MEST Journal
DOI 10.12709/mest.03.03.01.15

LINGUISTIC ANALYSIS AND MODELLING SEMANTICS OF TEXTUAL CONTENT FOR DIGEST FORMATION

Victoria Vysotska
Lviv Polytechnic National University, Lviv, Ukraine

Lyubomyr Chyrun
Lviv Polytechnic National University, Lviv, Ukraine

JEL Classification: **C02, C6, C61**

Abstract

This article suggests a method of functional and logistical processing of content as a stage of a content life cycle. Commercial content processing method describes the formation and rubrication of electronic digests, and simplifies a commercial content management technology. The main problems of functional processing of commercial content services were analyzed in this paper. The proposed method provides an opportunity to create information resource processing tools as well as implement commercial content management subsystems. This paper presents the generative grammar application in linguistic modelling and analysis for the textual commercial content. In the given article is functional logistic method of commercial content processing as the content life cycle stage in electronic commerce systems proposed. Description of syntax sentence modelling is applied to automate the processes of analysis and synthesis of texts in natural language. The method of commercial content processing describes the information resources forming in electronic commerce systems and automation technology simplifies the commercial content management. In the given article the main problems of electronic content commerce and functional services of commercial content processing are analyzed. The proposed model gives an opportunity to create an instrument of information resources processing in electronic commerce systems and to implement the subsystem of commercial content formation, management and support.

Keywords:
digest, content, content analysis, content monitoring, content search, information resource, electronic content commerce systems, generative grammar, structured scheme sentences, computer linguistic system.

Address of the corresponding author:
Vysotska Victoria
victana@bk.ru

1 INTRODUCTION

Research linguists in the sphere of morphology, morphonology, structural linguistics have identified different patterns for the word forms description (Berko, Vysotska, & Pasichnyk, 2009), (Bolshakova, Lande, Noskov, Klyshynskyy, Peskova, & Yahunova, 2011), (Gladky, 1985), (Gladky, & Melchuk, 1969), (Gladky, 1973), (Chomsky, 1956, 1959, 1961, 1962, 1963), (Chomsky, & Miller, 1958), (Chomsky, & Miller, 1963), (Chomsky, & Schuetzenberger, 1963). With beginning of the development of generated grammars theory linguists have focused not only on the description of the finished word forms, but also the process of their synthesis. In Ukrainian linguists is fruitful research in functional areas such as theoretical problems of morphological description, the classification of morpheme and word creative structure of derivatives in Ukrainian language, regularities for affix combinatory, modeling word-formative mechanism of the modern Ukrainian language in dictionaries of integral type, the principles of internal organization in words, structural organization of different verbs and nouns suffix, word creative motivation problems in the formation of derivatives, the laws of implementing morphological phenomena in Ukrainian word formation, morphological modifications in the verb inflection, morphological processes in word formation and adjectives inflection of modern Ukrainian literary language, textual content analysis and processing, etc.

This dynamic approach of modern linguistics in the analysis morphological level of language with focused attention researcher on developing morphological rules allows to effectively use the results of theoretical research in practice for the computer linguistic systems construction of textual content processing for various purposes (Berko, Vysotska, & Pasichnyk, 2009), (Bolshakova, Lande, Noskov, Klyshynskyy, Peskova, & Yahunova, 2011), (Braychevskyy, & Lande, 2005), (Clifton, 2009), (Grigoriev, & Lande, 2005), (Lande, Furashev, Braychevskyy, & Hryhorev, 2006), (Lande, 2005-2006), (Lande, & Braychevskyy, 2005-2006), (Lande, & Litvin, 2001), (Lande, & Morozov, 2004-2005), (Lande, & Furashev, 2006), (Lande, Furashev, Braychevskyy, & Hryhorev, 2006), (Lande, Furashev , & Grigor'yev, 2006), (Furashev, Lande, Grigor'yev, & Furashev, 2005), (Furashev, Lande, & Brajchevskiy, 2005), (Pasichnyk, Scherbyna, Vysotska, & Shestakevych, 2012). One of the first attempts to apply generated grammars theory for linguistic modeling belongs to A. Gladky and I. Melchuk (Gladky, 1973, 1985), (Gladky, & Melchuk, 1969). Experience and research of A. Gladky and Noam Chomsky (Chomsky, 1956), (Chomsky, 1959, 1961, 1962, 1963), (Chomsky, & Miller, 1958, 1963), (Chomsky, & Schuetzenberger, 1963) are applicable to the tools developing for textual content processing as information retrieval systems, machine translation, textual content annotation, morphological, syntactic and semantic analysis of textual content, educational-didactic system of textual content processing, linguistic support of specialized linguistic software systems, etc.

2 RECENT RESEARCH AND PUBLICATIONS ANALYSIS

At the present stage of development the need of development of general and specialized linguistic systems make active use of applied linguistics and computer information technology. Development of mathematical models for computer speech language systems support allows to realize such tasks of applied linguistics as analysis, synthesis of oral or written text content, description and indexing of text content, texts translation, creation of lexicographical databases, etc. Linguistic analysis of the text content consists of several processes of grapheme, morphological, syntactic and semantic analysis (Gladky, 1973, 1985), (Gladky, & Melchuk, 1969). For each of these stages appropriate models and algorithms were developed (Chomsky, 1956), (Chomsky, 1959, 1961, 1962, 1963), (Chomsky, & Miller, 1958, 1963), (Chomsky, & Schuetzenberger, 1963). An effective tool for linguistic modeling at syntactic and semantic level of language is a main part of combinatorial linguistics - theory of generative grammars, the beginning of which lays in the work of American linguist Noam Chomsky. He used the method of formal analysis of the grammatical structure of phrase to allocate syntactic structure (components) as the structure of the phrase, regardless of its value. The ideas of Noam Chomsky developed the Soviet linguist A. Gladky (1973, 1985), (Gladky, & Melchuk, 1969), using the concept of dependency trees and components systems for modeling language syntax. He proposed a method of syntax modelling using syntax groups that produce phrases components as units of dependency tree building – such representation combines the advantages of direct constituents method and dependency trees.

The advantages of using the simulation of generative grammars is the ability to describe not only the language syntax (rules of sentence formation forms words), but also morphemic (rules of words formation from morphs) and semantic (rules of meaningful sentences and texts formation) levels. It is used to automate the process of inflection/derivation, categorization or key words identification and forming text content digests. For example, when using automatic morphological synthesis computer system creates the necessary linguistic word forms based on requirements to word forms and morphemes databases.

Linguistic analysis of the content consists of three stages: *morphological*, *syntactic* and *semantic* (Bolshakova, Lande, Noskov, Klyshynskyy, Peskova, & Yahunova, 2011), (Gladky, 1973, 1985), (Gladky, & Melchuk, 1969). The purpose of morphological analysis is to obtaining basics (word forms without of inflections) with the values of grammatical categories (eg, part of speech, genus, number, case) for each word forms. There are the *exact* and *approximate* methods of morphological analysis (Bolshakova, Lande, Noskov, Klyshynskyy, Peskova, & Yahunova, 2011). In the exact methods use dictionaries with the basis of words or word forms. In the approximate methods use experimentally established links between fixed letter combinations of word forms and their grammatical meaning.

A dictionary using with word forms in the exact methods simplifies using the morphological analysis. For example, in the Ukrainian language solve the problem of the vowels and consonants letters alternation by changing the conditions of using the word (Bolshakova, Lande, Noskov, Klyshynskyy, Peskova, & Yahunova, 2011). Then for finding the words basics and grammar attributes use algorithms of search in the dictionary and selecting appropriate values. And then use morphological analysis provided the failure to locate the desired word forms in the dictionary. At sufficiently complete thematic dictionaries speed of textual content processing is high, but using the volume of required memory in several times more than using a basics dictionary. Morphological analysis with the use of the basics dictionary is based on inflectional analysis and precise selection of the word bases. The main problem here is related to homonymy the words basis. For debugging check the compatibility of dedicated bases in words and its flexion.

As the basis of approximate methods in morphological analysis determines the grammatical class of words by the end letters and letter combinations (Bolshakova, Lande, Noskov, Klyshynskyy, Peskova, & Yahunova, 2011),. At first allocate stemming from basis words. From ending word sequentially take away by one letter after another and obtained letter combinations are compared with a inflections list of appropriate grammatical class (Gladky, 1973, 1985), (Gladky, & Melchuk, 1969). Upon receipt of the coincidence of final part with words is defined as its basis. In conducting morphological analysis arise ambiguity of grammatical information determination, that disappear after parsing. The task of syntactic analysis is parsing sentences based on the data from the dictionary. At this stage allocate noun, verb, adjective, etc., between which indicate links in the form of dependency tree.

V – is finite not empty set, the alphabet; *T* – the subset of *V*, its elements are terminal (main) lexical units, terminals; *S* – the initial symbol ($S \in V$); *P* – is a finite set of productions (conversion rules) $\xi \to \eta$, where ξ and η – strings over *V*. The set $V \setminus T$ is denoted by *N*, its elements are non-terminal lexical units, non-terminals (Gladky, 1973, 1985), (Gladky, & Melchuk, 1969). Grammars are classified by the types of productions, which imposed certain restrictions.

1. Grammar G_0 is unlimited. Here ξ - random string that contains at least one non-terminal symbol, η - random string over V.
2. Context-sensitive grammar G_1. In the set of productions *P* exists production $\gamma\xi\delta \to \gamma\eta\delta$, $|\xi| \le |\eta|$ (but not in the form $\xi \to \eta$), then you can substitute ξ with ν only surrounded by strings $\gamma \ldots \delta$, i.e. in appropriate context.
3. Context-free grammar G_2. A non-terminal *A* on the left side of $A \to \eta$ production can be

replaced by string in random environment whenever it occurs, i.e. regardless of the context.

4. Regular grammar G_3. May occur productions $A \to aB$, $A \to a$, $S \to \lambda$, only, where, B – non-terminal, a - terminal, λ - empty string.

Terminal lexical units are word forms of natural language, nonterminal lexical units are syntactic categories, and terminal strings are correct expressions of the language (Chomsky, 1956, 1959, 1961, 1962, 1963), (Chomsky, & Miller, 1958, 1963), (Chomsky, & Schuetzenberger, 1963). Then production of the expression naturally interpreted as its syntactic structure, which is given in terms of generative grammar. The set of natural language expressions has a number of specific properties. Analyzing natural language expressions in the theory of formal grammars, they are considered as strings of word forms / morphemes as terminal lexical items. To set expression recognition algorithm exists, or submitted string is an expression of the language. Sets, for which recognition algorithms exists are recursive. But for the generation of natural language expressions, and only for them, for grammar impose restrictions on production: in production $A \to B$ string B is no shorter than string *A*; then while production strings are not reduced.

Grammar G_0 does not meet the specified limit - there are productions that reduce strings (Gladky, 1973, 1985), (Gladky, & Melchuk, 1969). However, language $L(G_0)$ is recursive. Languages generated by not contractile grammar, are easily recognizable. In the context-sensitive grammar G_1 exists a production $\gamma A\delta \to \gamma\eta\delta$, wherein at least one of the strings γ, different from Λ, and nonterminal A is substituted by string η surrounded only by γ and δ, i.e. in context. Language is context-sensitive, if there is at least one context-sensitive grammar that generates the language.

The term rules formation is borrowed from mathematical logic, where it refers to the rules of correct formulas construction. In logics a different type of rules is considered - the rules of conversion. They set certain correlation between valid formulas. Production rules are needed in the natural languages description. Setting transformation rules means a transition to a higher level language examination, namely the semantic level. Knowing language necessarily implies the ability to not only build the right phrase, but the switch from one sentence to another, or completely synonymous to it, or that differ from it by the sense of a certain amount, for example, to make an affirmative sentence negative or interrogative, to change active voice into passive, change the stylistic color of text, to express the same thought in different ways and so on. These features can not be stated in terms of grammars, and therefore raises the question of developing a formal system for transformation rules regarding natural languages. The corresponding problem was first stated in works of Noam Chomsky. His concept quickly gained the fame under the name of transformational grammar: an introduction of semantic level of language description. In fact, all invariant transformations usually makes sense, transformation – it is a change that preserves meaning. Thus, the transformation theory is a theory language synonymy. Description of synonymy in linguistics must take a central place. Hence the primary role of transformation appears. However, transformation does not belong to the same level as that of grammar: G_0 grammar is related to syntax and transformation - to semantic. That is insufficient to describe the grammar G_0 of language meaning is not true in the sense of grammar G_0 coverage semantic level. At the syntactic level grammar is fundamentally quite sufficient. Generative grammar is viewed within the formal theory. For transformations level of formalization is not achieved: transformation rules are not formulated in terms of a simple operation. The task of further formalization of transformations is very important. In the works of Noam Chomsky and several other authors (Gladky, 1985), (Gladky, & Melchuk, 1969), (Gladky, 1973), (Chomsky, 1956), (Chomsky, 1959), (Chomsky, 1961), (Chomsky, 1962), (Chomsky, 1963), (Chomsky, & Miller, 1958), (Chomsky, & Miller, 1963), (Chomsky, & Schuetzenberger, 1963) term generative grammar is used in two senses: in the broad - to refer to any system of formal rules describing the language, including transformation and morphonological components and narrow - to describe exactly grammars. In this work the term is always in a narrow sense. With such discourse transformation rules are beyond the generative grammar.

3 STATEMENT OF PURPOSE

Submission of syntactic structure in terms of generative grammar is often used in linguistics and studied many times in many different ways. It has won the right to exist both in theoretical terms and in experimental works (automatic translation or abstracting, etc.). Grammars while generating terminal strings, such as expressions of natural language, also giving their structure. Unlimited not shortens grammar has no longer the property of expressions comparison with their context-sensitive structure. In this grammar each time more than one lexical unit is replaced, but a group of them. In the derivation, the parent of each lexical unit cannot be uniquely, and so production rule applying is not converted in a context-dependent structure.

Linguistic software is used in many information systems (Berko, Vysotska, & Pasichnyk, 2009). Improving machine-human communication is an important current task, which is solved through the texts analysis and synthesis on the linguistic level. For this purpose, consider the process of linguistic phrases modelling in natural language by generative grammars.

For automated content retrieval and processing of textual content is of great importance to the presence/absence and frequency of occurrence of a particular category of linguistic units in the studied array of content. Quantitative calculation allows to draw objective conclusions about the direction of the content by the number of analysis units (key quotes) in the investigated arrays, for example, the number of positive/negative reviews on a certain type of product. Qualitative analysis allows to draw objective conclusions about the availability of desired linguistic units in the array of content and the direction of its context. Content search is performed not on the text content, but for its brief characteristics – the retrieval images (RIm), where the main text of content is served in terms of specialized information retrieval language. The procedure for the RIm provides indexing, semantic analysis of the main text content and translating it into information retrieval language (table. 1).

Table 1. The main stages in content-search operation

Operation name	*Operation description*
RIm Formation	Create, input, storage in RIm module
Query generation and SA	Create, input and storage in the queries module and SA of a user's query.
Content search	Comparison RIm of content with SA of user request
Content analysis	Quantitative and qualitative analysis of text content.
Result formation	The result of applying content analysis of positive content in range (0,7; 1] or (0,5; 1].
Decision making	The decision on issuing of the content according to the result of applying content analysis.
Content presentation	The issue of the content that corresponds to the information request of the user.

In the module are stored no texts content, but its RIm. To search the indexed content is used content analysis in information requests (Boiko, 2005), (Doyle, 2005), (Hackos, 2002), (Halverson, 2009), (McGovern, & Norton, 2001), (McKeever, 2003), (Nakano, 2002), (Rockley, & Cooper, 2002), (Stone, 2003), Salton, 1979), (Papka, 1999), (Zipf 1935, 1949). Information request translated into information retrieval language and supplemented to search for additional data, is a search available (SA). The degree of detail in the presentation of content in RIm of its central theme/subject and related topics/subjects is the depth of indexing. Automating this process allows you to ensure its unification, dismissing the part of staff from unproductive labour indexing content. Content search is provided by a set of semantic tools: information retrieval language, methods, content indexing/query and search. Based semantic tools is information retrieval language - specialized artificial language designed to describe the central themes/subjects and formal characteristics of the content, as well as to describe queries and search. In practice, one language is used for indexing content and the other for indexing information requests.

Content formatting is the process of indexing, semantic analysis, the basic definition of the content and convert it into XML format. The formatting of the content is performed manually by a moderator or automatically by means of content analysis. While indexing, examine the text content, determine its central theme and describe it in terms of information retrieval language. In the content section titles, as a rule, reveal a Central theme and subject, but the name is not always possible to identify the content. Natural language is not used as information retrieval through numerous grammatical inclusions, lack of structuring, ambiguity and greater redundancy, in particular, for the Ukrainian language 75-80 %. In information retrieval language among the major elements (table. 2) do not use synonyms and homonyms through their semantic ambiguity.

Table 2. The main elements of an information retrieval language

Element name	*Language unit characteristics*
Alphabet	Set of graphic characters to commit the words and expressions of the language.
Vocabulary	Set of interrelated linguistic units (words used in speech).
Grammar	Set of the rules of association of linguistic units in word, which is most effective means of building sentences.
Paradigm	Lexico-semantic group of words with subject-logical links based on semantic criteria.
Paradigmatic relations	Basic and analytical relation between words, with no depenence on the context in which they are used, generated with not linguistic, but logical relationships.
Syntagmatic relations	Linear relationships between words that are settled when combining words into phrases and phrases.
Identification rules index	Paradigms (vocabulary) and syntagmatics (grammar) of the language.
Statements unity	A statement is a sentence of natural language, but the reverse is not true.
Interphrase unit	United semantically and syntactically in the fragment. The core interphrase unity is a statement that is not subordinated to any other statement and saves sense when selecting from the context.
Blocks-fragments	Many interphrase unities that ensure the integrity of the text by semantic and thematic links.

The feasibility of using information retrieval languages depends on destination of search tools, technical tools, automation of information procedures and management. When designing information retrieval languages pay attention to the following points: the nature of the industry/theme for which you are developing language features of texts from the search array content; the nature of the information needs of users of electronic content commerce.

4 RESEARCH RESULTS ANALYSIS

The text content (article, comment, book, etc.) contains a lot of data in natural language, some of which are abstract. The text is considered as a sequence of iconic pieces, unified by content, the basic properties of which are informational, structural and communicative coherence/ integrity that reflects the content / structural nature of the text. The method of text content processing is a linguistic analysis of content (eg, comments, forums, articles, etc.). The process of the text content elaboration divides content on tokens using finite automates of linguistic analysis of natural language texts (Fig. 1).

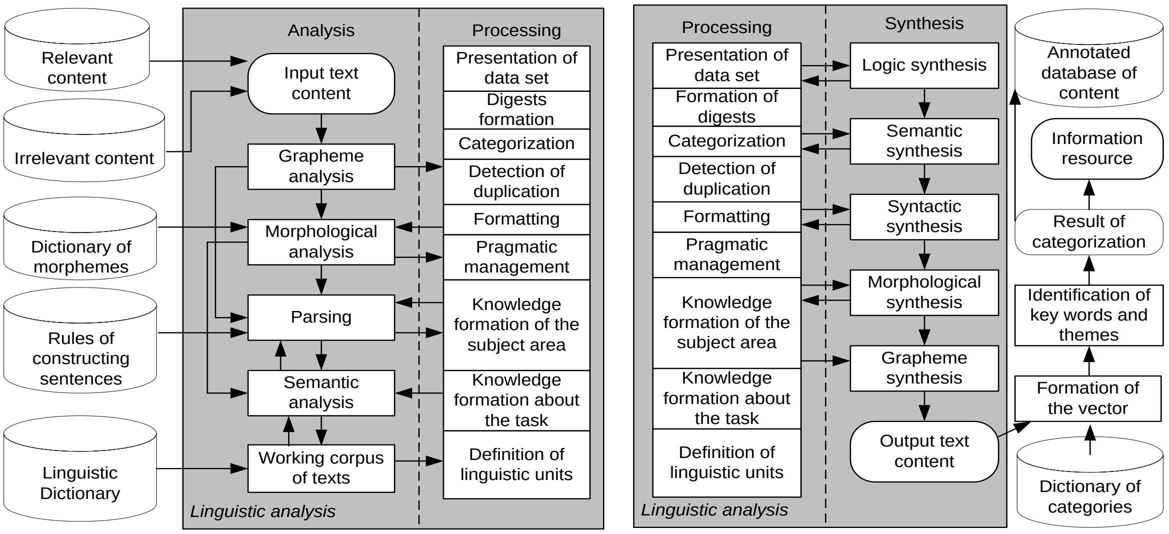


Fig. 1. Block diagram of the linguistic analysis of text content in the formation content

As functional, semantic and structural unity text content has rules of construction, detects patterns of meaningful and formal connections of constituent units. Connectivity of text content is shown by exterior structural indicators and formal dependence of the text components, and text content integrity - through thematic, conceptual and modal dependence of textual information. The integrity of text content leads to meaningful and communicative organization of the text, and text content connectivity - to form, structural organization of textual information. Let us consider not reductive grammar *G* with linguistic units and string length of the terminal linguistic units *p* of this grammar. Language L(G) is easily recognizable set by algorithm (alg. 1), which builds any constructions in the grammar starting with initial symbol S. The number of productions occurrences in the derivation is its length.

Algorithm 1. Language L(G) recognition

1st stage. Alternately apply S to grammar productions.

2st stage. Built string *x* verifying.

1st step. If yes, then go to 3st stage.

2st step. If not, then go to 1st stage.

3st stage. String *x* of *S* output.

In this algorithm, process of output can be infinite. To avoid this algorithm derived finite set is defined (alg. 2).

Algorithm 2. Recognition of L(G) language using the set of derivations M

1st stage. Analysis of string *x* with lengths *n*, derived from the initial symbol *S* of grammar *G*.

1st step. Calculate the number of strings from *S* to *x* (no string is repeated). Since the grammar *G* is not contractile, whereas none of the strings in this sequence is no longer than the string (length ≤*n*). Number of different strings length ≤*n* of *p* lexical units ≤

$$p^n + p^{n-1} + p^{n-2} + ... + p^2 + p^1 + p^0 = \frac{p^{n+1} - 1}{p-1} < p^{n+1}$$

when $p > 1$ (number of strings of length *n* of *p* lexical items equals to p^n, length of *(n-1)* of *p* equals to p^{n-1}, and so on; number of strings of 0 symbols is equal to $p^0 = 1$). At $p^{n+1} = P$ of various strings such sequences is not more than

$$P! + C_P^1 \cdot (P-1)! + C_P^2 \cdot (P-2)! + ... + C_P^{P-2} \cdot 2! + C_P^{P-1} \cdot 1!$$

Here the sum consists of p summands, its *k*-th summand equals to

$$C_P^k (P-k)! = {}^{P!}\!/_{k!(P-k)!} (P-k)! = {}^{P!}\!/_{k!} \leq P!\,,$$

where $C_P^k = {}^{P!}\!/_{k!(P-k)!}$.

The total sum is no more than $P! \cdot P < (P+1)! = (p^{n+1} + 1)! < (p^{n+2})!$.

2st step. Generate a sequence of strings from *S* to *x*. From the $(p^{n+2})!$ derived sequences set *M* is obtained.

2st stage. Construction of derivation set *M* of random string *x*.

3st stage. Check the resulting finite set of derivations *M*.

1st step. Find an appropriate derivation I set *M* or prove such derivation does not exist. If the derivation does not end in a string *x*, then go to 3st step.

2st step. If the derivation ends in a string *x*, then go to step 4.

3st step. If the end of the derivation set, then go to 5st stage, otherwise move to 3st stage.

4st stage. Formulating a positive answer: *x* is derived from *S*. Go to 5st stage.

5st stage. Formulating negative answer: *x* is not derived from *S*.

6st stage. Display the results.

Number of steps to form the set M to fins an appropriate derivation does not exceed $(p^{n+2})!$ and is great for natural languages. This calls large capacity for such a resource-intensive algorithm. It is therefore necessary to choose a set where the number of steps in the recognition is in that depending on the length of the string, and identify rather narrow class class of recursive sets. And assuming that the set of expressions is infinite, their processing rules are more homogeneous, which can reveal significant patterns of derivation.

For more precise derivation of string *x* of *S* must still add an additional constraint: each production $X \to Y$ on tne left side is Z_1CZ_2 (*C* - a lexical unit), and the right part (*Y*) - form Z_1WZ_2 (*W* - is a non-empty string). Then at every step of derivation is allowed to replace only one lexical unit. For any not contractility grammar an equivalent context-dependent grammar can be constructed, for example,

$$P_1 = \{AB \to BA\} \approx$$
$$\approx P_2\{AB \to 1B, 1B \to 12, 12 \to B2, B2 \to BA\}$$

Consistent use of these rules is equivalent to the use of rule $AB \to BA$ and replace them last not lead to the appearance of extra reduced, as lexical units 1 and 2 are new. In grammar *G* rules replace only one lexical unit (*C*). The left part of the production (*X*) does not necessarily consist only of a lexical unit. Around *C* may attend other lexical items (context), ie $X = Z_1CZ_2$. Then the production $Z_1CZ_2 \to Z_1WZ_2$ means a permit to replace C into W only in terms of context $Z_1..Z_2$ and without changing its location.

By introducing a new restriction (right side of any production containing not more than two lexical units) a new class of context-free grammars is formed, where the syntactic structure of expressions in constructing a tree with each structure is obtained not more than two branches. That expression always divided into two parts (eg, *nominal group* + *verb group*), each of these halves again divided in half and so on. But the binary representation of natural language expressions are not always satisfactory and natural in terms of meaningful linguistic interpretation. Criteria for selecting the appropriate descriptions are beyond theory: the choice is made on the basis of considerations relating to the specific objectives and the nature of the task. Since the number of lexical items in the right part of production is already minimal, impose restrictions on the nature of lexical units that replace (if the right side of each product consisted of one lexical unit or has the form *bB*, where – *b* terminal lexical unit, and *B* – syntactic category). This restriction specifies the string output, but requires a lot of power for computation.

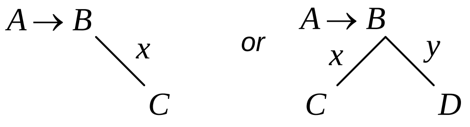
The grammar has the following important property in the linguistic aspect. The terminal symbols as word forms from natural language are interpreted. The supporting characters as syntactic categories (eg, V verb, S noun, A adjective, $\tilde{V}$ verb group, $\tilde{S}$ group noun), the initial character as R (sentence) and displayed terminal strings as a correct sentence of the language are interpreted. Unlimited grammar of type 0 are a special case of the general concept in generative grammar. But, they are certainly adequate for describing any natural language in full. Every natural language (set of correct sentences) is easily authentication set. This means the existence of fairly simple recognition algorithm of phrases correctness.

If natural language is recognized algorithm with with the specified restriction on the amount of memory, then it can be the generative grammar. Here, for displayed any terminal string of n length is a conclusion which none intermediate chain does not exceed the length of Kn (K is some constant). This grammar is *grammar with limited stretching* where capacitive signal function do not more than linear. For any grammar with limited stretching can build its equivalent grammar G_0. It is able to describe the set of correct sentences for any natural language, ie generate any correct phrases of this language, while do not generating any false phrases. Both structures, presented as examples unsuitability of context-free grammars, G_0 grammar easily is described.

The method disadvantages of generative grammar in three points are described.

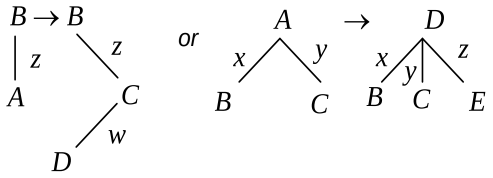
1. With their help, it is not possible naturally to describe phrases with discontinuous constituents.
2. The grammar G_0 contains only rules of linguistic expressions formation, such as word forms or phrases. The grammar specifies the correct expressions in contrast to incorrect.
3. Grammar G_0 are building sentences just with exactly certain word order (with the fact that this sentence should have in final form). In this case generated sentence is matched syntactic structure in the form of an ordered tree, ie a tree where between nodes (except subordination relation, defined by a tree) there is also a relation of linear order (to the right - to the left). Thus, in the syntactic structure of G_0 grammar are not separated two absolutely different by nature, though interconnected relationships: syntactic subordination and linear relative position. But possible describe the syntactic structure showing relation of syntactic subordination. As for the relation of linear order, it describes not the structure, but a phrase.

The words order depends on syntactic structure. It necessarily from its accounting is determined and thus is in relation to its somewhat derivative, secondary. It is expedient to modify the concept of generative grammar so that the left and right parts of substitution rules were not linearly ordered strings, eg as trees (no linear ordering), depicting the syntactic relation (Gladky, 1973, 1985), (Gladky, & Melchuk, 1969). Then the rules are as follows:



analogue of context-free rules

or



analogue of the rules for grammar of type 0

Lines with symbols represent the syntactic relations of different types. Letters $A, B, C, \ldots$ represent syntactic category NB: relative positions of symbols for one level of subordination does not play any role and in this scheme is randomly; $B\xleftarrow{x}A\xrightarrow{y}C$ means the same thing as $C\xleftarrow{y}A\xrightarrow{x}B$ (Gladky, 1973, 1985), (Gladky, & Melchuk, 1969), (Chomsky, 1956, 1959, 1961, 1962, 1963), (Chomsky, & Miller, 1958, 1963), (Chomsky, & Schuetzenberger, 1963). As a result the syntactic structures (not phrases) calculation in language is obtained. This calculation is part of the generative grammar. The rest of this grammar is the calculation that sets all possible its linear sequences of words for any given syntactic structure (including any other factors, eg with the mandatory accounting logical selection in the Ukrainian language, etc.). Then is removed the problem of discontinuous constituents (Gladky, 1985), (Gladky, & Melchuk, 1969), (Gladky, 1973). From the output sentence in a regular grammar can not get the natural presentation of immediate constituents structure with this sentence. That is, regular grammars provide some structure constituents like all grammars immediate constituents, however, these components are usually formal. In the analysis explore multilevel structure for textual content: a linear sequence of symbols; linear sequence of morphological structures; linear sequence of sentences; network of interconnected unities (Alg. 3).

Algorithm 3. The linguistic analysis of the textual content

1st stage. The grammatical analysis of textual content.

1st step. Separation of commercial textual content on sentences and paragraphs.

2st step. String separation of symbols into words.

3st step. Bold of digits numbers, dates, constant phrases and abbreviations.

4st step. Removing the non-text symbols.

5st step. Formation and analysis of a linear sequence of words with the special characters.

2st stage. Morphological analysis of the textual content.

1st step. Getting the basics (word forms from cut off their end).

2st step. Each word form is assigned a value of grammatical categories (the set of grammatical meanings: gender, case, declension, etc.).

3st step. Linear sequence formation of morphological structures.

3st stage. Parsing textual content.

4st stage. Semantic analysis of textual content.

1st step. Words are correlated with semantic classes from dictionary.

2st step. Selection of needed morphologically semantic alternatives for the sentence.

3st step. Linking words into a single structure.

4st step. An ordered set formation of superposition records from basis lexical functions and semantic classes. The result accuracy is determined by dictionary completeness/correctness.

5st stage. Referential analysis to form between phrasal unities.

1st step. Contextual analysis of content. According to his help is realized local reference resolution (this, that, it) and expression selection as the unity core.

2st step. Thematic analysis. Expression separation on the topic identifies thematic structure using, for example, content categorization and the digest formation.

3st step. Definition of regular frequency, synonyms and re-nomination of keywords; reference identification, ie the words ratio with the image object; implication availability, based on situational relations.

6st stage. Structural analysis of text. Prerequisites for using a high degree of unity terms coincidence, discursive unit, sentence of semantic language, expression and elementary discourse unit.

1st step. Basic set identification for rhetorical relations between unities content.

2st step. Nonlinear network construction for unities. A links set openness implies its enlargement and adaptation for the textual structure analysis.

Text construction is determined theme expressed information, terms of communication, task of messages and presentation style. With grammatical, semantic and compositional structure of content is related his stylistic characteristics that depend on individuality author and subordinated thematic/stylistic dominants of text. The main stages of the morphological characters determining for textual units: grammatical classes definition for words (speech parts) and principles of their classification allocation; part separation of the words semantics as morphology unit, set justification for morphological categories and their nature; the set description for formal tools that are attached to parts of speech and their morphological categories. This are used following methods for grammatical meaning expression: synthetic, analytical, analytical and synthetic.

Grammatical values are summarized through the same type of characteristics and are subject on partial values separation. For classes designation of similar grammatical meanings is used grammatical categories concepts. By the morphological values include the category of gender, number, case, person, time, method, class, type, united in paradigm for the text classification. The object of morphological analysis is the structure of words, inflection forms, ways for grammatical meanings expression. Morphological features for textual units are the research tools of communication between lexicon, grammar, using them in speech, paradigms (case forms words) and the syntagmatic (linear relationships of words, expression). The implementation of automatic coding of words in text (ie assigning them codes of grammatical classes) is associated with grammatical classification.

Morphological analysis includes the following steps: bases localization in word form; searching base in the dictionary of bases; the word forms structure comparison with data in dictionaries of the bases, roots, prefixes, suffixes, inflections. In an analysis identify the meaning of words and syntagmatic relations between content words. Analysis tools are a dictionary-based/inflections/ homonyms and statistical/syntactic combinations of words removing lexical homonymy, semantic analysis for nouns with non-prepositional constructions, tables for semantic syntactic combination of nouns/adjectives and component of prepositional structures, analysis algorithms for determination of the checks sequence and the appeals in the dictionary and the tables, words division system in text on inflection and base, thesaurus equivalences fot replacing equivalent words one/several new numbers ofconcepts that serve as identifiers for content instead of words based, thesaurus as a hierarchy of concepts for searching total/associated concepts for this concept, dictionaries service system.

5 METHOD OF ELECTRONIC DIGEST FORMATION

Information support is always needed when solving complex problems in any living environment (Berko, Vysotska, & Pasichnyk, 2009), (Bolshakova, Lande, Noskov, Klyshynskyy, Peskova, & Yahunova, 2011), (Braychevskyy, & Lande, 2005), (Grigoriev, & Lande, 2005), (Lande, Furashev, Braychevskyy, & Hryhorev, 2006), (Lande, 2005-2006), (Lande, & Braychevskyy, 2005-2006), (Lande, & Litvin, 2001), (Lande, & Morozov, 2004-2005), (Lande, & Furashev, 2006), (Lande, Furashev, Braychevskyy, & Hryhorev, 2006), (Lande, Furashev, & Grigor'yev, 2006), (Furashev, Lande, Grigor'yev, & Furashev, 2005), (Furashev, Lande, & Brajchevskiy, 2005), (Pasichnyk, Scherbyna, Vysotska, & Shestakevych, 2012). Meeting the information requirements is a mandatory requirement for innovation realization. At the same time, complexity of obtaining information affects efficiency and quality of solutions. The Internet can be considered as a large-scale mass media. Chaotic nature and existence, absence of clear periodicity maintaining and most sites updating as well as existing problems concerning effective information search do not allow using the Internet as a single solid mass media yet. Only certain network elements (which are often called network or the Internet media) are considered to be full-fledged mass media. Network mass media considered to be news portals with a certain updating frequency, electronic versions of printed periodical publications as well as newspapers and magazines existing in online electronic format. A constant growth of information production rates is one of the main features. Apart from increasing the amounts of information to the scales that render impossible its direct processing, there were a number of specific problems associated with information technology rapid development. Therefore, there is a quite powerful array of information (the Internet resources) for decision making in various public living environments, society and the individual on the one hand as well as a lack of information which is necessary for decision-making because of its dynamics, volumes, sources, and unstructured nature. Coverage and generalization of large dynamic information flows, which are continuously generated in the media, demands qualitatively new approaches.

The situation related to the sharp growth of information production rate has resulted in a number of problems:

- a disproportionate growth of information noise due to poor information structuring;
- an emergence of parasitic information (received as applications);
- an inconsistency between formally relevant information (which is thematically corresponding) and consumers' real needs;
- a multiple duplication of information (a typical example of which is publication of a certain message in various editions).

It is an undisputed fact that an exponential growth of information production rates significantly reduces the efficiency of information processing by traditional methods. The important data are duplicated on multiple websites, the number of which is increased by an exponential law. The text automated processing programs performing indexing, annotation, abstracting, fragmentation as well as other forms of information analysis and synthesis has been created since the beginning of computer era. At the same time, a niche of automated abstracting systems can not be considered completed. The majority of present annotate making processes are not effective; there is still a real need for scalable methodologies and programs. Nowadays there are many ways to solve the problem, which are clearly divided into two directions – quasi-abstracting and brief summary of source document content. Quasi–abstracting is based on extraction of document fragments, i.e. identifying of the most informative phrases and quasi-abstracts formation. Source material brief summary is based on selection of texts by using artificial intelligence methods, special information languages of the most essential information as well as new texts generation and source documents which are summarized by content. Using such an approach provides a possibility to get more sophisticated annotations which may contain information that supplements source text. Due to the reliance on formal representation of the source document semantics, such systems can be theoretically set to a very high compression ratio needed, for example, to send messages to mobile devices. Therefore it is possible to say that the main difference between the abstracting tools lies in the fact they form a set of extracts or a document brief summary. All existing industrial systems of Text mining class include automated abstracting tools, which are essential components of such systems. One of the basic procedures of this class – automated digest formation – is presented like automated abstracting which is based on a large number of documents. The documents, in which all input stream trends are reflected in the most clearly way, are selected to the digest. One can state that such digests should fully correspond information needs of the user whose request generates this incoming information flow.

Based on the abstract, which contains by its volume a small part of the source text, there is a possibility for users to make a reasonable opinion concerning source document spending significantly less efforts compared to complete reading. As a general rule, the volume of abstract used in automated abstracting should constitute from 5 % to 30 % of the source text. Preparing documents which represent the annotations from several sources (i.e. digests) offers even greater compression ratio.

The process of quasi-abstracting is reduced to extraction (removal) of the minimum relevant fragments from documents. At the same time, compared with a brief summary, it has some special feature which is based on synthetic surfactant relations analysis of lexical items presented in text. Quasi-abstracting focuses on specific fragments accentuation by phrase patterns comparing resulting in formation of blocks with the largest lexical and statistical relevancy. Frequency auto detection of certain words and combinations usage in the source document allows determining paragraphs and proposals, in which document theme is represented in the most accurate way. Final document formation is performed by selected fragments interaction. Thus the formed quasi–abstract gives the impression of a coherent text that significantly facilitates its perception. Nevertheless, in some cases, the quality of abstracting largely depends on genre of the text, which is being worked out. Quasi-abstract content-richness also depends on other features of the source text. It is practically impossible to create qualitative abstract from fragments of the source document for large texts, monographs or interviews without taking into account the semantic regularities. The basis of the quasi–abstracting analytical phase is a calculation of weighting coefficients for each text block in accordance with such characteristics as unit location in source, frequency of appearance in the text, frequency of key phrases usage along with some other indicators.

The process of indexing depends on the descriptor dictionary or information retrieval thesaurus. Descriptive dictionary is structured table with three columns: basis of words; descriptors sets assigned to each basis; grammatical features for descriptors. Indexing consists of an informative word combinations selection from the text; deciphering abbreviations; replacing words with base-descriptors on descriptor code; removal of homonymy. In digest formation using content analysis with regard to frequency weights of words from the concepts dictionary generated. The digest formation consists of algorithms of the concepts dictionary forming (alg. 4), of the content duplicate definition (alg. 5) and of digest create (alg. 6).

Algorithm 4. Concepts thematic dictionary formation

1st stage. Concepts frequency vocabulary formation.

1st step. Sequential selection of all words in input content.

2st step. Alphabetical-frequency vocabulary construction based on content categories.

3st step. Words normalization through automatic morphological analysis.

4st step. Alphabetical-frequency vocabulary modification.

5st step. Words assign of weight w (use frequency).

6st step. Insignificant words deleting from the alphabetical-frequency vocabulary ($W \leq k$, where k is threshold value of word extracting).

2st stage. The thematic dictionary choice as requested.

3st stage. Words weight adjustment of alphabetical-frequency vocabulary dictionary based on thematic dictionary.

4st stage. Words choosing $N = n$ with more w of frequency vocabulary (moderator determines $n = const$).

Algorithm 5. Duplicate content determination

1st stage. The initial data formation.

1st step. Moderator introduced of words in string $m = const$.

2st step. Moderator input strings unique coefficient $U = const$.

3st step. Coefficient limits formation of the keywords use ie $K = [a_1, a_2]$, where $a_1 = const$ and $a_2 = const$.

4st step. Content partitioning on n chains of m words.

5st step. Frequency calculation of k_i keywords use.

2st stage. Content duplicates determination.

1st step. Words strings comparison for all content.

2st step. Chains uniqueness coefficients calculation u_i.

3st step. Chains uniqueness coefficients comparison u_i with U. At $n^{-1}\sum u_i < U$ mark the content as unsuitable.

4st step. The frequency comparison k_i with coefficient K. If $k_i < a_1$ or $k_i > a_{21}$ then a content mark as unsuitable.

Algorithm 6. A digest create

1st stage. Content Select based on its weight.

1st step. Digest size C input.

2st step. The algorithm 1 implementation.

3st step. The weight consistent determine of each content as the weights sum of its individual words that $W = \sum w_i$.

4st step. The input content stream sort from the weights values.

5st step. Meaningful content duplicates definition for statistical criterion of text uniqueness $U \geq 0{,}9$ (alg. 5).

6st step. Content filter of unsuitable for digests building (when $W \leq l$, where l is content removal threshold value by the self-education rules of content structuring and moderating) and statistically substantial duplicates.

7st step. The choice of $V = q$ content with greater weight where $q = const$ and the moderator given.

2st stage. Digest text construction of selected content.

1st step. Dictionary construction of selected content (alg. 6).

2st step. Content analysis application to the text.

3st step. Sentences filtration that do not meet the semantic rules of content structuring and moderating.

4st step. Hypertext presentation formation of digest, its contents and a link to the original source.

3st stage. Generated text edit of digest.

1st step. The check amount of generated content c_i . If $c_i < C$, then 2st step, otherwise 4st stage.

2st step. Content delete from input stream for digest formation.

3st step. 1st -2st steps implementation.

4st step. Resulting append to pre-formed digest and goto 1st step.

4st stage. Digest text formation as a separate content and its maintaining in the database with reference on the source.

A digest is the annotated text based on the analysis of several documents. While compiling digests automated abstracting methods of the one document apply to an array with a large number of documents. The majority of document automated abstracting algorithms consist of three basic stages: source text analysis, significant fragments definition (suggestions or whole paragraphs) and conclusion formation. At the same time, a digest may also be considered as a source of annotated hyperlinks to the underlying documents. While forming out digests the usage of quasi-abstracting methods makes coherent text obtaining almost impossible. Combining abstracts of each document will contain an excessive amount of incoherent information. However, while forming out the auto abstract which consists of a certain number of source document announcements, and which is divided into subsections, the usage of above mentioned method is quite acceptable (Berko, Vysotska, & Pasichnyk, 2009), (Pasichnyk, Scherbyna, Vysotska, & Shestakevych, 2012). Content monitoring is information flow semantic analysis which is carried out in order to get required qualitative and quantitative sections. In contradistinction to content analysis, it is performed continuously in time. Continuous analytical processing of messages is the most characteristic peculiarity of this approach, which enables to extract facts from texts as well as detect new terms and generate various statistical reports. Nowadays the named tasks are covered by two major technologies – factographic information extraction from texts (Information Extraction) and profound analysis of texts (Text Mining) (Berko, Vysotska, & Pasichnyk, 2009). Content monitoring method is an adaptation of content analysis classical methods to the dynamic information files conditions, i.e. information flows from the Internet. In contradistinction to information, integration systems, which implement the idea of collecting and gathering all accessible information from both internal and external sources, can detect non-obvious regularities in documentary data arrays or texts – so called latent (hidden) knowledge. Systems of this class allow carrying out the analysis of large volumes of documents as well as creating index of notions and topics highlighted in these documents. A typical problem of content monitoring is charting of concept emergence dynamics over time (Berko, Vysotska, & Pasichnyk, 2009), (Pasichnyk, Scherbyna, Vysotska, & Shestakevych, 2012). The automated content monitoring technology has several important peculiarities:

- usage of publication key fragment as a forming unit of text information array;
- forming of publication key fragments is the union of two interrelated automated processes: analytical and synthetic processing as well as multilevel content analysis procedure of publication texts;
- indexation of key fragments publications is performed via multi-faceted classification.

The uniqueness of the proposed technology lies in uniting of content analysis semantic and quantitative methods. The sequence of stages of semantic problem analysis, which is investigated by specific information system, can be conditionally divided into semantic (qualitative) analysis of the publication totality and formal (quantitative) analysis of information arrays: index, bibliography and text array of publication key fragments (Berko, Vysotska, & Pasichnyk, 2009), (Pasichnyk, Scherbyna, Vysotska, & Shestakevych, 2012). There is a sufficient number of analytical systems presented in the software market focused on mathematical and statistical analysis of different quantitative and digital parameters. However a huge amount of text information contained in print media, information agencies news lines as well as the Internet content sites does not have any qualitative tool designated for analysis. This is precisely why working in this area acquires actuality.

The system’s final goal is material brief summary automatic production (so called electronic publication digest in the media), i.e. an extract of the most important information out of one or more documents and laconic information-rich reports generation based on them. The system should carry out information monitoring by getting large amounts of data as well as analyze, systematize them using automatic categories, accumulate information by indexing material and placing it into a database, solve the problem of content filtration and generate digests automatically. Selection of final unified compromise solution taking into account various criteria is sufficiently complex task during planning and decision-making process. So, it makes sense to present significant information selection problem in a hierarchical form using hierarchy analysis method. Information is selected by the algorithm which works out incoming text as well as identifies the most important information parts and fragments. Consequently, at the top level of hierarchy is located the goal – significant information selection. The goal specifying criteria are located on the second level: text basis, glossary completeness, key terms amount. Selection alternatives – incoming texts fragments – are located on the third level. (Fig. 2.)

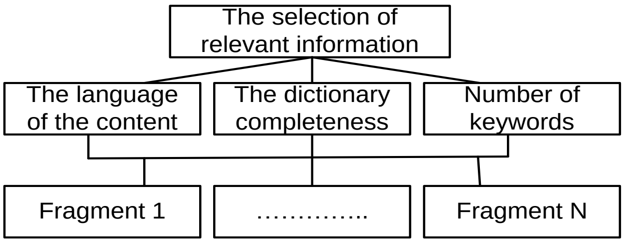

The selection of relevant information
The language of the content
The dictionary completeness
Number of keywords
Fragment 1
…………..
Fragment N

Fig. 2. Task layering

Use-case diagram represents various scenarios of interaction between actors (users) and precedents (use cases) as well as describes system functional aspects. The use-case diagram of automated formation and digest categorization system for electronic media is presented on Fig. 3. The article prepared by journalist is processed by a system in consequence of which terms statistical values are defined. Thematic classification makes it possible to include an article to a certain column. After that a digest is formed by Text Mining algorithms. In order to visualize system statistical aspects class diagram is plotted. System describing class diagram is presented on Fig.4. The class “content” is presented as a part of the class “analysis”, which is a part of the class “rubricator”. The class “glossary” is presented as a part of the class “rubricator” and “analysis”. The state diagram of automated formation and digest categorization system for electronic media is presented on Fig. 5. The finite state automation with simple states and transitions is presented on it. Activity diagram is a diagram showing some activity decomposition into its components (Fig. 5). Activity means treatment specification, which is performed, in coordinated sequential and parallel execution of subordinated elements – embedded types of activity as well as separate actions interconnected by streams that come from one node output to other. Algorithm schemes are activity diagram analogue.

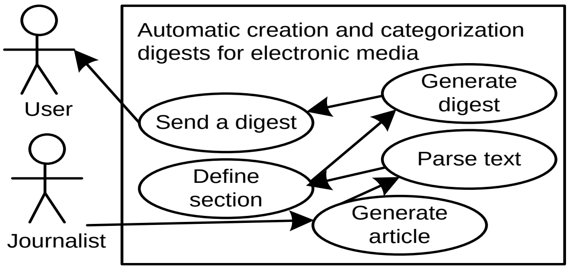


Fig. 3. System use-case diagram.

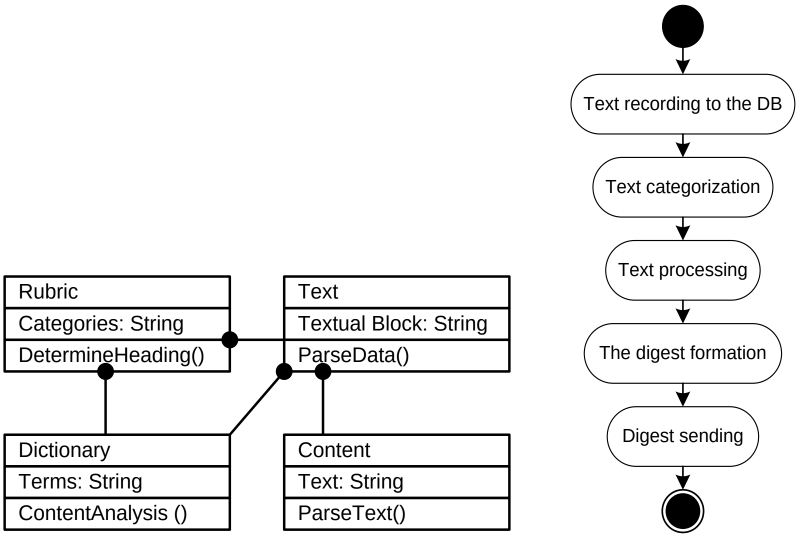


Fig. 4. Class diagram. State diagram

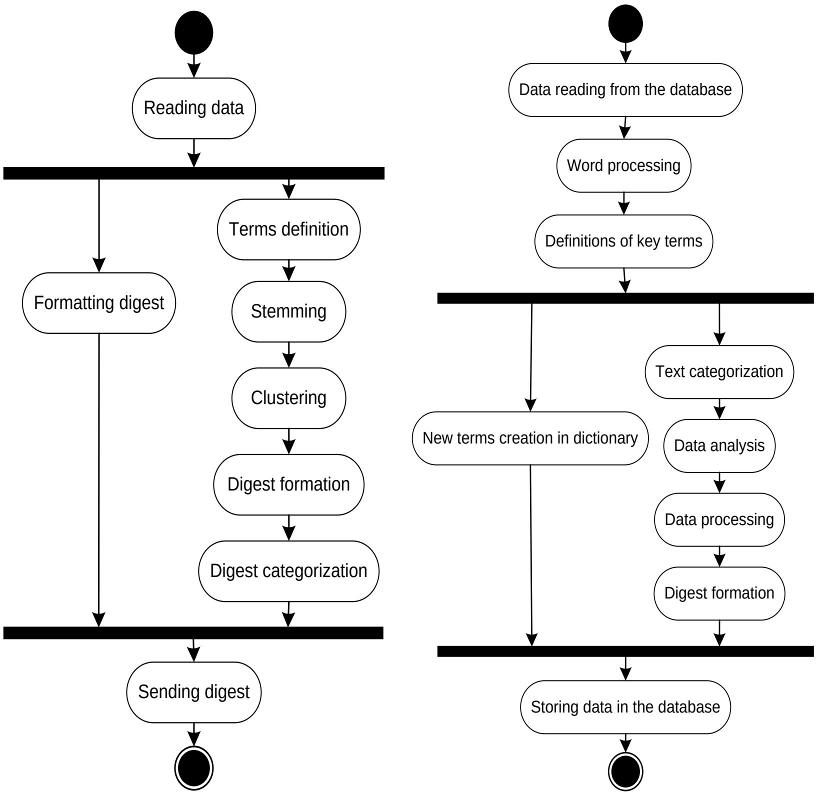


Fig. 5. Activity diagram of system operation algorithm

Message exchange (i.e. method call) between several objects in a specific time-delimited situation is presented on sequence diagram. Objects are instances of classes. The main emphasis in sequence diagrams is made on order and moments of time during which messages are sent to objects. In sequence diagrams the objects are represented using vertical dashed lines with the name of the object over them. The time axis has also vertical direction – it is facing down. Messages sent from one object to another will be marked by arrows with operation and parameters name (Fig. 6). Collaboration diagram (Fig. 7) is intended for the specification of interaction system structural aspects. Cooperation is convenient for design pattern modeling.

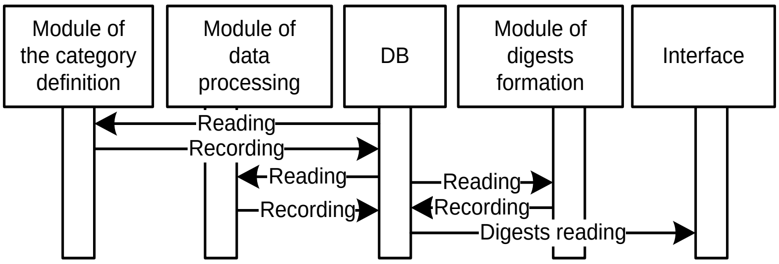

Module of the category definition
Module of data processing
DB
Module of digests formation
Interface
Reading
Recording
Reading
Recording
Reading
Recording
Digests reading

Fig. 6. Sequence diagram

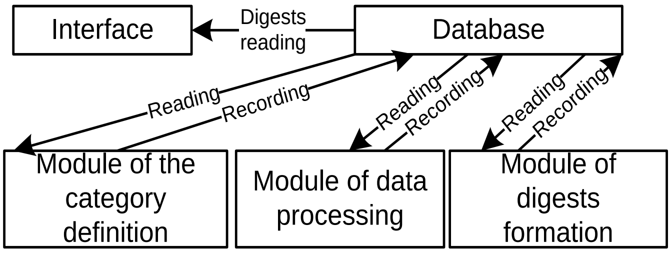

Interface
Digests reading
Database
Reading
Recording
Reading
Recording
Reading
Recording
Module of the category definition
Module of data processing
Module of digests formation

Fig. 7. Collaboration diagram

Web-monitoring module should allow circumventing user-specified pages and assigning download of web-pages updates. The data received should be read out by special modules which are oriented to receive information of such a type. All materials should be processed by categorization module after receiving and pre-processing. Initially construction and training of rubricator by expert is needed. The essence of education lies in teaching materials expert analysis including them to one or another rubric. Expert should specify relationship degree of text to one or another topic. Based on expert assessments categorization module should produce morphological and semantic analysis of text, specifying main thematic concepts as well as analyzing the structure of their location in the text. Data systematization in automatic mode is conducted on the basis of learning outcomes. Incoming text refers to one or more categories with affixing relationship degree. Material should pass indexation according to all words it contains after distribution by category process. This procedure provides flexible search capabilities both on the basis of materials as well as their content. System operation results are presented in the form of thematic digests which are automatically created. The simplest strategy of quasi-abstracts formation lays in profile change fixing moments, i.e. transitions from lower levels to higher and vice versa. It means that we react on beginning and the end of each cluster selecting the first and the last sentence for quasi-abstract. This strategy is typical for positional referencing methods (Berko, Vysotska, & Pasichnyk, 2009), (Pasichnyk, Scherbyna, Vysotska, & Shestakevych, 2012), which essentially boils down to keeping the initial and the final fragments in the text structure specified by author. Thus it is considered that the most important parts of scientific article text are introduction and conclusion, in every unit – initial and final paragraph, in every paragraph – the initial and final sentence. The remarkable thing is that not all research papers (let alone texts of other genres) are divided into sections and subsections, which is an additional argument in favor of developmental approach. The second strategy of quasi-abstracts formation is to determine weight of each sentence of text. The sentence weight is determined by the number of word form occurrences demonstrating clustering in random places in text. So, a sentence may not be a part of corresponding cluster. A version when sentence weight fixes diversity of clustered word forms presented in it (not their full amount) is more attractive (Berko, Vysotska, & Pasichnyk, 2009), (Pasichnyk, Scherbyna, Vysotska, & Shestakevych, 2012). Advantages of quasi-abstracting method lie in simplicity of implementation. However text block selection ignores relationships between them, and often leads to meaningless abstracts formation. Some sentences may appear missed or there may occur words or phrases that can not be understood without previous text, which is missed in the abstract. The attempts to solve this problem are mainly confined to exclusion such sentences from abstracts. The attempts to resolve references using linguistic analysis methods are made less frequently. The special approaches by which we can determine the presence of semantic gap are created in a number of human-machine interfaces. It is obvious that such an approach is not suitable for mass random word processing. As with the quasi-abstracting of separate text document, during the first phase of digest formation it takes place that the most important lexical items selection included in source documents array (incoming information flow) based on which system glossary is built up. Selection of the source documents from digest construction input array is carried out by taking into account their weights. The weight of each document is determined by taking into account separate words sum of weights normalized by the length of document included in this document. The phase of document selection intended to digest consists of following steps like definition of each document weight, sorting out incoming stream of documents by weight, definition of semantic duplicates of documents by statistical criteria, rejection documents unsuitable for digests formation (unacceptable types of documents, such as reviews), and semantic duplicates (which are detected by frequency alg. 7).

Algorithm 7. A digest formation

1st stage. Formation of digest which reflects main trend.

2nd stage. Removing documents that match trends as defined in previous step from incoming information flow.

3rd stage. Digest formation that reflects the main trend of the rest of the basic information flow.

4th stage. Obtained digests combining.

5th stage. Transition to the 2nd stage is carried out if necessary (based on resulting digest required volumes).

Let us consider the algorithm of electronic digests automatic formation and rubrication. The system receives data from the database as an array of articles. After that it verifies whether user-defined request is specified for text article processing. In case there is an article needed further processing, text articles lemmatisation is carried out sequentially in order to determine lexemes, stemming for rejection stop words and clusterization. As a result of such elaborations, will be collected the information about position and weight of word forms enabling to include the article to specified rubric and form digest.

6 CONCLUSIONS

Research of the mathematical methods application for textual information the analysis and synthesis in natural language for the development of mathematical algorithms and software for textual content processing is required.

Theory of generative grammar (proposed by Noam Chomsky) is modeling processes at the syntactic level of language. The structural elements in sentences describe syntactic constructions in textual content regardless of their content. The article shows the features of the sentences synthesis indifferent languages of using generative grammars. The paper considers norms and rules influence in the language on the grammars constructing course. The use of generative grammars has great potential in the development and creation of automated systems for textual content processing, for linguistic providing linguistic computer systems, etc.

In natural languages there are situations where the phenomenon that depend on context and described as independent of context (ie, in terms of context-free grammars). In this case description is complicated due to the formation of new categories and rules. The article describes features in the process of introducing new restrictions on data classes through the new grammar rules introduction. If the symbols number on the right side in the rules are not lower than the left then got not reduced grammar. Then at replacement only one symbol got a context-sensitive grammar. In the presence only one symbol in the left side of the rule got a context-free grammar. None these natural constraints on the left side rules apply is not possible. The theory application of generative grammars for solving problems of applied and computational linguistics at the morphology and syntax level allows to create a system of speech and texts synthesis, to create practical morphology textbooks and inflection tables, to conclude the morphemes lists (affixes, roots), to determine the performance and frequency for morphemes and the frequency of different grammatical categories realization in texts (genus, case, number, etc.) for specific languages. Developed models on the basis of generative grammars for linguistic functioning computer systems designed for analytical and synthetic processing of textual content in information retrieval systems, etc. are used. Is useful to introduce all new and new restrictions on this grammar, getting more narrow their classes. In describing the complex range of phenomena limit used means set of description, and the considering these features, which are served in general obviously insufficient. Research begin with minimum means. Whenever their are not enough (smaller portions) new means gradually are introduced. Thus possible to determine exactly what means can or can not use in the description of a phenomenon for understanding its nature.

While working in informational and analytical services and enterprises you have to deal with great diversity of information sources. These are electronic newspapers and other Internet resources. This article considers Ukrainian electronic mass media as well as their disadvantages, benefits, and services. The carried out researches of electronic mass media allow concluding of inexpediency use of human labor in the processes related to formation and ranking digests. A key part of this paper is also developing methods of formation and rubrication of electronic digests. The experience of implementing systems in different organizations has showed effectiveness and simplicity of adapting the system due to the developed instrument of automated digest formation and their rubrication. The universal data acquisition module allows you to automate completely the electronic information introduction from sources with bringing information to a common internal format, i.e. to minimize the routine work while entering text data.

WORKS CITED

Berko, A., Vysotska, V., & Pasichnyk, V. (2009). Systemy elektronnoyi kontent-komertsiyi. Lviv, Ukraine: NULP Publ.

Boiko, B. (2005). Content Management Bible. Indianapolis: Wiley Publishing.

Bolshakova, E., Lande, D., Noskov, A., Klyshynskyy, E., Peskova, O., & Yahunova, E. (2011). Avtomatycheskaya obrabotka tekstov na estestvennom yazyke y kompyuternaya lynhvystyka. Moscow, Russia: MYEM Publ.

Braychevskyy, S., & Lande, D. (2005). Sovremennye informatsionnye potoki. Nauchnotehnicheskaya informatsiya(11), 21-33.

Chomsky, N. (1956).Three models for the description of language (pp. 113-124). I.R.E. Trans. PGIT 2.

Chomsky, N. (1959). On certain formal properties of grammars (pp. 137-267, 393-395). Inf. and Cont.

Chomsky, N. (1961). On the notion «Rule of Grammar», Proc. Symp. Applied Math., 12. Amer.

Chomsky, N. (1962). Context-free grammars and pushdown storage, №65, Lab. of Electronics, M.I.T.

Chomsky, N. (1963). Formal properties of grammars (pp. 323-418). Handbook 2, ch. 12, Wiley.

Chomsky, N. (1962). The logical basis for linguistic theory, Proc. IX-th Int. Cong. Linguists

Chomsky, N., & Miller, G. (1958). Finite state languages (pp. 91-112). Information and Control 1.

Chomsky, N., & Miller, G. (1963). Introduction to the formal analysis of natural languages (pp. 269-322). Handbook of Mathematical Psychology 2.

Chomsky, N., & Schuetzenberger, M. (1963). The algebraic theory of context-free languages (pp. 118–161). North-Holland, Amsterdam.

Clifton, B. (2009). Google Analytics: Professional'nyj analiz poseŝaemosti veb-sajtov = Advanced Web Metrics with Google Analytics. M., Russia: Vil'jame Publ.

Doyle, B. (2005). Seven Stages of the CM Lifecycle. EcontentMag.com. Retrieved from EContent: http://www.econtentmag.com/Articles/ArticleReader.aspx?ArticleID=13554&AuthorID=155

Furashev, B., Lande, D., Grigor'yev , O., & Furashev, O. (2005) Elektronne informatsiyne suspilstvo Ukraini. Akademiya pravovih nauk Ukraine., Kiev, Ukraine: Inzhiniring Publ.

Furashev, B., Lande, D., & Brajchevskiy, S. (2005) Sistemnaya informatizatsiya izbiratelnyh i referendumnyh protsessov (pp. 11-15). OIKIT № 29, Kharkiv, Ukraine: NAKU HAI Publ.

Gladky, A. (1985). Sintaksicheskie struktury estestvennogo yazyka v avtomatizirovannyh sistemah obscheniya, Moscow, Russia:: Nauka Publ.

Gladky, A., & Melchuk, I. (1969). Elementy matematicheskoy lingvistiki, Moscow, Russia:: Nauka Publ.

Gladky, A. (1973). Formalnye grammatiki i yaziki, Moscow, Russia:: Nauka Publ.

Grigoriev, A., & Lande, D. (2005) Adaptivnyy interfeys utochneniya zaprosov k sisteme kontent-monitoringa InfoStream (pp.109-111). K:Dialog'2005 Publ.

Grigoriev, A., & Lande, D. (2005) Sistema monitoringa InfoStream – informatsionnoe prostranstvo iz odnih ruk. Postroenie informatsionnogo obschestva: resursy i tehnologii (pp. 17-20). Kyiv, Ukraine: UkrISTEI Publ.

Hackos, J. T. (2002). Content Management for Dynamic Web Delivery. Hoboken, NJ, USA: Wiley.

Halverson, K. (2009). Content Strategy for the Web. Reading, Mass: New Riders Press.

Hartman, E. (2006). Content management lifecycle poster. Retrieved from CM Professionals: http://www.google.rs/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=0CDsQFjAC&url=http%3A%2F%2Fwww.jrtcllc.com%2Ffiles%2Freference-files%2Fcm-poster%3Fformat%3Draw&ei=Iru2UtLrH8LeygPPw4D4DQ&usg=AF Q jCNHN0mXF1wved45chqXh7AhomQ0hpA&sig2=IuCGFtOmMepRZK0Ah3zUYw&

Lande, D., & Brajchevskiy, S. (2005) Opredelenie tematicheskoy napravlennosti zaprosov putem analiza nabora reytingovyh istochnikov (pp. 169-174). OIKIT № 29, Kharkiv, Ukraine: NAKU HAI Publ.

Lande, D., & Brajchevskiy, S. (2006) Prognozno-analiticheskie issledovaniya na osnove sistemy kontent-monitoringa InfoStream. Informatsiya, analiz, prognoz – strategicheskie rychagi effektivnogo gosudarstvennogo upravleniya (pp. 147-152) ., № 5. Kyiv, Ukraine: UkrISTEI Publ.

Lande, D. (2005) Nekotorye metody analiza novostnyh informatsionnyh potokov (pp. 277-287). Donetsk, Ukraine: IKVT , Donetsk National Technical University Publ., № 93.

Lande, D. (2005) Poisk znany v Internet. Moscow, Russia: Williams Publ.

Lande, D. (2005) Semanticheskiy veb. Telecom, № 6, pp. 60-65.

Lande, D. (2005) Skaner sistemy kontent-monitoringa InfoStream (pp. 53-58). OIKIT № 28, Kharkiv, Ukraine: NAKU HAI Publ.

Lande, D. (2006). Osnovy intehratsii informatsyonnyh potokov. Kyiv, Ukraine: Engineering Publ.

Lande, D., & Litvin, A. (2001) Fenomeny sovremennyh informatsionnyh potokov (pp. 14-21). Seti i biznes, № 1.

Lande, D., & Morozov, A. (2004) Chitayte novosti, batenka. CHIP, № 7, pp. 82-85.

Lande, D., & Morozov, A. (2005) Novostnoy Internet. Telecom , № 1-2, pp. 58-62

Lande, D., & Furashev, V. (2006) Voprosy postroeniya i ispolzovaniya mnogokriterialnoy modeli vybora istochnikov informatsii (pp. 76-85). OIKIT № 30, Kharkiv, Ukraine: NAKU HAI Publ.

Lande, D., Furashev, V., Braychevskyy, S., & Hryhorev, O. (2006). Osnovy modelirovaniya i otsenki elektronnyh informatsionnyh potokov. Kyiv, Ukraine: Engineering Publ.

Lande, D., Furashev , V., & Grigor'yev , O. (2006) Programno-aparatny kompleks informatsiynoyi pidtrymky priynyattya rishen. Kyiv, Ukraine: Inzhiniring Publ.

McGovern, G., & Norton, R. (2001) Content Critical. Upper Saddle River, NJ: FT Press Publ.

McKeever, S. (2003). Understanding Web content management systems: evolution, lifecycle and market. Industrial Management & Data Systems, 103(9), 686 - 692.

Nakano, R. (2002). Web content management: a collaborative approach. Boston: Addison Wesley Professional.

Papka, R. (1999) On-line News Event Detection, Clustering, and Tracking. Ph. D. Thesis, University of Massachusetts at Amherst, September

Pasichnyk, V., Scherbyna, Y., Vysotska, V., & Shestakevych, T. (2012). Matematychna linhvistyka. Lviv, Ukraine: Novyy Svit - 2000 Publ.

Rockley, A., & Cooper, C. (2002). Managing Enterprise Content: A Unified Content Strategy (Second ed.). Berkeley: Reading Mass New Riders Press. Retrieved from http://www.managingenterprisecontent.com/

Salton, D. (1979) Dinamicheskie bibliotechno-informatsionnye sistemy. Moscow, Russia: Mir Publ.

Stone, W.R. (2003). Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! IEEE Antennas and Propagation, 45(4). Retrieved from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1241310&url= http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D1241310

Zipf, G. (1949) Human Behavior and the Principle of Least Effort. Addison-Wesley.

Zipf, G. (1935) The Psychobiology of Language. Houghton-Mifflin.

Received for publication:	29.10.2014
Revision received:	19.11.2014
Accepted for publication:	09.12.2014

How to cite this article?

Style – **APA** *Sixth Edition:*

Vysotska, V., & Chyrun, L. (2015, Jan 15). Linguistic analysis and modelling semantics of textual content for digest formation. (Z. Čekerevac, Ed.) *MEST Journal, 3*(1), 127-148. doi:10.12709/mest.03.03.01.15

Style – **Chicago** *Fifteenth Edition:*

Vysotska, Victoria, and Lyubomyr Chyrun. 2015. "Linguistic analysis and modelling semantics of textual content for digest formation." Edited by Zoran Čekerevac. *MEST Journal* (MESTE) 3 (1): 127-148. doi:10.12709/mest.03.03.01.15.

Style – **GOST** *Name Sort:*

Vysotska Victoria and Chyrun Lyubomyr Linguistic analysis and modelling semantics of textual content for digest formation [Journal] // MEST Journal / ed. Čekerevac Zoran. - Belgrade : MESTE, Jan 15, 2015. - 1 : Vol. 3. - pp. 127-148.

Style – **Harvard** *Anglia:*

V ysotska, V. & Chyrun, L., 2015. Linguistic analysis and modelling semantics of textual content for digest formation. *MEST Journal,* 15 Jan, 3(1), pp. 127-148.

Style – **ISO 690** *Numerical Reference:*

Linguistic analysis and modelling semantics of textual content for digest formation. **Vysotska, Victoria and Chyrun, Lyubomyr.** [ed.] Zoran Čekerevac. 1, Belgrade : MESTE, Jan 15, 2015, MEST Journal, Vol. 3, pp. 127-148.