CLIP 1

Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache

Corpus Linguistics and Interdisciplinary Perspectives on Language

Bd./Vol. 1

Marek Konopka / Jacqueline Kubczak
Christian Mair / František Štícha
Ulrich H. Waßner (Hgg.)

Grammatik und Korpora 2009

Dritte Internationale Konferenz

Grammar & Corpora 2009

Third International Conference

Mannheim, 22.-24.09.2009

narr
VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Dischingerweg 5 · D-72070 Tübingen

Gedruckt auf chlorfrei gebleichtem und säurefreiem Werkdruckpapier.

Internet: www.narr.de
E-Mail: info@narr.de

Layout: Tröster, Mannheim
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6648-5

Inhalt / Contents

III. Methodologie korpuslinguistischer Grammatikforschung / Methodologies of corpus-linguistic Grammar Research

IV. Einblicke in die aktuelle Forschung / Insights into current studies

SVETLANA SAVCHUK

The *Russian National Corpus* as a Tool for Research on Grammatical Variability*

Abstract

The paper presents the *Russian National Corpus* (RNC) as a tool for research on grammatical variability. The RNC has all the necessary quantitative and qualitative characteristics to provide an adequate set of examples for various types of linguistic research. Being a representative collection of texts, the RNC reflects Russian language usage in two dimensions: 'horizontally' (in functional varieties) and 'vertically' (from a historical perspective).

Based on the corpus data, the problem of grammatical variation can be divided into three aspects:

1) the setting of the correlation of variants in contemporary language usage,
2) the study of the development that occurred during a certain period,
3) the comparison of these findings with the recommendations found in dictionaries and grammar manuals in order to evaluate the adequacy of these recommendations in real usage.

The capabilities of the RNC for linguistic investigation are demonstrated by means of corpus-based analyses of variants of genitive plural forms of masculine nouns (the names of those belonging to some branches of the armed forces have been chosen as an example).

1. Introduction

The corpus-based approach appears extremely useful and fruitful when researching grammatical variability and correlation between normative recommendations and real language usage. The following three aspects should be discussed in more detail:

1) If the focus is on the synchronic aspect, the co-existing variants should be studied concerning their distribution among different spheres of functioning, as well as social, professional variants of the language, etc.

* This work was supported by the Fundamental Research Program in the Department of History and Philology at the Russian Academy of Sciences (RAS) "Text in sociocultural environment: levels of historical, literary and linguistic interpretation" and by the Fundamental Research Program of the Presidium of the RAS (project "Russian language of the 18th century: corpus-based study of lexical and morphological variability").

2) When dealing with the problem in its diachronic aspect, some changes in usage, appearance and disappearance of the variants should be taken into consideration, as well as the changing of correlation between several variants, increasing and declining trends, and so on.

3) This aspect concerns the evaluation of the different variants with regard to a standard, or so-called 'codified norm' fixed in dictionaries and grammar manuals. A variant can be codified in the literary language or remain uncodified. There is a natural discrepancy: some variants can fluctuate over a long period of time while normative estimations of these variants change continually. Consequently, the evaluation of normative recommendations should also be regarded from synchronic and diachronic points of view, and in relation to real usage.

As far as statistical values are widely accepted as objective indicators of the current distribution of language phenomena, modern large text corpora seem to be a reliable tool for the research of linguistic norms and variation. Naturally, the reliability of this research depends on the parameters of the corpus: its capacity, coverage, and the linguistic information represented in it. The Russian National Corpus has all necessary quantitative and qualitative characteristics to provide an adequate set of examples for various types of linguistic research.

2. The Russian National Corpus: main parameters

A large group of specialists from Moscow, St. Petersburg, Voronezh and other Russian university centers have been creating the Russian National Corpus (RNC) within the program of the Russian Academy of Sciences since 2003. Although the project is still in progress, the corpus is already being used for research and educational purposes.

The RNC meets all the requirements for large contemporary text corpora, such as:

1) Large size
2) Representativeness
3) Linguistic annotation
4) Query tools

1) As far as the size of the corpus is concerned, the RNC contains approx. 170 million tokens at present (as of 2009).

2) The RNC is a representative corpus. It only includes complete texts of different language forms (spoken, written, and electronic) and of different functional spheres: fiction, journalism, memoirs, academic writing, administrative documents, religious texts, poetry, everyday dialogues, TV-programmes, broadcasts, etc.

3) The RNC is an annotated corpus: all texts are supplied with different types of linguistic annotation.

   Metatextual annotation refers to the text as a whole and includes information regarding the author's name, sex, age or date of birth, text characteristics (date of origin, functional sphere, text type, genre, domain), etc.

   Morphological annotation is performed automatically by a parser developed for modern Russian texts and based on Zaliznyak's *Grammatical dictionary of Russian* (1977 / 2003). The morphological information consists of four groups of tags:

   a) Lexeme (the lemma and the part of speech to which it belongs);
   b) Grammatical features of the lexeme (e.g., gender for nouns and transitivity for verbs);
   c) Grammatical features of the word-form (e.g., case for nouns and number for verbs);
   d) Information concerning non-standard forms of the lemma, orthographic variations, etc.

   Semantic annotation is performed automatically by 'Semmarkup', a software program by A. E. Poliakov which uses the semantic dictionary of the corpus. There are three groups of tags assigned to words:

   a) Class ('proper name', 'reflexive pronoun', etc.);
   b) Lexical and semantic features (thematic class of the lexeme, indications of causality or assessment, etc.);
   a) Derivational features ('diminutive', 'adjectival adverb', etc.).

   As a result, most words in a text are tagged with a number of semantic and derivational parameters such as 'person', 'substance', 'space', 'diminutive', 'verbal noun', etc.

   Sociological annotation is only specific to corpora of spoken language. It is assigned to different speakers' utterances and characterizes a word usage with regard to the sex and age of a speaker (if this information is available).

Sociological annotation allows a user to create his / her own sub-corpora by various parameters or their combinations: by a speaker's sex, age, or year of birth (this option is only available for movie transcripts), etc.

Accentological annotation is used in the Accentological Corpus. According to this annotation, each word is supplied with stress marks making it possible to carry out different kinds of search requests and retrieve data concerning stressed or unstressed word-forms in combination with grammatical and semantic features.

4) The corpus is available for all users at the following site: http://ruscorpora.ru. The search system is provided by the 'Yandex' server. Users can create their own subcorpora based on particular metatextual parameters and then run queries for words, grammemes and semantic features in various combinations, receiving contexts as query results.

3. The Russian National Corpus: composition and capabilities in the research of variation

The Russian National Corpus consists of the following subcorpora (as of 2009):

- Corpus of modern written texts (1950-2008): 97.4 million words
- Corpus of spoken language (1930-2008): 8.5 million words
- Corpus of written texts (18th century to the first half of the 20th century): 68 million words of which:
  - 26 million words are texts from the 19th century
  - 40 million words are texts from the first half of the 20th century
  - 2.6 million words are texts from the 18th century
- Poetry corpus: 3.2 million words
- Accentological corpus: 5.3 million words
- Dialect corpus: c. 200 000 tokens
- Parallel aligned corpus: c. 5.3 million words

Being a representative collection of texts, the RNC reflects Russian language usage in two dimensions: 'horizontally' (in functional varieties) and 'vertically' (from a historic perspective). Based on the corpus data, the problem of grammatical variation can be divided into three aspects:

1) the distribution of the correlation of variants in contemporary language usage;
2) the development of variants within a certain period;
3) the comparison of these findings with the recommendations of dictionaries and grammar manuals in order to evaluate the adequacy of these recommendations in real usage.

The capabilities of the RNC in linguistic investigations are demonstrated by two examples of corpus-based variant analyses of one of the 'weak points' of grammatical norm.

4. A corpus-based study of variants of genitive plural forms

4.1 Variants of genitive plural forms of masculine nouns

In modern literary Russian there are three variants of genitive plural masculine endings: *-ov*, *-ej*, zero (-Ø). The principle of selection, which was discovered by Jakobson (1956 / 1984: 135-140), has been adopted by grammarians and is used in grammatical descriptions: if there is a -Ø ending in the nominative singular, there is a non-zero ending in the genitive plural, and vice versa, a non-zero ending in the nominative singular involves a -Ø ending in the genitive plural.

According to this, generic forms with the ending *-ov* are standard for most masculine nouns with stems ending in a hard consonant or [j], and forms with the ending *-ej* are standard for masculine nouns with stem-final soft consonant or *ж*, *ш* (Shvedova (ed.) 1980: 498, Andrews 2001: 34). According to Zaliznyak (1967: 219), 97.3 % of masculine nouns have standard genitive plural forms with non-zero endings (Zaliznyak 1967: 219).

Genitive plural forms of masculine nouns with the -Ø ending are the exceptions to this rule because they have the same ending as nominative singular forms. According to Graudina (1976 / 2000), there are about 200 nouns with the -Ø ending in contemporary written and spoken language, which belong to several semantic groups:

1) Names of people as members of different associations – ethnic, military, political: gen. pl. *грузин* 'Georgians', *бурят* 'Buryats', *румын* 'Romanians'; *гусар* 'hussars', *драгун* 'dragoons', *кадет* 'Cadets'.

2) Names of some paired items: gen. pl. *чулок* 'stockings', *ботинок* 'boots', *брюк* 'trousers'.

3) Names of some measurement units in combination with numerals: gen. pl. *300 грамм* '300 grammes', *40 мегабайт* '40 megabytes', *20 рентген* '20 roentgens'.

4) Names of some fruit and vegetables: gen. pl. *баклажан* 'aubergines', *помидор* 'tomatoes', *гранат* 'pomegranates', *апельсин* 'oranges' (these forms are allowed as variants in colloquial speech).

The genitive plural of some nouns allows either the *-ov* or *-Ø* ending: *грамм-ов* and *грамм-Ø*, *помидор-ов* and *помидор-Ø*, *кадет-ов* and *кадет-Ø*, *гардемарин-ов* and *гардемарин-Ø*. This group is especially interesting for the study of variants because it includes words for which the process of variant competition is still in progress.

Contrary opinions exist on the correlation of variants in this 'weak point' of language norm. According to Markov (1992), genitive *-ov* forms have been gradually displacing *-Ø* forms since the 12th century and the process still is going on. According to another point of view, *-Ø* forms have become more active since the end of the 20th century and for this reason are considered to be the dominant variants in the observed group of nouns (Glovinskaya 2008).

The corpus-based study of the correlation of variants within the mentioned subgroups and hereafter within the whole group of nouns is likely to shed new light on the matter.

4.2 A corpus-based study of *-ov* and *-Ø* variants of genitive plural forms

As an example, the names of ranks belonging to some branches of the armed forces have been chosen because they form a finite list including 14 nouns. The variants of genitive plural forms of all these nouns were examined in written texts dating from four periods: the 18th century, the 19th century, and the first and the second half of the 20th century. Table 1 shows the total number of the relevant word-form and Table 2 demonstrates its frequency (items per million tokens).

Gen.pl. variant	**18th cent.**	**19th cent.**	**20th cent. (1st half)**	**20th cent. (2nd half)**
солдат 'soldier'	75	>1000	>2500	>4000
солдатов	1	5	11	6
партизан 'partisan'	0	7	>300	>350
партизанов	0	23	5	4
рекрут 'recruit'	23	96	5	1
рекрутов	1	26	20	26
*кадет*1 'cadet' (military)	1	40	54	10
*кадетов*1	8	9	13	25
*кадет*2 'Cadet' (party)	0	0	18	3
*кадетов*2	0	0	193	48
гренадер 'grenadier'	10	37	34	9
гренадеров	1	22	33	13
гардемарин 'midshipman'	0	1	13	2
гардемаринов	0	15	10	12
гусар 'hussar'	11	130	50	23
гусаров	1	38	15	10
карабинер 'carabineer'	5	4	0	0
карабинеров	2	6	7	11
драгун 'dragoon'	6	79	38	9
драгунов	0	18	3	1
кирасир 'cuirassier'	0	20	30	9
кирасиров	0	14	3	7
улан 'uhlan'	0	43	26	7
уланов	0	27	5	9
янычар 'janissary'	9	23	7	14
янычаров	1	2	5	0
рейтар 'rider'	0	4	0	7
рейтаров	0	1	2	8

Table 1: Total number of variants of gen.pl. word-forms in different subcorpora

Gen.pl. variant	18th cent.	19th cent.	20th cent. (1st half)	20th cent. (2nd half)
солдат 'soldier'	28.8	>38.5	>62.5	>41.1
солдатов	0.38	0.19	0.28	0.06
партизан 'partisan'	0	0.27	>7.5	>3.6
партизанов	0	0.88	0.13	0.04
рекрут 'recruit'	8.9	3.7	0.13	0.01
рекрутов	0.39	1	0.5	0.27
*кадет*1 'cadet' (military)	0.39	1.5	1.35	0.1
*кадетов*1	3.1	0.35	0.33	0.26
*кадет*2 'Cadet' (party)	0	0	0.45	0.03
*кадетов*2	0	0	4.8	0.49
гренадер 'grenadier'	3.8	1.42	0.85	0.09
гренадеров	0.39	0.85	0.83	0.13
гардемарин 'midshipman'	0	0.39	0.33	0.02
гардемаринов	0	0.58	0.25	0.12
гусар 'hussar'	4.2	5	1.25	0.23
гусаров	0.39	1.46	0.38	0.1
карабинер 'carabineer'	1.9	0.15	0	0
карабинеров	0.77	0.23	0.18	0.1
драгун 'dragoon'	2.3	3.03	0.95	0.09
драгунов	0	0.69	0.08	0.01
кирасир 'cuirassier'	0	0.77	0.75	0.09
кирасиров	0	0.54	0.08	0.07
улан 'uhlan'	0	1.65	0.65	0.07
уланов	0	1.03	0.13	0.09
янычар 'janissary'	4.2	0.88	0.18	0.14
янычаров	0.39	0.08	0.13	0
рейтар 'rider'	0	0.15	0	0.07
рейтаров	0	0.04	0.05	0.08

Table 2: Frequency (items per million tokens) of variants of gen.pl. word-forms in different subcorpora

The words on the above list (except *солдат* and *партизан*) are not frequent in modern texts and mainly belong to the passive vocabulary. The whole group can be subdivided into 3 subgroups according to the relation between *-ov* and *-Ø* variants of the genitive plural.

Subgroup 1 only includes two frequently used words: *солдат* 'soldier' and *партизан* 'partisan'. Each of them has only one codified variant with the *-Ø* ending. The variants with *-ov* (*солдатов*, *партизанов*) are sub-standard and mainly used in fiction for stylistic purposes.

Subgroup 2 includes the words *рекрут* 'recruit', *кадет*1 '(army) cadet', *кадет*2 'Constitutional Democrat, Cadet', *гренадер* 'grenadier', *гардемарин* 'midshipman', *гусар* 'hussar', *карабинер* 'carabineer'. The competition of variants has been continuing during the past three centuries, the process being especially active in the 20th century. The rates for each variant (defined as the ratio of the variants to the total number of genitive plural forms of each noun over different periods of time) are displayed in Figures 1 and 2.

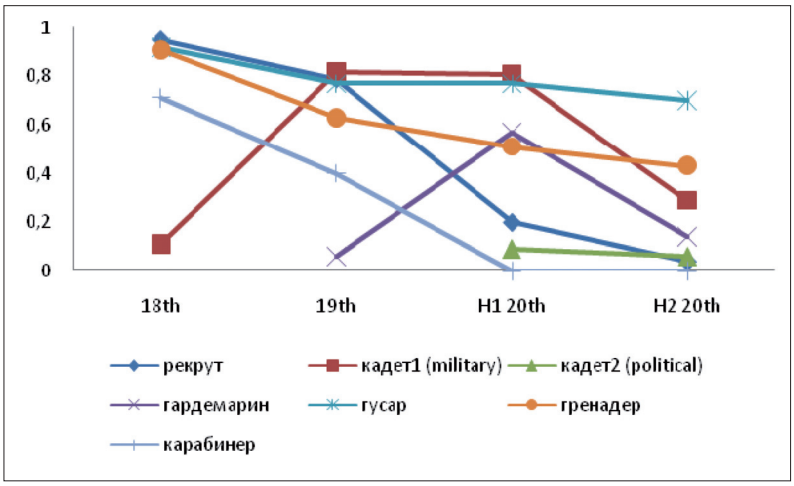

Figure 1: Variant ratio of genitive plural forms with the -Ø ending of masculine nouns in Subgroup 2

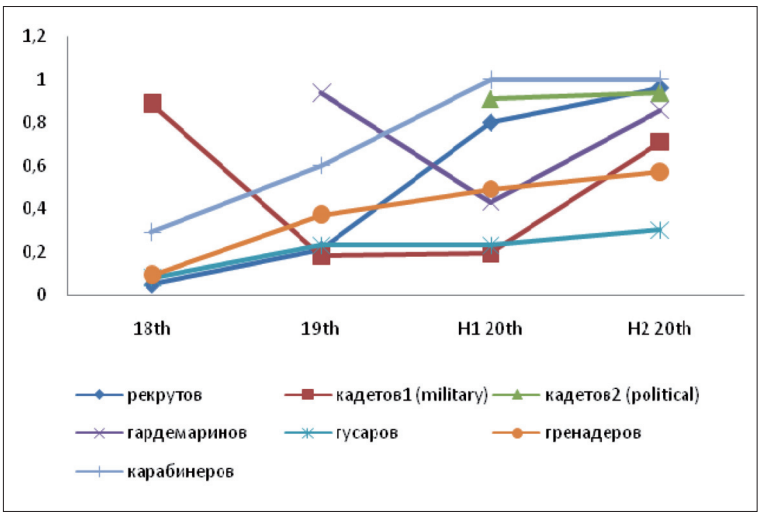

Figure 2: Variant ratio of genitive plural forms with the *-ov* ending of masculine nouns in Subgroup 2

As the diagrams show, the nouns in the second group have one point in common: the proportion of their variants with the -Ø ending dropped towards the end of the 20th century whilst the proportion of *-ov* variants increased.

Subgroup 3 includes the nouns *драгун* 'dragoon', *кирасир* 'cuirassier', *улан* 'uhlan', *рейтар* 'rider', *янычар* 'janissary'.

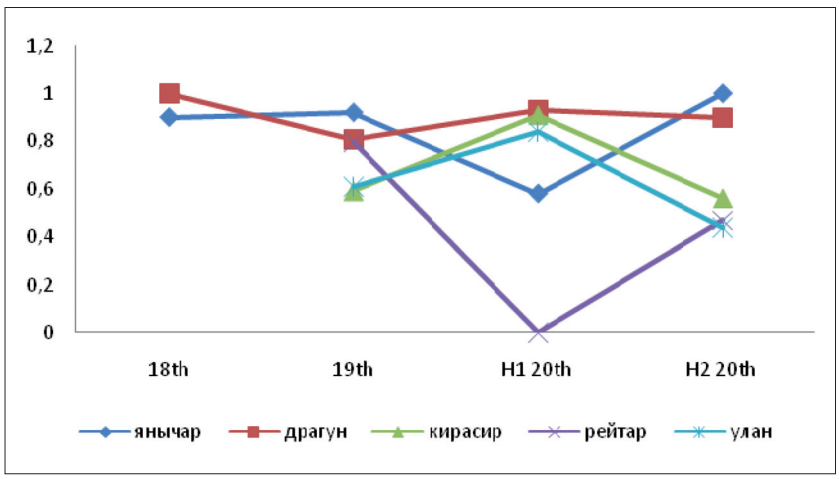

Figure 3: Variant ratio of genitive plural forms with the -Ø ending of masculine nouns in Subgroup 3

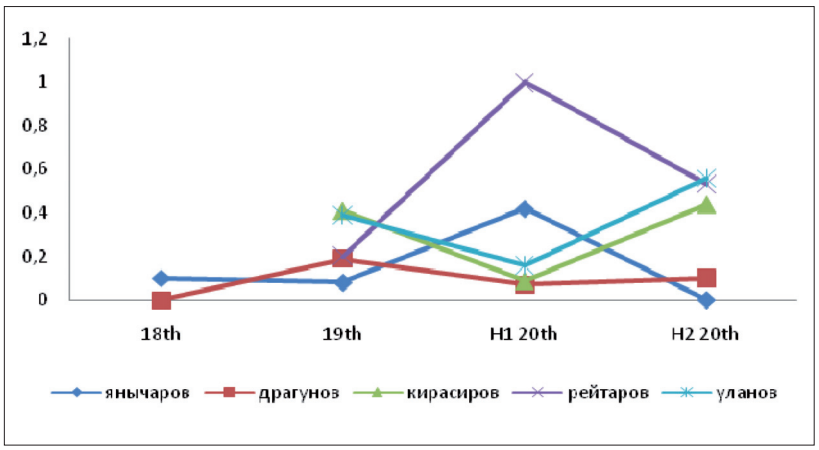

Figure 4: Variant ratio of genitive plural forms with the *-ov* ending of masculine nouns in Subgroup 3

On the whole, the nouns of the third subgroup are characterized by a lower rate of forms with the *-ov* ending and a higher rate of those with the -Ø ending. The quantitative relation between the two forms may be contrasting (as *янычар – янычаров*, *драгун – драгунов*) or may almost be equal (*кирасир – кирасиров*, *улан – уланов*). The second common peculiarity of nouns from Subgroup 3 (except for *рейтар*/ *рейтаров*) is that the proportion of their variants with -Ø and *-ov* endings remains nearly static for the period under study.

The grammatical classification of the studied nouns based on the gen.pl. variants ratio and its dynamics correlates with their lexico-semantic classification. Subgroup 1 includes two nouns from the active vocabulary. The gen.pl. variant competition of the word *солдат* was resolved by the 18th century; of the word *партизан*, by the beginning of the 20th century. The word *партизан* was naturalized in Russian in the 18th century with the meaning 'a strong supporter of a party, cause, or person'; the genitive plural is formed with the *-ov* ending. In contemporary Russian this meaning is considered outdated. The second meaning ('a member of an armed group formed to fight secretly against an occupying force') arose during the Patriotic War against Napoleon in 1812 and became commonly used; its genitive plural word-form has the -Ø ending.

Subgroup 2 includes words that are relatively infrequent in comparison with previous periods, but which are not out of use in contemporary language. Some of them, for one reason or another, even became more common at the end of the 20th century. This applies, for instance, to the following words:

a) *кадет*1 (cadet) 'a pupil of a military middle school' (due to the reestablishment of cadet corps in Russia);
b) *кадет*2 (Cadet) 'Constitutional Democrat' (from its abbreviated name for members of the Constitutional-Democratic Party founded in 1905)
c) *гардемарин* 'midshipman' and *гусар* 'hussar' which have become heroes of literature and cinema ("Midshipmen, forward" by Svetlana Druzhinina, "Hussar Ballad" by Eldar Ryazanov, or "Squadron of Flying Hussars" by Stanislav Rostotsky are very popular films), and so on.

The predominant form of the genitive plural in this group is the *-ov* form although in the 18th century, forms with the -Ø ending were prevalent (except for the noun *кадет*). In current usage, we may observe some regularity: if the word becomes more common, or a new sense develops, or a very old sense is revived in new contexts, the variant with the *-ov* ending is used. An exception to this rule is the noun *гусар* 'hussar': the -Ø form is still prevailing nowadays, although the gap between the frequency ratios of the two variants has become smaller.

Subgroup 3 includes words referring to the passive vocabulary. They fell out of use in the first half of the 20th century because the corresponding military branches were reorganized. Now these words are only used in historical contexts. A dictionary should be consulted when using these obsolete words that is why the preference of -Ø variants within this group remains rather constant during the whole period.

4.3 Corpus data vs. normative recommendations

Normative recommendations differ in different publications. The most authoritative sources are as follows:

Rozental' (1952 / 1977): *Practical stylistics of Russian* – -Ø variant for all nouns is recommended.

Graudina et al. (1976 / 2004): *Stylistic Dictionary of Variants* – -Ø variant for *гардемарины*, *гренадеры*, *рейтары*, *солдаты*, *уланы* and semantic rules of variant choice for the other words. The form *кадет* must be used for 'pupils of a military middle school', but *кадетов* when referring to 'members of the Constitutional Democratic Party'; and the forms *драгун*, *кирасир*, *янычар* with collective nouns – 'detachment', 'brigade', 'squadron', etc., but *драгунов*, *кирасиров*, *янычаров* when referring to individuals.

Zaliznyak (1977 / 2003): *Grammatical Dictionary of Russian*; Yes'kova (1994): *Short Dictionary of grammar difficulties* – for subgroup 1, only the variant with the -Ø ending is recommended. For subgroups 2 and 3, both variants are acceptable.

As can be seen from the corpus data, a slight expansion of the inflection *-ov* can be observed within the investigated group of nouns since the 18th century. This inflection is perceived as dominant for genitive plural forms of masculine nouns with stems ending in a hard consonant, which in turn leads to the unification of the plural case paradigm, in which masculine forms with the *-ov* ending are distinguished from all other forms with the -Ø ending (feminine, neutral, pluralia tantum).

Against the background of these data, the recommendations for the -Ø variant by Rozental' (1952 / 1977) and Graudina et al. (1976 / 2004) seem to be out of date: they are not supported by current usage and the corpus data. The semantic differentiation between 'collective' and 'individual' meaning seems to be too narrow, thus the underlying rules relying on them were violated as early as the 18th and 19th centuries, e.g., "[...] сам же взял с собою [...] суздальских шестьдесят *гренадеров*, сто мушкетеров, [...] и тридцать шесть воронежских *драгун*" (Suvorov 1786), "Миних послал вперед к Яссам Кантемира с трехтысячным отрядом волохов, *драгунов* и *гусар*, а сам следовал за ним"(Kostomarov 1862-1875).

As can be seen from the above, the permissive remark in Zaliznyak's *Grammatical Dictionary* (1977 / 2003) is most acceptable for contemporary normative manuals and grammar books. According to Zaliznyak, the variant with the *-ov* ending is generally recommended whilst the variant ending with -Ø is regarded as an option and appropriate, for instance, in archaized speech.

For more examples for using the Russian National Corpus as a tool for research on grammatical variability see Savchuk (2007), Savchuk / Grishina (2008), Kiseleva et al. (eds.) (2009).

5. Conclusions

The corpus approach for the study of variants in synchronic and diachronic aspects enables us to carry out a qualitative and quantitative analysis of units and constructions; to reveal trends in the relation between competing variants; to trace the development of new phenomena, and to amend lexicological descriptions and normative recommendations.

References

Andrews, Edna (2001): The Russian Reference Grammar. Internet: www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=6 (last visited: 11 / 2010).

Glovinskaya, Marina J. (2008): Aktivnyje protsessy v grammatike. In: Krysin, Leonid P. (ed.): Sovremennyj russkij jazyk. Aktivnyje protsessy na rubezhe XX i XXI vekov. Moskva: Jazyki Slavjanskich Kul'tur JSK.

Graudina, Ljudmila K./Ickovič, Viktor A./Katlinskaja, Lija P. (1976 / 2004): Grammaticheakaya pravilnost' russkoj rechi: Stilisticheskij slovar' variantov. Moskva: AST; Astrel'.

Jakobson, Roman (1956 / 1984): The relationship between genitive and plural in the declension of Russian Nouns. In: Waugh, Linda R. / Halle, Morris (eds.): Russian and Slavic Grammar: Studies 1931-1981. Berlin: Mouton, 135-140.

Es'kova, Natalja A. (1994): Kratkij slovar' trudnostej: Grammaticheskije formy. Udarenije. Moskva: Russkij jazyk.

Kiseleva, Ksenija et al. (eds.) (2009): Korpusnyje issledovanija po russkoj grammatike. Moskva: Probel-2000.

Markov, Vitalij M. (1992): Istoricheskaja grammatica russkogo jazyka: Imennoje sklonenije. Izhevsk: Izd-vo Udmurtskogo universiteta.

Rozental', Ditmar E. (1952/1977): Prakticheskaya stilistika russkogo jazyka. Moskva: Vysshaya shkola.

Savchuk, Svetlana (2007): Corpus-based investigation of language change: the case of RNC. In: Davies, Mark / Rayson, Paul / Hunston, Susan / Danielsson, Pernilla (eds.): Proceedings of the Corpus Linguistics 2007 July 27-30, University of Birmingham, UK. Internet: http://www.corpus.bham.ac.uk/corplingproceedings07/paper/181_Paper.pdf (last visited: 11/2010).

Savchuk, Svetlana / Grishina, Elena (2008): Variation in Russian. Dictionary project. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" Issue 7, 14. Moskva: RGGU [= Russian State University for the Humanities], 466-474.

Shvedova, Natalija Yu. (ed.) (1980): Russkaja grammatika. Moskva: Nauka.

Zaliznyak, Andrej A. (1977 / 2003): Grammaticheskij slovar' russkogo jazyka. Moskva: Russkije slovari.

Zaliznyak, Andrej A. (1967): Russkoje imennoje slovoizmenenije. Moskva: Nauka.