Everyday Speech of Russian Students: From Field Recordings to the ESC Speech Corpus (Data Processing and Representation)

Tatiana Sherstinova
HSE University-Saint Petersburg, Russia
tsherstinova@hse.ru

Irina Petrova
HSE University-Saint Petersburg, Russia
irina.petrova1011@gmail.com

Aleksei Melnik
HSE University-Saint Petersburg, Russia
melnik-a@russian-short-stories.ru

Karina Azarevich
HSE University-Saint Petersburg, Russia
kiazarevich@edu.hse.ru

Sofiia Chepovetskaia
HSE University-Saint Petersburg, Russia
soncheroz1503@gmail.com

Viktoria Deborina
HSE University-Saint Petersburg, Russia
vddeborina@edu.hse.ru

Abstract

This paper presents the current state of the ESC corpus, which is the largest collection (more than 2,000 hours of recordings) of spontaneous recordings of everyday Russian speech obtained in natural communicative environments. The data are collected using a long-duration speech recording methodology: volunteers record their everyday communication over extended periods of time, typically during a full "speech day", using professional audio recorders or high-quality mobile devices. The corpus data document the current state of youth everyday speech, including the active use of neologisms, emerging colloquial vocabulary, and elements of youth slang. The paper provides updated statistics on the current stage of corpus development and focuses on the methodological and technological aspects of corpus construction. In particular, the processing pipeline relies on diarization and automatic speech recognition systems for generating initial transcripts of the recordings. Another important stage of speech data preparation is the anonymization of personal information, as a substantial portion of everyday conversational recordings contains private and sensitive content. The ESC corpus can serve as a testbed for evaluating NLP tools for Russian. The resulting resource is expected to be valuable both for fundamental research on contemporary Russian spoken discourse and youth communicative practices, and for applied tasks such as the development and training of automatic speech recognition systems, language models, and other natural language processing technologies.

Keywords: corpus linguistics, spoken Russian, everyday speech, youth speech, speech corpora, speech technologies, speech data anonymization, oral discourse, speech processing.

1 Introduction

Everyday spoken speech is a truly unique type of linguistic material. It is the form of language we use in most communicative situations, helping us deal with everyday tasks both in private life and in professional settings. Its production does not require paper, writing tools, or any digital devices. This mode of communication has existed at all times, linking us with our distant ancestors and, most likely, remaining an essential means of communication for future generations as well.

Paradoxically, however, we know much less about spoken speech than about written language. Most grammars, dictionaries, and linguistic statistics are based on written texts, while only a relatively small number of studies dealing with spoken data rely on so-called "laboratory" speech — that is, language material recorded under

specific communicative conditions (in recording studios, over the phone, or in interview settings). The properties of such data differ considerably from those of natural spontaneous speech. As a result, everyday spoken language still lacks a sufficiently detailed theoretical and statistical description.

Despite the fact that in recent decades more spoken corpora have been developed (see, for example, (CLARIN 2026; Du Bois 2000; BNC 2007)), both the overall number of such resources and the volume of audio data they contain remain clearly insufficient for most of the world's languages. A notable exception is Japanese, where everyday language use has been the focus of linguistic research since the 1940s (Shibamoto 1987; Shibata 1983; Campbell 2004).

Until recently, the largest source of authentic spoken data for Russian has been the "One Day of Speech" corpus (ORD), recorded in 2007–2016 in Saint Petersburg using a continuous speech recording methodology. The corpus reflects everyday communication of 128 informants representing different social groups of the urban population—across gender, age, professional, and social status categories—which has made it possible to carry out a wide range of phonetic, lexical, grammatical, pragmatic, and sociolinguistic studies based on these data (Asinovsky et al. 2009; Bogdanova-Beglarian et al. 2016). However, language, as a living system, is constantly evolving, and material recorded ten years ago can no longer fully reflect its current state. This is especially true of youth speech.

This paper presents a new audio resource of everyday Russian speech, focused specifically on the speech of young people and students. It provides updated statistics on the current stage of corpus development and highlights key methodological and technological aspects of its construction.

2 ESC Speech Corpus: Methodology and Stages of Development

The ESC corpus of youth and student speech is a spoken corpus designed to document and study everyday Russian based on authentic spontaneous speech used in both private and professional communication (Sherstinova and Petrova 2024; Sherstinova et al. 2025). The development of the ESC corpus began in March 2023. It is based on the continuous speech recording methodology originally developed for the "One Day of Speech" (ORD) corpus (see above). Its key feature, however, is the focus on youth speech: the corpus primarily includes recordings of speakers aged 18–27. This age group — socially active and sensitive to current trends — makes it possible to capture the most up-to-date state of the language at a given point in time.

At the initial stage of the project, the main goal was to establish a stable workflow for continuous data collection. Over the past three years, a recording process has been set up that allows the corpus to be steadily expanded with new audio data. Informants participating in the project complete a set of documents, including: (1) informed consent to participate in the study; (2) consent to the processing of personal data; (3) a sociological questionnaire; (4) a speech day diary; (5) a psychological test; and (6) a passive vocabulary test.

For recording, informants are provided with professional audio devices — typically Zoom H1n or Tascam DR-05X recorders (output format: WAV, stereo, 16-bit, 44,100 Hz), which they carry with them throughout their regular day.

At present, the ESC corpus contains over 2,000 hours of original recordings collected from 165 participants, including 122 women and 43 men. The statistics as of early April 2026 are presented in Table 1. Thus, in terms of overall volume, the ESC corpus has already surpassed its predecessor, the ORD corpus, which contains approximately 1,500 hours of audio recordings.

Gender	**Informants (N, %)**		**Recording Hours (N, %)**	
Female	122	73.94%	1574	77.81%
Male	43	26.06%	449	22.19%

Table 1: Gender Distribution of ESC Recordings.

A noticeable gender imbalance can be observed, largely due to the conditions under which the ESC corpus has been developed. The project was initiated at the Faculty of Philology of HSE University, where the majority of students are women. The university setting accounts for the strong predominance of recordings from student informants aged 18–24 (85.45%). The average age of participants is 21.7 years.

For the vast majority of informants, the current place of residence is Saint Petersburg and the

Leningrad region (89.09%), where the recordings are collected. At the same time, geographic diversity is ensured by the fact that more than half of the informants (54.55%) indicated other regions as their place of birth. In addition, about 5% of participants are citizens of other countries (Kazakhstan, Belarus, Kyrgyzstan, Moldova, Iran, and Uzbekistan). Russian is the native language for almost all informants (97.58%). The corpus also includes bilingual student informants whose native languages are Belarusian, Moldovan, and Persian.

In addition to using professional audio recorders, participants were also allowed to use their smartphones as recording devices. This option was chosen by 32 informants (19.39%), including 8 who made recordings both with their own devices and with the recorders provided for all informants. Table 2 presents the distribution of recording devices.

Recording Method	Informants (N)	Average Recording Duration
Audio recorder	133	10:37:25
Smartphone	24	27:35:58
Both devices	8	16:27:13
Total	165	12:15:33

Table 2. Recording methods used by informants

On average, about 12 hours[1] of recordings were obtained per informant. The use of a smartphone as a recording device tends to increase the total duration of collected audio: up to 27.5 hours when recording only on a phone and up to 16.5 hours when combining both recording options. This difference in average values is mainly explained by the limited time for which audio recorders are provided to participants.

Informants may take part in the recording more than once. This option was used by 8 informants (half of whom used both types of devices). These participants account for 18.81% of the total recording time currently available in the ESC corpus. The largest amount of data obtained from a single informant is 134 hours of recordings, which corresponds to approximately 11 average "speech days"[2].

The second stage after collecting the audio data is its processing. At this stage, long audio files, with an average duration of about four hours, are reviewed by experts and manually segmented into macro episodes, which are "large segments unified by the place of communication, its conditions, and participants" (Sherstinova 2015) typically lasting 15–40 minutes. These macro episodes become the main unit of storage and description in the corpus.

During segmentation, extended fragments that do not contain speech are removed. In addition, since the primary focus is on "live" unprepared speech, fragments containing "secondary" (recorded or mediated) speech — such as background radio or film audio, or public transport announcements — may also be excluded if they are not involved in the ongoing interaction, that is, if they do not elicit any response or commentary from the informant and do not influence his/her speech behavior.

Thus, a typical 12-hour "speech day" of an informant, originally consisting of three four-hour audio files, is represented in the corpus as a set of *n* episodes, each accompanied by a brief description of the recorded situation. This description includes: (1) the type of communication; (2) the social roles and relationships of the speakers; (3) the location; and (4) a brief outline of each episode.

The episodes are then automatically transcribed using a pipeline that includes Whisper v3 for automatic speech recognition (ASR) and *pyannote* for diarization, i.e., the identification and separation of different speakers' utterances. The use of automated speech processing tools significantly speeds up the workflow. However, the quality of automatic transcription for recordings made in natural (often quite noisy) conditions remains imperfect, so the ASR output is subsequently reviewed and corrected by experts, being the most labor-intensive stage in corpus construction.

Before being published on the corpus website, the transcripts undergo an additional stage of data anonymization: all personal and sensitive information is replaced and encoded.

In this paper, we focus in particular on several aspects of corpus construction, including the use

[1] This figure reflects the total duration of audio recordings obtained from informants. Naturally, the amount of "actual speech" is significantly smaller and varies across informants.

[2] A speech day is defined as a continuous recording lasting between 8 and 16 hours that captures all of the informant's speech activity over the course of a day.

of speech recognition systems for generating initial transcripts, issues related to diarization, as well as the need for personal data anonymization and new approaches to morphological annotation.

3 Speech Diarization and ASR

To evaluate the quality of speech recognition for recordings made in real field settings, characterized by high levels of noise and overlapping speech, a series of experiments on speech diarization and ASR for spontaneous spoken language was conducted. These experiments assessed the performance of automatic diarization tools, i.e., methods for identifying and separating speakers in a speech stream. The selected models and associated tools were subsequently integrated into the overall audio data processing pipeline of the corpus. The analysis involved comparing the recognized text, automatically generated timestamps, and speaker labels with corpus transcripts produced manually by experts.

Since the ESC corpus does not yet contain a sufficiently large amount of expert-annotated speech transcripts, it was decided to conduct the initial experiments on data from the ORD corpus, which serves as a prototype and for which a substantial amount of expert-annotated transcripts is available. The reference dataset included 195 episodes of everyday speech recordings collected using the "speech day" methodology.

The expert annotation of the reference dataset was organized as a table in which each utterance is associated with a unique speaker identifier, a brief description of the episode, the name of the audio file, the start time, and the duration of the utterance. The transcripts include markers of syntagmatic and phrasal segmentation, as well as indications of pauses and paralinguistic phenomena (Bogdanova-Beglarian et al. 2016). Overlapping speech is represented within a single utterance, with multiple speakers marked using the symbols "@" or "#" (Asinovsky et al. 2009).

In contrast, the automatically generated data include speaker labels, along with start and end timestamps for each segment. Differences between the expert and automatic annotations concern speaker identification, timestamp format, and the representation of overlapping speech: in diarization output tables, speaker identifiers are reused within each audio file, timestamps indicate the beginning and end of each segment, and overlapping speech is represented as separate utterances with overlapping time intervals.

For evaluation purposes, the annotation formats were unified in order to eliminate differences that do not affect semantic content but may distort the metrics when comparing hypotheses and references. Thus, all punctuation was removed, all words were converted to lowercase, the letter "ё" was replaced with "е", and numerals (both in digits and written out) were replaced with the marker "NUM". At the same time, during the conversion of speaker labels, there was a risk of errors that could lead to incorrect calculation of evaluation metrics. Overlapping speech in the expert annotation was not transformed, as it is not possible in such cases to extract separate timestamps for individual speakers. In addition, occasional spelling errors are present in the reference annotation. These risks were minimized through additional testing of the processing algorithms and manual verification of aligned word pairs from the expert and automatic annotations.

The audio data processing pipeline for evaluation consisted of four stages: 1) normalization, 2) diarization and segmentation of audio into fragments, 3) speech recognition, and 4) metric calculation.

First, the audio files were converted to a unified format and resampled to the frequency required for ASR (16 kHz), and their volume levels were normalized. Next, diarization was performed to identify speakers and segment the audio accordingly. The resulting segments were saved as separate files and passed to the speech recognition model. Finally, the obtained transcripts were used to compute evaluation metrics for both diarization and recognition quality.

Standard metrics for evaluating diarization and speech recognition quality were used (Karpov, Kipyatkova 2012): DER (Diarization Error Rate), DWER (Diarization Word Error Rate), SAWER (Speech Attributed Word Error Rate), and WER (Word Error Rate) (Ali and Renals 2018; Munteanu et al. 2006). The WER, DWER, and SAWER metrics are based on counting substitutions, deletions, and insertions of words, whereas DER evaluates diarization quality at the level of time segments.

For speech recognition, the Whisper Large v3 model was used. Its performance and comparison with acoustic model developed by NTR in 2023

(NTR) were previously reported in (Sherstinova, Mikhaylovskiy, Kolpashchikova 2024). For diarization, we used *pyannote* (Bredin et al. 2020), an open-source framework whose models achieve DER values of 34.4 and 24 on the DIHARD and ETAPE datasets, respectively.

The results of DER calculations for the reference and test datasets show considerable variation in diarization quality. The minimum DER value was 10.3, while the maximum values approach 100. At the same time, for 10% of the recordings, the diarization error does not exceed 27.9. The mean DER across the dataset is also relatively high (68.3), with a median of 52.6. As expected, the best and most stable results are obtained for recordings with simpler conditions: minimal background noise, no more than one or two speakers, and clear, sufficiently loud speech.

The mean values of WER, DWER, and SAWER range from 46.1 to 64, with median values between 46.2 and 62.9. For less than 10% of the recordings, WER is 27.4, DWER is 24.0, and SAWER is 30.6. The 90th percentile reaches 62.6 for WER, 81.1 for DWER, and 96.8 for SAWER, with maximum values approaching 100. Although the minimum values (ranging from 8.4 to 13.3) indicate that some high-quality recognition results are achieved, the system — similarly to diarization — shows substantial variability in performance.

At present, 869 macro-episodes of the ESC corpus have been processed using diarization and ASR models, and their expert review is ongoing.

4 Anonymization of Personal Data in Transcribed Speech

Since recordings of everyday speech typically contain a large amount of personal information about informants and their interlocutors, the next important stage in the development of the ESC corpus involved establishing principles for anonymizing textual data prior to their online publication, in accordance with international data protection standards and Russian legislation. As in studies devoted to the ORD corpus (Sherstinova 2019), anonymization of ESC transcripts addresses two main tasks: removing or masking personal information that can be used to identify a speaker, and encoding obscene language.

For the ESC corpus, seven categories of data subject to anonymization were identified, and abbreviations were introduced for each of them:

- PROPN — given names,
- SURN — surnames,
- PATR — patronymics,
- ORG — organization names,
- STD — names of educational institutions,
- PLACE — toponyms,
- OBSC — obscene language.

Additionally, a separate category, NON, was introduced to mark elements subject to anonymization that appeared in automatically generated transcripts but were later identified as recognition errors or model hallucinations (Holtzman et al. 2020) and were therefore removed from the final edited version of the transcript.

The anonymization procedure was carried out using three different approaches, depending on the category of the word or phrase to which it was applied:

4.1 Proper Names

The anonymization of proper names, including surnames and patronymics, requires special attention, as names and surnames are not only the most frequent but also the most sensitive type of personal information that can be used to identify speakers. When anonymizing real given names, surnames, and patronymics, substitutes were selected so that the rhythmic structure of the utterance remained unchanged (i.e., preserving the number of syllables and the position of stress within the word). Each anonymized element was marked with the symbol %:

Marina Leonidovna → *Karina*% *Eduardovna*%
Vladimir Il'ich → *Arkadiy*% *Kuz'mich*%
Gordeev → *Avdeev*%.

For each episode, a separate list was created containing anonymized given names, surnames, and patronymics, along with their original counterparts. This was necessary to preserve the recognizability of individuals mentioned in the conversation: for example, if a person named *Marina* is mentioned in an episode and anonymized as *Karina*%, the same substitute is used consistently throughout that episode. Ideally, this approach should also be extended across all episodes within a single "speech day."

4.2 Names of Organizations, Places, and Educational Institutions

Names of organizations, locations, and educational institutions were replaced with a general category label enclosed in angle brackets:

Vy vse iz Griboedova[3]? → *Vy vse iz* <*university*>?

In addition to the categories <*university*> and <*organization*>, other frequent replacements include <*street*>, <*address*>, and <*toponym*>, the latter covering various types of geographical entities, as well as <*group*> or <*name*> (e.g., student associations or film clubs).

4.3 Obscene Language

In words belonging to obscene language, the entire root or part of the root was replaced with asterisks (*), corresponding to the number of letters, for example: *bl*t'*, *s**a*, etc. This approach masks offensive language while preserving the possibility of recognizing these words.

4.4 Statistics of Anonymization Results

At present, 4,304 utterances from 38 episodes have been anonymized and verified. Among them, 364 unique utterances (8.6%) contain at least one sensitive element. This means that sensitive data occur in approximately every 11–12 utterances. In total, 619 units were anonymized. Table 3 presents the absolute frequencies of these units in the corpus, along with their percentage relative to the total number of such units (%).

Category	Frequency	%	ipm
PROPN	335	54.12	3938
OBSC	154	24.88	1810
PLACE	50	8.08	588
PATR	29	4.69	341
SURN	25	4.04	294
STD	15	2.42	176
ORG	6	0.97	71
NON	5	0.81	–

Table 3: Distribution of lexical items subject to anonymization.

Thus, the vast majority of elements requiring anonymization belong to the category of proper names. Moreover, if given names, surnames, and patronymics are considered as a single category, their combined share reaches 63%. It should also be noted that elements related to obscene language account for almost 25%, making this the second most frequent category. This is followed by the category of toponyms, which accounts for about 8% of all anonymized elements.

These data were obtained from a sample of 85,071 tokens, which allows us to estimate the *ipm* (*instances per million*) of these elements in the text and thereby assess the workload involved in expert verification of the anonymization process. The corresponding values are presented in the right column of Table 3.

Further work in this direction includes developing approaches to automate the anonymization and masking of sensitive information with minimal involvement of experts. In particular, these procedures should allow for the automatic generation of substitute given names, surnames, and patronymics that preserve the syllabic structure of the original utterance.

5 Corpus Website Functionality

The ESC corpus website is currently operating in test mode at https://esc-corpus.ru/. To develop and test its functionality, a set of 38 communicative macro-episodes has been selected, containing 12.85 hours of speech data produced by 59 speakers. This collection reflects the diversity of everyday life of students in Saint Petersburg, including personal, professional, and social interaction. The corresponding transcripts contain 59,162 word tokens across 4,601 utterances (Sherstinova et al. 2025).

At present, corpus search is implemented at two levels: (1) the macro-episode level, which allows filtering speech data by communication conditions, participants' social characteristics, and type of interaction, and (2) the transcript level. The macro-episode level makes it possible to retrieve conversations based on these parameters and to view their full transcripts from beginning to end (rather than only immediate context windows, as is common in corpus linguistics).

[3] *Griboedova* is an informal name for one of the buildings of HSE University in Saint Petersburg.

Standard search is currently implemented by substring and keyword. In the near future, search by lemmas and morphological features is planned. The audio recordings themselves are not publicly available, since spoken data contain a large amount of personal information, including biometric data.

6 Morphological annotation of spoken speech

Morphological annotation is an essential component of any linguistic corpus, as it enables the tagging of parts of speech and grammatical features. However, when working with spontaneous spoken speech, the use of parsers designed for written texts often leads to inaccurate results. Models trained on written data are oriented toward normative language use, whereas spoken speech widely includes discourse markers and functionally reinterpreted forms that do not conform to these expectations, which leads to reduced accuracy in morphological analysis. The situation is further complicated by interruptions (incomplete words), hesitations, and other features of spontaneous speech.

Existing parsers (e.g., *MyStem*, *pymorphy*, *Stanza,* etc.) have been trained primarily on written, standardized texts, such as literature, blogs, news, and official documents, which further reinforces the mismatch described above. As a result, the distinctive features of spoken speech are underrepresented in the training data. Consequently, when analyzing spoken material with standard morphological parsers, the error rate increases: word boundaries may be identified incorrectly, and errors occur in part-of-speech tagging and grammatical feature assignment.

Therefore, one of the key practical tasks addressed using ESC data is the adaptation of existing automatic morphological annotation tools to the specifics of spoken speech. The ESC corpus, particularly its transcripts of everyday communication, provides a solid basis for developing such solutions. In particular, work is underway to adapt automatic morphological annotation for spoken data based on a manually annotated subcorpus.

A comparative evaluation of several widely used parsers for Russian (*MyStem*, *pymorphy3*, *Natasha*, *spaCy*, *Stanza*) was conducted on the corpus material. To ensure comparability, a unified tagging scheme was adopted, after which the automatically assigned tags were compared with a manually annotated gold standard. The evaluation was carried out using standard metrics, including accuracy, precision, recall, and F1-score, as well as a qualitative analysis of typical errors.

The most stable results were demonstrated by *MyStem* and *Stanza*; however, both systems show difficulties in tagging interjections, particles, and discourse markers, which play an important pragmatic role in conversational speech. Overall, the results confirm that models trained on written corpora tend to interpret spoken forms in line with the norms of standardized language, which leads to formally correct but functionally and pragmatically inaccurate analyses. This highlights the need to develop a morphological parser for Russian specifically designed for processing spontaneous spoken speech.

7 Conclusions

This paper presents the current state of the ESC corpus of everyday student speech and describes key practical aspects of its development, including the use of diarization models and speech recognition systems, which make it possible to speed up the addition of new speech transcripts to the corpus. Issues of personal data anonymization are also addressed. The ESC project is considered a long-term initiative, and the collection of audio data is planned to continue in the coming years.

Since the creation of the corpus is a multi-stage and labor-intensive process, its expansion proceeds unevenly across different stages. The fastest stage is the collection of new audio recordings, as it requires the least effort. In contrast, segmentation into macro-episodes and their annotation, which are necessary for preparing audio files for automatic transcription, are currently performed manually, which significantly slows down the growth of the final transcripts. Another bottleneck is the expert correction of automatically generated speech transcripts. Because both of these stages require expert involvement, the growth of text data lags far behind the accumulation of new recordings.

As shown in Section 6, a manually annotated subcorpus is used to adapt automatic morphological annotation to spoken speech. The data obtained make it possible to test alternative analysis strategies and, in the future, to develop a tool specifically designed for spontaneous discourse. This will make it possible to integrate

the corresponding modules into the corpus and its demonstration website.

In addition to textual transcripts of everyday speech, the original audio recordings represent a particularly valuable component of the corpus, as they make it possible to "hear" authentic student communication in Russian. These materials can be highly useful both for didactic purposes in teaching Russian as a foreign language and for speech technologies — for training speech systems and voice assistants to communicate in a natural way, taking into account the lexical, grammatical, phonetic-prosodic, and pragmatic features of real speech. However, since the speech signal constitutes a biometric characteristic that can often be used to identify a speaker, the publication of audio recordings in open access is only possible under conditions of substantial acoustic modification of the signal, which represents a specific form of anonymization.

The ESC corpus can also serve as a testbed for evaluating NLP tools for Russian. The resulting resource is expected to be valuable both for fundamental research on contemporary Russian spoken discourse and youth communicative practices, and for applied tasks such as the development and training of automatic speech recognition systems, language models, and other natural language processing technologies.

Acknowledgments

This work is an output of a research project HSE-BR-2025-032 implemented as part of the Basic Research Project at HSE University. The implementation of the project has been made possible thanks to the active involvement of HSE University philology students, who contribute to the corpus on a volunteer basis at all stages of its development. We also express our gratitude to all anonymous informants and their interlocutors who provided recordings of their everyday communication for the ESC corpus.

References

Tatiana Sherstinova and Irina Petrova. 2024. ESC Corpus of Spoken Russian: Everyday Student Conversations Captured through Continuous Speech Recording in Natural Communicative Environments. In Alexey Karpov and Vlado Delić, editors, Speech and Computer, 26th International Conference, SPECOM 2024, Belgrade, Serbia, November 25–28, 2024. Lecture Notes in Artificial Intelligence, volume 15299, pages 151–162.

Tatiana Sherstinova, Aleksey Melnik, Irina Petrova, Karina Azarevich, Valeria Melkozerova, and Sofia Chepovetskaya. 2025. "Okie dokie, here's the no-cap truth!": Everyday Russian Youth Speech in Corpus Representation (Structure and Application of the ESC Sound Corpus). Komp'juternaja Lingvistika i Intellektual'nye Tehnologii / Annual International Conference on Computational Linguistics and Intellectual Technologies, Dialogue 2025, volume 23, pages 331–344.

Tatiana Sherstinova, Nikolay Mikhaylovskiy, Evgeniya Kolpashchikova, and V. Kruglikova. 2024. Bridging Gaps in Russian Language Processing: AI and Everyday Conversations. In 35th Conference of Open Innovations Association (FRUCT), Tampere, Finland, April 24–26, 2024, pages 253–258.

Tatiana Sherstinova, Rostislav Kolobov, and Nikolay Mikhaylovskiy. 2023. Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level. In Proceedings of SPECOM 2023 / LNCS 14338/14339. Springer, pages 43–56.

Janet S. Shibamoto. 1987. Japanese Sociolinguistics. Annual Review of Anthropology, volume 16, pages 261–278.

Takeshi Shibata. 1983. Study of Language Behavior over 24 Hours [Issledovanie yazykovogo sushchestvovaniya v techenie 24 chasov]. Translated by V. M. Alpatov. In V. M. Alpatov, editor, Linguistics in Japan [Yazykoznanie v Yaponii]. Moscow: Raduga, pages 134–141.

Nick Campbell. 2004. Speech & Expression; the Value of a Longitudinal Corpus. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, editors, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, pages 183–186.

Reference Guide for the British National Corpus. http://www.natcorp.ox.ac.uk/docs/URG.xml. Last accessed: 10/01/2025.

Ari Holtzman et al. 2020. The Curious Case of Neural Text Degeneration. In Proceedings of

the International Conference on Learning Representations.

Natalia Bogdanova-Beglarian, Tatiana Sherstinova, Olga Blinova, Olga Ermolova, Ekaterina Baeva, Gregory Martynenko, and Anastasia Ryko. 2016. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. In A. Ronzhin et al., editors, Speech and Computer, Lecture Notes in Artificial Intelligence, volume 9811. Springer.

Aleksey A. Karpov and Irina S. Kipyatkova. 2012. Metodologiya otsenivaniya raboty sistem avtomaticheskogo raspoznavaniya rechi. Izvestiya vysshikh uchebnykh zavedenii. Priborostroenie, volume 55, number 11, pages 38–43.

Matthew Maciejewski, Dominik Klement, Ruizhe Huang, Matthew Wiesner, and Sanjeev Khudanpur. 2024. Evaluating the Santa Barbara Corpus: Challenges of the Breadth of Conversational Spoken Language. In Interspeech 2024, pages 2155–2159.

Whisper Model Homepage. https://openai.com/index/whisper/.

NTR Homepage. https://ntr.ai. Last accessed: 2026/04/10.

John W. Du Bois et al. 2000–2005. Santa Barbara Corpus of Spoken American English, parts 1–4. Philadelphia: Linguistic Data Consortium.

The Corpus of Spontaneous Japanese. https://clrd.ninjal.ac.jp/csj/misc/preliminary/index_e.html. Last accessed: 28/03/2025.

Awais Ali and Steve Renals. 2018. Word Error Rate Estimation for Speech Recognition: e-WER. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, volume 2, pages 20–24.

Cosmin Munteanu, Graeme Penn, Ronald Baecker, Elaine Toms, and D. James. 2006. Measuring the Acceptable Word Error Rate of Machine-Generated Webcast Transcripts. In Proceedings of the 9th International Conference on Spoken Language Processing.

Alexander Asinovsky, Natalia Bogdanova, Marina Rusakova, Svetlana Stepanova, Anastasia Ryko, and Tatiana Sherstinova. 2009. The ORD Speech Corpus of Russian Everyday Communication “One Speaker’s Day”: Creation Principles and Annotation. In Text, Speech and Dialogue, Lecture Notes in Computer Science, volume 5729.

Tatiana Sherstinova. 2015. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus. In A. Ronzhin et al., editors, Speech and Computer, Lecture Notes in Artificial Intelligence, volume 9319, pages 268–276.

One Speech Day Corpus Online. https://ord.spbu.ru/.

CLARIN Spoken Corpora. https://www.clarin.eu/resource-families/spoken-corpora.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: Neural Building Blocks for Speaker Diarization. In Proceedings of ICASSP 2020 — IEEE International Conference on Acoustics, Speech and Signal Processing, pages 7124–7128.