UDC 81:004.93

V. Vysotska
Lviv Polytechnic National University
Information Systems and Networks Department

GENERATIVE REGULAR GRAMMARS APPLICATION TO MODELING THE SEMANTICS OF SENTENCES IN NATURAL LANGUAGE

ЗАСТОСУВАННЯ ПОРОДЖУВАЛЬНИХ РЕГУЛЯРНИХ ГРАМАТИК ДЛЯ МОДЕЛЮВАННЯ СЕМАНТИКИ РЕЧЕННЯ ПРИРОДНОЮ МОВОЮ

This paper presents the generative grammar application in linguistic modelling. Description of syntax sentence modelling is applied to automate the processes of analysis and synthesis of texts in natural language.

Key words: generative grammar, structured scheme sentences, computer linguistic system.

Подано застосування породжувальних граматик у лінгвістичному моделюванні, застосовано моделювання синтаксису речення для автоматизації процесів аналізу та синтезу природномовних текстів.

Ключові слова: породжувальні граматики, структурна схема речення, комп'ютерна лінгвістична система.

Introduction

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. [1-2, 6-8]. Linguistic analysis of the text content consists of several processes of grapheme, morphological, syntactic and semantic analysis [3-5, 9-17]. For each of these stages appropriate models and algorithms were developed [3-5, 9-17]. 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 [9-17]. 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 [3-5], 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 [3-5].

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.

Relation of the presented issue with the important scientific and practical tasks

The methods and tools development for automatic processing of text of commercial content in modern information technology are important and topical [1-2, 6-8] (for example, systems of information

retrieval, machine translation, semantic, statistical, optical and acoustic analysis and synthesis of speech, automated editing, knowledge extracting from the text content, text content abstracting and annotation, textual content indexing, training and didactic, linguistic buildings management, instrumental means of dictionaries conclusion of various types, etc.). Specialists actively seeking new models of description and methods for automatic processing of text content [1-2, 6-8]. One of these methods is the development of general principles of lexicographic systems of syntactic type. It is important by these principles these systems construction of text content processing for specific languages [1-2, 6-8].

Research linguists in the sphere of morphology, morphonology, structural linguistics have identified different patterns for the word forms description [1-17]. 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 combinatorics, 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 [1–17]. One of the first attempts to apply generated grammars theory for linguistic modeling belongs to . Gladky and I. Melchuk [3–5]. Experience and research of Noam Chomsky [9–17], A. Gladky [3-5], M. Hross, A. Lanten, A. Anisimov, Y. Apresyan, N. Bilhayeva, I. Volkova, T. Rudenko, E. Bolshakova, E. Klyshynsky, D. Lande, A. Noskov, A. Peskova, E. Yahunova, A. Herasymov, B. Martynenko, A. Pentus, M. Pentus, E. Popov, V. Fomichev [2, 8] 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. [1–17].

Recent research and publications analysis

Linguistic analysis of the content consists of three stages: morphological, syntactic and semantic [2–5]. 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 [2–5]. There are the exact and approximate methods of morphological analysis [2]. 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 [2–5]. 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 [2]. 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 [2–5]. 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 [2–5].

As the basis of approximate methods in morphological analysis determines the grammatical class of words by the end letters and letter combinations [2]. 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 [2–5]. Upon receipt of the coincidence of final part with words is defined as its basis [2]. In conducting morphological analysis arise ambiguity of grammatical information determination, that disappear after parsing [2]. The task of syntactic analysis is parsing sentences based on the data from the dictionary [2–5]. At this stage allocate noun, verb, adjective, etc., between which indicate links in the form of dependency tree. Any tools of syntactic analysis consists of two parts: a knowledge base about a particular natural language and algorithm of syntactic analysis (a set of standard operators of text content processing on this knowledge) [2–5]. The source of grammatical knowledge is data from morphological analysis and various filled tables of concepts and linguistic units [2]. They are the result of the empirical processing of textual content in natural language of experts in order to highlight the basic laws for syntactic analysis. Table-based of linguistic units constitute configurations or valences sets (syntactic and semantic-syntactic dependencies) [2]. This is a lexical units list/dictionaries as instructions for every of them all possible links with other units of expression in natural language [2–5]. In implementing of the syntactic analysis should be achieved full independence of rules of tables data transform from their contents. This change of this content does not require algorithm restructuring.

The vocabulary *V* consists of finite not empty set of lexical units [2]. The expression on *V* is a finite-length string of lexical units with *V*. An empty string does not contain lexical items and is denoted by Λ. The set of all lexical units over *V* is denoted as V'. The language over *V* is a subset V'. The language displayed through the set of all lexical units of language or through definition criteria, which should satisfy lexical items that belong to the language [2]. Another is one important method to set the language through the use of generative grammar. The grammar consists of a lexical units set of various types and the rules or productions set of expression constructing. Grammar has a vocabulary *V*, which is the set of lexical units for language expressions building. Some of lexical units of vocabulary (terminal) can not be replaced by other lexical units.

Generative grammar *G* – is four *G* = (*V*, *T*, *S*, *P*), where *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*∈ *V*); *P* – is a finite set of productions (conversion rules) $x \to h$, where *x* and *h* – strings over *V*. The set $V \setminus T$ is denoted by *N*, its elements are non-terminal lexical units, non-terminals [1–17]. Grammars are classified by the types of productions, which imposed certain restrictions.

1. Grammar G_0 is unlimited. Here *x* – random string that contains at least one non-terminal symbol, *h* – random string over V.

2. Context-sensitive grammar G_1. In the set of productions *P* exists production $gxd \to ghd$, $|x| \le |h|$ (but not in the form $x \to h$), then you can substitute *x* with ν only surrounded by strings *g¼d*, i.e. in appropriate context.

3. Context-free grammar G_2. A non-terminal *A* on the left side of $A \to h$ 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 $A \to aB$, $A \to a$, $S \to l$, only, where, *B* – non-terminal, a – terminal, *l* – 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 [9-17]. 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 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 [3–5]. 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 $gAd \to ghd$, wherein at least one of the strings g, different from Λ, and nonterminal A is substituted by string h surrounded only by g and d, 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 [9–17]. 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 [1-17] 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.

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. 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.

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.

In the module are stored no texts content, but its RIm. To search the indexed content is used content analysis in information requests. 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.

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.

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). 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.

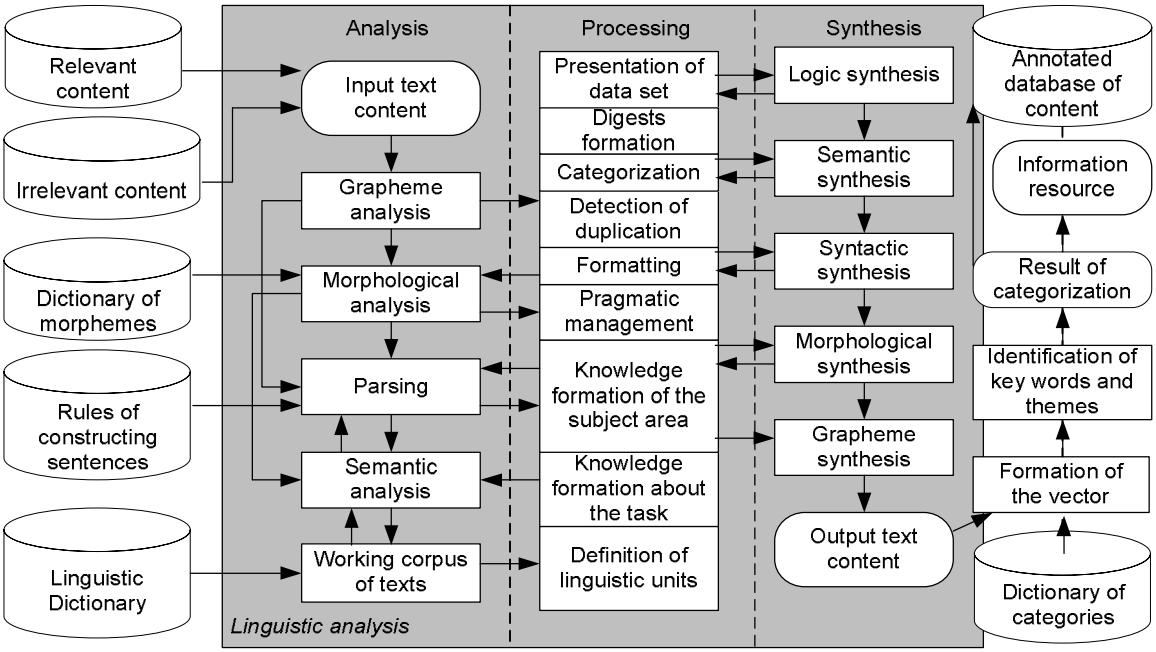


Fig. 1. Block diagram of the linguistic analysis of text content in the formation content

Algorithm 1. Language L(G) recognition

Phase 1: Alternately apply S to grammar productions.

Phase 2: Built string *x* verifying.

Step 1: If yes, then go to step 3.

Step 2: If not, then go to step 1.

Phase 3: 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.

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

Step 1: 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 $\leq n$). Number of different strings length $\leq n$ of *p* lexical units

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

$C_P^k (P-k)! = P! \big/ k!(P-k)! (P-k)! = P! \big/ k! \leq P!$, where $C_P^k = P! \big/ k!(P-k)!$. The total sum is no more than

$P! \cdot P < (P+1)! = (p^{n+1}+1)! < (p^{n+2})!$.

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

Phase 2: Construction of derivation set *M* of random string *x*.

Phase 3: Check the resulting finite set of derivations *M*.

Step 1: 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 step 3.

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

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

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

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

Phase 6: 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 $Z_1 C Z_2$ (*C* – a lexical unit), and the right part (*Y*) – form $Z_1 W Z_2$ (*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, $P_1 = \{AB \to BA\} \approx P_2\{AB \to 1B, 1B \to 12, 12 \to B2, B2 \to BA\}$ 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 $X = Z_1 C Z_2$. Then the production $Z_1 C Z_2 \to Z_1 W Z_2$ means a permit to replace C into W only in terms of context $Z_1 ... Z_2$ and without changing its location.

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.

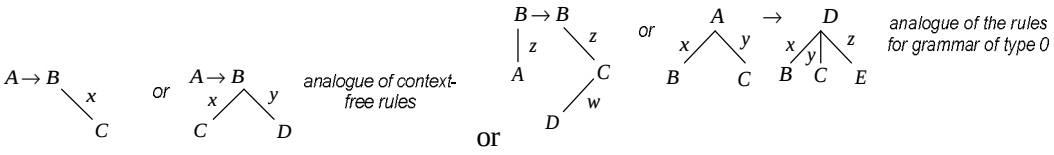
If natural language is recognized algorithm with with the specified restriction on the amount of memory, then it can be the generative grammar. Here, for displayed any terminal string of n length is a conclusion which none intermediate chain does not exceed the length of Kn (K is some constant). This 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 [3–5]. Then the rules are as follows:



Lines with symbols represent the syntactic relations of different types. Letters $A,B,C,...$ 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 $C \xleftarrow{y} A \xrightarrow{x} B$ [3-5, 9-17]. As a result the syntactic structures (not phrases) calculation in language is obtained. This calculation is part of the generative grammar [3-5]. 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 [3-5]. 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.

Stage 1. The grammatical analysis of textual content.

Step 1. Separation of commercial textual content on sentences and paragraphs.

Step 2. String separation of symbols into words.

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

Step 4. Removing the non-text symbols.

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

Stage 2. Morphological analysis of the textual content.

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

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

Step 3. Linear sequence formation of morphological structures.

Stage 3. Parsing textual content.

Stage 4. Semantic analysis of textual content.

Step 1. Words are correlated with semantic classes from dictionary.

Step 2. Selection of needed morphologically semantic alternatives for the sentence.

Step 3. Linking words into a single structure.

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

Stage 5. Referential analysis to form between phrasal unities.

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

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

Step 3. 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.

Stage 6. 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.

Step 1. Basic set identification for rhetorical relations between unities content.

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

The text realizes structural submitted activities through provides subject, object, process, purpose, means and results that appear in content, structural, functional and communicative criteria and parameters. The units of internal organization of the text structure are alphabet, vocabulary (paradigmatics), grammar (syntagmatic) paradigm, paradigmatic relations, syntagmatic relation, identification rules, expressions, unity between phrasal, fragments and blocks. On the compositional level are isolated sentences, paragraphs, sections, chapters, under the chapter, page etc. that (except the sentence) indirectly related to the internal structure because are not considered.

Content analysis for compliance thematic requests to $C_{Ct} = Categorization\left(KeyWords(C,U_K),U_{Ct}\right)$, where $KeyWords(C,U_K)$ is the keywords identify operator, $Categorization$ is content categorize operator according to the keywords identified, U_K is keywords identify conditions set, U_{Ct} is categorization conditions set, C_{Ct} is rubrics relevant content set. Digest set C_D formed by such dependence as $C_D = BuDigest\left(C_{Ct},U_D\right)$, where $BuDigest$ is digests forming operator, U_D is conditions set for the digests formation, C_{Ct} is rubrics relevant content set. With the help of a database (database for terms/morphemes and structural parts of speech) and defined rules of text analysis searching terms. Parsers operate in two stages: lexemes content identifying and a parsing tree creates (alg. 4).

Algorithm 4. Parser of textual commercial content.

Stage 1. Content C lexemes identification from set *V*.

Step 1. Terms string definition over *V* as a sentence.

Step 2. Nouns groups identification with bases dictionary V'.

Step 3. Verbs groups identification with bases dictionary V'.

Stage 2. Creating of a parsing tree from left to right. Each step of output is the deployment as one symbols of the previous string or it to others replacing, while other symbols are rewritten without change. It is obtained the component tree, or syntactic structure, if process of deployment, replacement or re-writing characters (*fathers*) connect lines directly with symbols that come out as a result of the deployment, replacement or re-writing (*descendants*).

Step 1. Nouns group deployment. Verbs group deployment.

Step 2. The implementation of syntactic categories of word forms.

Stage 3. The keywords set determination.

Step 1. The *Noun* terms determination (nouns, nouns word combinations or adjective with the noun) among the words set of commercial textual content.

Step 2. The *Unicity* uniqueness calculation for *Noun* terms.

Step 3. The *NumbSymb* value calculation (the characters number with no spaces) for *Noun* terms at $Unicity \geq 80$.

Step 4. The *UseFrequency* value calculation (frequency of keywords using). The *UseFrequency* frequency for terms with $NumbSymb \leq 2000$ is within the limits $(6;8]$ %. Frequency for terms with $NumbSymb \geq 3000$ is within the limits $[2;4)$ %. Frequency for terms with $2000 > NumbSymb < 3000$ is within the limits $[4;6]$ %.

Step 5. The values calculation of *BUseFrequency* (the frequency of keywords using at the beginning in the text), *IUseFrequency* (the frequency of keywords using in the middle of the text), *EUseFrequency* (the frequency of keywords using at the end in the text).

Step 6. Comparison of values *BUseFrequency*, *IUseFrequency* and *EUseFrequency* for priorities definition. Keywords with higher values *BUseFrequency* have higher priority than keywords with a higher value *IUseFrequency*.

Step 7. Keywords sorting according to their priorities.

Stage 4. The database filling of search image for content, i.e. attributes of *KeyWords* (keywords), *Unicity* (the keywords uniqueness ≥ 80), *Noun* (term), *NumbSymb* (the number of characters without spaces), *UseFrequency* (frequency of keywords using), *BUseFrequency* (frequency of keywords using at the beginning in the text), *IUseFrequency* (the frequency of keywords using in the middle of text), *EUseFrequency* (frequency of keywords using at the end in the text).

Based on the rules of generative grammar perform term correction under the rules of its use in context. The sentence define action limits of punctuation marks and links. The text semantics is due communicative task of information transfer. The textual structure is determined by internal organization of textual units and their relationship regularities. While parsing the text drawn in a data structure (eg, tree which corresponds the syntactic structure of the input sequence, and is best suited for further processing). After analysis textual block and term is synthesized a new term as a keyword of content topics by using base of terms and their morphemes. Next is synthesized terms for a new keyword formation by using base of structural parts of speech. The principle of keywords detection in content (terms) is based on Zipf's law. It is reduced to words choice with an average frequency of occurrence. The most used words are ignored through the stop-dictionaries. And the rare words do not include text. According a meaningful analysis of the content corresponds to the process grammatical data extraction from the word by grapheme analysis and the results correction of morphological analysis through the grammatical context analysis of linguistic units (Alg. 5).

Algorithm 5. The textual commercial content categorization

Stage 1. The division of commercial content on the blocks.

Step 1. Block presentation to the input of tree construction with commercial content blocks.

Step 2. New block creation in the blocks table.

Step 3. The newline characters accumulation.

Step 4. Checking on point availability before a newline character. If there is, then go to step 5. If do not, then begin the sequence saving in the table, the new block parsing and transition to step 3.

Step 5. Checking on availability of the end in the text. If the end of the text is, then go to step 6. If do not, then start the accumulated sequence saved in the table, the new block parsing and transition to step 2.

Step 6. Blocks tree getting on the output as a table.

Stage 2. The block division on sentence with structure preservation.

Step 1. The input is a table of blocks. The sentences table creation with link for field *ID_section* of *n*-to-1 type of blocks table.

Step 2. A new sentence creation in sentences table.

Step 3. The symbols accumulation to point, semicolon or newline character.

Step 4. Checking on availability of cuts. If the cut exists, then go to step 5. If do not, then start the sequence saving in the table, a new sentence parsing and transition to step 2.

Step 5. Checking on availability of the end in the text block. If the end of the text exists, then go to step 6. If do not, then begin a sequence saving in the table, new sentences parsing and transition to step 2.

Step 6. The sentences tree getting on the output as a table.

Step 7. Checking for the end of the text. If the end of the text exists, then go to step 8. If do not, then start the new block parsing and transition to step 1.

Step 8. The sentences tree getting on the output in the form of tables.

Stage 3. The sentences division for lexemes with indication of belonging to sentences.

Step 1. The lexemes table formation based on the sentences table with fields of *ID_lexemes* (unique identifier), *ID_sentence* (number equal to the code of the sentence with lexeme), *Lexemes_number* (number equal to the lexemes number in the sentence), *Text* (lexeme text).

Step 2. A sentence presentation to the input from the sentences table for parsing on lexemes.

Step 3. A new lexeme creation in the lexemes table.

Step 4. The symbols accumulation to point, spaces or end of a sentence and the seving in the lexemes table.

Step 5. Checking for the end of the sentence. If yes, then go to step 6. If not then accumulated sequence saving in the table, new lexeme parsing and transition to step 3.

Step 6. Conducting parsing based on data obtained on the output (Alg. 4).

Step 7. Conducting morphological analysis based on data obtained at the output.

Stage 4. The topics determination for the commercial content.

Step 1. The hierarchical structure construction of properties for each lexical unit with text that includes grammatical and semantic information.

Step 2. The lexicon formation with a hierarchical organization of properties types, where each type-descendant inherits and overrides the ancestor properties.

Step 3. Unification as a basic mechanism for constructing syntactic structures.

Step 4. The *KeyWords* set identification for the commercial content (Alh.4).

Step 5. The values set definition as *TKeyWords* (thematic keywords in the *KeyWords* set for commercial content), *Topic* (the theme for commercial content) and *Category* (commercial content category).

Step 6. The values set definition as *FKeyWords* (the frequency of keywords using in the textual commercial content) and *QuantitativeryTKey* (frequency of thematic keywords using in the textual commercial content).

Step 7. The values set definition as *Comparison* (the keyword using comparison with various topics). The values set calculation as *CofKeyWords* (coefficient of thematic content keywords), *Static* (coefficient of the statistical terms importance), *Addterm* (coefficient of the additional terms availability). Comparison of the content keywords set with the key concepts with topics. If there is a match, then go to step 9. If not, then go to step 8.

Step 8. A new category formation with a set of key concepts of the analyzed content.

Step 9. Assignment designated section of the analyzed commercial content.

Step 10. The values set calculation as *Location* (the coefficient of content location in the thematic section).

Stage 4. Filling the search images base for attributes as *Topic* (the theme of content), *Category* (content category), *Location* (the coefficient of content location in the thematic section), *CofKeyWords* (the coefficient of thematic keyword of textual content), *Static* (coefficient of statistical significance for terms),

Addterm (the coefficient of the additional terms availability), *TKeyWords* (the thematic keywords), *FKeyWords* (the frequency of using keywords), *Comparison* (the keywords using comparison of the different themes), *QuantitativeryTKey* (frequency of thematic keywords using in the text of commercial content).

The process of categorization through automatic indexing component in commercial content is divided into sequential blocks: a morphological analysis, a syntactic analysis, a semantic and a syntactic analysis of the linguistic structures and meaningful writing variation in the textual content.

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.

Conclusion

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.

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.

In the thesis known methods and approaches to solving the problem of automatic processing of textual content and selected advantages and disadvantages of existing approaches and results in the field of the syntactic aspects of computational linguistics is considered. In this paper the general conceptual framework of modeling inflectional processes of the text arrays creation is formed. The syntactic model and inflectional classification of lexical structure of sentences is proposed. Also in the theses lexicographical rules of syntactic type for automated processing of these sentences is developed. The proposed technique allows to achieve the highest parameters of reliability in comparison with known analogues. They also demonstrate the high efficiency of applied applications in the linguistics construction of new information technologies and research inflectional effects in natural language. The work has practical value, since the proposed models and rules can effectively organize the process of creating a lexicographical systems of textual content processing of syntactic type. The commercial content formation model implement in the form of content-monitoring complexes to content collection from data various sources and provide a content database according to the users information needs. As a result, content harvesting and primary processing its lead to a single format, classified according to the categories and he is credited tags with keywords.

1. Berko A. Electronic content commerce systems / A. Berko, V. Vysotska, V. Pasichnyk. – L.: NULP, 2009. – 612 p. 2. Большакова Е. И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О. В. Пескова, Е. В. Ягунова. – М.: МИЭМ, 2011. – 272 с. 3. Gladky A. Sintaksicheskie struktury estestvennogo yazyka v avtomatizirovannyh sistemah obscheniya / A. Gladky. – M.: Nauka Publ., 1985. – 144 p. 4. Gladky A. Elementy matematicheskoy lingvistiki / A. Gladky, I. Melchuk. – M.: Nauka Publ., 1969. – 192 p. 5. Gladky A. Formalnye grammatiki i yaziki / A. Gladky. – M.: Nauka Publ., 1973. – 368 p. 6. Lande D. Osnovy modelirovaniya i otsenki elektronnyh potokov / D. Lande, V. Furashev, S. Braychevsky, O. Grigorev. – K.: Inzhiniring Publ., 2006. – 348 p. 7. Lande D. Osnovy integratsii

informatsionnyh potokov / D. Lande. – Kiev: Inzhiniring Publ., 2006. – 240 p. 8. Pasichnyk V. Mathematical linguistic / V. Pasichnyk, Y. Scherbyna, V. Vysotska, T. Shestakevych. – L: Novy svit, 2012. – 359 p. 9. Chomsky N. Three models for the description of language / N. Chomsky. – I.R.E. Trans. PGIT 2, 1956. – P. 113-124. 10. Chomsky N. On certain formal properties of grammars, Information and Control 2 / N. Chomsky // A note on phrase structure grammars, Information and Control 2, 1959. – P. 137–267, 393–395. 11. Chomsky N. On the notion "Rule of Grammar" / N. Chomsky // Proc. Symp. Applied Math., 12. Amer. Math. Soc., 1961. 12. Chomsky N. Context-free grammars and pushdown storage / N. Chomsky // Quarterly Progress Reports, № 65, Research Laboratory of Electronics, M.I.T., 1962. 13. Chomsky N. Formal properties of grammars / N. Chomsky // Handbook of Mathemati-Mathematical Psychology, 2, ch. 12, Wiley, 1963. – P. 323–418. 14. Chomsky N. The logical basis for linguistic theory / N. Chomsky // Proc. IX-th Int. Cong. Linguists, 1962. 15. Chomsky N. Finite state languages / N. Chomsky, G.A. Miller // I nformation and Control 1, 1958. – P. 91–112. 16. Chomsky N. Introduction to the formal analysis of natural languages / N. Chomsky, G.A. Miller // Handbook of Mathematical Psychology 2, Ch. 12, Wiley, 1963. – P. 269–322. 17. Chomsky N. The algebraic theory of context-free languages / N. Chomsky, M.P. Schützenberger // Computer programming and formal systems, North-Holland, MR152391. – Amsterdam, 1963. – P. 118–161.

UDC 681.3:656.1

V. Mazur
Lviv Polytechnic National University, CAD Department

IMPROVEMENT OF CITY TRAFFIC NETWORK BASED ON AN ANALYSIS OF ITS FEATURES

УДОСКОНАЛЕННЯ ТРАНСПОРТНОЇ МЕРЕЖІ МІСТА НА ОСНОВІ АНАЛІЗУ ЇЇ ОСОБЛИВОСТЕЙ

The article deals with the features of city transport network and there is proposed measures on its improvement.

Key words: city transport network, 3-D model, typical element.

Розглянуто особливості транспортної мережі міста та запропоновано заходи з її вдосконалення.

Ключові слова: транспортна мережа міста, 3-D модель, типовий елемент.

Introduction

A further increase of the traffic flow intensity in conditions of the limited space and inadequate city transport network leads to the exasperation of traffic problems. The solution to these problems is particularly difficult in the central historical part of the cities of Western Ukraine, sated by a large number of intersections and narrow streets. Dense housing system and historical value of the buildings virtually eliminates the modernization of the road network on the basis of known standard approaches. Therefore, the development of measures to improve the transport network should be based primarily on the identification and consideration of its specific features. The development of specific methods and approaches to improving transport networks of the ancient city, is given, in our opinion, insufficient attention that causes the relevance of this work [1].

The objective of this work is the improvement of the transport network on the basis of its specific features. To achieve this goal the following tasks are solved: the identification of transport network models and determination of its critical cross-sections; the development of measures to improve the transport