Sociolinguistic variation in Slavic languages

Serge Sharoff, Nenad Ivanović

The Cambridge Handbook of Slavic Linguistics. 2024. Edited by Browne, Wayles and Sipka, Danko. Cambridge University Press

Abstract

This chapter discusses linguistic variation in Slavic languages by presenting an overview of the relationship between human communication in the society and the corresponding linguistic features. In this chapter we will focus on the parameters of variation according to the language user, such as age or dialects, and according to the language use, such as communicative functions or communication styles, e.g., politeness. In this chapter we are going to focus on both qualitative and quantitative methods for studying aspects of sociolinguistic variation. The examples are drawn from large corpora of two Slavic languages, Russian and Serbo-Croatian, with a particular focus on academic writing, news reporting and reporting personal experience in social media and from dictionaries and field studies.

1 Introduction

1.1 The notion of sociolinguistic variation

Studies of sociolinguistic variation provide the basis for investigating the relationship between the society and language (Holmes, 2013). This chapter outlines approaches to studying linguistic variation in Slavic languages by investigating both quantitative and qualitative aspects of variation with the focus on the relationship between human communication in the society and the corresponding linguistic features. Methodologically, in this chapter we will approach the notion of sociolinguistic variation according to the language user and the language use (Gregory, 1988):

> **variation according to the language user** covering such parameters of variation as the age, gender, region, socio-economic status and other sociologically distinctive features (Crystal, 2004, pp. 286, 364)
>
> **variation according to the language use** covering such parameters of variation as communicative functions, e.g., academic writing or regulatory legalese, and communication styles, e.g., politeness or adherence to language standards (Holmes, 2013).

Analysis of sociolinguistic variation starts from the assumption that language has its role in the process of shaping the society, its concepts and beliefs, while in turn it is shaped by them (Halliday, 1978). In this view, meanings to be exchanged in the society and linguistic constructions used to express those meanings are shaped by millions of interactions between the individuals on the micro-level. The totality of those interactions leads to establishing expectations within the society on the macro-level, both with respect to the language user, as a member of various social groups, and with respect to the language use, as adaptation which constrains features appropriate in specific social contexts. For example, the authors of academic articles come from different age groups, but the social context of sharing academic knowledge guides them in choosing linguistic constructions expected in this sphere of communication, rather than what is expected in their age group. In its turn the interface between the micro- and macro-levels is regulated (1) through formal language policies, which also change over time depending on the level of their acceptance in the society, and (2) through informal shaping of communication by influential speakers, such as writers, editors or educators acting as the gatekeepers setting the expectations on what is appropriate in a given social exchange. For example, the PhD advisors along with journal editors need to act as the gatekeepers for their students in the context of academic writing.

The theoretical framework we are going to use in this chapter takes into account concepts from both formal and functional linguistics. The focus of functional linguistics is aimed at investigating a link between how language functions in the society and which resources it has to express those functions (Halliday, 1985). This implies the need to deal with variation. This concerns variation in the linguistic functions, which evolve as the product of culture (in some societies the communicative function of academic writing is not sufficiently distinguishable from other kinds of communication), as well as variation in the lexicogrammatical resources as codified linguistic expectations (expressing an argument in favour of wearing masks in a Facebook forum is different from how the same argument is expressed in a research paper).

The very notion of sociolinguistic variation with respect to the user refers to 'identity' as a factor of social cohesion, so it is important to emphasize that this variation does not influence the communicative function of language as a whole, but rather works on the microlinguistic level(s), for example, with the sense of social identity of various social groups of speakers of a given language. Therefore, sociolinguistic variation can be analyzed from the linguistic standpoints communicating with the fact that language and its contextual

features can be analyzed as a code. These standpoints are: the language policy and language planning (with respect to the fact that much of this process is related to the spatial and social status of the language, (Mesthrie et al., 2009, pp. 373–374); also merging functional and discourse-oriented approaches.

Modern study of sociolinguistic variation entails a wide range of research ideas, both microlinguistic (the study of variation of the language elements, e.g., specific words, in different types of texts / spoken varieties) and macrolinguistic (the study of different types of language varieties, e.g., regional, as separate idioms).

The focus of formal linguistics is on the distinctive features of language variation, which can be divided into several types, most of which will be taken into account in this chapter: orthographical features (e.g., spelling differences in language variants); phonological features (e.g., the distinctive use of phonemes in dialects); grammatical features (such as word-formational, inflectional, constructional and other); lexical features (the choice of the vocabulary in relation to sociolinguistic factors); discourse features (e.g., the structural organization of a text) and so on (Crystal, 2004, pp. 7–9).

Generally speaking, there are two main determiners which play a role in controlling the direction of the discourse: global (e.g., topic markers and topic shifters), and local (exemplifiers, relators, evaluators and so on). Distinctive features which determine sociolinguistic markers can be found in both, on various levels of linguistic analysis.

While we cannot attempt to overview studies of sociolinguistic variation across the entirety of Slavic languages, we will examine modern **methods** for studying sociolinguistic variation. More specifically we will examine the typology of sociolinguistic layers with respect to variation according to the language user and the language use, while the examples will be provided by a contrastive study which compares variation parameters in two fairly distant Slavic languages, Russian and Serbo-Croatian. See also reviews of specific studies in Slavic sociolinguistic variation (Belikov and Krysin, 2001, Lauersdorf, 2009).

The research of sociolinguistic variation in Serbian linguistics dates back as far as the half of the 20th century. It has developed from dialectology as a study of geographical differences in speech: phonological, morphological, grammatical and lexical. Mainly in the second half of the 20th century, starting from the 1960-ies, it is noticed that the „main picture“ of dialectal distribution had to be supported by further language differentiation according to generation, sex, educational level and other social factors. In other words, besides the term „geographical space“, a term „social space“ in linguistic research hat to be considered, as well. The notion of „sociolect“ was born (Bugarski, 2009, p. 15).

In time, the notion of sociolinguistic variation in linguistic research gained autonomy, developing its own methodological apparatus. Inspired by Labov, Trudgill, Milroy and other researchers, Serbian sociolinguists introduced the term *variety*, as „any publicly pronounced form of a language: geographical, social, professional etc.“ (Bugarski 2009: 23). In this way, the horizontal dimension of language research, to which the term *dialect* was appointed, gained new, vertical dimension. In recent time, the term *vernacular* is being used instead of *variety.*

In Serbian linguistics, there were two main moments which marked the transition from 'classical' or 'rural' dialectology towards the research of language varieties (although the 'classical' dialectology continued to develop on its own until today). The first moment was rising interest for urbanization and the language of urban population, which can be traced back to 1920's, but has reached its full potential from the 1980's onwards, with P. L. Thomas, Lj. Rajić and other researchers (Bošnjaković, 2009, p. 49). Urbanization as a linguistic subject was analyzed both from dialectal and standardization viewpoints. These researches have developed from an assumption (which was later confirmed) that social changes and language changes are in causal relation (Jović 1978: 496). The second moment was (implicit or explicit) introduction of the term *variable* as a social marker causally related to language behavior. Later researches showed that sociolinguistic variables can be found along the lines of the following social markers: class identity, occupation,

sex, movability, age, region of origin, education, personality and affinity towards social networking (Rajić, 2009) . All of these variables were introduced in the publication *The Speech of Novi Sad* (Volume 1: theoretical foundations, phonetic features, 2009; Volume 2: morphosyntac, lexical and pragmatic features, 2011), a publication which fully revealed this type of sociolinguistic research both in Serbian and Slavic linguistics.

In conclusion, it can be said that, although research of sociolinguistic variation in Serbian linguistics does not follow the "waves" described in (Eckert, 2012), neither historically nor methodologically, it does not contradict them either. The first wave, according to which the term *variation* can be described in relation to "standard" language, more or less correlates with "classical" dialectological research of vernacular(s) in the first half of the 20th century. However, the understanding of the language variation without the polarization towards the language standard, which is the main feature of the second wave, and the rise of "social identities" in which the speakers place themselves through stylistic practice, which is the main feature of the third wave, both constitute the habitus of research of sociolinguistic variation in Serbian linguistics from the 1990's onwards.

1.1.1 1.2 Sources of evidence

In this description we will rely on two primary sources of information: quantitative research on the basis of large corpora and dictionary descriptions on the basis of fieldwork-based studies. This provides a complementary perspective, as large corpora offer an objective view of how linguistic constructions are actually used by a large number of speakers in a range of communicative situations. At the same time, corpus research focuses on written texts because of their availability in electronic form, so other sources are needed, such as field studies to focus on phenomena more widely manifested in spoken language.

1.1.2 Corpora

The first source comes from *monolingual corpora*, as they provide examples for qualitative analysis, as well as counts of detectable forms for quantitative analysis, which can be at the level of words, morphemes or syntactic constructions (for corpora with syntactic annotation). This helps in comparing the frequencies of forms with the functions expressed by those forms, even when the link between the forms and the functions is rarely one-to-one (Sharoff, 2017).[1] Corpora as collections of texts have been used since the computers became powerful enough to process large volumes of texts, see (Kučera and Francis, 1967). This has been applied to studying sociolinguistic variation, e.g., (Reppen et al., 2002) with one of the studies in that collection focusing on variation in expressing deontic or epistemic modality via *dolzhen, nado* or *nel'zja* in two Russian registers (news vs fiction) using the Uppsala corpus of Russian (de Haan, 2002). For a general overview of methods to study sociolinguistic variation using corpora, see (Andersen, 2010).

Since the beginning of the 2000s the amount of texts available on the Web democratised the process of collecting large corpora via crawling large samples (Sharoff, 2006, Baroni et al., 2009). Since a Web snapshot for a given language provides the closest approximation to creating comparable descriptions, the following Web corpora are used in this study:

hrWac a corpus of 1.3 billion words, 574,000 pages, produced by crawling Serbo-Croatian language websites from the .hr Internet domain (Ljubešić and Klubička, 2014) ;

ruWac a corpus of 2.5 billion words, 2 million pages, produced by crawling Russian language websites without restricting the Internet domains (Sharoff et al., 2017);

[1] See also Chapter X on Corpus Linguistic in this volume.

ukWac a corpus of 2 billion words, 2.5 million pages, produced by crawling English language websites from the .uk Internet domain (Ferraresi et al., 2008);

GICR a corpus of 20 billion words, which consists of Russian social media posts with information about the age, place of origin and gender of their authors, primarily from Livejournal.com and VK.com (Belikov et al., 2014).

Table 1: Region-specific subcorpora for Russian bloggers in Livejournal

%	Words	Location
37.29%	1598317700	NA
21.02%	900987080	Moscow
5.49%	235093320	St Petersburg
4.76%	203762440	Ukraine
1.69%	72221940	Israel
1.39%	59565660	Belarus
1.03%	44241820	USA
0.95%	40673500	Moscow region
0.80%	34392680	Yekaterinburg region
0.78%	33428360	Novosibirsk region
0.76%	32430020	Germany
0.51%	21809480	Samara region
0.47%	20061300	Latvia
0.44%	18765880	Estonia
0.43%	18363940	Krasnodar region
0.42%	18060980	Canada
0.40%	17047660	Rostov region
0.39%	16534700	Bashkortostan
0.38%	16467360	Chelyabinsk region
0.37%	15855280	Tatarstan
0.37%	15820000	Perm region

These corpora will allow us to capture variation through quantitative study of contexts. GICR from this list is particularly important as the language users in its texts can be described via demographic parameters, thus providing a text-external link to the lexicogrammatical features. Table 1 lists Russian administrative regions with the largest amount of texts in the Livejournal portion of GICR. Some regional indicators, such as Moscow, St Petersburg or the USA, are less interpretable with respect to their dialectal variation, as these are the destinations of mass migration. However, for smaller regions this corpus provides authentic examples from million-word corpora to study relevant features.

1.1.3 Dictionary descriptions

The second source comes from *monolingual descriptive dictionaries* as a special kind of linguistic manuals which provide a unique view to sociolinguistic variation. Given their primary task — to describe language on both syntagmatic and paradigmatic level — descriptive dictionaries by nature have in their sources a large variety of documents representing different types of texts or kinds of discourse. Furthermore, the metalanguage of a dictionary is in most cases structured so as to mark different kinds of sociolinguistic variation

(everything which is not “standard” to the lexicographer must be explicated in some way, usually by means of specific labels (e.g., “jarg.”(on), “nonlit.”(erary) etc.)). External indicators of sociolinguistic variation in dictionaries (elements of dictionary metalanguage) can be turned into variables and corelated with variables of internal indicators of sociolinguistic variation (e.g., the typological lexicogrammatical features of lexis which is being defined). Because of this, descriptive dictionaries – when they are being processed by quantitative methods – may provide useful additional/control information to the linguist who explores genre variation in the Web corpora and vice versa (Hanks, 2012).

A survey of the dictionary users (Šipka, 2021) shows that primary normative labels (those with primary purpose to exclude the word from the formal standard language variety, such as slang, colloquial, etc.) have a higher excluding effect than secondary normative labels (which mark something else, but have a secondary effect of excluding, e.g., facetious, obscene, etc.). This proves that descriptive dictionaries have an impact on the process of language standardization.

1.2 On different views on sociolinguistic variation

1. **The standardization view** Conception of “standard” or “literary”

language: This conception is based on the view that the contemporary language represents an entity which has undergone four main stages of language planning: selection, codification, implementation and elaboration, followed by other stages such as: acceptation, expansion, cultivation, evaluation (Haugen 1987, in: Mesthrie et al., 375 and Radovanović 2003: 190). The core of the “literary“ language is based on the rules governing the linguistic levels: orthographic, lexical, grammatical, and the like, which represent the language standard. On the other hand, the periphery of this language includes various kinds of atypical or non-standard language use. In other words, the central zone of the “standard” language represents the literary (written) language in standardized form; whereas, the peripheral zone is marked by various cases of vernacular, dialectal, obsolete or jargonistic language use.

This setting implies Schuhart’s view on relation between language of individual and language of collective (in: Belić 1951 and Kovačević 2014: 31–32). Individual participates in the building and institutionalizing the language of collective; collective serves as a corrective to the individual through accepting or not accepting its innovations. This “genetic” relationship is open to conceptualizing through various kinds of sociolinguistic models related to notions of “center/periphery” and “standard/non-standard”. Furthermore, this setting also implies that language users can “delegate” their language collective.

In historical overview it is important to conclude that the standardization view in its entirety is based on the notion of language cohesion among the members of one nation – therefore a national language is often perceived as a necessary condition for a nation to exist.

The main stages of language standardization can be explained in the case study of standard Serbian (Serbo-Croatian) and Russian languages from their origins to the present day. Contemporary Serbian is based on Shtokavian dialect, which is differentiated from other dialects (Chakavian and Kajkavian) by the pronunciation of the interrogative pronoun *što* (as opposed to *ča* and *kaj*, respectively). What is taken to be the basis of the language standard is the state of the Serbian language immediately following an overall language reform (orthographical, grammatical, lexical), which was undertaken by Vuk Stefanović Karadžić in the 19th century.

The modern Russian standard emerged first from the Central Russian dialects when Moscow

was gaining prominence over other Russian speaking regions during the ?? century, while in itself this dialect incorporated some features of the earlier Northern Russian dialects (e.g., the 'hard g' consonant) and Southern dialects (reduction of unstressed vowels). This was followed by its extensive expansion over the realm of the Russian empire towards Siberia and the Far East. The literary standard was also heavily influenced by borrowings from a range of European languages, such as Polish, Dutch, French and German (Timberlake, 1993).[2]

1. **The view from communicative functions** In addition to variations

coming from the user, variations in the sociocultural context of communication shape the linguistic constructions in various ways. First, the sociocultural context sets expectations on the kinds of communicative intentions suitable in a given situation. To describe them we need a typology of communicative functions. Second, it sets expectations on the kinds of linguistic constructions appropriate for expressing the communicative intentions, this is what Halliday (1978) refers to as "register". Thus, we need a typology of registerial features.

The choices available for the realization of communicative intentions are instantiated in individual texts according to individual preferences of their authors. However, statistics obtained from a large corpus, anonymizes these individual preferences and provides a map of how the system of language connects the communicative functions with their realizations. The intersubjective nature of stylistic expectations also leads to the relative stability of typical genre-related ways of linguistic realizations of communicative intentions. Following Bakhtin's wording: "each separate utterance is individual … but each sphere in which language is used develops its own relatively stable types of utterances" (Bakhtin, 1986, p. 60) .

Therefore, we will address in this chapter the communicative functions available in existing corpora following the framework of (Sharoff, 2018) followed by statistical analysis of register features associated with those functions following the framework of (Biber and Conrad, 2009). We start from the assumption that the kinds of sociocultural contexts are mostly compatible across modern societies (Francophone, Slavic or otherwise), especially when considering data from the Web corpora. However, the parameters of sociolinguistic variation across languages differ with respect to the following factors:

variation in communicative functions If corpora are collected from different sources or via different pipelines, such as newspapers or social media, the distribution of communicative functions is likely to be different. The composition of available corpora can also differ because of sociocultural differences in the frequencies of the functions, such as the preferences for the balance of argumentation or factual reporting in newspapers.

standard ways for expressing communicative functions Usually functions are associated with ways for their realisation acceptable in the society. However, some cultures can lack codified lexicogrammatical features for realising specific functions. For example, reporting in citizen journalism does not necessarily follow the established journalistic conventions, thus shifting the registerial choices. Alternatively, the gatekeepers are likely to influence the frequencies of registerial features accepted in the prestigeous genres, such as academic writing or fiction.

language-specific linguistic features Finally, the functions can be similar and well codified, while they can be expressed via language-specific mechanisms. For example, the communica-

[2] See also Chapter X on the linguistic history of Slavic languages in this volume.

tive function of narration is commonly associated with the higher rate of temporal adverbials and verbs in the past tense, unless a culture prefers expressing some kinds of narrative reporting in the present tense in the form of "historical present". Similarly, the argumentative texts are often characterized by the higher rate of explicit causation markers and emphatics. However, in certain cultures their use might be discouraged by the gate keepers.

1. **The interpersonal view** This view falls into the scope of research

of the "ethnography of communication". Dell Hymes developed a checklist of dimensions of sociolinguistic awareness that are involved when speakers communicate in particular speaking communications: genre, topic, purpose (or function), setting, key (emotional tone), participants, message, act sequence, rules of interaction, norms of interpretation (Hymes, 1971, Hymes, 1974).

More specifically, we will discuss methods for investigating such kinds of variation as:

• with respect to the language standardization model:

• regional variation - division according to:

• center : periphery (literary vs vernacular language)

• standard language : dialects

• temporal variation

• social variation (jargon and slang)

• idiolectal variation

• with respect to registers (communicative functions and their lexicogrammatical features):

• argumentative

• news reporting

• personal reporting

• academic writing

• with respect to interpersonal context of language use:

• politeness

• code switching

2 Variation with respect to the language standardization model

Variation with respect to the language standardization model implies the existence of levels of sociolinguistic variation which can be presented in terms of their "peripheral" relation to the language standard. In other words, the nature of these levels relies on their linguistic deviation from the "standard language", the "standard" being understood as a set of criteria for selecting the correct language expression in the society. In many Slavic languages, definitely in the two languages described here, these so-called standard varieties are privileged over the other regional and social varieties. Having in mind that the notion of the language standard does not present one homogeneous whole, it is understandable why this type of variation correlates to social and language stratification. The most common types of variation in Slavic languages are: 1) regional

(division according to various forms of regional and/or vernacular language use); 2) social (which implies that different social levels of speakers share common linguistic markers); 3) temporal (division according to the time of language use); and 4) idiolectal (which puts different kinds of "incorrect" language use related to the sense of identity of the speaker(s)). However, this division should be taken conditionally, given that, linguistically, these types of variation do not form isolated wholes, but rather overflow into each other.

2.1 Regional variation

In the sociolinguistic sense, "regional variation" represents complex notion which combines different linguistic approaches to the notion of "region".

The most common approach treats the "region" in a geographical sense, as a set of language features pertinent to the territory where a certain idiolect or dialect is spoken. This sense also entails the fact that "regional" use of language concerns the lexis naming the objects and terms from various aspects of everyday life, i.e., the names of the seasons, plants, folk remedies, domestic and wild animals, fruit, agricultural tools, church paraphernalia, terms of common law, etc.

Features of regional variation in a geographical sense function on the broad range of linguistic contexts, from the constructional (accentual, phonological, morphological …) to the lexical-semantic. For instance, in the *Dictionary of the Serbian Academy* the same label, "pokr." ("regional") is applied in wide range of cases: to the lexemes used specifically in certain regions (e.g., *lotnjak* (n) pokr. … "olive oil of a good quality" (Poljica); *nikolča* (n) pokr. … "national dance" (srednji Timok); to the regional phonetic variants of the lexemes of the standard language (e.g., *mrmnjati* (v) pokr. … „mrmljati" (to murmur); *nedomak* (prep) pokr. … „nadomak" (within reach)); and to the culturally specific meanings or constructions of the polysemous words (e.g., *kuća* (n HOUSE) … pokr. „dečja igra" (a children's game); *negodovati nekoga* (v RESENT + Gen. [Anim.]) … pokr. „osuđivati nekoga" (to judge someone)), etc.

A narrower subtype of regional variation can be considered a dialectal variation, which is directed towards systemic changes in the spoken language. Dialectal variation is most commonly analyzed through linguistic atlases, the collections of linguogeographical maps which enable their users to analyze areal dissemination of dialectal features which belong to different levels of language use (Miloradović 2012: 141). These atlases often represent the results of international projects, and their making is measured in decades. The most important atlases in Slavic world are: European linguistic atlas – ALE, Slavic linguistic atlas – OLA and Carpathian dialectological atlas – OKDA; there are also national projects, such as the project named Serbian dialectological atlas – SDA (earlier, Serbo-Croatian dialectological atlas). The central result of linguistic atlases is the production of linguistic charts, which represent phonetic, morphological, syntactic, semantic and lexico-derivational variation of languages in one or several areals (see: http://slavatlas.org).

As dialectal language feature can be realized on different levels of the organization of the language structure, this also implies various issues regarding the dialects (as linguistically complex forms) in relation to sociolinguistic perception of (literary) language and its variation. The research implemented by V. Karlić and S Šakić (Karlić–Šakić 2019) analyzes the language of literature written by Serbian writers from Croatia whose works were published after 1991 by the Serbian Cultural Society Prosvjeta's publishing house in the edition *Mala plava biblioteka*. The paper shows that lexical choices between "Serbian" and "Croatian" idiom employed by the writers (*sveštenik – svećenik, hrišćanin – kršćanin* and so on), especially when it represents the dialogue between the characters, can be context-dependent.

Social media corpora like GICR for Russian provide information about author profiles to study examples of sociolinguistic variation. For example, GICR can be used to test the regional distribution of such phenomena as names of professions (*himichka, himitsa* 'chemistry teacher') or food (*rasstegaj, rastjagaj* 'a specific kind of pies') with the differences clearly depending on the origin of the speaker (Belikov, 2010). At the

same time, modern society is characterized by extensive demographic movements, which dilute the specific geolinguistic features for the destinations of mass migration, as well as by innovations, which create new phenomena enriching the previously known distributions of dialectal features. For example, GICR shows a clear association of new words or senses with specific regions, such as *multifora* 'plastic wallet' or *svechka* lit. 'candle', in the sense of 'high-rise tower' (Belikov et al., 2014).

Development of mass transportation as well as socio-political upheavals of the 20th century have led to extensive demographic movements, especially in the context of the use of Russian in the Soviet Union, for example, mass movements to Siberia and Kazakhstan in the 1950s and 1960s, or the mass migration to the Moscow region in the 1990s. This all has led to very extensive contacts across the regional varieties resulting in the lack of well-defined dialectal features known from the rural communities due to the population mixing. Therefore, from the corpus viewpoint, it is difficult to provide definitive features which can describe the language of the destinations of mass migration as they are indicated in the current profile in Social Media accounts.

2.2 Social variation

In relation to the language standard there is usually notion of two kinds of social variation: jargon (occupational registers) and slang (subcultural and youth registers).

Jargon variation usually implies the existence of special registers related to various scientific and cultural fields (finance, politics, medicine etc.). Lexical units belonging to these registers can exist as separate, terminological units, or they can be integrated into the standard language either by terminologization (left (n) … "leva ruka" (the left hand) → pol. "members of revolutionary or liberal parties") or by determinologization (constitution (n) … pol. "the main law which determines rights and duties of citizens in a state" → "a set of physical or mental characteristics").

On the other hand, there are several types of divisions in the slang variation. The most common are the following: divison by the age of a speaker (slang as a characteristic feature of language of the youth), divison by the social status of a speaker (slang as a subcultural entity with the task to provide mutual understanding of speakers belonging to the same social structures), and division by speaker's education (Bugarski, 2006, Vujović and Alanović, 2011). The main linguistic characteristics of slang are in the tendency towards imaginative and vivid lexical creations, which opens several systemic possibilities for the analysis of its distinctiveness. These characteristics are: assigning new, metaphorical meanings to existing words (*krvav* (adj) (bloody) … "odličan, izvanredan" (excellent, remarkable)); distortion of the rules of language creation trough permutation of voices in a word (*vozdra* instead of *zdravo* (hello)); shortening of words or even just using their initials (*profa* (professor); *za dž* (za džabe, for free)); wide adoption of lexical borrowings (*picikato* (Ital. pizzicato "gentle"), *pis* (Eng. piece "small amount of drugs") and so on (Ibid, 22).

The slang features in Slavic languages are mostly researched in urban areas, where several non-language factors take place: change of socio-political concepts, technical and scientific progress, development of educational activities etc. These research studies show that sociolinguistic variation can be investigated through morphosyntactic features of words as well. The most prominent examples are the variation of lexical doublets and various case constructions according to age and occupation of the speaker (e. g. *ostareti – ostariti, zbog toga – radi toga,* etc.) (Vujović and Alanović 2011: 46–51).

However, the most productive features of slang in Serbian at the word level stay in the mechanisms of word formation. For instance, there is a significantly large number of expressive suffixes which, in the standard language, have fairly marginal, obsolete or informal use, but are highly productive in slang usage: -džija (tabadžija, tupadžija), -uša (uspijuša, bilderuša) etc. In addition to this, a part of the lexical inventory of slang is made by compounding and blending (radoholičar (workaholic), čedovišta (sweetchildmonsters)

etc.). For a detailed list of suffixes and compounds in Serbian slang see in: Bugarski 2006: 238–274 and 275–280). In addition to this, in the urban environment it is common for residents to develop different derivational models of gentilics (*Pejtonac – Pejtočanin – Pejtončan – Gradpejtonac – Gradić Pejtonac* etc.) (Štasni and Ajdžanović, 2011).

2.3 Temporal variation

2.3.1 Descriptive studies

In relation to the vertical, time-related, dimension of the "standard language", there are layers which can be characterized as "archaic" or "obsolete"; and "new".

The "archaic" layer: lexis which belonged to the "higher" styles in the history of Serbian literary language develops anachronous relation to the language standard (in Serbian, usually from the Slaveno-Serbian period: *otvečanije, vinodelije*): today it can be used with highly expressive function

The "obsolete" layer: lexis, idioms, meanings obsolete in relation to better choices (plav (adj. "blue") in the meaning of "koji ima svetliju nijansu, svetao" ("lighter"): 'yellow can be more or less blue') etc. It can be mixed with the dialect (in obsolete dialectal words).

The "new" layer can be attributed both to lexis and its meanings (e.g., new lexical borrowings (*peč* (patch), *strimovati* (to stream) …): in the standardizational sense, it signifies that the word is not yet accepted.

Variation with respect to the time of language use (diachronic level of sociolinguistic variation): there is also a model, to which this use is related. The model represents the "modern" or "present-day" usage of the standard language. In relation to this model, language can be "archaic" or "new" (with further subclassification in both levels, e.g., "archaic" and "recently obsolete"). Differences can be established on the levels of word-formation, word origin, syntax and so on.

2.3.2 Corpus evidence

The primary source of corpus evidence comes either from historical corpora, such as the Russian National Corpus (Sharoff, 2005) or from social media corpora annotated with author profiles, such as GICR (Belikov et al., 2014). Historical corpora are suitable for detecting variation over large time intervals, while at the same time they are limited by the availability of their sources, such as fiction and legal texts predominantly used in the historical part of the RNC, as the language of formal writing can be preserved much better, but it shows only a small portion of language use, which is also very much influenced by the gatekeepers at a specific time.

On the other hand, social media have made it possible to capture the language of spontaneous interaction in everyday life, while at the same time they are limited with respect to the relatively recent time period. Any sizeable social media collections are available from 2000s and their authors are mostly aged from 18 to 60, see the distribution of the number of blog posts for the authors who have explicitly provided their year of birth in GICR 1. A number of data points in a study of this kind can be expanded further via automatic age prediction for the authors who have not provided their age (Nguyen et al., 2016), but this has not been attempted for Slavic languages yet.[3]

Another limitation of relying on corpora is that they often exhibit topical biases leading to predictable prevalence of specific topics, such as dating and education for the authors younger than 22 or history and illnesses for those older than 55. Nevertheless, corpora allow detection of more interesting patterns, in particular through associating **grammatical** properties with age differences. This can be done by building a

[3] See also information about text classification and author profiling in Chapter XX on Computational Linguistics in this volume.

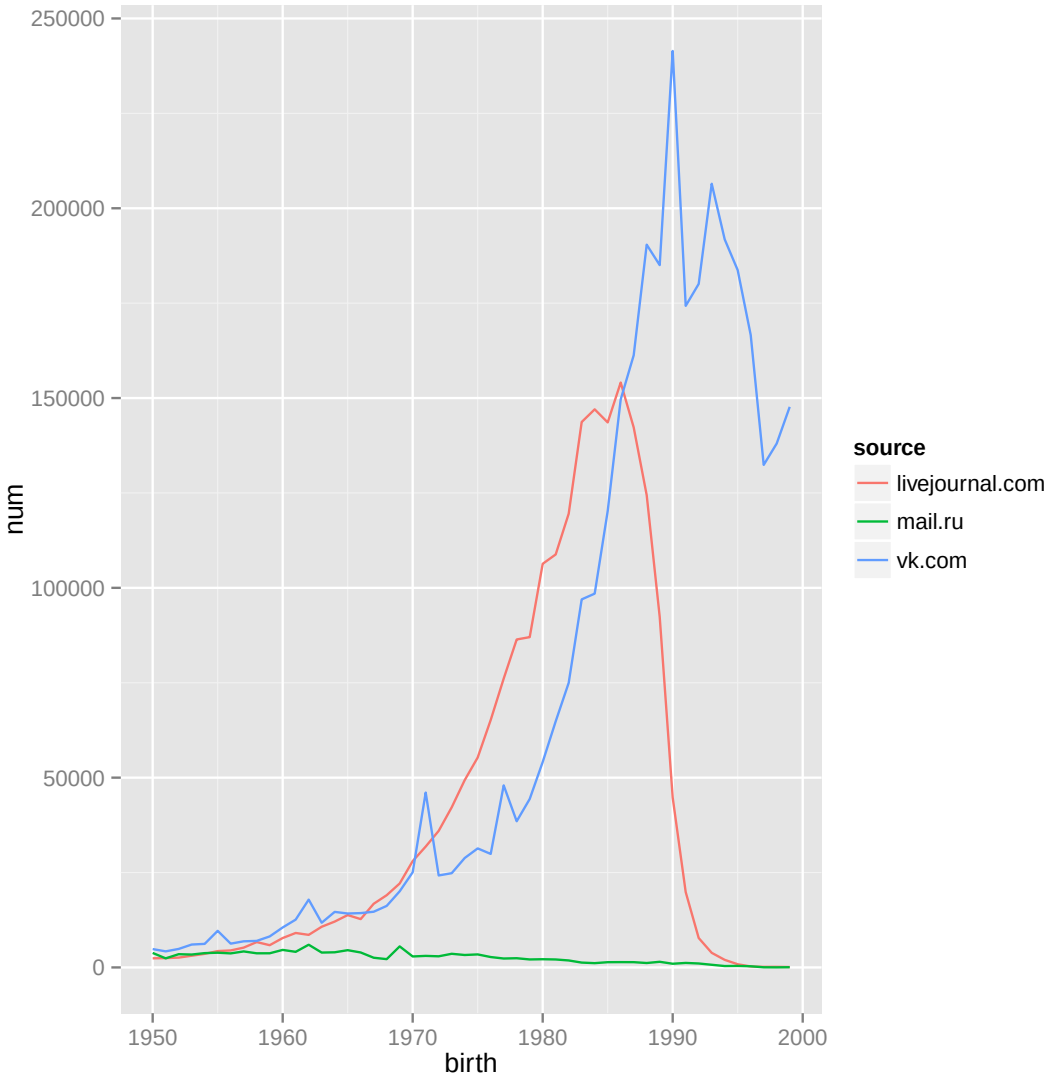

Figure 1: The number of blog authors in GICR with respect to the years of their birth

statistical model predicting the author's age on the basis of selected lexicogrammatical features, such as those from (Biber, 1988). For example, a linear Support Vector Regression model (Cherkassky and Ma, 2004) can predict the author's age in GICR using the grammatical features with the mean absolute error of 8.5 years. This model on the basis of modern GICR data for Russian associates higher positive weights (leading to positive correlation with age) for such features as the rate of third person pronouns, hedges, amplifiers and nominalizations, while higher negative weights (leading to negative correlation with age) are exhibited by such features as the rate of first person pronouns, relative subordinate clauses and omissions of *chto* ('that') after verbs of reporting. This suggests greater acceptance of more personal, less formal communication in the blog posts of younger writers.

2.4 Idiolectal variation

In Slavic languages, idiolectal variation implies idiosyncrasies. These idiosyncrasies function as transitional forms between language of collective and language of individual(s) (levels of "langue" and "parole"). This phenomenon is most frequently researched on lexical and semantic levels of language structure.

On lexical level, it is noticed that speakers which belong to certain regional, culture, educational or age groups use non-standard or sub-standard phonetic realisations of certain words signifying notions in everyday language use. These phonetic realisations are marking the identity of abovementioned groups. For instance, forms like: *bicikli* (instead of *bicikl*, 'bicycle'), *utornik* (instead of *utorak*, 'Tuesday'), *šaraf* (instead of *šraf*, 'screw') etc. are considered to be used more frequently among the older Serbian native speakers from Novi Sad. Idioms in active language use also fall into this classification, for instance *izvoditi kerefeke* instead of *izvoditi nešto* ('to act out'), or *mani me* instead of *ostavi me na miru* ('leave me be'). Researchers of this level agree that variables of age and the sense of local identity are the most dominant in this division (Štrbac and Vujović, 2011). Idiolectal deviation from the language standard can also be marked by morphonological irregularities. In Serbian, these irregularities can be related to deviation from orthoepic norm (*beleti* instead of *beliti, izvrsan* instead of *izvrstan*), to the point of (more or less widely accepted) incorrect pronunciation (*gledaoc, nosioc* instead of *gledalac, nosilac*).

The same applies to the idiosyncrasies on lexical level of the language structure. Non-standard or sub-standard realisations here imply that the language users of certain groups use loan-words, archaisms etc. which have the same meaning as 'standard' words, but with different connotation (for instance, *kanda* instead of *izgleda*, 'it seems so', or *špacirati se* instead of *šetati se*, 'going out'). When used in speech, these forms are used as recognizable features of their users' identity.

The most known idiosyncrasies to lexicographers are hapax legomena and the uncommonly used words or potential words. Both types are a result of non-standard procedures in the word formation of nouns, adjectives, adverbs and verbs. Idiosyncrasies are commonly used in the language of literature, where they serve as a way of expressing the personality of their author as well as his thought. In Serbian, individualisms can arise as a result of analogy to more common words (*zrakoproliće* ("beamshed") in analogy to *krvoproliće* ("bloodshed")), or by connecting or compounding the base of the word with unusual or unexpected derivational affix or a word (*mladoženjstvo* ("groomness")). Potential words, on the other hand, usually arise by analogy, and signify penetration of new derivational models into standard language (*detovati* ("to spend time as a child"), *kolumbovati* ("to spend time acting as Columbus")).

The conclusion to be drawn from this is that variation with respect to the language standardization model does not imply the existence of whole new idioms which deviate from the standard language, but rather a set of recognizable landmarks on different levels of language structure (morphonological, semantic, word-formation, pragmatic…). Overall, word-formation and lexical levels seem to be the most active in this division.

3 Variation with respect to communicative functions

3.1 Distribution of communicative functions

As mentioned by Douglas Biber, "language may vary across genres even more markedly than across languages" (Biber, 1995). Nowadays it is relatively easy to collect very large samples from the Web to build representative corpora. However, there is often a lack of understanding of the composition of those corpora with respect to their inherent variation in terms of communicative functions codified via genres.

From the viewpoint of the categories for describing the genres, a large number of labels is needed to account for a large number of different kinds of texts in a large corpus. However, from the viewpoint of their annotation and meaningful comparison, we needs to be organised in smaller groups for the immense variety of variety. It is difficult in practice to compare corpora using even the 70 genres from the BNC typology, so many genre variation studies focus on smaller subsets of 10-15 genres (Lee and Swales, 2006, Szmrecsanyi, 2009). At the same time, even the full BNC genre set is far too small to describe variation with respect to very common Web genres, such as personal blogs or discussion forums.

Another difficulty comes from the fact that the Web is a reasonably unconstrained publication medium, so many Web pages are produced without explicit gatekeepers such as editors or reviewers. In the end, Web corpora often contain examples of genre hybridism, for example, citizen journalism, which combines news reporting with personal observations, thus blending established genre categories.

In this study we will follow a topological approach to describe the variation of communicative functions by following the framework from (Sharoff, 2018). With the use of a small number of categories, the communicative functions for each text in a Web corpus can be analysed with respect to how similar the text is to prototypes. The communicative functions common on the Web are:

A1.Argumentative To what extent does the text try to persuade the reader? (For example, *argumentative blog entries, newspaper opinion columns*).

A8.News To what extent does the text appear to be an informative report of events recent at the time of writing? Information about future events can be considered as reporting too. (For example, *reporting newswire story*).

A11.Personal To what extent does the text report a first-person story? (For example, *diary entries, travel blogs*).

A12.Promotion To what extent does the text promote a product or service? (For example, *adverts, promotional publications*).

A14.Academic To what extent does the text report academic research? (For example, *academic research papers*).

The prototypes are given in the lists of examples, so that a citizen journalism blog entry can be judged as serving the same function as a reporting newswire story irrespectively of its place of publication, i.e., a blog instead of a newspaper. At the same, hybridisation of genres is represented via assessing the presence of several functions in the same text. For example, a citizen journalism post can blend the function of news reporting with personal reporting as in a private diary entry.

3.1.1 Variation with respect to similarity to prototypes

While each text can be assessed with respect to its communicative functions manually, assessing them across the entirety of the Web corpora requires development of automatic classifiers. Table 2 presents the results of automatic classification of the respective corpora with respect to selected communicative functions using a neural model (Sharoff, 2021). As the model can predict hybrid categories, the counts are given for the function with the strongest presence (according to the automatic classifier).

Table 2: Distribution of communicative functions in three Russian corpora as compared to English

FTD		GICR-lj		GICR-news		ruWac		ukWac
A1.Argumentation	4.69%	480113	28.64%	605223	18.20%	222741	18.27%	366412
A4.Fiction	1.22%	124942	0.12%	2624	3.23%	39526	1.37%	34943
A8.News	2.08%	212712	64.67%	1366632	5.77%	70689	11.78%	236233
A11.Personal	81.07%	8305825	2.29%	48414	44.29%	542111	4.45%	89199
A12.Promotion	1.99%	204261	0.66%	13977	5.34%	65417	21.52%	547013
A14.Academic	0.57%	58636	0.29%	6172	4.77%	58410	2.62%	52516

The composition of the corpora differs in both expected and unexpected ways. Reporting personal experience (A11) is by far the most common communicative function for postings in social media (GICR-lj). This is followed by argumentative texts (A1), primarily concerning discussions on a wide range of topics, for example, politics, parenting or entertainment. The profile of the GICR news segment demonstrates a combination of factual reporting and argumentative texts expressing analysis and opinions at the ratio of two to one, as its sources consist of informative newswires (e.g., lenta.ru or rosbalt.ru). Any presence of fiction, promotion or academic texts in GICR-news comes from errors of the classifier. Analysis of composition also shows an estimate of the amount of reposting in GICR-lj, primarily from news and fiction, which obscures the demographic features of the authors for studying variation with respect to the user by means of social media.

In contrast to well-curated sources of GICR, the corpora produced by wide crawling of websites (ruWac and ukWac) contain a substantial portion of promotional texts, which reflects one of the major functions of the Web as the medium of business transactions, especially in the case of ukWac in English. As ukWac was constrained with respect to its top-level domain name (.uk), while ruWac was not, ruWac contains a substantial portion of texts from livejournal.com, blogspot.com and hiblogger.net, which provide sources of personal reporting, making it fairly different from ukWac and more similar to GICR.

3.1.2 Cases of hybridisation

Individual authors contributing to their blogs have freedom in expressing their thoughts by combining several communicative functions, so that a personal blog entry can follow the style of traditional mass media with additions of personal diaries and argumentation about the state of affairs, thus instantiating an example of citizen journalism. In turn, many online news outlets are likely to employ some of the genre techniques of citizen journalism since this genre has gained popularity on the Web. With respect to the social media sources (GICR-lj), 11% of texts are detected as hybrids with the majority of them being hybrids of A1 and A11 or vice versa, for example, political argumentation supported by personal stories.

In addition to complete hybridisation, when the entire text is aimed at expressing two or more communicative functions, many texts also have quanta of specific communicative functions. This situation is especially common in the interviews, which can include clearly separate bits, such as aimed at reporting or

evaluation (Kibrik, 2013). Quantisation of communicative functions can be easily detected via human interpretation, but since the granularity of automatic predictions is at the text level, this situation is not captured by the automatic classifiers.

3.2 Lexicogrammatical features

After completing our analysis of variation in terms of communicative functions, we can zoom in to describe variation in terms of lexicogrammatical features. A set of features suitable for automatic extraction from corpora has been proposed by Douglas Biber (1988); this has been adapted to Russian (Katinskaya and Sharoff, 2015, Sharoff, 2021). The features include:

Text-level features such as:

- average word length
- type/token ratio (TTR)

Part-of-speech (POS) features such as:

- past tense verbs
- wh-words in English and the corresponding question words in Russian
- nominalisations (nouns ending in *-tion, -ness, -ment* in English, *-tsija*, *-st', -nie* in Russian)

Syntactic features such as:

- adjectives in the attributive function
- subordinate clauses

The POS and syntactic features of this kind can be reliably extracted using modern tools such as udpipe (Straka et al., 2016). Even though the original set was initially proposed by Biber for English, it matches many properties of Russian (and other Slavic languages). For example, the major parts of speech, subordinate clauses or nominalisations provide useful indicators for the functional varieties. Even when the classes differ from English, they can be functionally equivalent. For example, wh-words do not exist under the same class in Slavic languages. However, their translations (*kto, gde, kogda…* in Russian) are likely to be a useful indicator of functional variation, though not necessarily identical to how they are used in English.

Information about the predicted communicative functions in a corpus from Table 2 can be used in assessing their registerial profile. Table 3 presents the lexicogrammatical features most closely associated with the respective communicative functions in ruWac. The absence of + and - in this table indicates the absence of statistically significant correlation, their presence indicates the degree of correlation with respective communicative functions (Sharoff, 2021).

Each communicative function has its own profile. For example, the Russian argumentative texts (A1) in comparison to other functions show a higher rate of conjunctions, nominalisations, WH-pronouns and attributive adjectives. In total this indicates the most typical features associated with expressing argument in Russian. The unexpected registerial profile of the promotional texts (A12) concerns their similarity to fairly formal texts with the higher rate of attributive adjectives, subordinate clauses, denser noun phrases, longer words and a higher Type-Token Ratio (TTR). For example, the following sentence from a typical promotional text in Russian:

Table 3: Comparison of register features in Russian using ruWac

Features	A1.argument	A4.fiction	A8.news	A11.personal	A12. promotion	A14.academic
Type-Token Ratio		+ +	+ +	- - -	+ +	
Word length		- - -	+ + +	-	+ +	
Adverbs			- -		+ +	- -
Conjunctions	+ +	-	+			
Discourse particles		-		+ +	- -	-
Nouns		+		- -	+ +	
Nominalisations	+ +	- -		- - -		+ +
Prepositions		-	+ +	-		-
Pronouns, 1p	- -	-	- - -	+ + +		- - -
Pronouns, 2p		+	- -		+ +	- - -
Pronouns, 3p		+ + +	- -	- -	- -	- -
Pronouns, WH-	+ +			-		
Verbs, past		+ + +	+ +	+ +	- -	- -
Verbs, present		+ +		+		+
Attributive adjectives	+ +		- -		+ +	
Negation	+				- -	- -
Subordinate clauses	+	-		-	+	

> *Luchshie partnerskie programmy dlya zarabotka na vashem sayte s oplatoy za kliki, procenta s prodazh, SMS partnerki…* ('The best partnership programmes for earning on your site with payment through clicks, sales royalty, SMS partnerships…')

has a long list of noun phrases and a range of different, often infrequent, words without repetitions (leading to a higher TTR).

A table of this kind can be also used to investigate differences between reasonably similar communicative functions. Examples from fiction (A4) are often used as a substitute for representing everyday language. At the same time, this table shows how Russian fiction is different from personal reporting in blogs (A11) by the distribution of discourse particles, nouns, pronouns and TTR. The differences indicate much better planning of linguistic constructions employed by fiction authors (e.g., the higher density of nouns) and the use of more varied lexicon (TTR) in comparison to more spontaneous, often repetitive, interaction in personal blog entries. At the same time, personal blog entries exhibit a much higher rate of first person pronouns and discourse particles which indicate less objective reporting and lesser formality.

The functions of A4, A8 and A11 are all related to a more general communicative function of narration about past events. From the viewpoint of features, this is expressed in their registerial profile by the rate of verbs in the past tense. At the same time, they exhibit important differences from each other. For example, texts classified as A8 (news reporting) show a higher rate of prepositions (which are often used for expressing spatiotemporal circumstances which are especially important for news reporting) and a lower rate of pronouns, which indicate a very different context of narration in the case of news reporting in comparison to either fiction or personal reporting.

The registerial profile of academic writing is clearly different from other functions, including the argumentative texts which are coming from newspaper opinion columns and argumentative blogs. Academic texts demonstrate considerably lower rates of adverbs, prepositions, verbs in the past tense and negations (unlike other argumentative texts), which indicate the lack of the narrative context of academic argumentation. Also, lower rates of negations can be linked to the aim for obtaining positive knowledge (what *is* true) and the greater formality of academic writing. Surprisingly, TTR, which often correlates with formality and

difficulty, does not emerge as a defining feature of Russian academic texts, possibly formal academic texts often use repetitive constructions.

According to the corpora of Serbian scientific texts of XVIII and XIX centuries, academic discourse in Serbian language is characterized by large number distinctive features, for example: present tense is being used more than other tenses (… *Ova škodljiva trava/zadušuje konoplju* …); the verbal forms are being expressed in personal way (IIIrd person/plural: … *Sve navedene pojave/svrstavamo u dve klase… u radu ćemo istražiti* …); de-agentivized (passive) constructions 'should + infinitive' in texts that give guidelines to the readers for something: … *seme /treba* dobro *osušiti* … *treba napisati* šest jednačina …/; nominalized constructions are being frequently used: … *promene se najčešće dešavaju pri /spavanju* …/, and so on (Stojanović 2014). These features lead to expressiveness of contemporary style of Serbian scientific writing (ibid.: 85–119).

4 Variation with respect to communication media

There is also an aspect of variation which is not often addressed in traditional accounts of sociolinguistic variation. Digital communication in addition to providing a better window into sociolinguistic research (such as the GICR corpus discussed above) also resulted in changes in the use of language. The relevant technological innovations on the Web include all forms of user-generated content, such as Wikipedia or review publication websites, as well as social media. They all offer more opportunities for individual contributions with relatively little threshold for participation in comparison to traditional communication and distribution channels. At the same time, the new affordances are first developed and then accepted by a small proportion of the population, primarily within the younger, more affluent and better educated groups. In the case of Russian this has led (among other things) to the much greater use of both borrowing and calques from English, for example, *sayt* ('website'), *klik* ('click'), *partnerki* ('partnerships'). A little later some of the innovations gain enough popularity to lead to their standardization, for example, *sayt* belongs to the top 500 words in the general corpus of ruWac, as this word can be used in the most formal discourse contexts:

> *Materialy conferencii budut razmeshcheny na ofitsial'nom sayte konferentsii v vide elektronogo sbornika i razmeshcheny v RINTS* ('The conference materials will be published on the official website of the conference as an electronic publication and will be included in the Russian Science Citation Index.')

Because of the difference in the alphabet between English and Russian, expression of some borrowings can be made easier by combining the Cyrillic and Latin characters when the more difficult combinations of Latin characters are rendered in English, thus creating a mix of spellings, for example, *call-центр* (the call center), *web-сайт* (the website), *Android'ом* (with Android).

The new ways of using language also lead to the emergence of new cultural phenomena, which can start their relatively unhindered development in the absence of gatekeepers (at least at the initial stage of their development). For example, many orthographic features have been borrowed into Russian social media from English, such as an extensive use of smileys, punctuation combining exclamation and question marks (!?!?), repeated characters (*Daaaa!!*, 'yeeees!!'), texts written in the upper case or with strikethroughs (Piperski and Somin, 2013).

On the Russian Internet, there is also a case of creative spelling in the subculture known as *padonki* (lit. 'scoundrels' with the first vowel misspelled, *o* → *a*), also known as *Olbanskiy yazyk* (Olbanian language, this time misspelling *A* → *O*). This is related, though not directly borrowed from the cacography tradition

of making deliberate spelling or writing mistakes, usually with the purpose of mockery. What is specific to the Russian subculture case is a very extensive development of rules for violating the expected norms of Russian spelling, such spelling *a* as *o* and vice versa, turning *v* to *ff*, etc., often with the aim of expressing phonetically a similar string in a way which is maximally different from the accepted orthographic norms (Krongauz, 2013).

References

Andersen, G. (2010). How to use corpus linguistics in sociolinguistics. In O'Keeffe, A. and Mccarthy, M., editors, *The Routledge Handbook of Corpus Linguistics*, pages 547–62. Routledge, Abingdon.

Bakhtin, M. M. (1986). *Speech genres and other late essays*. University of Texas Press. Translated by Vern W. McGee, edited by Caryl Emerson and Michael Holquist.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Belikov, Vladimir Kopylov, N., Selegey, V., and Sharoff, S. (2014). Variational corpus statistics using author profiles. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo.

Belikov, V. (2010). Methodological novelties in sociolinguistic lexicography in the 21 century. In Mustajoki, A., Protassova, E., and Vakhtin, N., editors, *Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian*, pages 32–49. Slavica Helsingiensia. (in Russian).

Belikov, V. and Krysin, L. (2001). *Sociolinguistics*. RSUH. (in Russian).

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Bošnjaković, Ž., editor (2009). *Govor Novog Sada, sv. 1: fonetske osobine*. Filozofski fakultet, Novi Sad.

Bugarski, R. (2006). *Žargon*. XX vek.

Bugarski, R. (2009). Teorijske osnove urbane dijalektologije (theoretical foundations of urban dialectology). In Bošnjaković, Ž., editor, *Govor Novog Sada, sv. 1: fonetske osobine*, pages 13–30. Filozofski fakultet, Novi Sad.

Cherkassky, V. and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, 17(1):113–126.

Crystal, D. (2004). *The Cambridge Encyclopedia of the English Language*. Cambridge University Press, Cambridge.

de Haan, F. (2002). Strong modality and negation in Russian. In Reppen, R., Fitzmaurice, S. M., and Biber, D., editors, *Using corpora to explore linguistic variation*, pages 91–110. John Benjamins Publishing, Amsterdam.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Ferraresi, A., Zanchetta, E., Bernardini, S., and Baroni, M. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *The 4th Web as Corpus Workshop: Can we beat Google? (at LREC 2008)*, Marrakech.

Gregory, M. (1988). Generic situation and register: a functional view of communication. In *Linguistics in a systemic perspective*, pages 301–330. John Benjamins, Amsterdam.

Halliday, M. (1978). *Language as Social Semiotic: The social interpretation of language and meaning*. Blackwells, Oxford.

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. Edward Arnold, London.

Hanks, P. (2012). Corpus evidence and electronic lexicography. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, pages 57–83. Oxford University Press.

Holmes, J. (2013). *An introduction to sociolinguistics*. Routledge, Abingdon.

Hymes, D. (1971). *On Communicative Competence*. University of Pennsylvania Press, Philadelphia.

Hymes, D. (1974). *Foundations of Sociolinguistics. An Ethnographic Approach*. University of Pennsylvania Press, Philadelphia.

Katinskaya, A. and Sharoff, S. (2015). Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. In *Proc BSNLP*, Sofia.

Kibrik, A. (2013). Discourse structure and communicative intentions: a study of Russian TV interviews. In Volodina, M., editor, *Mediensprache und Medienkommunikation im interdisziplinären und interkulturellen Vergleich*, pages 223–245. Institut für Deutsche Sprache.

Krongauz, M. (2013). *Samouchitel Olbanskogo (Teach yourself Olbanian)*. Corpus.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

Lauersdorf, M. R. (2009). Slavic sociolinguistics in north america: Lineage and leading edge. *Journal of Slavic linguistics*, pages 3–59.

Lee, D. and Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for specific purposes*, 25(1):56–75.

Ljubešić, N. and Klubička, F. (2014). {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proc 9th Web as Corpus Workshop at EACL'14*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Mesthrie, R., Swann, J., A., D., and Leap, W. L. (2009). *Introducing Sociolinguistics*. Edinburgh University Press, Edinburgh.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Piperski, A. and Somin, A. (2013). Liturativy v russkom internete: semantika, sintaksis i tehnicheskie osobennosti bytovaniya (strikethrough on the russian web: semantics, syntax and technical issues). In *Proc Dialogue, Russian International Conference on Computational Linguistics*.

Rajić, L. (2009). Gradski govori (urban vernaculars). In Bošnjaković, Ž., editor, *Govor Novog Sada, sv. 1: fonetske osobine*, pages 31–45. Filozofski fakultet, Novi Sad.

Reppen, R., Fitzmaurice, S. M., and Biber, D., editors (2002). *Using corpora to explore linguistic variation*, volume 9. John Benjamins Publishing.

Sharoff, S. (2005). Methods and tools for development of the Russian Reference Corpus. In Archer, D., Wilson, A., and Rayson, P., editors, *Corpus Linguistics Around the World*, pages 167–180. Rodopi, Amsterdam.

Sharoff, S. (2006). Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Sharoff, S. (2017). Corpus and systemic functional linguistics. In Bartlett, T. and O'Grady, G., editors, *The Routledge Handbook of Systemic Functional Linguistics*, pages 533–546. Routledge.

Sharoff, S. (2018). Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95.

Sharoff, S. (2021). Genre annotation for the web: text-external and text-internal perspectives. *Register studies*, 3:1–32.

Sharoff, S., Goldhahn, D., and Quasthoff, U. (2017). *Frequency Dictionary: Russian*, volume 9 of *Frequency Dictionaries*, chapter Corpus, pages 9–14. Leipziger Universitätsverlag. Uwe Quasthoff, Sabine Fiedler, Erla Hallsteindóttir (editors).

Stojanović, A. (2014). *Stereotipnost naucynogo teksta*. Međunarodna asocijacija ”Stil”.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proc LREC 2016*, Portorož, Slovenia.

Szmrecsanyi, B. (2009). Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change*, 21(3):319–353.

Timberlake, A. (1993). Russian. In Comrie, B. and Corbett, G. G., editors, *The Slavonic Languages*. Routledge.

Šipka, D. (2021). Normative usage labels: a case study in croatia. In *Lexicography and Lexicology in the Light of Current Issues*, pages 175–188. Belgrade: Serbian language institute of SASA.

Štasni, G. and Ajdžanović, M. (2011). Stanovnici novog sada i petrovaradina: derivacioni modeli etnika i upotrebna norma (residents of novi sad and petrovaradin: Derivational models of gentilics and the norm of usage). In Vasić, V. and Štrbac, G., editors, *Govor Novog Sada, sv. 2: morfosintaksičke, leksičke i pragmatičke osobine*, pages 78–120. Filozofski fakultet, Novi Sad.

Štrbac, G. and Vujović, D. (2011). Između dijalekta, mesnog govora i standardnog jezika: leksički dubleti i sinonimi (between dialect, local speech and standard language: Lexical doublets and synonims). In Vasić, V. and Štrbac, G., editors, *Govor Novog Sada, sv. 2: morfosintaksičke, leksičke i pragmatičke osobine*, pages 192–209. Filozofski fakultet, Novi Sad.

Vujović, D. and Alanović, M. (2011). Tipična morfosintaksička obeležja u govoru novog sada (typical morpho-syntactic features of novi sad vernacular). In Vasić, V. and Štrbac, G., editors, *Govor Novog Sada, sv. 2: morfosintaksičke, leksičke i pragmatičke osobine*, pages 39–55. Filozofski fakultet, Novi Sad.