Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

**DOI: 10.37943/AITU.2020.75.91.002**
**UDC: 004.891.2**

**O. Kravchenko**
Candidate of Technical Sciences, Department of Information Systems and Technology
kravchenko_ov@gmail.com, orcid.org/0000-0002-9669-2579
Taras Shevchenko National University of Kyiv, Ukraine

**Zh. Plakasova**
Senior Lecturer, Department of Automated Systems Software
zh.plakasova@chdtu.edu.ua, orcid.org/0000-0003-3911-2600
Cherkassy State Technological University, Ukraine

**M. Gladka**
Assistant, Department of Information Systems and Technology
miragladka@gmail.com., orcid.org/0000-0001-5233-2021
Taras Shevchenko National University of Kyiv, Ukraine

**A. Karapetyan**
Candidate of Technical Sciences, Department of Information Technology Design
anait.r.karapetyan@gmail.com., orcid.org/0000-0002-7412-3252
Cherkassy State Technological University, Ukraine

**S. Besedina**
Candidate of Technical Sciences, Associate Professor, Department of Information Technologies
besedina_sv@ukr.net, https://orcid.org/0000-0002-5391-643X
Bohdan Khmelnytsky National University of Cherkasy, Ukraine

# APPLICATION OF INFORMATION TECHNOLOGIES FOR SEMANTIC TEXT PROCESSING

**Abstract:** An expert system for text analysis based on the heuristic knowledge of an expert linguist is proposed. Methods of linguistic analysis of the text through the use of computer technology have been further developed. Data verification was performed on the example of the Germanic language group. The algorithm of the system operation is given. The sequence of actions of the text analysis process is described.

Research relates to the subject of computational linguistics and helps to automate text analysis processes. The main purpose of the research is to improve the machine's understanding of the semantic structure of the text by finding current connections between the main members of the sentence, current connections between secondary members of the sentence, the best concept of the current word and the function of the current word in the sentence.

Semantic networks are used in the software solution. The Java programming shell, such as NetBeans IDE 8.1, and the CLIPS shell, were used to create the software product. The main logical connections and structure of the program are described in the article.

Methods and relations are considered on the example of the Germanic group of languages. All languages of the Germanic group are similar because they have a direct line of words, which makes them even more similar: subject + predicate + subordinate clauses.

Thus, to reflect the structure of the Germanic group of languages, it is sufficient to consider one of them. Namely, English, as it is the most common (1.5 billion people), international, has

DOI: 10.37943/AITU.2020.75.91.002



the largest vocabulary among the group (500 thousand words) and, in our opinion, the most complex.

**Key words:** computer linguistics, semantics, text, method, expert system, machine text analysis.

**Introduction**

Language is a system of sound and graphic signs, which arose at a certain level of human development, is developing and has a social purpose; language rules normalize the use of signs and their functioning as a means of human communication [1].

There are human, formal, and animal language. From the point of view of the science of linguistics, the main methods of language research are descriptive, comparative-historical, comparative and structural. Research methods are also used to study two "slices" of language: diachrony and synchrony. Linguists have established the kinship of languages in cases where linguistic unity has disintegrated no more than 5-10 thousand years. They were united into language families. Some researchers have tried to establish a more remote genetic relationship of languages [2].

We will verify the research data for the Germanic language group – a group of related languages of the Indo-European language family. More than 550 million people use the languages of the German group.The most common of these are English, German and Dutch. Over the last 300 years, German has become international and is now the state language in more than 70 countries, including Dutch in five, German in six, and English in fifty-four [3].

A typical sequence of actions for the text analysis process is presented by reflecting the change in the consciousness of the recipient depending on the gradual receipt of the text and the impact of experience on the analysis of information. The use of computer technology helps to speed up the text analysis process, but requires prior analysis and automation [4, 5].

**Analysis of literary tribute and problem statement**

Interesting for the study are text posts in both news outlets and social networks.

Information in the form of text – a fairly common form in which the researcher-sociologist receives data for analysis. The text is one of the most "saturated" forms of source data, as it has no shortcomings of artificially created and formalized "reality" of questionnaires with a limited choice of answers. At the same time, textual data is one of the most inconvenient to analyze, as it requires a lot of time when it comes to representative research [6]. Therefore, the question arises about the use of IT technologies in the semantic analysis of the text.

The idea of automating the systematic analysis of texts arose long ago in Western Europe and the United States [7]. Harvard General Inquirer, the first widely used automated content analysis program, still works, rewritten from the original IBM PL/1 language in Java [7].

In describes the implementation of two algorithmic approaches for modeling relationships coreferent text. Machine learning of the system for determining and analyzing relationships using the method of maximum entropy and using the method of reference vectors allows experiments with marked text corpora. These methods have shown the high accuracy of the co-reference analysis system. The reference vectors method demonstrates higher estimates of accuracy, completeness and, accordingly, F-measures than the maximum entropy method. However, it should be noted that learning the method of reference vectors requires much more time [8].

The use of graphic logical form as a semantic representation of text comprehension is described in [9]. This bridges the gap between highly expressive «deep» notions of logical forms: fine semantic encodings such as word-sense and semantic relations.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

The authors determine the evaluation indicator and use it to evaluate the efficiency of the TRIPS analyzer on the general task points [9].

Logical representation of semantic rules of analysis allows you to build algorithms for contextual intelligence: from text to actual data. That is, extract the value from the unstructured text and put it in context using a simple API. These principles are used in expert analytical systems [10].

Today, there are online tools that allow you to partially analyze text messages. One such open source is a system using Dandelion API technology. The system works even on short and incorrect texts in English, French, German, Italian, Spanish and Portuguese [11].

As you can see, there are many techniques for semantic analysis of the text, but they are still not enough. This is due to the problems of the ontology of research in engineering and the complexity of the analysis of text notifications. Text notifications convey characteristics only by the specifics of phrases, and audio information also contains emotional reflection.

**The purpose and objectives of the study**

The study aims to improve methods of semantic analysis by usage of information technology.

To achieve this goal, the following tasks were set:

- analysis of the main methods of semantic analysis of the text;
- creation of an expert system of semantic analysis of the text;
- verification of data on the example of the analysis of the German language group.

**Review of methods of semantic analysis of the text.**

In the field of artificial intelligence, work is being done to create knowledge representation languages, that is, computer languages focused on the organization of descriptions of objects and ideas. The main criteria for the presentation of knowledge are logical adequacy, heuristic power and naturalness of notation.

Synchronous study involves the analysis of linguistic phenomena in one time of language development: at the present stage or in a certain historical period.

At diachronic or different time studying, it is supposed to trace all way that passed a certain structural element of language (sound, word, and sentence).

The essence of the descriptive method is to inventory and systematize language units. This method has practical significance: it connects linguistics with social needs. Descriptive grammars of different languages, explanatory, orthographic, orthoepic and other normative dictionaries created on its basis.

The task of the comparative-historical method is to reveal the laws and laws according to which related languages developed in the past. Historical and comparative-historical grammars of languages and etymological dictionaries created on its basis.

The structural method is used in the study of the structure of language, and its purpose is to know the language as a holistic functional structure, the elements of which are correlated by a strict system of connections and relations.

The structural method is implemented in four methods of linguistic analysis: distributive analysis; analysis by direct components; transformational analysis; component analysis.

Linguists have established the kinship of languages when linguistic unity disintegrated no more than 5,000 to 10,000 years ago and have united them into language families.

Comparative analysis of these methods is performed on the principle of division into two groups:

- 1 Empirical systems (or databases) – these include Example-Based Machine Translation, Statistics-Based Machine Translati.

DOI: 10.37943/AITU.2020.75.91.002



- 2 Heuristic systems (or Rule-Based Machine Translation) – these include: direct computer translation, transfer, interlingua.

The analysis shows (Table 1) that the Data-based system is inferior to the RBMT system due to the need for large calculations that require powerful hardware, ease of use and literal translation.

The RBMT system, unlike Data-based, does not require powerful hardware, provides acceptable quality of the overall content of the translated fragment. Using the ability to connect external dictionaries expands the potential of the program and especially when working with special vocabulary.

The RBMT system is able to activate an improved version of the translation, taking into account the adjustments made to the original version.

Table 1. Comparative analysis of methods of semantic analysis of the text

| Name of the system | Advantages | Shortcomings |
|---|---|---|
| Data-Based system | 1) the presence of a three-dimensional body, gives an improving and accelerating operation of the program and does not require additional actions;<br>2) good quality of translation of the text of a certain subject;<br>3) Translation resembles the work of a human translator. | 1) strong attachment to the body;<br>2) the lack of an equivalent in the body limits the ability to make changes and improve the quality of the translated fragment;<br>3) it is not possible to predict the final result of the translation;<br>4) lack of work when translating "according to the rules";<br>5) requires powerful hardware. |
| RBMT system | 1) the ability to make changes to the original text, therefore, improves the quality of the translated text;<br>2) does not require powerful software;<br>3) grammatical rules are used in translation;<br>4) predictability of the translation result;<br>5) Acceptable quality of translation of general texts. | 1) high requirements for special knowledge from the average user;<br>2) requirements for large investments by developers;<br>3) an excess of literalism. |

Because the computer cannot understand the state of things in the world as a human being, it needs to present all the information in a formal way. Thus, ontologies are a kind of model of the world around them, and their structure is such that they can be easily machined and analyzed.

**Software implementation of expert system of semantic analysis of the text**. A typical sequence of actions for the text analysis process is presented on Fig. 1, 2 using a block diagram of the activity. They reflect the changes in the consciousness of the recipient depending on the gradual receipt of the text and the impact of experience on the analysis of information.

When designing the diagram of use (Fig. 3) we define actors (actors), and then – actions of actors.

Actors in the system:

1) expert linguist:

- analyzes the current content;
- change the current word connections;
- change the current meaning of the word;
- change the current word function;
- editing the knowledge base of the expert system.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

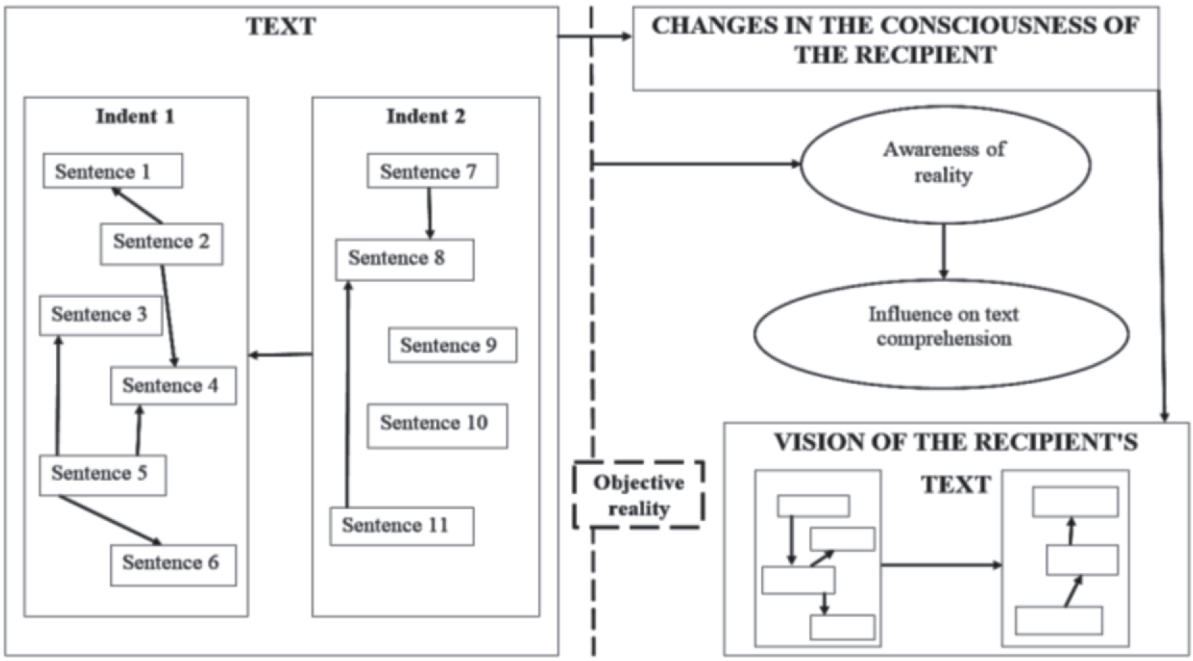


Fig. 1. The scheme of perception of the text

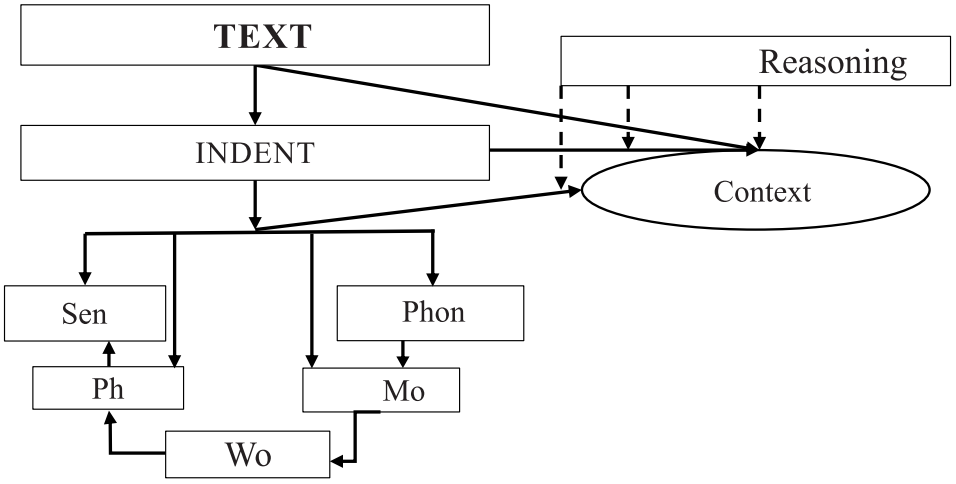


Fig. 2. The scheme of finding the content

2) users:
- analyzes the current content;
- change the current word connections;
- change the current meaning of the word;
- change the current word function.

The block diagram of the interaction of the expert linguist and users with the system is presented in Fig. 3.

DOI: 10.37943/AITU.2020.75.91.002



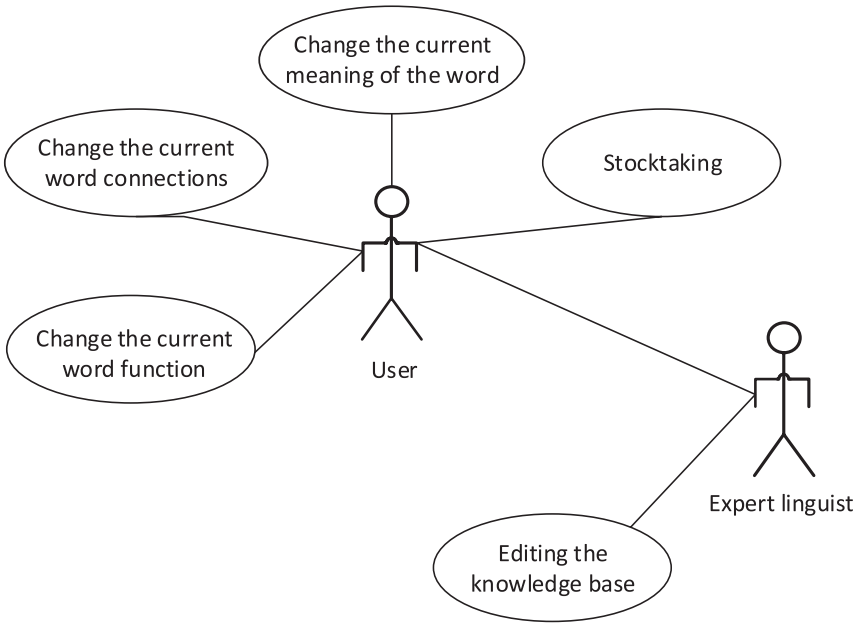


Fig. 3. Block diagram of the interaction of actors with the expert system

The main purpose of the set of tasks is to improve the machine's understanding of the semantic structure of the text. This makes machine translation better and satisfies the guarantee of content transfer.

To achieve this goal to solve the following tasks:

1) finding current connections between the main members of the sentence;

2) finding current connections between subordinate clauses;

3) finding the best concept of the current word;

4) finding the function that performs the current word in the sentence.

The main goals of the program are:

1) creating a system of functional dependencies on parts of speech to reflect the relationships between words;

2) providing the possibility of self-learning system to create a semantic network between specific concepts;

3) providing the ability to add and delete information about the relationships between concepts in the knowledge base.

The initial data is processed by entering text in the workspace, which describes the content of the text (data by words and data by syntactic constructions). Consider the data structure contained in the workspace.

Data by words:

1) word form;

2) identification part-of-speech tag of the word.

Data on the structure of the text:

1) identification syntactic tag of the punctuation mark.

2) After reading the input data, the expert system performs all the necessary operations for semantic analysis of the text.

The source data is a list of phrases, ie data that contains information about semantic connections in the text. The list of phrases is represented by the following data:

1) the main word and its function in the sentence;

2) questions from the main word to the dependent;

3) dependent word and its function in the sentence.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

Output data format:

| |
|---|
| [sentence № open ] |
| {part № open} |
| (word <function> – question? – word <function> ) |
| ……. |
| (word <function> – question? – word <function> ) |
| {part close} |
| [sentence close] |

Fig. 4-12 describe the structure of the data fields of the expert system of semantic analysis of the text.

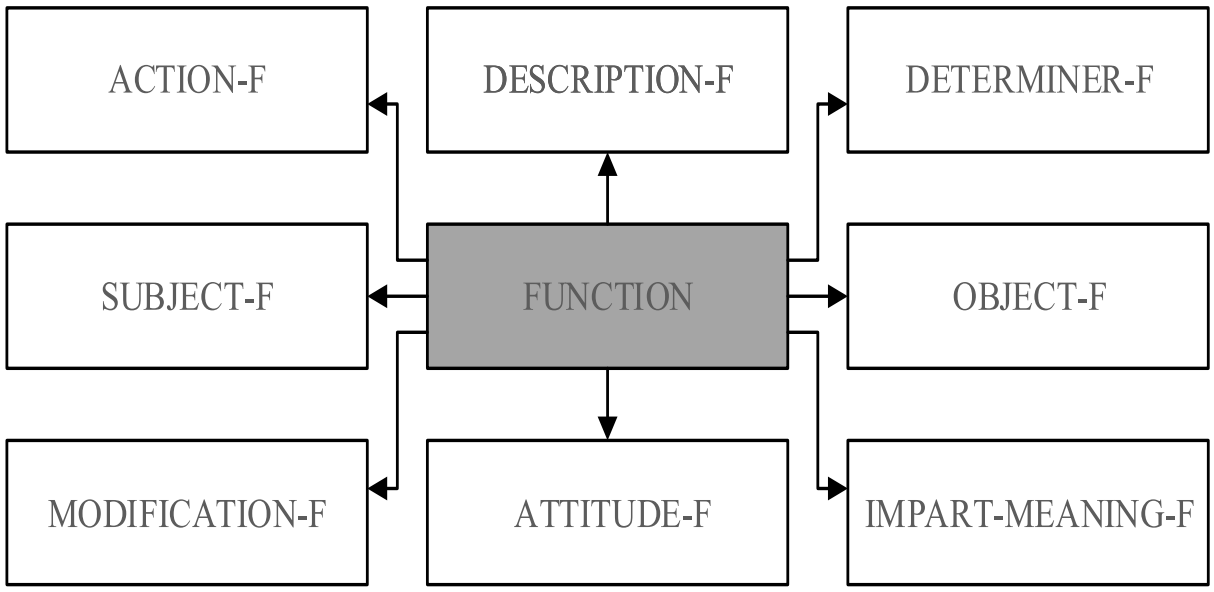


Fig. 4. The structure of abstract classes “Function”

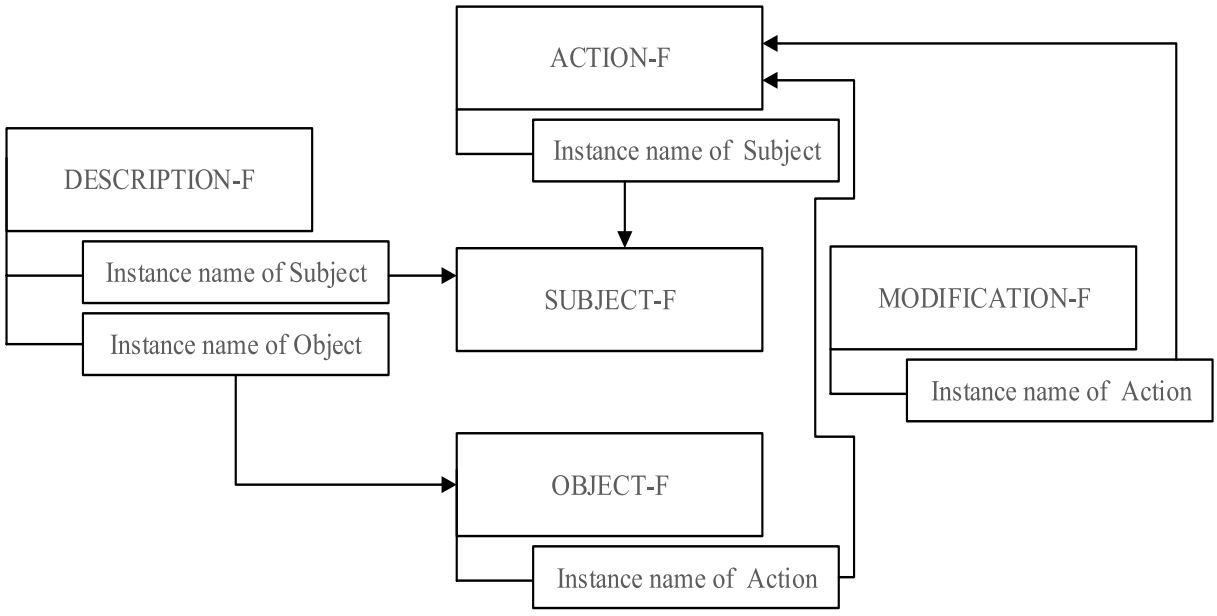


Fig. 5. Fields and relationships between abstract subclasses of «Function»

DOI: 10.37943/AITU.2020.75.91.002



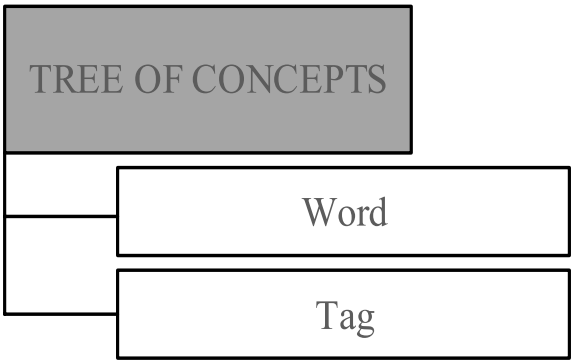


Fig. 6. Fields of the abstract superclass «Tree Of Concepts»

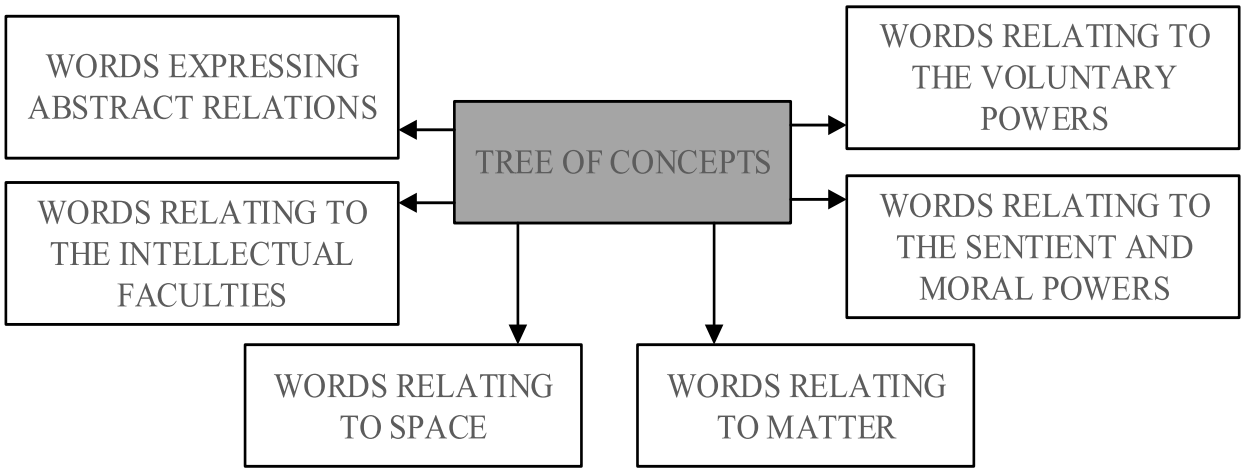


Fig. 7. The structure of abstract classes «Tree Of Concepts»

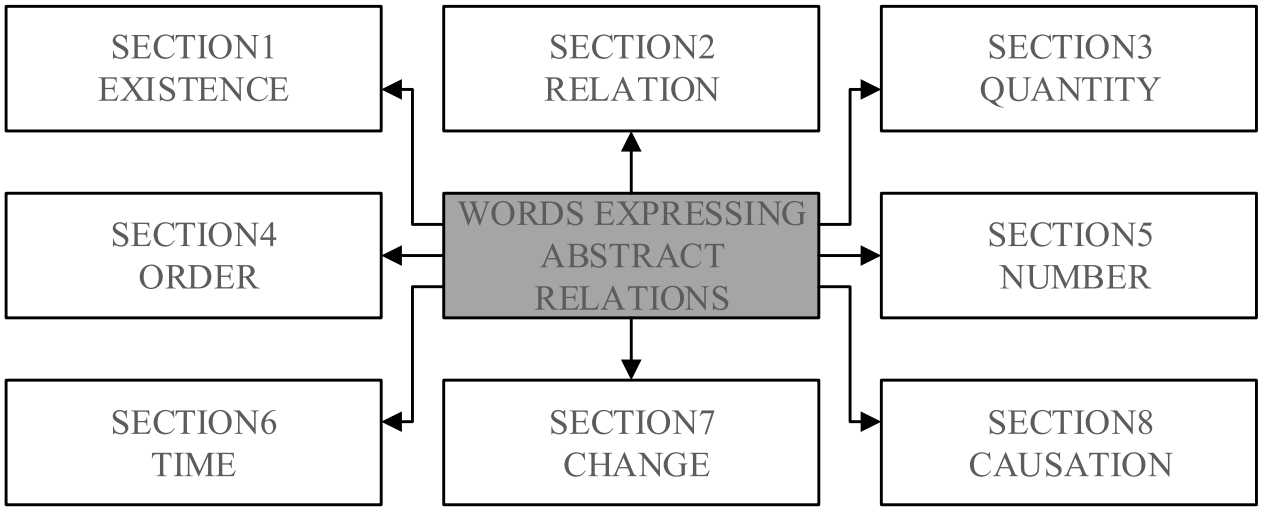


Fig 8. The structure of abstract subclasses «Words Expressing Absract Realtions»

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

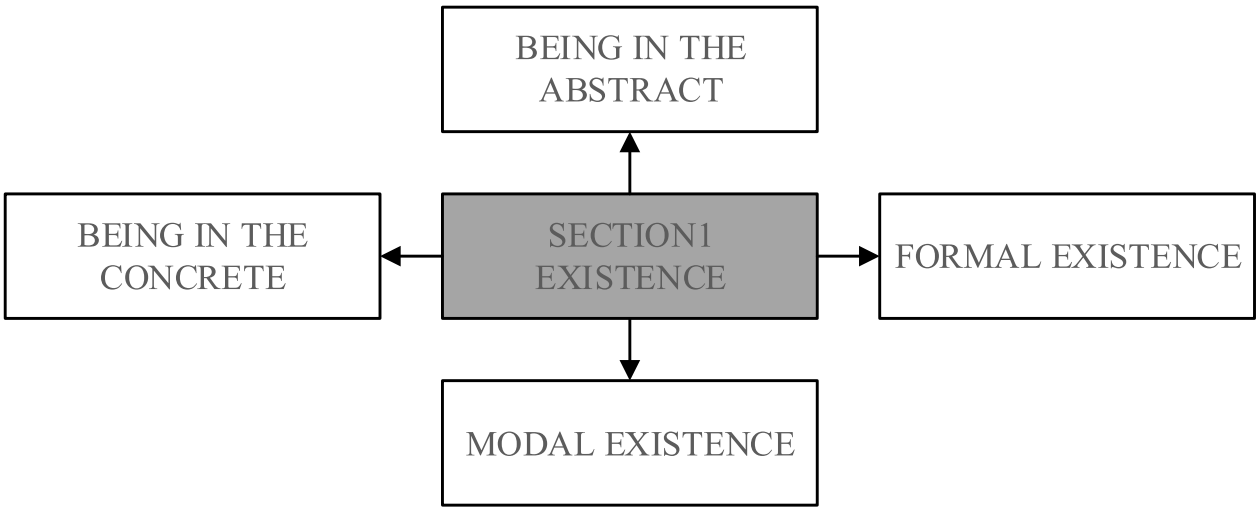


Fig. 9. The structure of abstract subclasses «Section1 Existence»


Fig 10. The structure of abstract subclasses «Being In The Abstract»

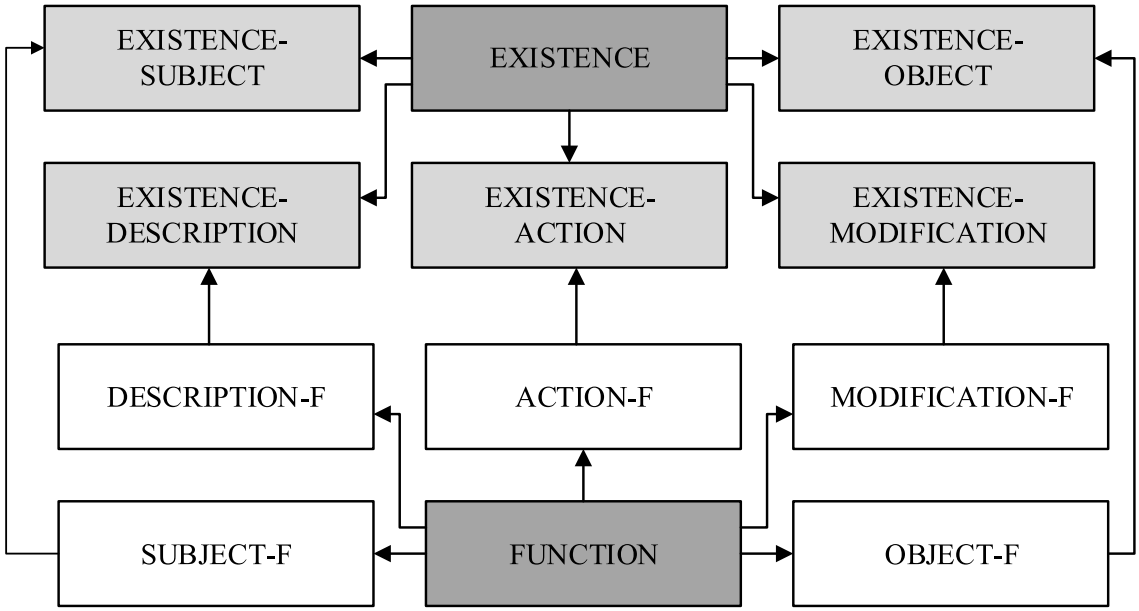


Fig 11. The structure of specific classes for storing information

DOI: 10.37943/AITU.2020.75.91.002



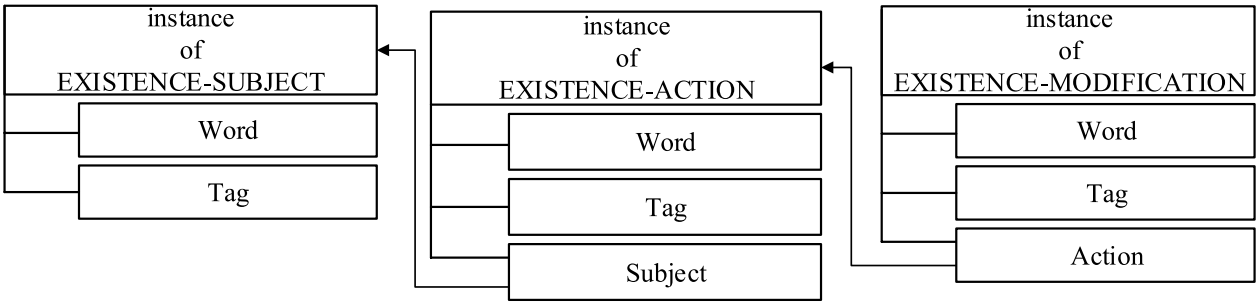


Fig 12. Field structure and relationships between class instances

Fig. 13 shows a diagram of a logical model that reflects the main links in the expert system:
- fact grouping templates;
- facts that are in the system;
- rules for solving problems;
- storage of accumulated information.

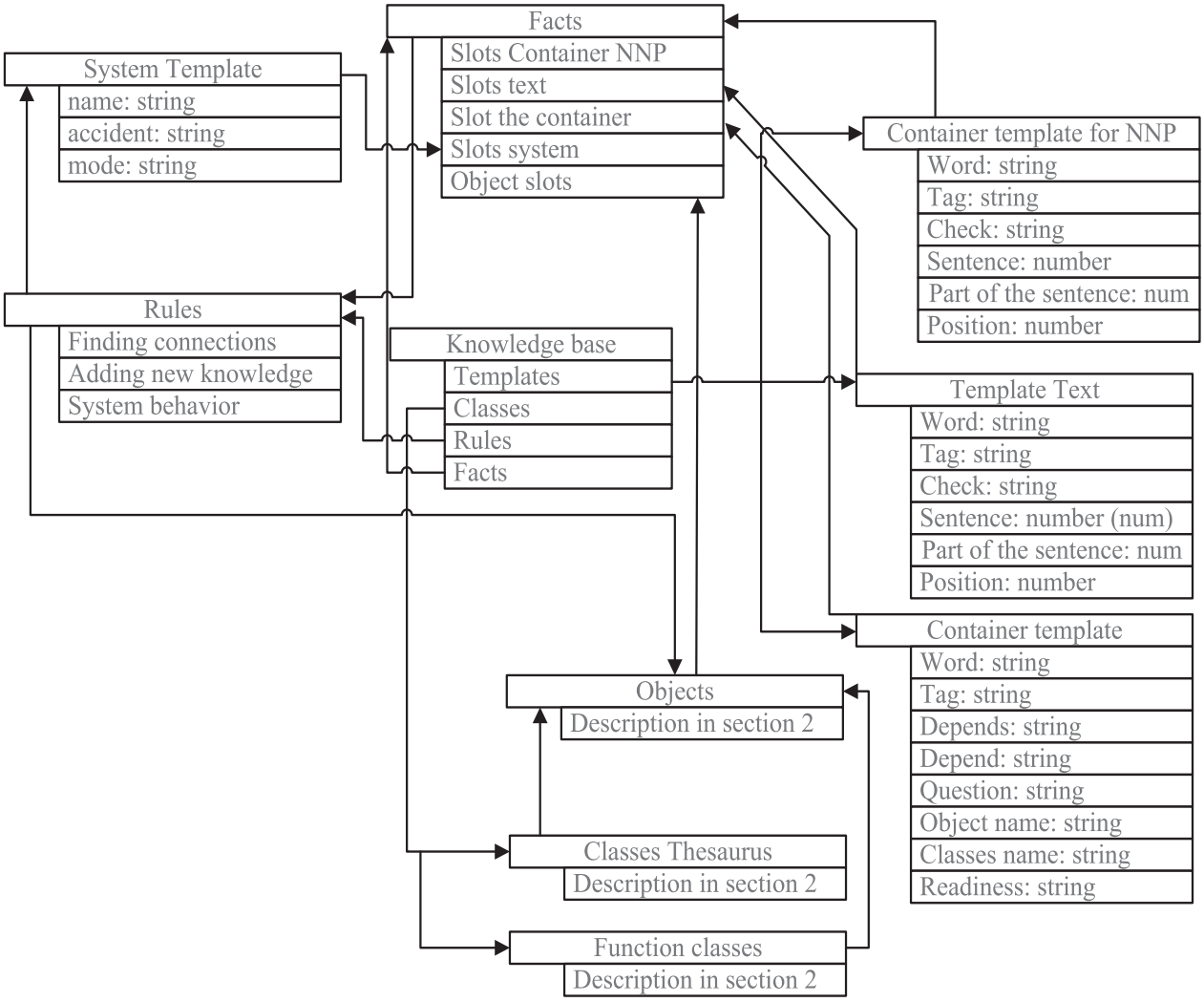


Fig. 13. Logical model of the expert system

Fig. 14 shows a diagram of the structure of the program which shows the components used in the set of tasks, and the relationships between them.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

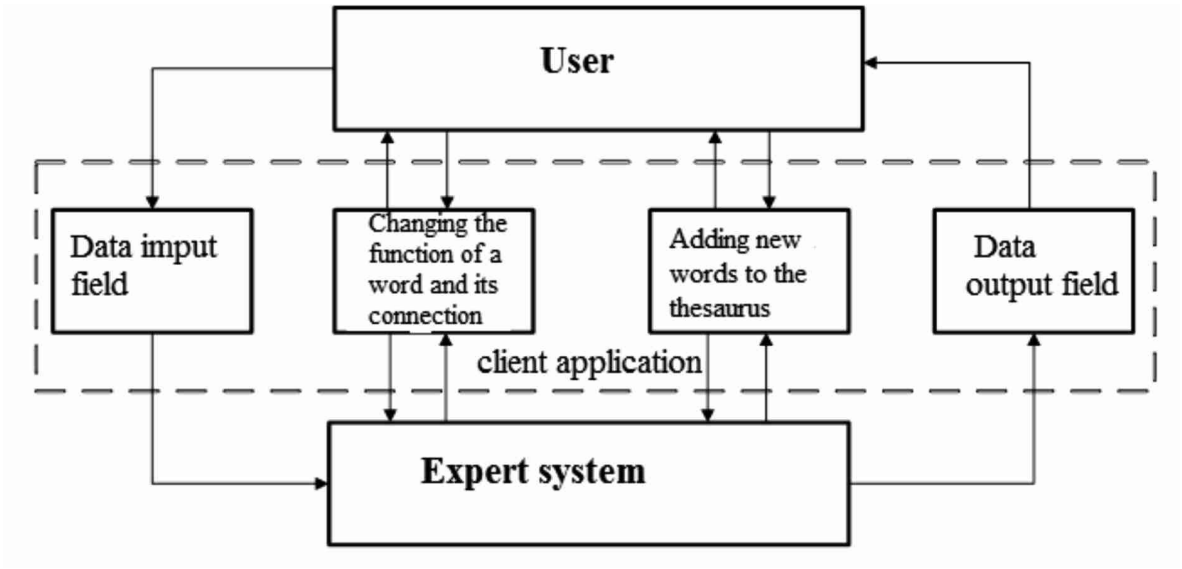


Fig 14. The structure of the expert system

The main components in the system are the user, the client application and the expert system on CLIPS.

**Research results**

Methods and relations are considered on the example of the Germanic group of languages. Germanic languages are a group of related languages of the Indo-European language family. The languages of the German group are used by more than 550 million people. The most common of these are English, German and Dutch. Over the last 300 years, German has become international and is now the state language in more than 70 countries, including Dutch in five, German in six, and English in fifty-four [12].

Consider some of the characteristics of this group:

1) a characteristic tendency to analysis;

2) the opposition of the general and genitive case, so his relationship is expressed mainly in the order of words and prepositional constructions;

3) the system of definite and indefinite article;

4) a wide system of time forms;

5) two-member category of collateral (asset-liability).

The structure of a simple sentence is characterized by a tendency to fix the order of words (Tables 2-4), especially verbs-predicates (solid word order in English, frame construction in German and Dutch).

Inversion is observed in interrogative, motivational and adjunct sentences. There are patterns of word placement in adverbial sentences (especially in conjunctionless conditional).

All languages of the Germanic group are similar because they have a direct word order, which makes them even more similar: subject + predicate + subordinate clauses.

DOI: 10.37943/AITU.2020.75.91.002



Table 2. Word order in an English sentence

| Circumstance time and place | Subject | Predicate | Addition | | | Circumstance | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Indirect | Direct | Prepositional | Reasons | Image of action | Places | Time |
| | I | Wrote | my client | a letter | | | | | |
| | He | Writes | | a letter | to my father. | | | | |
| | I | Write | my client | a letter | | | | in New York | tomorrow |
| Tomorrow | I | Write | my client | a letter | | because of the price | with pleasure | | |

Table 3. Word order in a German sentence

| Circumstance time | Subject | Predicate | Subject | Circumstance time | Inderect additin | Circumstance | | Addition | | Circumstance place |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Reasons | Image of action | Direct | Prepositional | |
| | Ich | schreibe | | | meinem Klienten | | | einen Brief | | |
| | Er | schreibt | | | | | | einen Brief | an seinen Vater | |
| | Ich | schreibe | | Morgen | meinem Klienten | | | einen Brief | | Nach New York |
| Morgen | | schreibe | Ich | | manem Klienten | wegen des Preises | geme | einen Brief | | |

All languages of the Germanic group are similar because they have a direct line of words, which makes them even more similar: subject + predicate + subordinate clauses.

Table 4. Word order in a Dutch sentence

| Circumstance time | Subject | Predicate | Subject | Circumstance time | Inderect addition | Circumstance | | Addition | | Circumstance place |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Reasons | Image of action | Direct | Prepositional | |
| | Ik | schrijf | | | een client | | | een brief | | |
| | Zij | schrijft | | | | | | een brief | naar zijn vader. | |
| | Ik | schrijf | | morgen | een client | | | een brief | | Naar New York |
| Morgen | | schrijf | ik | | een client | vanwege de prijs | graag | een brief | | |

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

From Tables 2-4 we select the following patterns:

- the order of words in the sentence is fixed;
- circumstances may change their location depending on the language (in English, the circumstances of place and time may be in the first or last place in the sentence, in German and Dutch – in the first or third place);
- the circumstance of the place is always in the last place.

**Discussion of results**

For a clearer indication of the similarity of word order in these languages, we present the table (Table 5).

As can be seen from the table (Table 5), in all languages in the first and second place are the main members of the sentence: subject and predicate, followed by secondary: addition and circumstance. Changing the order of words in a sentence can completely change its meaning.

Table 5. Similarity of word order in the sentences of the Germanic language group

| Language | Subject | Predicate | Addition | | Circumstance |
|---|---|---|---|---|---|
| | | | Indirect | Direct | |
| English | I | write | my client | a letter | in New York. |
| German | Ich | schreibe | meinem Klienten | einen Brief | Nach New York |
| Dutch | Ik | schrijf | een cliënt | een brief | Naar New York. |

The analysis of the structures of the Germanic language group, the number of temporal forms, cases and more is performed. Thus, to reflect the structure of the Germanic group of languages, it is sufficient to consider one of them. Namely, English, as it is the most common (1.5 billion people), international, has the largest vocabulary among the group (500 thousand words) and, in our opinion, the most complex.

**Conclusion**

Thus, we have analyzed the methods of semantic analysis of the text. It is established that the Data-based system is inferior to the RBMT system due to the need for large calculations. The RBMT system is able to activate an improved version of the translation, taking into account the adjustments made to the original version.

Changes in the consciousness of the recipient depending on the gradual receipt of the text and the impact of experience on the analysis of information are reflected.

The description of the data field structure of the expert system of semantic analysis of the text, its logical structure and the structure of the program are given.

Methods and relations were considered on the example of the Germanic group of languages. It is established that to reflect the structure of the Germanic group of languages it is enough to consider one of them, namely English. It is the most common.

In the future, it is planned to apply the developed expert system of text analysis for posts on social networks. This will help to conduct sociological research about notifications.

DOI: 10.37943/AITU.2020.75.91.002



## References

1. Kocherhan, M.P. (2020, June 6). Vstup do movoznavstva. Resource access mode: https://pidruchniki.com/1222090548043/dokumentoznavstvo/vstup_do_movoznavstva(in Ukrainian).
2. Karpilovs'ka, A.E. (2006). Vstup do prykladnoyi linhvistyky: komp'yuterna linhvistyka. – Donets'k: Yuho-Vostok, 187. (in Ukrainian)
3. Kenzhaev, A.D. (2020, June 6) Machine translation: history and modernity. Resource access mode: https://lomonosov-msu.ru/archive/Lomonosov_2014/2568/2200_72719_187154.pdf. (in Russian).
4. Meyye, A. (2016). Osnovnyye osobennosti germanskoy gruppy yazykov.Per. s fr. Izd. Stereotip.URSS, 168.
5. Slovari i sistemy mashinnogo perevoda (2020). Resource access mode: http//www.itland.com.ua/products/sect.php.section.
6. Ivanov, O. V. (2009). Komp'yuternyy kontent-analiz: problemy ta perspektyvy vyrishennya. Metodolohiya, teoriya ta praktyka sotsiolohichnoho analizu suchasnoho suspil'stva, 15, 335-340.
7. Monroe, B.L., & Schrodt, P. A. (2008) Introduction to the Special Issue: The Statistical Analysis of Political Text. Political Analysis, 16, 351-355.
8. Marchenko, O.O. (2015). Systema analizu koreferentnykh zv'yazkiv u tekstakh// Shtuchnyy intelekt, №3-4 [Electronic resource]: Resource access mode: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/117200/01/Marchenko.pdf?sequence
9. Deep Semantic Analysis of Text James F. Allen1,2 Mary Swift1 Will de Beaumont [Electronic resource]: Resource access mode: https://www.aclweb.org/anthology/W08-2227.pdf (accessed 01.06.2020)
10. Klapur, A. (2007) Semantic analy p sis of text and speech [Electronic resource]: Resource access mode: https://www.cs.tut.fi/sgn/ arg/klap/introduction-semantics.pdf (accessed 01.06.2020)
11. Dandelion API. Semantic Text Analytics as a service. [Electronic resource]: Resource access mode: https://dandelion.eu/. (accessed 01.06.2020)
12. Nikolayeva, S.Yu. (2018) Zmist fakhovoho vyprobuvannya do aspirantury zi spetsial'nosti 011 osvitni/pedahohichni nauky dlya spetsializatsiyi "Teoriya ta metodyka navchannya: hermans'ki/romans'ki movy"// International Scientific and Practical Conference World Science. И-во: ROST (Dubai), 4(30), 52-59.