24. Razaq, A. A big data analytics based approach to anomaly detection [Text] / A. Razaq, H. Tianfield, P. Barrie // Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies – BDCAT '16, 2016. – P. 187–193. doi: 10.1145/3006299.3006317
25. Perlovsky, L. Dynamic Logic Machine Learning for Cybersecurity [Text] / L. Perlovsky, O. Shevchenko // Advances in Information Security. – 2014. – p. 85–98. doi: 10.1007/978-3-319-10374-7_6

Розроблено методи автоматичного аналізу тексту на основі декларативного представлення правил синтаксичної сполучуваності та програмного розподілення аналітико-синтетичної обробки природно-мовного тексту в системах машинного перекладу. Програмна реалізація експерементально доводить, що застосування розроблених методів зменшує кількість помилок семантичного характеру в середньому на 14–16 % у порівнянні з відомими системами машинного перекладу

Ключові слова: система машиного перекладу, автоматичний аналіз тексту, аналітико-синтетична обробка тексту

Разработаны методы автоматического анализа текста на основе декларативного представления правил синтаксической соединяемости и программного распределения аналитико-синтетической обработки естественно-языкового текста в системах машинного перевода. Програмная реализация експерементально подтверждает, что применение разработанных методов уменьшает количество ошибок семантического характера в среднем на 14–16 % по сравнению с известными системами машинного перевода

Ключевые слова: система машинного перевода, автоматический анализ текста, аналитико-синтетической обработка текста

UDC 519.76:81'22
DOI: 10.15587/1729-4061.2017.92021

DEVELOPMENT OF KNOWLEDGE-ORIENTED SYSTEM OF MACHINE TRANSLATION BASED ON THE ANALYTIC-SYNTHETIC TEXT PROCESSING

L. Lytvynenko
Postgraduate student*
E-mail: l.lytvynenko@gmail.com
O. Nikolaievskyi
Postgraduate student*
E-mail: a1.n1@yandex.ru
V. Lakhno
Doctor of Technical Science, Associate Professor**
E-mail: lva964@gmail.com
E. Skliarenko
PhD, Associate Professor*
E-mail: sigma.inet@gmail.com
*Department of Information Systems and Mathematical Sciences***
Department of Managing Information Security*
***European University
Academika Vernadskogo blvd., 16 V,
Kyiv, Ukraine, 03115

1. Introduction

A constant growth of the volume of text information (TI), associated with the use of the Internet, leads to an increase in the need for automatic text processing of TI. The quality requirements for processing, primarily based on the use of modern information technologies, are at the forefront. Unfortunately, high quality software in the tasks of synthetic-analytical processing of multilingual text information in machine translation systems (MTS) exists only for narrow subject areas and cannot be easily adapted to a wide range of tasks. In addition, existing solutions mostly require post-editing and are oriented to professional translators, rather than ordinary users.

The relevance of present work is in the study of method of automatic syntactic analysis (ASA) of the text based on declarative representation of the rules of syntax combinability and on the method of software distribution of analytical-synthetic processing of the natural language text (NLT) at MTS.

2. Literature review and problem statement

As shown by the analysis of theoretical and practical work in the field of MTS development, a lifetime problem of automatic translation is polysemy and uncertainty, the solution to which involves computer modeling of the process of understanding NLT, particularly evident for the Slavic languages due to rich morphology [1].

Today, three complex models for building formal semantics of NLT are known [2–5].

Model [2] was developed at Stanford University (United States) and has the title "semantics of advantages"; it is

development of research by the Cambridge linguistic school. This model works for texts as a whole on the logical-semantic principles and does not imply the use of morphological or syntactic analysis of text, which actually predetermines its purely theoretical nature.

Studies of the experimental systems based on the model of "semantics of advantages" are presented in articles [3–5]. But similar studies are not aimed at creating applied software for MTS and the studies are limited to the needs and goals of particular customer.

The work on studying the types of conceptual structures and their language correspondences started within the framework of exploring artificial intelligence [6, 7]. The result of this work was the model of "conceptual dependences". The model has not been implemented in software so far.

Model "Sense ⇔ text" (MST) [8] represents language as a multilevel model of transforming semantic content into the text and vice versa. The model focuses on computer processing of texts, however, it is most developed for the Russian language. A number of systems were implemented based on MTS, including the machine translation system "ETAP" [9]. A common disadvantage of these systems and MTS as a whole is that they do not take into account the regularities of describing knowledge at the sign level of the text representation. This, in turn, leads to serious errors of the text semantic analysis. In addition, pragmatic analysis is conceptually implemented only as knowledge about the world (subject area).

Modeling of the understanding process in the "knowledge-oriented" model [10, 11] implies development of the means of recognition and formalization of knowledge, contained in the source text, the interpretation of this knowledge regarding a specific task (target orientation), and the synthesis of results of knowledge interpretation in the resulting NLT.

A pragmatic analysis of the existing models [12–15] comes down, as a rule, to the means of recognition of knowledge of the subject area [16, 17], while the knowledge about certain applied task is not considered, and this is, in fact, the determining factor when the NLT is analyzed by a specialist [18, 19].

According to authors [20, 21], efficient methods of automatic text analysis and quality linguistic support to MTS is applied only in the narrow subject areas.

As authors of study [22] believe, the results of most DSS, based on the existing models of automatic text analysis require correction by a qualified interpreter. This makes auch DSS inconvenient for ordinary user.

The main disadvantages of ASA methods for MTS, according to authors [23, 24], include the need for post-editing because a word-for-word translation distorts the text content, and pragmatic features of the text are not taken into account.

Numerous studies and publications [16, 18, 25], associated with the development of MTS, indicate that there is a need to develop and research into new methods of ASA of text based on the declarative representation of rules of syntax combinability and the software distribution of analytical-synthetic processing of NLT in PS of machine translation.

3. The aim and tasks of the study

The aim of present work is to improve the quality of translation in MTS through the development of methods for analytical-synthetic processing of multilingual text information based on the knowledge-oriented approach.

To achieve the aim, the following tasks were to be solved:

– to develop a method for the formalization, extraction and analytical-synthetic processing of the knowledge contained in NLT, at the level of semiotic system of text representation;

– to develop a method for the analytical-synthetic processing of knowledge contained in NLT, at the level of linguistic system of text representation;

– to develop software for the implementation of proposed methods in a knowledge-oriented machine translation system and experimentally explore effectiveness of the developed software for an array of texts on military-technical subjects in the developed SMP.

4. Method for the analytical-synthetic processing of knowledge contained in natural language text at the sign level of its representation

The proposed approach to developing methods for the recognition, extraction and formalization of knowledge contained in NLT is based on the following conceptual provisions:

– a source natural language text is semantically and logically coherent;

– a text is a representation of three interrelated systems: semiotic system, linguistic system and the system of knowledge about the world (subject area), the text coherence is provided by graphical means of text organization, linguistic and extralinguistic means;

– all these means in the text are the encoding tool for the knowledge about the world (subject area); the objects that have the correspondent lexical equivalents of concepts, relations and characteristics of the concepts and relationships are regarded as the elements of the real or abstract world.

Accordingly, a linguistic processor should provide for the processing of NLT at all levels of text organization:

– semiotic (sign);

– at the level of language organization (this level includes morphological, syntactic and semantic levels of processing the source NLT);

– pragmatic, that is, at the level of reflection of knowledge about the world in the source NLT.

The core of linguistic processor (LP) is the algorithms of processing NLT, which include algorithms for accessing and processing a linguistic database (LDB) and a knowledge base (KB) on SA. Traditionally, LP in the systems for automatic processing of textual information involves two blocks: analysis and synthesis. Working with knowledge on SA, contained in NLT, requires advanced means of interpretation of language units in terms of knowledge itself (concepts) in both SA and regarding the target orientation on the applied problem (knowledge of applied problem). This led to the development of LP, which would include three separate blocks: analysis, interpretation and synthesis. Structural-logical schematic of linguistic processor, which reflects the essence of NTL processing, is shown in Fig. 1.

According to Fig. 1, we will consider the structure of grapheme level of analysis and synthesis of the document (stages are highlighted in red, dictionaries – in blue). At the level of line analysis, the symbol of the end of the line (Enter) serves as the main fragment identifier. After this stage, we have selected text fragments and may proceed to the analysis of lexemes.

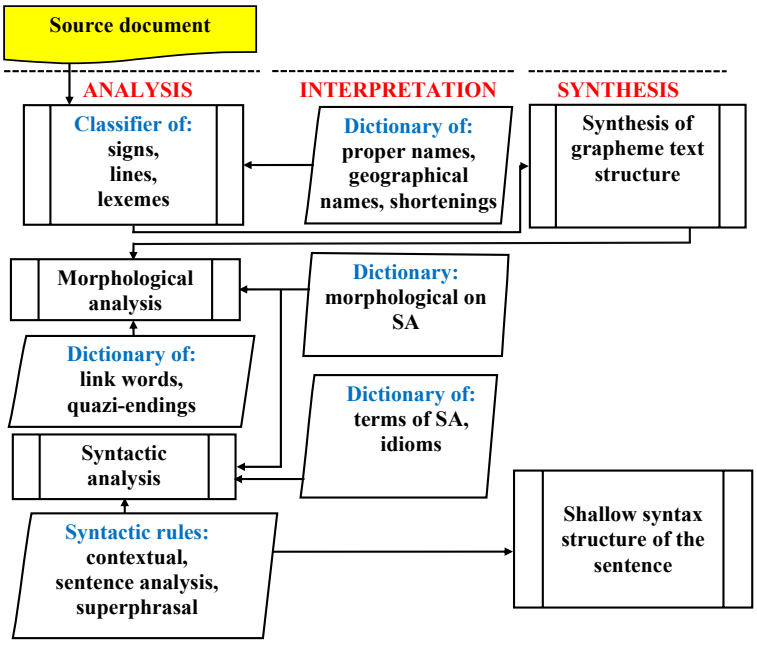

Fig. 1. Structural-logical schematic of linguistic processor for the knowledge-oriented machine translation system

We will introduce the following concepts:

Grapheme (G) := <Symbol ∪ Space (Sp)>;

Symbol (S):= <Letter ∪ Figure ∪ Special character ∪ Syntactic sign ∪ Brackets ∪ Mathematical symbol >;

Letter (L) := <Latin alphabet ∪ Cyrillic alphabet ∪ ‘>;

Figure (F) := <1 ∪ 2 ∪ 3 ∪ 4 ∪ 5 ∪ 6 ∪ 7 ∪ 8 ∪ 9 ∪ 0>;

Syntactic sign (SS) := <, ∪ . ∪ : ∪ ; ∪ “ ∪ - ∪ ? ∪ ! >;

Latin alphabet (L) := <Latin alphabet capital (LC) ∪ Latin alphabet low case (LL)>;

Cyrillic (C) := < Cyrillic capital (CC) ∪ Cyrillic low case (CL)>;

Special character (SC) := <№ ∪ % ∪ / ∪ @ ∪ # ∪ $ ∪ & ∪ * ∪ \>;

Brackets:= <[∪] ∪ (∪) ∪ { ∪ }>;

Mathematical symbol:= <+ ∪ < ∪ > ∪ = ∪ == ∪ ≤ ∪ ≥ ∪ ± ∪ ≠>;

LC := <Consonant (LCC) ∪ Vowel (LCV)>;

LL := <Consonant (LLC) ∪ Vowel (LLV)>;

CC := < Consonant (CCC) ∪ Vowel (CCV) ∪ Ь ∪ Ъ>;

CL := < Consonant (CLC) ∪ Vowel (CLV) ∪ ь ∪ ъ>;

LCC := <Q ∪ W ∪ R ∪ T ∪ P ∪ S ∪ D ∪ F ∪ G ∪ H ∪ J ∪ K ∪ L ∪ Z ∪ X ∪ C ∪ V ∪ B ∪ N ∪ M>;

LCV:= <E ∪ Y ∪ U ∪ I ∪ O ∪ A >;

LLC:= <q ∪ w ∪ r ∪ t ∪ p ∪ s ∪ d ∪ f ∪ g ∪ h ∪ j ∪ k ∪ l ∪ z ∪ x ∪ c ∪ v ∪ b ∪ n ∪ m>;

LLV:= <e ∪ y ∪ u ∪ i ∪ o ∪ a ∪ й ∪ ё >;

CCC:= <Ц ∪ К ∪ Н ∪ Г ∪ Ш ∪ Щ ∪ З ∪ Х ∪ Ф ∪ В ∪ П ∪ Р ∪ Л ∪ Д ∪ Ж ∪ Ч ∪ С ∪ М ∪ Т ∪ Б>;

CCV:= <У ∪ Е ∪ А ∪ О ∪ Я ∪ Ю ∪ Й ∪ Ё>;

CLC:= <ц ∪ к ∪ н ∪ г ∪ ш ∪ щ ∪ з ∪ х ∪ ф ∪ в ∪ п ∪ р ∪ л ∪ д ∪ ж ∪ ч ∪ с ∪ м ∪ т ∪ б>;

CLV:= <у ∪ е ∪ а ∪ о ∪ я ∪ ю >;

The following features make the basis of the classifier of graphemes: the sign type (number, letter, syntactic sign, special character, etc.), belonging to the alphabet (Latin, Cyrillic, including Russian, including Ukrainian), letter size (lowcase, capital), phonetic features (vowel, consonant). By the lexeme we imply word forms or word combinations that have some grapheme representation and perform differential semantic function.

According to Fig. 1, the next step is knowledge immersion in SA, all lexemes L02, which follow the symbols [?!], are checked in the dictionary of proper names, incorporated into linguistic software of the pre-morpheme analysis. If the word is not found there, the class of lexeme (CL) is changed to L01.

This is followed by the analysis of compound lexemes – selection of more general language lexemes on the basis of simple lexemes. The rules of formation of compound lexemes are shown in Table 1.

Table 1

Rules of formation of compound lexemes

Class code (CC)	Name	Rule	Example
Lk1	Words in inverted comas	«[L01-L33]+»	«McDonnel helicopter»
Lk2	Words in brackets	([L01-L33]+)	(MTS)
Lk3	Compound name	L07 L02	ВПО-133 Vepr
Lk3	Compound name	L07 Lk1	ВПО-133 «Vepr-K»
Lk3	Compound name	Lk1 L07	«Hurricane» ВМ-27

There are also rules that cannot be represented in Table 1 and have procedural description. In particular, the rule of definition of the context link: If there is lexeme Lk2, consisting of one lexeme L21, lexemes L02 that follow or precede it are the explanation of the context link.

The next step distinguishes *syntagms*, which are defined as a sequence of lexemes, located between the syntactic punctuation signs. Syntagms are distinguished to facilitate analysis at further levels of analysis, in particular, semantic analysis.

Once lexemes and syntagms have been recognized, the transition to the sentence recognition takes place. By sentence we imply logically and semantically complete statement, included in a specific discourse, organized in accordance with grammar of the source language and having a completed grapheme organization.

Development of rules for the context (sign) environment of a full stop allows distinguishing sentences according to their semantic finality in most cases. In cases where a full stop performs a function of the end of the sentence, it is referred to the lexeme class “KP”. According to the rules, given in Table 2, the following sentence types are distinguished: a language sentence (LS), a title-sentence (TS), service sentence (SS).

Finally, after defining the boundaries of sentences, an analysis of boundaries and types of fragments is run. In case the first LS of one fragment starts with a lowercase letter and the last LS of the previous fragment is a blank line, such fragments are united. The second rule indicates that two fragments are necessary to unite, there is no marker KR at the end of the last LS of the first fragment is formed completely.

Table 2

Rules for determining sentence types

Sentence type	Rule	Example
TS	Text is centered, given in bold and is not more than 15 words long	«Requirements to the MTS organization»
SS	Text alignment is set to the right edge	May, 09, 1945
LS	Neither of the previous rules was applied	The main task when analyzing MTS is the formation of the concept structure of the text

The grapheme organization of the source text takes place at the very last stage, when the text has been translated. The purpose of this procedure is to define the capital and lowercase letter for proper names.

The procedure of transliteration, depending on the class of a lexeme that is being recognized and target orientation of the model of knowledge about the decisive applied task, has three algorithms: direct transliteration, combined transliteration, and preserving the original spelling of a lexical unit.

Direct transliteration is used for such lexical units as proper names of the companies that are given in the text in quotation marks. For example: the English name of the company "*MacDonnal Helecopter*" in Russian will be transliterated as "*Макдоннел гелеkоптер*".

Combined transliteration is used for the proper names that indicate surnames and names. Only the initial form is transliterated, and the required declensions take place with the help of synthetic word-changing dictionary of quasi-endings, which is general for morphological synthesis of the source text. For example, for the full women's name *Hillary Clinton*, the Ukrainian translation in genitive case will be *Хіларі Клінтон*, and for the men's name *Bill Clinton*.

The procedure for preserving the original spelling of lexical unit is used for such lexemes as designation (for example: *BTR-60, AR-670-1*), complex designation (for example: *AN/PRC-77 radio – AN/PRC-77 радіо*. Usually, only the part of a lexical unit, which includes proper designations, abbreviations, has no corresponding context directly in the text and no correspondents in the dictionary, remains unchanged. This procedure may be used for proper names in quotation marks if it is determined by the target orientation in the model of knowledge about applied task. Sometimes, if a customer wants to obtain the information, for example, about a certain company, it is better to run additional search by the original form, since any transliteration distorts the original name.

At the stage of interpretation of pre-morpheme processing of a text, certain classes of lexical units, such as a name, a title, designation, abbreviation, shortenings, etc, are checked on the model of knowledge about the world (SA). The purpose of this stage is to distinguish the classes of lexemes that can coincide with the class of language lexemes by the format of presentation.

Semantic lexeme type does not depend on the grapheme class of a lexeme, so abbreviations and capitalized lexeme may belong to one and the same semantic type.

The following semantic types of lexemes may be distinguished:

1 – *the geographical name*. This class includes spatial data, such as the names of cities, seas, oceans, rivers, lakes, continents, etc. The necessity of introducing the dictionary of geographical names is caused by the fact that these names in the text are given without determining lexemes, since they define encyclopedic knowledge (that is, it is to be known). We did not include the names of the countries into this class, because for our SA these names have a political context, which actually caused their entering another semantic class – political name;

2 – *the historical name*. This class includes well-known names of historic events;

3 – *the name*. The necessity of introduction of this semantic type is caused by the fact that in English texts little-known names are given next to the name. This allows, on the one hand, identifying that this is a person and incorporating two lexemes into one indivisible concept, on the other hand, defining the category of genus for a name allows achieving greater accuracy when translating into Russian or Ukrainian;

4 – *the entity*. This class includes known titles of organizations, institutions, types of armed forces, etc.;

5 – *the unit of measurement*. This class includes shortenings that define measurement units and the names of the months and days of the week for the English language, which in the text are written with capital letters;

6 – *the name, which is not translated*. This class includes the names of organizations, institutions, surnames, street names, etc., which are transmitted by means of another language entirely according to the rules of transliteration;

7 – *the position*. This class includes titles of position that are written with capital letters;

8 – *the political name*. The necessity of introducing this semantic type is caused by the fact that the names of the countries in our context (SA: military-political texts) are considered as geopolitical objects, rather than geographical names;

9 – *non-defined semantic type*. This semantic type is assigned when a lexeme is suitable for none of the enumerated semantic types. If there is a significant number of such lexemes, the classifier of semantic types needs expanding.

The second position of the semantic code defines semantic characteristics of a lexeme in comparison with the world. Thus, the following meanings of semantic types of lexemes are distinguished:

1 – *time*. The characteristic of time determines the lexeme of the corresponding semantic type in time;

2 – *space*. The characteristic of space determines the lexeme of the corresponding semantic type in space;

3 – *time-space*. This characteristic is common for some complex units of measurement (for example: km/h);

4 – *quantity*. The characteristic that refers exclusively to the estimation of quantity;

5 – *object*. The characteristic that defines the specificity (objectness) of a lexeme of the corresponding semantic type;

6 – *person*. The characteristic that defines a person (official, etc.);

7, 8 – reserve characteristics.

9 – *others*.

The characteristic of a lexeme of the corresponding semantic type, which does not match any of the above classes.

Table 3 presents the semantic parameterization of dictionary units, which were revealed when analysing the Russian, English and Ukrainian texts on special military subject matter. By contents, text files contain general reference information (for example, a dictionary of names, a list of units

of measurement, a dictionary of geographical names, etc.), such files reflect generally accepted knowledge about the world and, as a rule, are not accompanied by any explanatory context. For example, the names of the countries, famous cities usually are not accompanied by such lexical determinants, like (city) *Moscow*, (country) *Ukraine*, (President) *Clinton*, etc.

In addition, there are specific lexical units, which are generally accepted in a given subject area (e. g.: *омбр (rus.)* is a separate mechanized brigade). For this purpose, the model sample of texts of the given subject area is analyzed and the database (DB) of the appropriate language is updated.

Table 3

Semantic parameterization of lexemes at the level of text organization as a sign system

Code of semantic class	Examples	Interpretation
12!	*Asia, Africa, Europe, Mediterranean sea, Pacific Ocean*	Geographical name: is characterized by space
21	*World War II, World War I*	Historic event: is characterized by time
22!/21!	*Brest peace*	Historic event: is characterized by time and space
35!	*Taras, Martha, Alexander*	Name: person
45!	*the National Security Council, the Central Intelligence Agency*	Institution
51!	*January, Monday, min.*	Unit of measurement: is characterized by time
52!	*Cm, km, mm*	Unit of measurement: is characterized by space
55!	*омбр (rus.)*	Unit of measurement: structural subdivision
59!	MHz	Unit of measurement: characteristic is not defined
65!	*Supreme Council, Duma*	Proper name that is being transliterated
76!	*President, Supreme Commander-in-Chief*	Post: person

The DB of proper names and shortenings, compiked in this way, in turn, is a component of software for automated pre-morpheme analysis of new NLT.

5. Method for analytical-synthetic processing of text at the syntactic level of language system

Research in the framework of theory of applied linguistics for MTS may be separated into two approaches: building universal Grammar – a tool suitable for working with any language, and building a formal model that best covers the linguistic component L (T) for the NLT in a particular language.

General, formal mathematical models and their software implementation [16] are not able to cover all the complexity and diversity of the group $G_L: = \cup L_i(T)$, the group consisting of concepts $L_i(T)$ for different languages. The application of generative model typically leads to the loss of the proper syntactic representation, or to a combined explosion. In terms of Linguistics, the substantiation of the existence of problems lies in the phenomenon of homonymy and in the length of the links between syntactic units.

Syntactic analysis in the proposed knowledge-oriented approach generally involves 3 stages:

1) context-syntactic analysis (building syntactic compounds);

2) syntactic analysis of simple sentences (building a tree of syntactic subordination);

3) inter-phrase syntactic analysis.

The first stage of the context-syntactic analysis.

The tasks of this stage of ASA in MTS include:

– distinguishing syntactic compounds by grammatical features, obtained at the stage of morphological analysis;

– elimination of morphological homonymy, obtained at the stage of morphological analysis;

– definition of the word combinations that are terms and set concepts in a given subject area.

To solve the first problem, the declaratively assigned sets of the context-syntactic rules (coordination, subordination, parataxis) were developed, which in the case of meeting the initial conditions point out how word combinations are formed within the syntagm (syntagms are defined at the pre-morpheme level of the text analysis [9]). The rules fall into the categories: rules of coordination, subordination and parataxis. The rules are applied from the end of a sentence, and if any rules have already worked, the next word is the main word in the group of the previously applied rule. The format of the rule of coordination, subordination and parataxis are shown in Tables 4–6, respectively.

In Table 4 the following designations are used: 1*, 2*, 23*, 12* – codes of the lexico-grammatical classes that were obtained at the stage of morphological analysis and respectively meaning the noun, adjective, preposition, past participle; sign "+" means the existence of the same values in the respective positions of the lexico-grammatical classes. In the column "conditions" the entry «2*:=¬12*» means that the rule may be applied if the sets with lexical-grammatical class 12* are missing in the word form 2*. This is required in order to avoid false syntactic compounds, such as *"conferred to the officer"* from the above given example of the sentence. The column "rule" determines which procedure to apply. Thus, M1 means that for classes 1* and 2* we apply the procedure for crossing all sets of grammar codes that accompany these lexemes. We will show the work of rules M1 by the example of the Russian sentence: *На стекло стекло варенье.*

Table 4

Syntactic rules of coordination of classes (CL)

What CL	With what CL	gender	number	case	number by order of the main word	syntactic type	conditions	rule	examples
2*	1*	+	+	+	(2)	C1	2*:=¬12*	M1	*state bodies*
23*	1*	any	any	+	(1)	C2	any	M1	*kind of form*
...	...	...	...	...	...	...	...	...	...

After morphological analysis, the sentence enters ASA in the form:

*на 23*004000000/23*006000000/*
*стекло 1*311000001/1*314000000/8*314020000/*
*стекло 1*311000001/1*314000000/8*314020000/*
*варенье 1*311000001/1*314000000/*

The second line of Table 4 meets the format of the rule, the application of M1 to the first and the second lexeme gives a definite morphological information, so the operation of crossing by feature "case" (the third position after *) yields the word combination *на (23*004000000/) стекло (1*314000000).*

Table 5

Syntactic rules of subordination of classes (CL)

Which CL	With what CL	Main word No.	Syntactic type	Rule	Example
1*	1*--2	(1)	S1	M2	*norm of law, keeping peace*
L09	1*--2	(1)	S2	M2	*17000 military men*
1*	1*--5	(1)	S2	M2	*governing the country*

In Table 5, in the second column, entry «1*--2» means that the case of lexeme of class 1* in the 3rd position may take the meanings either 2 or 5 (third line). Rule M2 determines that the second lexeme retains only the set of grammatical information with the appropriate meaning of the case, the first lexeme retains its all sets of meaning.

Table 6 defines the rules of прилягання that operate only on the syntactic compatibility of lexico-grammatical classes.

Table 6

Syntactic rules of parataxis of classes

Which CL	With what CL	Main word №	Syntactic type	Conditions	Rule	Example
9*	10*	(1)	П1	any	M3	*continue doing*
14*	10*	(2)	П2	any	M3	*influence considerably*
14*	12*	(2)	П2	any	M3	*already aimed*
...	...	...	...	...	...	...

After working out the rules for contextual syntactic analysis, the formed word combinations are checked (the problem of pragmatic interpretation) in the dictionary of terms and set word combinations in a given subject area. If the matches are found, such word combination is further considered as one lexeme, it is brought to the original form, but the set of grammar meanings is preserved. The original form is required in order to search for a word combination in the translation electronic dictionary, and saving the code of the text meaning is necessary for the synthesis in a language of translation.

The task of the next stage is to define the main parts of the sentence: subject, predicate, and secondary parts: object, preposition object, attribute, etc. For determining the main parts of the sentence, the rules, indicated in Tables 7, 8 are applied.

Separating the main parts of the sentence begins with determining the predicate. Once the predicate has been found in the sentence, candidates for the subjects are sought for according to the rules, given in Table 8.

Table 7

Rules for determining the predicate

Which class	With what class	Main word No.	Syntactic type	Example
12*	10*	(1)	Pred3	is able ... to apply
21*	10*	(1)	Pred2	слід перенести
9*	any	(1)	Pred1	performs

Table 8

Rules for determining the subject

Which class	With what class	gender	number	Number by order of the subject	Syntactic type	Example
1*--1	12*	+	+	(1)	Sub2	The pistol was designed
1*--1	9*	any	+	(1)	Sub2	President performs

The rules for determining the secondary parts of the sentence are presented in the format similar to the rules of coordination (attribute) and subordination (object, preposition object).

At the stage of the inter-phrase syntactic analysis for lexical connectors from the previous sentence, the words, which they replace, are defined. It is important for translation adequacy, since, for example, connector *він* is translated into English as *he* only in case it replaces the word that denoted a person, in other cases it is translated as *it*.

At the basis of multilingual machine translation lies the knowledge-oriented technology, the essence of which is the complex solution of problems of automation of elimination, submission and processing the knowledge from SA, contained in multilingual text sources. The features of the analysis of the NLT are determined by the orientation to the formation of the concept structure, i.e., to automatical knowledge extraction from multilingual texts and their pragmatic interpretation in terms of the applied problem. In this case, the text is considered as an object of different levels of analysis: as a sign system, as a grammatical system, and as a system of knowledge about the world (problem area) [8, 12, 16].

The core of LP are the algorithms of processing NLT, which includes the algorithms of accessing and processing linguistic database (LDB) and knowledge base (KB) on SA. Traditionally, LP in machine translation systems comprises two blocks: analysis and synthesis. Working with SA knowledge, contained in the NLT requires advanced means of interpretation of language units in terms of knowledge (concepts) themselves in both SA and relatively targeted orientation to the applied problem (knowledge of applied problem). This caused the development of the LP, which would include three separate blocks: analysis, interpretation and synthesis. The theoretical and practical basis of creating such LP lies in the systems of artificial intelligence systems to support natural speaker's interface [6, 9, 16]. But, it should be noted that problems of interpretation in the support systems of human and computer interaction considerably differ from the problems of processing NLT.

A distinctive feature of processing NLT is the involvement of different kinds of knowledge at each of its stages. The essence of the problems that are resolved by the linguistic processor at each level of text organization is presented in Fig. 3. Inclusion of the interpretation block as independent one at each level of the analysis of the source text allows, on the one

hand, distributing processing both at the level of the presentation of linguistic data, and at the stages of their processing. On the other hand, the introduction of the interpreter can make logical-semantic (by means of formal logic) knowledge processing independent on a certain source language. "Intermittence" of processing lies in the fact that at each stage of the textual information processing there is the immersion of the obtained results of processing into the knowledge about the world. This approach allows turning the disadvantages into disadvantages in the limits of the first two approaches. This is evident in the fact that every stage of processing is modeled as an independent module that allows getting more precise information. Thus, at the stage of morphological processing of immersion into DB of SA makes it possible to determine more precisely the grammatical characteristics of these lexical units, such as a *name, a title*, etc. For example, in the English language for the full name *Martha Browner* it is possible to define the category of gender only by the interpretation of lexeme *Martha* on KB of the SA, where for the first word form *Martha* will be determined: *female gender*, animated (human). Due to the work of the interpreter, the number of errors decreases both in the process of analyzing the incoming text and in the process of synthesis. In addition, the simulation of volumetric (that is, three-dimensional) information processing allows a considerable reduction of the processing time.

The concept of LP design is based on the following fundamental positions:

– a natural language text is the representation of three interrelated objects of analysis: semiotic system, grammar structure of certain language and knowledge of the world;

– fragments of knowledge, which are described in natural language texts reflect the state of professional (or, in the general case, logical-semantic) penetration in SA, rather than particular natural language.

Therefore, the implementation of the knowledge-oriented technology of building multilingual machine translation system involved the development of LP, which is to provide the processes of automatic recognition, formalization and processing knowledge of Ppg, contained in the NLT.

The proposed structural-logical scheme of the linguistic processor in addition to traditional processing modules implies: analysis and synthesis, including the interpreter as an independent module, performing the functions of representing the knowledge of SA with the language tools in the process of constructing a conceptual structure by NLT and synthesis of native language texts by their conceptual structure. This allowed, on one hand, improving the quality of the linguistic analysis at the morphological, syntactic and semantic stages, on the other hand, reducing the time of software text processing due to the parallel data processing.

6. Software implementation and experimental research into methods of analytical-synthetic processing in the knowledge-oriented machine translation system

When developing software, the following modules were distinguished.

The module of the grapheme analysis of the text is used for the preliminary analysis of NLT; the main objective of this module is to analyse NLT at the graphic level of the text representation: selection of structural fragments of the text, the headings, etc., as well as for further transfer to the lexeme and semantic processing level.

The module of morphological analysis (APM PARADIGM), Fig. 2, is used to create KB of a language. A professional linguist operates these components. Here we have another advantage of the modular use – for assessing the effectiveness of presented technology for certain text models, it is possible to substitute the implementation of the module of morphological analysis with the module that exactly identifies morphological information for the given texts, thus the assessment of other modules will be preformed.

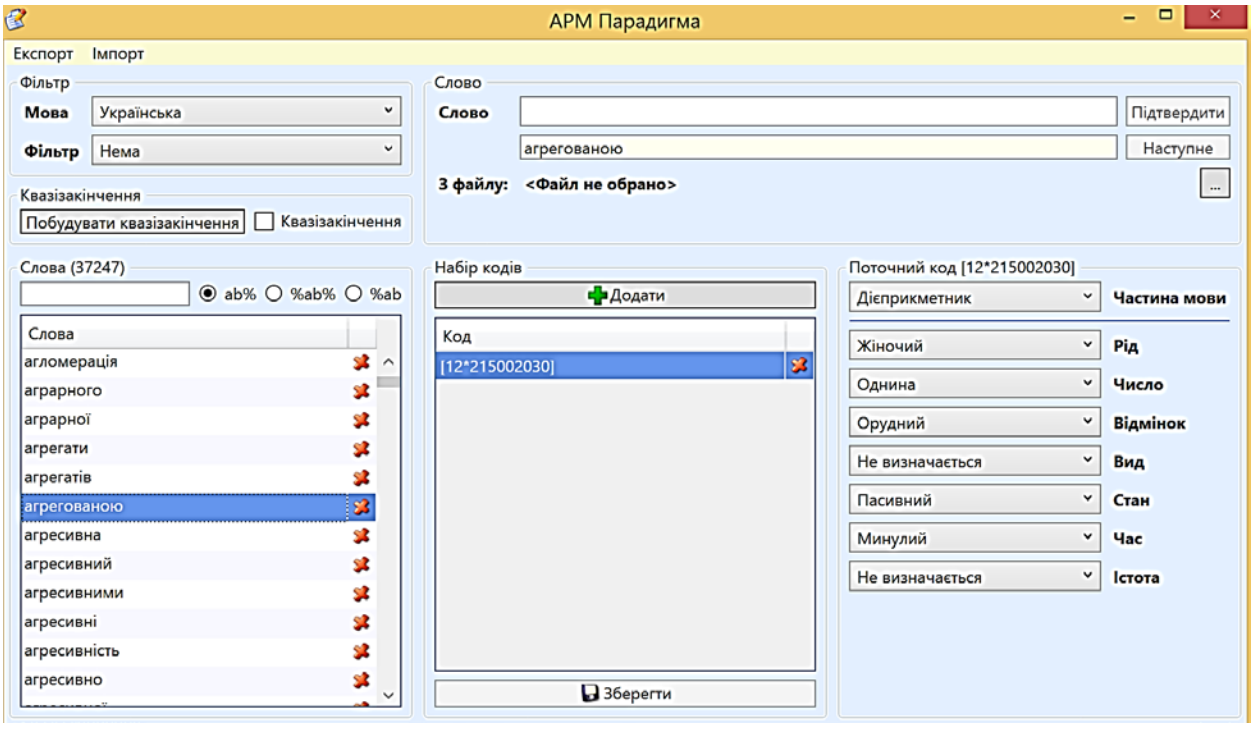

Fig. 2. General view of APM PARADIGM

The system APM PARADIGM allows entering and editing morphological information about the words, as well as providing the functional for additional features to build the components of linguistic support. Data in the system are presented in the form of matrix of lexico-grammatical classes and a list of grammatical categories, which introduce a single format for data entering for the Russian, Ukrainian and English languages. The classification, put into the basis of automated morphological analysis, is focused on the fact that the results serve as source data for automatic syntactic, lexical and semantic analyses of several languages.

The system APM PARADIGM is an interface for the introduction of morphological information for the objects that are added to the system. APM PARADIGM has also the functions of building the components of linguistic support – dictionaries of quazi-endings and link words, and a number of additional features that help a linguist to perform a high quality work with data and receive a variety of characteristics of the source data.

When entering the information for a particular word, a linguist does not work with digital encoding, but with the interface elements that display the appropriate lexical-grammatical categories and allow selecting only the morphological features, which will be fixed in advance for a specific lexical-grammatical category.

The module of lexeme analysis of the text is used for dividing the proposed NLT into the structures, which are mainly universal for each language, which causes greater universality of the algorithm.

The relationships, which specify certain parameters for the structure (words) and form the basis for further processing and translation, are established between some of the structures. At this stage it is possible to apply special dictionaries (abbreviations, geographical names, etc.).

The module of semantic analysis (APM EXPERT), Fig. 3, is meant for the final organization of phrases with determination of their semantic meaning. It is possible to involve special dictionaries (dictionary of idioms, etc.).

The developed structural schematic allows the implementation of the knowledge-oriented approach in SMF, as well as meeting all the above mentioned requirements and is scalable as for the quantity of data, users, and, most importantly, the load on the individual modules and adding new structures.

System APM EXPERT allows entering and storing semantic information that is intended for automated formation of encyclopaedic knowledge about the world. For each object, added to the system, a linguist identifies a semantic class, according to which it is possible to assign additional characteristics, such as “person”, “number”, etc. The selection of such features is transformed to the numeric form, this encoding is the data representation in the APM. The system allows entering the information for the English, Ukrainian and Russian languages. The following semantic classes are determined: special, which includes geographical names (names of countries, capitals, continents, oceans, seas, rivers, etc); temporal, which includes English names of months of the year, days of the week, as in English texts; they will always be spelled with capital letters. The semantic classes in MTS also include: names (proper names serve for the correct determination of gender when translating from English into Russian or Ukrainian); posts (in the text they are presented as proper names; it is necessary to distinguish them for further correct translation); proper names of institutions; units of measurement (spatial, temporal, and others).

The result of the work of a linguist in the system APM EXPERT, Fig. 3, is a dictionary of proper names, general for the source language and a dictionary of proper names in a particular subject area for the source language.

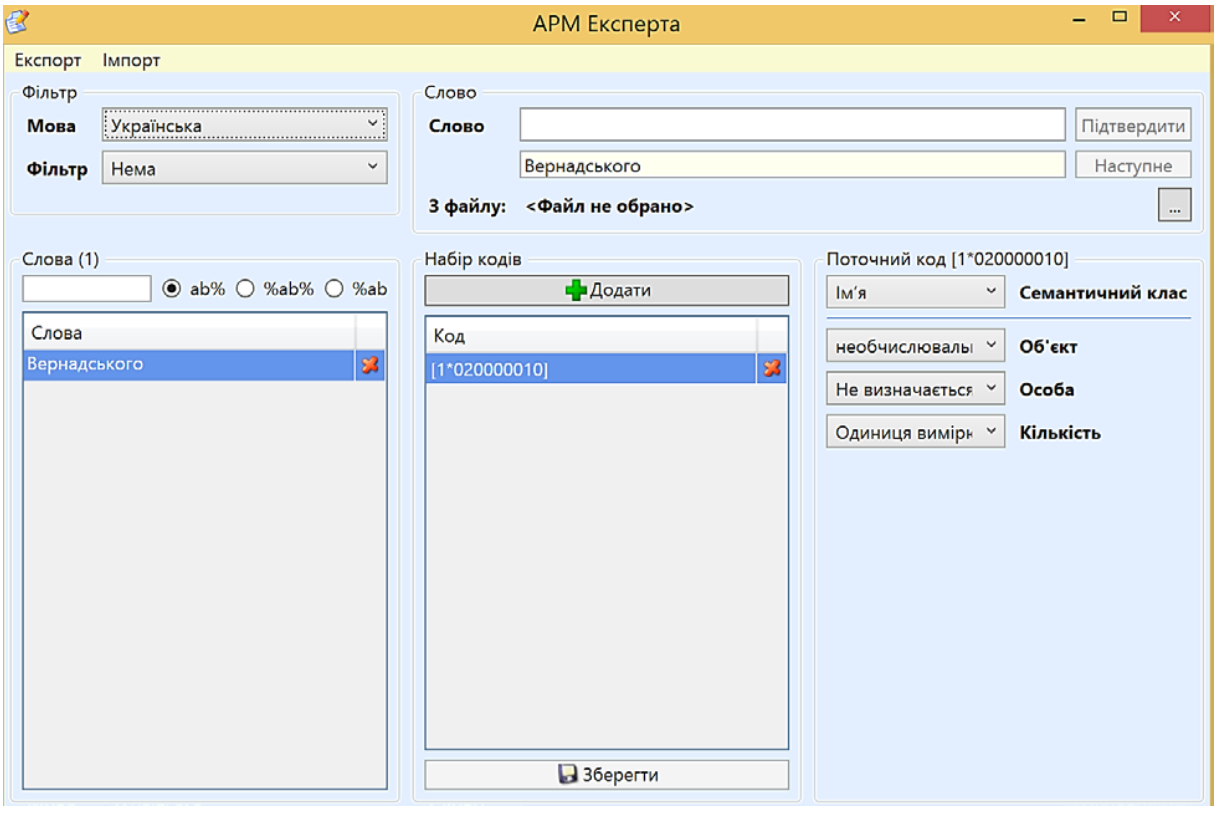

Fig. 3. General view of APM EXPERT

Actually, at the stage of processing a word, which is necessary to add to the system when entering the relevant information, a linguist selects the proposed semantic category from the list that displays the natural information representation. APM EXPERT automatically puts the correspondent numeric code – such approach allows significant reduction of the number of errors and speeding up entering the information because a linguist works with natural objects, rather than with numeric codes.

The special feature of the APM EXPERT is that the source data may include both relevant translation dictionaries and NLT.

The developed software product supports English, Russian and Ukrainian languages.

7. Discussion of results of testing the methods for MTS and the prospects for further research

Results of the research are implemented in the framework of the scientific research work "Technological principles for developing a knowledge-oriented multilingual machine translation system", cipher "ASTRA". The developed software systems were tested experimentally at the Military Institute named after Taras Shevchenko, Kyiv National University, Ukraine, and applied for devising methodological materials for training at the Institute of Philology of Taras Shevchenko Kyiv National University, Ukraine.

It should be noted that in case of downloading any NLT on the assigned subject –matter, the APM EXPERT automatically receives only capitalized words, abbreviations, contractions, which are recognized by their spelling at the stage of pre-morpheme analysis (for example, mln, km/h, *о-в, ртбр (rus.)*) and the words that may present shortenings or other classes that are not transferred to the stage of morphological analysis. Automatic detection of "suspicious" words in the text is achieved by the fact that the APM EXPERT combined with APM PARADIGM, and the words of the text are firstly checked based on all word forms of the correspondent language in the database.

Directions for further research in this area may include expanding and improving both the developed positional-digital morphological code and the models of semantics in order to extend the universal approach to the majority of inflectional languages. Practical development is the application of results, obtained in the systems for automation of processing NLT – search engines, machine translation systems, etc.

8. Conlusions

1. A method for automated syntactic text analysis based on declarative representation of the rules of syntactic combinability was developed. The method is based on the synthesis of tables of syntactic rules. The tables were designed not only for context analysis (rules of coordination, subordination and parataxis), but also for defining the subject, the predicate, secondary parts of the sentence, as well as for super-phrase syntax combinations.

2. A method for software distribution of analytical-synthetic processing of natural language texts in machine translation systems was developed. The method takes into account the conditions of transition to parallel data processing both at the level of processing tasks (analysis of the source text, its pragmatic interpretation, synthesis by means of another language), and depending on the data type.

3. The method developed for analytical-synthetic processing of miltilingual texts (Russian, Ukrainian, and English) was implemented in software in the form of applications. Experimental research into the developed software for the texts of military subject area revealed a decrease in errors of semantic nature by 14–16 % on average in comparison with the known MTS. A decrease in the number of errors is due to the automated text processing at the level of the sign system and due to the introduction of super phrase synthesis.

References

1. Toldova, S. Evaluation for morphologically rich language: Russian NLP [Text] / S. Toldova, O. Lyashevskaya, A. Bonch-Osmolovskaya, M. Ionov // International Conference on Artificial Intelligence (ICAI). – USA: ACM, 2015. – P. 300–306.
2. Freitag, M. Jane: Open Source Machine Translation System Combination [Text] / M. Freitag, M. Huck, H. Ney // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. – 2014. doi: 10.3115/v1/e14-2008
3. Clark, E. M. Sifting robotic from organic text: A natural language approach for detecting automation on Twitter [Text] / E. M. Clark, J. R. Williams, C. A. Jones, R. A. Galbraith, C. M. Danforth, P. S. Dodds // Journal of Computational Science. – 2016. – Vol. 16. – P. 1–7. doi: 10.1016/j.jocs.2015.11.002
4. Evans, J. A. Machine Translation: Mining Text for Social Theory [Text] / J. A. Evans, P. Aceves // Annual Review of Sociology. – 2016. – Vol. 42, Issue 1. – P. 21–50. doi: 10.1146/annurev-soc-081715-074206
5. Dunham, J. LingSync and the Online Linguistic Database: New Models for the Collection and Management of Data for Language Communities, Linguists and Language Learners [Text] / J. Dunham, G. Cook, J. Horner // Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages. – 2014. doi: 10.3115/v1/w14-2204
6. Wahl, H. A Generic Software Framework for Intelligent Integrated Computer-Assisted Language Learning (iiCALL) Environment [Text] / H. Wahl, R. Galler, W. Winiwarter // Lecture Notes in Computer Science. – 2015. – P. 264–270. doi: 10.1007/978-3-319-25515-6_26
7. Ghosh, S. Part-of-speech Tagging of Code-Mixed Social Media Text [Text] / S. Ghosh, S. Ghosh, D. Das // Proceedings of the Second Workshop on Computational Approaches to Code Switching. – 2016. doi: 10.18653/v1/w16-5811
8. Mel'cuk, I. A. Dependency Syntax: Theory and Practice [Text] / I. A. Mel'cuk. – NY: SUNY, 1988. – 428 p.

9. Apresian, J. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT [Text] / J. Apresian, I. Boguslavsky, L. Iomdin et. al. // Conference on Meaning-Text Theory. – Paris: Ecole Normale Superieure, 2003. – P. 279–288.
10. Zamarujeva, I. V. Komp'juterna model' rozuminnja pryrodno-movnoi' tekstovoi' informacii' [Text] / I. V. Zamarujeva // Problemy programmirovanija. – 1999. – Issue 2. – P. 96–102.
11. Zamarujeva, I. V. Znannja-orijentovanyj pidhid do avtomatyzacii' informacijno-analitychnoi' dijal'nosti [Text] / I. V. Zamarujeva, A. O. Ros', O. Ju. Gubajdulin et. al. // Problemy programuvannja. – 2000. – Issue 1-2. – P. 601–614.
12. Moroz, A. V. Automatic creating test as one of the processing tasks of natural language texts [Text] / A. V. Moroz // Eastern-European Journal of Enterprise Technologies. – 2012. – Vol. 2, Issue 2 (56). – P. 14–17. – Available at: http://journals.uran.ua/eejet/article/view/3658/3430
13. Turian, J. Word representations: A simple and general method for semi-supervised learning [Text] / J. Turian, L. Ratinov, Y. Bengio // The 6 Association for Computational Linguistics. – Sweden: ACM, 2010. – P. 384–394.
14. Klementiev, A. Inducing Crosslingual Distributed Representations of Words [Text] / A. Klementiev, I. Titov, B. Bhattarai // Conference on Computational Linguistics (COLING): 24th international conference. – Bombai: ACL, 2012. – P. 1–15.
15. Zou, W. Y. Bilingual Word Embeddings for Phrase-Based Machine Translation [Text] / W. Y. Zou, R. Socher, D. Cer, C. D. Manning // Conference on Empirical Methods in Natural Language Processing: 12th international conference. – USA: ACL, 2013. – P. 1–6.
16. Lytvynenko, L. Module of syntactical module for analysis of natural language texts [Text] / I. Zamaruieva, L. Lytvynenko // The Advanced Science. – 2013. – Issue 1. – P. 57–60.
17. Feldman, R. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data [Text] / R. Feldman, J. Sanger. – Cambridge: Cambridge University Press, 2006. – 423 p. doi: 10.1017/cbo9780511546914
18. Danchenkova, S. I. Automatic text classification in the system of concepts lexical ontology [Text] / S. I. Danchenkova, V. N. Polyakov // Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki. – 2010. – Vol. 152. – P. 255–267.
19. Wahl, H. Natural language processing technologies for developing a language learning environment [Text] / H. Wahl, W. Winiwarter, G. Quirchmayr // Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services – iiWAS '10. – 2010. doi: 10.1145/1967486.1967546
20. Cohn, T. Machine translation by triangulation: Making effective use of multi-parallel corpora [Text] / T. Cohn, M. Lapata // Meeting of the Association for Computational Linguistics: 45th annual conference. – USA: ACL, 2007. – P. 728–735.
21. Bazrafshan, M. Semantic Roles for String to Tree Machine Translation [Text] / M. Bazrafshan, D. Gildea // Meeting of the Association for Computational Linguistics. – USA: ACL, 2013. – P. 419–423.
22. Furstenau, H. Semi-Supervised Semantic Role Labeling via Structural Alignment [Text] / H. Furstenau, M. Lapata // Computational Linguistics. – 2012. – Vol. 38, Issue 1. – P. 135–171. doi: 10.1162/coli_a_00087
23. Xiong, D. Modeling the Translation of Predicate-Argument Structure for SMT [Text] / D. Xiong, M. Zhang, H. Li // Meeting of the Association for Computational Linguistics: 50th annual conference. – USA: ACL, 2012. – P. 902–911.
24. Nikolaievskyi, O. Components of Lingware for Automatic Morphological Analysis in Kwnoeledge-Oriented Machine Translation System [Text] / O. Nikolaievskyi // The Advanced Science. – 2013. – Issue 5. – P. 32–36.
25. Nikolaievskyi, O. Procedure of Forming Dictionary of Quasi-Inflections Basing on Wordform Dictionary [Text] / I. Zamaruieva, O. Nikolaievskyi // The Advanced Science. – 2013. – Issue 8. – P. 61–65.