

Google Books Ngram as an Instrument of Teaching Foreign Language

Galeev Timur

Kazan (Volga region) Federal University
Kazan, Russia
tigaleev@kpfu.ru

Solovyev Valery

Kazan (Volga region) Federal University
Kazan, Russia
maki.solovyev@mail.ru

Abstract—Described in the article quantitative study of the evolution of forms of the inflectional paradigm of verbs of the unproductive class I (*mer'-it*→*mer'aj-et* —,to measure“ in Russian) analysis gives international students the opportunity to regularly adjust and update their knowledge in the field of verbal variance. Based on the data obtained through the Google Books case (6.7 billion word forms), managed to describe the main pattern of change in the frequency of competing forms.

Keywords—verb; paradigm; N-gram; Google Books; unification; Russian as a foreign language; learning a foreign language.

I. INTRODUCTION

In the constantly changing Russian shaping for centuries, the process of the unification of verbal stems in the paradigms of the present time in which the unproductive verb classes (I unproductive class: *мер-ит*, *мер-ят* — „to measure“ in Russian) are gradually superseded by productive verb classes (I productive class: *меря-ет*, *меря-ют* —,to measure“ in Russian). Currently, there is no theory describing, explaining and predicting the evolutionary dynamics of variable structures, including redundant verbs.

The occurrence value for the morphological system of verbs in recent years is studied very actively [4]. So, with the help of statistical methods, it has been proven that the more frequency English verbs are less prone to regularization than less frequency (Fig. 1).

Класс частотности по Lieberman et al.	Глагол	Всего употреблений (N)	Ошибочных употреблений прошедшего времени (P)	P/N
2	give	214 000 000	116 000	0,000542
3	seek	23 500 000	258 000	0,010979
4	arise	2 220 000	212 000	0,095495
5	wring	99 800	13 900	0,139279
6	slink	90 900	87 500	0,962596

Tab. 1. Frequency versus Resistance.

A similar study with made using traditional linguistic approaches was performed for the German language [1]. These works confirmed the actual data of the intuitively obvious assumption that more frequency words retain the inflectional type, and less frequency of words tend to change under the influence of analogy.

Ref. [5] on the example of verbs with variation in the type of *хнычет/хныкает* showed that the inflectional paradigm has a radial structure, i.e. it is possible to distinguish the centre and the periphery, and the elements of the linear order paradigm: 3Sg>3Pl>communio>1 and 2 Sg and Pl>imperatives>gerunds. The conclusion is made on the basis of a detailed study of the frequency of occurrence of the matched options for all inflectional forms according to the RNC. It turned out that although in General, the Russian language in these verbs has been a shift from form-a to form-aj, verbs 3Sg longer retain their original form, and the more peripheral it is easier moving to a new. So the verb *хнычет* is still used more often than *хныкает*, but *хнычущий* less frequently than *хныкающий* (Fig. 1).

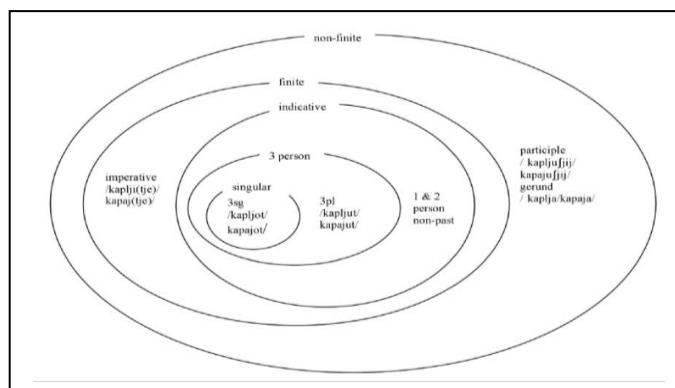


Fig. 1. Paradigm structure.

In the theory and practice of methods of teaching Russian as a foreign language (here in after RFL) study selected based on comparative studies [8] variant (competing) forms of the modern Russian language, so necessary for work with foreign students of advanced learning stage (II-III certification levels B2-C1), until recently, was carried out without using scope databases [2],[7].

In order to deeper understand the related language, it is necessary to consider not only the facts of the synchronic but often also the diachronic character that allows one to recreate common language roots and trace the history of their development in each language. “Cultural and ethnic identity of

each language community develops over a long historical time and has origins in the past" [3].

This work will be tested the following hypothesis: if two verb forms completely synonymous, one of them is gradually replacing the other, however, this process takes place unevenly within the verbal paradigm. The object of research is suffix al changes in excess of the verbal paradigms (мерит/меряет). The subject of our study is that exposure to these forms of 3Sg and 3Pl the unification of the productive type (Fig. 2).

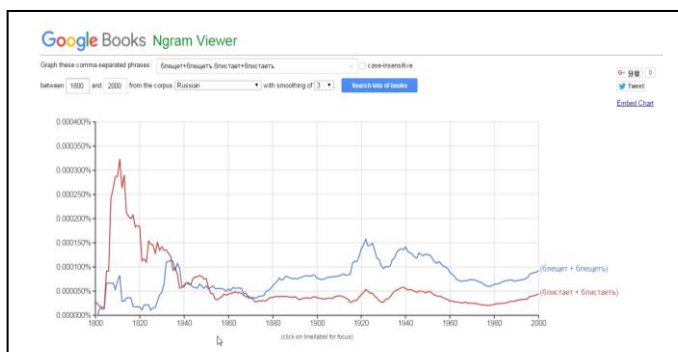


Fig. 2. Example of theResistance.

The aim of the study is to reveal patterns of evolution of various forms of the center of a radial model of the verbal paradigm – 3Sg and 3Pl of the present time. Specific objectives: the allocation of cases change from one form to another over the past 200 years; obtaining numerical characteristics of evolutionary changes for the most "conservative" element of the inflectional verbal paradigm, a comparison of the frequency of usage of 3Sg and 3Pl forms with the others.

II. METHODS.

In order to study the evolution of the lexemes semantics which present some interest the electronic library with Google Books is used with the search service Ngram Viewer (<http://books.google.com/ngrams>), containing the texts, the earliest of which date back to the beginning of the XVIIIth century., with the total volume of more than 67 billion words in Russian language, which 200 times more than the volume of the Russian Language National Corpus (RLNC). It should be noted that the degree of Google Books corpus material diversity is different from RLNC, which also contains periodicals, blogs, speech, etc. along with the book texts. The rest of it seems to be very balanced, that is, it contains the texts of different genres, spheres of functioning and different subjects.

In this work, the first examples of this corpus use are presented in sociological, historical and cultural studies. In the future, it is used in many works to study the evolution of various aspects of society in the context of the language vocabulary evolution. To study the evolution of variant forms is proposed to use a quantitative method. Based on the data of the case "Google Books" (Google Books, then – GB) providing the service Ngram, performing a search for books published mostly from 1800 to 2000, will be built the graphs of the variation of the frequency of 14 pairs finite and

definitive forms (personal form – 6, of the sacrament – 4, the participle – 2 imperative – 2) constituting the redundant paradigm of the 50 verbs.

Fig. 3. Example of the Resistance.

сыпать		
	Ед. ч.	Мн. ч.
1 л.	сыплю/сыплю	сыплю/сыплю
2 л.	сыплешь/сыплешь	сыплете/сыплете
3 л.	сыплет/сыплет	сыплот/сыплот
Активное причастие настоящего времени		сыплющий/сыплющий
Активное причастие прошедшего времени		сыпавший/—
Деепричастие несовершенного вида		сыпля/сыпля
Императив		сыпл(и)те/сыпл(и)те

In recent years scope diachronic corpus "Google Books", Russian-speaking part of which consists of 6.7 billion word forms, is still poorly used, development of a methodology for its use is a very important task in connection with the accompanying potential complications. First, the "peripheral" forms (participles, gerunds) are less common in spoken language, namely, it is a testing ground for linguistic experiments. However, the peculiarities of the style of the book contribute to the accumulation of interest to us empirical material that makes the language dynamics based on the data of GB immaculate illustrations for verbs of the studied type. Secondly, the majority of the shift of verbs is presented in the spoken language. The reason for this can be considered the desire of the speaker to replace the questionable word form, selecting a synonym, or avoid it altogether by changing the sentence structure. And as the "variable" verbs mostly belong to the literary style of speech, then GB will be the best tool to study the question of variability.

III. RESULTS AND DISCUSSION

A. Prototypical forms

The study was first obtained frequency characteristics of the functioning of the redundant verbs of the specified type. It was built 446 charts changes in the frequency of usage. 93 graphs describe the diachronic changes of word pairs 3Sg and 3Pl.

In parallel a classification of verb pairs according to dynamics of the frequency of their use. More than half of the cases (55%) unproductive form dominates productive, changing norms is not expected (колышет more often than колышет) (Fig. 4). Almost 12% of cases were detected only form the basis of the unproductive type (varies). In 10% of cases unproductive form becomes more productive relative frequency, which is 100years ago was more frequent (движет began to be used more often than двигает) (Fig. 5). About 5% of the graphs illustrate the decline in the frequency of both forms in the XX century while maintaining the unproductive type as dominant (алчет still more often than алкает) (Fig. 6). Comparing the data on the forms 3.with other forms, it is

possible to conclude that nepodvijnosti forms of 3Sg and 3Pl. In 41 of the 50 paradigms (82% of cases) in these forms preserved the old type of declension. Other forms of productive declension class are much more common. Approximately 30% of cases the type of declension forms of the 3Sg and 3Pl (кудахчет, кудахчут) does not match the type of the decline in other forms (кудахтаю, кудахтая, кудахтающий, кудахтай).



Fig. 4. Knees versus swaying *Движеть* versus *движает*.



Fig. 5. Alchette versus is hungry *Алчеть* versus *алкает*.

In a much smaller number of graphs (less than 10%) of the productive type during the 2nd half of the XX century replacing unproductive form type (form dripping supplanted by a form of drips). Interestingly, the archaic verb can make it less "sustainable" unification: in a third of cases the change of norms takes place against the background of the decline of the frequency (клеплет→клепает) (Fig. 7). In other charts (8.5 percent) is the same frequency of competing forms exist in parallel (лазит/лазает) (Fig. 8).

Fig. 6. Claplet versus rivets *Клеплет* versus *клепает*.

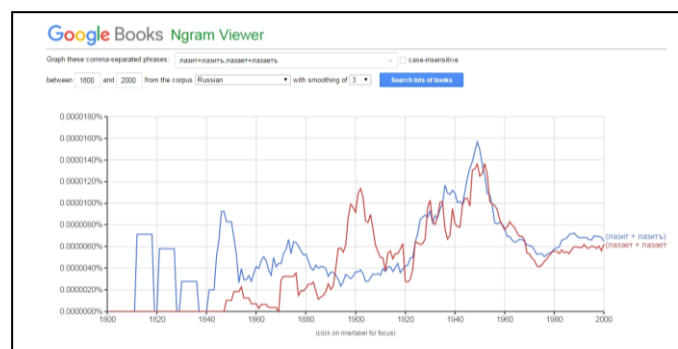


Fig. 7. She climbs versus climbing. *Лазит* versus *лазает*.

B. Communication of frequency and "conservatism".

The second phase of the study was to compile a list of the frequency of these verbs. If the group E. Lieberman to work



with the list of more than 200 verbs used method of ranking, in our study, 50 of paradigms more appropriate to compare the 10 least frequent verbs (up to 3 thousand occurrences over 200) with the same number of most frequent verbs (30 – 400 thousand occurrences per 200 years).

In the dynamics of changes in the frequency 8 of the 10 most frequent verbs prevails a tendency to preserve prototypical form. The opposite dynamics is observed among the rarest of verbs: 7 out of 10 verbs underwent unification by I productive class. Thus, we can conclude that the most frequent verbs are more conservative.

IV. CONCLUSIONS

The experiment showed that conservative verbs are more frequent: the more often a verb is used, the more it retains its original form. Inside the verbal paradigm, the most conservative are 3Sg and 3Pl forms: they have the highest resistance to change and unification.

It was possible to make a classification of verb pairs according to dynamics of the frequency of their use on the basis of the biggest diachronic Corpora.

Graphs and tables supporting these findings are available at <https://cloud.mail.ru/public/HSht/ry3ewo3oF>

ACKNOWLEDGMENT

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University and to RFBR (project №16–06–00165 A).

REFERENCES

- [1] R. Carroll, R. Svare, J. Salmons, "Quantifying the evolutionary dynamics of German verbs," *Journal of Historical Linguistics* № 2 (2), 2012, pp. 153–172.
- [2] T. Galeev, "Application of linguistic typology in training of Russian for foreign students. How can language typology databases be useful in studying languages?," *EDULEARN14 Proceedings*, 2014, pp. 5995–6005.

- [3] I.V. Erofeeva, "The Methodology of Teaching Russian as A Foreign Language to Slavonic Speaking Students," *Procedia Social and Behavioral Sciences*, vol. 186, May 2015, pp. 1095-1100.
- [4] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang and M.A. Nowak, "Quantifying the evolutionary dynamics of language," *Nature*, vol. 449, 2007, pp. 723–716.
- [5] T. Nessel, L. Janda, "Paradigm structure: Evidence from Russian suffix shift," *Cognitive Linguistics*, vol. 21(4), 2010, pp. 699–725.
- [6] V.D. Solovyev, A.A. Kibrik, "How can computer technologies help linguistic typology?," *Herald of the Russian Academy of Sciences*, vol. 85, issue: 1, 2015, pp. 33-39.
- [7] M. Varlamova, E. Palekha and A. Miftakhova, "Visual Listening in theory and practice of effective foreign language teaching," *INTED: 8th International Technology, Education and Development Conference*, March 2014, pp.7129-7133.
- [8] V. Koprov, *Variant forms in the Russian language: the textbook for foreign students*. Voronezh: Voronezh State University, 2001.