear.

However, annotators should not try to reproduce the non-standard pronunciation of a speaker. You should use a standard spelling even if the word is non-standard. Words that are OBVIOUSLY pronounced incorrectly should be marked with an asterisk (see above). Exceptions from this rule are

- the word что, which is frequently pronounced as че or шо
  - these non-standard pronunciations should be written as you hear them
  - Nota bene: the standard pronunciation is што, both this pronunciation and the non-standard pronunciation что should be written using standard spelling что
- and the word сейчас, which is frequently pronounced as щас
  - this non-standard pronunciation should be written as you hear it
  - the standard pronunciation сейчас should be written using standard spelling сейчас

Ну и че? Они ж с Магадана. И этим все сказано.

14. Proper names

Annotators should make notes of the places in audio recordings where personal information is used, i.e. proper names that can identify the speaker, such as personal names, names of cities, addresses, etc. These notes will be used to “de-identify” the recordings in the future, that is to erase all personal information from the recordings and transcripts.

In order to complete this task, please fill out a table for each audio file. Time stamp reflects the time in the recording when personal information begins.

Do not keep track of the names of dolls, dogs, or cats. Proper names used during games, such as names of cities said when playing “Goroda” can also be left out. If you are unsure whether to make a note for the name or not, do note it!

Please use the audio file names in naming the MS Word file with the corresponding table.

<table>
<thead>
<tr>
<th>Time Stamp</th>
<th>Proper nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:00:15</td>
<td>Алисса</td>
</tr>
<tr>
<td>0:20:05</td>
<td>Германия</td>
</tr>
<tr>
<td>21:02:13</td>
<td>Алисса</td>
</tr>
<tr>
<td>1:05:01</td>
<td>Бавария</td>
</tr>
</tbody>
</table>