**BONCH-OSMOLOVSKAYA — Medium**

<0 figures>

**Medium Data Method for Cultural Studies: The Case of Gender Studies in the Russian National Corpus**

Bonch-Osmolovskaya, A.

The main aim of the paper is to show how the data of the Russian National Corpus can be used to explore nonlinguistic objects. The National Corpus is created as a well-balanced, marked-up text collection оf almost 300 million documents of different ages (18th–21st centuries), followed with rich metadata. Being one of the best national linguistics corpora it can also be regarded as a store of cultural memory and social reflection. Then the quantitative analysis of the change of frequency of some conceptually important lexeme or qualitative analysis of its various contexts in different periods can lead us to some new analysis of cultural trends and social processes, focusing primarily on their reflection in social consciousness.

It should be emphasized that though some lexical change can be explained by extralinguistic reasons, the study is based on texts that are not thematically connected with this reason. For example, we observe an abrupt increase of the use of lexeme ‘woman’ in the second half of the 19th century, which surely should be associated with the first wave of emancipation in Russian society. But these are not manifestos, but mostly novels that are being analysed. The frequency increase for ‘woman’ in the 1850s signals a change in readers’ interest. This change and its connection with social processes and events are the objects of investigation with the methods of corpus analysis, distant reading, and culturomics.

The paper doesn’t claim to make some profound sociocultural research, but the main focus will be on methodology. I will introduce the medium data approach and will argue that it can be used for cultural research, and that it allows solving some of the well-known problems of big-data analyses in textual data. First I will observe the culturomics approach and the benefits of National Corpus use instead of Google Books. Second I will concentrate on one case of medium data cultural research: gender naming in the 19th and 20th centuries in Russia and the Soviet Union. I will show how context analyses can shed light on unexpected data emission, and how competition of lexemes reflects changes of social consciousness.

Culturomics was first declared as a new scientific method in a 2010 paper in *Science* magazine titled ‘Quantitative Analyses of Culture Using Millions of Digitized Books’ by a group of scientists headed by Jan-Baptiste Michelle. Though very influential and inspiring, the method also has been widely criticized. The two main problems of the method mentioned most often are the trivial results and the dirty, nonreliable data. The methodological problems of culturomics research seem to be the other side of its benefits: the big textual data only allows comparing the queries known in advance. We can observe the effect of censorship by the decrease in Mark Chagall mentions, but we have no instrument to compare the change of semantic context of Chagall’s mentions that accompanied the beginning of content restrictions. That means that we can notice how the political changes are reflected in culture, but we still lack the instrument to explore these processes thoroughly.

In contrast to the influential trend of Big Data, I propose a concept of nedium data. Medium data is the amount of data that allows for quantitative and qualitative studies. The main characteristics of the medium data are

• The reliability of sources, which metadata can be filtered manually.

• The sufficiency of the data amount for reliable statistical measures.

• The possibility of additional semantic mark-up.

The medium data concept serves to oppose the current practices where computational methods tend to ignore the complexity of the humanitarian sphere. I argue that the quality of the research can benefit much from the contamination of statistical and computational methods, with expert manual analyses possible only with very pure, precise data of not a tremendous amount. Although the primary data filtration will in this case be a matter of the researcher’s responsibility, this situation really doesn’t differ much from any natural science case, when the researcher has to provide the specific conditions of the experiment.

Though the Russian Natural Corpus is much smaller than the Google Books corpus, its dataset surely has some advantages. First, every document has rich metadata. This allows counting not only the edition date but also the creation date, which can be quite important for studying the Soviet period, as its published texts can demonstrate a significant lexical bias due to censorship. Second, the mistakes of object recognition are very rare in RNC. The morphological mark-up, which plays an important role in a morphologically rich language such as Russian, is diverse and multivariate in RNC, while it is limited to POS-tagging in Google n-grams. The complex morphological mark-up gives an opportunity to make distant queries, which are targeted to represent syntactic relations. For example, the query that consists of the verb ‘to steal’ plus the noun in the accusative extracts the change of consumption needs for different periods (what people steal in different times is what they actually need). Their comparison can inspire some new level of social and cultural research. The textual collection and the markup of the Russian National Corpus thus give us an example of the medium dataset that can be used not only for language investigation but also for social research.

How can we benefit from a medium dataset? The research can focus not only on frequency change but also on qualitative context variation. It is not the words themselves but the semantic concepts (synsets, that are different in every period) that are being studied. Medium data allow disambiguation of semantic polysemy, which is usually impossible with big datasets and sometimes can cause damaged results. The semantically close contexts can be merged into classes that enable more transparent and still reliable analyses. The most significant difference between culturomics and medium data can be formulated as follows. Culturomics research results in an overall graphic that often demonstrates quite trivial dependencies. The medium data method allows for treating the graphics not as a diagnosis but as symptom, which serves as an impulse for further research.

Finally I will focus on one model case of the medium data method. I will compare the frequency of the words ‘man’ and ‘woman’ in the 19th and 20th centuries. The comparison of two frequency graphs shows that the word ‘woman’ is much more frequent than the word ‘man’ in the both centuries. Is it that women are more often written about? Or are men referred to with some other lexical means? If we compare the two words ‘muzhik’ and ‘baba’, which in the 19th century are used as gender terms of low class, we get the opposite picture: males are more frequent than the females. The answer can be drawn out of the context analyses. The reference of males in general is rarely direct in the 19th century, but mostly implicit together with some specific lexical means, characterizing age, social, or professional status. This contrasts much with female references, for which the gender idea is much more important than the social occupation. I will also follow the changes in male and female word usage that take place in Soviet and post-Soviet epochs.