Appendix

We now provide supplementary materials. In Appendix A, we provide the text of instructions given to workers on the Toloka crowdsourcing platform. Appendix B provides a screenshot of the interface; details of the filtering procedure mentioned in Section 3 are given in Appendix C. In Appendix D, we provide additional details on the T5-based models. Finally, Appendix E reports the analysis of errors made by humans and algorithms on our data.

A Task Instruction

Figure 3 displays the task instruction presented to the workers. HTML template of the instruction is also available on our GitHub data release.

Transcription Rules

- All recordings have a clearly identifiable speaker – you need to transcribe their speech only. If there is some background speech, ignore it.
- Your transcription must use only letters and apostrophes ('). Do not use digits and any punctuation marks (including the question mark "?") except the apostrophe.

Important Details:

1. If you hear a number, spell it out in words (e.g., 19 -> nineteen).
2. Even when the grammar rules and speaker's intonation suggest the punctuation mark, omit it (except the apostrophe). We need to obtain texts without punctuation marks.
3. Use the apostrophe according to the grammar rules:
   - To show omissions of letters (e.g., constructions like I'm, don't, shouldn't).
     Importantly: listen carefully to what the speaker says. If they use a full form (e.g., do not), you must also write the full form. If they use a shortform (e.g., I'm), you should also write the short form.
   - To show possession (e.g., Ivan's pen)
   - To form plurals of letters/abbreviations etc. (e.g., She got two A's in the biology exam)

Pipeline

Follow a simple pipeline to perform the task:

1. Play the audio and carefully listen to the speech.
   Important: You must listen to the complete audio record to submit your response.
   Technical Difficulties: If the audio does not play, or there is no voice in the recording, or any other technical difficulty arises, answer "No" to the "Does the audio play properly?" question and proceed to the next task (skip all steps below).
   Hint: It is best to use headphones to perform this task – you will hear the speech better.
2. Transcribe the audio and type the transcription into the text field
   Important: You must follow the transcription rules outlined above.
   Hint: You can play the audio multiple times and pause it at any point. Please do your best to produce a high-quality transcription.
3. Carefully check your transcription for typos and mistakes
   Important: Speech in the recordings should be grammatically correct. But if you are sure that the speaker makes a mistake, do not fix it. Your goal is to provide accurate transcriptions.
   Hint: Listen to the audio in full one more time to double-check your transcription.

Exam

To start working on the main task, you have to pass a qualification exam. In the exam, you need to transcribe 10 audios and you will be accepted to the main task if the quality of your transcriptions is high enough.

FAQ

Q: Do I get paid for the tasks I had technical difficulties with?
A: Unfortunately, no. We will reject these tasks, but it won't hurt your status or payment for other tasks.

Q: I have technical difficulties with all audios; what to do?
A: Try to use another browser (we recommend Google Chrome).

Q: I do not understand some words in the audio; what to do?
A: Some audios are hard to transcribe, and it is ok. Listen to it once more and write your best guess if you do not know for sure.

Q: When will I get paid?
A: We try to review the tasks within several hours, but occasionally it may take up to a day.

Q: All my tasks are rejected. Why?
A: We have some spam detection rules in place and reject users who spam. Additionally, we have some ground truth transcriptions and may block a worker if they consistently supply clearly inadequate transcriptions. If you believe that we made a mistake, shoot us a note.

Figure 3: Task instruction presented to the workers.

B Task Interface

Figure 4 demonstrates the interface of the task available to workers. In order to activate the text field to enter the annotation, a worker needs to positively answer the first question. The negative answer to the first question indicates a technical issue and enables the worker to skip the question.

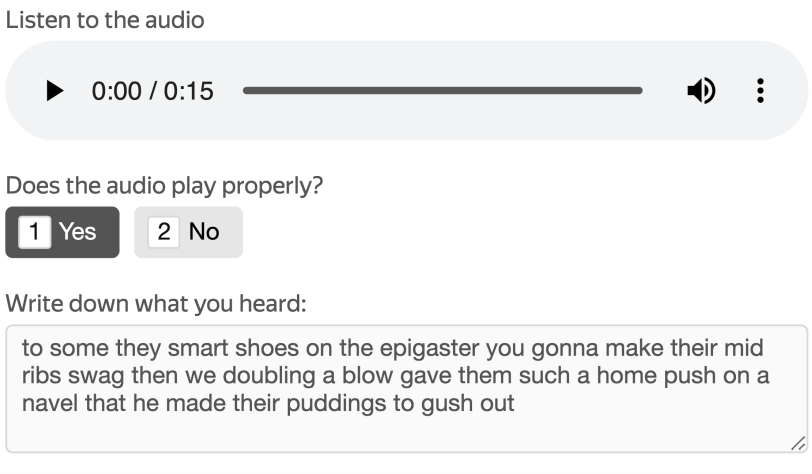

Figure 4: Task interface. There was a single audio-annotation task on each page, and a worker could complete as many tasks as they like (subject to availability of tasks).

C Additional Filtering of the RUSNEWS Dataset

In order to align the format of sentences in the RUSNEWS dataset to the format of the LIBRISPEECH ground truth, we perform the following filtering steps on the RUSNEWS dataset.

- *Remove sentences that contain digits.* From the workers' standpoint, numerals and spelled-out numbers in the original texts do not make a difference because they are pronounced in the same way. However, to compute the accuracy of annotations (both crowdsourced and aggregated), we need to compare a target text against the ground truth text. Thus, it is crucial to ensure that numbers in the ground-truth texts are written in a consistent manner. To ensure this consistency, we remove sentences that contain digits.
- *Remove sentences that contain specific abbreviations.* Some texts in the source contain abbreviations that are special for the Russian written language (e.g., a Russian equivalent of the word "doctor" is sometimes shortened to "d-r"). Such abbreviations are not used in the spoken language so we remove sentences that contains them from the pool.
- *Remove sentences that contain letters from non-Russian alphabets.* Finally, some sentences in the initial pool contain words from other languages (e.g., names of companies). We also remove such sentences from the pool because these cannot be properly annotated using Russian alphabet.

In practice, instead of removing sentences that fall in the aforementioned categories, one could alternatively pre-process sentences to convert numerals to the spelled-out form and adjust instructions to account for abbreviations and non-Cyrillic letters. However, careful implementation of such changes requires a non-trivial amount of work and we do not take this route in this paper as it is orthogonal to our main contributions.

D T5, a Transformer-based Model

In this section, we give a high-level overview of the T5 model — a workhorse for comparisons we described in Section 6.3. Additionally, we provide details on the fine-tuning procedure.

T5 Model. Our evaluations in Section 6.3 rely on a pre-trained T5-large model [41] — a Transformer-based model that is designed to solve various text-to-text tasks. By using T5, we reduce the problem of aggregation of crowdsourced transcriptions to a more studied text summarization problem. In our case, the input to the model consists of concatenated transcriptions provided by crowd workers for a particular recording, and the T5 model outputs a single final transcription.

Details of Fine-Tuning The main power of the T5 model is its ability to quickly adjust to new tasks: starting from initial weights provided by Google (`https://huggingface.co/t5-large`), one can utilize the knowledge of pre-trained T5 while letting the model to pick up the new task. In this work, both T5 (ST + FT) and T5 (FT) models are fine-tuned on the corresponding data for 8 epochs with 10% of data set aside for the validation purposes. Specifically, to train the T5 (FT) model, we use the original weights provided by Google as the initialization. For T5 (FT + ST), we started with the weights reported by Pletenev [39]. The fine-tuning procedure directly follows the summarization training approach by HuggingFace.[7]

E Error Analysis

In order to further understand the difference between the performance of aggregation algorithms, we now conduct an analysis of errors made by human annotators and aggregation algorithms discussed in Section 6. Specifically, we rely on automated and manual analysis and evaluate (i) the impact of errors in human annotations on the performance of the algorithms and (ii) the causes of errors made by crowd workers.

Automated Analysis We begin from the automated analysis. For each example in the CROWDSPEECH test-clean dataset, we compute the WER scores of each human annotation and the output of each algorithm under comparison. We then treat transcriptions with zero WER as correct (+) and those with non-zero WER as incorrect (−). To study the effect of the quality of human transcriptions on the accuracy of aggregation algorithms, we now compare the performance of the algorithms with breakdown by three different settings:

- **All Correct** All workers provided the correct transcription
- **Has Correct** At least one worker provided the correct transcription
- **All Incorrect** All the workers provided an incorrect transcription

Similar to the experiments conducted in Section 6, we compare the performance of the baseline and novel methods with Oracle that always produces the best transcription submitted by the workers.

Results of the comparisons are summarized in Table 6. Let us now make several important observations:

- **All Correct** First, we observe that none of the methods introduced errors if the crowd unanimously agreed in their correct response.
- **Has Correct** Next, as we increase the difficulty of the task, methods start to behave differently with T5 demonstrating better performance than other methods.
- **All Incorrect** Finally, in the most challenging category, RASA, HRRASA, and Oracle always produce an incorrect result with non-zero WER. In contrast, ROVER and T5 perform better and sometimes are able to recover correct transcriptions even when all transcriptions contain mistakes.

These observations corroborate the results presented in Section 6 (Table 4 and Table 5) and demonstrate promise of methods that are not limited to crowd-generated transcriptions (ROVER and especially T5).

As a side remark, we note that RASA and HRRASA showed identical performance in this experiment, while the former outperforms the latter in terms of the WER metric (Table 4). Inspection of the data reveals that HRRASA tends to choose transcriptions with a higher WER than RASA when they both make errors, thereby being scored worse on the WER metric.

[7] `https://github.com/huggingface/transformers/tree/master/examples/pytorch/summarization`

Table 6: Comparison of the automated algorithms with breakdown by the human performance. Specifically, for each of the three regimes (All Correct, Has Correct, All Incorrect), we compare how often aggregation methods are able to come up with perfect transcriptions. Comparison is executed on the test-clean subset of CROWDSPEECH.

Crowd	ROVER		RASA		HRRASA		T5 (ST)		Oracle	
	+	−	+	−	+	−	+	−	+	−
All Correct	46	0	46	0	46	0	46	0	46	0
Has Correct	1,055	482	1,085	452	1,085	452	1,146	391	1,537	0
All Incorrect	63	974	0	1,037	0	1,037	157	880	0	1,037

Manual Analysis Next, we proceed to the manual analysis with a goal of understanding the causes of human errors. In that, we condition on a subset of hard recordings: those, that caused non-zero WER in outputs of all algorithms under consideration and sample 100 transcriptions from this subset for manual analysis. In the analysis, the authors of this paper independently classified human errors into three predefined categories:

- **Task Difficulty** Some recordings were lengthy or contained unexpectedly difficult lexemes, such as rare words or proper nouns. Thus, the first cause of mistakes is the task difficulty.
- **Violation of Instructions** The second category corresponds to various violations of instructions: incorrect punctuation marks, wrong formatting of numbers, incomplete transcriptions – all these mistakes were classified in this category.
- **Homophones** Finally, a transcription could be grammatically correct and meaningful, but contain homophones or verbs in wrong tenses that impact the WER score. We classified such cases in a separate category.

In the sample of 100 transcriptions that we annotated, homophones caused 55 errors, difficult or lengthy recordings caused 28 errors, and instruction breaks caused the remaining 17 errors. We evaluated the inter-rater agreement of our analysis using Krippendorff's alpha for nominal scale and found that our annotation is reliable in classifying error causes ($\alpha = 0.81$).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]