Investigating the Degree of Adequacy of the Relations in the Concept Structure of Students using the Method of Latent Semantic Analysis

Senia Petrova Terzieva, Preslav Ivanov Nakov, Sneja Handjieva

Abstract. *The research on the effects of study is hindered by the possibilities of the techniques and methods of registering, measuring and assessing the actually formed knowledge as information represented in the memory with the appropriate correlation among its units. The problem has been solved by the use of the latent semantic analysis for comparison and assessment of scientific texts and knowledge, expressed by the students in the form of free verbal statements.*

Key words: *latent semantic analysis; notional structures, content analysis.*

INTRODUCTION

The researches on the results of study and knowledge acquisition in the modern theory of education are based on the interdisciplinary approaches, methods and procedures. Specialised research techniques for registering and analysis of the effects of study have been developed in this respect as well as maximum accurate and objective ways of measurement; formalisation, processing and analysis of the achievements have been sought.

A very significant feature of knowledge as an object of research is the verbal form, which is predominant. This fact poses a number of limitations of the research on the information about the degree of acquisition and the study of the levels of processing are considerably impeded during the analysis of the results of the achievements that are registered in the traditional methods of control. A major step towards the overcoming of this problem is the use of latent semantic analysis of the survey in the process of study and knowledge building [2,3].

The role of the "scheme" of the cognitive structures in the processes of information perception and processing as well as decision making have been derived from the ideas of the cognitive paradigm in the theory of study. That is why in the appraisal of the effects of study the built notional structures and their features are the starting point. The most significant criterion of adequacy in this particular case is the correspondence of the logical correlation between notions or information units in the input information.

In the present research we aim at the establishing of the level of processing of the study information under the conditions of specific organisation of the content of one specialised technological subject from the course in Technology of Organic Synthesis at the UCTM.

METHOD

Web format educational texts have been developed and structured as key terms and additional information as well as especially developed schemes for representing the connections between the notions in compliance with the concept of the models of semantic memory [1]. The structuring of the scientific knowledge is done on the basis of content analysis. Thus the basic levels of interrelation between the notions and information units were defined in an expert way (16 notions central to the subsequent cognitive actions and are a new object of acquisition) [9]. A scheme representation of the basic hierarchical interrelations in the input information was obtained as a result of the content analysis Figure 1. The study content is organised in a Web site that contains two forms of presentation – schemes of key terms and additional information. The scheme forms reflect the hierarchical and proportional interrelations between the units of scientific information.

Each student works independently with the study content and him or her have been given series of preliminary questions and tasks whose solution should be found on the basis of perceived information. The method of the interview is chosen as a means of feedback on the result of the processing since an unconventional format for reproduction and maximum degree of free expression of all verbal variants of presentation of the acquired knowledge is pursued. A plan-thesis of the interview including 16 questions directed to the knowledge expression for the researched information units (fixed in notions) has been developed.

The results' assessment is done on the basis of the LSA method, which allows the verbal information to be assessed. Analysis of data allows an assessment of the closeness between the notions in the input study information and the one, which has been achieved in the students' system of knowledge.

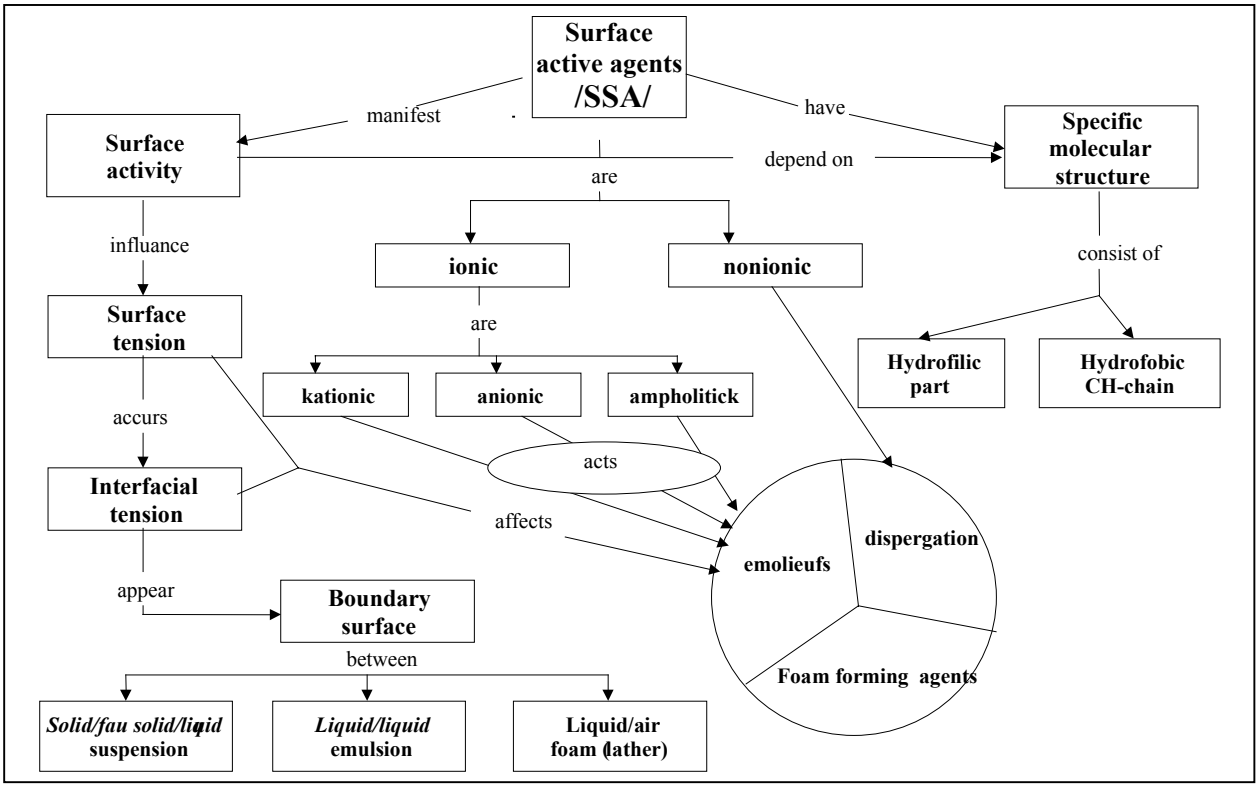


Fig. 1. Generalized scheme of the connections between the researched key terms from the study information

ASSESSMENT AND RESULTS

The assessment was done on the basis of Latent Semantic Analysis (LSA). It is a powerful statistical technique for indexing, extraction and analysis of text information that has been used successfully in different spheres of human cognition during the past decade. The method is completely automated and does not use any preliminary compiled dictionaries, semantic nets, knowledge data base, conceptual hierarchies, grammatical, morphological and syntactic analysers, etc. The analysis is based on the hypothesis that latent interrelations governing the entirety of mutual limitations exist between the separate words and the generalised context (sentences, paragraphs and whole texts) in which they occur. Their discovery and correct treatment allows LSA to cope successfully with synonymy and to some extent with polysemy: the two hardest nuts to crack in the statistical processing of text information.

LSA is a two-stage process including training and result analysis. The training phase begins with the formation of a frequency matrix of word's occurrences in the

documents. The matrix is submitted to logarithmic and entropy transformations followed by singular value decomposition. This results in the compression of the source space in much smaller one where we have only a limited number of significant factors (generally between 50 and 400). Thus, each term or document is associated a vector of reduced dimensionality, e.g. 100.

The second phase is the analysis phase. Most often this includes the study of the proximity between a couple of documents, a couple of terms or between a term and a document. A simple mathematical transformation permits to obtain the vector for a non-indexed text. The proximity degree between two documents as well as between two terms can be calculated as the dot product between their normalized LSA vectors. The usage of other measures is also possible, e.g.: Euclidean and Manhattan distances, Minkowski measures, Pearson's coefficient etc. (see [5,6,7,8] for details)

The implementation of the method of LSA for the information corpora and their research was done in several stages:

- Recording the whole information into the study texts, the basic notions that are not represented in complete definitions in separate files (documents), and coding after which they are subjected to LSA. An additional processing of the texts was needed at this stage in order to record all the symbols contained in the texts in comprehensive verbal form
- Research of the relations within the system of notions in the input texts (according to the theoretical data about the content of the notions and the whole text information) Thus the matrix of closeness is produced. Figure 2 illustrates the matrix of closeness between pairs of notions in 5 different colours corresponding to the 5 different intervals of closeness: black (87,5%–100%), dark blue (75%–87,5%), grey (62,5%–75%), light grey (50%–62,5%) and white (0%–50%). It defines the character of the output space of information and the structure of the notions within.

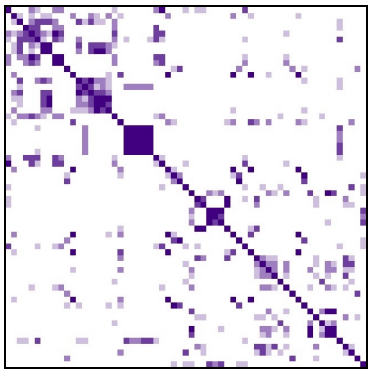
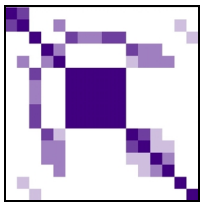


Fig. 2. Initial space of closeness between the concepts in the study texts

- Recording all answers supplied by the students in the interview in separate files (documents) and the relevant coding so that they can be subjected to LSA.
- Defining the relations between the 16 notions. This is the system of notions as indicative to the present research and correspondingly central to the semantic space of the study texts. Their relations are defined as a sub-matrix of the initial one, optionally called *theoretical matrix*. LSA allows their closeness to be identified with the help of definitions /documents/ fig. 3 and only as terms for the notions fig. 4, i.e. as key words without the content of the documents. The matrix of fig. 4 is based only on the covert relations between the notions captured by LSA and shows the degree of closeness between the notions. The matrix of fig. 3 uses again the covert inner structure, but also shows the closeness between whole definitions and not just between notions, i.e. complete definitions (texts) and not separate notions are compared here. Further on we will work with the definitions since this yields better results.
- Calculating the closeness between the theoretical definitions of notions and those supplied by the students. Thus, the picture that emerges represents the relations between the theoretical definitions and the statements expressed by the students in response to the given questions. It is presented in mean values in Table 1, which shows the generalised degree of closeness between the input information and the one, received as feedback. The modules of difference between the notions (absolute

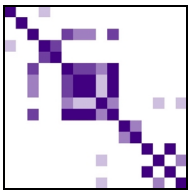
deviation) compared to the original (theoretical) correlation matrixes are calculated for the same data, Table 2.

Fig. 3. 16 researched key notions – Sub-space and the sub-matrix of the *definitions* of the basic notions in the original space



1,000	0,867	0,000	0,199	0,404	0,000	0,091	0,147	0,000	0,000	0,000	0,000	0,000	0,066	0,361	0,112
0,867	1,000	0,000	0,472	0,690	0,000	0,031	0,073	0,000	0,000	0,271	0,255	0,233	0,290	0,554	0,183
0,000	0,000	1,000	0,191	0,000	0,843	0,738	0,747	0,857	0,787	0,012	0,203	0,121	0,000	0,221	0,591
0,199	0,472	0,191	1,000	0,709	0,000	0,000	0,000	0,000	0,000	0,763	0,699	0,699	0,474	0,000	0,313
0,404	0,690	0,000	0,709	1,000	0,000	0,000	0,000	0,000	0,000	0,519	0,734	0,627	0,730	0,201	0,000
0,000	0,000	0,843	0,000	0,000	1,000	0,902	0,907	0,994	0,977	0,000	0,207	0,000	0,000	0,208	0,331
0,091	0,031	0,738	0,000	0,000	0,902	1,000	0,995	0,909	0,947	0,048	0,196	0,000	0,000	0,139	0,496
0,147	0,073	0,747	0,000	0,000	0,907	0,995	1,000	0,912	0,947	0,023	0,184	0,000	0,000	0,169	0,481
0,000	0,000	0,857	0,000	0,000	0,994	0,909	0,912	1,000	0,980	0,000	0,213	0,000	0,003	0,236	0,397
0,000	0,000	0,787	0,000	0,000	0,977	0,947	0,947	0,980	1,000	0,000	0,234	0,000	0,000	0,108	0,344
0,000	0,271	0,012	0,763	0,519	0,000	0,048	0,023	0,000	0,000	1,000	0,829	0,507	0,250	0,000	0,290
0,000	0,255	0,203	0,699	0,734	0,207	0,196	0,184	0,213	0,234	0,829	1,000	0,547	0,511	0,008	0,074
0,000	0,233	0,121	0,699	0,627	0,000	0,000	0,000	0,000	0,000	0,507	0,547	1,000	0,723	0,000	0,398
0,066	0,290	0,000	0,474	0,730	0,000	0,000	0,000	0,003	0,000	0,250	0,511	0,723	1,000	0,057	0,000
0,361	0,554	0,221	0,000	0,201	0,208	0,139	0,169	0,236	0,108	0,000	0,008	0,000	0,057	1,000	0,222
0,112	0,183	0,591	0,313	0,000	0,331	0,496	0,481	0,397	0,344	0,290	0,074	0,398	0,000	0,222	1,000

Fig. 4. 16 researched key notions – Sub-space and the sub-matrix of the basic notions (*notions themselves*) in the original space



1,000	0,245	0,000	0,572	0,292	0,000	0,119	0,164	0,000	0,000	0,000	0,000	0,000	0,310	0,000	0,020
0,245	1,000	0,154	0,163	0,105	0,010	0,000	0,000	0,014	0,000	0,054	0,000	0,000	0,000	0,000	0,000
0,000	0,154	1,000	0,000	0,000	0,828	0,676	0,679	0,443	0,799	0,000	0,080	0,000	0,000	0,000	0,000
0,572	0,163	0,000	1,000	0,755	0,000	0,000	0,000	0,000	0,000	0,144	0,116	0,302	0,055	0,249	0,000
0,292	0,105	0,000	0,755	1,000	0,000	0,000	0,000	0,000	0,000	0,135	0,413	0,000	0,263	0,000	0,000
0,000	0,010	0,828	0,000	0,000	1,000	0,864	0,875	0,651	0,994	0,000	0,094	0,000	0,000	0,000	0,000
0,119	0,000	0,676	0,000	0,000	0,864	1,000	0,998	0,600	0,898	0,000	0,102	0,000	0,000	0,000	0,094
0,164	0,000	0,679	0,000	0,000	0,875	0,998	1,000	0,601	0,904	0,000	0,068	0,000	0,000	0,000	0,083
0,000	0,014	0,443	0,000	0,000	0,651	0,600	0,601	1,000	0,668	0,000	0,000	0,000	0,601	0,000	0,625
0,000	0,000	0,799	0,000	0,000	0,994	0,898	0,904	0,668	1,000	0,000	0,114	0,000	0,000	0,000	0,000
0,000	0,054	0,000	0,144	0,135	0,000	0,000	0,000	0,000	0,000	1,000	0,732	0,000	0,000	0,000	0,000
0,000	0,000	0,080	0,116	0,413	0,094	0,102	0,068	0,000	0,114	0,732	1,000	0,067	0,000	0,000	0,000
0,000	0,000	0,000	0,302	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,067	1,000	0,000	0,945	0,206
0,310	0,000	0,000	0,055	0,263	0,000	0,000	0,000	0,601	0,000	0,000	0,000	0,000	1,000	0,000	0,674
0,000	0,000	0,000	0,249	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,945	0,000	1,000	0,187
0,020	0,000	0,000	0,000	0,000	0,000	0,094	0,083	0,625	0,000	0,000	0,000	0,206	0,674	0,187	1,000

Table 1. Closeness between the theoretical *definitions* and those supplied by the students (generalized results from the questions)

Въпрос	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**Средно**
мин.	0,757	0,872	0,342	0,613	0,664	0,803	0,575	0,596	0,496	0,766	0,789	0,945	0,085	0,832	0,518	0,684	**0,646**
Макс.	0,977	0,960	0,813	0,967	0,977	0,989	0,909	0,901	0,943	0,999	0,972	0,984	0,894	0,948	0,932	0,944	**0,944**
ср.ар.	0,876	0,923	0,662	0,849	0,922	0,904	0,812	0,786	0,822	0,910	0,923	0,970	0,808	0,875	0,822	0,844	**0,857**
ср.ст.	0,876	0,932	0,683	0,903	0,956	0,893	0,853	0,808	0,804	0,887	0,940	0,974	0,866	0,861	0,866	0,843	**0,872**

Table 2. Module of the difference between the theoretical *definitions* and those supplied by the students from the experimental group (generalized responses to questions)

Въпрос	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**средно**
мин.	0,023	0,040	0,187	0,033	0,023	0,011	0,091	0,099	0,057	0,001	0,028	0,016	0,106	0,052	0,068	0,056	**0,056**
Макс.	0,243	0,128	0,658	0,387	0,336	0,197	0,425	0,404	0,504	0,234	0,211	0,055	0,915	0,168	0,482	0,316	**0,354**
ср. ар.	0,124	0,077	0,338	0,151	0,078	0,096	0,188	0,214	0,178	0,090	0,077	0,030	0,192	0,125	0,178	0,156	**0,143**
ср.ст.	0,124	0,068	0,317	0,097	0,044	0,107	0,147	0,192	0,196	0,113	0,060	0,026	0,134	0,139	0,134	0,157	**0,128**

These data have a generalising character and may be used as an indicator on the overall effect of the educational influences on the students' groups. We can perceive to which extend each notion has found an adequate contemplation in the students' groups. The minimum and maximum values show the greatest closeness of preserved information about the notion and the most incomprehensive and inaccurate reflection of the content. The achievements as a whole can be seen either in the row of mean arithmetic values or they are indicative of how the students' responses during the interview correlate with the theoretical data about the notions. In the experimental group

these results are within the range from 0.662 to 0.970, which demonstrates a high degree of correspondence between the theoretical data and the practically acquired information. The data about the mean arithmetic module of difference are with the corresponding values ranging from 0.338 to 0.030. It is of significant importance to note here that these mathematical expressions of the degree of closeness between the theoretical and practical data have been obtained due to the transformation specific for LSA. In other words the relation to the whole bulk of scientific information in the texts and the definitions of the 58 key terms lies behind every number and this means that it is impossible to expect complete correspondence. The measurement of the data received in the LSA should be analysed in more details in order to reveal the essence of the results. On one hand they are very indicative of the results of the innovative educational technology, i.e. they verify the ideas laid in the experimental education, which accomplishes the cognitive strategies for optimisation of the organisation of the study environment. In practice we reveal the correspondence between the basic notional relations in the semantic space and those of the students. On the other hand the research proves the potential of LSA for revealing latent correlation. This is the reason the research manages to achieve its objective for assessment of the deep levels of information processing. In an analysis based on the traditional methods of research of learning data results only for the "surface" level have been obtained i.e. the correlation between the information units represented in the memory cannot be assessed. In this particular research of a free verbal expression during an interview we manage to subject to processing text forms of knowledge and thus achieve a more detailed characteristic of the actually built system of notions and their correlation.

DISCUSSIONS AND FUTURE EXPLORATION

The obtained results are encouraging and they demonstrate the indisputable advantages of the LSA in identifying and quantity assessment of the qualities of acquired knowledge. New experiments, which will help us to assess better the role of LSA, are planned including other methods as well.

BIBLIOGRAPHY

1. Gerganov, E. ,(1987) Memory and thought, Sofia.
2. Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E., (1997). 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 (pp. 412-417). Mawhwah, NJ: Erlbaum.
3. Landauer, T. K., Laham, D., & Foltz, P. W., (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10,(pp. 45-51). Cambridge: MIT Press.
4. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
5. Laudauer T., Foltz P., Laham D. Introduction to Latent Semantic Analysis. Discourse Processes, 25, pp. 259-284.
6. Nakov P. Latent Semantic Analysis for Bulgarian literature. In Proceedings of Spring Conference of Bulgarian Mathematicians Union. Borovetz. 2001.
7. Nakov P. Getting Better Results with Latent Semantic Indexing. In Proceedings of the Students Prenetations at ESSLLI-2000, pp. 156-166, Birmingham, UK, August 2000.

– –

8. Nakov P. Latent Semantic Analysis of Textual Data. In Proceedings of CompSysTech'2000, Sofia, Bulgaria. June 2000.
9. Terzieva, S., Handjieva, S., 2000, Cognitive Approach to Design Learning Process, 3-th Working Conference on Engineering Education 21-st Century. pp. 250 –254. 17-20 April 2000, Sheffield, England.

Senia Petrova Terzieva, University of Chemical Technology and Metallurgy, ++ 359 2 6254 462, senia@uctm.edu
Preslav Ivanov Nakov, Sofia University "St. kliment Ohridski", 088/373-609, preslav@rila.bg, preslav@rocketmail.com
Dr. Sneja Valkanova Handjieva, University of Chemical Technology and Metallurgy 6254359, sneja@uctm.edu