Text Semantic Analysis applied for Educational and Training Purposes

Jozef Stašák, Jaroslav Kultan

Abstrakt Príspevok sa zaoberá uplatnením sémantickej analýzy textu v procese vzdelávania. Obsahuje prehľad prístupov k riešeniu danej problematiky – teda sémantickej analýzy textu písaného v prirodzenom jazyku (ďalej len TNL text). Prístup založený na uplatnení latentnej sémantickej analýzy sa pokladá za najvýznamnejší, ktorý sa používa na dané účely. Avšak existujú aj iné prístupy a metódy. Jednou z iných metód je sémantická analýza TNL textu na báze riešenia základnej rovnice pre lingvistické modelovanie biznis procesov (ďalej len rovnice PBPL) aplikovanej na sémantickú analýze TNL Textu. V príspevku sú taktiež prezentované výsledky takéhoto riešenia, ktoré možno aplikovať na rôzne typy testov.

Abstract The paper deals with text semantic analysis applied within educational and training processes. It contains an overview of approaches and methods concerned to semantic analysis of text written in a natural language (TNL text). The Latent Semantic Analysis (LSA) is considered to be the most significant method applied for the above-mentioned purposes. However, there are several different approaches methods as well. On the other hand, the TNL text semantic analysis is considered to be a process and we propose a new TNL text semantic analysis approach based on PBPL Equation solution, which is usually applied for process modeling. Subsequently, the PBPL Equation solution for TNL semantic analysis results are presented together with their application in TNL text content understanding, which might create basis for different types of tests.

Kľúčové slová: text v prirodzenom jazyku, TNL text, sémantická analýza textu, rovnica PBPL, riešenia pre rôzne typy testov.

Key words Text written in natural language, TNL text, text semantic analysis, PBPL Equation, different test type solution

JEL classification: D83

1. Introduction

In general, a content of most documents is written in a natural language represented by set of logical sentences, while the following thesis might be postulated: "Any written or spoken (pronounced) idea is considered to a natural language logical sentence" and the text content represented via those types of sentences is considered to be a *"text in natural language" hereinafter known as TNL text* (Stašák, 2004). A set of syntactic, grammar and semantic variables represent the TNL content. However, this paper deals with TNL text sematic aspects and variables only, while applying semantic analysis to training and education play a role principle importance related to that paper's content On the hand, the TNL text semantic analysis might be considered to be a business process, where the Principle Business Process Linguistic Equation - PBPL Equation (Stašák, Andrejčák, Sláviková, Grell, 2014) could be applied from TNL text semantic analysis quantification point of view as well.

The paper's main goal is to prepare a proposal for application of PBPL equation for quantification and modeling the text semantic analysis process, while that modeling results should be applied to training and education purposes.

In order to achieve the main goal, the following partial goals should be postulated and fulfilled:

- To discover a set of appropriate formulas and algorithms, which should enable applying the PBPL Equation for deriving of the Principle TNL Text Content Analysis Linguistic Equation (PTCAL Equation).
- To prepare a proposal for the (PTCAL Equation) design and implementation for TNL text semantic analysis.
- To prepare a proposal for application of quantification and modeling results achieved with the use of PTCAL Equation for education and training purposes.

However, the above-mentioned goal structure determines the paper's structure as well. A review of Text Content Semantic Analysis principles, approaches and methods closely related to their application for educational and training purposes is described within first part of the paper. The problems concerned to PBPL Equation for deriving of the Principle TNL Text Content Analysis Linguistic Equation (PTCAL Equation) and the PTCAL Equation application for semantic analysis of texts utilized in providing educational and training activities are described in the second part of the paper.

2. Text Content Semantic Analysis applied for Educational and Training Purposes - Qualitative and Quantitative Aspects

2.1 Terms, Principles and Approaches

Latent Semantic Analysis (LSA) is a mathematical/statistical technique for extracting and representing the similarity of meaning of words and passages by analysis of large bodies of text. It uses singular value decomposition, a general form of factor analysis, to condense a very large matrix of word-by-context data into a much smaller, but still large-typically 100-500 dimensional-representations (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). The right number of dimensions appears to be crucial; the best values yield up to four times as accurate simulation of human judgments as ordinary co-occurrence measures.

However, latent semantic analysis (LSA) can be very suitable for extracting quantitative information from nonnumeric marketing data proposed by Inigo Arroniz. He developed a scalable methodology that is capable of extracting information from extremely large volumes of nonnumeric data. The proposed methodology integrates concepts from information retrieval and content analysis to analyze textual information. This approach avoids a pervasive difficulty of traditional content analysis, namely the classification of terms into predetermined categories, by creating a linear composite of all terms in the document and, then, weighting the terms according to their inferred meaning. In the proposed approach, meaning is inferred by the collocation of the term across all the texts in the corpus. It is assumed that there is a lower dimensional space of concepts that underlies word usage. The semantics of each word are inferred by identifying its various contexts in a document and across documents (i.e., in the corpus). After the semantic similarity space is inferred from the corpus, the words in each document are weighted to obtain their representation on the lower dimensional semantic similarity space, effectively mapping the terms to the concept space and ultimately creating a score that measures the concept of interest (Arronis, 1997, 27-36).

Qualitative content analysis Qualitative content analysis - as one of today's most extensively employed analytical tools, content analysis has been used fruitfully in a wide variety of research applications in information and library science (ILS). Similar to other fields, content analysis has been primarily used in ILS as a quantitative research method until recent decades. Many current studies use qualitative content analysis, which addresses some of the weaknesses of the quantitative approach.

Qualitative content analysis has been defined as:

- a research method for the subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes or patterns (Hsieh & Shannon, 2005, 1278),

- an approach of empirical, methodological controlled analysis of texts within their context of communication, following content analytic rules and step by step models, without rash quantification" (Mayring, 2000, 2), and any qualitative data reduction and sense-making effort that takes a volume of qualitative material and attempts to identify core consistencies and meanings" (Patton, 2002, 453).

The above-mentioned three definitions illustrate that qualitative content analysis emphasizes an integrated view of speech/texts and their specific contexts. Qualitative content analysis goes beyond merely counting words or extracting objective content from texts to examine meanings, themes and patterns that may be manifest or latent in a particular text. It allows researchers to understand social reality in a subjective but scientific manner[1] ,[2] .

Semantic analysis broadly refers to using computers to determine and analyze the meaning of natural language. Semantic analysis is a subfield of the more general fields of natural language processing (NLP) and computational linguistics. Typical semantic analysis problems include (Robson, Ray, 2012)

- Word sense disambiguation:
  - Determine the meaning of a particular word in a given context, e.g. in the sentence "a tank is parked in the exhibit hall," does "tank" refer to an Army vehicle or a tank of water?

- Topic detection:
  - Determine the topics discussed in a passage of text. A simpler form of topic detection is keyword generation, i.e. automatically associating a useful set of keywords with a document.
- Semantic Similarity:
  - Find resources that are related in meaning to a given document or text passage.

1 Yan Zhang, Y- and Wildemuth, B.M.: Qualitative Analysis of Content
https://www.ischool.utexas.edu/~yanz/Content_analysis.pdf

2 An Introduction to content analysis
http://www.depts.washington.edu/uwmcnair/chapter11.content.analysis.pdf

This approach is applied for extracting and modeling the semantic information content of different documents to support semantic document retrieval, especially WEB oriented documents , where Existing HTML mark-up is used only to indicate the structure and lay-out of documents, but not the document semantics. As a result web documents are difficult to be semantically processed, retrieved and explored by computer applications. Existing information extraction system mainly concerns with extracting important keywords or key phrases that represent the content of the documents. The semantic aspects of such keywords have not been explored extensively. The authors propose an approach meant to assist in extracting and modeling the semantic information content of web documents using natural language analysis technique and a domain specific ontology. Together with the user's participation, the tool gradually extracts and constructs the semantic document model which is represented as XML. The semantic models representing each document are then being integrated to form a global semantic model. Such a model provides users with a global knowledge model of some domains (Shahrul Azman Noah, Lailatulqadri Zakaria, Arifah Che Alhadi, 2013).

Concept based approach

Information retrieval systems traditionally rely on textual keywords to index and retrieve documents. Keyword-based retrieval may return inaccurate and incomplete results when different keywords are used to describe the same concept in the documents and in the queries. Furthermore, the relationship between these related keywords may be semantic rather than syntactic, and capturing it thus requires access to comprehensive human world knowledge .Concept-based retrieval methods have attempted to tackle these difficulties by using manually built thesauri, by relying on term co-occurrence data, or by extracting latent word relationships and concepts from a corpus. In this article we introduce a new concept-based retrieval approach based on Explicit Semantic Analysis (ESA), a recently proposed method that augments keyword-based text representation with concept-based features, automatically extracted from massive human knowledge repositories such as Wikipedia. Our approach generates new text features automatically, and we have found that high-quality feature selection becomes crucial in this setting to make the retrieval more focused. However, due to the lack of labeled data, traditional feature selection methods cannot be used, hence we propose new methods that use self-generated labeled training data. The resulting system is evaluated on several TREC datasets, showing superior performance over previous state-of-the-art results (Egozi, Markovitch, Gabrilovich, 2011).

Text based approach is closely related to the analysis of film content, while we ask three questions, when considering that approach:

- What information do these texts provide about film content and how do they express it?

- How can machine-process able representations of film content be extracted automatically in these texts?
- How can these representations enable novel applications for analyzing and accessing digital film data?

In order to answer these questions, the authors have analyzed collocations in corpora of audio description scripts (AD) and screenplays (SC), developed and evaluated an information extraction system and outlined novel applications based on information extracted from AD and SC scripts. The authors found that the language used in AD and SC contains idiosyncratic repeating word patterns, compared to general language. The existence of these idiosyncrasies means that the generation of information extraction templates and algorithms can be mainly automatic. We also found four types of event that are commonly described in audio description scripts and screenplays for Hollywood films: Focus_of_Attention, Change_of_Location, Non-verbal_Communication and Scene_Change events. We argue that information about these events will support novel applications for automatic film content analysis (Vassiliou, 2006).

Ontology based Text Semantic Analysis

Organization of control knowledge is closely related to problem of selecting the type of questions, testing trajectory formation mode and methods of answer verification.
We propose the concept of intellectual testing system based on the domain knowledge ontology in order to solve this problem. The planimetry ontology is used as the domain knowledge ontology.
The following types of questions are offered for knowledge quality control:

- Test questions, when several answer variants are pre-defined and one them shall be selected

The following types of questions are offered for the control of knowledge quality:

- test questions of closed form, i.e., when several variants of the answer are suggested, one of which is correct and should be selected;
- test questions of open form, i.e.
  without suggested variants of answers (such questions are useful for evaluation of knowledge of terms, definitions, notions, etc.);
- without suggested variants of answers (such questions are useful for evaluation of knowledge of terms, definitions, notions, etc.);
- questions situation tests, i.e., set of test assignments
  designed for solution of problematic situations (a geometrical task).

Pre-linguistic processing of the source text – analysis of the text morphology and syntax needed to separate appropriate terms (classes, subclasses, properties and relations)

- Text formal understanding constructing an ontological graph.

Semantic Networks versus Document Content Semantic Analysis

Semantic Network Analysis deals with the abstraction gap by using objects that are as close to the text as possible in the extraction and aggregating these along a hierarchy in the querying step.. However, this context is kept as broad as possible in the extraction phase. The choice of how to aggregate the concrete objects to the abstract theoretical concepts is a much stronger commitment to a context, but this step is reversible, making the data useful for other analyses using other contexts compatible with the initial choice of objects.

Similarly, the complexity gap is dealt with by postponing the irreversible reduction of complex structures into unstructured variables to the queries used to answer the research question. This way, the coders map the rich structure of text to the relational structure of the network representation. If required by the research question, the queries map this relational structure to a set of unstructured variables, for example the frequency of specific patterns of one or more relations.

This network representation is combined with relevant background knowledge about the objects in the network. This background knowledge contains both dictionary knowledge such as 'a politician is a person,' and encyclopedic knowledge such as the fact that Bill Clinton was president from1993to2001. This combined data set of media data and background knowledge is then queried to answer the research question.

The general domain context contains the broad assumptions necessary for extracting the network of objects, specifically the choice of objects and relations that are extracted, and the textual features that are used to extract those relations manually or automatically. The inner oval represents the more specific research context, containing the assumptions made to infer an answer to the research question. These assumptions include the way objects are categorized using the background knowledge, the operationalization of the research question in terms of patterns in the combined data set of media data and background knowledge, and the interpretation of these patterns in terms Semantic Network Analysis of the social context of the message.

The advantage of Semantic Network Analysis for sharing and reusing data can be explained in terms of these contexts: Since the extraction of the Semantic Network depends only on the domain context, that network can be used to answer any research question whose specific research context is contained in that domain context. In other words: to share data between researchers, their choice of which concrete objects to measure has to be compatible, but they can differ in how these objects are aggregated or how the network is queried to answer the research question. The advantage of Semantic Network Analysis for coding can also be seen in these terms: Since coding relies only on

minimal assumptions about choice of vocabulary and relations, coders are not burdened with interpreting categorisation schemes or reducing complex textual phenomena to single variables. Automatic extraction is not necessarily easier, but since the extraction method is linked to the domain context rather than to the specific context of a research question, the method is more general and does not need to be retrained or reconstructed for different research questions as long as the assumptions made in the domain context remain valid.

The separation between extraction in the domain context and querying in the research context also allows us to position Semantic Network Analysis in the manifest–latent and quantitative–qualitative debates. In the former, since the extraction of the network should be done as close to the text as possible, it should be restricted to the manifest content. In the querying, however, latent content can be exposed by inferences based on the manifest content (such as transitivity of relations), by inference based on background knowledge (such as relations between objects that are assumed to be general, e.g. that inflation is bad for the economy), or by causal inferences assumed in the query operationalization of the research question (e.g. inferring motives of sources from their messages). In the quantitative–qualitative debate, the main difference between the two positions is that quantitative analysis requires codes with a definition and meaning that is fixed before reading the text, while qualitative analysis builds the code structure and meaning in an interactive process while (re)reading the text. In the extraction phase, the domain context fixes the vocabulary and the meaning of the contained objects. However, the aggregation to more abstract objects and interpretation of patterns of relations happens in the research context, and this can be conducted in an interactive fashion, by conducting queries, inspecting the results and the original texts on which these results are based, and refining the definitions of the concepts and patterns used in the query. In this approach, the text is first coded using descriptive codes, similar to the extraction of a representational network.

The text is then recoded using interpretive codes, which place a layer of meaning on the original codes, and can be seen as an interactive parallel to the categorisation of objects using background knowledge. Finally, the text is coded using pattern codes to indicate interesting patterns and interaction, which is similar to the querying of the network representation.

As noted by Pleijter (2006), many qualitative Content Analysis studies would benefit from a more rigorous description of the used analysis procedure, so it would be very interesting to explore whether Semantic Network Analysis can play a role in increasing that rigour without unduly limiting the hermeneutic analysis.

For Semantic Network Analysis to be possible, it is required that the concept to be constructed can be expressed as a pattern on the extracted network. Although the exact operationalization is beyond the scope of this thesis, it is plausible that the constructs mentioned above can be expressed as patterns on a network: studying Agenda Setting requires determining the visibility of issues, which can be done by counting the total number of links connecting the issue node to other nodes in the network. Evaluations of

actors and issues can be operationalized as the connections of those actors and issues with the Ideal or with positive values. Second-level Agenda Setting uses the visibility of associations between objects and their attributes, which is the frequency of links between these objects and attributes. Horse Race news is represented by the positive and negative links connecting the real world with the studied actors. Issue positions look at the evaluations of issues in texts from a political source such as manifestoes, or at relations between actors and issues in other texts. Political conflict is naturally operationalized as negative relations between political actors. Frames are defined by Entman (1993) as salience of selected attributes which are also operational is able as the edges between the framed issues and these attributes. These observations make it plausible that networks are a good way of representing media messages; and successful applications of Semantic Network Analysis point in the same direction (Kleinnijenhuis et al., 2007, 366-384) **(**Atteveldt, 2008, 27-36**).**

2.2 Text Content Semantic Analysis results related to providing Educational and Training Activities (application of Latent Semantic Analysis (LSA) method)

The **Latent Semantic Analysis (LSA) method** basic principles are described within section 2.1. Now, I would like to give a short overview how that method might be applied for educational a training purposes and activities, while learning from reading are considered to be of the above-mentioned activities.

Learning from text, however, is not the same as remembering the text. (Kintsch,1994, 294-303) argued that a central feature of learning from text is linking up the textual information with prior knowledge. The new information must be integrated with prior knowledge both for current comprehension and for later use in new situations.

Thus, learning presupposes suitable prior knowledge to which the to-be-learned information can be linked. If there is no relevant knowledge basis at all, this integration cannot take place -learning fails. Although texts can be memorized, they will remain isolated memory episodes, inert knowledge that is not available for new tasks.

Now we shall have a look at application of, Latent Semantic Analysis (LSA) for providing a practical alternative for matching students differing in background knowledge within structured texts of varying topical sophistication (Wolfe, Schreiner, Foltz, Kintsch, Launder, 1998).

As a knowledge domain, we selected the functioning of the human heart. This is a topic about which college students have some background knowledge and in which large variations in background knowledge might be expected. As instructional texts we chose four texts, each of which was 17 paragraphs long. These texts varied widely in sophistication. Text A, the easiest text, was a chapter from a children's book about the human body. Text B was taken from an adult-level general introduction to the human circulatory system. Text C was written for undergraduate introductory anatomy students.

Text D, the most difficult text, came from a medical school text and contained information a layperson couldn't ordinarily be expected to know.
Each participant was randomly assigned to one of these texts and the amount of learning was determined for each student by comparing performance on pre- and post reading knowledge assessment tasks designed to reflect any knowledge gained by reading. Knowledge was assessed in two ways: by means of a short-answer questionnaire, and by means of an essay the student wrote on the functioning of the heart.

Domain knowledge was measured in three ways: as the score a student received on the questionnaire prior to reading the text (pre- questionnaire), the grade on the essay written prior to reading (pre-essay), and the cosine between the LSA representations of the student's prior essay and a standard college-level textbook chapter on the functioning of the heart (cosine pre-essay. standard). We chose Text C as a standard text, since it appeared to be most representative for the heart knowledge of college students, though Text B could have been used Domain knowledge was measured in three ways: as the score a equally well, with very similar results (Text A was too easy, Text D too difficult, which resulted in the range of cosines being more restricted). Amount of learning was operationally defined in two ways: as the proportion of possible improvement in the scores on the questionnaire from before to after reading (learn-questionnaire), and as the proportion of possible improvement in the grades the student's essays received before and after reading (learn-essay). Another way of looking at the relationship between background knowledge and learning efficiency is to plot the latter as a function of how closely related a particular student's knowledge is to the text he or she actually read. The LSA cosine between the essays a student wrote and the text he or she read provided such a measure (cos pre- essay.textread). However, there are more other results as well, while the results are described and discussed in (Wolfe, Schreiner, Foltz, Kintsch, Launder, 1998, 309-336).

An example related to application of Latent semantic indexing approach

When considering a text content semantic analysis we have to consider the vocabulary problem in human-computer interaction. Most approaches to retrieving textual materials depend on a lexical match between words in users' requests and those in or assigned to database objects. Because of the tremendous diversity in the words people use to describe the same object, lexical matching methods are necessarily incomplete and imprecise (Furnas et. all. 1999, 1753-1806). The latent semantic indexing approach tries to overcome these problems by automatically organizing text objects into a semantic structure more appropriate for matching user requests.

This is done by taking advantage of implicit higher-order structure in the association of terms with text objects. The particular technique used is singular-value decomposition, in which a large term by text-object matrix is decomposed into a set of about 50 to 150 orthogonal factors from which the original matrix can be approximated

by linear combination. Terms and objects are represented by 50 to 150 dimensional vectors and matched against user queries in this "semantic" space. Initial tests find this completely automatic method widely applicable and a promising way to improve users' access to many kinds of textual materials, or to objects and services for which textual descriptions are available (Dumais, Furnas, Landauer, Deerwester, Harshman, 1990).

2.3 Quantification of Text in Natural Language

The previous sections of that contribution deal with a literary overview concerned to semantic text analysis, where Latent semantic text analysis plays a role of great importance. On the other hand, this section deals with quantification of text in natural language (TNL), which is much closed to semantic text analysis as well.

Any TNL document consists of semantic subsets, which creates its basic structure elements. These semantic subjects are called fragments and contain one or more TNL logical sentences. When considering a TNL logical sentence to be pronounced or written idea, we may also postulate the statement that one or more TNL logical sentences can represent one or more knowledge. With respect to that issue we are allowed to say that each TNL logical sentence consists of *objects* and appropriate *semantic relations.*

3. The principle TNL text content analysis linguistic equation (PTCAL equation)

3.1 The (PTCAL Equation) existence issues and derivation assumptions and principles

Consideration no. 3.1

Let us consider a semantic content text analysis to be a business process or a process actually. In general, any business process might be quantified via Principle Business Process Linguistic Equation – PBPL Equation, the general form of which is postulated via formulas (3.1a – 3.1d)

$$\{Petx\ (i, j')\} \otimes \{P(i, j)\} = \{Tbex\ (i, j'')\} \otimes \{Retx\ (i, j''')\} \quad (3.1a)$$

$$\{Petx\ (i, j')\} \otimes \{P\ (i, j)\} = \{Res1\ (i, j''')\} \quad (3.1b)$$

$$\{Tbex\ (i, j'')\} \otimes \{Retx\ (i, j''')\} = \{Res2\ (i, j'''')\} \quad (3.1c)$$

$$\{Res1\ (i, j''')\} = \{Res2\ (i, j''')\} \quad (3.1d)$$

Let us try applying it to the process denoted as the TNL text content analysis represented by a set of logical TNL sentences, which describes a selected event or appear. **Our task is to explain a semantic content of that TNL text.** With respect to the above-mentioned task requirement, the following issues should be postulated:

Issue no.1

A set of words represents any TNL text logical sentence (hereinafter know as sentence), while they might be divided in two groups. The ***first group*** represents the words, which play a role of principle importance from the text semantic analysis point f view and might be approximated via elements of {Petx (i, j)} linguistic set elements[3]
The **second group** represents the words, which play a role of supplementary importance from the text semantic analysis point f view and might be approximated via elements of {Pe (i, j)} linguistic set elements. On the other hand, the semantic content of selected TNL logical sentence might be approximated and quantified via equation (3.1b), while the **{Res1 (i, j'')}** linguistic set

Issue no.2

Any TNL logical sentence V_{TNL} consists of elements, which create an integral part of {V (i, j)} set might be approximated via linguistic set **{Res1 (i, j''')}** and formula (3.2a) and (3.2b) might be postulated

$$V_{TNL} = \{V (i, j)\} \qquad (3.2a)$$

Where i=1….n - number of TNL logical sentences, the V_{TNL} text consists of
j=1….m1 – number of TNL words, the individual logical sentence consists of

With respect to the above-mentioned issues formula (3.3) may be postulated

$$V_{TNL} = \{V (i, j)\} = \mathbf{\{Res1 (i, j''')\}} \qquad (3.3)$$

Where i=1….n and has the same meaning as postulated within issue no. 2, j'''=1….m_2 and represents a number of Res1 set elements, which approximate a semantic meaning of any TNL logical sentence denoted as V_{TNL} . On the other hand, the Res1 set create basis for definition of an appropriate semantic function S_F1 (i), see also formula (3.4)

$$S_F1 (i) = \prod_{m_2} \mathbf{\{Res1 (i, j''')\}} \qquad (3.4)$$

[3] Petx (i, j)}, where i=1….n – number of a logical sentence, which creates an integral part of the above-mentioned TNL text and j=1…..m_1 a number of this type words contained in that sentence.

j=1

It means, any TNL logical sentence semantic meaning might be approximated via appropriate semantic function value S_F1 (i), see also formula (3.4a),

$$S_F1\ (i) = \prod_{j=1}^{m_2} \{V\ (i, j)\} \qquad (3.4a)$$

while a semantic meaning of the entire TNL text, which consists of may be approximated via formula (3b)

$$S_{VTNL} = \prod_{I=1}^{n} S_F1\ (i) \qquad (3.4b)$$

Consideration 3.2

Semantic meaning approximation via set of other logical sentences

Consideration no.3.1 deals with a problem of semantic meaning value determination, while that value is closely related to TNL logical sentences {V (i, j)}, which create, an integral part of V_{TNL} text.

On the other hand, Consideration no.3.2 is concerned to approximation of V_{TNL} text via other TNL logical sentences and to estimate how accurately or exactly the original V_{TNL} might be approximated via new (other) TNL logical sentences.

In order to fulfill the above-mentioned goal, let us consider a TNL text, which consists of other TNL logical sentences creating an integral part of $\{V_{exp}\ (i, j)\}$ set. We shall try estimating how exactly or accurately may a semantic meaning of TNL text denoted as V_{TNL} text approximated via new sentences contained within $\{V_{exp}\ (i, j)\}$ set, while the $\{V_{corel}(i, j)\}$ elements should represent a tightness measure as well.

$$V_{TNL} = \{V\ (i, j)\} = \{V_{exp}\ (i', j')\} \otimes \{V_{corel}(i'', j'')\} \qquad (3.6)$$

Consideration 3.3

Theorem 3.1

Let us consider a TNL text denoted as V_{TNL}, which consists of a set of TNL logical sentences creating the {V (i, j)} set content, while i=1….n is number of TNL logical sentences stored within {V (i, j)} set and $j=1….m_{1,}$ it means that any TNL logical

sentence consist of m_l words having an appropriate semantic meaning and creating an entire semantic meaning of i^{th} TNL logical sentence. However, a semantic meaning of the above-mentioned TNL text can be approximated by a set other TNL logical sentences denoted as $\{V_{exp}\,(i', j')\}$ set as well, while the $\{V_{corel}(i'', j'')\}$ set contains elements, which represent a tightness measure of that approximation (see also formula 3.6).
With respect to the above-mentioned assumptions and assumptions postulated within Consideration no.1 and Consideration no.2, the following assertion might be postulated:

The semantic meaning of $\{V_{exp}\,(i', j')\}$ set can be approximated via set of Tbe terms contained within {**Tbex (i, j'')**} set and the $\{V_{corel}(i'', j'')\}$ set can be approximated via set of Ret terms contained within {**Retx (i, j''')},** while formula (3.7) may be postulated:

$$V_{TNL} = \{V\,(i, j)\} = \{V_{exp}\,(i', j')\} \otimes \{V_{corel}(i'', j'')\} = \mathbf{\{Tbex\ (i,\ j'')\}} \otimes \mathbf{\{Retx\ (i,\ j''')\}} \quad (3.7)$$

It means, a TNL text content semantic analysis may be considered to a business process and the PBPL Equation solution result can be applied for quantification purposes related to that business process.

Proof
In order to provide a proof of 3.1 theorem let us apply equations (3.1b), (3.1d) (3.3) and (3.6) as an outgoing point. With respect to these equations formulas (3.8) (3.9) may be postulated

$$\{V\,(i, j)\} = \{Res2\,(i, j'''')\} \quad (3.8)$$
$$\{Tbex\,(i, j'')\} \otimes \{Retx\,(i, j''')\} = \{Res2\,(i, j'''')\} \quad (3.9)$$
$$\{V\,(i, j)\} = \{Tbex\,(i, j'')\} \otimes \{Retx\,(i, j''')\} \quad (3.10)$$

With respect to equation (3.6), equations (3.11) and (3.12a), (3.12b) may be postulated

$$\{V\,(i, j)\} = \{V_{exp}\,(i', j')\} \otimes \{V_{corel}(i'', j'')\} = \mathbf{\{Tbex\ (i,\ j'')\}} \otimes \mathbf{\{Retx\ (i,\ j''')\}} \quad (3.11)$$

$$\{V_{exp}\,(i', j')\} = \mathbf{\{Tbex\ (i,\ j'')\}} \quad (3.12a)$$
$$\{V_{corel}(i'', j'')\} = \mathbf{\{Retx\ (i,\ j''')\}} \quad (3.12b)$$

The equation (3.6) or (3.11) is considered to be a solution of PBPL Equation for semantic approximation of original TNL text via set of terms to be explained Tbe(i, j), while the Retx (i, j''') contains elements, which indicate a measure approximation. However, the (3.7) or (3.11) equation is denoted as **"**The principle TNL text content analysis linguistic equation (PTCAL Equation)".

3.2 The (PTCAL equation) design and implementation for TNL text semantic analysis

The formulas 3.11, 3.12a and 3.12b postulated within section 3.1 and derived based on Theorem 1 represent a solution of general PBPL equation for the process of TNL text semantic analysis, which that solution is described in a general form as well. Now, there is a question *"How to apply those formulas design and implementation for TNL text semantic analysis?"* In order to answer that question let us consider the TNL text, which deals with "latent semantic indexing" (LSI) approach, which the student should read and answer the test questions. The correct answer should confirm, he/she did or did not understand that text properly. The test questions are created and generated based o the following philosophy:

1. Any text is written because of appropriate reason and has its own *mission statement*. In that case the TNL logical sentence "The "latent semantic indexing" (LSI) approach" represents the TNL text mission statement, while that sentence semantic meaning might be quantified via set of linguistic sets represented by formula (3.7).
2. On the other hand, a set of appropriate *objectives* represents that TNL text content, which the set is closely related to the above-mentioned mission statement, while those objectives should be quantified via Tbex (i, j) sets and formulas (3.13a) and (3.13b) might be postulated

   Tbex (1, m_1) =" The problems of word-based access by treating the observed word to text-object association data as an unreliable estimate of the true"

   Retx (2, m_2) = "Larger pool of words that could have been associated with each object"

   Where
   m_1 – is a number of words contained in the sentence i=1
   m_2 – is a number of words contained in the sentence i=2

3. Now a set of appropriate actions should be postulated in order to enable understanding the objectives. At first an appropriate an appropriate action value should be postulated and after that an appropriate relation type (see also formulas 3.14a and 3.14b)

Petx (1, m_3) = Action value = "we propose tries to overcome" (3.14a)
Pe (1, m4) = Relation type= "Objective – Actions" (3.14b)

4. After having read the TNL text mission statement and objectives, we ought to have the main idea concerned to the TNL text semantic content.

5. The principal layout of that approach related to TNL semantic content understanding is shown in Fig.3-1.

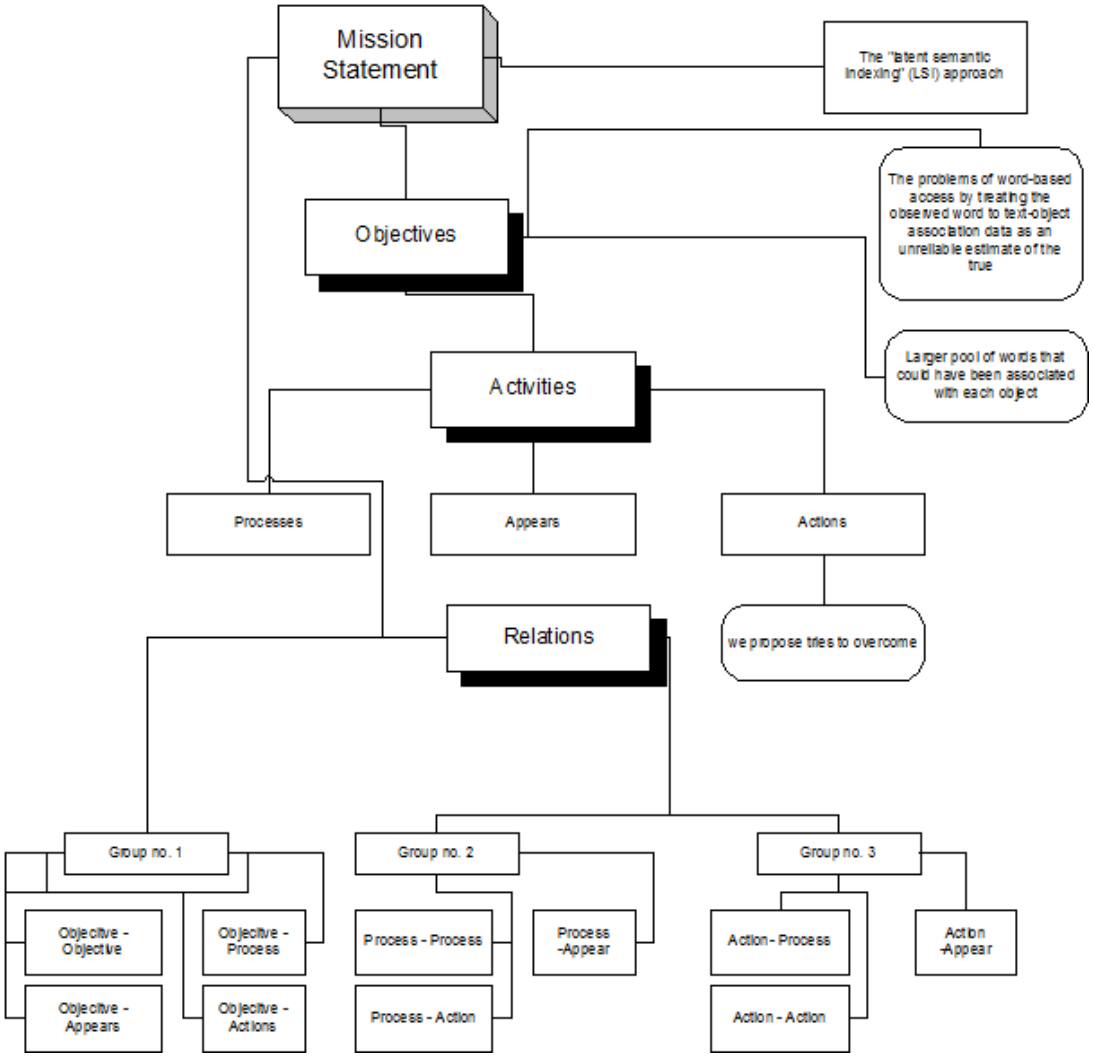

Fig. 3-1
The principal layout of the described approach related to TNL semantic content understanding

3.3 Application of (PTCAL equation) solutions for analysis of TNL texts closely related to education and training materials

The previous section deals with (PTCAL equation) design and implementation for TNL text semantic analysis and contains a concept description closely related the above-mentioned purposes. This section deals with a proposal for communication of authorized user (student or teacher) when providing tests based the article, which the student is reading and after that answering test question in order to confirm if he/she understood the article content correctly.

The user inserts a text string concerned the read article mission statement and the system compare it with Problem Mission Statement (PMST2) Data. If the user's answer is correct he/she is allowed to insert further data concerned to the article objectives. This procedure is denoted as S-Procedure a might be repeated in a loop if necessary. After that, a set of appropriate activities, which contain processes, appears and actions are being activated. The V1 variable value is being increased and after having completed the test this value is compared with value VX contained within Problem Objective (POBJ2) Data. This comparison indicates a percentage of understanding concerned to the read article. The principal schema related to *Application of* **(**PTCAL equation) solutions for analysis of TNL texts closely related to education and training materials is shown in Fig.3-2.

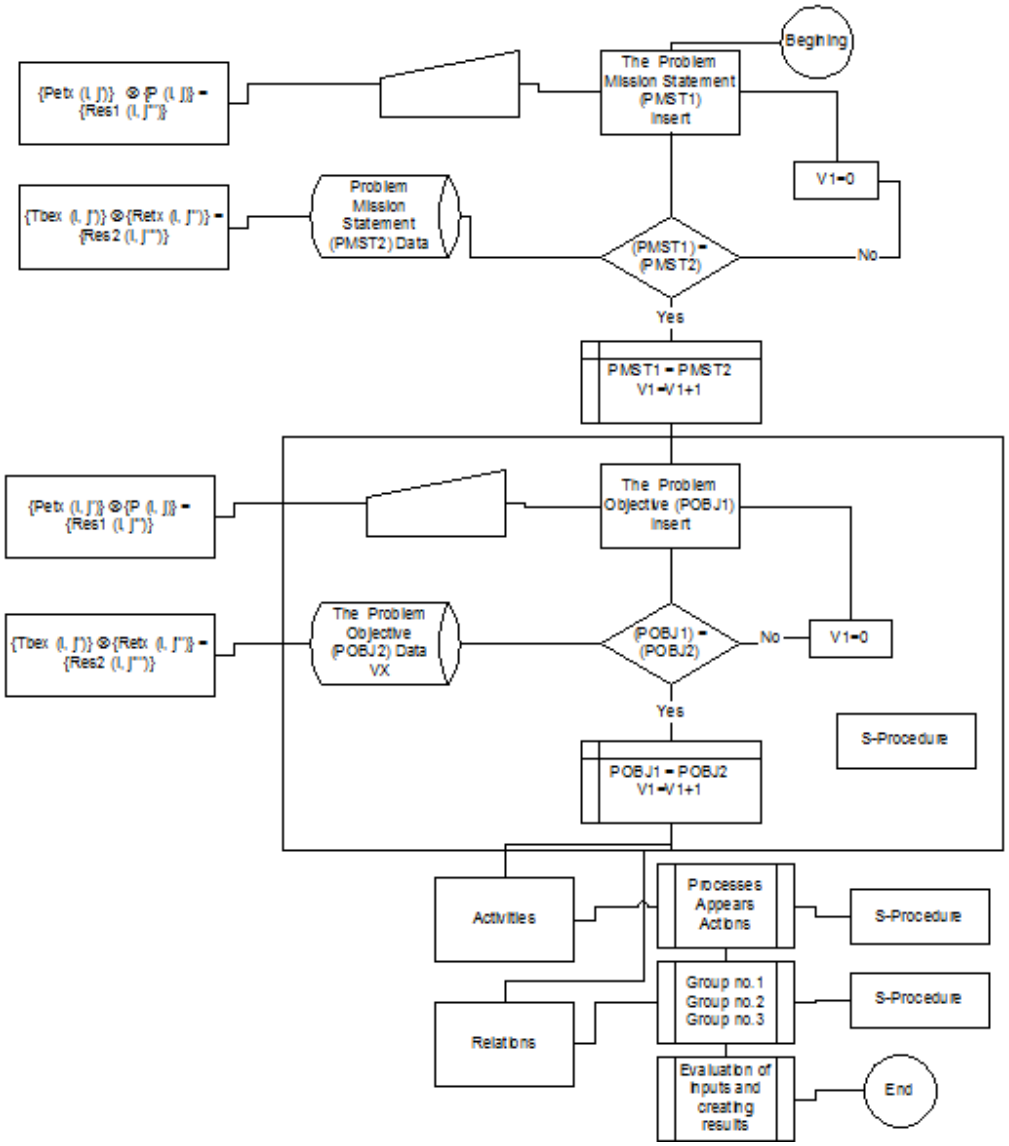

Fig. 3-2
The principal schema related to Application of (PTCAL equation) solutions for analysis of TNL texts closely related to education and training material.

4. Conclusion

The paper's main goal was to prepare a proposal for application of PBPL equation for quantification and modeling the text semantic analysis process, while that modeling results should be applied to training and education purposes. With respect to the above-mentioned paper's main goal and appropriate partial aims, we have defined a way of

PBPL solution for TNL text semantic analysis process. This solution is denoted as *(*PTCAL equation) and creates basis design for TNL text semantic analysis concept, which create basis for implementation of adequate application programs applied in testing of text understanding, which the user (student or teacher) has read before. This paper content represents a set of TNL text quantification rules, which might be applied for finding new solutions of PBPL equation and their application for testing evaluation.

References

ARRONIS, I. 2007. Extracting Quantitative Information from Nonnumeric Marketing Data: An Augmented Latent Semantic Analysis Approach http://www.cfmemory.org/cdm4/results.php?cisoop1=any&cisofield1=cisosearchall&cisoroot=/etd&cisobox1=nonnumeric

ATTEVELDT, W. 2008. Semantic Network Analysis Techniques for Extracting, Representing, and Querying Media Content, Book Surge Publishers, Charleston SC, 2008 p.27-36, ISBN: 1-4392-1136-1

DEERWESTERS, S., DUMAIS, S.T.,. FURNAS, G.W. LANDAUER, T.K., HARSHMAN, R. 1990. Indexing by Latent Semantic Analysis, 1990 http://biblioteca.universia.net/autor/Susan%20T.%20Dumais.html

DUMAIS, S, T. , FURNAS, W.G., LANDAUER, K.,T., DEERWESTER, S., HARSHMAN, H 1990. Using Latent Semantic Analysis to Improve Access to Textual Information http://www.wortschatz.unileipzig.de/~sbordag/aalw05/.../dumais88using.pdf

EGOZI, O., MARKOVITCH, S., GABRILOVICH, E. 2011. Concept-Based Information Retrieval Using Explicit Semantic Analysis http:// www.cs.technion.ac.il/~ofere/.../concept-based-ir-esa-tois11.pdf

ENTMAN, R. M. 1993. Framing: Toward Clarification of a Fractured Paradigm. Journal of Communication, 1993, 43(4):51–58.

HSIEH, H.F., SHANNON, S.E. 2005. Three Approaches To Qualitative Content Analysis. Qualitative Health Research, 2005 15(9), 1277-1288.

FURNAS, G.W., LANDAUER, T.K., GOMEZ, L.M., DUMAIS, S.T. 1999. Statistical Semantics: Analysis of the Potential Performance of Key-Word Information Systems. Bell System Technical Journal, 1999, 62(6), 1753-1806

KABENOV, D.- MURATKHAN, R.-SATYBALDINA, D.- RAZAHOVA, B. 2013. International Journal of Philosophy Study (IJPS) Volume 1 Issue 1, March 2013, www.seipub.org/ijps

KHAKHALIN G., KURBATOV S., NAIDENOVA K., LOBZIN A. 2012. Integration of the Image And Nl-Text Analysis/Synthesis Systems. In Rafael Magdalena (Eds): Intelligent Data Analysis For Real-Life Applications: Theory And Practice, Igi Global, 2012.

KINTSCH, W. 1994. Text Comprehension, Memory, And Learning. American Psychologist , 49 , 294-303.

KLEINNIJENHUIS, J., VAN HOOF, A. M. J., OEGEMA, D., DE RIDDER, J. A. 2007. A Test of Rivaling Approaches to Explain News Effects: News on Issue Positions of Parties, Real-World Developments, Support and Criticism and Success and Failure. Journal of Communication, 57(2):366–384.

KRIPPENDORFF, K. 2004. Content Analysis: An Introduction to its Methodology (SecondEdition). Sage Publications, Thousand Oaks, CA.

MAYRING, P. 2008. Qualitative Content Analysis. Forum:Qualitative Social Research, 1(2). Retrieved July 28, 2008, From http://217.160.35.246/Fqs-Texte/2-00/2-00mayring-E.pdf, 1-10

MILES, M, HUBERMAN, A. 1994. Qualitative Data Analysis: an Expanded Sourcebook. Sage, London, 1994
Patton, M.Q. 2002. Qualitative Research And Evaluation Methods. Thousand Oaks, CA: Sage.

PLEIJTER, A. 2006. Typen En Logica Van Kwalitatieve Inhoudsanalyse In De Communicatie-Wetenschap (Dissertation). Tandem Felix, Ubbergen, 2006 The Netherlands.

ROBSON, R, RAY, F. 2012. Applying Semantic Analysis to Training, Education, and Immersive Learning Interservice/Industry Training, Simulation, and Education Conference (I/Itsec) http://eduworks.com/wp-content/uploads/.../applyingsemantics.pdf

SHAHBAZOVA, SH. 2011. Application of Fuzzy Sets for Control of Students Knowledge. Application Comput. Math., Vol.10.1, Special Issue, 2011; pp.195-208

SHAHRUL AZMAN NOAH, LAILATULQADRI ZAKARIA, ARIFAH CHE ALHADI Extracting and Modeling the Semantic Information Content of Web Documents to Support Semantic Document Retrieval, http://crpit.com/confpapers/crpitv96noah.pdf

STAŠÁK, J. 2004. A Contribution to Semantic Text Analysis In: Electronic Computers And Informatics ECI 2004, The University of Technology Košice, Department of Computers and Informatics of FEI, 22-24. 9.2004 Košice – Herľany, SR, P.132-144, ISBN 80-8073-150-0

STAŠÁK, J., ANDREJČÁK, I., SLÁVIKOVÁ, G., GRELL, M., 2015. Business Process Modeling Linguistic Approach – Problems of Business Strategy Design – Neural Networks (Printed)

STAŠÁK, J. 2006. Modeling of Text Semantic with the use of Fuzzy Sets Ekonomičny Vistnik NTUU, KPI, 2006 (3), P.376 – 384

STAŠÁK, J. 2011. How Image and Text Semantic Analysis Systems May be applied for Educational and Teaching Purposes – Acta Technologica Dubnicae, 2011, Č.1, S.1-18, ISSN 1338-3965

VASSILIOU, A., 2006. Analysing Film Content: A Text-Based Approach Analysing Film Content: A Text-Based Approach, University Of Surrey Guildford, Surrey, Gu2 7xh, UK, 2006 http://www.bbrel.co.uk/pdfs/av_phd.pdf

WOLFE, B.W.M., SCHREINER, M.E., FOLTZ, P.W., KINTSCH, W., LAUNDER, T.K. 1990. Learning from text: Matching Readers and Texts by Latent Semantic Analysis Discourse Processes, 25, 309-336. http://www.lsa.colorado.edu/papers/dp2.wolfe.pdf

Address of authors:

RNDr. Jozef Stašák, PhD.

Vysoká škola technická a ekonomická
v Českých Budějoviciach
Okružní 10

370 01 České Budějovice

E-mail: jozefstasak@yahoo.com

Dr. Ing. Jaroslav Kultan, PhD.

Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky
Dolnozemská 1

852 35 Bratislava

E-mail: jaroslav.kultan@euba.sk