THE PROBLEMS OF FORMALIZATION OF LITERARY TEXTS IN HIGHER SCHOOL

Bogun M.V. (Kharkiv)

Abstract. *The paper concerns problematics in the methods of formalization of literary texts from the point of view of thematic modelling. A number of possible ways of text formalization are being discussed setting up the questions to solve these problems.*

Key words: *computer linguistics, formalization, thematic modelling, deep learning, neural networks*

ПРОБЛЕМИ ФОРМАЛІЗАЦІЇ ЛІТЕРАТУРНИХ ТЕКСТІВ У ВИЩІЙ ШКОЛІ

Анотація*. Стаття присвячена проблематиці у сучасних методиках формалізації літературних текстів з точки зору тематичного моделювання. В статті висвітлюються різні види формалізації текстів ставлячи перед читачем запитання спрямовані на розв'язування вищеозначеної проблеми.*

Ключові слова: *комп'ютерна лінгвістика, тематичне моделювання, глибоке навчання, нейронні сіті.*

Digital research of the literature is to some extent customary because now there are large amounts of data in different fields of knowledge and people have learned to manage and to study it. The following question arises: Can we do the same with the literature? We are able to analyze large amounts of data relating to astronomy, physics, biology and genetics. But literature is a complex aspect, complex object, and perhaps it could be also somehow explored with the use of computers to understand what is happening there, what kind of sophisticated trends exist there that are not always obvious to readers.

But first of all, when we talk about this, we need to understand what the formalization is, because the science is primarily trying to simplify its object, to divide it into a number of parameters and to make numerals, and then to count them using the computer because it is its main function. The difficulty of the digital study of literature is that it is not very clear how a literary work can be formalized. And any natural science starts from formalization.

In physics for instance the task becomes much simpler: there is some object such as a star, a planet or a quasar, and scientists watch it and know what parameters from those they watch (e.g., luminosity, or the position in the sky) are important and which are unimportant. For example, spectrum or spectral class of stars are, probably, important, and some other aspects (e.g., a person who is watching a star) are not very important. And this is not applied to literature, because science in any case is modelling, object simplification, and we do not know exactly what we can simplify without losing the content of the literary work and what we can convert into numbers and what not. And it is really a very big problem, which is still not resolved, and scientists are now in a position to find some solution to this problem.

The question of how a literary work can be formally described has been dealt for a very long time. Yet in the 1920s, the famous literary formalism was partly involved in this problem. It was found that folklore works can be formalized much better and literary works did not give such results, although those schemes that were applied by Vladimir Propp to the fairy tale in his time, were also tried to apply to author's literary works. But here we may see great diversity naturally found in the author's literature.

Propp analyzed many of his country's folk tales and identified common themes within them. He broke down the stories into morphemes (analyzable chunks) and identified 31 narratemes (narrative units) that comprised the structure of many of the stories. But still he has been both lauded for his structural approach and criticized for his lack of sensitivity to subtle story elements such as mood and deeper context.

In the 1960 a new trend appeared, some kind of an update of the old tendency. Scientists again began to look for opportunities of formalization of literary works, to search for some schemes and structures. This was generally the time of the progress of natural sciences, and computers appeared at the same time. There was no significant progress on this way for many reasons, some of them have already disappeared from the horizon at our time. Among them we can name computer capacities, which have been increased significantly since that time. Also now we have electronic access to the digital texts which may be analyzed automatically.

And now we see a new renaissance of this research. But the difficulties that we have mentioned, remain. What can we convert into numbers and what we cannot? Philologists are looking for some senses in the literary work, they are very difficult to be transferred into numbers, to make some numerical parameter. And computer may deal with some elemental facts, such as, for example, words. Words can be counted, and if we take any corpus of literary texts, we can look for the tendencies existing in the words with the help of which art senses are implemented in the literary work.

Indeed, in recent years such researches appeared, trying to determine what a postmodern novel differs from similar literary work with – e.g., modern ones at the same period of time. And there are methods of computer linguistics, which allow to calculate how much one text is similar to the other. For example, we may collect all postmodern novels with the other ones that we do not consider postmodern, and then find out what words indicate special content relevant to the style.

Another aspect is more formalized and represents what characterizes the poetic speech. Poetic language is organized rhythmically, it has stressed and unstressed syllables. And the problem is how we can explain to the computer what distinguishes pentameter from chorea. In case of success we will easily be able to formalize the parameters important for the organization of the text, such as cadency, meter, to calculate what is more often used in poems, and what is used more rarely, and to make conclusions on this basis. And when we have a lot of texts and we can analyze them with a computer, it would be interesting to find some great tendencies in them, imperceptible for close reading, familiar to philologists.

Philologists are able to read the texts attentively – it's called close reading. And what will happen if we collect all the novels of the XX century and try to extract some trends and patterns from them? Franco Moretti published a book "Distant Reading". It is some opposition to close, attentive reading, maybe even "abstract reading". That means we digress from the text, try to extract some information from it and examine it using large amount of data. Analysis of the data represents the area which is very important for modern life, not only for science. We take a huge amount of information

and try to analyze it statistically. It is very similar to the philology, because here and there we are trying to find some non-trivial patterns that are not visible at first sight, and in one case the statistics helps us, and the other, we have not decided yet how to use it.

But if we let a computer analyze a text with some understandable way, it may show some things that are not visible to the reader at first sight. For example, if we analyze dialogues and speech by the heroes of some great novels, we will find out that, for instance a certain kind of characters typologically combined in some groups may use the verbs of the same type in their speech, but a different kind of characters may express themselves in another way. Though the speech was written by the same author, by one and the same person, and it should not be so different. Computational linguistics allows us to classify characters by their speech. This is something that is difficult to do in the process of slow reading, especially if we have a big novel.

Another important problem concerns the subject of the text, thematic modelling. We understand the meaning of the text very well when we read it, but it is hard to explain the meaning to the computer. Nevertheless, due to mathematicians and computer scientists, we are gradually approaching to the solution of these problems, and we are now able to explain to a computer where the difference between one theme and another lies, for example between the themes of nature and love.

The dynamics of their development in a literary text, invisible in the large amount of material, is now available to us and we can see its signs in more text typologically. For example, there are such tendencies that at the beginning of the novel we are talking about something good, and by the end of the novel we are slipping into something tragic. These are the categories that were important to the medieval literary criticism, because, as we know, in the Middle Ages the comedy was a story that started badly but ended well, and it was not connected with something funny.

Digital technologies, which we deal with, are not yet able to reach for the problems important for literary criticism. They are the matters of sense, abstract matters, matters of higher functions of the nervous system. Here, the computer is still a very stupid device and can deal only with outer ways of realization of these matters, that is,

with words, their combinations, with their distribution in the text. Of course, this is still not enough, but, probably, the further machine learning will develop, the so-called deep learning, neural networks, the closer we will be to understanding how the text creates the impression on a man, when tension arises, or when, on the contrary, a man loses attention for the text and starts to think about something else. We can finally explain to the computer how to deal with the major, almost reflex things related to reading of fascinating works.

So the extraction of the motives and computer trying to recognize where "the strains" in the text which capture the reader are and where the relaxed moments are – this is what the digital humanities will deal with in the near future. But in fact they are just new methods, which humanitarian scientists prefer, and, in contrast to the natural sciences, where the leading scientists are those who discover something, in the humanities not discoveries determine the significance of a scientist, but the extent to which he is able to invent new methods and engage others' ones. Therefore, the involvement of digital technologies in the near future will give impetus to the humanities.

References

1. Bettina Fischer-Starcke. Corpus Linguistics in Literary Analysis. Jane Austen and her Contemporaries. Corpus and Discourse. Bloomsbury Academic, 2010. – 240 pp.
2. Franco Moretti. Distant reading. Verso, 1 edition. / London, New York, 2013. – 224 pp.
3. Propp, V. (1927). Morphology of the Folktale. Trans., Laurence Scott. 2nd ed. Austin: University of Texas Press, 1968.