Latent Semantic Analysis for Notional Structures Investigation

Abstract. The research on the effects of study is hindered by the limitations of the techniques and methods of registering, measuring and assessing the actually formed knowledge. The problem has been solved using latent semantic analysis for comparison and assessment of scientific texts and knowledge, expressed in the form of free verbal statements. Education at higher schools has the specific objective to develop knowledge and experience both of which have two fundamental dimensions: the first is expertise training in a well-defined occupational or disciplinary domain, and the second — learning strategies and skills to be an effective learner. Various trends for stimulation of deep learning, transferring in practice the achievements of the cognitive psychology, have been developed during the last decade. Here we present a research on the cognitive activity of university students and its results in the dimension of declarative knowledge. In practice a comparative analysis is made between the input system of notions from the learning texts and the formed mental structures of the students. The research includes a sequence of actions and procedures for: facilitation of the formation of stable concepts structures (preparation of learning materials, its content, structure and visual presentation, organisation of learning, etc.); feedback output on the preservation of knowledge of certain number of key notions; and assessment of manifested knowledge. The data used is verbal - learning texts, linguistic descriptions of notions contained in them and all these are rendered in an open format by the people observed while posing indirect questions. The nature of the processed material (input stimuli and preserved knowledge), decided on the application of Latent Semantic Analysis (LSA) as a research method on the information data. This statistical technology permitted the formation of a model of semantic connections between the researched notions in the output and the general representation of the results.

Latent Semantic Analysis for Notional Structures Investigation

Abstract. *The paper presents a comparison method for input system of notions from the learning texts and the formed mental structures of students based on latent semantic analysis. The data used is verbal – learning texts, linguistic descriptions of notions contained in them and all these are rendered in an open format by respondents while posing indirect questions. This statistical technology permitted the formation of a model of semantic connections between the researched notions in the output space against whose background is made an assessment of the individual achievements and the general representation of the results.*

1. Introduction

The main idea behind the modern learning strategies is that the development of the educational environment should be oriented towards the creation of conditions which should provide: stability of the cognitive structures and steadiness of the system of knowledge contained in them; flexible implementation of well structured knowledge bases; development of metacognitive skills, of heuristic methods and strategies of analyses of the problems as well as transformation.

This directs the research towards the depths of the learning processes and information acquisition and more specifically defines the interdisciplinary character of this kind of studies. We present here the results of the study of the notional structures and their acquisition in the declarative memory based on the *latent semantic analysis (LSA).* The experiment was conducted with students from the University of Chemical Technology and Metallurgy (UCTM) during the actual learning process.

In order to characterize the level of learning and the degree of structuring the system of notions, the assessment should be based on a more detailed system of criteria with the relevant interpretation indices. With respect to this two basic criteria for semantic structuring with their relevant indices were adopted as follows:

First Criterion – Knowledge of the notions' content

Indices: restricted or incorrect rendition of the content of the notion; accurate and exhaustive definition of the notion; knowledge of the relation of the notion to other notions; application of the attributes of the notion

Second Criterion – mastering the connections between notions.

Indices: establishing the existing connections; recognition of the degree of subordination; anticipation of subsequent connections.

The recording of the assessment data and especially their processing to a compatible form is impossible to do with the standard methods of assessment and knowledge. The presence of LSA as a method of processing and analysis of the verbal information permits the research of these phenomena.

2. Latent Semantic Analysis (LSA) – description of the method

LSA is a powerful statistical technique for indexing, extracting and analyzing of textual information, which has been applied successfully in different spheres of human cognition during the last decade, LSA (1990-2000). The method is completely streamlined and does not use any preliminary compiled dictionaries, semantic nets, knowledge databases, conceptual hierarchies, grammatical, morphological or syntactic analyzers, etc. The main idea is that there exists a set of mutual dependecies (implying mutual limitations) between the separate words and the generalized context (sentences, paragraphs and texts) in which they appear. Their uncovering and correct processing permits LSA to cope successfully with synonymy and to a certain extent with polysemy. Laudauer, Foltz, Laham, (1990).

LSA is a two-stage process including education and analysis of indexed data. During the education stage LSA conducts an automatic indexing of documents. The process begins with the building of a matrix X whose columns correspond to the separate documents and its rows — to words. The frequency of appearance of word i in document j is recorded in position (i, j). Built in that way the matrix X is decomposed as the product of three matrixes D, T (orthogonal) and S (diagonal) so that $X=DST^t$. Then follows the deletion of the major part of the columns and rows of D, S and T so that matrix $X'=D'S'T'^t$ resulting from their multiplication is least squares best-fit approximation. Thus, a considerable degree of compression of the input space is achieved which results in small number of meaningful factors (usually between 50 and 400). Therefore a vector of low dimensionality, for example 100, represents each document and each word.

Analysis of the data while using the index is done during the second stage. The degree of closeness between a pair of documents, pair of words or a word and a document is the most frequent object of research. A vector corresponding to the text, not belonging to the indexed documents can be obtained using simple mathematical transformation. Other appropriate measures such as Euclid's and Manhattan distance, Minkowski's measure, Pearson's coefficient and others may also be used.

This type of analysis is an object of research and at the same time it is extensively applied for discovering semantic connections, Wolfe, Schreiner and others (1998), Foltz, Kintsch, Landauer (1998), Glenberg & Robertson (2000). Its possibilities were assessed in a comparative analysis of the knowledge of the respondents. Because LSA is a method in which information is initially input (figuratively speaking – the system is educated) and subsequently is researched and assessed as a structure, it is tested in the circumstances of real life assessment procedures and an environment close to the reactions of the

respondents. A significant compatibility in the reactions of LSA and the students is discernible in texts for reverse reproduction.

In the scientific sphere research with LSA has been made on learning from texts, which aim is the solving of the following tasks: to disclose to what extent or how the knowledge in a certain sphere can be successfully measured; to define the empirical difference between the prior knowledge of the students by using conventional methods which influence their ability to learn from different texts; to define how efficiently LSA provides the opportunity to outline the existence of a background of prior knowledge and to anticipate the acquired knowledge.

Despite its exceptional possibilities LSA poses certain questions related to its application. The method is directly related to the lexical idiosyncrasies of the language and they should be allowed for in the development of computerized programmes for each language Another major issue is the so called problem of symbols. Different solutions have been considered in this aspect. The first one depends on perception and is related to the recognition of symbols as part of the acquired language. This, in its turn, defines the aspect of learning which leads to the building of schemes of cognitive abstract and arbitrary symbols with specific relations both to objects and among themselves. In some of the cases it is possible to show the realization of these connotations and their comprehension in certain situations. But still the question of how a parallel reflection of all the symbols and their separating and decoding is achieved in a certain situation considering the variation of its sensory and situational peculiarities remains to be answered. Here the answers lies in the plain of logical operations and relations. We can view the system of symbol managing as resulting from the (product of) logical processing or some calculations, i.e. similarly to how LSA defines correspondence through vectors and their cosines. But how the system can recognize what the symbol signifies, i.e how it considers itself? The decision is clearly again in the building of a scheme of cognitive symbols based on their relation to real life or in the searching for abstract descriptions transferable to objects through systematic symbolic. The problem of specific language acquisition in science and the application of symbols within the context of different domains or problems of knowledge is being resolved in this aspect.

The second decision offered by Landauer and Dumais (1997) reduces the problem to abstract systems such as the mathematical expressions. Their suggestion is to decode the stream of words and the other sensory modulations en mass in order to achieve layering of additional information which may be juxtaposed and assessed.

One of the tenets of the symbol based problem suggests a decision neglecting the condition that meaning and connotations are based on abstract symbols in their relations to the objects they signify.

The method of LSA has been developed for the Bulgarian language and has shown its possibilities in a series of research studies, Nakov (2000 – 1, 2000 – 2 and 2000 – 3) and has been accompanied by the overcoming of lexical difficulties bearing extremely heavily on our language.

In principle the application of LSA is connected with the solving of certain problems of the natural language part of which are implicit in the nature of the method while others, depending on the specific research case, are overcome only within the researched volume of textual information .

An important stage in the work with LSA is the transformation to a basic form of all the words in a text. It is of paramount importance to consider the forms of one and the same word as identical in the construction of the initial matrix *X* which represents the semantic space. This is especially pertinent for a strongly inflexional language like Bulgarian. There are specialized products for this purpose (primarily for English and for other popular European languages) but they were not used in this particular case for several reasons. Typical of these tools is that they work with approximated precision which leads to the possibility of noise interference and deviation of the results of the experiment. Apart from this and despite the fact that similar research has been made such software product is not offered for the Bulgarian language at the moment.

3. Research method

Learning texts, whose structure is subjected to the models of semantic memory, are developed. The texts are assigned to students as individual tasks in paper and as a website. After the cognition of the scientific information students have the possibility to apply the acquired information to series of questions and then they proceeds to the reproduction procedure

The method of the interview has established itself as a feedback source. A series of questions specifically directed to the content of the research notion have been devised. The primary objective in their wording was to trigger a response in an indirect way i.e. not to prompt a direct reproduction. For the content of several of the notions which exist in an explicitly expressed gender-class relation this proved impossible and there direct questions and answers about classifying attributes were obtained, a fact that enhanced the image of the received data.

The interview as a procedure was conducted individually by the researcher in the absence of the lecturer in the specific subject on the explicit condition that it has no bearing on the examination results and will not be included in the final mark.

4. Experiment

Quantive expressions of the theoretical definitions of notions and those supplied by the students are found with the help of LSA. Thus an image illustrating the relation between the theoretical definitions and the students' assertions expressed in their answers to the questions is projected.

On the other hand the closeness between all possible pairs of theoretical definitions of notions and their corresponding pairs supplied by the students is calculated. In other words here we measure to what extent the theoretical closeness of notions *a* and *b* corresponds to their closeness according to the students' answers. Thus an image illustrating the relation between pairs of theoretical definitions and their corresponding pairs of students' assertions expressed in their answers to the questions is projected.

Fig. 1 represents a graphic "map" of the closeness of 16 key notions: according to the theoretical definitions and according to the knowledge of two of the students. The darker colour corresponds to a greater degree of closeness.

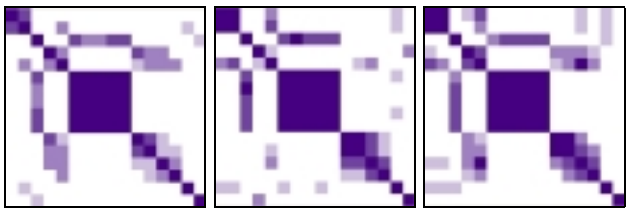

Fig.1. *Closeness of 16 key notions: according to the theoretical definitions (right) and according to two different students*

Table 1 represent a generalized data about 25 students participating in the research. It gives the opportunity to compare the modules of difference of the matrix corresponding to each student in relation to the original correlation matrix. The table represents generalized data according to questions and the average as well as the mean value of digression according to theoretical data. Each column of the table is divided in two and shows the digression for both the notions and the relations between two pairs of notions.

The presented values construct a basis for comparison and analysis of the individual achievements in subsequent tasks for applying this volume of declarative knowledge. As a whole they demonstrate the

quality of knowledge according to the defined criteria most significant to our experiment, which may be summarized as follows:

- High degree of adequacy of the notions and the relations kept in the memory. This is illustrated by the mean digression between the individual and the theoretically defined data, which is 0.121 for the relations between pairs of notions and 0.143 — for separate notions.

- Strongly expressed dependence of notions existing in gender-class relation. This is revealed in the research of theoretical data and those of the respondents.

- More strongly consolidated scheme compared to the knowledge of particular notions, which can be seen from the lower degree of digression for all students in the pairs of notions: 0.085 against 0.129 for separate notions. This points to a differentiation both in the level of knowledge acquisition and in their character.

- Closer results for the respondents for notions as a structure. This is clearly seen in the gap between the individual achievements of the respondents. In the results for pairs of notions it is less than the for the separate notions. Therefore the acquired knowledge can be described as a system reflecting the logical connections between notional knots, and the system itself – as equal to the theoretically defined one.

Student	**Average**		**Mean value**	
	Couple	**Single notion**	**Couple**	**Single notion**
1	0.127	0.158	0.084	0.156
2	0.105	0.121	0.076	0.139
3	0.096	0.107	0.062	0.112
4	0.107	0.142	0.073	0.121
5	0.137	0.145	0.105	0.138
6	0.113	0.118	0.085	0.121
7	0.110	0.115	0.077	0.109
8	0.100	0.109	0.077	0.113
9	0.113	0.173	0.082	0.091
10	0.115	0.156	0.084	0.120
11	0.148	0.179	0.124	0.197
12	0.112	0.105	0.078	0.103
13	0.103	0.126	0.056	0.133
14	0.146	0.174	0.108	0.152
15	0.138	0.157	0.094	0.147
16	0.113	0.116	0.084	0.104
17	0.120	0.128	0.079	0.103
18	0.110	0.137	0.076	0.117
19	0.116	0.173	0.083	0.098
20	0.132	0.164	0.084	0.125
21	0.149	0.172	0.114	0.180
22	0.130	0.142	0.085	0.124
23	0.105	0.120	0.067	0.121
24	0.148	0.187	0.102	0.153
25	0.138	0.152	0.097	0.148
Average	**0.121**	**0.143**	**0.085**	**0.129**

Table. 1. *Module of the difference between the theoretical definitions and those supplied by the students*

5. Discussion

The principal conclusion of the conducted research study is about the place of LSA in the method of processing verbal material; working with natural language increases considerably the possibilities for measuring and analysis of the effects of learning.

With respect to the task presented here – assessment of the notional structures – the applied method yields unquestionably good results and they can be viewed in two aspects — the efficiency of LSA as a method and the quality of the research method and procedures.

The basic conclusion about the efficiency of the educational components of the model around which is organized the learning of the researched educational content is that it gives results in the direction anticipated by us. This is also supported by the data from the examination of the values of closeness of notions as a general structure, which are higher than those for closeness between the separate notions (according to the theoretical and experimental data). On one hand this is the result of the possibilities of LSA to uncover latent connections and on the other the physical characteristics of the notion structures built by the students. It is from the object of cognition — specialized scientific information for which logical relations with knowledge presented in the past are typical - that different strategies for building constructions of notions are discovered. In practice the data about the structures of the students show the relation between the existing knowledge (built positions) and the perception of incoming information but the effects of the support of the comprehension of the inner logical structure of the input information are also added. The fact that students make a description, which discloses the cause and effect relations beyond the definitions, is indicative of the degree of stability of their own structures as a logically founded constructions. This should be the basis of a subsequent effective learning and solving of problems from specific spheres.

6. Prospective studies

The obtained results about the degree of closeness between the structure of notions in the learning texts and that of the students is a material for various future research studies and analysis both with respect to the construction of a learning environment and in the area of procedures for selection and processing of the results of learning and the ways of applying LSA in the methods of scientific experiments.

7. Literature

Berry M., Do T., O'Brien G., Krishna V., and Varadhan S. (1993), SVDPACKC (Version 1.0) User's Guide. April.

Deerwester S., Dumais S., Furnas G., Laundauer T., Harshman R., Indexing by Latent Semantic Analysis, Journal of the American Society for Information Sciences. 1990. 41, pp.391-47.

Laudauer, T., Foltz, P., Laham, D., Introduction to Latent Semantic Analysis. Discourse Processes, 1990, 25, pp. 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E., How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto & P. Langley (Eds.), Proceedings of the 19th annual meeting of the Cognitive Science Society 1997, (pp. 412-417). Mawhwah, NJ: Erlbaum.

LSA (1990-2000), see http://lsa.colorado.edu

Nakov P. (2000-1) Getting Better Results With Latent Semantic Indexing. In Proceedings of ESSLLI'2000, Birmingham, 2000.

Nakov P. (2000-1) Getting Better Results with Latent Semantic Indexing. In Proceedings of the Students Prenetations at ESSLLI-2000, pp. 156-166, Birmingham, UK, August 2000.

Nakov P. (2000-2) Latent Semantic Analysis of Textual Data. In Proceedings of CompSysTech'2000, Sofia, Bulgaria. June 2000.

Nakov P. (2000-3) Web Personalization Using Extended Boolean Operations with Latent Semantic Indexing. In Lecture Notes in Artificial Intelligence — 1904 (Springer). Artificial Intelligence: Methodology, Systems and Applications. 9th International Conference, AIMSA 2000, Varna, Bulgaria, September 2000.

Paskaleva E., Nakov P., Angelova G., Racheva P., Mateev P.: Matching Text Meaning for Information Retrieval Tasks in Inflexional Languages, Fifth TELRI European Seminar in Corpus Linguistics: How to Extract Meaning from Corpora, Ljubljana, Slovenia, September 22-24, 2000.