An ASR-to-Universal-Dependencies Pipeline for Morphosyntactic Speaker Profiling of Russian Spontaneous Speech: A Case Study of Convicted Violent Offenders

Aleksandra Tsygankova
Saint Petersburg, Russia
tsygankova.aleksandra.v@gmail.com

Abstract—**The automatic profiling of individual speaker characteristics based on spontaneous speech is useful in several areas at once — in speech technologies, corpus sociolinguistics and applied tools for forensic linguistics. For the Russian language, this task is still limited by the fact that there are almost no processing procedures that have been tested on unprepared institutional speech. This article describes a semi-automatic pipeline comprising: automatic speech recognition (Whisper), manual correction and separation of the speaker's speech, lemmatisation and part-of-speech tagging (MyStem), and syntactic parsing using the Universal Dependencies model (Stanza, russian-syntagrus). The result is a set of morphosyntactic features of the speaker: pronoun distribution, tense and aspect distribution and the proportion of constructions with reduced grammatical subject expression. The pipeline was applied to a compiled corpus of Russian spoken monologue by serial and/or mass violent offenders, comprising 39,062 tokens. The corpus included nine speech recordings of eight speakers taken from publicly available video recordings. The feature values obtained were compared with data from the "One Day of Speech" corpus of everyday Russian speech using the χ^2 test, and the consistency of the feature values was examined on two recordings of the same speaker. The automatic speech recognition error rate was assessed separately: the word error rate lies between 4.4% (insertions excluded) and 7.5% (upper bound), and the bias in feature values introduced by this error does not exceed 2% for the frequency of first-person pronouns and averages 4.4% (maximum 14.9%) for the egocentricity index; the other features have not yet been checked for this bias. Within the corpus the feature values differ noticeably from one recording to another, and for most features the two recordings of the same speaker give similar values, although for several of them they do not. Since the corpus includes only eight speakers and a single genre, the results describe specific recordings of this genre; they cannot be extended to offenders as a group, and the stability of the features for an individual speaker still has to be tested on a larger amount of data.**

I. Introduction

The automatic description of a speaker's individual characteristics based on the features of their spontaneous speech is in demand in speech technologies, corpus sociolinguistics and applied tasks in forensic linguistics. For the Russian language, this task is complicated by the fact that there are still virtually no reproducible processing procedures that have been tested on unprepared institutional speech.

The linguistic study of the speech of individuals who have committed violent crimes, based on Russian-language data, remains a virtually unexplored field. In international scholarship, however, this area has been actively developing since the early 2010s. Works by J. Hancock and co-authors [1], B. Gawda [2], M. Woodworth and co-authors [3], L. E. Cano Gómez [4], and C. Bellara [5] propose a detailed methodology for the quantitative description of the speech of speakers with a criminal history. Most of these studies were conducted using English-language data, with the sole exception being B. Gawda's study, which was carried out in Polish. B. Gawda herself notes that "there may be differences in linguistic characteristics between English-speaking and Polish-speaking individuals with psychopathic traits" [2]. In other words, the transferability of results between languages is not assumed by default. To date, no corpus studies combining lexico-statistical and grammatical levels of analysis have been carried out on Russian-language data.

Lexico-statistical indicators (frequency dictionaries of lemmas, word forms and parts of speech) answer the question of which units the speaker chooses more frequently. Grammatical indicators (the distribution of pronouns, tense and aspect forms, and means of expressing the subject) answer the question of how the speaker positions themselves in relation to what is being described. These two levels complement one another, and describing speech at both of them captures its specific characteristics more fully.

This paper describes a semi-automatic procedure for processing speech data, from publicly available video recordings to the profiling of speaker characteristics, as well as the results of its application to a newly compiled corpus of Russian spoken monologue by serial and/or mass violent offenders. This is a descriptive pilot study. The word "profile" is used below to refer to the vector of feature values obtained for a single recording, without assuming that these values are a stable trait of the speaker or of any group of speakers. The contributions of this work are as follows:

- a reproducible pipeline: automatic speech recognition (Whisper medium), manual correction, lemmatisation and part-of-speech tagging (MyStem), syntactic parsing using the Universal Dependencies model (Stanza, russian-syntagrus), feature extraction, and comparison with a reference corpus using the χ^2 test;

- a description of each recording by seven morphosyntactic features (frequency of "ya" [I] and "my" [we], egocentricity index, proportion of perfective aspect, imperfective present-to-past ratio, proportion of passive constructions, proportion of predications without an explicit subject);
- a corpus of Russian spoken monologue comprising 39,062 tokens (nine recordings, eight speakers), compiled from publicly available video recordings;
- a preliminary check of the consistency of feature values across two recordings of the same speaker, made in different years, with approximate confidence intervals for the proportion features;
- a quantitative assessment of the automatic speech recognition error rate and its impact on the extracted features; this impact was measured for the two pronoun-based features.

II. Literature review

In international research, the study of the speech of individuals convicted of violent crimes is primarily conducted within the framework of forensic and criminological psycholinguistics and draws, amongst other things, on quantitative and automated methods of text analysis — the same family of methods that underlies quantitative stylistics [6] and authorship attribution [7]. The methodology of these studies is based on the idea that a speaker's speech behaviour is regarded as a set of formal parameters that can be measured — for example, the frequencies of lexemes belonging to certain semantic classes, grammatical forms and function words — rather than simply as a qualitatively described set of stylistic features. Many of these parameters are also thought to lie beyond the speaker's conscious control.

The work of J. Hancock, M. Woodworth and S. Porter, based on 52 narratives by individuals convicted of murder, is the basic study in this field. For the present study, their result on psychological distancing matters most. Speakers in the study group "used more past tense forms of content verbs (e.g. stabbed) but fewer present tense forms (e.g. stab) than participants in the control group" [1, pp. 107–110]. The same conclusion is reiterated in the work by M. Woodworth and co-authors, which is aimed at law enforcement officers: function words are produced unconsciously, so computer analysis of "the subtle aspects of their language represents a new avenue for uncovering important information about their behaviour" [3].

Further development in this area has involved the use of more sophisticated statistical methods. Whilst the study by J. Hancock and his co-authors compared two predefined groups, B. Gawda's research employed cluster analysis, which enables narrative types to be identified without first dividing the data into groups. On a sample of 600 spontaneous narratives, the *k*-means method identified six combinations of linguistic features, two of which are statistically associated with a high level of psychopathic traits [2]. More recent studies utilise neural network-based natural language processing tools. For example, in C. Bellara's master's thesis, the speech of serial killers is analysed using tone classification, emotion recognition based on RoBERTa-class models, and psycholinguistic "profiling" using the LIWC dictionary [5]. It was precisely the counting of function words, primarily pronouns, that formed the basis of automated systems for psycholinguistic speech analysis, such as LIWC [8]. The present study extends this logic from the lexical level to the morphosyntactic level.

Despite differences in data and methods, the studies reviewed yield some similar or even common findings: emotional detachment from the event being described, including that expressed grammatically through the predominance of the past tense; a high frequency of self-references using pronouns; an "instrumental-rational" structure in the account of the crime (cause-and-effect constructions); reduced focus on the social sphere; and disruptions to coherence and fluency. In terms of the methods employed, these studies are united by a quantitative, automated analysis of speech characteristics and features.

At the same time, this field of research has certain limitations which influence the formulation of the objectives of this study. Firstly, the majority of studies have been conducted using English-language material. By contrast, the Russian-language speech of individuals who have committed violent crimes is virtually absent from similar studies. The closest Russian-language precedent is a qualitative study of euphemisms in the speech of professional killers [9], which does not use corpus-statistical methods. Secondly, foreign researchers have focused primarily on the lexical and tonal characteristics of speech (semantic word classes, emotional tone), whilst the grammatical level — for example, the distribution of tense and aspect forms, pronouns and voice constructions — is addressed only in a fragmentary manner and, as a rule, does not become a subject of analysis in its own right. Thirdly, the majority of studies focus on the speech of individuals with a medical diagnosis of "psychopathy" (according to the PCL-R scale). In the present study, the object of study is defined not in clinical terms, but in legal and social terms, through the fact of a conviction for a violent crime.

The reference corpus of spoken Russian is the "One Day of Speech" (ORD) corpus, created at the Faculty of Philology of Saint Petersburg State University. The methodological basis of the ORD is "the production of audio recordings in conditions as close as possible to natural ones, for which a method of multi-hour monitoring and continuous 24-hour recording of all speech produced by informants is used" [10, p. 101]. To date, the corpus "contains around 1,250 hours of audio, over 2,800 communicative macro-episodes, and includes speech data from 128 informants <…> as well as over 1,000 of their main interlocutors", whilst "the volume of the corpus's transcriptions exceeds 1 million word occurrences" [10]. Analysis of the ORD data shows that the frequency characteristics of speech are sensitive to the speaker's social roles: "the application of statistical parameters to analyse the frequency vocabulary of informants in the Russian language corpus of everyday communication 'One Day of Speech' has revealed indicators that mark roles in asymmetrical relationships" [11, p. 71]. This motivates the use of frequency-based methods for describing the speech of individual speakers.

III. Corpus

The research material consisted of a corpus of spoken language compiled independently specifically for this study, based on audio and video recordings of interviews available in

the public domain on the Internet. The criteria for selecting the material were: (1) the speaker having been convicted under articles of the Criminal Code of the Russian Federation that classify offences as violent; (2) the existence of a recording featuring the convicted person themselves, containing detailed spoken discourse rather than merely brief answers to questions from an interviewer or investigator; (3) the technical quality of the audio signal being sufficient for subsequent automatic transcription.

Based on these criteria, the corpus includes speech recordings of eight individuals convicted of serial and/or mass violent crimes. Eight of the recordings are video interviews posted on documentary and journalistic channels on YouTube, and one is an excerpt from a 2000 documentary film posted on the RuTube platform. For one of the speakers, the corpus includes two different excerpts from two separate interviews recorded in different years. These are used later to check the consistency of features (Section VII). Speakers are designated by the codes S1–S8, and the two samples from the same speaker are designated S3-1 and S3-2.

TABLE I. COMPOSITION OF THE CORPUS

Recording	Tokens	Share, %
S1	13,153	33.7
S2	10,299	26.4
S3-1	5,378	13.8
S4	3,064	7.8
S5	2,245	5.7
S6	1,786	4.6
S3-2	1,301	3.3
S7	1,107	2.8
S8	729	1.9
Total	39,062	100

The total corpus size was 39,062 tokens. This size makes it possible to build frequency lists for the corpus as a whole, but it is small for comparisons between speakers; in addition, the distribution of the corpus across speakers is uneven. Two speakers (S1 and S2) account for around 60% of the total corpus volume (33.7% and 26.4% respectively), whilst the speech of S8, S7 and S3-2 is represented by fewer than 1,500 tokens each. This will need to be taken into account when interpreting the results, as quantitative estimates for speakers with a small subcorpus (primarily S8) are less statistically reliable than those for speakers with a larger sample size.

The corpus recordings also differ in terms of several extralinguistic conditions of speech production (Table II), which are used below when interpreting the obtained feature values.

IV. PROCESSING PROCEDURE

The processing of the material is organised as a sequence of steps, from the original video recording to the profiling of the speaker (Fig. 1). Each step is described below, along with the versions of the tools used.

TABLE II. EXTRALINGUISTIC CHARACTERISTICS OF THE CORPUS RECORDINGS

Recording	Type of offence	Mode of commission	Type of interview
S1	serial	acting alone	formal
S2	serial	organised group	formal
S3-1	serial	acting alone	formal
S3-2	serial	acting alone	author-hosted
S4	single, mass-casualty	acting alone	formal
S5	single, mass-casualty (gangland score-settling)	group (with accomplices)	author-hosted
S6	single, mass-casualty	acting alone	formal
S7	serial	group (with an accomplice)	formal
S8	serial	acting alone	documentary film

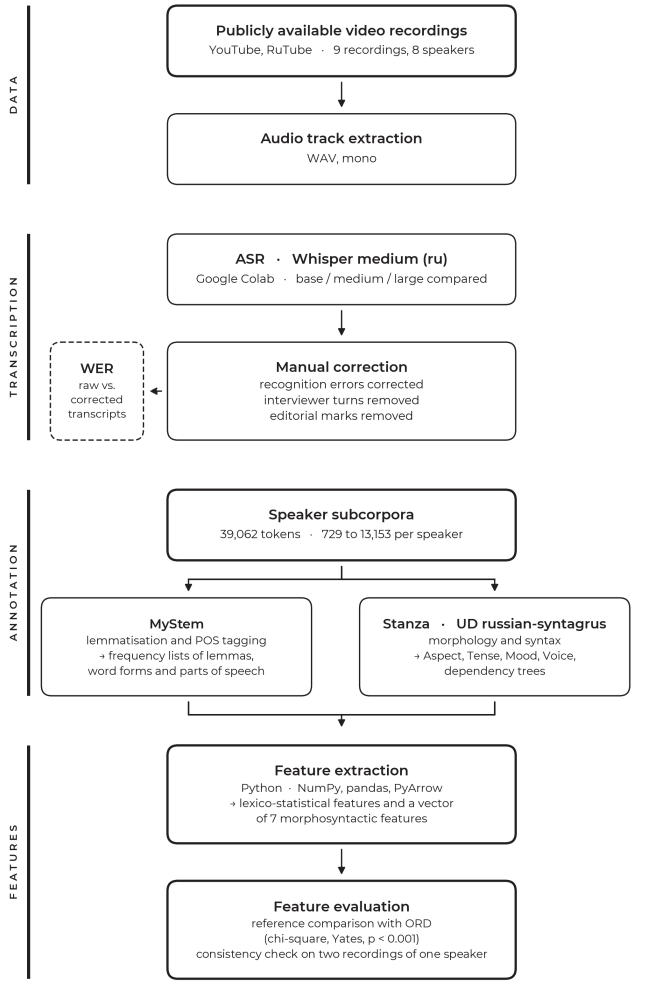


Fig. 1. Processing pipeline: from a publicly available video recording to a speaker feature vector

A. Automatic speech recognition

Automatic transcription of audio tracks was carried out using the Whisper automatic speech recognition system [12], [13], developed by OpenAI. The "medium" model was used for transcription, and Google Colab [14] served as the execution environment. The choice of the "medium" model, rather than the "base" or "large" models, was based on preliminary testing on several test fragments. The "base" model exhibited noticeably lower recognition quality, whilst the "large" model, according to the test results, offered virtually no significant improvement over the "medium" model, yet required considerably more time and computational resources.

TABLE III. COMPARISON OF WHISPER MODELS ON TEST FRAGMENTS

Model	Recognition quality	Time and resources	Decision
base	noticeably lower than medium	lowest	not used
medium	sufficient for subsequent manual correction; WER 4.4% (Table IV)	moderate	chosen for processing the whole corpus
large	no practically significant gain over medium	substantially more time and computing resources	not used

The models were compared as a preliminary qualitative assessment using test data. A quantitative evaluation (Section IV.C) was carried out for the selected "medium" model.

B. Manual correction of transcriptions

The transcriptions generated by Whisper were further edited manually. At this stage, three tasks were addressed simultaneously: (1) correcting errors in automatic recognition (restoring correct word forms where Whisper had substituted homonyms or phonetically similar units, restoring and standardising punctuation, and splitting or merging utterances in accordance with the actual audio of the recording); (2) removing from the text all utterances not made by the speaker (questions and comments from the interviewer or presenter, remarks by other participants in the conversation, and fragments of ancillary communication); (3) removal from the text of administrative notes and fragments not relating to the substantive content of the recording (long pauses, unintelligible segments).

The second of these steps is fundamental to further analysis, as it ensures that only the speaker's own speech is included in the corpus and allows each resulting text to be regarded as a speech "portrait" of a specific speaker, reflecting the institutional communicative situation of the interview.

C. Evaluation of automatic recognition quality

All subsequent analysis is based on transcriptions, so any recognition errors are carried directly over into the feature values. To measure this, Whisper's transcriptions were compared with manually corrected versions of the same recordings.

A direct comparison of the two files would be incorrect, as, in addition to correcting recognition errors, manual correction also removed the interviewer's remarks from the text. The original files therefore contain a significant amount of material that is absent from the corrected version, not due to any specific recognition error. The following procedure was applied. Both texts were standardised (lower case, punctuation removed, the letter "yo" normalised to "e") and split into words, after which a word-by-word alignment was constructed whilst preserving the order (difflib.SequenceMatcher, autojunk=False). Sections present only in the raw file correspond either to deleted remarks by the interviewer or to insertions by the recognition system. As it is not possible to distinguish automatically between these two cases, they were excluded from the count. Errors were defined as substitutions (S) — words recognised incorrectly — and deletions (D) — words from the corrected version that were absent from the raw transcript. The word error rate (WER) was then calculated as

$$\mathrm{WER} = (S + D) / N, \quad (1)$$

where N is the word count of the corrected version.

TABLE IV. WHISPER MEDIUM WORD-LEVEL RECOGNITION ERROR (ALL NINE CORPUS RECORDINGS)

Recording	*N*, words	Substitutions *S*	Deletions *D*	WER, %
S1	13,153	193	21	1.6
S2	10,299	389	307	6.8
S3-1	5,378	158	119	5.2
S4	3,064	111	87	6.5
S5	2,245	47	7	2.4
S6	1,786	89	39	7.2
S3-2	1,301	27	5	2.5
S7	1,107	17	3	1.8
S8	729	70	41	15.2
Total	39,062	1,101	629	4.4

Pooled over the nine recordings, the WER calculated with (1) is 4.4%, with a substitution rate of 2.8%. This represents the lower limit, as it does not include insertions made by the recognition system. If we count as insertions all short fragments present only in the raw file, the upper limit would be 7.5%. However, manual verification shows that the vast majority of such fragments are short remarks by the interviewer ("da" [yes], "pochemu" [why], "kto eto" [who is that]). The true value therefore lies closer to the lower limit. The variation between recordings is significant and may be explained by their origin. The smallest WER is found in the longest and technically highest-quality recording (S1 — 1.6%), whilst the largest is in S8 (15.2%), an excerpt from a 2000 documentary film and, at the same time, the shortest subcorpus. In S8, poor recording quality coincides with a small volume, so the two effects cannot be separated, and the values for this recording should be treated with particular caution.

What is more important for evaluating the pipeline is how this error affects the features. Feature values were calculated twice: using the aligned section of the raw transcription and using the corrected version. The frequency of first-person pronouns varies by no more than 2.0% (on average by 0.7%). The egocentricity index varies by an average of 4.4%, with the maximum deviation (14.9%) occurring in S6, a recording with one of the highest WER values and a small volume. By way of comparison, the variation in these same features across speakers is a factor of 2.81 for the frequency of "ya" [I] and 6.4 for the

egocentricity index. For these two features, recognition errors are thus much smaller than the variation between recordings. The tense, aspect, voice and subject-expression features have not been checked in the same way: they depend on the parser and may react to recognition errors differently, so the effect of recognition errors on them remains unknown. It should also be borne in mind that the estimate was made on the aligned part of the transcripts, i.e. without insertions. The main benefit of the manual stage of recording processing is the separation of the speaker's speech from that of the interviewer, rather than the correction of word forms.

D. Linguistic annotation

Following manual correction, each text was annotated at two linguistic levels. Lemmatisation and part-of-speech tagging were carried out using MyStem [15], [16]. The resulting data were used to construct frequency dictionaries of lemmas, word forms and parts of speech. Morphological and syntactic parsing was carried out using the Stanza software pipeline [17], [18] based on the Universal Dependencies model [19] for the Russian language (russian-syntagrus [20]). The result of this processing is the representation of each sentence as a syntactic dependency tree, specifying for each word form its lemma, part of speech, morphological features (categories such as Aspect, Tense, Mood, Voice, etc.) and syntactic role within the sentence.

All post-processing stages (constructing tables of speakers' grammatical characteristics, calculating χ^2, and exporting tables) are implemented in Python using the NumPy [21], [22], pandas [23], [24] and PyArrow [25] libraries.

E. Extracted features

At the output of the pipeline, two sets of features are produced for each speaker. The first is lexico-statistical (frequency dictionaries of lemmas, word forms and parts of speech, normalised as instances per million (IPM), and their comparison with a reference corpus). The second comprises seven morphosyntactic features (Table XI): frequency of "ya" [I] in IPM, egocentricity index, frequency of "my" [we] in IPM, proportion of perfective aspect, imperfective present-to-past ratio, proportion of passive constructions, and proportion of predications without an explicit subject. All features are calculated automatically from the manually corrected transcripts, using the tagger and parser output. The parser-derived values have not been validated against manual annotation (see Section VIII). Pronoun frequencies are lemma frequencies normalised per million tokens as counted in Table I; the tense and aspect features are proportions of the verb tokens to which the parser assigns both tense and aspect (finite forms, participles and converbs; N in Table IX), whereas the passive, subject-expression and reflexive features are proportions of all tokens tagged VERB by the parser, infinitives included (N in Table X). Both sets are smaller than the MyStem part-of-speech class V in Table VI.

V. Lexico-statistical features

The lemmatised corpus contains 39,062 tokens, 4,047 unique lemmas and 7,432 unique word forms. 1,892 lemmas (46.8% of the frequency dictionary) are hapaxes, i.e. they occur only once in the corpus; a further 1,378 lemmas (34.0%) have a frequency of between 2 and 5; and only 55 lemmas (1.4% of the frequency dictionary) have a frequency higher than 100. On the modelling of low-frequency vocabulary in Russian see [26].

These figures relate to the corpus as a whole. For individual speakers, a direct comparison is not entirely accurate: both the TTR and the proportion of hapax terms decrease as the volume of text increases. Subcorpus S8 (729 tokens) will therefore yield higher values than subcorpus S1 (13,153 tokens), regardless of the actual lexical richness of the speech. This is also evident in the collected data (Table V): the TTR decreases from 0.565 for the speaker with the smallest text volume to 0.260 for the speaker with the largest. Therefore, alongside the raw TTR, a standardised indicator (STTR) was calculated. The text is divided into windows of 500 word forms. The TTR is calculated in each window and averaged, whilst the proportion of hapaxes is averaged across the same windows. Only these values are comparable across speakers.

Table V. Lexical diversity by speaker

Speaker	N tokens	Distinct word forms	TTR	500-word windows	STTR	Hapax in window, %
S1	13,153	3,417	0.260	26	0.548	71.8
S2	10,299	2,476	0.240	20	0.525	68.1
S3-1	5,378	1,635	0.304	10	0.537	71.7
S4	3,064	1,053	0.344	6	0.535	71.7
S5	2,245	888	0.396	4	0.560	75.5
S6	1,786	627	0.351	3	0.465	67.6
S3-2	1,301	570	0.438	2	0.568	75.3
S7	1,107	505	0.456	2	0.529	72.4
S8	729	412	0.565	1	0.586	72.7

Following standardisation, the variation decreases sharply and the STTR falls within the range of 0.465–0.586. No correlation with subcorpus size can be observed, which means that the metric performs as intended. The lowest lexical diversity is observed in S6 (0.465 with a median of 0.537). This is the only notable deviation in lexical diversity across the corpus, and it is consistent with other features of his speech (one of the lowest frequencies of "ya" [I], the highest proportion of predications without an explicit subject). Formally, the highest value is found in S8 (0.586), but this was obtained from just a single window and cannot be considered reliable. The proportion of hapax within the windows behaves in the same way and provides no additional information.

The distribution of parts of speech (Table VI) reveals certain features typical of spontaneous spoken language. The most frequent part of speech is the verb (19.5%), which is slightly higher than the proportion of nouns (17.9%). This ratio, with verbs predominating, is characteristic of spoken language in particular, as nouns usually predominate in written texts, especially in journalism and non-fiction prose. At the same time, the overall proportion of pronouns is high: pronoun-nouns (SPRO) account for 14.2%, adjectival pronouns (APRO) for 5.2%, and adverbial pronouns (ADVPRO) for 4.4%. Taken together, almost 24% of all tokens in the corpus are pronouns. The proportion of particles (PART) is also high — 9.96%. In contrast, however, there are few adjectives (A) in the corpus — 3.67%.

TABLE VI. PART-OF-SPEECH DISTRIBUTION OF THE CORPUS

No.	POS	Abs. frequency	Rel. frequency, %	IPM
1	V	7,608	19.48	194,767
2	S	6,993	17.90	179,023
3	SPRO	5,546	14.20	141,979
4	PART	3,892	9.96	99,636
5	PR	3,421	8.76	87,579
6	CONJ	3,397	8.70	86,964
7	ADV	2,718	6.96	69,582
8	APRO	2,035	5.21	52,097
9	ADVPRO	1,722	4.41	44,084
10	A	1,433	3.67	36,685
11	NUM	180	0.46	4,608
12	ANUM	99	0.25	2,534
13	INTJ	18	0.05	461

Among the pronouns, the first-person singular pronoun “ya” [I] ranks first, accounting for 5.20% of all tokens in the corpus, with an IPM of 51,994. One word in every nineteen in the corpus under study is the pronoun “ya”. The frequency list of word forms elaborates on this observation: with a total lemma frequency of 2,031, the nominative case form “ya” accounts for 1,430 occurrences, the genitive/accusative case form “menya” for 334, and the dative case form “mne” for 238. Of all 2,031 occurrences of the first-person pronoun, more than 70% are in the nominative case, functioning as the subject. This indicates that speakers in the corpus under study prefer constructions in which they themselves act as the grammatical subject of the action, rather than as the object or addressee.

A comparison of the frequency characteristics of the corpus under study with data from everyday spoken Russian is carried out on the basis of the published upper zone of the ORD frequency list [27]. The comparative material from the ORD consists of a dictionary of word forms, which was compiled from a sample comprising 232,370 tokens (152 macro-episodes of speech from 209 speakers), so the comparison is also carried out at the level of word forms rather than lemmas. The significance of the differences was assessed using the χ^2 test with Yates’s correction [28] at $p < 0.001$ (critical value 10.83). Since 50 word forms are compared, this threshold corresponds to a Bonferroni-corrected family-wise level of 0.05 (0.05 / 50 = 0.001). The test assumes that tokens are independent observations, whereas in both corpora they are clustered within speakers, and in the corpus under study S1 and S2 alone contribute about 60% of the tokens. For this reason, the χ^2 values are used here, together with the frequency ratio, as a descriptive measure of how strongly the two samples differ, and are not treated as inferential tests about a population of speakers. On frequency profiling as a technique for comparing corpora see [29].

Overall, the frequency profiles of the two corpora show considerable similarity. In the ORD frequency list, the top ten word forms account for 20.69% of all tokens [27, p. 619]. In the corpus under study, this figure stands at 20.65%. The overall composition of the upper zone also coincides. The differences, however, relate primarily to the ranking within this upper zone and specific proportions, and can be divided into three groups (Table VII).

TABLE VII. COMPARISON OF THE CORPUS WITH THE ORD (MOST REPRESENTATIVE WORD FORMS)

Word form	Corpus, %	ORD, %	Ratio	χ^2
“ya” [I]	3.66	2.56	1.43	152.6
“menya” [me]	0.86	0.48	1.78	85.6
“my” [we]	0.65	0.45	1.44	27.1
“bylo” [was]	0.69	0.25	2.78	202.0
“vot” [here, well]	0.93	2.33	0.40	311.4
“da” [yes]	1.05	2.22	0.47	226.8
“a” [and, but]	0.83	1.95	0.43	235.0
“ty” [you, singular]	0.22	0.81	0.27	160.0
“tam” [there]	0.89	1.47	0.60	83.7
“nu” [well]	2.08	2.33	0.89	8.9 (n.s.)
“mne” [to me]	0.61	0.56	1.09	1.6 (n.s.)
“eh” (hesitation)	0	0.74	0	—
“ugu” [uh-huh]	0.01	0.62	0.02	—

The first group is technical: the absence of markers of filled hesitation and feedback. The ORD frequency list includes the word forms “eh” (filled hesitation, 0.74%) and “ugu” [uh-huh] (feedback from the listener, 0.62%). In the corpus under study, these units are virtually absent (0 and 4 instances respectively). This discrepancy can be explained by two factors. Firstly, the Whisper automatic speech recognition system does not always transcribe filled hesitations, as it is trained on “clean” speech data. Secondly, “ugu” in the corpus under study would predominantly have referred to the interviewer’s remarks, which were removed during the manual correction stage.

The second group is dialogical: a reduced frequency of dialogue markers. In the corpus under study, the frequency of the following word forms is significantly lower than in the ORD: “vot” [here, well] (0.93% versus 2.33%, $\chi^2 = 311.4$), “da” [yes] (1.05% versus 2.22%, $\chi^2 = 226.8$), “a” [and, but] (0.83% versus 1.95%, $\chi^2 = 235.0$), “ty” [you, singular] (0.22% versus 0.81%, $\chi^2 = 160.0$), and “tam” [there] (0.89% versus 1.47%, $\chi^2 = 83.7$). Their reduced frequency is explained by the genre-specific nature of the corpus under study. Interviews tend to be forms of monological communication in which the speaker narrates at length about their own experiences.

The third group is thematic: an increased frequency of certain words, linked to the subject matter and narrative mode. It is precisely this group that presents the greatest research interest. In the corpus under study, the frequency of the following word forms is significantly higher than in the ORD: “ya” [I] (3.66% versus 2.56%, $\chi^2 = 152.6$), “menya” [me] (0.86% versus 0.48%, $\chi^2 = 85.6$), “my” [we] (0.65% versus 0.45%, $\chi^2 = 27.1$), and the past tense form of the copula “bylo” [was] (0.69% versus 0.25%, $\chi^2 = 202.0$). The total frequency of first-person singular pronoun forms (“ya”, “menya”, “mne”) in the corpus under study is approximately 5.13%, compared with 3.60% in the ORD. At the level of individual recordings (Table VIII), the frequency of the lemma “ya” is 3.10% for S2, which is below the ORD figure, 4.14% for S6, and as high as 8.71% for S4; seven of the nine recordings exceed 4.5%. As the lemma count includes all case forms, this comparison is only approximate. The excess is thus not produced by the two largest subcorpora alone; moreover, in one of them (S2) it is absent.

When interpreting these differences, it is important to bear in mind that the ORD is a multi-genre corpus. It comprises informal everyday conversations (62.5%), professional

conversations (19.7%), "customer service"-type communication (9.9%), educational communication (7.2%) and public speeches (0.7%) [27]. The corpus under study, by contrast, is homogeneous in terms of genre and pertains to a single, very narrow communicative situation (interviews). Therefore, some of the identified differences (particularly those relating to the second group, dialogical markers) are more likely to be characteristics of this specific genre rather than differences between the speakers. The same applies to the third group: a first-person retrospective account of one's own actions will contain many instances of "ya" and of past-tense forms regardless of who the speaker is. Until a matched corpus of interviews with non-offenders is available, the differences from the ORD cannot be attributed to the speakers as a group.

VI. Morphosyntactic features

The main output of the pipeline falls into three groups: pronoun features, tense and aspect features, and features of the grammatical expression of the subject.

A. Pronoun features

Table VIII. Frequency of pronouns by speaker, IPM

Speaker	ya [I]	my [we]	ty [you sg]	vy [you pl]	on [he]	ona [she]	oni [they]	sebya [self]
S4	87,141	8,486	1,632	326	17,624	7,833	23,825	2,285
S3-2	66,872	3,075	0	5,380	5,380	4,612	3,843	2,306
S7	65,944	3,613	903	32,520	9,033	1,807	1,807	0
S1	57,705	8,515	5,626	2,357	8,059	9,123	8,667	2,281
S3-1	55,783	1,673	372	12,644	8,553	4,649	5,578	5,950
S5	50,780	6,682	8,909	8,463	9,800	9,800	8,463	2,673
S8	45,267	0	1,372	6,859	17,833	0	6,859	2,743
S6	41,433	1,680	1,120	560	17,357	0	5,599	560
S2	30,974	23,012	1,651	1,942	27,284	3,884	6,991	1,165

The pronoun "ya" [I] is the most frequent in all, without exception, individual "portraits" of the speakers, which is to be expected in a first-person autobiographical narrative. However, its frequency varies considerably between speakers: from 30,974 IPM for S2 to 87,141 IPM for S4, i.e. by a factor of 2.81. The mean across recordings is 55,767 IPM, and the median is 55,783 IPM. This difference shows that the frequency of "ya" varies considerably between recordings of the same genre. However, a single recording per speaker is not enough to determine whether it is a stable characteristic of the individual.

Another important parameter of pronominal speech analysis, which is calculated in this study, is the ratio of the frequency of the first-person singular pronoun to the total frequency of third-person pronouns ("on" [he], "ona" [she], "oni" [they]). This could be termed, for example, the "egocentricity index" of the narrative. Based on the value of this index, the speakers can be provisionally divided into three groups. The first group comprises those speakers in whose speech the first person is markedly prominent (S7 (index 5.21) and S3-2 (4.83)). The second group comprises speakers with a moderate, balanced ratio of the pronoun "ya" to third-person pronouns: S3-1 (index 2.97), S1 (2.23), S8 (1.83), S5 (1.81), S6 (1.80), S4 (1.77). The third group consists of a single speaker — S2, with an index of 0.81. This is the only instance in the corpus where the narrative is centred not on the speaker's own "ya", but on third persons.

S2's speech has another unique feature compared with the other speakers. Her frequency of the pronoun "my" [we] is 23,012 IPM, whilst for all other speakers it ranges from 0 (S8) to 8,515 IPM (S1). This sharp deviation can generally be explained by the context, as S2 was a member of a criminal gang, and the pronoun "my" in her speech functions precisely as a referential unit — denoting the group of accomplices. This is, however, a single case: it shows how the pronoun features can reflect the content of an account, and no general conclusion can be drawn from it.

B. Tense and aspect features

Table IX. Tense and aspect distribution (proportions of the total number of verb tokens)

Speaker	N verbs	IPFV	PFV	Past	Pres.	Fut.	IPFV pres./past
S4	530	0.46	0.54	0.75	0.21	0.05	0.80
S1	1,921	0.56	0.44	0.68	0.26	0.07	0.85
S2	1,588	0.56	0.44	0.69	0.26	0.05	0.82
S6	244	0.44	0.56	0.71	0.23	0.05	1.12
S7	145	0.62	0.38	0.54	0.43	0.03	2.21
S5	340	0.59	0.41	0.52	0.42	0.06	2.57
S8	96	0.59	0.41	0.49	0.44	0.07	2.80
S3-1	713	0.65	0.35	0.42	0.49	0.09	3.02
S3-2	165	0.70	0.30	0.36	0.55	0.10	3.60

The combination of "perfective aspect + present tense" is virtually non-existent in Russian, as the perfective form almost always expresses the future tense and is classified as such in the Universal Dependencies annotation scheme. Consequently, the "future" column consists mainly of perfective forms, whilst the "present" column consists almost entirely of imperfective forms.

The proportion of perfective verb forms varies between speakers, ranging from 0.30 (S3-2) to 0.56 (S6). The highest proportion of the perfective aspect is observed in S6 (0.56) and S4 (0.54): in these speakers' accounts the narrative is structured predominantly as a sequence of separate, completed actions. The lowest proportion of the perfective aspect, however, is found in both samples of S3 (0.35 and 0.30): his speech, by contrast, tends to present actions as ongoing, repetitive and unlimited in duration.

The most revealing parameter of the tense and aspect analysis of speech turned out to be the ratio of the frequency of the imperfective present to that of the imperfective past. Based on the value of this parameter, the corpus can be divided into two groups. The first group comprises speakers with a ratio of less than 1.2 — S4 (0.80), S1 (0.85), S2 (0.82) and S6 (1.12). Their accounts are structured predominantly as retrospective narratives in the past tense, in which the proportion of the past tense is approximately 0.68–0.75. The second group comprises speakers with a ratio above 2 — S7 (2.21), S5 (2.57), S8 (2.80), and both samples of S3 (3.02 and 3.60). For these speakers, the imperfective present strongly predominates over the imperfective past, and the total proportion of the present tense reaches 0.42–0.55. This distribution indicates the active use of

present tense forms when describing past events, that is, the "narrative present" or the "habitual present" [30].

The predominance of the past tense, characteristic of the first group, is consistent with the observation by J. Hancock and co-authors that a higher proportion of past tense forms is associated with the speaker's psychological detachment from the event being described. The predominance of present tense forms in the second group, however, allows for a dual interpretation (narrative present or habitual present) and requires a contextual analysis, which was not carried out in this study. The proportion of imperative forms ranges from 0% to 4.7% and shows no significant differences between speakers, so this feature has not been included in the final analysis.

C. Features of the grammatical expression of the subject

This group describes the speaker's speech in terms of the relative frequency of constructions that lower the communicative status of the agent-subject or eliminate it from the syntactic structure of the utterance altogether. This study considers three types of such constructions, namely: passive constructions (the grammatical passive, in which the agent is downgraded in rank or not expressed), predications without an explicit subject, and reflexive forms ending in "-sya" / "-s'". For each speaker, the proportion of each type of construction out of the total number of verb tokens was calculated (Table X).

TABLE X. GRAMMATICAL EXPRESSION OF THE SUBJECT (PROPORTIONS OF THE TOTAL NUMBER OF VERB TOKENS)

Speaker	N verbs	Passive	No explicit subject	Reflexive forms
S7	181	0.028	0.481	0.149
S4	658	0.014	0.460	0.147
S1	2,294	0.071	0.467	0.161
S5	402	0.050	0.483	0.144
S6	299	0.010	0.656	0.167
S3-1	979	0.022	0.547	0.150
S3-2	224	0.031	0.500	0.080
S8	120	0.092	0.433	0.075
S2	1,996	0.041	0.517	0.118

The indicator "proportion of predications without an explicit subject" warrants a separate comment. Russian is a pro-drop language [31, p. 319], and when the referent is highly activated, the speaker regularly uses reduced reference, including "anaphoric zero" — an unstated subject that can be reconstructed from the context [32]. Consequently, the indicator combines two distinct phenomena simultaneously: impersonal constructions proper, and sentences with a zero but reconstructible subject. It also counts non-finite forms (infinitives, participles and converbs), which normally have no subject of their own. As a result, its absolute values are high across all speakers (corpus median: 0.483), but cannot be interpreted directly as the "proportion of impersonality": it is not the absolute values of this indicator that are informative in terms of content, but rather significant deviations from the corpus median.

The proportion of passive constructions in the entire corpus is small (median: 0.031), which is generally logical, as the passive voice is, in principle, infrequent in Russian spoken language. The most reliable value is that of S1 — 0.071 of 2,294 verb tokens, i.e. 163 forms tagged as passive. This is more than double the corpus median. Taking into account the caveat regarding predications without an explicit subject, the only notable deviation is the S6 indicator, which is 0.656 compared to a corpus median of 0.483. This is also consistent with another characteristic of his speech — one of the lowest frequencies of the pronoun "ya" in the corpus (41,433 IPM).

A tentative observation from the analysis of subject expression is that different speakers may use different grammatical means to achieve the same communicative goal — reducing their own subject role in the description of actions. Thus, S6 reduces the prominence of the subject primarily through predications without an explicit subject (and a concomitant reduction in the frequency of "ya"), whilst S1 does so through passive constructions (with an "average" frequency of "ya"). This observation still needs to be verified manually, as the passive proportions for the smaller subcorpora rest on only a few instances each, and the parser's assignment of voice and subject has not been checked by hand.

D. Final set of features

The three groups of features together form a "speaker × feature" matrix of dimension 9 × 7 (Table XI). This is the main output of the pipeline. The matrix is calculated automatically from the corrected transcripts and gives a compact description of each recording.

TABLE XI. MATRIX OF MORPHOSYNTACTIC FEATURES

Speaker	ya, IPM	Egoc. index	my, IPM	PFV share	IPFV pres./past	Passive	No subject
S7	65,944	5.21	3,613	0.38	2.21	0.028	0.48
S4	87,141	1.77	8,486	0.54	0.80	0.014	0.46
S1	57,705	2.23	8,515	0.44	0.85	0.071	0.47
S5	50,780	1.81	6,682	0.41	2.57	0.050	0.48
S6	41,433	1.80	1,680	0.56	1.12	0.010	0.66
S3-1	55,783	2.97	1,673	0.35	3.02	0.022	0.55
S3-2	66,872	4.83	3,075	0.30	3.60	0.031	0.50
S8	45,267	1.83	0	0.41	2.80	0.092	0.43
S2	30,974	0.81	23,012	0.44	0.82	0.041	0.52

Across the nine recordings, the feature values differ noticeably: S4 exhibits the highest frequency of "ya" in the corpus (87,141 IPM), but, unlike S7 and S3, uses it not in the "present" but in the retrospective narrative mode. Among the speakers with an adequate verb sample, S1 has the highest proportion of passive constructions (0.071); the nominally higher value for S8 (0.092) rests on 11 forms out of 120 verb tokens. S6 is characterised by both a low frequency of "ya" (41,433 IPM) and a record proportion of predications without an explicit subject (0.66). S2 has the most distinctive profile in the entire corpus. She is the only one to centre her discourse not on her own "ya" but on third persons, and also exhibits a record frequency of the pronoun "my" (23,012 IPM, compared with a range of 0–8,515 for the others), which is consistent with the group nature of the events she describes.

VII. STABILITY OF FEATURES

The corpus allows a limited check of consistency: it includes two speech samples from the same speaker (S3) taken from two different interviews, recorded in different years and processed

independently of one another. If the features characterise the speaker rather than sampling noise, the values for these two recordings should be consistent. A single pair of recordings is not enough to establish stability; it only shows whether the values are compatible with it.

Overall, the two samples agree in the tense and aspect features: both show the lowest proportion of perfective aspect in the corpus (0.35 and 0.30), the highest ratio of imperfective present to imperfective past (3.02 and 3.60), and a relatively similar proportion of the passive voice (0.022 and 0.031). In terms of pronoun usage, the two samples agree only in direction: a high frequency of "ya" (55,783 and 66,872 IPM — one at the corpus median, the other above it), and a predominance of the first person over the third (index 2.97 and 4.83). The proportions of predications without an explicit subject (0.547 and 0.500) are similar, whilst the proportion of reflexive forms differs markedly (0.150 and 0.080). Approximate 95% Wilson confidence intervals were calculated from the counts behind Tables IX and X. They are 0.31–0.38 and 0.24–0.38 for the perfective share, 0.46–0.53 and 0.47–0.62 for the present-tense share, and 0.52–0.58 and 0.44–0.56 for predications without an explicit subject, and in each case the two intervals overlap. For reflexive forms the intervals (0.13–0.17 and 0.05–0.12) do not overlap, so sample size alone does not explain this difference.

Lexical diversity is also consistent. The standardised TTR for the two samples is 0.537 and 0.568, with a corpus range of 0.465–0.586, whilst the proportion of hapax within the windows is 71.7% and 75.3%.

The discrepancy in the "egocentricity index" (2.97 versus 4.83), given that the speaker is the same, is large: according to this index, the two recordings of the same speaker fall into different groups of the classification given in Section VI.A. The type of interview (formal versus author-hosted) may play a part here, but with a single pair of recordings this cannot be separated from sampling variation. The frequency of "my" also differs (1,673 and 3,075 IPM). The pronoun-based features should therefore, for the time being, be treated as characteristics of a recording.

Overall, the tense and aspect features agree across the two recordings, whilst the pronoun-based features and the proportion of reflexive forms differ. The former may therefore reflect properties of the speaker's speech, although a single pair of recordings does not demonstrate this; a resampling analysis within the longer recordings and further pairs of recordings from the same speaker are required.

VIII. Limitations and ethical considerations

When interpreting the results obtained, the limitations of the study must be taken into account. Firstly, the size of the subcorpora for individual speakers varies significantly: the figures for S8, S7 and S3-2 were derived from fewer than 1,500 tokens each and are therefore less statistically reliable. Secondly, the metric "proportion of predications without an explicit subject" cannot be interpreted directly as the "proportion of impersonal constructions", as it combines impersonal constructions with sentences featuring a zero subject that can be inferred from the context, and it also counts non-finite verb forms. Thirdly, the morphosyntactic analysis is based on automatic tagging, the accuracy of which, though high in modern Russian-language models, is still not absolute. The quality of automatic recognition also varies across corpus recordings (WER ranging from 1.6% to 15.2%), with the highest WER occurring in the shortest recording (S8). Fourthly, the analysis records the distribution of grammatical forms throughout the speaker's text, but does not localise them within specific thematic segments of the narrative. Fifthly, all the material belongs to a single institutional genre (interviews), and some of the identified differences from the ORD are attributable to this, rather than to the characteristics of the speakers. Sixthly, the comparison with the ORD pools tokens across speakers and relies on the χ^2 test, which assumes independent observations. Before any difference can be related to a population, a speaker-level analysis and a matched corpus of interviews with non-offenders are required. Seventhly, the consistency of the features was examined on one pair of recordings; confidence intervals are given for the proportion features of that pair, and no resampling analysis was performed. Eighthly, the WER estimate excludes insertions, the effect of recognition errors was measured for two of the seven features, and the parser-derived features were not validated against manual annotation. Finally, the pipeline is not fully automatic, since recognition errors are corrected and the speaker's speech is separated by hand.

Ethical considerations. Speakers in the article are identified by codes. Names, nicknames and other identifying details are not provided. All recordings used have been made publicly available by their copyright holders. The object of study is defined in legal and social terms — through the fact of a conviction having been handed down — rather than clinically (diagnostic scales, in particular the PCL-R, are not used in this study, and no claims are made regarding the speakers' mental state). The method described is not a diagnostic tool and is not intended to predict criminal behaviour, assess guilt or select individuals on the basis of speech characteristics. With eight speakers and a single communicative genre, the obtained feature values describe specific recordings rather than any group of individuals. The observed differences between speakers are interpreted as differences in speech strategies within a specific communicative situation.

IX. Conclusion and future prospects

This paper presents a semi-automatic pipeline for processing spontaneous Russian spoken monologue, from audio recording to the representation of a set of the speaker's morphosyntactic features, and the results of its application to a corpus of spoken monologue by serial and/or mass violent offenders comprising 39,062 tokens (nine recordings, eight speakers).

A lexico-statistical analysis of the corpus revealed that the upper zone of its frequency list consists of the core elements expected in Russian spoken language: pronouns, particles, conjunctions and prepositions. Verbs are the dominant part of speech (19.5%), with a slight predominance over nouns (17.9%), whilst the total proportion of pronoun classes (around 24%) and particles (9.96%) is high. A comparison with data from the ORD corpus revealed 30 statistically significant differences (at $p < 0.001$) out of the 50 word forms analysed in the upper zone. Most of these differences can be explained by genre and recording conditions, and therefore cannot be attributed to the speakers.

The morphosyntactic feature values differ between recordings: the frequency of "ya" [I] varies from 30,974 to 87,141 IPM, i.e. by a factor of 2.81; tense and aspect analysis divides the corpus into "past narrators" (imperfective present-to-past ratio 0.80–1.12) and "present narrators" (2.21–3.60); features of subject expression suggest that different speakers may use different grammatical means to achieve the same communicative goal — reducing their own subject role in the description of actions.

The quality of the automatic recognition stage has been quantified: the WER lies between 4.4% and 7.5%, and the resulting bias in feature values does not exceed 2% for the frequency of first-person pronouns and averages 4.4% for the egocentricity index, considerably smaller than the variation of these two features between recordings. The other five features have not yet been checked.

A comparison of two recordings of the same speaker showed agreement in the tense and aspect features and noticeable differences in the pronoun-based ones. This observation is preliminary and cannot be regarded as a validation of the method. The results of the study describe specific recordings of a single genre and should not be read as profiles of speakers or of offenders.

Areas for further work: validating the parser-derived features against manual annotation; reporting resampling-based intervals for all features and extending the recognition-error check to all seven of them; a comparison with a matched corpus of interviews with non-offenders; expanding the corpus by including a larger number of speakers and ensuring a more even distribution of subcorpus sizes; conducting a comparative analysis with subcorpora of spoken monologue based on the ORD and SAT (Sbalansirovannaya annotirovannaya tekstoteka, "Balanced Annotated Text Collection") [33] corpora; carrying out a qualitative analysis to identify the grammatical strategies revealed in thematic areas of the narrative; automating the stage involving the separation of the speaker's speech from that of the interviewer, which, together with the correction of recognition errors, remains a manual step in the pipeline and, as the WER assessment has shown, contributes more to the result than the correction of recognition errors.

The pipeline is not restricted to the speech of convicted prisoners: all that is required at the input is a recording of sufficient quality, and the output is a comparable table of features, so in principle it can be applied to other groups of speakers and other institutional genres — such as political interviews, court hearings or medical consultations — provided that the transcripts are corrected by hand as they were here. In corpus sociolinguistics it gives a first quantitative description of new material without manual morphosyntactic annotation, although the parser output still has to be validated, and in speech technologies it helps to select and balance recordings when a spoken corpus is being compiled. Forensic linguistics is the more delicate case. The procedure can check whether the features of two recordings by the same speaker agree, but no further: with seven features it provides no basis either for identifying the individual or for drawing conclusions about their condition. The procedure for assessing how far recognition errors seep into the features can be applied separately: it is not tied to the language or to a particular set of features, although here it was carried out for two features only, and it will be useful for any work that calculates metrics based on automatic transcriptions.

REFERENCES

[1] J.T. Hancock, M.T. Woodworth, and S. Porter, "Hungry like the wolf: a word-pattern analysis of the language of psychopaths", *Legal and Criminological Psychology*, vol. 18, no. 1, 2013, pp. 102–114.

[2] B. Gawda, "The differentiation of narrative styles in individuals with high psychopathic deviate", *Journal of Psycholinguistic Research*, vol. 51, 2022, pp. 75–92.

[3] M. Woodworth, J. Hancock, S. Porter, R. Hare, M. Logan, M.E. O'Toole, and S. Smith, "The language of psychopaths: new findings and implications for law enforcement", *FBI Law Enforcement Bulletin*, July 2012. Web: https://leb.fbi.gov/articles/featured-articles/the-language-of-psychopaths-new-findings-and-implications-for-law-enforcement.

[4] L.E. Cano Gómez, *Register analysis and linguistic manipulation in the murderers of the true crime series "I am a killer"*, bachelor's thesis, Universidad de Cádiz, Cádiz, Spain, 2020.

[5] C. Bellara, *Studying the linguistic and psychological markers in the speech of serial killers*, master's thesis, Purdue University, Fort Wayne, IN, 2025.

[6] G.Ya. Martynenko, *Osnovy stilemetrii* [Foundations of stylometry], Leningrad: Leningrad University Press, 1988. (in Russian).

[7] E. Stamatatos, "A survey of modern authorship attribution methods", *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, 2009, pp. 538–556.

[8] Y.R. Tausczik and J.W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods", *Journal of Language and Social Psychology*, vol. 29, no. 1, 2010, pp. 24–54.

[9] N.N. Sappa, "Evfemizmy kak psikhologicheskoe opravdanie v yazyke professional'nykh ubiyts" [Euphemisms as psychological justification in the language of professional killers], *Vestnik Chelyabinskogo gosudarstvennogo universiteta*, no. 37 (328), Filologiya. Iskusstvovedenie, iss. 86, 2013, pp. 102–104. (in Russian).

[10] N.V. Bogdanova-Beglarian, O.V. Blinova, G.Ya. Martynenko, and T.Yu. Sherstinova, "Korpus russkogo yazyka povsednevnogo obshcheniya Odin rechevoy den' (ORD): tekushchee sostoyanie i perspektivy" [The Russian everyday speech corpus "One Day of Speech" (ORD): current state and perspectives], *Trudy Instituta russkogo yazyka im. V.V. Vinogradova*, no. 3 (21), 2019, pp. 100–109. (in Russian).

[11] T.I. Popova, "Parametry chastotnogo slovarya v asimmetrichnykh i simmetrichnykh sotsial'nykh rolyakh govoryashchego" [Parameters of the frequency dictionary in asymmetrical and symmetrical social roles of the speaker], *Sotsio- i psikholingvisticheskie issledovaniya*, iss. 6, 2018, pp. 71–74. (in Russian).

[12] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision", in *Proc. 40th International Conference on Machine Learning (ICML 2023)*, PMLR, vol. 202, 2023, pp. 28492–28518.

[13] OpenAI, Whisper: automatic speech recognition system, 2022. Web: https://github.com/openai/whisper.

[14] Google, Google Colaboratory, 2024. Web: https://colab.research.google.com.

[15] I. Segalovich, "A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine", in *Proc. International Conference on Machine Learning; Models, Technologies and Applications (MLMTA'03)*, Las Vegas, 2003, pp. 273–280.

[16] Yandex, MyStem, 2023. Web: https://yandex.ru/dev/mystem/.

[17] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C.D. Manning, "Stanza: a Python natural language processing toolkit for many human languages", in *Proc. 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 101–108.

[18] Stanford NLP Group, Stanza: a Python NLP package for many human languages, version 1.8.2, 2024. Web: https://stanfordnlp.github.io/stanza/.

[19] J. Nivre, M.-C. de Marneffe, F. Ginter, et al., "Universal Dependencies v2: an evergrowing multilingual treebank collection", in *Proc. 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille: European Language Resources Association, 2020, pp. 4034–4043.

[20] K. Droganova, O. Lyashevskaya, and D. Zeman, "Data conversion and consistency of monolingual corpora: Russian UD treebanks", in *Proc. 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, Oslo: Linköping University Electronic Press, 2018, pp. 52–65.

[21] C.R. Harris, K.J. Millman, S.J. van der Walt, et al., "Array programming with NumPy", *Nature*, vol. 585, 2020, pp. 357–362.

[22] NumPy Developers, NumPy, version 1.26.4, 2024. Web: https://numpy.org/.

[23] W. McKinney, "Data structures for statistical computing in Python", in *Proc. 9th Python in Science Conference*, 2010, pp. 56–61.

[24] The pandas development team, pandas: Python data analysis library, version 2.2.2, 2024. Web: https://pandas.pydata.org/.

[25] The Apache Software Foundation, PyArrow: Apache Arrow Python bindings, version 16.1.0, 2024. Web: https://arrow.apache.org/docs/python/.

[26] O.V. Blinova, "Nizkochastotnye slova v russkom yazyke i podkhody k modelirovaniyu obshcheyazykovoy chastotnosti" [Low-frequency words in Russian and approaches to modelling general-language frequency], *Sotsio- i psikholingvisticheskie issledovaniya*, iss. 7, 2019, pp. 7–13. (in Russian).

[27] T.Yu. Sherstinova, "Naibolee upotrebitel'nye slova povsednevnoy russkoy rechi (v gendernom aspekte i v zavisimosti ot usloviy kommunikatsii)" [The most frequent words of everyday Russian speech (in the gender aspect and depending on the conditions of communication)], in *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Proc. International Conference "Dialogue 2016"*, Moscow: RSUH, iss. 15 (22), 2016, pp. 616–631. (in Russian).

[28] F. Yates, "Contingency tables involving small numbers and the χ^2 test", *Journal of the Royal Statistical Society*, Supplement, vol. 1, no. 2, 1934, pp. 217–235.

[29] P. Rayson and R. Garside, "Comparing corpora using frequency profiling", in *Proc. Workshop on Comparing Corpora (ACL 2000)*, Hong Kong, 2000, pp. 1–6.

[30] E.V. Paducheva, *Semanticheskie issledovaniya: Semantika vremeni i vida v russkom yazyke; Semantika narrativa* [Semantic studies: the semantics of tense and aspect in Russian; the semantics of narrative], 2nd ed., Moscow: Yazyki slavyanskoy kul'tury, 2010. (in Russian).

[31] V.A. Plungian, *Vvedenie v grammaticheskuyu semantiku: grammaticheskie znacheniya i grammaticheskie sistemy yazykov mira* [Introduction to grammatical semantics: grammatical meanings and grammatical systems of the world's languages], Moscow: RSUH, 2011. (in Russian).

[32] A.A. Kibrik, *Analiz diskursa v kognitivnoy perspektive* [Discourse analysis in a cognitive perspective], Doctoral dissertation in the form of a scientific report, Moscow, 2003. (in Russian).

[33] N.V. Bogdanova-Beglarian, O.V. Blinova, K.D. Zaides, and T.Yu. Sherstinova, "Korpus Sbalansirovannaya annotirovannaya tekstoteka (SAT): izuchenie spetsifiki russkoy monologicheskoy rechi" [The corpus "Balanced Annotated Text Collection" (SAT): studying the specificity of Russian monological speech], *Trudy Instituta russkogo yazyka im. V.V. Vinogradova*, no. 3 (21), 2019, pp. 110–125. (in Russian).