HUMAN AFFAIRS, *8*, 1998, 2, 112–120

ON THE CORPUS OF LITERARY WORKS

Eduard KOSTOLANSKÝ
Department of Applied Mathematics and Computer Science, Faculty of Natural Science,
University of Sts Cyrill and Methodius, Nám. J. Herdu 2, 917 01 Trnava, Slovakia

Information technology (IT) penetrates into all areas of human activity. The automated analysis of literary texts aimed at problem identification and problem solving in these texts is a unique challenge to IT. The success of IT in the analysis of the works of art would confirm IT to be a breakthrough in technology also in the direction of the integration of natural, technical, and artistic knowledge and their optimum use in the evolution of society. The complexity of computer modelling of literary texts and the demand for the formation of extensive corpora require concentrated efforts which might be coordinated, for example, by UNESCO.

1. Introduction

The drive of data communication networks based on the fascinating achievements of computers and telecommunications is always gaining new areas: the texts of literary works of art have not avoided it either. And why should they avoid it, rather the other way round. It seems that it would be necessary to tame, at least a little, the self-conscious lady named information technology which is the heart of the mentioned data communication networks through its significant involvement in the world of the texts of works of art. We shall try to show that IT should be the element in the evolution of society which integrates two main forces of evolution, namely technical and scientific knowledge on the one hand and the knowledge encoded in works of art on the other. Without reducing the strength of art, we let artistic knowledge be represented by the knowledge encoded in the literary works of art. In addition to the concentration on decoding the artistic knowledge by means of information technology it is our intent to highlight the art knowledge as priority knowledge in further steering of the evolutionary spiral of society.

There is no reason to doubt that information technology is a work of many cultures of human society coming one after the other. Arguments supporting this statement are found in each elementary textbook of information science [5]. Intuitively, IT can be estimated as a huge instrument. The information technology could con-

tribute to problem solving 'whether the meditation, continuous accumulation of knowledge and its use in various forms and the orientation of the milestones of human culture is correct'. In relation to belles-lettres we are primarily interested in the massiveness of IT, namely, its massiveness in the good sense of the word. The fact that it is no problem for computers to store all new works in the digital form and to analyse and process them from various angles or at least to mediate them should also be increasingly challenging to belles-lettres.

2. Redundancy of specialized publications

Let us make a short excursion to the world of information aimed at reaching a better understanding of the strong tie between belles-lettres and information technology. More than one million technical articles and around sixty thousand scientific journals are published worldwide every year and their number is still increasing. About 300 thousand scientific monographs are published yearly. The statistical data on the proceedings from scientific conferences, on the achievements in biochemistry, medicine, mathematical theories, etc., would follow the same trend. The representation of particular scientific disciplines in these statistics is different. But generally, to keep abreast of the new knowledge achieved in the particular discipline is a serious problem for scientific workers. Various ways of keeping up with it within reasonable limits are searched. One of the ways to achieve familiarity with crucial results in a particular field is based on providing the titles of the contributions published in significant publications. The authors of contributions and their institutions create a supplementary source of information for the accumulation of new knowledge.

However, there is an astonishing relation of redundancy between the creation of the specific knowledge and its use. It is best documented by the mathematics of citations. According to statistics, out of one million works written per year, at least one third and not more than half of the works have never been cited at all. Today it makes more than 25 million works. Are they really redundant articles? Is it true that nobody except for the authors and reviewers has read them? Have they not influenced anything? Have they not inspired anybody, in fact have they delayed and represented a lot of money thrown out of the window? Are they a necessary side product of the second half of publications, each of which has at least one citation? The answer to these questions is probably not easy. It is also because many results or initiated problems, are revitalized after several decades and then a boom of their citations begins.

It is not our aim to analyse the issues of scientific literature further. The preceding look at this literature and some statistical data pertaining to it should serve as argumentation in favour of the construction of the information aggregate targeted at belles-lettres.

3. Artistic texts – a principal element in the evolutionary spiral

We start from the assumption that belles-lettres is addressed to a wider circle of users compared to scientific literature. It is generally accepted that belles-lettres is a picture of the period, the reflection of the life and thought of humans as individuals and the human within a team. We keep to the standpoint that in the evolution of human race, positive steps should grow and violent, bloody, and catastrophic evolutionary elements should diminish in proportion; if such steps can at least be considered as evolutionary. Such an attitude calls for an assessment of all the steps of humankind realized on its evolutionary spiral. It is not that this has not been done so far. But the disability of the exact and technical part of the progress of human society to reduce the mentioned negative evolutionary steps forces us to change principally the attitude towards the role of belles-lettres as a component of human evolution.

Let us try to speak about belles-lettres using language close to the terminology of information technology. Each literary work accepts knowledge from the surrounding world, which is artistically transformed and offered in a new form. The knowledge encompassed in literary works of art has been the object of interest for centuries and it is the same in the age of computers. This is knowledge or information of another type in comparison with the exact knowledge of natural and technical sciences. While the latter knowledge is concerned with the identification of the procedures ruling in nature, literary works accumulate chiefly knowledge on the state of human society and the individual, on the reasons that cause transition from one state into another and on the ways of realizing such transitions. Of course, the knowledge is associated with different structuralizations of society and with various modules of society. Since the pattern of positive advancement of human society is not known and a search for it is part of the existence of the human race, it is a duty to use all means available which offer at least some hope of uncovering the forces and laws of human behaviour. We assume that an artistic text is not only the bearer of the knowledge on the behaviour of human society in particular conditions and in a certain time, but it is also a strong factor controlling the behaviour of society. The aim should be achievement of decoding, and structuring the knowledge of the artistic text in such a form as to be able to launch the definition of the knowledge on the states of human society and the knowledge of the processes of the transition from one state into another. Let us say that we have built up procedures of the identification of these states and the steps towards the transition from one state to another. Then the analysis of literary works of art could predict possible future states of society. The forces which feel competent to administer individual countries would have, in the analytical and synthetic results of artistic texts achieved by means of information technology, an important information source which would have to be taken into account in their decision-making.

This is an ideal situation. But before we speak about such a significant computer application in the area of artistic texts, a lot of work has to be done. For instance, the

definition of the states of human society requires a demanding study so as to have a certain constructive device for their identification and decoding in artistic texts. Which states will be at issue? It will definitely be the states of good and evil, truth and untruth. And others? What will be the relations between these states? Definitely the states of addiction, subjection, mutual destruction. And which others? What is the precondition for the transition from one state into another? To what extent can we speak about certain forms of patterns for the 'calculation' of the states and relations between the states? These interesting questions are not new, but information technology offers means which enable contributions to their solution from new starting points. It should be emphasized that the problems are difficult and that even a partial result would be a success. This contribution is in principle an agitation for a solution to these problems. The courage for agitation is found in the conviction that problems introduced and solved by literary works of art have at least such a significance for human survival as the problems advanced and solved in the exact sciences. We thus ascribe an increasing importance in the development of the human spirit to belles-lettres.

Information technology represents in the first place the integration of scientific and technical knowledge. But it seems that the mentioned massiveness of IT also stimulates its integration with belles-lettres. The ambitions of IT as a technology, the core of which is the processing of knowledge, are also to identify and process the knowledge encoded in the texts of belles-lettres. Identifying and processing such knowledge could be shown to be a measure of the maturity of IT. On the other hand, one should reckon with the level of unique artistic knowledge in the texts of belles-lettres, the interpretation of which will be the domain of man. The outcome of the cooperation between IT and the arts should be a more clear-cut presentation of the (artistic) knowledge of literary texts and a promotion of their roles in everyday life. Our attitude is that such an integration of IT and belles-lettres will supply the administration of human society with new knowledge and simultaneously will ensure a priority role of artistic knowledge in the evolution of society. We start from the fact that knowledge and use of the known is the essential and natural human activity. The recognition of the strength of knowledge born from artistic texts will rank this knowledge among the primary knowledge used in directing the evolutionary spiral of society.

Problems formulated within the context of the literary works of art can have much more forms in comparison with the phenomenon of multiformity of the problems in the exact disciplines. The same situation arises in looking for and formulating problem-solving.

4. Corpora of artistic texts and targeting

It is useful to interpret this multiformity in belles-lettres as its targeting. Each book, each problem and its solution, metaphors or paraphrases have its addressees. However, the question how they reach the addressees remains open. Even if they

reach them, these are not able to read them. The extent of information available as literary works of art is large and it is beyond their capability to read all those books. The situation is analogous to that in scientific literature: time enforces us to select from all addressed books and read only those most valuable to us. We are able to formulate a selective criterion so that it can delimit a very small group of works we read. But what about other literary works of art? Is their fate similar to the fate of the majority of scientific publications? What have the evaluation schemes of scientific works and of literary works of art in common?

An opinion emerges intuitively that belles-lettres is richer in its essence. However, we do not want to centre on the measurement of the complexity of problems raised by science and belles-lettres. The fact at issue is to provide access to all formulations of and solutions to problems contained in the literary works of art for users/readers. It definitely is a complex problem.

We put aside the question of how to cope with the works from previous periods. However, the entire wealth of current belles-lettres can be made accessible in an up-to-date way through information technology. Let us remember that this technology based on computers and telecommunications enables processing, remembrance, and transfer of enormous information sets and thus also texts. Technically, there is nothing in the way to launching the creation of the text computer collection – a corpus of all literary works of art of particular national cultures. In addition to the primary goal – the use of knowledge conveyed through artistic writings in influencing the evolutionary spiral of society, which we pursue by introducing IT into the world of belles-lettres; the secondary aim has its adequate weight as well: to make the literary works of art accessible to as wide a public as possible. This will lead to increased influence for the works of art. Apparently, the introduction of literary works as corpora can optimize the access of the ordinary man to the wealth of ideas and contribute to an increase of the users' interest in these works. It is, for example, realistic to imagine the "reading" of all titles of a particular national belles-lettres (let us say, published in one year) with the aim to learn or enjoy the sections describing relations between mothers and children. The electronic form, for example, of the yearly production of belles-lettres would enable us to read the part (parts) of individual works, in which such a relationship is described. The part valued most by the readers/users should probably result in their deeper interest in the particular work. Moreover, the assumed, still developing litware (the term will be explained later) might enable the reader to add his/her view to the artistic text to enrich the original idea. Similarly, the reader could learn about the views of other readers of the text.

All this is based on the existence of the electronic form of literary texts. Access to it is thus a crucial question. The copyright and all matters associated with it are in the background. Maybe we are wrong, but we think there is no ban on any work of ancient literature being made accessible in the electronic form. We know a number of contemporary writers, highly appreciated and read, who give their

works as part of the computer corpus without any conditions. Ultimately, on deeper insight into the issue of the text corpus of the literary works of art, the question of managing royalties emerges.

We grasp at ancient and Renaissance literature and other works in order to learn and to contribute to our maturation. Since each literary work contains, figuratively speaking, at least one sentence worth publishing for the public, we are obliged to do so. Maybe time will introduce it as a decisive sentence or idea. If we reach for old masters of the literary works of art in order to relax, realizing that they also had to solve many problems, which we still run into, and to feel encouraged by brilliant, careful, compassionate, treacherous, sinful, penitent and other steps, with which they armed their heroes – problem solvers of the human soul and human society, it is our duty to get acquainted with the problems of today and procedures or attempts at their resolution submitted by contemporary masters of the pen. It appears topical, possible, reasonable, and efficient to set out to create an electronic corpus of literary works of art in individual countries. A corporate offensive should not have to be necessarily launched at full length, but it is necessary to approach the introduction of IT into belles-lettres systematically. We should say here that, with regard to the foregoing description of the problems, it would be acceptable to start speaking about this area as about literary informatics. The first problem of literary informatics is the problem of building up text corpora. The construction of such a corpus might be launched by the digitalization of the works which won the first prizes or other awards in various national competitions. Information on awards is also significant since it speaks about a certain hierarchy of values of a particular period.

5. IT vehicles for the domain of belles-letters

Literary science and other humanities interpret problems and their solutions which are artistically encoded in the particular literary works of art. The user/reader is a decisive and primary interpreter of works. In this article we want to adhere to a constructive introduction of problems, problem solving and formation of problems within literary works.

There are computer programs allowing creation of subcorpora from the main corpus, for example, a subcorpus of the works written by one author, subcorpus of socially-oriented works, subcorpus of love works, etc. Other programs extend their original text of corpus by grammatical information (e.g. the words are supplemented by the data on their part of speech, case, person, or function in a clause, etc.). Yet other programs could enrich the text of the work with the information which might be called interpretation of the original text from different logical levels.

We can expect, and this should be the core advantage of IT, that the analysis of the literary works of art will assert itself more significantly from the perspective of

various logics. And this is the quintessence of a constructive look at the problems introduced by and solved in the literary works of art. It is expected intuitively that it makes a sense to think about the logic of morals, religion, justice, about political logic, etc., and about the formalization of these logics. (The word logic might be replaced by the word arithmetic and we would talk about the arithmetic of morals, religion, etc.) Of course, this is associated with the definition of the basic concepts of these logics, their rules, values, etc. The products of literature can be interpreted in the light of these logics. The paraphrases of problems, ways of their solution, contribution to education and progress, or to the formation, deepening, or decay of human relations can be identified. In short, all components that are associated with the activities of the human spirit and are shaped by belles-lettres can be made precise, presented and better known and some can be newly discovered through the instruments of IT. Its advantage is to see exactly, as if mathematically, the problems encoded in the literary works of art and to see them as silhouettes in different levels of literature and in different levels of details in focusing on a particular work. But what is countable, should be counted, as soon as possible.

The particular technology for modelling and evaluating artistic texts will develop in parallel with the spread of the corpora. The essential idea of this technology is the modelling of the texts of the works of art as data structures which can be processed by computer. Processing will incorporate transformation of the text form of the work into its structured data form, identification of the ways (mainly morphological and syntactic) of the expression of the content, and identification of the content of the work. One can also reckon with the fact that the computerizable part of these processing methods will be specific and realized in the form of computer programs it will create a special software, which can be justifiably named literary software or abridged: litware.

It is realistic to assume the possibility of constructing the time sequence for the formation of particular problems depicted in the works, ways of their solutions, impacts of these solutions on life, up to the formation of a negative situation in the evolutionary spiral of society. We start from the existence of laws for literary works of art, which establish that belles-lettres is a significant agent in the evolutionary spiral of society and as such, it has to be in close relation to society. Information technology should help reveal at least partly this relation and use its knowledge in influencing the development of society.

6. Steps for the near future

The massiveness and availability of IT almost inspires its massive use in bringing the literary works of art closer to the reader. This computer-assisted way of making the literary works accessible can be implemented at different levels. In this paper we want to support the international level of the computer presentation of lit-

erary works. If the attitude presented here finds supporters, the first step in the construction of the international corpus of literary works of art and its analysis would be a challenge advanced to an appropriate international organization, e.g. UNESCO. This would be a continuation of the UNESCO programme focused on the creation of a corpus of representative works. The following step might be the preparation of a project for the creation of such a corpus under the auspices of UNESCO. The analytical part of the project proposal would assemble the present knowledge on the approaches to these problems followed by coordinated efforts to solve the projects of literary information science. If, for any reasons, the projects would not be formulated under the auspices of UNESCO organizations, small cultures would seemingly be forced to coordinate and put the means for resolutions of the projects of literary information science together. We think that the challenge for such projects might be implemented through the journal Human Affairs.

7. Conclusions

We can look at the described situation as a symptom of the period in which it will be possible to record every idea considered as essential or new, or, simply interesting by its author. We tried to argue in favour of information technology within the context of belles-lettres. We are convinced that information technology will bring revolutionary knowledge also in this area. The core of this knowledge should be the contribution to the modelling of the relations between the progress in technology and the progress of the human spirit. A vital precondition for such an approach is, however, creation of comprehensive literary corpora. That is why in addition to argumentation in favour of the adoption of information technology in preserving and analysing literary works, this contribution is a challenge to all authors and competent institutions and organizations to show good will and efforts towards creation of the model of belles-lettres that could be interpreted and evaluated by computer. The first unavoidable step is the construction of the corpora of literary works of art. We think that international organizations and institutions, such as UNESCO, might be helpful in organizing and preparing projects of literary information science. The building of the constructive models of artistic texts is to some extent a challenge to information science and mathematics. It does not start on the green meadow, however. Language inspired already the first computer designers and efforts endure to equip computers with communication instruments with the quality comparable to the communication among people. The achievements of the works in this direction result from computer modelling and language processing. Some methodological procedures, formal models and applications, the results of which can be used in the projects of literary information science were worked out within computer linguistics, where [2] artificial intelligence [4], development of information systems [3], software [6], and applications [1] can serve as examples.

The heart of the issue, which we have tried to elucidate in this paper, is the formation of the constructive models of artistic texts, which would be close to the intuitive vision of the artistic text and to present constructive procedures in literary science. In other words, we have to present information and procedures or their parts used in the analysis, interpretation, classification and assessment of artistic texts in the form of data structures and computer programs. It will be the base which can be further developed. Or it will be a source of impulses on how to proceed in order to achieve a base of new quality which would guarantee modelling of artistic texts supported by the mentioned logical calculuses oriented to operations like: a sum of texts, reduction of texts, identification of the core of the text, interpretation of the text, arrangement of the texts, etc. Our hypothesis is that such models of artistic text can be created. The hypothesis is based on the assumption that information technology is a great device with which we can launch the integration of the technical and knowledge stream of evolution on the one hand, with the artistic and religious evolutionary current on the other. In an extreme case, the outcome of these integration works will be a radical opening of the artistic text to ordinary man and people making decisions at various levels of administration. Consequently, the influence of the knowledge and intuition, which are encoded in artistic texts, will increase.

An immense processing strength of information technology is challenging us to make a wide coverage in the construction of artistic digitalized text sources. To reach the expected aim, a coordinated and systematic approach is necessary. It might be a unique opportunity for UNESCO. Along with the creation of the computer form of literary works, formation of models of the data structures and procedures of the processing of literary works of art will be deepened.

REFERENCES

[1] Copeland, C. – J. Durand – S. Krauwer – B. Maegaard (eds.): *Studies in Machine Translation and Natural Language Processing.* Vol. 1. *The EUROTRA Linguistic Specifications. Office for Office.* Publications of the CEC, 1991, Luxembourg.

[2] Bookman, L.: *Trajectories through Knowledge Space: A Dynamic Framework for Machine Comprehension.* Kluwer Academic Publ., 1994, Boston.

[3] Kemper, A. – G. Moerkotte: *Object-Oriented Information Management for Advanced Applications.* Prentice Hall, Englewood Cliffs, 1993, U.S.A.

[4] Kunz, J.: *Vererbung für Systementwickler.* Grundlagen und Anwendungen. Vieweg, 1995, Wiesbaden.

[5] Sanders, D.: *Computer concepts and applications with BASIC.* McGraw-Hill Co., 1987, Hamburg.

[6] Stroustrup, B.: *Die C++ – Programmiersprache.* Addison-Wesley Pu., 1986, München.