






Everyday Conversations: A Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level

Tatiana Sherstinova¹ , Rostislav Kolobov² , and Nikolay Mikhaylovskiy^{2,3} 

¹ HSE University, Saint Petersburg, Russia 190068
tsherstinova@hse.ru

² NTR Labs, Moscow, Russia 129594
{rkolobov,nickm}@ntr.ai

³ Higher IT School of Tomsk State University, Tomsk, Russia 634050

Abstract. The study examines the outcomes of automatic speech recognition (ASR) applied to field recordings of daily Russian speech. Everyday conversations, captured in real-life communicative scenarios, pose quite a complex subject for ASR. This is due to several factors: they can contain speech from a multitude of speakers, the loudness of the conversation partners' speech signals fluctuates, there's a substantial volume of overlapping speech from two or more speakers, and significant noise interferences can occur periodically. The presented research compares transcripts of these recordings produced by two recognition systems: the NTR Acoustic Model and OpenAI's Whisper. These transcripts are then contrasted with expert transcription of the same recordings. The comparison of three frequency lists (the expert transcription, the acoustic model, and Whisper) reveals that each model has its unique characteristics at the lexical level. At the same time, both models perform worse in recognizing the following groups of words typical for spontaneous unprepared dialogues: discursive words, pragmatic markers, backchannel responses, interjections, conversational reduced word forms, and hesitations. These findings aim to foster improvements in ASR systems designed to transcribe conversational speech, such as work meetings and daily life dialogues.

Keywords: ASR Systems · Acoustic Model · Whisper · Field Recordings · Everyday Conversations · Dialogues · Russian Language · Vocabulary · Word Lists · Backchanneling · Discursive Words · Pragmatic Markers · Interjections · Conversational Word Forms · Hesitations

1 Introduction

Speech recognition technologies are currently advancing at an unprecedented pace. Utilizing neural networks, these systems can produce high-quality transcripts of spoken language in real-time, as exemplified by automatically generated subtitles for YouTube videos. However, challenges persist when dealing with everyday, real-world speech dialogues that encapsulate private interactions among individuals. Unlike curated YouTube

recordings intended for mass consumption [15], these sound recordings are derived from authentic communication contexts. Speakers may be at a significant distance from the recording device, the number of interlocutors can greatly fluctuate, and the recordings often contain various background noises reflecting the aural realities of daily life. Frequently, these include unrelated speech elements, such as background television, radio, or simultaneous conversations among other groups of people. The need for collecting this type of field speech data spans both scientific research and practical applications, such as monitoring speech for medical or other specialized purposes. They also serve as invaluable training data for AI systems, enabling them to communicate more authentically or simulate real-world scenarios. The Japanese ESP corpus (from the JST/CREST Expressive Speech Processing project) [6], the BNC [20], and the Russian ORD corpus [3], which forms the basis for this research, are notable examples of databases containing such everyday speech conversations.

The measure of a speech recognition system’s quality is its alignment with human auditory perception. Hence, expert human transcription is deemed the “gold standard”, providing the reference point for key performance indicators of speech recognition systems, including the Word Error Rate (WER), Character Error Rate (CER), and other associated metrics [1, 16, 23]. In this study, we examine how two modern Automatic Speech Recognition (ASR) models of different types handle these challenges. Our analysis employs a lexical-statistical approach, comparing the frequency lists of words derived from each system.

2 Data and Method

2.1 Data

The study is based on 195 macro-episodes of everyday speech communication taken from the ORD corpus [5, 22], obtained from 104 participants. The data set includes everyday (both casual and institutional) communication dialogues featuring men and women from all age groups (young, middle-aged, and seniors), and across various professional sectors: laborers involved in manufacturing or construction, service industry workers, educators, law enforcement officers, creative professionals, office workers, IT specialists, engineers, humanities scholars, and natural scientists.

The total volume of speech data consists of approximately 300,000 instances of word usage. An anonymized version of this sample is publicly accessible on the ORD corpus website [17]. The macro-episodes used in this study capture the full spectrum of everyday communication—the recordings were made in a variety of settings, such as homes, workplaces (offices and manufacturing sites), universities, medical facilities, shops, cafes, restaurants, and outdoor public spaces [21].

Expert transcription of the recordings was carried out using the ELAN multimedia annotation software [9]. This process involved several iterations where the transcriptions were reviewed and corrected by at least three experts. We employed an “error correction” method, where each expert would correct the errors made by the previous one, retaining the prior transcription for reference if needed. However, ORD expert transcription method does have one minor drawback—the text is presented on a linear scale, which can oversimplify the representation of a multichannel speech signal. Furthermore, not

all fragments of the speech signal were decipherable by the experts. In instances where the spoken content was unclear, the experts marked the section with an asterisk followed by the letter H (for “hard to decipher”).

2.2 Method

Transcripts were obtained for the same 195 audio recordings of speech episodes using two speech recognition systems: the NTR Acoustic model and the OpenAI Whisper system.

NTR Acoustic Model is a non-autoregressive headless variant of Conformer [8] which uses CTC loss instead of Transducer and is based on NVIDIA NEMO Conformer-CTC large [11]. Conformer-CTC has a similar encoder as the original Conformer but uses CTC loss and decoding instead of RNNT/Transducer loss, which makes it a non-autoregressive model. This model uses the combination of self-attention and convolution modules to achieve the best of the two approaches, the self-attention layers can learn the global interaction while the convolutions efficiently capture the local correlations. The self-attention modules support both regular self-attention with absolute positional encoding, and also Transformer-XL’s self-attention with relative positional encodings. Figure 1 presents Conformer-CTC encoder architecture.

In addition to the original NVIDIA NEMO Conformer-CTC training datasets Mozilla Common Voice 10.0 [13] (28 h), Golos-crowd (1070 h) and fairfield (111 h) [14], Russian LibriSpeech (RuLS) [4] (92 h) and SOVA—RuAudiobooksDevices (260 h) [12] and RuDevices (75 h) [12] NTR Acoustic Model is trained on a private Russian part of NTR MediaSpeech dataset (1000 h) [15].

Whisper [19] is an encoder-decoder audio-to-text Transformer [24] that ingests 80-channel logmagnitude mel-spectrogram computed on 25-ms windows with a stride of 10 ms from audio sampled at 16,000 Hz. The encoder processes this input representation with a small stem consisting of two convolution layers with a filter width of 3 and the GELU activation function [10] where the second convolution layer has a stride of two. Sinusoidal position embeddings are then added to the output of the stem after which the encoder Transformer blocks are applied. The transformer uses pre-activation residual blocks [7], and a final layer normalization is applied to the encoder output. The decoder uses learned position embeddings and tied input-output token representations [18]. The encoder and decoder have the same width and number of transformer blocks. Figure 2 summarizes Whisper architecture.

Whisper is trained with multitask approach on a large and diverse dataset [19]. The training tasks included transcription, translation, voice activity detection, alignment, and language identification. The training dataset was constructed from audio that is paired with transcripts on the Internet. Several automated filtering methods were used to improve the transcript quality. These included removing ASR-generated transcripts, ensuring that the spoken language matches the language of the transcript according to CLD2, and removing low-quality data sources based on WER of the initial model.

Thus, for the purpose of this research, two ASR systems of entirely different nature were selected. OpenAI Whisper is essentially a multilingual language model with an acoustic input. It possesses extensive knowledge about languages and written speech and is partly involved in translating spoken language into written form. As such, its

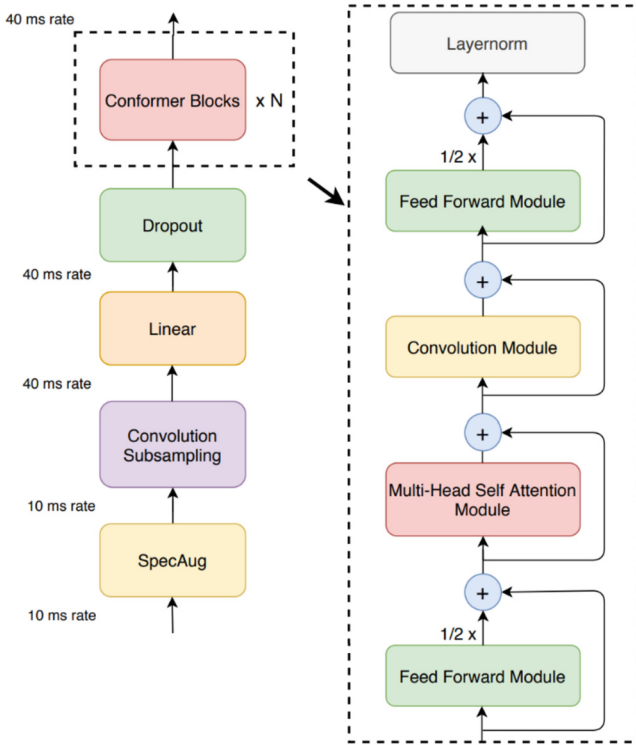


Fig. 1. Overall architecture of the encoder of Conformer-CTC.

errors relative to expert translations are often considered as “excessive literarization” mistakes. In contrast, the NTR Acoustic Model has no understanding of language but acts as a transcription tool. It transcribes exactly “what it hears”, without considering the accuracy of spelling. A significant portion of its errors could be classified as “illiterate”. Consequently, these two models can be regarded as two opposing poles for ASR, with all other models occupying intermediate positions. Therefore, we can anticipate that the recognition results of the other ASR systems would fall somewhere in the middle of this spectrum. It is for this reason that we decided to use OpenAI Whisper and NTR Acoustic Model these systems for this study.

The models’ performance was evaluated based on the Word Error Rate (WER) across 195 speech episodes. The NTR Acoustic Model yielded an average WER of 65%, with a single speech episode achieving the best performance of 30% WER, and the poorest performance reaching 99%. On the other hand, the Whisper system had a lower average WER of 49%, with the best-performing episode yielding a 7% WER and the worst reaching 99%. These numbers are an indication of how complicated the ORD dataset is even for the best contemporary speech recognition systems.

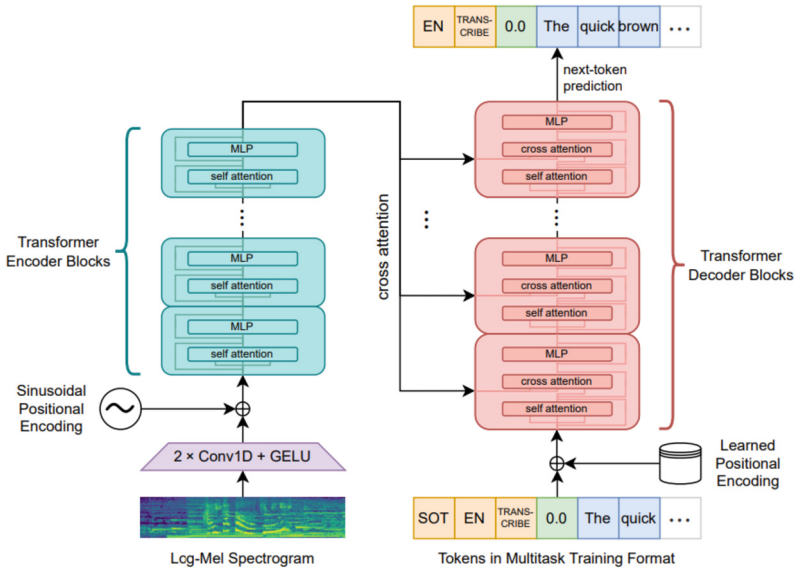


Fig. 2. Whisper architecture overview (from [19]).

To assess the models from a lexical perspective, we employed AntConc [2], a corpus management tool, to create three frequency dictionaries. The referent dictionary represented the transcriptions considered as the “gold standard”, which were produced by human experts. The second and third dictionaries comprised the recognition results from the NTR Acoustic Model and the Whisper system, respectively.

One notable aspect of the expert-generated transcriptions is their treatment of overlaid speech. When speakers’ volumes were sufficiently loud, the experts discerned multiple speech channels, recording them linearly as a sequence of utterances. Therefore, simultaneous speech from different speakers is transcribed as consecutive utterances. Thus, it becomes inherently impossible to achieve a 100% alignment between the performance of the recognition system and the expert transcription, since the considered recognition systems do not distinguish background speech during overlapping speech.

By comparing the frequency dictionaries, we were able to identify the lexical items that the models recognized more or less effectively.

3 Word Lists Statistics

3.1 Comparison of Frequency Lists

Table 1 presents the primary statistics of the obtained frequency dictionaries. Primarily, we note the distinctly different volumes of transcriptions – the expert transcription contains 302,681 words (tokens), the NTR Acoustic Model (AM) produced 274,305 words, and the OpenAI Whisper Model (WM) transcribed only 236,179 words. The WM contains the fewest words because it appears to clean the text from repetitions, hesitations, and other “garbage” elements of spontaneous speech.

Table 1. General statistics of transcripts obtained.

Transcription	Tokens	Types
Expert transcript	302,681	31,311
AM	274,305	36,846
WM	236,179	28,060

The dictionary volume (the number of different types of words) also varies slightly. It has maximum for the Acoustic Model and minimum for the Whisper Model, which is trained on internet content. As a result, the AM seems to “hear” more variety, while the WM, as a variant of a large language model, only uses high-probability vocabulary.

The Spearman rank correlation coefficient shows high values, indicating a strong agreement between the models and the expert transcriptions. The correlation coefficients are 0.78 for the Expert Model (EM) and Acoustic Model (AM), 0.84 for the EM and Whisper Model (WM), and 0.78 between the AM and WM. This implies that the Whisper system aligns more closely with the expert transcription in terms of ranked word frequencies than the Acoustic Model does.

Table 2 showcases the top 25 words for the generated frequency dictionaries. Normalized frequency is given in ipm (items per million).

The word ‘*ja*’ (‘*I*’) turned out to be the most frequently used across all models. However, its occurrence is slightly higher in the WM transcriptions, potentially due to the model’s propensity to automatically complete incomplete sentences, such as changing ‘*ushla*’ (‘*have left*’) to ‘*ja ushla*’ (‘*I have left*’). The same pattern can be observed in the recognition of the second most frequent personal pronoun ‘*ty*’ (‘*you*’), as well as ‘*on*’ (‘*he*’), and ‘*ona*’ (‘*she*’).

Both models have difficulties recognizing the discourse word ‘*nu*’ (‘*well*’). This could be attributed to the fact that it commonly occurs at the beginning of sentences and is frequently reduced to the consonant [n]. Since it doesn’t carry any lexical meaning, it might be overlooked by the language model. The models also often struggle to identify accurately ‘*vot*’ (‘*well*’/‘*here you go*’), another commonly used pragmatic marker in everyday spoken Russian. This difficulty could be attributed to the fact that ‘*vot*’ often undergoes phonetic reduction, turning into [ot] or even [ʌt]. Moreover, ‘*vot*’ often does not carry a specific lexical meaning, acting instead as a pragmatic marker [25], which could further complicate its recognition by the models.

Both models show a lesser frequency for the particle ‘*da*’ (‘*yes*’) compared to the expert transcription. This could be because this word is often used as a backchannel response, typically as an agreement during an ongoing conversation. As the models tend to disregard overlapping speech, this discrepancy arises. A similar trend is observed for other forms of backchanneling like ‘*ugu*’ (‘*uh-huh*’) and ‘*aga*’ (‘*aha*’). The particle/conjunction ‘*a*’ is in a similar situation. At the same time, both ASR models provides more occurrences of conjunction ‘*cto*’ (‘*what*’), which is more characteristic for written and official language.

Thus, it can be seen that the frequency dictionaries differ significantly not for all words, but only for some, which the models “hear” more or less often compared to the expert transcription. Let’s consider the most frequent words of this kind.

Table 2. Top 25 words in transcripts obtained.

Expert transcript			Acoustic Model			Whisper Model		
Rank	Word	Norm.Fr	Rank	Word	Norm.Fr	Rank	Word	Norm.Fr
1	<i>ja</i>	26,999	1	<i>ja</i>	26,372	1	<i>ja (I)</i>	28,212
2	<i>nu</i>	26,946	2	<i>ne</i>	24,294	2	<i>ne</i>	25,366
3	<i>ne</i>	24,527	3	<i>nu</i>	23,889	3	<i>chto</i>	22,297
4	<i>da</i>	23,619	4	<i>chto</i>	22,016	4	<i>nu</i>	21,793
5	<i>a</i>	21,353	5	<i>vot</i>	20,889	5	<i>da</i>	20,057
6	<i>vot</i>	21,353	6	<i>da</i>	19,566	6	<i>v</i>	19,223
7	<i>chto</i>	21,207	7	<i>i</i>	18,308	7	<i>a</i>	18,888
8	<i>i</i>	19,189	8	<i>v</i>	16,985	8	<i>i</i>	17,885
9	<i>v</i>	18,670	9	<i>a</i>	16,886	9	<i>vot</i>	17,203
10	<i>to</i>	16,598	10	<i>to</i>	16,788	10	<i>to</i>	16,360
11	<i>eto</i>	16,503	11	<i>eto</i>	16,336	11	<i>eto</i>	16,051
12	<i>tam</i>	14,348	12	<i>na</i>	13,496	12	<i>tam</i>	13,079
13	<i>u</i>	12,422	13	<i>tam</i>	13,405	13	<i>na</i>	12,673
14	<i>na</i>	12,300	14	<i>u</i>	11,349	14	<i>u</i>	12,643
15	<i>tak</i>	11,140	15	<i>tak</i>	11,053	15	<i>tak</i>	10,335
16	<i>vse</i>	10,566	16	<i>kak</i>	10,368	16	<i>kak</i>	10,136
17	<i>kak</i>	10,354	17	<i>vse</i>	10,302	17	<i>ty</i>	9,815
18	<i>ty</i>	9,773	18	<i>ty</i>	8,585	18	<i>s</i>	7,994
19	<i>s</i>	8,471	19	<i>s</i>	8,046	19	<i>on</i>	7,837
20	<i>on</i>	7,572	20	<i>on</i>	6,854	20	<i>vse</i>	7,350
21	<i>net</i>	6,512	21	<i>net</i>	6,526	21	<i>net</i>	7,283
22	<i>ona</i>	6,168	22	<i>ona</i>	6,070	22	<i>ona</i>	6,982
23	<i>ugu</i>	6,105	23	<i>mne</i>	5,968	23	<i>mne</i>	5,860
24	<i>mne</i>	5,676	24	<i>est'</i>	5,702	24	<i>est'</i>	5,780
25	<i>est'</i>	5,663	25	<i>po</i>	5,282	25	<i>men'a</i>	5,064

3.2 Words that Are Least Recognized

Words that are poorly recognized by the models are intriguing because they are frequently omitted during automatic recognition. Analyzing the lists of the least recognized frequent

words showed a significant overlap for both models. The following groups of words were found to be poorly recognized:

- 1) Discursive words and pragmatic markers, along with their components: “nu” (well), “vot” (well, you know), “govorit” (he/she says), “govorju” (I say), “koroche” (in short/basically), “tipa” (like), “eto samoe” (you know), “v obschem” (in general), etc.
- 2) Backchannel responses: “ugu” (uh-huh), “aga” (yeah), “da” (yes).
- 3) Interjections: “oj” (oh), “blin” (damn), “b...d” (f...k), “akh” (ah), “ekh” (eh), etc.
- 4) Conversational forms of literary words: “naverno” (probably), “chtob” (in order to), etc.
- 5) Hesitations and unfinished words: “m” (um), “n” (un), “e” (er), etc.

All these speech elements characterize informal, everyday verbal communication, particularly in a dialogic form. These results for Whisper can likely be attributed to the lack of substantial training data on everyday dialogic speech, resulting in a more formal and “literary” quality of the generated transcriptions. Moreover, for both models, the difficulty in recognizing these words may arise from their occurrence in weak positions—at the beginning or end of phrases, amid the speech of the interlocutor (especially for backchannelling), and when they serve rhythmic functions [26],—leading to their weaker acoustic representation as they lack substantial semantic content. The exclusion of these speech elements by the systems does not have a significant impact on the core content of the utterances but does lead to a loss of the specific colloquial nuances in the transcriptions of oral speech.

Furthermore, consistent differences in the performance of each system are observed. The recognition errors in the Acoustic Model (AM) can be explained by the strong reduction of frequent words in real speech, such as “ty” (you), “on” (he), “tam” (there), “ili” (or), “teb’a” (you), “vidish” (see), and similar words. An interesting characteristic of Whisper is the lack of verbal representation for numerals (represented as numbers) and the limited, albeit irregular, use of the letter “Ė” (jo), which accounts for the reduced occurrence of words like “vsjo” (everything), “*jeshche*” (still/yet), “jejo” (her).

Table 3 provides two lists of the top 25 words/tokens with the most significant discrepancies in frequency compared to the expert transcription. The lists are ordered by decreasing differences in relative frequency (ipm), and they also include information on differences in rank orders and relative differences expressed in %.

4 Words Not Found in the Expert Transcription

Lastly, the words that were “recognized” by a specific model but are missing in the expert transcription are of special interest. The Table 4 shows the top 25 most frequent words in this category.

The NTR Acoustic Model provides a fairly extensive list of frequent tokens that are not present in the expert transcription. In the vast majority of cases, these are reduced forms of common words, such as “govor(yu)/(it)” (I say/he/she says), “(poni)maesh” (you understand), “ponyat(no)” (I see), “slusha(y)” (listen), “chelove(k)” (man/person), “(nor)mal’no” (normal), etc. For example, “govor” is a shortened form of “govoryu”/“govorit” (I say/he/she says):

- *Kuda on vyshla?* (Where did she go?)
- *Mam, ya govoryu, po delam ochevidno* (Mom, I'm speaking to you, obviously, on business). (ordS72–11¹)

Table 3. Frequent words worst recognized by models.

Acoustic Model					Whisper Model				
#	Word	Rank dif	IPM dif	Relat dif	#	Word	Rank dif	IPM dif	Relat dif
1	<i>a</i>	-4	4466	21	1	<i>nu</i>	-2	5153	19
2	<i>da</i>	-2	4053	17	2	<i>ugu</i>	-51	4382	72
3	<i>nu</i>	-1	3056	11	3	<i>vot</i>	-3	4149	19
4	<i>ugu</i>	-14	2405	39	4	<i>da</i>	-1	3562	15
5	<i>m</i>	-112	1917	73	5	<i>vsjo</i>	-4	3215	30
6	<i>v</i>	1	1685	9	6	<i>a</i>	-2	2464	12
7	<i>ty</i>	0	1187	12	7	<i>m</i>	-206	2244	86
8	<i>u</i>	-1	1074	9	8	<i>jeshche</i>	-9	1347	29
9	<i>tam</i>	-1	944	7	9	<i>i</i>	0	1304	7
10	<i>i</i>	1	880	5	10	<i>tam</i>	0	1269	9
11	<i>k</i>	-22	723	30	11	<i>chego</i>	-46	984	48
12	<i>on</i>	0	719	9	12	<i>zh</i>	-134	902	69
13	<i>n</i>	-1018	671	92	13	<i>tak</i>	0	805	7
14	<i>aga</i>	-30	629	33	14	<i>oy</i>	-37	782	38
15	<i>ya</i>	0	627	2	15	<i>n</i>	-18306	729	99
16	<i>oy</i>	-20	532	26	16	<i>dvadtsat'</i>	-307	695	78
17	<i>vot</i>	1	463	2	17	<i>aga</i>	-34	667	35
18	<i>s</i>	0	425	5	18	<i>govorit</i>	-16	582	22
19	<i>zh</i>	-37	419	32	19	<i>pyat'</i>	-111	565	58
20	<i>fon</i>	-388	414	76	20	<i>veroyatno</i>	-18301	534	99
21	<i>chtoby</i>	-143	380	57	21	<i>chtoby</i>	-448	533	80
22	<i>verojatno</i>	-221	345	64	22	<i>jejo</i>	-29	511	32
23	<i>e</i>	-3047	345	94	23	<i>t</i>	-1508	502	92
24	<i>o</i>	-28	336	23	24	<i>s</i>	1	477	6
25	<i>ili</i>	-10	330	14	25	<i>tri</i>	-61	455	43

¹ Hereinafter, the code of the macro-episode of the ORD corpus.

It is worth noting that among the frequent tokens, there are forms that can be traced back not to one but to several completely different words. For instance, “mas” may stand for “maslo” (butter/oil), “master” (master), and even “most” (bridge):

- *Ya, naprimer, sama ne pokupayu natural(noe), ya pokupayu rastitel(noe) mas(lo)* (For example, I myself don't buy natural, but vegetable oil) (ordS11–06).
- *Gde vot eta to ulitsa vykhodit na Liteynyy, kakaya ot tsirka tam cherez mas (=most) perezhaesh'* (Where is this street connecting to Liteynyy prospect, the one from the circus, where you cross the bridge) (ordS84–22).

The compiled lists of real words' transcriptions along with their corresponding orthographic representations can serve as a valuable resource for creating a dictionary of reduced forms. This dictionary can showcase the various alterations and distortions that occur in words during spontaneous speech. Such a resource is of great significance for conducting phonetic research on spoken language and for enhancing the training of language models.

Regarding the Whisper model, the frequency list is dominated by words containing the letter “Ě” (yo). As mentioned earlier, this recognition model incorporates this letter, unlike the acoustic model. However, Whisper uses “Ě” irregularly. For example, the word “*jeshcho*” (yet/still) with this letter appeared in the transcriptions only half as often as without it (337 vs. 761). It's worth noting that in the expert transcription used for the study (as well as the version for the ORD website), all instances of “Ě” were replaced with “E” despite the original ORD corpus, which includes it.

Furthermore, Whisper demonstrates good English language comprehension and recognizes the names of many brands and computer terms, such as *Photoshop*, *USB*, *Microsoft Office*, *Wi-Fi*, which were originally written in Russian in the expert transcription. These words have also been included in this list.

From Table 4, it is also evident that among the frequent words recognized by Whisper, there are words that do not actually exist in the language (e.g., “shvejn”, “vodonyuchka”). However, upon examining the transcriptions, it becomes apparent that these occurrences should be considered as system glitches—these words are repeatedly present just in a single file, and the transcription of that file has little resemblance to reality.

Let's take, for example, the usage of the word “vodonyuchka”. All instances of this word are related to a single episode. This recording represents a noisy speech of a man addressed to his car, which is experiencing some technical issues. In the original file, there were significant pauses between the utterances, which were replaced with 2 ms pause insertions. The system incorrectly recognized two initial words of the recording—“*batareechka*” and “*batareyka*” (both meaning “battery”)—as “*vodonyuchka*” (no sense) and then erroneously “recognized” this new word in 22 other places:

Expert transcript:—*Batareechka, batareyka. Ya tebya ponyal. Vse. Tu tu tu. [Profanity]... Nu nakonets-to! (Little battery, battery. I understood you. That's all. Tu tu tu. [Profanity]. Finally!)* (ordS74–02).

Whisper Model transcript:—*Vodonyuchka, vodonyuchka. Ya tebya ponyal. (I understood you) Vodonyuchka, vodonyuchka. Vodonyuchka, vodonyuchka. [Profanity]... Vodonyuchka, vodonyuchka* (ordS74–02).

Another glitch in the system was identified during the analysis of the word “shveyn”. In this instance, Whisper generated a transcription fragment with a repetitive loop of the

Table 4. Words Not Found in the Expert Transcription.

Acoustic Model				Whisper Model			
#	Word	Frequency	IPM	#	Word	Frequency	IPM
1	<i>b...d'</i>	29	106	1	<i>vs'o</i>	725	3070
2	<i>govor</i>	24	87	2	<i>jeshcho</i>	337	1427
3	<i>mas</i>	21	77	3	<i>jejo</i>	116	491
4	<i>boy</i>	17	62	4	<i>ona</i>	89	377
5	<i>mayesh'</i>	16	58	5	<i>cho (чѐ)</i>	82	347
6	<i>ponyat</i>	15	55	6	<i>id'ot</i>	67	284
7	<i>mer</i>	14	51	7	<i>shvejn*</i>	61	258
8	<i>uga</i>	12	44	8	<i>ru</i>	41	174
9	<i>slushayu</i>	10	36	9	<i>kub</i>	40	169
10	<i>dtsat'</i>	9	33	10	<i>vodonyuchka*</i>	22	93
11	<i>mala</i>	9	33	11	<i>prishol</i>	22	93
12	<i>stav</i>	9	33	12	<i>chom</i>	22	93
13	<i>glyad'</i>	8	29	13	<i>poyd'om</i>	20	85
14	<i>yer</i>	8	29	14	<i>poyd'ot</i>	19	80
15	<i>sen</i>	8	29	15	<i>dolzhen</i>	19	80
16	<i>khom</i>	8	29	16	<i>prichom</i>	18	76
17	<i>chelovek</i>	8	29	17	<i>ber'ozy</i>	15	64
18	<i>agu</i>	7	26	18	<i>nashol</i>	14	59
19	<i>shay</i>	7	26	19	<i>r</i>	13	55
20	<i>gal</i>	6	22	20	<i>l'ova</i>	13	55
21	<i>ed'e</i>	6	22	21	<i>mojo</i>	13	55
22	<i>krasno</i>	6	22	22	<i>poshel</i>	12	51
23	<i>lega</i>	6	22	23	<i>prid'ot</i>	11	47
24	<i>mali</i>	6	22	24	<i>savushkinu</i>	11	47
25	<i>mal'no</i>	6	22	25	<i>zhivjot</i>	10	42

phrase “ya ponyal, chto dlya shveyn, a v remonte tozhe cherez nikh” (I understood that for shveyn, and in repairs too through them – quite a meaningless phrase), which was repeated 61 times and has nothing to do with the original transcript.

“Vodonyuchka” and “shvejn” repetition cases, along with any other phrase repetitions, represent typical instances of degeneracy observed in texts generated by large language models [27]. If a word appears in the text for any reason, the probability of the model generating it again often increases. This phenomenon explains the repetition of peculiar words in the text. Hence, “vodonyuchka” was an initial recognition error, and then the model degeneratively propagated it through repetition. These examples

highlight the need for a preliminary filtering process to exclude files that exhibit such malfunctions, ensuring a more accurate comparison of frequency dictionaries.

Additionally, another distinctive feature of Whisper is its active use of colloquial forms, such as “chyo” (what), while the expert transcription utilizes the word “chto” in the same context. For example:

— *Chyo, pryam na Ladogu poedesh’?... Nu chyo tam khoroshego?... Kamney net, ogromnykh etikh valunov, a tak ya ne znayu, chyo tam khoroshego v etoy Ladoge...* (What, are you going straight to Ladoga lake?... Well, what’s good there?... There are no big rocks, those huge boulders, but I don’t know what’s good about Ladoga...).

— *Chyo ty nesyosh’?* (What are you talking about?)

— *A chyo?* (And what?)

— *Chyo ty nesyosh’?* (What are you talking about?) (ordS88–03)

Moreover, both ASR models utilize the colloquial form of greeting “zdraste” (hello), which was not employed by the experts in their transcriptions. As a result, some of the discrepancies between the ASR transcripts and the expert model can be attributed not to deficiencies in these systems but rather to the variability in orthographic representation (“zdraste” instead “zdravstvujte”), as well as the usage of the letter “Ë” (yo).

5 Conclusions

The study utilized two ASR systems with distinct characteristics: 1) OpenAI Whisper, a multilingual language model with acoustic input, which exhibits extensive knowledge of languages. Its transcriptions tend to be excessively literary and contain more written language features than the original audio, and 2) NTR Acoustic Model, lacking language understanding but acting as a transcription tool by transcribing exactly “what it hears”, without considering spelling accuracy. Other ASR systems for the Russian language would likely yield results falling between those presented.

The research showed that each of the considered systems has its own peculiarities and regular recognition errors. On the lexical level, certain groups of words are poorly recognized by both systems, such as discursive words, pragmatic markers, backchannel responses, interjections, conversational reduced word forms, and hesitations. The challenging recognition of these words can be attributed to various factors, ranging from the lack of sufficiently representative training corpora of spontaneous dialogues to the presence of these words in prosodically weak positions (e.g., used for rhythmic functions). The omission of these speech elements by ASR systems generally does not have a significant impact on the overall comprehension of the message conveyed in the utterance. However, it does lead to the transcripts lacking the specific colloquial nuances of spoken language.

Some differences between ASR transcripts and the expert model are not due to shortcomings in the recognition systems but stem from the variability in the orthographic representation of spoken words and the specific nuances of their pronunciation in each context. Considering this aspect, the WER values mentioned in Sect. 2.2 might be somewhat overstated.

Upon comparing the overall performance of the systems, the Whisper model generally produces more “easily readable” text, resembling written language, and requires less manual correction for most practical purposes (e.g., for publishing interviews recorded on a dictaphone). However, some manual correction remains necessary as the transcripts may contain disruptions like word repetitions and even looped phrases and sentences. Furthermore, when using Whisper, it is important to take into account the likelihood of degeneracy during text generation by large language models. If a word appears in the text for any reason, the probability of the model generating it again is heightened. This phenomenon explains the repetition of peculiar words in the text.

On the other hand, the transcripts generated by the NTR Acoustic Model are less familiar to educated readers, but they more accurately reflect the pronunciation peculiarities of words and phrases. The frequency lists of words obtained with the NTR model, along with references to their orthographic representation, can serve as valuable material for constructing a dictionary of reduced forms, showcasing word distortions in spontaneous speech. This provides essential material for phonetic studies of oral speech and further ASR model training designed for everyday conversations.

Acknowledgements. This article is an output of a research project “Text as Big Data: Modeling Convergent Processes in Language and Speech using Digital Methods” implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

References

1. Ali, A., Renals, S.: Word error rate estimation for speech recognition: e-WER. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Vol. 2: Short Papers. Melbourne, Australia. Association for Computational Linguistics, pp. 20–24 (2018). <https://aclanthology.org/P18-2004.pdf>
2. AntConc. <http://www.laurenceanthony.net/software.html>. Accessed 1 Sept 2023
3. Asinovsky, A., Bogdanova, N., Rusakova, M., Stepanova, S., Ryko, A., Sherstinova, S.: The ORD speech corpus of Russian everyday communication “One Speaker’s Day”: creation principles and annotation. In: Matoušek, V., Mautner, P. (eds) Text, Speech and Dialogue. TSD 2009. Lecture Notes in Computer Science, vol 5729, pp. 250–257. Springer, Berlin, Heidelberg (2009)
4. Bakhturina, E., Lavruxhin, V., Ginsburg, B.: A toolbox for construction and analysis of speech datasets. Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks, **1** (2021)
5. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Ermolova, O., Baeva, E., Martynenko, G., Ryko, A.: Sociolinguistic extension of the ORD corpus of Russian everyday speech. Ronzhin, A., et al. (eds.), *SPECOM 2016*, Lecture Notes in Artificial Intelligence, LNAI, **9811**, pp. 659–666. Springer, Switzerland (2016)
6. Campbell, N.: Speech & expression; the value of a longitudinal corpus. *LREC* **2004**, 183–186 (2004)
7. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint [arXiv:1904.10509](https://arxiv.org/abs/1904.10509) (2019)
8. Gulati, A., et al.: Conformer: convolution-augmented transformer for speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 5036–5040 (2020)

9. Hellwig, B., Van Uytvanck, D., Hulsbosch, M., et al.: ELAN — Linguistic Annotator. Version 4.9.3 [in:] <http://tla.mpi.nl/tools/tla-tools/elan/> Linguistic Annotator ELAN. <https://tla.mpi.nl/tools/tla-tools/elan/>
10. Hendrycks, D., Gimpel, K.: Gaussian Error Linear Units (gelus). arXiv preprint [arXiv:1606.08415](https://arxiv.org/abs/1606.08415) (2016)
11. https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_ru_conformer_ctc_large
12. <https://github.com/sovaai/sova-dataset>
13. <https://voice.mozilla.org>
14. Karpov, N., Denisenko, A., Minkin, F.: Golos: Russian dataset for speech research. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2, 1076–1080 (2021). <https://doi.org/10.21437/Interspeech.2021-462>
15. Kolobov, R., et al.: Mediaspeech: Multilanguage ASR Benchmark and Dataset. arXiv, [arXiv:2103.16193](https://arxiv.org/abs/2103.16193) (2021)
16. Morris, A.C., Maier, V., Green, P.: From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In: Eighth International Conference on Spoken Language Processing, pp. 2765–2768 (2004). https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2004/i04_2765.pdf?ref=https://githubhelp.com
17. One Speech Day corpus online. <https://ord.spbu.ru/>
18. Press, O., Wolf, L.: Using the output embedding to improve language models. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 157–163, Valencia, Spain, April 2017. Association for Computational Linguistics (2017)
19. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning, in Proceedings of Machine Learning Research 202, 28492–28518 (2023)
20. Reference Guide for the British National Corpus. <http://www.natcorp.ox.ac.uk/docs/URG.xml>
21. Sherstinova, T.: Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus. Ronzhin, A., et al. (eds.) SPECOM 2015. Lecture Notes in Artificial Intelligence, LNAI. Vol. 9319. Springer, Switzerland, pp. 268–276 (2015)
22. Sherstinova, T.: The structure of the ORD speech corpus of Russian everyday communication. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS (LNAI), vol. 5729, pp. 258–265. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04208-9_37
23. Von Neumann, T., Boeddeker, C., Kinoshita, K., Delcroix, M., Haeb-Umbach, R.: On word error rate definitions and their efficient computation for multi-speaker speech recognition systems. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1–5 (2023). <https://ieeexplore.ieee.org/abstract/document/10094784/>
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5999–6009 (2017)
25. Bogdanova-Beglarian, N., Blinova, O., Sherstinova, T., Troshchenkova, E.: Russian pragmatic markers database: developing speech technologies for everyday spoken discourse. In: 26th Conference of Open Innovations Association FRUCT, FRUCT 2020; Yaroslavl; Russian Federation, pp. 60–66 (2020)
26. Bogdanova-Beglarian, N., Sherstinova, T., Kisloshchuk, A.: On the rhythm-forming function of discursive units. Perm University Herald. Russian and Foreign Philology 2(22), 7–17 (2013)
27. Holtzman, A., et al.: The curious case of neural text degeneration. In: Proceedings of the 2020 International Conference on Learning Representations, p. 2540 (2020)