Richard ZMĚLÍK
Palacký University in Olomouc
DOI: 10.14746/bo.2023.3.8

Lubomír Doležel and his Legacy to Digital Humanities[1]

Keywords: Lubomír Doležel, theory of information, cybernetics, quantitive methods in linguistic and literary studies, digital humanities in literary studies

Abstract

The study focuses on that part of Lubomír Doležel's scientific legacy which is connected with the application of quantitative and statistical methods in linguistics and literary science. In the early 1960s, the influence of information theory and cybernetics began to be strongly felt in the context of these disciplines. Alongside Jiří Levý, Lubomír Doležel was one of several linguistic researchers who creatively rethought the possibilities of these disciplines, among others, in literary science. This interest lasted, to varying degrees of intensity, until the 1980s. Although this phase of Doležel's scholarship is usually neglected at the expense of his other works, for which he eventually became known as a world-class literary theorist, their importance is confirmed today, when quantitative methods are beginning to be used again in literary scholarship in general in the context of the digitial humanities. This study focuses not only on Doležel's legacy to today's literary scholarship in the digital humanities, which works with these methods, but above all highlights those of his ideas that, seen through today's lens, have become literally timeless in this context.

1.

Lubomír Doležel is world famous primarily as a theoretician of fictional worlds, or rather as the founder of fictional semantics. He also has considerable merit in the fields of narratology and literary interpretation. However, it is no secret that he began his scientific career not as a literary theorist, but as a linguist who focused his interest mainly on stylistics, initially on the functional analysis of semi-direct speech. From the very beginning, Doležel's professional interest was de facto both linguistical and literary, which, after all, corresponded to the tradition of the Prague structuralist school, which he was significantly inspired by.

At the beginning of the 1960s, he was professionally and organizationally involved in the then completely new and promisingly developing discipline, mathematical linguistics, that benefited from the stimuli of cybernetics and information theory, which after the Second World War also began to have a significant profile in linguistics and gradually their influence began to be reflected in the context of literary studies (e.g. Pavel Vašák, Jiří Levý or Eduard Petrů). Doležel recalls this period in his memoirs, where he states:

> It must be said that both Levý and I were passionate about the application of mathematical methods in the study of language and literature. The main impetus for this enthusiasm was the pursuit of the highest possible exactness and explicitness in the formulation and application of the theory. The 1960s was, among other things, a time when the demand for exactness in all humanities and social sciences was asserted in the world and then also in our country. In the ideal of the exactness of theories and methods we saw freedom from ideological chatter and rapprochement with the ‚real', i.e. with the natural sciences. We wanted to strive to overcome the doctrine of ‚two cultures' and to realize the ideal of ‚unified science' (Doležel 2013, pp. 111–112).

What inspired Doležel to start dealing with this research area resulted from of two basic circumstances: firstly, the still relatively young researcher was here influenced by the allure of new and progressive methods that promised to bring linguistics, and potentially also literary studies, closer to the limits of exactitude, which had always been the domain of the natural sciences. The second circumstance was the fact that the epistemological stimuli that the theory of information and cybernetics provided to these fields (linguistics and later literary studies) corresponded very well with the initial basic assumptions of the structuralist theory. The common denominator was mainly the empha-

[1] The article was created on the occasion of the centenary of the birth of Lubomír Doležel (1922–2017) and was presented at the Brno Colloquium for the 100th birthday of Lubomír Doležel held in October 2022.

The preparation and publication of the article was made possible thanks to the financial support of the Faculty of Arts of Palacký University in Olomouc in the years 2022–2024 from the Fund for the Support of Scientific Activities (FPVC2022/20) .

sis on the sign system and its structural arrangement, while quantitative or statistical analysis, as methods associated with information theory and cybernetics, approached structural organizations primarily from the point of view of their formally empirical manifestation. In doing so, they emphasized not only the uniqueness of the sign, as a carrier of information or as matter that can be coded into a binary numerical system for its machine transmission, but also of the entire sign system, which was analyzed with the use of the mentioned methods.[2] Ultimately, however, information theory was not only about the rules of the formal structural arrangement of the system; the question of the meaning of a linguistic sign was also reflected by this theory. As, among other things, Max Bense notes in his book *Theory of Texts* (1962, in czech translation in 1967), between the degree of organization, or there is a mutual relationship between the disorder of the system and the meaning of linguistic signs and entire structural systems, which can be expressed statistically using Shannon's expression for entropy.[3]

From the viewpoint of information theory, meaning is given by the degree of orderliness. In other words, the meaning is shaped by convention, i.e. the probability that is given depending on the frequency of the element (sign). The higher the probability of the occurrence of an element in the system, the more stable the meaning, not the information (!) and the lower the level of information, and vice versa.[4] This relationship can be written symbolically as follows:

[2] Doležel notes in his own memories that Sgall's group „set themselves a challenging task – to mathematically (algebraically) model the structures of natural language. […] Because I intuitively felt that language style is a probabilistic phenomenon, I chose statistics and probability theory as my mathematical methods." (Doležel 2013, p. 112)

[3] Formula for calculating entropy is: $H = -\sum_{i=1}^{N} p_i log_2 p_i$, where p_i is the probability of a specific element of a certain structure defined by the relation: $p_i = \frac{f_i}{N}$, where f_i is frequency of item and N is size of text.

[4] Significant thinking about the possibilities of cybernetics and information theory both in artistic creation and, for example, in aesthetics took place in the 1960s at the

$$Z_{freq+}^{(-inf,\ +sem)} \qquad Z_{freq-}^{(+inf,\ -sem)}$$

A number of works by Lubomír Doležel from the 1960s also belong to the framework of these contexts. Probably Doležel's first contribution to the issue of quantitative and statistical methods in linguistics is the study *Předběžný odhad entropie a redundance češtiny* [*Preliminary estimation of the entropy and redundancy of written Czech*], published in 1963 in the journal *Slovo a slovesnost* [*The Word and Verbosity*]. In it, Doležel addresses the issue of „the distribution of the frequencies of graphemes and their digram combinations in texts" (Doležel 1963, p. 165), which he solves using entropy. In the abstract of the treatise he formulates the basic goals and results:

> The theoretical meaning of the treatise consists in assessing two basic questions of the linguistic interpretation of entropy and redundancy: 1. Are these values characteristics of the language as a whole or of individual language styles? 2. Do different languages show similarities or rather differences in terms of these characteristics? On the basis of the results, **numerical entropy**, interpreted as a characteristic of language, and **predictive entropy, considered a characteristic of language styles,** are tentatively distinguished in the state (Doležel 1963, p. 165).

Doležel's aim was to formulate the initially intuitive knowledge about the stylistic distinction of speech using exact evidence in the form of objective evidence. Regularities and rules that are realized in the given stylistic manifestations at the macrostructural level as invariant regularities of the given system.

Not long after that, Doležel received an offer from Jiří Levý for another publication, this time in the prepared foreign anthology *Mathematik und Dichtung* (1965), for which he wrote the study *Zur sta-*

University of Stuttgart. Names associated with this important initiative, the influence of which was also evident in Czechoslovakia (see the translation of Bense's book into Czech and its publication in 1967, L. Doležel then reviewed it positively in the magazine Česká literatura), are the following names: Max Bense, author of visual poetry Georg Nees, Reinhard Döhl, Franz Mon et al. Outside the circle of this school, Umberto Eco, for example, dealt with these stimuli in an interesting way in the book *The Open Work* (1967).

tisticians Theorie der Dichtersprache. The anthology was edited by Helmut Kreuzer and Rul Gunzenhauser, who were associated with the Stuttgart School milieu in the 1960s. And other methodologically similar studies followed in the subsequent years. In the same year that Doležel's German study was published, he published another article in the magazine *Slovo a slovesnost* [*The Word and Verbosity*] under the title *Model stylistické složky jazykového kódování* [*Model of the stylistic component of language coding*] (1965) and the text entitled *Kybernetika a jazykověda* [*Cybernetics and Linguistics*] (1965) in the collection *Kybernetika ve společenských vědách* [*Cybernetics in the Social Sciences*]. The aforementioned publications were published shortly after Doležel became a member of the Department of Mathematical Linguistics at the Institute for the Czech Language of the Czechoslovak Academy of Sciences, which was also established in 1962 on his initiative. As a scientific editor and translator, Doležel participated in publishing of translated articles on mathematical linguistics, information theory and cybernetics in linguistics under the title *Teorie informace a jazykověda* [*Information Theory and Linguistics] (1964).*

At this point, however, I would like to mention in particular the study *Pražská škola a statistická teorie básnického jazyka* [*The Prague School and the Statistical Theory of Poetic Language*] from 1965, which Doležel published in the magazine *Česká literatura* [*Czech Literature*]. In it, he presented a more general model of statistical analysis of texts, which enables texts to be classified according to their stylistic characteristics. For the needs of such analysis, Doležel starts from several categories that reflect the stylistic properties of the text with regard to the possibilities of expressing these properties quantitatively, or statistically. As a result, the goal was to convert the obtained values into a binary code for the needs of further machine processing.

Specifically, there are six categories or characteristics: The so-called M-characteristics and B-characteristics relate to the issue of standardization and updating of language expression. Doležel defines them as follows:

> A communication standard is an average language the stylistic characteristics of which are estimated by averaging the characteristics obtained from non-artistic text selections. We will therefore distinguish the characteristics of the average communicative language (communication standard), which we will label M-characteristics, and the characteristics that statistically significantly deviate from the average – B-characteristics (Doležel 1965d, p. 106).

O-characteristics and E-characteristics define the degree of constancy and text variability. And again in the words of Lubomír Doležel,

> A stylistic characteristic will be called objective (O-characteristic) if its values remain statistically constant in the entire set of texts that is assigned to the communication circle of the given language. [...] A stylistic characteristic will be called subjective (E-characteristic) if its values in the set of texts of a certain communication circle show significant differences, but remain statistically constant in the set of texts of a certain speaker (Doležel 1965d, p. 106).

The last two are the S-characteristic and the N-characteristic indexing the stationarity and non-stationarity of the text, while „A stylistic characteristic will be called stationary if it satisfies the known conditions of stationarity in its time course: $M\,O(t_1) = M\,O(t_2) = M\,O(t_3) = \ldots = \mu$" (Doležel 1965d, p. 106). Texts that bear the signs of M-characteristics (average values, correspond to the communication standard), O-characteristics (stylistic objectivity) and S-characteristics (stationarity) are standardized texts in the given communication circuit, set of texts and vice versa (see Tab 1).[5]

Tab. 1: Doležel's example of binary coding of text according to individual characteristics. In this case, it is an example of two extreme situations. A = average value, B = objectivity, C = stationarity. 0 = statistically insignificant, 1 = statistically significant (Doležel 1965d, p. 107).

	A	**B**	**C**
standardised language	0	0	0
updated language	1	1	1

[5] Doležel adds to this period: „So I proceeded in the spirit of multifunctional linguistics of the Prague School, but at the same time I went beyond its methodological framework when I proposed to study functional differences in language communication statistically. I even assumed that by thoroughly measuring

The above mentioned study is also interesting as in it Doležel spoke in general about the Czechoslovak tradition in the application of quantitative methods in literary studies. Understandably, the author refers to the interwar Prague structural school that in its approach to these methods was de facto pre-statistic as Doležel puts it, since „the real apparatus of mathematical statistics was not applied here either in the definition or in the analysis of poetic language. The main theorems of the theory are statistical in content, not in formulation." (Doležel 1965, p. 104)[6]

2.

Although after 1968 Doležel's personal and professional path already took a different direction, he still spoke about the issues of quantitative and statistical methods in linguistics in the 1980s, when he responded to Alvaro Ellegård's article *Genre style, individual styles, and authorship identification* delivered at the 52nd Nobel Symposium in the year 1982. From Doležel's reaction, I select the following part in particular, which can be considered absolutely fundamental, especially from today's perspective of one of the areas of digital humanities (DH), which is focused on quantitative and statistical methods in literary studies. I consider these words of his to be timeless, as they accurately express the basic meaning and goal of any literary scholarship that deals with quantitative and statistical methods, which is currently es-

representative samples of the communicative bond, the language norm could be affected and then the poetic language could be studied as a deviation from the statistical characteristics of this population." (Doležel 2013, p. 112)

[6] It is also worth mentioning what Doležel states in connection with his paper devoted to issues of mathematical linguistics and statistics, which he presented at the Slavistic Congress in Sofia in 1964. In his memoirs, Doležel recalls the reluctance shown by Jan Mukařovský towards his paper, which according to in Doležel's words, in the 1960s, he considered mathematical methods in linguistics – and we can add that also in literary science – to be a relic. This dismissive attitude was apparently the result of Mukařovský's self-criticism and renunciation of structuralism in the early 1950s. (cf. Doležel 2013, pp. 89, 116–117)

pecially true in the field of DH oriented towards quantitative and corpus methods and tools in literary scholarship:

> We can master and use the most sophisticated statistical and probabilistic techniques, but this fancy equipment will continue to yield dubious results if the epistemological goals and theoretical foundations of quantitative text theory remain vague or primitive. **It is especially imperative to clarify the relationship between qualitative statements and quantitative statements about textual phenomena**. **To put this task in operational terms, we have to develop carefully controlled procedure for moving from qualitative to quantitative text descriptions and vice versa.** In the short time which is alloted to me I cannot do more than to outline briefly the problems connected with two such strategies (both of which are generally known in quantitative investigations), namely SCALING and INTERPRETATION. **The first procedure can be characterized as transformation of qualitative properties into quantitative data, while the second one is conversion of numerical data into structural description** These two strategies are indispensable for any empirical theory; the neglect of their foundations in the text study is, in my option, a major cause of our difficulties, misconceptions and misunderstanding. […] There are, in principle **two possibilities of interpreting numerical data: in terms of qualitative properties and in terms of quantitative structures**. […] In a qualitative interpretation, the data are taken as INDICATORS (indices, symptoms), i.e. their values (or difference in values) are interpreted as signaling the presence of certain qualitative (formal) properties, relations or taxonomies (Doležel 1982, pp. 540–541, 543; bolds R. Z.).

The above stated formulations, or the knowledge they represent, is also one of the key ones for current research in the field of DH, which is focused on the application of quantitative methods and models in literary studies. In my opinion, the initial distrust or even rejection of such approaches (and not only) in Czech literary studies stemmed from a number of prejudices and misunderstandings. Although today these methods, as well as the material to which they can be applied, are at an incomparably more advanced stage of development than in the 1960s, what remains decisive and essential is the necessity to provide the output values of these methods with a relevant interpretation. Like Doležel, others were also aware of this apparent obviousness, including Pavel Vašák in the 1980s, who in the book *Metody určování autorství* [*Methods of determining authorship*] (1980) formulates this fact more than clearly:

When it comes to the relationship between mathematics and literary science (linguistics and other social science disciplines) at all, I do not believe that in the future there will be any mathematical literary science, similar to e.g. existing mathematical physics, biometrics, etc., in the end, even in these fields, it is necessary to give mathematical results an appropriate physical, biological, etc. interpretation, similarly, even the existing field of mathematical linguistics is not a branch of mathematics, but of linguistics (Vašák 1980, p. 51).

And with this final quote, I would also like to conclude a small glimse back at Lubomír Doležel and his contribution to contemporary quantitative[7] research in literary studies.

References

Teorie informace a jazykověda. (1964). Praha: Nakladatelství Československé akademie věd.

B e n s e, Max. (1967). *Teorie textů*. Praha: Odeon.

D o l e ž e l, Lubomír. (1963). *Předběžný odhad entropie* a *redundance psané češtiny*. *Slovo a slovesnost*, 24, pp. 165–175.

D o l e ž e l, Lubomír. (1965a). Zur statistischen Theorie der Dichtersprache. *Mathematik und Dichtung: Versuche zur Frage einer exakten Literaturwissenschaft.* München: Nymphenburger Verlagshandlung GmbH, pp. 275–294.

D o l e ž e l, Lubomír. (1965b). Model stylistické složky jazykového kódování. *Slovo a slovesnost* 26, pp. 223–235.

D o l e ž e l, Lubomír. (1965c). Kybernetika a jazykověda. *Kybernetika ve společenských vědách*. Praha: Nakladatelství Československé akademie věd, pp. 267–180.

D o l e ž e l, Lubomír. (1965d). Pražská škola a statistická teorie básnického jazyka. *Česká literatura* 13, pp. 101–113.

D o l e ž e l, Lubomír. (1982). Discussion of Alvar Ellegård's Paper ‚Genre Styles, Individual Styles, and Autorship Identification'. *Text Processing. Text Analysis and Generation. Text Typology and Atribution. Proceedings of Nobel Symposium 51.* Stockholm: Almqvist & Wiksell International, pp. 539–551.

D o l e ž e l, Lubomír. (2013). *Život s literaturou: vzpomínky a rozhovory*. Praha: Academia.

[7]Analogously, this also applies where, for example, quantitative methods are combined with cartographic methods, which is an example of literary cartography. Even in the case of cartographic mapping of, for example, fictional topography, this type of model must be given a literary interpretation.

E c o, Umberto. (2009) *Otevřené dílo*. Praha: Argo.

E l l e g å r d, Alvar. (1982). Genre Styles, Individual Styles, and Autorship Identification. Text Processing. *Text Analysis and Generation. Text Typology and Atribution. Proceedings of Nobel Symposium 51*. Stockholm: Almqvist & Wiksell International, pp. 519–537.

V a š á k, Pavel. (1980). *Metody určování autorství*. Praha: Academia.