JOURNAL *of* SLAVIC LINGUISTICS

VOLUME 29 · ISSUE NO. FASL SPECIAL ISSUE

Inferring stress placement variability from a poetic corpus

ALEXANDER PIPERSKI & ANTON KUKHTO
HSE University *MIT*

ABSTRACT

In this paper, we analyse stress placement variation in the Poetic subcorpus of the Russian National Corpus. We measure the extent of variation observed in the data and show that the rate of stress variation in poetry written in Modern Russian has been declining since the middle of the 19th century. We suggest that variation observed in poetry reflects variation in the language overall, although this link requires much further investigation. Nevertheless, we maintain that poetry is a useful resource in the investigation of stress placement, especially when other resources are not available.

KEYWORDS corpus linguistics · poetry · Russian · stress · variation

1 INTRODUCTION: STRESS IN RUSSIAN

Stress in Modern Russian belongs to the class of unbounded lexical stress systems (Goedemans & van der Hulst, 2013), where the accentual properties of individual morphemes (or "diacritic weight") determine the placement of surface stress, cf. *moróz* 'frost' – *moróz-a* (gen. sg.) with stress fixed on the stem vs. *durák* 'fool' – *durak-á* (gen. sg.) with stress shifting between the stem and the desinence.[1] Stress is also contrastive, e.g. *múka* 'torment' – *muká* 'flour'. On top of that, there exist phonological factors that influence the position of default stress, see Mołczanow et al. (2019) and references therein. This system has received a lot of attention in the phonological literature, although, needless to say, many puzzles remain. For extensive analyses, see Jakobson (1963), Halle (1973), Zalizniak (1985), Melvold (1989), Alderete (1999); for secondary stress, see Gouskova (2010), Gouskova & Roon (2013). What is less fully understood is variation in stress placement, namely occurrences of a given morphological form that show stress on different syllables rather than uniform stress placement, which Russian exhibits quite often and which is what this study addresses.

Variability in the position of stress in Russian is found across the entire lexicon and is not restricted to a particular morphological class, cf. *pródal ~ prodál* 'sold (masc.)', *tvórog ~ tvoróg* 'cottage cheese', *glubokó ~ glubóko* 'deeply', *pólny ~ polný* 'full (pl.)'. Many speakers of Modern Russian are well aware of certain instances of accentual variation, which are sometimes subject to debate and considerable stigmatisation, e.g. in the case of *zvónit ~ zvonít* 'calls (3 sg.)', where the innovative variant with initial stress is often regarded as unacceptable. This system, or at least some of its parts, is also in flux. Even if we look at the recommendations of recent pronunciation dictionaries separated only by 30 years, Avanesov (1983) and Kasatkin (2012), we can observe certain changes. Below are past tense forms of the verb *prodat'* 'to sell' in these two pronouncing dictionaries, Avanesov (1983) in (1) and Kasatkin (2012) in (2); translation of the labels is ours.

(1) *pródal* and acceptable *prodál* 'sold (masc.)'
prodalá incorrect *pródala*, *prodála* 'sold (fem.)'
pródalo **and acceptable** *prodálo* **not recommended** *prodaló* 'sold (neut.)'
pródali and acceptable *prodáli* 'sold (pl.)'

[1] All examples from Russian are romanised. Throughout the paper, we mark the position of primary stress with the acute accent; stress is not normally reflected in the Russian orthography.

(2) *pródal* and acceptable *prodál* 'sold (masc.)'
prodalá incorrect *pródala*, *prodála* 'sold (fem.)'
pródalo **and** *prodálo* **acceptable new** *prodaló* 'sold (neut.)'
pródali, acceptable new *prodáli* 'sold (pl.)'

The origins of stress variation in Modern Russian can be traced back to Old Russian stress and the subsequent phonological changes. Old Russian had a system of stress assignment based on lexical accent. As in Modern Russian, morphological conditioning, i.e. accentual properties of individual morphemes, played an important role in determining the position of stress. The most complete and thorough description of the Old Russian system is provided by Zalizniak (1985, 2014).

Without going into the details of Old Russian, we give an example of stress assignment in the past participle forms, which gave rise to the Modern Russian past tense, of the verbs *dati* 'to give', *prodati* 'to sell', and *vydati* 'to give away' in (3). The stem itself does not have an underlying diacritic stress, which is shown in (3-a) by the fact that an underlyingly stressed feminine desinence *-á* attracts stress (in accordance with the Basic Accentuation Principle, the leftmost accented morpheme surfaces as stressed, see Kiparsky 2010 and references therein). In the rest of the forms, as well in the corresponding forms in (3-b), we see the default initial stress, which signals the absence of underlying stresses in the morphemes of these forms. In (3-c), the prefix *vy-* is always stressed by virtue of it having an underlying accent and being the leftmost morpheme.

(3) a. *dá-l-ŭ* (m.) – *da-l-á* (f.) – *dá-l-o* (n.) – *dá-l-i* (pl.) 'gave'
b. *pró-da-l-ŭ* (m.) – *pro-da-l-á* (f.) – *pró-da-l-o* (n.) – *pró-da-l-i* (pl.) 'sold'
c. *vý-da-l-ŭ* (m.) – *vý-da-l-a* (f.) – *vý-da-l-o* (n.) – *vý-da-l-i* (pl.) 'gave away'

This brief example demonstrates that even in Old Russian some paradigms had significant oscillation in stress placement that was due to the mechanism of surface stress computation based on the accentual properties of individual morphemes. Diacritic weight, however, is unstable and can be overruled by paradigmatic factors and phonological change. Thus, in Western and Southwestern dialects of Russian, stress definalization (*prodalá* → *prodála*) and/or analogical stress shift to the beginning of the word (*pródala*) can be observed (Zalizniak, 1985). In other dialects, analogical stress shift to the end of the word is found (*prodál*, *prodaló*). These and other changes have obscured the regular Old Russian pattern and led to the emergence of stress variation.

2 STRESS VARIATION: QUESTIONS AND SOURCES OF EVIDENCE

Stress placement variation is a widespread phenomenon in Modern Russian. There are multiple factors that influence this variation. We discuss some of those factors below, but the list is most probably not exhaustive.

First, different social variables, such as education and age, correlate with the position of stress, which is what sometimes makes debates about stress placement in the public domain quite heated. For instance, the more conservative dative plural form of the word *sreda* 'Wednesday' with final stress, *(po) sredám* 'on Wednesdays' is often regarded as a social marker of more "educated" speech as opposed to the more innovative variant with initial stress. Examples like this can easily be multiplied, although there is at present a lack of thorough sociolinguistic investigations (one promising source is Dobrushina & Staferova 2018). Second, speaker's occupation has been shown to influence stress placement in professional lexicon (Zalizniak, 2012). Across various professions, stress is known to shift to the final syllable in relevant lexical items, e.g. *kompás* 'compass' (sailors) vs. the standard *kómpas*, *iskrá* 'spark' (electricians) vs. the standard *ískra*, *kraný* 'water taps' (plumbers) vs. the standard *krány*. Third, style and situational context can influence the choice of stress location, sometimes even consciously for some speakers. For instance, the form *proréktory* 'vice-chancellors' is more likely to be used in a more formal context, whereas *prorektorá* with final stress is more likely to occur informally in a professional setting. Then, for certain lexical items, stress variation appears to accompany semantic divergence, e.g. *kúrit sigaretu* 'smokes (3 sg.) a cigarette' vs. *kurít fimiam* 'burns (3 sg.) incense'. A divergence of this kind is observed by Say (2020) in constructions with the preposition *po*, e.g. *po króvi* vs. *po kroví* 'through the blood', where forms with final stress

express the idea of "fixed location". Finally, there are cases of geographical variation, i.e. different stress placement in different dialects and regional varieties of Russian, e.g. *spála* 'slept (fem.)' in Ukrainian Russian and Southern Russian vs. the standard *spalá*.

The cases exemplified above are predominantly instances of inter-speaker variation, where forms with different stress placement are used by different speakers depending on their age, dialect, educational attainment, etc. Although comprehensive studies of such cases remain a task for the future, some information about them can be found in pronouncing dictionaries (Avanesov, 1983; Kasatkin, 2012). There are also cases of intra-speaker variation, namely situations where the same speaker uses variable stress in a certain form in their speech, e.g. *odnovrémenno ~ odnovreménno* 'simultaneously'. At first glance, variants like these might appear to be in free variation. Upon closer investigation, however, it turns out that this kind of variation is governed by multiple subtle factors. For instance, see Kukhto & Piperski (2020) for rhythmic constraints influencing the position of stress in past-tense verb forms, e.g. *pródal dáču* 'sold (sg. m.) a country house' vs. *prodál pal'tó* 'sold (sg. m.) an overcoat' with alternating stress. Again, most of these factors are not currently understood sufficiently.

There are multiple questions pertaining to stress variation in Modern Russian. Which forms vary, i.e. which accentual classes exhibit variation? What are the factors that influence this kind of variation? How does it reflect and impact accentual change? How are the two connected? Is it persistent over time?

In this study, we are focusing on the last issue, namely the scope of stress variation over time. The main question we are targeting is whether the proportion of words and word forms with variable stress stays the same or changes over time; and if there is change, what the direction is. This initial investigation could open up the path to asking more detailed questions about the nature of stress variation in Russian.

Since the question we are aiming to address is a diachronic one, we are inevitably going to be restricted in resources. A natural approach to the study of variation is through experiment, which allows one to focus narrowly on specific phenomena, yet this approach is obviously not going to be available to us. We could turn to spoken corpora, which could provide naturally produced data. Unfortunately, the corpora that are currently available, such as the Spoken subcorpus of the Russian National Corpus (RNC), see Grishina (2006), contain only small amounts of texts by the same speaker and do not allow to carry out an investigation of stress variation. However, there does in fact exist a sufficiently large corpus of Russian that contains information about stress placement and could fit our purposes, namely the Poetic subcorpus of the RNC[2] (see also Korchagin 2008 for the properties of this subcorpus and the possibility of using it to study accentology). In the next section, we argue that this corpus is indeed suitable for our purposes and, more generally, that poetry is a useful source of information about stress.

3 POETRY AS A RESOURCE FOR THE STUDY OF STRESS VARIATION

Using poetry as a source for studying phonology is, of course, no innovation of this paper (see, for instance, Jakobson 1960). Poetic rhyme has been used as evidence of judgments of perceptual similarity or identity (Steriade 2009 and references therein); there is also a long tradition of using meter for the study of syllable weight and other phonological phenomena (Ryan, 2014), to give but a couple of examples. However, we need to justify the use of Russian poetry as represented in the Poetic subcorpus of the RNC for the study of stress variation. For one thing, it provides a wealth of naturally produced (albeit more consciously controlled than in usual speech production) data with many texts by the same speaker. The size of the corpus is 11m tokens as of this writing, which makes it suitable for detecting variation and quantifying the results. What needs to be made clear is the connection between poetic beats and linguistic stresses.

Russian classical poetry employs the alternation of stressed and unstressed syllables, representing a syllabotonic (also known as accentual-syllabic or syllabic-accentual) system. This type of versification has prevailed in Russian poetry since the mid-18th century (Gasparov, 2000). As Kolmogorov & Prokhorov (1968) show, a fundamental property of this system is that the stress of

[2] The corpus can be freely accessed at: http://ruscorpora.ru/search-poetic.html.

a polysyllabic word should be aligned with the strong position in the verse. Below is an example of iambic tetrameter from the *Ode … on the Taking of Khotyn* (1739) by Mikhail Lomonosov, as represented in the RNC, with strong positions indicated by acute accents; translation by Harold Segel.

(4)

Vračébnoj dáli mné vodý:	'They gave me healing waters there:
Ispéj i vsé zabúd′ trudý;	"Do drink and all your work forget;
Umój rosój Kastál′skoj óči,	Your eyes bathe with Castalian dew;
Črez stép′ i góry vzór prostrí	Your gaze extend cross steppe and hills,
I dúx svoj k tém stranám vperí,	And guide your spirit to those spots
Gde vsxódit dén′ po témnoj nóči.	Where day ascends upon the darkness." '

Every single strong position in this fragment coincides with a lexical stress. This is an ideal case; in practice, dibrachs and spondees (feet with no stresses and two stresses, respectively) can occur and polysyllabic words can contain more strong positions than stresses. Nevertheless, the correspondence between beats and stresses is generally observed.

A typical objection to the claim that poetry can be used to infer stress placement involves the notion of poetic license. It is sometimes said that poets bend the rules of conventional grammar and, among other things, put stresses wherever they please in order to conform to the meter even when those stresses do not correspond to the usual stress placement of the language. If that were so, it would undermine our efforts to infer any facts about stress variation in Russian by recourse to poetic corpora. We would like to argue against this claim and maintain that the effect of poetic license is negligible given the amount of data we have at our disposal. Consider the following example from a poem by Marina Tsvetaeva (*Ljudi na dušu moju l′stjatsja*, 1916); translation is ours.

(5)

Zváli – rávno, nazyváli – rázno,	'(They) called (me) equally, called variously,
Vse nazyváli, niktó ne názval.	everyone called (ipf.), no one called (pfv.).'

The form *názval* 'called (masc.)' sounds off to Modern Russian speakers, and it may seem that it was invented for the sake of rhythm and rhyme only. However, the study of old dictionaries and accentuated texts shows that this is just a manifestation of the original mobile stress pattern, for instance, cf. *nánjal* 'hired (masc.)' – *nanjalá* 'hired (fem.)'.

Next, consider two well-known examples, from a poem by Alexander Pushkin (*The Poet and the Crowd*, 1828) and a fable by Ivan Krylov (*The Grasshopper and the Ant*, 1808); translations by Philip Nikolaev and the authors respectively.

(6)

Ty pól′zy, pól′zy v nëm ne zriš′.	'Yet in his form you see no good.
No mrámor sej ved′ bog!.. tak čto že?	That marble is a god! So what?
Pečnój goršók tebé dorózе:	You much prefer your cooking pot,
Ty píšču v nëm sebé varíš′.	Because therein you cook your food!'

(7)

Poprygún′ja Strekozá	'The bouncy grasshopper
Léto krásnoe propéla;	sang all beautiful summer long;
Ogljanút′sja ne uspéla,	before she knew it (*lit.* she had no time to look around),
Kak zimá katít v glazá.	winter came (*lit.* as winter rolls into her eyes).'

The stresses in *varíš′* 'cook (2 sg.)' and *katít* 'rolls (3 sg.)' again sound unusual to a Modern Russian speaker. However, these are exactly the stresses for these words indicated in the *Dictionary of the Russian Academy* (1789–1794). Likewise, the vast majority of examples in our sample where the stress deviates from modern usage are either found in accentuation dictionaries (Modern Russian: Avanesov 1983; Kasatkin 2012; Old Russian: Zalizniak 2014)[3] or conform to native speaker intuitions (both authors are native speakers of Russian), or both.

[3]We have not checked every single form in our sample against these dictionaries in order to quantify the scope of poetic license more precisely. That remains a desideratum. However, due to the data annotation procedure, which we describe in more detail in Section 4, every example was filtered manually. Forms that diverged from our intuitions and were not found in the dictionaries were vanishingly rare and, moreover, seemed to be characteristic of particular more experimentally inclined poets like, for instance, Aleksey Vernitsky (b. 1970).

The view that poetry does not allow arbitrary stress placement is shared by many literary scholars specializing in metrics. To give but one example, consider the following quote from Georgy Shengeli (1940, 6), one of the most prominent Russian verse theorists of the 20th century: "Stress cannot be arbitrarily shifted away from the syllable where it is naturally placed in a given word in its grammatical form; one cannot say *čelóvek* 'person' instead of *čelovék*. In the old days, the stresses ... were not necessarily the same as today. Pushkin pronounces *muzýka* 'music', and we say *múzyka*. When one encounters such stresses, one should not assume that this is an artificial stress shift" (translation from Russian is ours).

We conclude from this brief discussion that the effect of poetic license is negligible for our purposes, namely using poetry to study stress variation in Russian. There remain concerns and the need for caution in drawing conclusions due to the fact that poetry is obviously a specific type of text that has multiple properties of its own. For example, one could argue that, even if poetry does not allow forms with random stress locations, it is still not representative of day-to-day Russian and may have different distributions of variable stresses. This is a valid concern which needs to be taken into account and which we return to in Section 6. With these provisos in place, we can now move on to the discussion of data collection and annotation methods that we used.

4 DATA COLLECTION AND ANNOTATION

We have now settled on syllabotonic poetry from the Poetic subcorpus of the RNC as our main source of data for the study of stress variation in Russian. To illustrate the technique of data collection that we used, let us consider the following two examples; translations by Charles Johnston and the authors respectively.

(8)
I golosóv nestrójnyj gul '[A] hum of voices grows, falls still
Terjáetsja, i karavány lost in the distance, and the tinkling
***Idút** zvenjá izdaleká.* caravan bells sound far away...'

(9)
***Ídut** vse polkí mogúči,* 'All regiments march, mighty,
Šúmny kak potók, loud as a stream,
Strášno-médlenny, kak túči, frightfully slow, like thunderclouds,
Prjámo na vostók. directly eastwards.'

These fragments come from two poems by Mikhail Lermontov, which were both written in the year 1841. In *The Demon* (8), there is final stress on the verbal form *idút* 'go (3 pl.)', whereas in *The Dispute* (9) there is initial stress on the very same form, *ídut* 'go (3 pl.)' (which is nowadays antiquated). Examples like this, namely a single morphological form used by the same speaker (poet) with different stress placement, but not necessarily within the same year, is what we were after. By restricting the domain of our search for variable forms to corpora of individual poets, we are aiming to detect instances of intra-speaker stress variation and control for various factors of inter-speaker variation, such as regional or educational factors, and so on. While this is going to be the focus of our study, we will briefly mention another viable option, namely detecting variation within a certain time frame, in Section 6.

Our data collection procedure consisted of the following steps.

1. We manually selected a sample containing 20 poets and extracted all texts by those poets marked as purely syllabotonic in the Poetic subcorpus of the RNC; the list of poets with their dates of birth is given in Table 1 below.[4] As we have seen, syllabotonic verse fixes the number of stresses and syllables within a line or stanza. There exist other types of verse in Russian (accentual, syllabic, vers libre), which do not allow to establish the position of stress with certainty and had to be excluded.

2. For all word forms in the collected sample, we automatically detected all instances of stress variation using the annotation provided in the corpus. For instance, having collected all syllabotonic texts by Mikhail Lermontov, we detected for each form, such as *idut* 'go (3 pl.)', whether it has

[4]This sample is not balanced, reflecting certain properties of the Poetic subcorpus of the RNC, e.g. the predominance of male authors and the under-representation of present-day poetry due to copyright restrictions.

Poet's name	Year of birth	Corpus size (tokens)	Variable forms (types)
Lomonosov	1711	53,210	57
V. Maykov	1728	50,938	45
Küchelbecker	1797	13,420	10
Pushkin	1799	182,014	187
Yazykov	1803	59,008	41
Lermontov	1814	125,883	121
A. Maykov	1821	107,696	135
Mey	1821	38,544	57
Grigoryev	1822	39,654	38
V. Ivanov	1866	103,357	117
Kuzmin	1872	57,742	63
Gumilev	1886	57,389	61
G. Ivanov	1894	40,130	27
Lugovskoy	1901	42,072	17
Poplavsky	1903	34,797	42
Kornilov	1907	23,185	13
Tvardovsky	1910	101,448	43
Simonov	1915	51,332	30
Samoylov	1920	58,178	15
Gandlevsky	1952	13,890	5

Table 1: Poets, corpus sizes, and numbers of forms with variable stress placement

accentual variants or not. If it does, which is the case for *idut*, we counted the number of occurrences for each accentual variant; in this case, *idút* × 18, *ídut* × 2. Note that there is no word meaning disambiguation in the RNC, and part-of-speech-tagging ambiguity is left unresolved, so homophonous forms can be classified as accentual variants at this stage and need to be filtered out later. It is also worth mentioning that we study word forms rather than paradigms; thus, if a sample contains forms *xolmám* 'hill (dat. pl.)' and *xólmax* 'hill (loc. pl.)', this will go unnoticed.

3. Finally, we manually filtered out erroneously detected instances of variable stress. Some of the most common types of false positives were homonyms, e.g. *gotóv* 'ready' vs. *gótov* 'Goth (gen. pl.)'; markup errors, e.g. syllabic or accentual poems classified as syllabotonic; and typos resulting in markup inconsistencies. Overall, the precision of the automatic detection of stress variation was quite low (across 20 poets, only 6% to 30% of types automatically identified as variable turned out to exhibit actual variation), so significant filtering was unavoidable. Since the corpus is not disambiguated, as we have mentioned, cases like *bérega* 'shore (gen. sg.)' ~ *beregá* 'shore (nom./acc. pl.)' were also extracted as potential examples of variable stress and had to be excluded by hand.

To illustrate the results of the procedure outlined above, we show all words with variable stress extracted from the works of Sergey Gandlevsky (b. 1952), represented by a relatively small corpus in the RNC (103 texts; 13,890 tokens; 7,172 word forms types total); the list is given in (10).

(10) *Forms with variable stress in the works of Sergey Gandlevsky*
- a. *kládbišče* 'cemetery' × 1 ~ *kladbíšče* × 3
- b. *póutru* 'in the morning' × 1 ~ *poutrú* × 1
- c. *p'jány* 'drunk (short form, pl.)' × 1 ~ *p'janý* × 1
- d. *srédam* 'Wednesday (dat. pl.)' × 1 ~ *sredám* × 1
- e. *tótčas* 'at once' × 1 ~ *totčás* × 1

Table 1 provides a summary of overall corpus sizes for individual poets in our sample (in tokens) and numbers of forms with variable stress placement (in types) in their corpora.

%	N	V(N)	K(N)	K(N)/N	K(N)/V(N)
10	5,738	3,120	2	0.035%	0.064%
20	11,476	5,466	5	0.046%	0.091%
30	17,214	7,482	15	0.087%	0.20%
40	22,952	9,160	16	0.07%	0.17%
50	28,690	10,742	23	0.08%	0.21%
60	34,428	12,302	41	0.12%	0.33%
70	40,166	13,613	44	0.11%	0.32%
80	45,904	14,961	48	0.10%	0.32%
90	51,642	16,105	57	0.11%	0.35%
100	57,390	17,245	61	0.11%	0.35%

Table 2: Proportion of word form types with variable stress in the corpus of Gumilev; % is the size of the sample relative to the size of the corpus, N is the number of tokens, V(N) is the number of types, K(N) is the number of types with variable stress placement

Poet	Mean K	Poet	Mean K	Poet	Mean K
Lomonosov	7.066	Mey	8.915	Poplavsky	8.804
V. Maykov	5.148	Grigoryev	7.093	Kornilov	3.455
Küchelbecker	6.895	V. Ivanov	6.477	Tvardovsky	3.603
Pushkin	6.732	Kuzmin	5.365	Simonov	3.237
Yazykov	6.435	Gumilev	4.595	Samoylov	2.185
Lermontov	6.086	G. Ivanov	3.477	Gandlevsky	2.963
A. Maykov	5.764	Lugovskoy	2.686		

Table 3: Mean number of variable forms (K) per 10,000 forms

5 DATA ANALYSIS

As Table 1 shows, the number of forms with stress placement variation depends on corpus size; compare 10 variable types in 13,420 tokens in the works of Küchelbecker vs. 187 types in 182,014 tokens in the works of Pushkin. This presents a problem for the comparability of individual corpora in our sample: larger corpora contain more variation just because of their size. This becomes even more obvious when one examines Table 2, which shows the number of forms with variation (K) across random differently-sized samples from the corpus of Nikolay Gumilev (1886–1921).

The reason why the proportion of form types with variable stress increases in larger slices of a corpus is fairly intuitive. Imagine a situation where randomly selected 30% of the corpus include several instances of *tótčas* 'at once' but not a single instance of *totčás*; the latter is only found in the remaining 70% of the corpus. As a result, in the 30% of the corpus the adverb *totčas* is not counted as exhibiting variable stress placement. If instead of different portions of the same corpus we are comparing different corpora, this problem makes the results of such a comparison uninterpretable.

There are at least two principal ways of solving this problem. One is to find a mathematical relationship between the number of words with variation and corpus size (Piperski & Kukhto, 2016).The other is to reduce all corpora to the same corpus size for compatibility without loss of data. We follow the second path and implement the following procedure: 1) take a random sample of 10,000 tokens; 2) count the number of word form types with variable stress in this sample; 3) repeat 1000 times; 4) calculate the mean number of word form types with variable stress across these 1000 samples. This gives us a measure of the rate of variability in the corpus of an individual poet. To go back to our example with *tótčas* and *totčás* 'at once', some of these 10,000-token samples will lack *totčas* completely, some will only include *tótčas* and some will only include *totčás*, but some will include both. Corpora that have more stress variation will show higher averages. The results are shown in Table 3.

With these means, we now have an estimate of the extent of stress placement variability present in the corpus of an individual poet. We can now turn to the main question we are asking in this

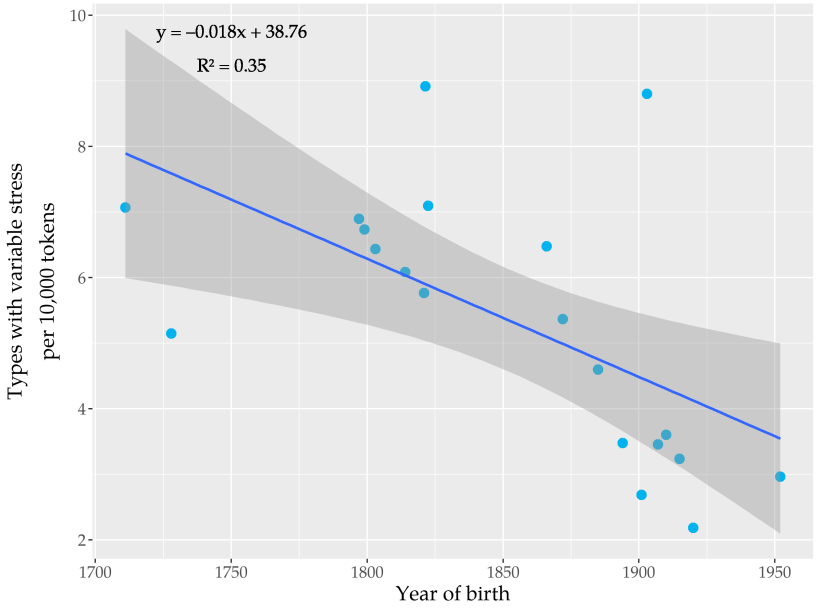

Figure 1: Types with variable stress per 10,000 tokens against the year of birth; linear approximation

paper, namely whether the rate of stress variation in the Poetic subcorpus of the RNC changes over time and what the direction of this change is if it does. The means shown in Table 3 can be plotted against the year of birth of each poet in the sample (see Table 1). The results are shown in Figure 1.

Figure 1 shows a clear trend: the more recently a poet is born, the less variation their corpus exhibits. However, approximation by a quadratic function quadratic function provides a better fit to the data ($R^2 = 0.49$ as opposed to $R^2 = 0.35$ for a linear function). The graph of the quadratic function is shown in Figure 2.

One can notice that in both graphs there is a conspicuous outlier, namely a poet born after 1900 whose corpus in fact shows a lot of variation with the second highest mean K. This outlier is Boris Poplavsky (1903–1935). He was born in Moscow; both his parents were of Lithuanian origin. He was mostly active in emigration, in Constantinople and Paris, where he died. A feature of his poetic technique is play on stress as a stylistic device. For instance, consider the following fragment from 1923–1930 in (11); translation is ours.

(11)

I v lilovoj áure aúre...	'And in a lilac aura...'
Vyšla v nebo Láura Laúra.	'Laura went out into the sky.'

The two words *aura* 'aura' and *Laura* 'Laura' each appear two times with different stresses within the same line (unlike in English, both are trisyllabic in Russian). This is not a unique example of this type from Poplavsky, nor is it indeed unique to Poplavsky. This shows that a better understanding of individual poetic style might improve our ability to infer stress placement patterns in the language in general on the basis of data from poetry. We leave this discussion for the future. Nevertheless, the overall trend is present no matter whether we consider Poplavsky or not, though the quadratic approximation without Poplavsky provides a noticeably better fit ($R^2 = 0.73$), as shown in Figure 3.

6 DISCUSSION

The analysis in Section 5 confirms that there is change in the rate of stress placement variation in the Poetic subcorpus of the RNC: the amount of variation decreases from the end of the 18th century up till the first half of the 20th century; given that these are years of birth of the poets in our sample, we can say that stress variation has been declining since mid-19th century. Before we draw our final conclusions, recall a question from before, namely why collect data by poet rather than by

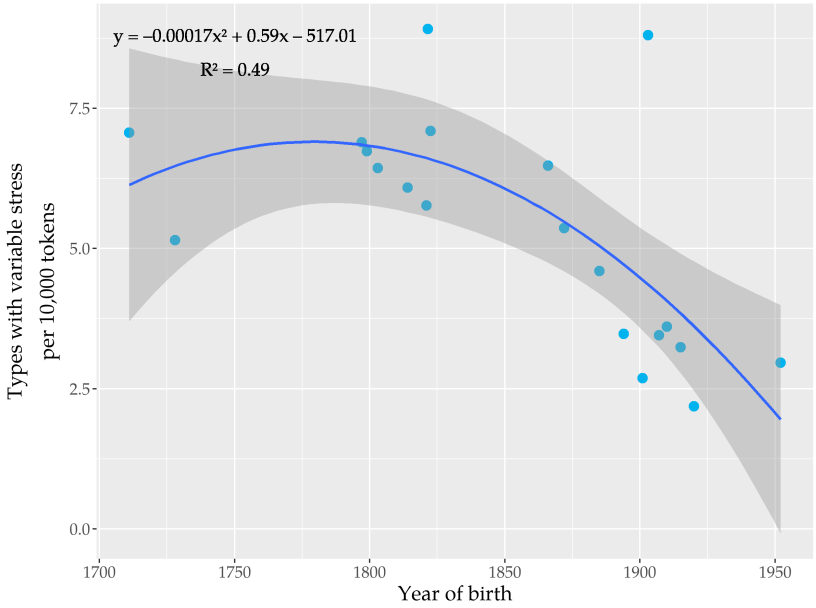

Figure 2: Types with variable stress per 10,000 tokens against the year of birth; quadratic approximation

Stress 1	Gumilev	['86; '96]	Stress 2	Gumilev	['86; '96]	Same D
bély	67%	71%	*belý*	33%	29%	1
blédny	33%	33%	*bledný*	67%	67%	1
vídny	67%	57%	*vidný*	33%	43%	1
vysóko	86%	74%	*vysokó*	14%	26%	1
vysótax	67%	12%	*vysotáx*	33%	88%	0
...	...	...	...	...	...	...

Table 4: Agreement between Gumilev and [1886; 1896]; D stands for dominant form

decade or year. This question can be reformulated for the purposes of our current discussion as follows: are individual poets representative of their generation?

We leave a more complete investigation of this issue for the future, but nonetheless we would like to suggest that intra-speaker variation reflects the language a speaker has acquired during their childhood and youth. Without going into the issues of language acquisition, we arbitrarily select the period of 10 years starting from the birth of a speaker to represent such a period. That is to say, if a poet was born in the year X, their pattern of intra-speaker variation must be similar to the pattern of variation in texts from [X; X+10]. The source of texts is going to be the same in our case, the RNC.

To test this hypothesis, we turn again to the example of Nikolay Gumilev. Gumilev was born in 1886, so we are going to compare stress variation in his texts with variation found in the decade from 1886 to 1896 (the sizes of the two corpora are 57,390 tokens and 344,902 tokens respectively). The absence or presence of a particular variable form is not going to be indicative in this case: all gaps might be accidental, especially given the difference in the sizes of the two corpora. Hence we need a different criterion to test agreement between the two corpora. The method we use relies on the notion of the dominant form. For a given form that exhibits stress placement variation in both corpora, if Gumilev's language reflects the language of the selected decade, the dominant variants, that is variants that account for the majority of occurrences of this form, should be the same in both corpora. Examples are shown in Table 4.

Of the 28 words for which there is sufficient data (namely, variation of the same form in

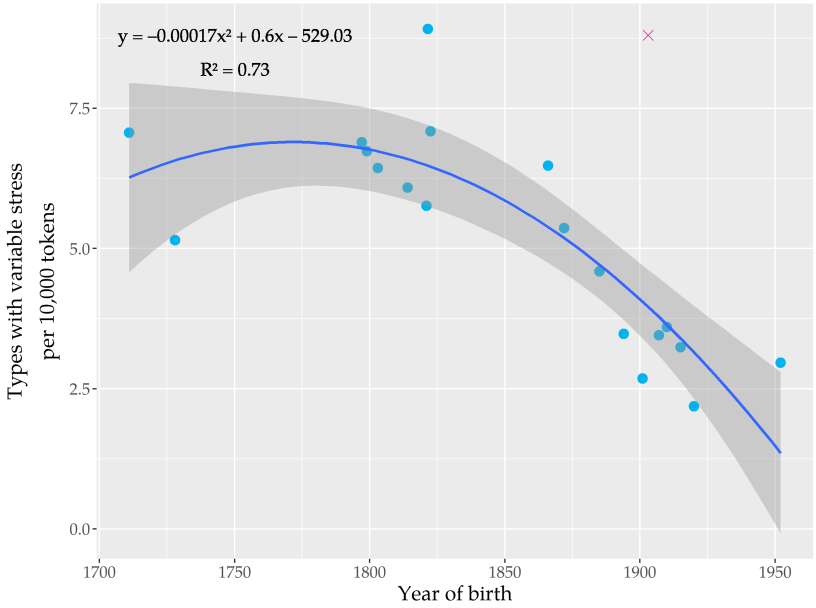

Figure 3: Types with variable stress per 10,000 tokens against the year of birth, Poplavsky excluded; quadratic approximation

both corpora), 24 have the same dominant form. This can hardly be due to chance (binomial test: $p = 0.00018$), and we take this measure to indicate that the language of a particular poet is representative of the language of the decade of their birth, hence changes in the rate of stress placement variation across the poets in our sample are indicative of the changes in the language and poetry overall, not just individual styles.

Naturally, certain questions remain regarding this rather simplistic measure. First of all, ideally the time range for comparison should be determined empirically rather than chosen arbitrarily. Second, it might also be that this measure would give a positive result for any or at least many poet-decade pairs, which might undermine its interpretability. Third, this model adopts the apparent time hypothesis in its strict form, implying that linguistic usage does not change over the course of one's life (Milroy & Gordon, 2003, 35–38). As one of the reviewers points out, this assumption is a simplification; indeed, an individual's linguistic usage can change throughout adulthood (see discussion in Milroy & Gordon 2003). We make this assumption due to the nature of our data. As discussed in Section 4, we counted in all forms that exhibit variable stress placement even where instances with different stress positions did not occur within the same year. The reason is that the amount of variable forms in the corpus is not enough to trace any individual changes that might have occurred in a poet's usage. Therefore, we have to assume that a speaker's system of stress placement remains stable over the course of their life and reflects the usage at the time when they acquired this system.

One other, and probably more straightforward, way to check whether the trend observed in the data from individual poets does not just reflect peculiarities of individual style is to test whether a similar trend can be observed in the data collected by year, where, as discussed above, there are more intervening external factors in the sample, i.e. inter-speaker variation. For this, we used the same procedure as outlined in Section 4 to collect data on stress placement variation from the year 1800 till 1950 with an increment of 5 years (that is, 1800, 1805, 1810, ..., 1945, 1950).

Recall that, as discussed in Section 5, the number of word form types with variable stress strongly depends on corpus size. We used a simple Monte-Carlo simulation to arrive at an appropriate measure of variability, but in fact one can calculate the expected count of types with variation based on the size of the original corpus N and frequencies of variant forms in this corpus. Let us state the question as follows: if a word occurs f_1 times with one stress and f_2 times with another stress

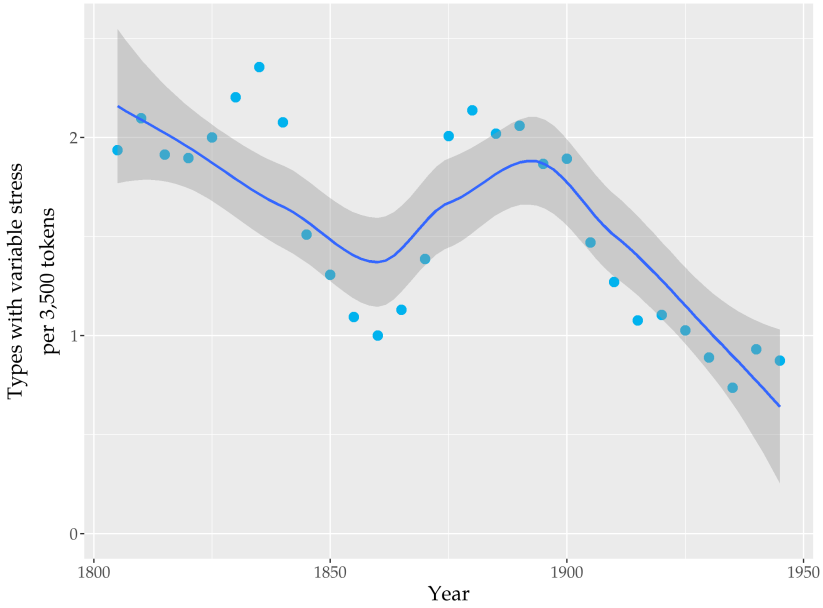

Figure 4: Types with variable stress per 3,500 tokens

in a corpus of size N, what is the probability that we will observe stress variation in this word in a sample of size S (which shows how much this word contributes to the expected count of types with variation)? In order to observe stress variation, we need to encounter each variant form at least once. Thus, we must calculate $P(stress_1 \in sample)$, i.e. the probability of $stress_1$ occurring in the sample, $P(stress_2 \in sample)$, i.e. the probability of $stress_2$ occurring in the sample, and the probability of the joint event $P(stress_1 \in sample \wedge stress_2 \in sample)$.

(12) $P(stress_i \notin sample) = (1 - f_i/N)^S$
$P(stress_i \in sample) = 1 - (1 - f_i/N)^S$
$P(stress_1 \in sample \wedge stress_2 \in sample) = (1 - (1 - f_1/N)^S)(1 - (1 - f_2/N)^S)$

The sum of these probabilities for all form types is the expected count of form types with stress variation.[5] We used this measure to establish the rate of stress variation from 1800 to 1945 in the data we collected. We held $S = 3{,}500$ based on the size of the smallest corpus within the range (3,966 tokens in 1850). The results are given in Figure 4, where moving averages are presented, i.e. the data for 1805 is actually the mean of the results for 1800, 1805, and 1810.

The decline in the rate of stress variation can still be observed in these data, confirming that our initial result is not a reflection of individual poetic techniques and styles. Note, however, that in Figure 4 we see two periods of decline in stress variation: one in the mid-19th century and another in the first half of the 20th century. Again, the question of whether these changes are due to general tendencies within Russian or tendencies particular to Russian poetry is worth asking. Recall that due to the presence of interfering factors, results of the analysis by year are less readily interpretable, e.g. the years showing the dip in variation might contain texts by authors from a narrower range of social and geographical backgrounds. Presumably, this effect might combine both the general trend towards the reduction in the rate of stress variation over time and poetic preferences of different periods. At present, we do not have a more insightful explanation and, to reiterate, only note that the overall trend that we observed in the data from individual poets is roughly replicated in these data as well.

One final note concerning the data by year is that these data provide an additional argument against the effect of poetic license. In (13), we give forms with stress variation from the years 1845

[5] These formulae are based on a simplifying assumption that the two events $stress_1 \in sample$ and $stress_2 \in sample$ are independent. This is not strictly correct, but the error introduced by this assumption is negligible.

(126 texts; 21,730 tokens containing more than one syllable and at least one stress mark), 1850 (57 texts; 3,960 tokens), and 1855 (106 texts; 13,347 tokens).

(13) a. **1845**: *vetvjami*, *vetvjax*, *vysoko*, *gluboko*, *daleko*, *detjam*, *dolžno*, *idut*, *inače*, *koni*, *men'šim*, *(na) nebo*, *nuždu*, *nuždy*, *podnjalsja*, *(po) polu*, *probuditsja*, *purpurnoj*, *(na) serdce*, *sestram*, *sčastliv*, *totčas*, *utra*, *xolma*
b. **1850**: *vysoko*, *(na) nebe*
c. **1855**: *vysoko*, *gluboko*, *dolžno*, *zvezdami*, *znamena*, *idut*, *manjat*, *(na) nebe*, *obnjal*, *operšis′*, *pomost*, *razdalsja*, *rvalsja*, *(iz) sadu*, *statui*, *totčas*

Note that these sets are intersecting: *vysoko* and forms of *nebo* with prepositions are found in all three; *gluboko*, *dolžno*, *idut*, *totčas* in two out of three. If poetic license were the main effect accounting for the variation of stress placement, we would expect the variation to occur in random forms and would not expect to see the same set of forms exhibit stress variation in different samples.

To sum up, we hope to have shown that the rate of stress placement variation in Russian poetry, based on a sample from the Poetic subcorpus of the RNC, decreases over time. Following the discussion in the present section, we would like to suggest that this generalisation applies not only to Russian poetry but to the Russian language in general. However, caution is necessary in drawing this conclusion. A reviewer points out that it requires further assumptions that need not be true, primarily the crucial assumption that variation patterns observed in poetry are representative of variation patterns observed in "regular" speech. Arguing against the prevalence of poetic license, we hope to have convinced the reader that poetry does not casually exhibit ungrammatical forms, i.e. instances of stress placement that are never found outside of a given poem. What we have not shown is that the language use found in poetry faithfully represents language use in general and that whatever changes happen in poetic language reflect changes in the rest of the language. In other words, the language of poetry might exhibit forms that are rarely found in contemporary non-poetic use and are deemed to be archaic, elevated, or otherwise suboptimal by the speakers. For example, there exist certain accentual variants of some words that might only pertain to poetic style, e.g. *dalëko* 'far away' as opposed to the form *dalekó*, much more widespread nowadays. And even the extent to which the language of poetry deviates from non-poetic language can change over time, thus obscuring the relationship between the two even further.

We cannot exclude the possibility that the trend we have found is driven at least in part by a decline in stylistic variation in poetry rather than by language change. Factors such as the establishment and increasing codification of the literary standard, as well as (self)censorship or poetic trends could have influenced this pattern as well. All in all, while it is plausible that language change underlies the trend we observe in the language of poetry, it might not be the only or even the main factor contributing to this trend. At present, we are not in a position to prove that language change is indeed the primary factor determining the decline of stress variation in Russian poetry. One might approach this issue by showing that the variation found in present-day poetry reflects variation in the present-day language. For instance, going back to the example in (10), three out of five words with variable stress in the corpus of Gandlevsky, namely *kladbiščе* 'cemetery', *p'jany* 'drunk (nom. pl.)', and *totčas* 'at once', are cited with variable stress in Kuznetsov (2014). One more word, *sredam* 'Wednesday (dat. pl.)', has variable stress in Kasatkin (2012). For contemporary poetry, it might even be possible to check the stress variation it exhibits against spoken or accentuated corpora. However, the only viable option for the earlier periods is comparison with the available dictionaries since stress is not indicated in Modern Russian orthography. While such dictionaries exist, they are not many, do not cover all of the relevant time frame, and naturally rely on the judgment of the compilers rather than corpus evidence, especially before the second half of the 20th century. Given these difficulties, we leave such an investigation for the future.

7 CONCLUSIONS

Based on a sample of 20 poets (with years of birth ranging from 1711 to 1952) from the Poetic subcorpus of the Russian National Corpus, we have shown that the amount of stress placement variation in Russian poetry decreases over the last three centuries. With caution, this trend can be

projected onto the Russian language in general, although the connection between stress variation in poetry and "regular" language requires further investigation. Nevertheless, we maintain that this test case demonstrates that poetic corpora are a useful and valid resource for the study of lexical stress and its variation in Russian and beyond, especially when other sources are lacking.

The next step in this investigation is to analyse data from more poets and years to test the central claim on a larger set of data. This line of inquiry also opens a range of connected questions, such as what the direction of accentual change is in this data and what drives this change, or whether certain grammatical classes are more susceptible to the change than others, or whether there is a connection between the rate of change and the extent of variation if a particular class. These and other directions for this strand of research (for instance, suggested in Zalizniak 2015) remain for the future.

8 ACKNOWLEDGEMENTS

The authors would like to thank the audiences at FASL 28, the editors of this issue, and the two anonymous JSL reviewers. All mistakes are our own.

9 CONTACT

Alexander Piperski — apiperski@gmail.com
Anton Kukhto — kukhto@mit.edu

REFERENCES

Alderete, John. 1999. *Morphologically governed accent in Optimality Theory*. Amherst: University of Massachusetts dissertation.

Avanesov, Ruben I. (ed.). 1983. *Orfoèpičeskij slovar' russkogo jazyka [A pronouncing dictionary of Russian]*. Moscow: Nauka.

Dobrushina, Nina R. & Daria A. Staferova. 2018. Variational Studies Repository. Available at https://vastry.ru/. NRU HSE, Moscow.

Gasparov, Mikhail Leonovich. 2000. *Očerk istorii russkogo stixa [A concise history of the Russian verse]*. Moscow: Fortuna Limited.

Goedemans, Rob & Harry van der Hulst. 2013. Weight-sensitive stress. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info/chapter/15. Accessed on 2019-11-18.

Gouskova, Maria. 2010. The phonology of boundaries and secondary stress in Russian compounds. *The Linguistic Review* 27(4). 387–448.

Gouskova, Maria & Kevin Roon. 2013. Gradient clash, faithfulness, and sonority sequencing effects in Russian compound stress. *Laboratory Phonology* 4(2). 383–434.

Grishina, Elena Aleksandrovna. 2006. Spoken Russian in the Russian National Corpus (RNC). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa: European Language Resources Association.

Halle, Morris. 1973. The accentuation of Russian words. *Language* 49(2). 312–348.

Jakobson, Roman. 1960. Closing statement: Linguistics and poetics. In T. Sebeok (ed.), *Style in Language*, 350–377. Cambridge, MA: MIT Press.

Jakobson, Roman. 1963. Opyt fonologičeskogo podxoda k istoričeskim voprosam slavjanskoj akcentologii [An attempt at a phonological approach to the historical issues of Slavic accentology]. In *American Contributions to the 5th International Congress of Slavists.* Vol. 1: *Linguistic contributions*, 153–178. The Hague: Mouton.

Kasatkin, Leonid L. (ed.). 2012. *Bol'šoj orfoèpičeskij slovar' russkogo jazyka [A big pronouncing dictionary of Russian].* Moscow: AST-Press.

Kiparsky, Paul. 2010. Compositional vs. paradigmatic approaches to accent and ablaut. In Stephanie W. Jamison, H. Craig Melchert & Brent Vine (eds.), *Proceedings of the 21st Annual UCLA Indo-European Conference*, 137–181. Bremen: Hempen.

Kolmogorov, Andrey N. & Alexandr V. Prokhorov. 1968. K osnovam russkoj klassičeskoj metriki [Toward the foundation of Russian classical metrics]. In Boris S. Meylakh (ed.), *Sodružestvo nauk i tajny tvorčestva [The symbiosis of sciences and the secrets of creativity]*, 397–432. Moscow: Iskusstvo.

Korchagin, Kirill. 2008. Poètičeskij podkorpus nacional'nogo korpusa russkogo jazyka kak akcentologičeskij istočnik [Poetry subcorpus of Russian National Corpus as an accentological source]. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2008"*, vol. 9, Moscow: Izd-vo RGGU.

Kukhto, Anton & Alexander Piperski. 2020. Lexical stress variation and rhythmic alternation in Russian: A pilot study. *Linguistic Variation* 20(1). 33–55.

Kuznetsov, S. A. (ed.). 2014. *Bol'šoj tolkovyj slovar' russkogo jazyka [A big explanatory dictionary of the Russian language].* Saint Petersburg: Norint. 1st edition, 1998.

Melvold, Janis L. 1989. *Structure and stress in the phonology of Russian*. Cambridge, MA: MIT dissertation.

Milroy, Lesley & Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.

Mołczanow, Janina, Ekaterina Iskra, Olga Dragoy, Richard Wiese & Ulrike Domahs. 2019. Default stress assignment in Russian: Evidence from acquired surface dyslexia. *Phonology* 36(1). 61–90.

Piperski, Alexander & Anton Kukhto. 2016. Intra-speaker stress variation in Russian: A corpus-driven study of Russian poetry. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"*, vol. 15, 540–550. Moscow: Izd-vo RGGU.

Ryan, Kevin. 2014. Onsets contribute to syllable weight: Statistical evidence from stress and meter. *Language* 90(2). 309–341.

Say, Sergey S. 2020. *Grjaz'*, *pyl'* i *krov'* v poètičeskom korpuse: semantizacija akcentovok v konstrukcijax s predlogom *po* ['Dirt', 'dust', and 'blood' in the poetic corpus: Semantic divergence of stress patterns in constructions with the preposition *po*]. In et al. Kibrik, A. A. (ed.), *VAProsy jazykoznanija: Megasbornik nanostatej. Sbornik statej k jubileju V. A. Plungjana [A festschrift for Vladimir A. Plungian]*, 116–120. Moscow: Buki Vedi.

Shengeli, Georgy A. 1940. *Texnika stixa [Versification technique].* Moscow: Sovetskij pisatel'.

Steriade, Donca. 2009. The phonology of perceptibility effects: The P-Map and its consequences for constraint organization. In K. Hanson & Sharon Inkelas (eds.), *The Nature of the Word: Essays in Honor of Paul Kiparsky*, 151–179. Cambridge, MA: MIT Press.

Zalizniak, Andrei A. 1985. *Ot praslavjanskoj akcentuacii k russkoj [From Proto-Slavic to Russian accentuation].* Moscow: Nauka.

Zalizniak, Andrei A. 2012. Mexanizmy èkspressivnosti v jazyke [Mechanisms of expressivity in language]. In et al. Apresjan, Ju. D. (ed.), *Smysly, teksty i drugie zaxvatyvajuščie služety. Sbornik statej v čest' 80-letija Igorja Aleksandroviča Mel'čuka [Meanings, texts, and other exciting things. A festschrift on the occasion of the 80th anniversary of Igor A. Mel'čuk]*, 650–664. Moscow: Jazyki slavjanskoj kul'tury.

Zalizniak, Andrei A. 2014. *Drevnerusskoe udarenie: Obščie svedenija i slovar' [Old Russian stress: General remarks and a dictionary]*. Moscow: Jazyki slavjanskoj kul'tury.

Zalizniak, Andrei A. 2015. Èpizod iz istorii russkogo udarenija [An episode from the history of Russian stress]. Lecture given at the 17th Summer School in Linguistics, Dubna, July 9, 2015. Available at http://www.mathnet.ru/php/seminars.phtml?option_lang=rus&presentid=11930.