Logico-Linguistic Model of Ukrainian Text

Nataliia Darchuk, Kryvyi Sergii, Victor Sorokin

Taras Shevchenko National University, 14, Taras Shevchenko Boulevard, Kyiv, 01601, Ukraine

Abstract

The purpose of the article is to develop a methodology for analyzing a Ukrainian-language text with two components, linguistic and logical, both of which are based on the formal apparatus of both linguistic and logical-model analysis. An example of a formal apparatus for the presentation of procedural knowledge is the computer grammar AGAT as an integral computer model of the Ukrainian language, in which the ontological system works in a complementary mode to the epistemological aspect. Its model – an active text analysis machine – hierarchically solves all the necessary tasks similarly to a human linguist, but it does so according to the rules of computer grammar, which consists of two sections according to the objects of description – morphology and syntax, as well as semantics as the final stage automatic text analysis.

Keywords

Ontograph, logical-linguistic analysis, dependency graph, syntactic-semantic relations, descriptive logic, area of interpretation, concept

1. Introduction

Knowledge acquisition from natural language texts is one of the prevalent challenges in artificial intelligence. A natural language text is the object of computational linguistics research and the subject of language and speech modeling. Based on the formal apparatus of linguistic data analysis, it is closely connected to logic, psychology, mathematics, artificial intelligence, and cybernetics. Every computational model related to the analysis of natural language texts denotes a generation and processing of declarative and procedural knowledge [1, p.75]. Analyzing such knowledge requires describing denotational and operational semantics to answer the following questions: a) what is generated or calculated? b) in which way is it generated or calculated.

It must be noted that natural language texts should be analyzed in two stages: the linguistic (syntactic and semantic) stage, as well as formal, logical modeling stage. Semantic module must be present in both stages. They are closely connected: the more accurate the results of the first analysis stage are, the better its translation into the formal logical language is.

In this paper, we aim to develop an analysis methodology of Ukrainian language texts using two components, linguistic and logical, both of which are based on the formal apparatus of linguistic and logical modeling analysis. The computational grammar AGAT is an integral computational model of Ukrainian is an example of formal apparatus of procedural knowledge representation. In AGAT, the ontological and the gnoseological aspects complement each other, functioning together. The model, an active text analysis automaton, hierarchically solves all necessary tasks analogously to a human linguist, but does so according to the rules of computational grammar. The grammar comprises two parts, morphology and syntax, as well as semantics as the final stage of text analysis.

2. Related Works

For the creation of linguistic modules for natural language text analysis, two main approaches are currently used: the first is based on rules [2], and the second is the engineering approach called "machine learning" [3; 4]. The first approach is linguistic, as it represents linguistic information in formal rules,

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, Lviv, Ukraine, April 20-21, 2023.
EMAIL: NataliaDarchuk@gmail.com (N. Darchuk); sl.krivoi@gmail.com (S. Kryvyi); victor.sorokin@gmail.com (V. Sorokin)
ORCID: 0000-0001-8932-9301 (N.Darchuk), 0000-0003-4231-0691 (S. Kryvyi), 0000-0002-3637-0535 (V.Sorokin)

sometimes embedded in the program code or in a specially created formal language. The rules are formulated by linguists themselves. Within the machine learning approach, the source of linguistic information is not the rules, but the selected texts of the problem domain. Among the methods used in this approach are supervised, unsupervised and bootstrapping learning. Supervised learning is most commonly used when building a mathematical and software model of a machine classifier that can recognize different classes of text units (words, word combinations, etc.) or texts themselves [5]. The learning is based on general regularities that are inherent in natural language texts based on data from the training sample, so knowledge is both declarative (rules) and procedural (machine learning).

Both methods have advantages and disadvantages. Creating rules is a laborious process, but it is deeply linguistic, taking into account even partial complex cases, which are quite numerous in diverse texts. Rules are declarative, easy to understand, and easy to modify depending on the results of the module's work. Machine learning does not require manually creating rules, which shortens the development time of systems, but classifiers are opaque and hard to interpret linguistically. Therefore, the AGAT grammar is chosen as the basis of the system for automatic processing of Ukrainian text.

A logical approach to the analysis of natural language texts is considered in many works, which can be divided into the following directions: a) search for coreferences in the text [6]; b) construction of specialized parsers for highlighting the semantic properties of the text [7; 8]; c) a direction that partially includes directions a) and b) and is oriented towards obtaining knowledge from the text [9; 10]; formally logical and ontological direction [11;12], and the direction of transformational analysis of texts [13; 14].

The scientific novelty, theoretical and practical value of the results

Natural language processing is one of the main computer science tasks today. This is largely due to the desire of humankind to overcome language barriers and also due to the dozens of practical tasks, such as: methods of automatic translation, referencing and annotation, real-time speech recognition, including natural language commands, automatic search, constructing responses to questions, detecting and correcting grammatical errors, building natural language dialogue systems, text coherence checking, sentiment analysis, etc. Any developments in this field deepen theoretical linguistic knowledge and solve practical tasks, as they are mostly linguistic, related to the definition of parts of speech, lemmatization, building dependency trees, coreference resolution, named entity recognition, establishing structural and semantic incompleteness of sentences, detecting connections and relationships between language units. The scientific novelty of the results of this research lies in the combination of knowledge from natural language texts and powerful mathematical logic apparatus, which allows representation, analysis and knowledge extraction from unstructured natural language texts. In Ukraine, there are no similar research studies.

3. Methods

Methods of structural linguistics are used in linguistic analysis modules: distributive analysis, constituency and dependency tree construction, and component analysis. Automatic morphological analysis module uses the distributive method, automatic syntactic analysis module relies on constituency and dependency trees, and automatic semantic analysis module utilizes component analysis.

The logical component utilizes the results of formal grammar analysis. Full automation of the logic-modeling stage encounters the problem of choosing a formal logical language in which knowledge obtained on the first stage is presented and studied, and depends on the complexity of the input text T. This problem is solved in the following way: usually a first-order predicate language is chosen for working with knowledge, as it is expressive enough and has well-developed algorithmic tools. This choice is also confirmed by the fact that the selection of higher-order logical languages has a high complexity of the analysis process and insufficiently developed tools for logical inference.

4. Experiment

Automation of linguistic research is associated with the creation of systems for automatic processing of written Ukrainian text. The stages of computer language analysis are:

- tokenization - segmentation of letter sequences into words and sentences;
- morphological analysis - part-of-speech and categorical grammatical information;
- syntactic analysis – automatic construction of trees of sentence dependencies, the result of which is also a marked tree of subordination, attribution to each pair of words of the type of

syntactic connection and syntactic-semantic relations at the level of the morphological way of expression of the "owner";

- semantic analysis - determining the meanings of individual sentences or their parts;
- logic-model analysis, i.e. translation of the input text into the language of mathematical logic in order to identify contradictions, illogicalities in the expression of meaning and the possibility of obtaining information relevant to the request from it.

Each analyzed text is a separate file in XML format, which contains morphological information about all the word forms of the text (the lemma and its set of grammatical features), as well as the syntactic structure of each sentence in the form of a dependency graph. All branches of the tree are marked with names of syntactic relations (coordination, subordination, conjunction), semantic-syntactic relationships (6 of them: subject, object-direct and indirect, attributive, adverbial, and completive; and 6 conjunctive ones: identical-conjunctive; contrasting-conjunctive; comparative-conjunctive; explanatory-conjunctive; joining-conjunctive; separating-conjunctive). The analysis modules use a morphological dictionary containing 200,000 lemmas and a syntagm grammar, which includes hundreds of rules. The syntactic-semantic annotation, however, is built automatically, and its results are necessarily corrected.

The following figure illustrates the stages of text analysis:

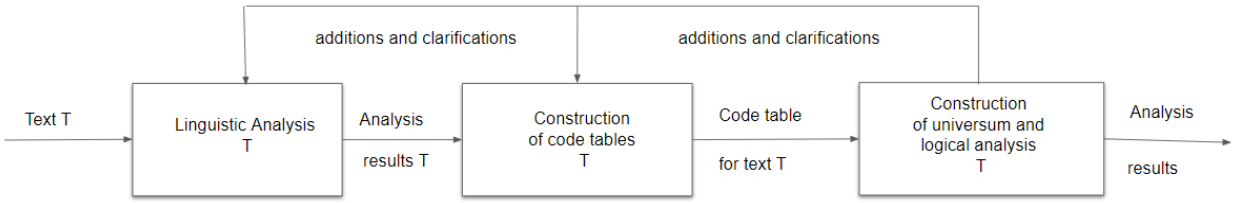

Figure 1: The stages of logical-linguistic analysis of the text T

Stage 1. Morphological analysis of the given text *T* in order to construct a dictionary of words and word forms for text T and division of the set $L = \{L_1, L_2, L_3, L_4\}$ into classes (parts of speech and categorical grammatical characteristics). A lexical analysis is provided with the establishment of deverbatives, deadjectives, and denominatives.

Stage 2. Construction of a set of objects *D*, based on the results of automatic syntactic analysis of text *T* and stage 1 results. At this stage, the terms combining several words, anaphoric connections, etc. are found.

Stage 3. Comparison of the set of objects D with the data of the information and search thesaurus

Stage 4. Construction of an ontograph based on the set of objects D (construction of relations R_L) sing classes $L = \{L_1, L_2, L_3, L_4\}$. The ontograph of the text is built on the basis of sentence ontographs by applying conjunction and simplification rules.

Based on the stages of **logical-linguistic analysis** presented in the figure above, the following algorithm can be presented:

LOGICAL-LINGUISTIC ANALYSIS OF TEXT (T)

Input: Initial text T.

Output: Results of queries to knowledge base of the text T.

Method:

Algorithm start

1. Enter the initial text T;
2. Carry out syntactic-semantic analysis of T;
3. Based on the results of the analysis of T, construct a table (i) of codes of classes of text T;
4. Based on the table (i) of codes, construct a universe B for text T;
5. Give an interpretation of the universe using an information-search thesaurus.
6. Carry out logical analysis of the universe.

6.1. Check the obtained facts for inconsistency.

6.2. If the facts are not inconsistent, then enter them into the knowledge base and generate answers to queries to this knowledge base.

Algorithm end

5. Results

The analysis of a scientific text on marketing is presented. The length of the text "Marketing Distribution Policy" is 4142 tokens; it contains 204 sentences. We will illustrate all stages of the automatic text analysis using both short and long sentences as an example.

1) *Головним у маркетинговій політиці розподілу є формування відповідних каналів .*
2) *Важливість цього питання визначається такими обставинами : вибраний канал розподілу справляє принциповий вплив майже на всю маркетингову програму підприємства; формування каналу розподілу передбачає укладення тривалих комерційних угод з його суб'єктами , які потім дуже важко змінити , нехай навіть вони й будуть помилковими ; між суб'єктами каналів часто виникають конфлікти , які погано відбиваються на результатах збутової діяльності підприємства ; користувач каналами розподілу (продуцент товарів) часто тією чи іншою мірою втрачає безпосередній контроль над ринком збуту.*

Figure 2 shows the first sentence and the dependency table automatically built by the program.

Головним (АЦ) у (ПП) маркетинговій (АН) політиці (КП) розподілу (ЙР) є (ГЮ) формування (ЛИ) відповідних (АЕ) каналів (ЙЕ

Головним	у	adjectival subjunctive compound
у	політиці	prepositional compound
політиці	маркетинговій	noun compound without preposition
політиці	розподілу	noun compound without preposition
є	формування	coordination communication
є	Головним	connection + adjective
формування	каналів	noun compound without preposition
каналів	відповідних	noun compound without preposition

Figure 2: A fragment of the automatic construction of word combinations in the sentence 1)

The table in the figure contains sentences with morphological annotations, as well as information about the part-of-speech and categorical features of words. The table has three columns: the first column is the main member of the binary phrases, the second column is the subordinate member of the phrase, and the third column is the syntactic information about the type of phrase. This makes it possible for the program to create an alphabetical frequency dictionary of text word combinations upon completion of the program. Alphabetical frequency dictionaries have been built for specific lexemes and classes of words. Among the most frequent nouns are "канал" (103), "розподіл" (89), "товар" (58), "споживач" (49), "посередник" (35), "ринок" (32), "підприємство" (28), "товаровиробник" (25), "рівень" (24), "продукція" (22) and so on.

Figure 3 demonstrates a graphical representation of a dependency tree created by automatically inverting the dependency table.

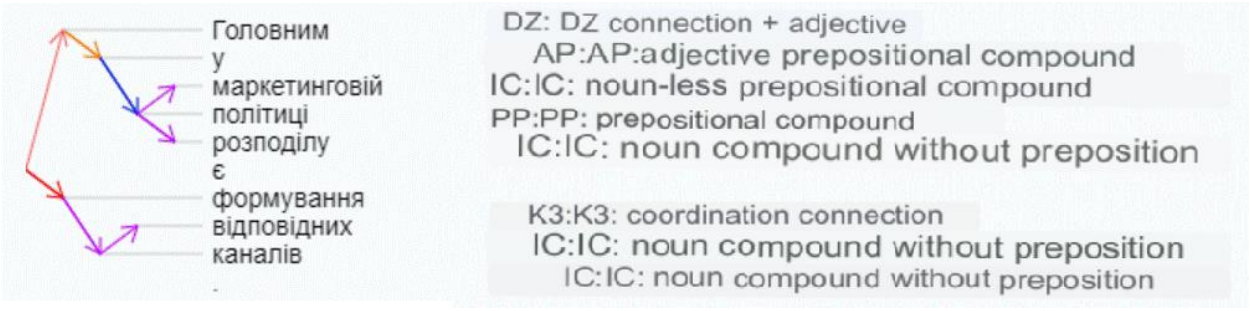

Figure 3: The dependency graph of the sentence 1)

The dependency tree consists of nodes and edges, where nodes represent the words, and edges illustrate the relations between head words and dependents of a phrase. Aside from that, additional information on types of relations between nodes is given. This makes it possible to describe the configuration, form, and outer parameters of the sentence. However, this is not enough to present the structure of the sentence. The information about the type of relations between the constituents of the phrase and semantic-syntactic relations is automatically applied to the set of tree edges. This helps with analyzing complex correlations between semantics and its formal representation, as the text is parsed automatically based on the formal features of its units. Thus, automatic syntactic analysis of the sentence is done on two levels: 1) for each phrase, the program determines its syntactic type based on the morphological features of its head; 2) syntactic relation type is determined for each edge of the graph

Figure 4: Graph with syntactic-semantic relationships.

In addition to relationships between words in a sentence, we observe another, more important type of ordered relationships - relationships between groups of words, or word combinations, and for their representation, a formal structure of another type is needed - the constituent structure. By analyzing the sentence in Figure 3, 4, intuitively, we can divide it into segments that have a hierarchical structure, in which some have a common part, that is, one part is included in another. The sentence is automatically divided into segments that form a hierarchical structure:

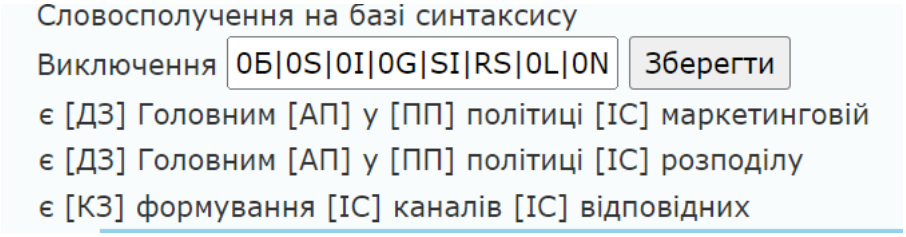

Figure 5: The system of constituents (propositions) of the sentence

If the constituents have a common part - one completely falls within the other - the system of constituents is considered the formal model of the sentence. The constituent "є головним у маркетинговій політиці розподілу" is combined with the dominant one "є формування відповідних каналів" with a subordinate link, because the predicative pair "є формування" is the constructive center of the sentence, and "є головним…" expands the group of predicate. Studies have shown that the constituent includes not only individual words, but also nested "complex" constituents, for example, an adverbial clause. Therefore, the list of constituents demonstrates that not only individual words are the syntactic units in the sentence, but also whole word combinations or **groups**: [[Головним [у] [маркетинговій [політиці]] розподілу] // [[є] формування] [відповідних каналів]]. Note that constituents cannot overlap, but can "nest". This means that if one word or a group of words simultaneously is a part of two or more constituents, one of them completely envelops the other one. Following up, it is determined which sets of such structural units (constituents) belong to the same grammatical class. The structure of constituents illustrates this, represented in the form of a labeled tree. They add up in the structure of components, forming a system of sentence components.

Considering the content part of the sentence constituents, I. R. Vykhovanets notes that semantic researchers often associate the semantic organization of a sentence with its formal organization, using the

concept of semantic sentence structure, qualifying it as the meaning of the sentence, presented in a generalized form taking into account those elements of meaning that are outlined by the sentence's form [15, p. 121]. The objective content of the sentence is best reflected in the concept of a **proposition**. This is a stable core, a constant of the sentence, which reflects the structure of the described situation. The linguist notes that the structure of the proposition is determined by the predicate. The predicate indicates the nature of the situation - in our case this is the root of the tree - and the corresponding places for objects - the participants of the situation - these are the actants, arguments, represented by groups of subject and predicate, quality and functions of which are determined by the predicate. And only the semantic nature of the predicate determines the number and roles of the actants. Thus, two aspects of study are relevant: the semantics of predicate words and the semantic roles of actants. The first aspect - the semantics of predicate words - already has its form as semantic domains of verb classes, predicative adverbs, to which a semantic class number is assigned, but semantic roles require more research.

Automatically obtained constituents can be considered as the raw input for forming n-ary predicates, as the primary way of expressing its content. The semantics of the predicate is determined by a certain semantic class to which the predicate belongs. Therefore, we can determine the ways of expressing propositions (propositions of movement, speech, sound, mental sphere, emotions, etc.), the number of actors, etc. Using a large corpus material, subcorpora of constituents associated with certain propositions can be formed, from which one can distinguish the constituents which is most frequently used for a specific proposition.

The aforementioned sentence is simple, extended, and declarative. Let's analyze its content. It is about the marketing distribution policy, but the noun-deverbative "розподілу" with the meaning of "placement" requires clarification: distribution of what? (see The Dictionary of the Ukrainian Language: in 11 volumes). The subject group "формування відповідних каналів" is also incomplete, since the word "канали" is used in a figurative sense as "means and ways of achieving something." Therefore, from a semantic and pragmatic point of view, this sentence is poorly constructed.

The semantic-logical model of the aforementioned sentence constitutes a predicative-argumentative structure of the following type. Check Table 1:

Table 1

Predicative-argumentative structure sample

Parent node	Child node	Syntactic relation	Type of relation	Semantic class of the parent	Semantic class of the child
Головним	у	Adjectival prepositional phrase	Object-adjective	O0ocinka0posit ,rqual	
У	політиці	Prepositional phrase	Prepositional		t0imen,r0rel
політиці	маркетинговій	Noun phrase	Attributive	t0imen,r0rel	dr0imen, r0rel
політиці	розподілу	Noun phrase	Possessive	t0imen,r0rel	T0put
є	головним	Link verb + adjective		V0poss	O0ocinka0posit ,rqual
є	формування	Coordinative	Subjective	V0poss	der0v,t0activity
формування	каналів	Noun phrase	Possessive	der0v,t0activity	t0activity
каналів	відповідних	Noun phrase	Attributive	t0activity	R0rel,dr0men

All the information contained in the columns of the table is obtained automatically: columns 3-4 (see Figure 3), and 5-6 are the result of automatic semantic analysis with the assignment of taxon codes according to the synoptic scheme (see http://www.mova. info/wnetschema2.aspx0).

The predicate structure = argument1 + argument2 + argument3 ... argument n. Let's make a few remarks about the way the predicate structure is presented: the members of the predicate pair are separated by a sign (–); it is separated from the argument structure by a sign (=), arguments are joined by a sign (+). The ways

of expressing all members of the predicate structure are lexemes of meaningful parts of speech (nouns, verbs, adjectives). Service words (prepositions, conjunctions, particles) are not presented.

The predicate pair (ФОРМУВАННЯ КАНАЛІВ) – ACTION/БУТИ ГОЛОВНИМ) = argument 1 (МАРКЕТИНГОВА ПОЛІТИКА) + argument 2 (ПОЛІТИКА РОЗПОДІЛУ)

Sentence 2

Важливість цього *питання визначається такими обставинами : вибраний канал розподілу справляє принциповий вплив майже на всю маркетингову програму* ***підприємства; формування*** *каналу розподілу передбачає укладення тривалих комерційних угод з його суб'єктами , які потім дуже важко змінити , нехай навіть вони й будуть помилковими ; між суб'єктами каналів часто виникають конфлікти , які погано відбиваються на результатах збутової діяльності підприємства ; користувач каналами розподілу (продуцент товарів) часто тією чи іншою мірою втрачає безпосередній контроль над ринком збуту.*

This sentence follows the first one in the text. It is a complex sentence with a compound-complex sentence structure. The sentence consists of 70 words and eight predicative parts. Figures 5a-5e represent fragments of the dependency graph, which is automatically constructed based on types of syntactic and semantic-syntactic relations, which allows us to simplify the sentence, identify coreference links and restore the entities, and prepare for logical analysis.

Figure 6

From the **first** predicative part (Figure 6), two propositions are automatically extracted: "важливість цього питання" and "визначається такими обставинами". The demonstrative pronoun "цього" indicates a connection with the previous sentence on the semantic level. Its antecedent is "формування відповідних каналів". The pronoun "такими" actualizes the noun "обставини", and there is a colon after it.

The predicate structure of the predicative part (Figure 6) looks like this:

The predicate pair (ВАЖЛИВІСТЬ (ФОРМУВАННЯ КАНАЛІВ)) – action (ВИЗНАЧАЮТЬСЯ) = argument 1 (ОБСТАВИНАМИ)

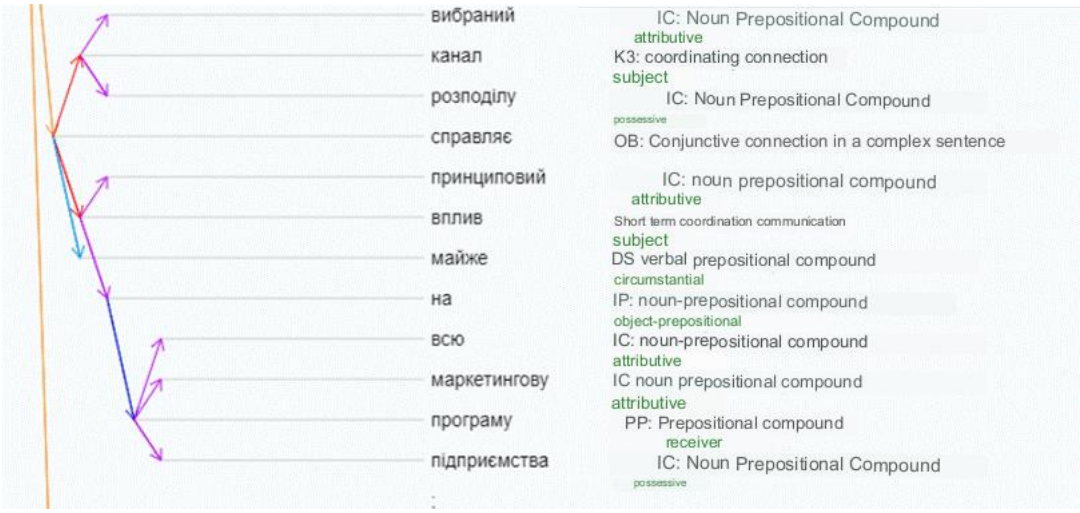

Figure 7

From the **second** predicative part (Figure 7), two propositions are extracted:

"канал справляє вплив на програму підприємства" and "справляє вплив на програму маркетингову".

The predicate pair (КАНАЛ РОЗПОДІЛУ) – action (СПРАВЛЯЄ ВПЛИВ) = argument 1 = (ПРОГРАМУ ПІДПРИЄМСТВА) + argument 2 = (ПРОГРАМУ МАРКЕТИНГОВУ)

From the **third** predicative part, three such propositions are extracted:
"формування каналу розподілу передбачає"; "передбачає укладення тривалих комерційних угод з його суб'єктами"; "які потім дуже важко змінити".

The predicate pair (ФОРМУВАННЯ КАНАЛУ РОЗПОДІЛУ) – action (ПЕРЕДБАЧАЄ УКЛАДЕННЯ) = argument 1 = (ТРИВАЛИХ КОМЕРЦІЙНИХ УГОД) + argument 2 = (ЙОГО СУБ'ЄКТАМИ) + argument 3 = predicative pair (0) – action (ВАЖКО ЗАМІНИТИ); argument 4 = (ЯКІ)

Upon the construction of a predicative pair, the antecedents of the possessive pronoun "його" and its conjugate word "які" remain unclear. It is hard to understand whether the author is speaking in regards to "угоди, які важко потім замінити", or "суб'єкти комерційних угод". Such logical errors have an impact on the final result of linguistic and logical analysis of a scientific text.

From the **fourth** predicative part *нехай навіть вони й будуть помилковими* , one proposition is extracted: The predicate pair (ВОНИ) – action (НЕХАЙ БУДУТЬ ПОМИЛКОВИМИ) It is complicated to determine the antecedent of the pronoun "вони".

From the **fifth** predicative part *між суб'єктами каналів часто виникають конфлікти, які погано відбиваються на результатах збутової діяльності підприємства*, such propositions are extracted:

The predicate pair (КОНФЛІКТИ) – action (ВИНИКАЮТЬ); argument 1 = (СУБ'ЄКТ КАНАЛІВ) = the predicate pair (ЯКІ) – action (ВІДБИВАЮТЬСЯ); argument 2 (РЕЗУЛЬТАТ ДІЯЛЬНОСТІ) + argument 3 (ЗБУТОВОЇ ДІЯЛЬНОСТІ) + argument 4 (ДІЯЛЬНОСТІ ПІДПРИЄМСТВА)

From the **sixth** predicative part *користувач каналами розподілу (продуцент товарів) частотією чи іншою мірою втрачає безпосередній контроль над ринком збуту*, such propositions are extracted:

The predicate pair (КОРИСТУВАЧ КАНАЛАМИ РОЗПОДІЛУ) – action (ВТРАЧАЄ КОНТРОЛЬ); argument 1 = (КОРИСТУВАЧ= ПРОДУЦЕНТ ТОВАРІВ) + argument 2 (КОНТРОЛЬ НАД РИНКОМ ЗБУТУ)

Such method of formulating a predicative structure allows to identify the elements of text that can be disregarded as insignificant from the point of view of meaning, such as: (вплив) ***майже*** (на); (які) ***потім дуже*** (важко…); ***нехай вони будуть помилковими***; ***Часто*** (виникають); (які) **погано** (відбиваються); ***Часто тією чи іншою мірою.*** The simplification of text occurs through the use of particles, qualitative adverbs, which are specified by a list. In contrast, coreferential connections, which restore the antecedent of the text, are important both for the transmission of the text's meaning and as a linking element between the sentences of the text.

An important part of linguistic processing is the dictionary component - the thesaurus of terms of the subject area (SO) "Marketing" of the information-search type, where each term is represented in a network, the nodes of which are terms, and the arcs are relationships between terms (http://www.mova.info/thes_nl.aspx). Automatic comparison of the predicate structures obtained from the text with the terminological network of the thesaurus is the basis for building an ontograph for the above sentences (see Figure 8)

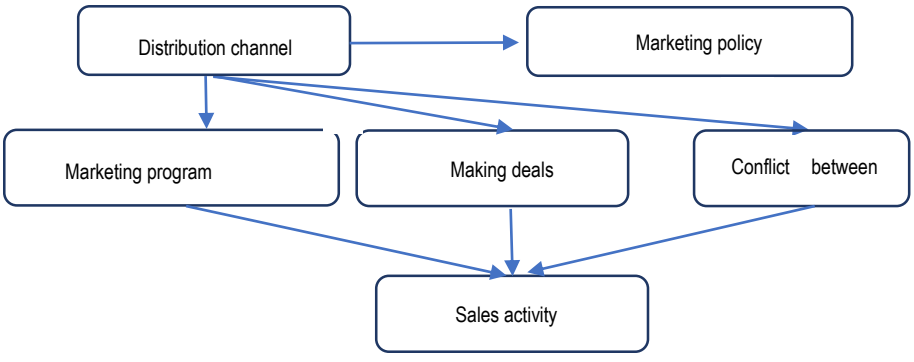

Figure 8: The ontograph of the text

R1 – predicate «бути головним», R2 – predicate «впливати», R3 – predicate «передбачати», R4 – predicate «провокувати», R5 – predicate «бути результатом».

This concludes the linguistic syntactic-semantic analysis and leads to logical analysis.

Logical analysis of the results of linguistic text processing

The purpose of logical analysis is to verify the consistency and logical compatibility of facts. To perform logical analysis of the results of linguistic syntactic-semantic analysis, it is proposed to use the tools of **descriptive logics and ontologies,** because descriptive logics have a reliable algorithmic basis, and the use of ontologies is a direct way to build a knowledge base. The knowledge base accumulates knowledge obtained from the original text, but this knowledge must be consistent (compatible) in a logical sense. The results of linguistic analysis of the input text presented in the form of an ontograph of this text, can be checked for inconsistency. To do this, it is necessary to interpret the objects of the universe and determine the formal logical language in which the checks of the properties of the acquired knowledge and the generation of consequences that arise from this knowledge will be performed. Currently, the most suitable descriptive logic for performing such tasks is ALC-logic and some of its extensions. For this logic, algorithms for generating consequences and checking the consistency of a set of knowledge represented by formulas of this logic have been developed. Now, let us consider the formal definitions.

The creation of ontology-like systems is based on the concept of **ontology**.

Definition 1. An ontology is defined as an ordered triple

$$O = (X, R, F),$$

where X is a finite set of concepts,, R – s a finite set of binary (semantic) relations defined on X, i F– is an interpretation function on a domain D of elements from X and R, such that $F : X \cup R \to D.$ For example, X (concept)= {(МАРКЕТИНГОВА) ПОЛІТИКА, (МАРКЕТИНГОВА) ПРОГРАМА, РОЗПОДІЛ ТОВАРУ, ЗБУТОВА ДІЯЛЬНІСТЬ, КОМЕРЦІЙНА УГОДА, СУБ'ЄКТ КАНАЛУ РОЗПОДІЛУ, ПІДПРИЄМСТВО, etc.}; relations (roles) = {**R1 – «бути головним» , R2 – «впливати», R3 –«передбачати», R4 –«провокувати», R5 – «бути результатом»}.**

If A = 'канал розподілу', B = 'маркетингова політика розподілу', then the formula AR_1B means: *'канал розподілу'* **is prevalent** in «*маркетинговій політиці розподілу».*

When constructing an ontology, the subject area (SA) is specified, to which the concepts from X and the relations from R pertain. In this case, it is Economics (marketing). The specification of the SA is necessary for defining the interpretation of F. The relationship of F with the SA may introduce additional corrections to the definition of F. These additional corrections are described by the axioms A of the given SA and the restrictions R_c, which have the form of additional definitions (clarifications, limitations on possible values, etc.), and properties from the interpretation area D of the given ПО. Thus, we arrive at a refined definition of the ontology for a specific SA.

Definition 2. A refined ontology is defined as an ordered quadruple $O = (X, R, F, A (D, R_c))$, where X is a finite set of concepts (terms), *R s a finite set of binary (semantic) relations defined on* X, i F is an interpretation function on a domain D of elements from X and R, and $A(D,R_c)$ are additional constraints $\{R_c\}$that are described by axioms A on the domain D.

The difference between definitions 1 and 2 lies in the following [1].

a) The set of concepts X (in this case, terms) in definition 1 is oriented towards the problem (Economics) to be solved, while in definition 2 this set is specified (the subject area is "Marketing") and should be as complete as possible for the given SA and should be constructed using automated means (from dictionaries and texts).

b) The set R in definition 1 is established by experts in the relevant subject area, while in definition 2 it should be executed on the set D, built using automated means and verified for consistency by the logical deduction system.

c) The interpretation function F in definition 1 is chosen by the user according to their professional competence, own or reference information, and in definition 2 this function is formed based on general sources of text information such as encyclopedias, dictionaries, results of syntactic and semantic analysis, etc. For example, for the authors of this work, using definition 1 is sufficient from the perspective of their competence.

d) The set of axioms A in definition 2 describes additional specific definitions of concepts from D and limitations on the interpretation of R_c for a given SA.

Therefore, it is necessary to define the subject area because the same concepts in different SA may have different meanings. The system input are texts related to the given SA (for now only in Ukrainian, although the further development of the system is planned by including other languages, in particular English).

The interpretation area of concepts and ontological relationships is represented by a set of concepts X of text T, on which the terminology of the SA is built, in which these concepts and a set of semantic relationships R between concepts are interpreted. For the set X, the interpretation area X^F is divided into classes (for example, proper/common names, names of individuals, abstract/concrete names, expertise, etc.)

[10]. This division uses the results of syntactic-semantic analysis, which builds classes of concepts by their types. In addition, syntactic dependencies between sentence members are found by this analysis and illustrated in the form of an acyclic graph. Syntactic dependencies, by a certain relationship between sentence members, carry certain semantic information that is used to detect semantic features and potential semantic links between lexical units. Detection of semantic features is not done according to rules (they simply do not exist), but depends on the goal of the analysis, on researchers, and on the developers' skills.

The attributive language *AL* is the basis for the descriptive logics (DL). *AL* contains the set of **atomic concepts** CN and the set of **atomic roles** RN (binary relationships on CN) [10,11]. More complex concepts and relationships are built using **constructors**.

The semantics of concepts and relationships is built according to set theory, and the following concept constructors are used: union of concepts, existential quantifier, numerical restriction, and negation of any concept. The semantics of the concept language is a fragment of the first-order predicate language.

The extension of the AL language by some subset of constructors gives a specific descriptive logic. If we add to the AL language the constructor of negation (C - complement), called the **complement** of the concept, we get the **ALC** logic. This logic forms the core of the entire family of descriptive logics.

The formal description of the syntax and semantics of the ALC logic can be found in [11], and so we do not introduce these concepts here, but instead return to our example, specifically to Figure 8. This figure represents the results of the linguistic analysis of the input text T. The onthograph on the Figure 8 accumulates the set of concepts C and the set of relations R.

Indeed, in this case we have:

C = {C_1 = канал-розподілу, C_2 = маркет-політ.-розподілу, C_3 = марк.-прогр.підпр,
C_4 = уклад-комер-угод, C_5 = конфліктн-ситуац, C_6 = збут-діял-підпр,…},

R = {R_1 (C_1, C_2), R_2 (C_1, C_3), R_3 (C_1, C_4), R_4 (C_1, C_5), R_5 (C_3, C_6), R_5 (C_4, C_6), R_5 (C_5, C_6),}.

These sets may not actually be atomic. To ensure that these are indeed atomic concepts and roles, we need an interpretation of these sets. The interpretation will determine which concepts are atomic and which are derived from atomic concepts, and the same will be done for roles.

Therefore, the information presented in Figure 8 is a high-level **partially interpreted ontology** template, which after clarification of the subject area and interpretation is transformed into an ontology in which logical analysis is performed and after logical analysis, an ontological knowledge base is constructed. How does this happen? Let's consider our example.

Example 1. Let's consider a given SA and the interpretation of concepts that appear in the text T about the marketing policy of a company:

Objects = {C_1 = канал-розподілу, C_{11} = легальний, C_{12} = нелегальний, С6 = контроль-над-ринком-збуту, C_5 = розв’язання-конфл-ситуацій, C_7 = розподіл-товарів, C_8 = контрабанда, C_9 = наркотрафік, C_2 = маркет-політика, C_3 = програма-маркет, C_4 = комерц-угода, C_{21} = підприємство, C_{22} = товари, C_{23} = користувач-каналами, C_{24} = суб'єкти}

Let the terminology of the given SA be:

Канал ≡ легальний ⊔ нелегальний,

Легальний ≡ тривала-комерц-угод ⊔ нетривала-комерц-угода ⊔ контроль-над-ринком-збуту ⊔ розподілу-товарів ⊔ розв’язання-конфл-ситуацій,

Нелегальний ≡ контрабанда ⊔ наркотрафік ⊔ розв’язання-конфл-ситуацій,

Програма-маркет ≡ програма-підприємства ⊔ тривала-комерц-угода ⊔ нетривала-комерц-угода ⊔ збутова-діяльність-підпр,

Підприємство ≡ виготовлення-лікарських засобів ⊔ виготовлення-косметики,

Товари ≡ {серцеві, діабетичні, ортопедичні} ⊔ {шампуні, гелі, креми}.

From this terminology, such atomic concept sets can be extracted:

CN = {C_3 = програма-підпрємсва, C_4 = комерц-угод, C_{41} = нетр-комерц-угод, C_6 = збутова-діяльність-підпр, C_8 = контрабанда, C_9 = наркотрафік, C_5 =розв’яз-конфл. ситуацій, C_9= серцеві, C_{10}= діабетичні, C_{11}= ортопедичні, C_{12}= шампуні, C_{13}= гелі. C_{14}= креми}.

The ontograph from Figure 8 gives us the set of roles RN (binary relations) of concepts:

DN = {R_1 (C_1, C_2), R_2 (C_1, C_3), R_3 (C_1, C_4), R_4 (C_1, C_5), R_5 (C_3, C_6), R_5 (C_4, C_6), R_5 (C_5, C_6)}.

It is clear that the set DN can be expanded with additional relations, for example, one can add the relation R_1 (C_1, C_9).

End of example 1.

The terminology allows us to record general knowledge about concepts and roles, but in addition, it is also necessary to record knowledge about specific objects or individuals. For example, we need to understand to which concept they belong and how they are connected to each other. This is found in that

part of the knowledge base, which is called the system of facts about individuals or *ABox*. For this purpose, in addition to the set of atomic concepts CN and the set of atomic roles RN, a finite set of IN - individual names is introduced.

For example, if we return to the initial text (sentences):

1)Головним у маркетинговій політиці розподілу є формування відповідних ***каналів*** .

2)Важливість цього питання визначається такими обставинами : вибраний ***канал*** *розподілу справляє принциповий вплив майже на всю маркетингову програму* ***підприємства****; формування* ***каналу розподілу*** *передбачає укладення тривалих комерційних угод з* ***його*** *суб'єктами* ***, які*** *потім дуже важко змінити , нехай навіть вони й будуть помилковими ; між суб'єктами каналів часто виникають конфлікти, які погано відбиваються на результатах збутової діяльності підприємства; користувач каналами розподілу (продуцент товарів) часто тією чи іншою мірою втрачає безпосередній контроль над ринком збуту.*

we see these concepts are not clearly defined. The absence of clarity in the formulation of the concept-term "політика розподілу", "канал розподілу". This is partly compensated by interpretation and terminology, but not fully. The semantics of the word "розподіл" is semantically limited and requires clarification:

1) In the phrase "Важливість цього питання", the noun "питання" does not correlate with the previous sentence, where this word is not actualized and there is no question mark at the end of the sentence.
2) The term "обставини" encompasses various concepts, including: "канал розподілу товарів"; "процес формування каналу" and "укладення угод", while the information is vague and unclear.
3) The fact presented in "користувач каналами розподілу часто тією чи іншою мірою втрачає безпосередній контроль над ринком збуту" is unclear and not necessarily true.
4) Forming channels for distributing their own products, the enterprise cannot but find answers to three questions.

Despite the absence of clearly defined concepts and their semantic meanings, logical analysis can be performed on partially interpreted ontology. This is because with the tools of logical-mathematical analysis, we process not words and their compatibility, but the compatibility of concepts represented by codes of corresponding concepts. This explains the presence of arrows on Figure 1, which are labeled "Refinement and Specification" and relate to both semantic and logical refinement. Logical refinement may be required by terminology (terminology will be contradictory if its axioms are contradictory) and facts. If terminological contradictions are resolved at the syntactic level, factual axioms require resolution by logic, which is used in such analysis.

5. Discussion

An obvious property of any approach to resolving a specific task is its potential for automation and efficiency. The complexity of algorithms used in linguistic analysis is assessed as follows: the morphological module processes 1000 word forms in 1.2 seconds, the syntactic-semantic module in 10-20 seconds, and the complexity of algorithms for checking the compatibility of a knowledge base and working with ontologies for language ALC (in the aforementioned algorithm, steps 4-6) belongs to the class of PSPACE-hard complexity [11]. Such activities are supported by tools such as OWL and Protégé based on the ALC description logic and its extensions, which are specifically designed for the creation of ontologies and knowledge bases [16]. Full automation of knowledge base construction currently seems problematic, as certain details must be clarified and added by experts in the software.

This opinion is held by the majority of developers of ontological knowledge bases

This last assessment leads to a certain skepticism in the community of practitioners and, in particular, linguists, programmers, and knowledge base administrators. Upon such criticism and skepticism, the response emerges from the described approach and the possibility of its implementation in practice. Full automation at this stage appears somewhat problematic, as certain details must be clarified by an expert in the field. The expansion of ontology and the knowledge ontological base and the clarification of its concepts, related to a specific subject area, should be done by the appropriate expert or experts in this subject area.

6. Conclusions

Based on the above, the following conclusions can be drawn. The combination of linguistic (semantic-syntactic) analysis and logical-modeling and ontological paradigm allows us to assert that the process of acquiring knowledge and consequences from these inferences can be significantly automated. The significance of the proposed method is seen in the perspectives of development as both linguistic and logical analysis of the input text. It is necessary to use the method of automatic construction of an information-searching thesaurus of a certain subject. Therefore, the task is to develop as many thesauruses as possible for different fields of science and technology [2]. The projection of a thesaurus onto a specific text of a certain subject will help create a semantic network of the text, then the combination of syntactic-semantic relations with logical thesauruses will be the starting point for applying the logical-modeling method.

7. References

[1] O.L. Semotyuk, Modern technologies of linguistic research, Lviv, 2011, pp. 151.

[2] N.P. Darchuk, Computer annotation of Ukrainian text: results and prospects. Kyiv: Education of Ukraine, 2013, 543 p.

[3] D.V.Lande, I.Yu.Subach, A.Ya. Gladun, Processing of extremely large data sets (Big Data): tutorial, Kyiv: KPI named after Igor Sikorskyi, Polytechnic Publishing House, 2021. ISBN 978-966-2344-83-7

[4] S.Lai, K. S. Leung, Y.Leung, SUNNYNLP at SemEval-2018 Task 10: A Support-Vector-Machine-Based Method for Detecting Semantic Difference using Taxonomy and Word Embedding Features. – Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, P.741-746. http://doi.org/10.18653/v1/S18-1118

[5] S.L.Kryvyi, N.P. Darchuk, O.I.Provotar, Ontology-like systems for the analysis of natural language texts. J. "Problems of programming", 2018, No. 2-3, P.132-139 (proceedings of the international conference J."UKRPROG-2018". DOI: 10.15407/pp2018.02.132

[6] L. Xu, J. D. Choi, Revealing the Myth of Higher-Order Inference in Coreference Resolution, Proceedings of the 2020 , Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, P. 8527–8533.

[7] K.Mrini, F.Dernoncourt, T.Bui, W.Chang, N. Nakashole, Rethinking Self-Attention: An Interpretable Self-Attentive Encoder-Decoder Parser, 2019.

[8] W.Che, Y. Lui, Y.Wang, B. Zheng, T. Liu, Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. – Proceedings of the {CoNLL} 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. – Association for Computational Linguistics, 2019, P. 55-64.

[9] J. Zha n, H.Zhao, Span Model for Open Information Extraction on Accurate Corpus. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 2020, P. 9523-9530.

[10] H. Hoherchak, N. Darchuk, S.Kryvyi, Representation, analysis, and extraction of knowledge from unstructured natural language text, Cybernetics and Systems Analysis, 2021, Volume 57, N 3., P. 164–183. https://doi.org/10.1007/s10559-021-00373-7

[11] F. Baader, D.Calvanese, D.L.McGuinness and other, The Description Logic Handbook, Cambridge, University Press, 2007, pp 601.

[12] H. Hoherchak, Knowledge Based and Description Logics Applications to Natural Language Texts Analysis, Proceedings of the 12th International Scientific and Practical Conference of Programming (UkrPROG 2020), 2021, Volume 2866, P. 259-269.

[13] D.Rothman, Transformers for Natural Language Processing (2nd addition), publishing Packt, 2021, pp 384.

[14] M.-W. Devlin, K. Chang Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT, 2019, P. 4171–4186.

[15] I.R. Vyhovanets, Grammar of the Ukrainian language. Syntax, Kyiv,: Lybid, 1993, pp 365.

[16] OWL Full,OWL DL and OWL Lite. – http:// www.w3.org/TR/owlquade/#Sublanguage-def.