


DISTANT READING OF NAÏVE POETRY: CORPORA COMPARISON AS RESEARCH METHODOLOGY

Bonch-Osmolovskaya, Anastasia 

National Research University Higher School of Economics Moscow, Russian Federation

Orekhov, Boris 

National Research University Higher School of Economics Moscow, Russian Federation

Category: Long Paper

Session: 7

Date: 2014-07-11

Time: 11:00:00

Room: 321 - Amphipôle

The method of distant reading - term proposed by Franco Moretti (Moretti 2005 , Moretti 2013) , or – using another term – macroanalysis (Jockers 2013) has become a mainstream in digital humanities in last couple of years. The general idea of the approach is to gain new knowledge about literary and cultural processes with the help of digital and quantitative models applied to all sorts of language or literary resources. In our paper we follow the distant reading method focusing on the phenomenon of naïve poetry – poetical opuses, composed by non-professional poets and distributed on special web-sites. One of them – Russian stihi.ru (*stih*i means 'verses') – has nowadays become a giant collection of dilettant literature with more than 5000 authors, about 21 mln of works and everyday update. It, thus, can be regarded as an extraordinary cultural linguistic resource made by crowd sourcing. At the same time Russian National Corpus resources possess a unique resource - Poetic corpus (Grishina et al. 2009; <http://ruscorpora.ru/search-poetic.html>). In contrast to stihi.ru, Russian Poetic Corpus has been collected and marked up by the team of experts in linguistic and literary studies and it presents Russian poetical classics from the 18th century till the early 1930th.

Comparison of the two poetic resources of naïve and classical poetry gives us an excellent possibility to use quantitative analysis to get some promising insights. We can understand more about the nature of literary imitations and epigone writings, the foundations and circulations of literary canon, the mechanisms of prosaic/poetic language shifts and many other topics related to sociology of textual culture that could not be studied with traditional methods. Some steps taken in this direction are presented in this paper.

We look first of all at frequency measures in both resources and analyze the revealing fluctuations of different word frequencies. We use the frequency list of Russian National Corpus (called below the general frequency list) as a controlling benchmark, that helps us to separate the words, which are most frequent in common Russian Lexicon, from those, which get high frequency exclusively in Russian verses, high or naïve. Then we make a qualitative analysis of the nouns that occur in the top 100 of each frequency list. We identified a range of semantic domains that can be expressed by these nouns and compared the domains of each poetical resource. As a result we defined three main strategies of naïve composition.

Preparation

For our research we have taken a sample of 50 mln word usage from naïve poetic corpus, which makes a representative corpus of more than 54 thousand authors. As our main aim was to find out what poetic patterns from high poetry are apt to be borrowed and imitated, we decided to extract a subcorpus of most typical imitating poetry. We searched for the authors who would bear in mind high poetical examples and try to go with them in their composition. The sorting has been made automatically. First we extracted all the bigramms from the high poetical corpus, and then we took only those documents from the naïve corpus, which a) have at least 50% bigrams that coincide with the bigrams of the high poetry list, b) are longer than 20 bigramms. Our final sample consisted of almost 9 mln word usage. The high poetic corpus has about 8 mln word usage. We lemmatized all the words in both corpora. After lemmatization the naïve poetic corpus consisted of 84 thousand lemmas

Methodology

We conducted three analyses based on the comparison of the three frequency lists: the frequency list from naïve poetic sample, the frequency list from Russian Poetic Corpus and the general Russian frequency list based on Russian National Corpus. In the first experiment we compared the change of ranks of very high frequent words in the naïve sample relatively to high poetic and general lists. Secondly we considered outliers of naïve poetry frequency list: those words that demonstrate dramatically different frequency behavior. The last experiment consisted in relating all the top 100 nouns of each frequency list to semantic domains, that they are most probably used for. Then the contents and the variety of each domain in each list has been analysed.

Results

The interpretation of the resource comparison results can be summarized by defining three basic strategies in naïve poetry: imitation, self-actualization and naming. Each of the strategies will be illustrated below by data examples.

1. The top frequency list of naïve poetical resource shows interesting deviation both from high poetical corpus list and general frequency list (see table 1)

Word (in Russian)	Word (translation)	Position rank in naïve list	Position rank in high poetical list	Position rank in general frequency list
И	and	1	1	1
Я	I	2	3	5
не	not	3	4	3
в	in	4	2	2
ты	you	5	7	33
то	this	6	5	23
что	that	7	11	9
быть	be	8	10	6
на	on	9	6	4
как	as	10	9	19
с	with	11	8	8
мы	we	12	13	18
а	but	13	17	10
мой	my	14	15	60
но	but	15	14	16
так	so	16	27	30
за	for behind	17	22	24
любовь	love (noun)	18	52	307
любить	love (verb)	19	66	181

As we can see from the table, the naïve poetry demonstrates important lexical features, some of them are specific, and some of them are typically poetic, being shared with the list of high poetry frequency. We observe an interesting tendency at the very top, where personal pronouns I and you displace the most common propositions in and on from the second and the fourth positions correspondingly. Both pronouns I and you can be considered as lexical traits of poetical discourse. But the naïve poetry shows higher ranks for both of them (2 vs 3 in high poetry, and 5 vs. 7 in high poetry correspondingly). We see increase of frequency of those words which are already indicative for high poetical frequency list. The tendency to intensify specific poetical lexical features can be called the imitative strategy. The rank shift of the words love (noun and verb) is even more straightforward manifestation of the same strategy. While in general frequency list those words are not even in top 100, they occupy 52 and 66 positions in high poetical list and so far being an etalon of poetical shift they become the most frequent words in naïve poetry

2. The table below shows 5 nouns which have the biggest difference of ranks between naïve poetry list noun frequency of the Russian Poetical corpus (see table 2)

word	word (translation)	Position rank in naïve list	Position rank in high poetical list	Position rank in general frequency list
------	--------------------	-----------------------------	-------------------------------------	---

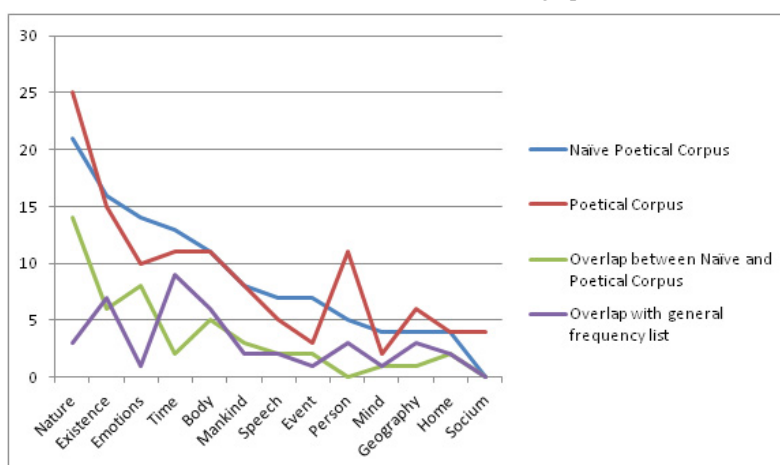
русское слово	английское слово	971	30827	3400
фото	photo	971	30827	3400
сигарета	cigarette	939	28957	2109
проблем	problem	610	12572	197
мама	mum	330	2125	309
девчонка	babe, gal	865	5030	2563
	(derivative from girl)			

These words behavior is various: some of them are more frequent comparing the general frequency list (*photo, cigarette*), some of them stay on roughly the same rank (*mum*), some are more rare than in general, but still show immense difference with the high poetical list (*problem*). The gap between naïve and poetic frequency list signals that there are some semantic zones where naïve poetry seems to be independent from the classical poetic canon. This trend can be defined as a self-actualization strategy which is in some sense opposite to the imitative strategy.

3. We took top 100 nouns of every list and compared their lexical distribution. The nouns had been grouped into abstract semantic domains. Some words could be associated with several domains due to their polysemy. As a result we have identified 13 semantic domains, 12 of them are shared between naïve poetry, high poetry and common frequency lists and the 13th is not presented in the naïve poetry list. The domains we have defined are as follows:

Mankind (everything that may characterize a person: *soul, beauty, name, heart, strength* etc.), Body, Emotions, Mind, Existence (*God, world, truth, time, fate* etc.), Speech, Person (*father, son, friend, enemy* etc.), Event (*love, happiness, past, disaster* etc.), Time, Nature, Geography (*road, hill* etc.), Home (*window, door* etc.). The 13th domain which is found only in the high poetical list and in the general frequency list is Social and it includes such words as people, labor, fame in the poetical list, and state, money, law etc. in the general list.

Analysis of the overlaps and varieties of the naïve and high poetical lists showed differences in the elaboration of the domains in two corpora. The general frequency list helped to draw out the words that are commonly frequent and their presence in the list cannot be understood as a signal of the poetic concentration on the domain. The results are demonstrated on the graph below:



As we can see from the graph, there are three zones of the naïve poetical sample that demonstrate high lexical variety of frequent nouns in comparison to the poetical corpus. These are Emotions, Event and Speech. Most of the words of those domains are not frequent in general lexicon. The lexical multiplicity can be explained by extensive strategy: the naïve poets do not use sophisticated verbal apparatus to express the conceptual space of the verse, but prefer straightforward lexical naming (*pain, wish, encounter, grief, love, question, answer* etc.)

References

Grishina E., Korchagin K., Plungian V., Sitchinava D. *Poeticheskij korpus v ramkah NKRYa: obschaja struktura i perspektivy ispol'zovanija* 'Natsional'nyj korpus russkogo jazyka: 2006-2008. Novye rezul'taty i perspektivy'. Saint Petersburg, 2009. P. 71-113. [Poetic Corpus in RNC: general structure and using perspectives]

Moretti, Franco. *Graphs, Maps, Trees: Abstract models for a literary history*. Verso, 2005.

Moretti, Franco. *Distant Reading*. Verso, 2013

Jockers, Matthew L. *Macroanalysis*. University of Illinois Press, 2013