Quantitative Approaches to the Russian Language

Edited by
Mikhail Kopotev, Olga Lyashevskaya
and Arto Mustajoki

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Contents

PART III

PART IV

1 Russian challenges for quantitative research

Mikhail Kopotev, Olga Lyashevskaya, and Arto Mustajoki

1. Introduction

The Russian language, despite being one of the most studied in the world, until recently has been little explored quantitatively. After a burst of research activity in the years 1960–1980, quantitative studies of Russian vanished. They are now reappearing in an entirely different context. Today, we have large and deeply annotated corpora available for extended quantitative research, such as the Russian National Corpus, ruWac, ruTenTen, to name just a few (websites for these and other resources will be found in a special section in the References). The present volume is intended to fill the lacuna between the available data and the methods that can be applied to studying them.

Our goal is to present current trends in researching Russian quantitative linguistics, to evaluate the research methods vis-à-vis Russian data, and to show both the advantages and the disadvantages of the methods. We especially encouraged our authors to focus on evaluating statistical methods and new models of analysis. New findings concern applicability, evaluation, and the challenges that arise from using quantitative approaches to Russian data. The goal of this volume is therefore twofold: a) to address the topic of quantitative analysis of the Russian language, and b) to present an evaluation of methods applied to Russian data.

2. Main features of the Russian language

The Russian language is the mother tongue of 163.8 million people around the world, the majority of whom (around 130 million) live in the Russian Federation. The language belongs to the large Indo-European family; it shares origins with English, French, Italian, Greek, Persian, Hindi, and some 50 other tongues. The closest related languages to Russian are Slavonic: East Slavonic (Ukrainian, Belorussian), West Slavonic (Polish, Czech, Slovak), and South Slavonic (Bulgarian, Serbian, Slovenian), and Baltic (Latvian and Lithuanian). Russian uses the Cyrillic alphabet (for instance, стиль ‘style’, рок ‘rock’). It shares various grammatical features with the Indo-European linguistic community, but it has also many differences.

2.1 Morphology

Russian is a morphologically rich language. *The MULTEXT-East morphosyntactic specification* (Erjavec et al., 2010), which is intended to create a unified cross-language annotation scheme, defines 156 POS-specific morphosyntactic values for Russian as compared to 80 for English, for example, and 191 for Hungarian (see http://nl.ijs.si/ME/V4/msd/html/msd-ru.html).

The **Stress** in Russian can fall on any syllable; cf. *prínter* ‘*printer*’, *proféssor* ‘professor’, *inženér* ‘engineer’. The stress is movable in the sense that different morphological forms of a lexeme may have different syllable structures: *stol* ‘table-NOM’, *stolá* ‘table-GEN’; it can also differentiate morphological forms: *proféssora* ‘professor-GEN-SG’, but *professorá* ‘professor-PL.NOM’.

Nouns have three basic genders, as they do in German – masculine, feminine, and neuter – but they lack articles. The gender is also marked in adjectives, predicatives, and some verbal forms. There are six main cases in Russian: nominative, accusative, genitive, dative, instrumental, and prepositional/locative. The oblique case forms are nearly always manifested by non-zero endings: *student* ‘male student’:

Case	*Singular*	*Plural*
nominative	student	student+y
accusative	student+a	student+ov
genitive	student+a	student+ov
dative	student+u	student+am
instrumental	student+om	student+ami
prepositional	student+e	student+ah

There is often a regular ambiguity whenever two case endings merge: *studentk+i* ‘female student’ stands for SG-GEN and PL-NOM, while *studentk+e* stands for SG-DAT and SG-PREP. In a certain group of nouns, even more endings can merge: *stepi* ‘steppe’ – SG-GEN, SG-DAT, SG-PREP, PL-NOM, and PL-ACC. Because of the regular ambiguity, especially in written texts, some forms are a challenge to disambiguate, even in context; consider the following:

On	*ne*	*videl*	*syna*
he-nom-sg	not-prtcl	see-pst-1-cg	**son**-acc-sg/**SON**-gen-sg

‘He did not see (his) son’

All oblique cases are also used with prepositions: ***v*** *universitet* (SG-ACC) ‘to the university’; ***v*** *universitete* (SG-PREP) ‘in the university’; ***iz*** *universiteta* (SG-GEN) ‘from the university.’

The Russian language marks the **animacy** of nouns. For historical reasons, masculine nouns in the singular and all nouns in the plural denoting “living” objects make a special instance in the accusative, which coincide with the genitive form, while other nouns have an accusative that fits the nominative forms.

Adjectives normally agree with nouns; they have **case, gender**, and **number**, as well as **comparative and superlative forms** (the latter not provided in Table 1.1). Typologically more rare are **long and short forms**, where the latter are exclusively reserved to mark a predicate role.

Table 1.1 Declension of the adjective *kreativnyj* ‘creative’

	Long form, Singular			*Long form, Plural*	*Short form, Singular*			*Short form, Plural*
	Masculine	*Feminine*	*Neuter*		*Masculine*	*Feminine*	*Neuter*	
Nominative	kreativnyj	kreativnaja	kreativnoje	kreativnyje	kreativen	kreativna	kreativno	Kreativny
Genitive	kreativnogo	kreativnoj	kreativnogo	Kreativnyx				
Dative	kreativnomu	kreativnoj	kreativnomu	kreativnym				
Accusative	kreativnyj/ kreativnogo	kreativnuju	kreativnoe	kreativnyx/ kreativnyje				
Instrumental	kreativnym	kreativnoj	kreativnum	kreativnymi				
Prepositional	kreativnom	kreativnoj	kreativnom	kreativnyx				

The Russian adjectives should agree with their syntactic heads:

kreativnyj	klass
creative-NOM-M-SG	class-NOM-M-SG
‘creative class’	

kreativnym	klassom
creative-INS-M-SG	class-INS-M-SG
‘creative class’	

rieltory	kreativny
realtor-NOM-PL	creative-PL-SHORT
‘realtors (are) creative’	

Russian **Verbs** possess a rich morphology. According to (Janda and Lyashevskaya, 2011), there are 189 word-forms for each regular verb. As with other Slavonic languages, most Russian verbs have two **aspectual variants**, imperfective and perfective. In addition, Russian has a relatively large group of biaspectual verbs, which, depending on the context, convey both perfective and imperfective meaning. The pure aspectual pair can be roughly interpreted as follows:

Imperfective aspect

Vladimir	guglil	press-reliz
Vladimir-NOM	google-IMP.PST.SG	press-release-ACC.SG
‘Vladimir was googling a/the press release’		

Perfective aspect

Vladimir	poguglil	press-reliz
Vladimir-NOM	google-PERF-PST-SG	press-release-ACC-SG
‘Vladimir has googled a/the press release’		

Verbs have person forms marked in present and future tenses, while the past tense is marked with the so-called preterit, which is historically rooted in a short participle and morphologically close to adjectives. Thus, only presence and future, both simple and analytical, conjugate in a narrow sense (that is, by having a person marking):

Table 1.2 Conjugation of the verb *guglit’* ‘to google’

Person	*Present tense, Imperfect*	*Simple Future tense, Imperfect*	*Analytical Future tense, Perfect*	*Preterit, Imperfect*	*Preterit, Perfect*
ja ‘I’	guglju ‘I am googling’	budu guglit’ ‘I will google’	poguglju ‘I will google’	on guglil ‘he googled’	on poguglil ‘he has googled’

Person	*Present tense, Imperfect*	*Simple Future tense, Imperfect*	*Analytical Future tense, Perfect*	*Preterit, Imperfect*	*Preterit, Perfect*
ty 'You'	gugliš'	budeš' guglit'	poguglíš'	ona guglila	ona poguglila
	'You are googling'	'You will google'	'You will google'	'she googled'	'she has googled'
on(a) 'he/she'	guglit	budet guglit'	poguglit	ono guglilo	ono poguglilo
	'he/she is googling'	'he/she will google'	'he/she will google'	'it googled'	'it has googled'
my 'we'	guglim	budem guglit'	poguglim	oni guglili	oni poguglili
	'we are googling'	'we will google'	'we will google'	'they googled'	'they have googled'
vy 'you'	guglite	budete guglit'	poguglite		
	'you are googling'	'you will google'	'you will google'		
oni 'they'	gugljat	budut gulit'	pogugljat		
	'they are googling'	'they will google'	'they will google'		

Verbs also have non-conjugated Gerund and Infinitive forms:

Table 1.3 Non-conjugated forms of the verb *guglit'* 'to google'

	Imperfect	*Perfect*
Gerund, Past tense	gugliv '(while) googling'	pogugliv 'having googled'
Gerund, Present tense	guglja '(while) googling'	poguglja '(after) googling'
Infinitive	guglit' 'to google'	poguglit' 'to google'

Participles have Active and Passive forms with the adjectival declension (see above), including long and short forms, but excluding comparative and superlative ones. The short forms are used exclusively with a predicate head.

Russian, as other Slavonic languages, developed **reflexive verbs** that are a salient category associated with changes in the structure of argument constructions. Slightly simplifying the issue, this category basically marks reflexive meaning with the suffixes *-sja*, which is diachronically derived from the reflexive pronoun *sebja* '-self'. In Modern Russian, the suffix is ambiguous between reflexive and passive reading, as in:

> *guglit'* google-INF-ACT 'to google'
> *goglit'-**sja*** google-INF-REFL 'to google oneself'/to-google-PASS 'to be googled'

2.2 Syntax

Unlike most Indo-European languages, Russian belongs to the BE languages, which means that the verb *byt'* 'to be' is mainly used in those morphological forms and constructions where we expect the verb 'to have' in English. For example, the possessive construction is usually built by using *byt'* 'be' and a preposition:

U	Vladimira	est'	Ferrari
u-PREP	Vladimir-GEN-SG	**be-3-sg**	Ferrari-NOM

'Vladimir has a Ferrari'

Russian favors syntactic zeros; it has pro-dropping and zero verbs, including copula. Also, it has constructions without nominative subjects; cf:

ØPro	guglju	press-reliz
ØPro-1-SG	google-PRAES-1-SG	press-release-ACC-SG

'I am googling a press release'

Vladimiru	Øcop	holodno
Vladimir-DAT	ØCop-PRAES-3	cold-PRED

'Vladimir is cold'

Russian is an SVO language, but the word order is much more flexible than in English. Thus, in certain contexts all of the following clauses are possible:

Vladimir press-reliz guglil
Press-reliz Vladimir guglil
Guglil Vladimir press-reliz
Vladimir guglil press-reliz
'Vladimir was googling a press release' (The last example shows the neutral and most common word order.)

There are strict rules for agreement in Russian. The forms of the predicative and adjective depend on the form of the noun. The verb controls the case of a noun or the preposition; for example:

Vladimir	guglit	novyj	press-reliz	**na**	kompjutere
Vladimir-N-SG	google-3-SG	new-SG-ACC	press-release-SG-ACC	on-prep	computer-sg-prep

'Vladimir is googling a new press release on a computer'

3. Quantitative (corpus) studies in Russian (2000–2010s)

3.1. Russian corpora

The re-launch of quantitative research in Russian took place since the dawn of the twenty-first century, not surprisingly shortly after the emergence of digital Russian corpora (Kopotev and Mustajoki, 2008). Among the first Russian corpora annotated morphologically were the Uppsala corpus of Russian texts (later incorporated into the Russian Corpora in Tübingen [Lönngren, 1993]), the MSU newspaper corpus (Vinogradova et al., 2001), and the Helsinki annotated corpus HANCO (Kopotev and Mustajoki, 2003), which was also annotated syntactically. Today, the Russian language as a source for computational linguistic studies is highly developed in terms of available corpora and tools for natural language processing (NLP); it is also well represented in both national computational linguistic landscapes (see the "Dialogue" conferences at www.dialog-21.ru/en) and international collaboration (see Erjavec et al., 2010).

In 2004, the **Russian National Corpus** (RNC) became available. Having developed since then into a functional and extensively annotated resource, the RNC is now comparable both in its size and scientific value to the ANC, BNC, Czech, Polish, and other national corpora. Its core collection includes manually selected samples of written texts of various genres (ca. 600 million words) and a 7.6-million-word corpus of spoken texts. Both written and spoken parts are annotated morphosyntactically (lemmatization, POS tagging, deep morphological annotation) and semantically (lexical classes). Rich metadata and sophisticated search options, such as multiword expressions, tag repetition in adjacent tokens, and stress marking, are the crowning strokes of this monumental resource. Two spin-off projects of the RNC – the SynTagRus treebank (Boguslavsky et al., 2000, 2009) and the FrameBank (Lyashevskaya and Kashkin, 2015) – are manually annotated with syntactic dependencies and lexical functions markup (SynRagRus) and with semantic roles (Framebank). Not least, the RNC includes a unique, to our knowledge, corpus of Russian poetry, providing the ability to search by meter and rhyme type.

Other big data resources for Russian include:

- the 1.3-billion-token ruWac, the Russian portion of the project Web as a Corpus (Sharoff, 2006; Sharoff and Nivre, 2011);
- the 14.5-billion-word ruTenTen, a member of the commercial TenTen corpus family (Jakubíček et al., 2013);
- the 50+ billion word Integrum database (Mustajoki, 2006);
- the 1.2-billion-token Araneum Russicum corpora (Benko, 2014);
- the recently announced 19.8-billion-token General Internet Corpus of Russian (GICR; see Belikov et al., 2013).

With only one exception (the Integrum database), all of these corpora are morphologically tagged and available for downloading; additionally, ruWac is automatically annotated with dependency trees, whereas ruTenTen is parsed for shallow syntactic patterns. Furthermore, large text collections (both plaintext and POS-tagged), such as lib.ru, lib.rus.ec, and the Russian Wikipedia dump, are widely in use in linguistic research.

Considerably smaller corpora are intended for linguistic research in specific subfields, of which the following are worth mentioning:

- L1 and L2 acquisition and heritage corpora: Russian CHILDES (Voeikova, 1995); RuLeC/RLC – Russian Learner Corpus (Alsufieva et al., 2012; Klyachko et al., 2013); the Corpus of Russian Student Texts (CoRST; Zevakhina and Dzhakupova, 2015); RusLTC – the Russian Learner Translator Corpus (Kunilovskaya and Kutuzov, 2014);
- Russian dialects and historical corpora: Ustya River Basin corpus (von Waldenfels et al., 2014); Kazan University digital library of Russian folk dialects (Kulsharipova and Ibragimov, 2013); database of Russian Folk Dialects (Krylov and Ter-Avanesova, 2012);
- Speech corpora: the ORD (Russian abbreviation for *One Speaker's Day*) corpus, (Asinovsky et al., 2009; Sherstinova, 2009); the corpus of Russian professionally-read speech CORPRES (Skrelin et al., 2010); the Prosodically Annotated Corpus of Spoken Russian PrACS-Russ (Podlesskaya, 2015); the corpus of Pear film-based stories Russian CLiPS (Khudyakova et al., 2016);
- corpora of non-verbal communication: MURCO (Grishina, 2010); the Russian Emotional Corpus (Kotov, 2012).

To access specific linguistic phenomena, a researcher can also use:

- the 1-million-word RSTB Treebank, the 20-million-word RUS-Treebank, and two UD-Russian corpora annotated with syntactic dependencies (Toldova et al., 2015; Kuznetsov, 2016; Droganova and Lyashevskaya, in manuscript);
- the .2-million-word Russian Coreference Corpus (RuCor), annotated for anaphoric chains (Toldova et al., 2016); a similar project is OpenCorpora (Protopopova et al., 2014);
- the Russian Paraphrase Corpus, containing 7,000 paraphrased pairs of news titles (Pronoza et al., Forthcoming);
- a few collections annotated for named entities (Starostin et al., 2016; Vlasova et al., 2016);
- the 6-million-word Corpus of Russian Twitter, which is supplied with such pragmatic attributes as the relevance and importance of a posting and the strength of its impact on the reader (Rubtsova and Zagorulko, 2014).

The list of corpora provided in this section is by no means exhaustive; there are a number of others, for example, historical and parallel corpora (see Mitrenina, 2014; Zakharov, 2014; Mikhailov and Cooper, 2016 for more details). There are numerous other corpora created for NLP and information retrieval purposes, but which are not publicly available and, thus, beyond the scope of our review. However, Russian data are well represented in many forms and formats. The next section is dedicated to the question of how to deal with the material further.

3.2. Corpus-based resources

Quantitative studies usually involve analysis of certain linguistic features contrasted with control conditions. The frequency distributions observed in a sample are compared with the overall frequency patterns obtainable by scanning a general corpus. Many studies in Russian lexicon and grammar use the RNC-based *Chastotnyj slovar' sovremennogo russkogo jazyka* (*A Frequency Dictionary of Russian;* Lyashevskaya and Sharoff, 2009), which reports occurrences of the 50,000 most frequent lemmas. The *Dictionary* allows a user to trace changes in the distribution of words in time and registers (fictional domains), as well as to obtain so-called "idiosyncratic word lists," meaning those whose frequency is register-biased. Besides *A Frequency Dictionary*, there is a ruWac-based *The frequency dictionary for Russian* (Sharoff et al., 2014), which contains 5,000 lemmas, and a *Statisticheskij slovar' jazyka russkoj gazety* (*A Statistical Dictionary of Modern Russian Newspapers* (Shaykevich et al., 2008), which includes the 104,000 most frequent lemmas used in this domain. The further steps in this direction will be the *Frequency grammar of Russian* (Kopotev, 2008), which summarizes the frequencies of Russian morphological categories, and the *Lexico-grammatical frequency dictionary* of Russian, which shows the distribution of grammatical forms in the inflectional paradigm of Russian nouns, adjectives, and verbs (Lyashevskaya, 2013).

A valuable source for investigating word combination properties is known as *n-grams*, that is, word sequences of length *n*, e.g., bigrams or trigrams. The RNC provides the lists with raw frequencies for 1-, 2-, 3-, 4-, and 5-grams, which are found in the main corpus two, three, or more times (the so-called RNC n-grams). In addition, a list of relation-based n-grams (sequences of words connected syntactically) can be accessed via the RNC Sketches service. A Google n-grams viewer likewise provides data on both window-based n-grams (co-occurrences within a window of a given size) and syntactic n-grams (dependency tree fragments) extracted from the large Russian Google Books collection. Several services make tools available online for n-gram searches pre-aggregated with various automatic methods:

- the SketchEngine service includes a Russian portion available upon registration (see ruTenTen in the References);
- Basic tools for collocation extraction (MI, t-score, log-likelihood) are available on the University of Leeds' Russian corpora page (see ruWac in the References);
- a tool for relation-based (syntactic) collocations extraction is under development within the CoSyCo project (Lukashevich et al., 2016);

- the CoCoCo service provides a tool for tied rankings of both colligations (which means mutually occurring lexical items *and* grammatical features) and collocations (Kopotev et al., 2015; see also Pivovarova et al., in this volume);
- semantic distances between lemmas calculated from word context similarities are available via the RusVectores webpage (see Kutuzov and Kuzmenko, in this volume).

3.3. Corpus-based quantitative Russian studies

This review is visibly corpus-oriented, since experimental, that is, psycholinguistic papers usually follow a standard language-independent research design and are thus of less interest in terms of methods used. Recent advances in corpus linguistics, along with the development of statistical approaches to data analysis and statistical software, have profoundly affected almost all fields of linguistic research (see Facchinetti, 2007; Divjak and Gries, 2012; and Janda, 2013, among many others). In the next sections, we present a selection of works that, since 2000, have contributed to exploring quantitative methods as applied to various Russian language phenomena. Below, we first address the concept of the linguistic profile, which has largely been explored using Russian data and is seen to contribute most to modern linguistics. Second, we review some basic statistical tests before turning to more elaborate multivariate models. For more general details, readers who are not familiar with particular statistical techniques are invited to consult one of the many available textbooks on statistics for linguists (Levshina, 2015; Gries, 2013; and Baayen, 2008, among others).

2.3.1. Linguistic profiles

An important notion of the "behavioral profile" was initially introduced in the first decade of the 2000s by Dagmar Divjak in her works (some of which were co-authored with S.T. Gries) on Russian synonyms (Divjak, 2004, 2010; Divjak and Gries, 2006, 2008, etc.). For this concept, Divjak drew on Patrick Hanks's idea of "lexical profiling" (Hanks, 1996) and the notion of a "lexicographic portrait" developed in the Moscow Lexicographic School (Apresjan, 2000). Basically, Apresjan's approach presupposes that the contexts of a lexeme should be systematically studied at the following levels:

- Grammar: inflection and its underpinning and constraints for a certain lexeme;
- Collocations: mutual preferenses for lexeme/token co-ocurancies;
- Constructions: the ability to form case frames and other syntactic patterns.

Based on corpus data, Apresjan (2004) demonstrated that synonyms systematically differ at all levels and are not completely interchangeable, since their substitution is only possible in a small number of contexts. As a part of her research, Divjak studied a group of Russian near-synonyms for the word with the meaning "trying" based mainly on corpus data. The distance between synonymic pairs was calculated by taking into account the frequencies of 87 manually annotated variable categories, thus yielding the hierarchical clustering of synonyms, comparable (but not identical) to the results obtained by Apresjan. The notion of *behavioral profile* proposed by Divjak includes all levels of linguistic annotation feasible for a moderately time-consuming manual analysis (numerous weeks or months).[1]

In research conducted by the CLEAR group located at the University of Tromsø, Norway, this approach has been continued at the more individual levels of profiling. Laura Janda and Olga Lyashevskaya (2011) and Yulia Kuznetsova (2013) studied the **grammatical profiles** of Russian verbs, which are understood as "a relative frequency distribution of the inflected forms of a word in a corpus" (Janda and Lyashevskaya, 2011, p. 719). Examples of research on **constructional profiling** can be found in Soloviev and Janda (2009) and Sokolova et al. (2012). The former studied frequencies of prepositions used with the near-synonymous nouns *grust', pečal', toska* 'sadness.' In the latter, the choice between two valency patterns is modeled on verbs of loading (*gruzit' telegu/na telegu* 'to load a wagon/ on a wagon'). This is known as the locative alternation, which is predictable by morphosyntactic parameters of a verb form, that is, aspect, prefixes used, finite vs. participle form, etc. One more type of profile – **the lexical profile** – is studied in Kuznetsova (2013), where the relative distribution of lexeme variables within a construction is compared to the overall distribution of the same class of lexemes in the corpus. This tactic is close to what Michael Stubbs (2001) also calls a lexical profile. One other approach to profiling is the **semantic profile**, which is the relative frequency distribution of word senses across a corpus, which, in reference to George Lakoff's works, is called a **radial category profile** (Endresen et al., 2012). In her work, Anna Endersen showed that two verbal prefixes *vy-* and *iz-* share a single network of meanings (roughly 'out of'), yet highlight different parts of this network, with more concrete senses being overrepresented in *vy-* rather than in *iz-*.

The list of profiling types can easily be continued by taking into account word order, syntactic and semantic roles, the narrator's viewpoint, or other kinds of linguistic features. Finally, because they are categorized into groups, the lexical items can be generalized according to their profiles, which include grammatical, constructional, and lexical profiles taken together (Janda and Lyashevskaya, 2013). Having various features marked in the corpus, researchers can work not only with a "bag of words," but with a more powerful and finely tuned "bag of tags" to catch the differences in word usages or texts or semantic clusters or individual contexts – all in quantitative terms that are explicable and invaluable.

2.3.2. Basic significance testing

Descriptive quantitative analysis has a respectable tradition in Russian linguistics, featuring numerical data and basic visualizations, such as bar plots, line plots, and scatter plots. It is also rather standard to run the **χ^2-test** to estimate the statistical significance of differences found in data distribution. For example, a recent study based on the Russian multimedia corpus MURCO (Grishina, forthcoming) shows that the deictic words *vot* 'this', *von* 'over there', *eto* 'this, that', etc. tend to be accompanied by a pointing gesture made with an index finger when the speaker is referring to a single object. In contrast, when the speaker is indicating multiple objects, these words tend to be followed by an open palm gesture. The author shows that the differences in gesture cannot be attributed to chance. However, the association between the number of objects being pointed at and palm orientation is not straightforward; although the observed occurrences of palm-down gestures in reference to multiple objects are higher than expected by chance, the χ^2-test fails to reach significance, since the probability of obtaining the same result or even more extreme values in the case of another random sample taken from the population is considerably high (p = 26%).

At the same time, there is growing awareness in the research community of the limited reliability of χ^2 (Gries, 2005; Perry, 2005; Bergsma, 2013; see also a general overview of the problem in behavioral science in Sharpe, 2015); with the increasing sizes of corpora, statistical significance becomes too easy to achieve. Since the measures provided by the χ^2-test are strongly correlated with the sample size n, the χ^2-statistics will always indicate statistical significance when the sample size is large enough (Gries, 2005; Perry, 2005; Bergsma, 2013). A more sophisticated approach can be found in Dickey and Janda (2009), who calculate the effect size using **Cramer's *V*** (also referred to as φ in the case of 2 × 2 tables), a measure that standardizes χ^2 against the sample size and the dimensions of the contingency table. In their study of the distribution of semelfactive markers (i.e., having meanings "occurring once") across the morphological classes of verbs, Dickey and Janda report the statistical significance of the distribution and a large effect size (V = 0.83). Thus, they statistically prove that one cannot obtain the observed distribution by chance, given the sample size of 389 and a 6 x 2 table. Another method of calculating the effect size, **Cohen's *d*** and **Cohen's *h***, is used in Berdichevskii's (2011) study of electronic communications such as chats and emails. Of particular interest is that, while the words are significantly shorter and brackets are significantly underrepresented in chats, the effect size is almost nonexistent in both cases ($h < 0.1$), which suggests that statistical significance was achieved due to the large size of the sample. In contrast, some other factors, such as sentence length and the use of capital letters, are both statistically significant and clearly demonstrate a large/medium effect size.

As alternative measures of statistical significance, the **odds ratio**, the **Fisher exact test**, and the exact binomial test **p-values** have proven reliable for Russian data. For example, the **odds ratio** is applied in (Gorokhova, 2013), who shows that the probability of coming up with taxonomically related speech errors (slips of the tongue), such as substituting *užin* 'dinner' for *zavtrak* 'breakfast,' are much higher than the probability of observing non-taxonomically related pairs, such as *kover* 'carpet' and *pol* 'floor.'

The Fisher exact test is especially useful when data are sparse and/or have an intrinsic bias. This is precisely the case when a particular context pattern or a particular lexical unit is taken under investigation. In performing the collostructional analysis of the verbs of disappearance attracted by the possessive construction (e.g., *U menja issjaklo terpenie* ‘My patience has run out’), Kuznetsova (2013) compares the frequency of a given word in the construction against a) the total frequency of the construction, b) the frequency of the word in all verbal constructions, and c) the frequency of all verbal constructions. In such instances, the expected frequency of the word in the construction is often much less than five. For example, the verb *issjaknut’* ‘run out’ occurs in the possessive construction 11 times, which is much more than the expected one time. Fisher’s exact test shows that the probability of obtaining 11 or more in the cell of the contingency table given the same marginal totals is close to zero ($p=1.260e^{-07}$). In the same way, Fisher’s test for *ubit’* ‘to kill’ estimates the probability as low as $p=0.03$ (which means that the verb is repelled from interaction with the possessive construction); for *skryvat’sja* ‘to disappear,’ the test provides p-values above the 0.05 threshold (which means being neutral to the construction).

Kizach (2012) used **the exact binomial test** to investigate the hypothesis that length differences affect word order so that the shortest argument tends to be placed before a longer one (in adversity impersonal constructions such as *Lodku[Object] uneslo[Impersonal Verb] vetrom[Instrument]* ‘The boat.$_{ACC}$ was carried away by the wind.$_{INS}$’). Given the overall size of the corpus sample, Kizach estimates the chance of random word allocation to be 1/6 (16.7 percent or 21 times in this particular case, given the six possible word orders of O, impV, and I). However, the principle “the shorter element is placed first” was observed in 67 percent of cases, which is significantly more often than expected (successes=85, n=127, $p<0.0001$).

As for non-categorical, continuous variables, the **standard error** and the 95 percent **confidence interval** are reported to be the most often utilized (typically shown as error bars on the plots). Both measures estimate the possible difference in means if samples are randomly selected from the population they represent. The confidence interval is a range of values around that statistic that are believed to contain, with a probability of 95 percent, the true value of that statistic (Field, 2013). For example, Nagy et al. (2014, p. 64), in studying the English influence on the phonetics of heritage Russian speakers in Toronto, Canada, found that, across three generations of speakers, the voice onset time in the word-initial voiceless stops /p/, /t/, /k/ in stressed position moves away from the monolingual Russian short lag and toward the long lag characteristic of English. Since for each stop, the confidence intervals calculated in each generation do not overlap, this can informally signal the statistical significance of the difference in VOT means. Further, estimates that are more formal, whether or not the group means are statistically different, are usually done using a **t-test** (Akhutina et al., 1999; Ionin and Wexler, 2002) and various types of **ANOVA** tests (e.g., Riazantseva, 2001;Winawer et al., 2007; Pavlenko and Driagina, 2007).

2.3.3. Multivariate modeling

Non-quantitative research studies have suggested that there is typically more than one factor affecting the behavior of a given linguistic unit. To access the relative impact of several factors and their possible interaction, two main approaches are mostly exploited, regression and classification. In both cases, the output (the values of a response variable, e.g., the levels of a grammatical category, mean reaction time, etc.) is modeled on a set of predictors (such as the values of other grammatical categories, the choice of lexeme, sentence length, and so on).

In corpus studies, categorical variables are common as both predictors and responses. The **binary logistic regression** model presented in (Sokolova et al., 2012; see Section 2.3.1 above) may obtain two response variables, such as two alternating constructions of the verb *gruzit'* 'load,' and obtain several input variables, such as verb prefix, finiteness, ellipsis of a participant. The categorical predictors are decomposed into a vector consisting of 1s and 0s, depending on whether a given value is in the input. The model assigns weight to each element in the vector according to the strength of its influence on the probability of the output. The resulting probability estimate is a sum of weighted variables transformed with logistic function so that it fits in the interval between 0 and 1 (if probability is less than 0.5, then the default construction is predicted; if it is more than 0.5, then the alternative construction is predicted). Logistic regression analysis of the Russian data shows that the choice of construction is not random, with a statistically significant relationship between the construction and all three predictors. In addition, analysis reveals interaction between the prefix and the participle form, which means that their use is correlated. Overall, the model with the optimal weights yields an accuracy of 89 percent, which means that only 11 percent of the responses have been assigned the wrong class.

If there are more than two possible outcomes from which the model has to choose, then another form of regression analysis is needed. In Divjak and Arppe (2013), **polytomous** (multinomial) **logistic regression** is used to predict the choice for one of the synonyms *pytat'sja*, *starat'sja*, *silit'sja*, *probovat'*, *poryvat'sja*, and *norovit'* (all roughly meaning 'to try') in a sample of contexts. Arppe's model is built following the one-vs-rest approach, in which separate binary classifiers are trained to predict the odds of each outcome against all other classes. As a result, different sets of predictors are revealed to be either for or against each synonym. As an example, *starat'sja* prefers the gerund form; it governs the imperfective infinitive, which most often semantically indicates high control; the subject of *starat'sja* refers to a human being, and the sentence is used as a declarative. In contrast, the verb is rated low when in the past tense or when a dependent infinitive refers to motion. However, the drawback of the polytomous classifiers is that their accuracy is not very high (cf. 52 percent reported by Divjak and Arppe, 2013, p. 42); thus, in many cases, no particular choice is strongly preferred.

Baayen et al. (2013) examined a **conditional inference tree model** for the choice between two alternating Russian verb constructions, theme-object and goal-object. This non-parametric classification approach evaluates all possible splits provided by input variables and selects the best split ("decision") to partition the data into two subsets that are the most homogeneous with respect to the outcome. The data are then partitioned recursively to form a decision tree. To make the model more robust vis-à-vis the characteristic properties of a particular corpus sample not intrinsic to the population, a large number of bootstrap samples[2] are constructed, and the model makes its ensemble prediction by voting on the results of individual decision trees. Baayen et al. (2013) obtained an accuracy of 89 percent, with the main split and many secondary splits associated with the choice of prefix and two other intermediate splits induced by finiteness and reduced construction, respectively. Thus, in this case, both regression and classification trees models provide converging evidence for the association between the choice of syntactic construction and three other explanatory variables. At the same time, the models can partly diverge in assessing the relative strength of predictors, such as factor significance in regression analysis and variable importance in the classification trees method (see further Levshina, 2015, pp. 266, 292). Classification trees are applicable even when there are more than two possible outcomes. For example, Endresen and Janda (2013) present a classification tree model for acceptability judgments of nonce words with the prefixes *o*- and *u*-, where five scores (A–E) are treated as categorical response data.

Yet another type of classifier, **naive discriminative learning** models, has recently been developed as an attempt to bring a more cognitively plausible perspective to multivariate statistics. Baayen et al. (2013) assessed the performance of this approach on several datasets for Russian verbs, including one in which the choice between the prefixes *pere*- 'over, trans-, re-' and *pre*- 'over, across, very' is predicted, given the aspectual relationship between verbs with and without a prefix, a possible shift in transitivity, a possible prefix stacking effect, and the lexical group to which the verbs belong. Baayen's model builds up a two-layer network in which the choice of response variable is driven exclusively by the distributional properties of input variables. In each learning event, certain equations are applied to update the weights on the links from the sets of cues to possible outcomes. The model is designed to strengthen the weight on outcomes that are rarely attested, i.e., while scanning data in the sample. At the same time, it places less weight on common outcomes, since the weights on the links from frequent cues to less frequent outcomes are unlearned (weakened) each time they are not seen together. In the case of two Russian constructions presented in Baayen et al. (2013), the lexical group was shown to be the most important predictor, with aspectual relationship, prefix stacking, and shift in transitivity being less important. As with any other network models, naive discriminative learning requires a large amount of annotated data (a ten-fold cross-validation is used to avoid over-learning). Since the annotated datasets usually available in quantitative research hardly attain such a size, it is not surprising that the naive discriminative learning model loses in accuracy to the logistic regression and classification tree models (e.g., an accuracy of 84 percent, 96 percent, and 96 percent, respectively, on data for the choice between *pere*- and *pre*-).

Mixed-effect modeling is particularly suitable for repeated measures, which may occur both in experimental and in corpus-based studies. In experimental linguistics, this is a common approach for analyzing repeated stimuli and multiple answers by each speaker. As a result, the response variable may be influenced by speakers' individual preferences and performance (e.g., in reaction time) or by certain properties of the stimuli. Since subjects and stimuli in an experiment are usually randomly sampled from a much larger population of speakers and stimuli, it is important to distinguish between variability related to subjects and stimuli (random effects) and variability driven by main linguistic predictors (fixed effects). In quantitative corpus research, words and text sources can be modeled as random variables.

Janda et al. (2010) is an example of the logistic mixed-effect modeling of alternating two Russian verb forms, e.g.: ‘WAVE.3.SG’: *mahaet and mašet.* In their research, three predictors (grammatical form, place of articulation of the consonant preceding the suffix, and the log-transformed frequency of the verb) serve as the fixed effects; by-verb random contrasts for grammatical form are treated as a random effect to predict the alternation. Janda et al. validate the model used with two others: a logistic model without any random effect assumptions and a model with a random effect for individual verbs. Akaike’s information criterion (AIC), which measures goodness of fit, is reported to be the worst (4,729) for the ordinary logistic regression model and still unsatisfactorily high for the random effect modeling (1,395), though it registers 522 when the complex random variance (individual verb preferences depending upon which paradigm slot is considered) is taken into account.

To conclude, while quantitative analysis is a growing field of research in Russian linguistics, there are many topics in grammar and lexicon that have yet to be covered. The objective of this volume is to broaden the range of issues under consideration and to provide more examples of quantitative approaches to the Russian language. The volume is comprised of nine articles and, together with this Introduction, represents state-of-art research in Russian quantitative linguistics.

4. The contributions

The research presented here is organized into four parts around the following topics:

- Part I: Introductory Chapters
- Part II: Topics in Semantics
- Part III: Topics in the Lexicon-Grammar Interface
- Part IV: Topics in Language Acquisition

Among the introductory articles, **Maria Khokhlova** reviews the main corpus resources for study of the Russian language. She discusses a topic currently of broad interest in modern corpus linguistics, namely, how linguistic phenomena are distributed across corpora of different sizes. She begins with an overview of the large Russian corpora, which serves as detailed introduction to the main resources available to researchers. Khokhlova then looks at samples of 18.28 billion, 1.25 billion, and 187.97 million tokens in order to examine token collocability by exploring quantitative properties of ten grammatical relations. The results obtained for high- and low-frequency Russian nouns were compared with data published in *A Frequency Dictionary of Modern Russian* (Lyashevskaya and Sharoff, 2009). The analysis shows that the nouns listed there differ from those archived in the corpora.

Part II, **Topics in Semantics**, opens with a chapter by **Olga Lyashevskaya**, **Maria Ovsjannikova**, **Nina Szymor**, and **Dagmar Divjak**. It reports on the Russian part of a larger survey of Slavic modal words and investigates the structurally diverse domain of modality. The article elucidates the role of formal and semantic context of modal words in a new way. The availability of large corpus data paves the way for study of the empirical reliability of existing classifications originally proposed by philosophers. An important property of the modal words is that they are largely ambiguous, developing new modal meanings both diachronically and from the synchronic point of view.

A study conducted by **Anastasiya Lopukhina**, **Konstantin Lopukhin**, and **Grigory Nosyrev** begins with a well-known observation by G.K. Zipf, namely, that there is a strong correlation between word frequency and polysemy. Even though WordNet and SemCor contain information about sense frequency in English, word sense frequency distribution is a neglected area in Russian linguistics. To fill this lacuna, the authors developed and evaluated a model based on semantic vectors. The model was first trained unsupervised on large corpora and then supplied with contexts and collocations from *Aktivnyj slovar' russkogo jazyka* (*the Active Dictionary of the Russian Language*; Apresjan, 2014). The frequency estimation error of the model is between 11–15 percent, depending on the corpora used. Word sense frequency distributions for 440 nouns are available online for further consultation.

The third contribution in this part, by **Andrey Kutuzov** and **Elizaveta Kuzmenko**, is based on the diachronic investigation of lexical data. It deals with using distributed neural embedding models to observe lexical changes in the Russian language on a large scale. The case presented employs models trained on three sub-corpora of the RNC: texts produced before Soviet times (before 1917), during Soviet times (1917–1990), and after the fall of the USSR (after 1991). By focusing on nouns and adjectives, the authors calculated the most semantically similar words and then the changes that have taken place in the meanings of these words over time. Several methods for calculating the overlap of semantic associates are proposed and evaluated; computing Kendall's τ coefficient proved to be the best method, both when tested on artificially generated data and on a manually compiled gold standard. As a result, a list of several thousand nouns and adjectives that have undergone semantic shifts, either after 1917 or after 1991, is available online.

Part III begins with an article by **Alexander Piperski**, who turns our attention to the **Lexicon-Grammar Interface**. Piperski deals with Russian biaspectual verbs, which can be used to convey both perfective and imperfective meaning. The author proposes three quantitative methods for determining the status of a biaspectual verb: by estimating the relative frequencies of its perfective and imperfective gerunds, by classifying its grammatical profile (i.e., the frequencies of major classes of Tense/Aspect/Mood forms) using the k Nearest Neighbors algorithm, and by conducting an experiment on the perception of the inherent aspect of biaspectual verb forms.

Collocations is the topic of the article by **Lidia Pivovarova**, **Daria Kormacheva**, and **Mikhail Kopotev**. This research focuses on empirical collocations, and its main goal is to examine closely the methods for empirical collocation extractions that are widely used in corpus-based studies, sometimes without proven efficiency. The research indicates that a t-score gives the best ranking in two evaluations (one with a dictionary and another with native speakers' responses), with log-likelihood and Dice not far behind. In general, the evaluations, although each has its own limitations, lead to similar results, which can be taken into account in future research.

Anastasia Bonch-Osmolovskaya sets out to demonstrate how quantitative corpus methods used in linguistics research may help to rank different realizations of the same phenomenon: the use of dative subjects in predicative and adjective constructions. The core idea of the research is to study the distribution of dative subject constructions with predicative and adjective forms that potentially can be used in such constructions, i.e., the tendency of the construction to be used in explication or omitting the dative subject. While usually, the predicates are classified on the basis of whether they can potentially be used with a dative subject, the author studied the trends for explicit use of the dative (or prepositional beneficiary arguments) among the "dative subject predicates." She shows that the frequency rates of the real use of dative subjects can be very different with different predicates. Finally, data from the eighteenth and twenty-first centuries are compared and hierarchical clustering used to reveal diachronic trends.

The fourth and final part, **Language Acquisition**, contains two articles. **Alexei Korneev** and **Ekaterina Protassova** investigate the literal language proficiencies of primary school children. Finnish and Russian are typologically different and use different script systems (Roman vs. Cyrillic) and handwriting principles (printing vs. cursive). Their study attempted to measure written language proficiency in the first stage of literacy, after alphabetization has been mastered. A tablet-based system of handwriting assessment allowed the researchers to record and precisely score the process of handwriting. With an rmANOVA test used as a basic statistical tool, the researchers came to the conclusion that the language of the environment might support language skills, but training in a different language and in a different script supports the quality of writing.

The final contribution, by **Robyn Orfitelli** and **Maria Polinsky**, examines the application of the grammaticality judgment task (GJT), which has been widely used to elicit monolingual speakers' metalinguistic knowledge of their language. When applied to non-native populations, the metalinguistic awareness that the GJT requires may lead speakers to *perform* differently from native speakers, irrespective of their offline *comprehension* of the structure in question. The article discusses data from two new studies comparing the performance of Russian heritage language speakers on GJTs and comprehension tasks. It suggests specific limits on the use and interpretation of acceptability tasks and proposes alternative testing measures that avoid the GJT's excessive demands on metalinguistic decision-making.

To make the reading of this volume easier, Table 1.4 summarizes the inventory of sources and quantitative methods used here.

Table 1.4 Statistical measures used in the articles included in the present volume.

Article	*Data sources (incl. corpora)*	*Quantitative methods*
Khokhlova: Big data and word frequency: measuring the consistency of Russian corpora	ruWac, ruTenTen, (RNC)	Log-likelihood score, Spearman's correlation coefficient
Lyashevskaya et al.: Looking for contextual cues to differentiating modal meanings: a corpus-based study	RNC	multiple correspondence analysis, shaded mosaic plots (incl. χ2 test), polytomous regression modeling, classification and regression trees
Lopukhina et al.: Automated word sense frequency estimation for Russian nouns	ruTenTen, RNC	Distributional semantic models (skip-gram word2vec with negative sampling), spherical k-means clustering; mapping between context clusters and dictionary senses, cluster-map, sense-vec
Kutuzov and Kuzmenko: Two centuries in two thousand words: neural embedding models in detecting diachronic lexical changes	RNC	Distributional semantic models (Continuous Bag of Words, CBOW), nearest neighbors sets comparison metrics: Jaccard similarity index, Kendall's, τ Normalized Discounted Cumulative Gain, PageRank, Relative Neighborhood Graphs
Piperski: Between imperfective and perfective: quantitative approaches to the study of Russian biaspectual verbs	RNC, native speakers' judgments	χ^2 test, Mann – Whitney U test, One-Rule classifier (Holte, 1993). k Nearest Neighbors classifier
Pivovarova et al.: Evaluation of collocation extraction methods for the Russian language	RNC, I-Ru; lists from A Russian-English Collocation Dictionary, native speakers' judgments	Five collocation metrics: t-score, Log-likelihood score, pointwise MI, Dice score, weighted frequency ratio; inter-annotator agreements: Fleiss Kappa, Krippendorf's Alpha
Bonch-Osmolovskaya: Russian predicative constructions with dative subject: a quantitative analysis	RNC	Hierarchical clustering
Korneev and Protassova: Measuring bilingual literacy: challenges of writing in two languages	experimentally elucidated data	t-test; Pearson product-moment correlation coefficient (Pearson's r); repeated measures ANOVA (rmANOVA); Tukey's Multiple Comparison test
Orfitelli and Polinsky: When performance masquerades as comprehension: grammaticality judgments in experiments with non-native speakers	heritage and native speakers' judgments and experimentally elucidated data	t-test, mixed-effect models

5. Internet resources

1. RNC: Russian National Corpus (accessed June 1, 2016).
   http://ruscorpora.ru
2. ruWac: Russian Internet Corpus (accessed June 1, 2016).
   http://corpus.leeds.ac.uk/ruscorpora.html
3. ruTenTen: Russian web corpus (accessed June 1, 2016).
   www.sketchengine.co.uk/rutenten-corpus
4. Araneum Russicum: Russian Web corpora (accessed June 1, 2016).
   http://sketch.juls.savba.sk/aranea_about/_russicum.html
   (A larger Russicum Maximum corpus is available upon separate request).
5. Integrum database (accessed June 1, 2016).
   www.integrumworld.com/services.html
6. GICR: The General Internet-Corpus of Russian (accessed June 1, 2016).
   www.webcorpora.ru/en
   Some parts of the corpus are accessible with Serge Sharoff's search engine at:
   http://corpus.leeds.ac.uk/ruscorpora.html
7. Russian CHILDES (accessed June 1, 2016).
   http://childes.psy.cmu.edu/browser/index.php?url=Slavic/Russian
8. RuLeC/RLC: Russian Learner Corpus (accessed June 1, 2016).
   http://web-corpora.net/RussianLearnerCorpus/search
9. CoRST: Corpus of Russian Student Texts. (accessed June 1, 2016).
   http://web-corpora.net/learner_corpus
10. RusLTC: the Russian Learner Translator Corpus. (accessed June 1, 2016).
    http://rus-ltc.org
11. Database on Russian Folk Dialects (accessed June 1, 2016).
    www.ruslang.ru/agens.php?id=krylov_dialect
12. The Ustya River Basin corpus (accessed June 1, 2016).
    www.slavist.de/Pushkino
13. Digital Library of Russian Folk Dialects (accessed June 1, 2016).
    http://dialekt.rx5.ru
14. MURCO (accessed June 1, 2016).
    http://ruscorpora.ru/search-murco.html
15. SynTagRus (accessed June 1, 2016).
    http://ruscorpora.ru/search-syntax.html
16. Russian Corpora in Tübingen (incl. Uppsala corpus, Corpus of Russian interviews, other corpora; accessed June 1, 2016).
    www.lingexp.uni-tuebingen.de/sfb441/b1/korpora.html
17. HANCO: the Helsinki Annotated Corpus (accessed June 1, 2016).
    www.ling.helsinki.fi/projects/hanco
18. MSU corpus of Russian newspapers of the end of the 20th century (accessed June 1, 2016).
    www.philol.msu.ru/~lex/corpus
19. AOT corpus (based on lib.ru data, accessed June 1, 2016).
    www.aot.ru/search1.html

20. RSTB: Russian Syntactic TreeBank (accessed June 1, 2016).
    http://otipl.philol.msu.ru/~soiza/testsynt/res00/duo.php
21. RUS-Treebank (accessed June 1, 2016).
    http://corpus-i.compling.net/res01/rtb.php
22. UD-Russian (UD-Russian-Google and UD-Russian-SynTagRus converted into the Universal Dependencies format; accessed June 1, 2016).
    http://lindat.mff.cuni.cz/services/pmltq/#!/treebanks
23. RuCor – Russian Coreference corpus (accessed June 1, 2016).
    http://ant0.maimbava.net
24. OpenCorpora named entities collection (accessed June 1, 2016).
    http://opencorpora.org/ner.php
25. PSI RAS collections tagged for named entities and events (accessed June 1, 2016).
    http://ai-center.botik.ru/Airec/index.php/ru/resources/collections
26. Corpus of Russian Twitter (accessed June 1, 2016).
    http://study.mokoron.com
27. RNC n-grams (accessed June 1, 2016).
    http://ruscorpora.ru/corpora-freq.html.
    Online service: http://ruscorpora.ru/search-ngrams_2.html
    n-grams distribution over time: http://ruscorpora.ru/ngram.html
28. RNC Sketches (accessed June 1, 2016).
    http://ling.go.mail.ru/synt
29. CoCoCo: Collocations, Colligations, and Corpora. (accessed June 1, 2016).
    http://cococo.cs.helsinki.fi
30. RusVectōrēs: distributional semantic models for Russian (accessed June 1, 2016).
    http://ling.go.mail.ru/dsm/en
31. The ORD corpus. (accessed June 1, 2016)
    http://model.org.spbu.ru
32. A Russian Paraphrase Corpus. (accessed June 1, 2016)
    www.paraphraser.ru

Notes

[1] This approach was criticized by Kuznetsova (2013), who argues that multi-level comparison may generate noise, thus obscuring important differences. According to Kuznetsova, the main risk is “multicollinearity” when “one of the variables in a multilevel profile is by necessity correlated to another variable within that profile” (Kuznetsova, 2013, p. 16). See, however, (Milin et al., forthcoming).

[2] A bootstrap sample is a sample of the same size as the original. In a bootstrap sample, examples are drawn randomly from the original sample with replacement so that some items were sampled twice or more, while others were missing.

[3] The sources listed here refer to publications in English whenever possible, whether or not an extended version in Russian is available. The foremost English publications are informative and/or contain further references.

References[3]

Akhutina, Tatiana, Andrei Kurgansky, Maria Polinsky, and Elizabeth Bates. "Processing of grammatical gender in a three-gender system: Experimental evidence from Russian." *Journal of Psycholinguistic Research* 28:6 (1999): 695–713.

Alsufieva (Yatsenko), Anna A., Olesya V. Kisselev, and Sandra G. Freels. "Results 2012: Using flagship data to develop a Russian learner corpus of academic writing." *Russian Language Journal* 62 (2012): 79–105.

Apresjan, Juri. *Systematic Lexicography.* Translated by Kevin Windle. Oxford: Oxford University Press, 2000.

Apresjan, Juri (ed.). *Novyj objasnitel'nyj slovar' sinonimov russkogo jazyka* [The New Explanatory Dictionary of Russian Synonyms]. 2nd edition. Moscow, Wien: Languages of Slavic culture, 2004 [Wiener Slawistischer Almanach. Sonderband 60].

Apresjan Juri (ed). *Aktivnyj slovar' russkogo jazyka* [The Active Dictionary of the Russian Language]. Vol. 1. Moscow: Languages of Slavic Culture, 2014.

Asinovsky, Alexander, Natalia Bogdanova, Marina Rusakova, Anastassia Ryko, Svetlana Stepanova, and Tatiana Sherstinova. "The ORD speech corpus of Russian everyday communication 'One speaker's day': Creation principles and annotation." *Proceedings of TSD 2009*. *LNCS 5729*. Heidelberg: Springer (2009): 250–257.

Baayen, R. Harald. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge: Cambridge University Press, 2008.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova, and Tore Nesset. "Making choices in Russian: Pros and cons of statistical methods for rival forms." *Russian Linguistics* 37:3 (2013): 253–291.

Belikov, Vladimir, Alexander Piperski, Vladimir Selegey, and Serge Sharoff. "Big and diverse is beautiful: A large corpus of Russian to study linguistic variation." *Proceedings of the 8th Web as Corpus Workshop* (*WAC-8*)/*International Conference on Corpus Linguistics*. Lancaster: Lancaster University, July (2013).

Benko, Vladimir. "Aranea: Yet another family of (comparable) web corpora." In *Text. Speech and Dialogue*. *17th International Conference*. *TSD 2014*, edited by Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala. Brno, Czech Republic, September 8–12, 2014. Springer International Publishing (2014): 257–264.

Berdichevskii, Alexander. "E-mail vs. chat: The influence of the communication channel on the language." *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference* "*Dialogue*". Moscow: RGGU, May 25–29 (2011): 89–97.

Bergsma, Wicher. "A bias correction for Cramér's V and Tschuprow's T." *Journal of the Korean Statistical Society* 42 (2013): 323–328.

Boguslavsky, Igor, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Kreidlin, and Nadežda Frid. "Dependency treebank for Russian: Concept, tools, types of information." *Proceedings of the 18th International Conference on Computational Linguistics* (*COLING*). Saarbrücken: Universität des Saarlandes (2000): 987–991.

Boguslavsky, Igor, Leonid Iomdin, Svetlana P. Timoshenko, and Tatiana I. Frolova. "Development of the Russian tagged corpus with lexical and functional annotation". *Metalanguage and Encoding Scheme Design for Digital Lexicography*. *Proceedings of MONDILEX Third Open Workshop*. Bratislava: Štúr Institute of Linguistics (2009): 83–90.

Dickey, Stephen M., and Laura A. Janda. "*Xoxotnul*, *sxitril*: The relationship between semelfactives formed with *-nu-* and *-s-* in Russian." *Russian Linguistics* 33:3 (2009): 229–248.

Divjak, Dagmar. *Degrees of Verb Integration: Conceptualizing and Categorizing Events in Russian.* PhD dissertation, KU Leuven, Belgium, 2004.

Divjak, Dagmar. *Structuring the Lexicon: A Clustered Model for Near-Synonymy*. Berlin: Mouton de Gruyter, 2010.

Divjak, Dagmar, and Antti Arppe. "Extracting prototypes from exemplars. What can corpus data tell us about concept representation?" *Cognitive Linguistics* 24:2 (2013): 221–274.

Divjak, Dagmar, and Stefan Th. Gries. "Ways of trying in Russian. Clustering behavioral profiles." *Journal of Corpus Linguistics and Linguistic Theory* 2:1 (2006): 23–60.

Divjak, Dagmar, and Stefan Th. Gries. "Clusters in the mind? Converging evidence from near synonymy in Russian." *The Mental Lexicon* 3:2 (2008): 188–213.

Divjak, Dagmar, and Stefan Th. Gries (eds.). *Frequency Effects in Language Representation*. Berlin/Boston: Walter de Gruyter, 2012.

Droganova, Kira, and Olga Lyashevskaya. *From Tagset to Tagset: Influence of Morphological and Syntactic Input on Syntactic Parsing*. Unpublished MS, Moscow. 2017.

Endresen, Anna, and Laura A. Janda. "What is a possible word? Evidence from Russian factitive verbs." *Talk presented at the 12th International Cognitive Linguistics Conference (ICLC)*, June 23–28, 2013 at the University of Alberta in Edmonton, Alberta, Canada.

Endresen, Anna, Laura A. Janda, Julia Kuznetsova, Olga Lyashevskaya, Anastasia Makarova, Tore Nesset, and Svetlana Sokolova. "Russian 'purely aspectual' prefixes: Not so 'empty' after all?" *Scando-Slavica* 58:2 (2012): 229–289.

Erjavec, Tomaž, Ivan Deržanski, Dagmar Divjak, Anna Feldman, Mikhail Kopotev, Natalia Kotsyba, Cvetana Krstev, Aleksandar Petrovski, Behrang QasemiZadeh, Adam Radziszewski, Serge Sharoff, Paul Sokolovsky, Duško Vitas and Katerina Zdravkova. *MULTEXT-East Non-Commercial Lexicons 4.0.* Slovenian Language Resource Repository CLARIN.SI (2010). Available at http://hdl.handle.net/11356/1042 (accessed June 1, 2016).

Facchinetti, Roberta (ed.). *Corpus Linguistics 25 Years On*. Amsterdam: Rodopi, 2007.

Field, Andy. *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock 'N' Roll*. London: Sage Publications, 2013.

Gorokhova, Svetlana. "Some factors that determine the outcome of lexical competition in language production: A corpus-based analysis of Russian speech errors." In *Yearbook of the German Cognitive Linguistics Association*. Vol. 1, edited by A. Stefanowitsch and Doris Schoenefeld. 25–38. Berlin/Boston: Mouton de Gruyter, 2013.

Gries, Stefan Th. *Statistics for Linguistics With R: A Practical Introduction*. Berlin/Boston: Walter de Gruyter, 2013.

Gries, Stefan Th. "Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff." *Corpus linguistics and linguistic theory* 1:2 (2005): 277–294.

Grishina, Elena. "Multimodal Russian Corpus (MURCO): First steps." *Proceedings of LREC*. Valletta, Malta, May 19–21 (2010): 2953–2960.

Grishina, Elena. *Russkaja žestikuljacija s lingvisticheskoj tochki zrenija (korpusnyje issledovanija)* [Russian Gestures From the Linguistic Perspective: Corpus Studies]. Moscow: Languages of Slavic Culture [Forthcoming].

Hanks, Patrick. "Contextual dependency and lexical sets." *International Journal of Corpus Linguistics* 1:1 (1996): 75–98.

Holte, Robert C. "Very simple classification rules perform well on most commonly used datasets." *Machine Learning* 11:1 (1993): 63–90. doi:10.1023/A:1022631118932

Ionin, Tania, and Kenneth Wexler. "Why is 'is' easier than '-s'?: Acquisition of tense/agreement morphology by child second language learners of English." *Second Language Research* 18:2 (2002): 95–136.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. "The TenTen corpus family". *Proceedings of the International Conference on Corpus Linguistics*. Lancaster, July (2013): 125–127.

Janda, Laura A. (ed.). *Cognitive Linguistics – The Quantitative Turn: The Essential Reader*. Walter de Gruyter, 2013.

Janda, Laura A., and Olga Lyashevskaya. "Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian." *Cognitive Linguistics* 22:4 (2011): 719–763.

Janda, Laura A., and Olga Lyashevskaya. "Semantic profiles of five Russian prefixes: *po-, s-, za-, na-, pro-*." *Journal of Slavic Linguistics* 21:2 (2013): 211–258.

Janda, Laura A., Tore Nesset, and R. Harald Baayen. "Capturing correlational structure in Russian paradigms: A case study in logistic mixed-effects modeling." *Corpus Linguistics and Linguistic Theory* 6:1 (2010): 29–48.

Khudyakova, Maria, Mira Bergelson, Yulia Akinina, Ekaterina Iskra, Svetlana Toldova, and Olga Dragoy. "Russian CliPS: A corpus of narratives by brain-damaged individuals." *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data From People With Various Forms of Cognitive/Psychiatric Impairments* (RaPID-2016), May 23 (2016): 22–26.

Kizach, Johannes. "Evidence for weight effects in Russian." *Russian Linguistics* 36:3 (2012): 251–270.

Klyachko, Elena, Timofey Arkhangelskiy, Olesya Kisselev, and Ekaterina Rakhilina. "Automatic error detection in Russian learner language." *Proceedings of Corpus Analysis With Noise in the Signal* (*CANS 2013*). Available at http://ucrel.lancs.ac.uk/cans2013/abstracts/Klyachko%20et%20al.pdf (accessed June 1, 2016).

Kopotev, Mikhail. "K postroeniju chastotnoj grammatiki russkogo iazyka: padežnaja sistema po korpusnym dannym" [Towards the frequency grammar of Russian: The case system based on the corpus data]. In *Instrumentarij rusistiki: korpusnyje podhody* [=Slavica Helsingiensia, 34], edited by A. Mustajoki, M. Kopotev, L Birjulin, E. Protassova, 136–151. Helsinki: Yliopistopaino, 2008.

Kopotev, Mikhail, and Arto Mustajoki. "Printsipy sozdaniya Helsingskogo annotirovannogo korpusa russkikh tekstov (HANCO) v seti Internet" [Guiding principles behind the Helsinki annotated corpus HANCO]. *Nauchno-tekhnicheskaya Informatsiya*, Series 2, 22:6 (2003): 33–37.

Kopotev, Mikhail, and Arto Mustajoki. "Sovremennaja korpusnaja rusistika" [Modern Russian corpus linguistics]. In *Instrumentarij rusistiki: korpusnyje podhody* [=Slavica Helsingiensia, 34], edited by A. Mustajoki, M. Kopotev, L Bityuilin, E. Protassova, 5–24. Helsinki: Yliopistopaino, 2008.

Kopotev, Mikhail, Llorenc Escoter, Daria Kormacheva, Matthew Pierce, Lidia Pivovarova, and Roman Yangarber. "CoCoCo: Online Extraction of Russian Multiword Expressions." In *The 5th Workshop on Balto-Slavic Natural Language Processing,* 43–45. Sofia: INCOMA Ltd, 2015.

Kotov, A., and E. Budyanskaya. "The Russian emotional corpus: Communication in natural emotional situations." *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue*. Moscow, 11:18 (2012): 296–306.

Krylov, Sergei, and Alexandra Ter-Avanesova. "Elektronnye bazy danykh po russkim narodnym govoram" [Digital databases of Russian folk dialects], Moscow (2012). Available at www.ruslang.ru/agens.php?id=krylov_dialect (accessed June 1, 2016).

Kulsharipova, R. E., and T. I. Ibragimov. "Elektronnaja biblioteka russkikh narodnykh govorov Kazanskogo universiteta: vozmožnosti primeneniya, informacionnyj potencial" [Kazan University digital library of Russian folk dialects: Potential for use and awareness raising]. *Meždunarodnyj žurnal ehksperimental'nogo obrazovanija* 5 (2013): 95–96.

Kunilovskaya, Maria, and Andrey Kutuzov. "Russian learner translator corpus: Design, research potential and applications." In *Text, Speech and Dialogue*, edited by P. Sojka, A. Horák, I. Kopeček, K. Pala, 315–323. Springer International Publishing, 2014.

Kuznetsov, Ilya. *Avtomaticheskaya razmetka semanticheskikh rolej v russkom jazyke* [Semantic Role Labeling for Russian]. PhD thesis, Moscow State University, Moscow, 2016.

Kuznetsova, Julia. *Linguistic Profiles: Going From Form to Meaning via Statistics*. Mouton de Gruyter, 2013.

Levshina, Natalia. *How to Do Linguistics With R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins, 2015.

Lönngren, Lennart. *Častotnyj slovar' sovremennogo russkogo jazyka* [The Frequency Dictionary of Modern Russian] [=Studia Slavica Upsaliensia 32]. Uppsala: Uppsala University, 1993.

Lyashevskaya, Olga. "Chastotnyj leksiko-grammaticheskij slovar': prospekt proekta" [Lexico-grammatical frequency dictionary: A preliminary design]. *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue*. Moscow, 12:19 (2013): 478–489.

Lukashevich, N. Y., Klyshinsky E. S., Kobozeva I. M. "Lexical research in Russian: Are modern corpora flexible enough?" *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue*. Vol. 1, Moscow, 15:22 (2016): 427–440.

Lyashevskaya, Olga, and Egor Kashkin. "FrameBank: A Database of Russian Lexical Constructions." *Proceedings of AIST 2015*, CCIS. 542:1–11 (2015): 350–360. Heidelberg: Springer.

Lyashevskaya, Olga, and Serge Sharoff. *Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Nacional'nogo korpusa russkogo jazyka* [Frequency Dictionary of Modern Russian (based on the RNC data)]. Moscow: Azbukovnik, 2009.

Mikhailov, Mikhail, and Robert Cooper. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*. Routledge, 2016.

Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević, and R. Harald Baayen. "Towards cognitively plausible data science in language research." *Cognitive Linguistics* 27:4 (2016): 507–526.

Mitrenina, Olga. "The corpora of old and middle Russian texts as an advanced tool for exploring an extinguished language." *Scrinium* 10:1 (2014): 455–461.

Mustajoki, Arto. "The Integrum database as a powerful tool in research on contemporary Russian." In *Integrum: tochnye metody i gumanitarnye nauki*, edited by Galina Nikiporets-Takigava, 50–76. Moscow: Letnij sad, 2006.

Nagy, Naomi, Nina Aghdasi, Y. Kang, A. Kochetov, D. Denis, A. Motut, and J. Walker. "Heritage Russian variation and change in Toronto." *Journal of Slavic Linguistics* 20 (2014): 269–286.

Pavlenko, Aneta, and Viktoria Driagina. "Russian emotion vocabulary in American learners' narratives." *The Modern Language Journal* 91:2 (2007): 213–234.

Perry, Fred L., Jr. *Research in applied linguistics: becoming a discerning consumer.* Mahwah, NJ; London: Laurence Erlbaum Associates, 2005.

Podlesskaya, Vera. "A corpus-based study of self-repairs in Russian spoken monologues." *Russian Linguistics* 39:1 (2015): 63–79.

Pronoza, E., E. Yagunova, and A. Pronoza. "Construction of a Russian paraphrase corpus: Unsupervised paraphrase extraction." Information Retrieval. 9th Russian Summer School, RuSSIR 2015, Saint Petersburg, Russia, August 24–28, 2015, Revised Selected Papers, edited by P. Braslavski, I. Markov, P. Pardalos, Y. Volkovich, D.I. Ignatov, S. Koltsov, O. Koltsova. 146–157. Berlin, Heidelberg: Springer, 2016.

Protopopova, E. V., A. A. Bodrova, S. A. Volskaya, I. V. Krylova, A. S. Chuchunkov, S. V. Alexeeva, V. V. Bocharov, and D. V. Granovsky. “Anaphoric annotation and corpus-based anaphora resolution: An experiment.” *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue.* Vol. 1, Moscow, 13:20 (2014): 562–571.

Riazantseva, Anastasia. “Second language proficiency and pausing a study of Russian speakers of English.” *Studies in Second Language Acquisition* 23:4 (2001): 497–526.

Rubtsova Yuliya V., and Yury A. Zagorulko. “An approach to construction and analysis of a corpus of short Russian texts intended to train a sentiment classifier.” *Bulletin of the Novosibirsk Computing Center, Series: Computer Science* 37 (2014): 107–116.

Sharoff, Serge. “Creating general-purpose corpora using automated search engine queries.” In *WaCky! Working Papers on the Web as Corpus*, edited by Marco Baroni and Silvia Bernardini, 63–98. Bologna: Gedit Edizioni, 2006.

Sharoff, Serge, and Joakim Nivre. “The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge.” *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference* “*Dialogue*”. Moscow (2011): 657–670.

Sharoff, Serge, Elena Umanskaya, and James Wilson. *A Frequency Dictionary for Russian: Core Vocabulary for Learners*. Abingdon, UK/New York: Routledge, 2014.

Sharpe, Donald. “Your Chi-Square Test is statistically significant: Now what?” *Practical Assessment, Research & Evaluation* 20:8 (2015): 8.

Shaykevich, Anatoly, Vladislav Andryuschenko, and Natalia Rebetskaya. *Statisticheskij slovar’ jazyka russkoj gazety (1990-e gody)* [A Statistical Dictionary of the Language of Russian Newspapers (1990s)]. Vol. 1. Moscow: Languages of Russian Culture Press, 2008.

Sherstinova, Tatiana. “The structure of the ORD speech corpus of Russian everyday communication.” *Proceedings of TSD 2009. LNCS 5729*. Heidelberg: Springer (2009): 258–265.

Skrelin, Pavel A., Nina B. Volskaya, Daniil Kocharov, Karina Evgrafova, Olga Glotova, and Vera Evdokimova. “A Fully Annotated Corpus of Russian Speech.” *Proceedings of LREC* (2010): 109–112.

Sokolova, Svetlana, Olga Lyashevskaya, and Laura A. Janda. “The locative alternation and the Russian ‘empty’ prefixes: A case study of the verb gruzit’ ‘load’.” In *Frequency Effects in Language: Linguistic Representations*, edited by Dagmar Divjak and Stefan Th. Gries, 51–86. Berlin: Mouton de Gruyter, 2012.

Soloviev, Valery D., and Laura A. Janda. “What constructional profiles reveal about synonymy: A case study of Russian words for SADNESS and HAPPINESS.” *Cognitive Linguistics* 20:2 (2009): 367–393.

Starostin, Anatoly S., Victor V. Bocharov, Svetlana V. Alexeeva, Anastasiya A. Bodrova, Alexander S. Chuchunkov, Stanislav S. Džumaev, Irina V. Efimenko, Dmitry V. Granovsky, Viktor F. Khoroshevsky, Irina V. Krylova, Maria A. Nikolaeva, Igor M. Smurov, Svetlana Y. Toldova. “FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian.” *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference “Dialogue”*. Moscow (2016): 702–720.

Stubbs, Michael. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell, 2001.

Toldova, Svetlana, Yulia Grishina, Alina Ladygina, Maria Vasilyeva, Galina Sim, and Ilya Azerkovich. “Russian coreference corpus.” In *Input a Word, Analyze the World*, edited by Francisco Alonso Almeida, Ivalla Ortega Barrera, Elena Quintana Toledo, and Margarita E. Sanchez Cuervo, 107–124. Cambridge: Cambridge Scholars Publishing, 2016.

Toldova, Svetlana, Olga Lyashevskaya, Anastasia Bonch-Osmolovskaya, and Maxim Ionov. "Evaluation for morphologically rich language: Russian NLP." *Proceedings on the International Conference on Artificial Intelligence* (ICAI'15) (2015): 300–306.

Vinogradova, V. B., Olga V. Kukushkina, Anatoly A. Polikarpov, and Svetlana O. Savchuk. "Kompjuternyj korpus tekstov russkikh gazet kontsa 20-go veka" [The Computer Corpus of Russian Newspapers of the End of the 20th Century]. 2001. Available at www.philol.msu.ru/~lex/corpus/corp_descr.html (accessed June 1, 2017).

Vlasova, Natalia, Natalia Lando, Elena Suleymanova, and Igor Trofimov. "Situations-1000: A tagged corpus for event extraction from texts." *Talk Presented at Computational Linguistics and Intellectual Technologies Dialogue, Moscow, Russia, June 1–4, 2016*. Available at www.dialog-21.ru/media/3458/vlasovanaetal.pdf (accessed June 1, 2016).

Voeikova, Maria. *Russkaia rech v sisteme CHILDES* [Russian Speech in CHILDES]. Saint-Petersburg: Herzen University Press, 1995.

von Waldenfels, Ruprecht, Michail Daniel, and Nina Dobrushina. "Why standard orthography? Building the Ustya River Basin corpus, an online corpus of a Russian dialect." *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference* "*Dialogue*". Vol. 1, Moscow, 13:20 (2014): 720–728.

Winawer, Jonathan, Nathan Witthoft, Michael C. Frank, Lisa Wu, Alex R. Wade, and Lera Boroditsky. "Russian blues reveal effects of language on color discrimination" *Proceedings of the National Academy of Sciences* 104:19 (2007): 7780–7785.

Zakharov, Victor. "Corpora of the Russian language." *Proceedings of Text, Speech, and Dialogue Conference* (*TSD 2014*). *LNCS* 8082 (2014): 1–13.

Zevakhina, Natalia, and Svetlana Dzhakupova. "Corpus of Russian student texts: Design and prospects". *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference* "*Dialogue*". Moscow (2015). Available at www.dialog-21.ru/digests/dialog2015/materials/pdf/ZevakhinaNADzhakupovaSS.pdf (accessed June 1, 2016).