Journal of Theoretical and Applied Information Technology
15th March 2017. Vol.95. No 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MODELS AND TOOLS FOR AUTOMATIZATION OF THE LINGUISTIC RESEARCH

[1]LAKHNO V.A., [2]NIKOLAIEVSKYI O.Y., [3]SKLIARENKO E.V., [4]LYTVYNENKO L.O.

[1,2,3,4] Department of Managing Information Security, European University, Ukraine

E-mail: [1]lva964@gmail.com, [2]a1.n1@yandex.ru, [3]sigma.inet@gmail.com, [4]l.lytvynenko@gmail.com

ABSTRACT

This article is focused on solving problem of building tools for automatization of linguistic research, particular for the morphological and semantical multilingual analysis. The results obtained are of great importance for the development of linguistic research. Developed a procedure for the automatic creation of dictionary quasi-inflections of wordforms as linguistic components providing automatic morphological analysis. Developed a system of the automatic workplace AWP PARADIGM that allows to automate the process of building the components of lingware for automatic morphological analysis in machine translation system, as well as to provide functional completeness. Developed system of AWP EXPERT, which allows you to automate the process of building the components of lingware for automatic premorphemic analysis and definition of the semantic information in machine translation system (SMT). The developed software product support English, Russian and Ukrainian languages. It is proved that the proposed method of encoding lexical-grammar information and approach to automatic morphological analysis using quasi-inflections has significant advantages in comparison with existing methods.

Keywords: *Linguistic Information; Natural Language Processing; The Data Processing System; Quasi-Inflections; Automation Of Morphological Data Processing; Semantic Information Encoding.*

1. INTRODUCTION

The modern stage of development of information technologies contributes to the scope of application of information technology in systems of automatic analysis (SAA) of natural language texts (NLT).

For such systems, expanding the range of problem solving tasks and increases the number of schedules and software products that solve similar problems that leads to exponential growth of the volume processable linguistic data.

However, there are certain problems in the field of the development of machine translation systems and SAA of NLT. First, lingware automated data-processing systems (ASP) NLT has a narrowly specialised focus and is not suitable for use in other systems. Secondly, a variety of automation systems of linguistic researchers are not comfortable for expert linguist and cannot in a single format to work with several languages, which, as a consequence, hampers and implementation toolkit.

However, the theoretical achievements in the field of structural, applied and mathematical Linguistics, as well as the intelligent computer SMT, can talk about the general patterns of the functioning of the NLT, which can ensure their use in various systems, automatic processing of NLT. Previous analyses [1, 6, 10] focused on one certain language or describe approach without possibility for good automatization further researches.

The general problem of raising the efficiency of automatic processing NLT causes the relevance of scientific tasks, that is solved and placed within the framework of this study scientific and methodological substantiation of development tools automate the building components of lingware ASP NLT computer.

2. FORMULATION OF THE OBJECTIVES OF THE STUDY

The purpose of the research is to improve the efficiency of computer systems automatically NLT processing by creating a toolkit of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

automation construction of components of linguistic support for ASP NLT.

To achieve the goal of the work necessary to solve the following tasks:

- develop the method and model of the encoding of the lexical-grammar information to implement automatic morphological analysis and synthesis based dictionary quasi-inflections;
- to develop a method of encoding semantic information to implement automatic semantic analysis of lexical units NLT in ASP;
- develop lingware to create a dictionary of the quasi-inflections for word forms, as components of linguistic automatic morphological analysis in ASP and SMT NLT.

3. REVIEW OF THE LITERATURE

The development of linguistic processors in ASP NLT is dedicated to numerous studies. The most common in CL (in computer linguistics) morphological dictionaries used for morphological analysis in their dictionary article presented morphological information corresponding to the word grammar.

Depending on the organization of the linguistic processor [1, 2] in the dictionary can be added grammar information as well. There are dictionaries, in which it is presented and more extensive information about the words [3]. In a number of linguistic processors used dictionaries of synonyms. A relatively new type of dictionaries-dictionaries of paronymies, the superficially similar words, which differ within the meaning [4]. According to the authors of [5] the most complex types of lexical resources in ASP NLT is the thesaurus and ontology. The spread of thesaurus is associated with the solution of tasks of information retrieval [6]. The concept of the thesaurus has closely related the concept of ontology [7].

Now ways of encoding the lexical-grammar information to word forms are based on the integral representation of grammar information to word forms

The technique of integrated digital encoding of grammar information [8] is designed for building a dictionary of quasi-inflections in automated systems of classification and indexing of abstracts. The disadvantage of the method of encoding a priory the incompleteness of the selected grammar categories.

Another method of integral letter coding lexical-grammar information described in work [9]. The developers used a combined method. The disadvantage of this method can be called a limited its use, so it is effective for building paradigm word forms, for example in automated systems, stacking grammar dictionaries. In addition, the combined use of dictionaries quasi-inflections and quasi-base makes an automatic morphological analysis of new words. [15, 16]

A common disadvantage of the methods of the integral coding lexical-grammar information is a highly specialized field of their use, not adaptable for a variety of tasks the ASP NLT

Numerous studies and publications [10, 18], related to the development of ASP NLT, indicate that there is a need to develop and research new methods, models and tools automate the building components of linguistic software in computer ASP NLT [17]. This article will cover new approaches to automatization of morphological and semantical analysis as a part of linguistic processor for ASP NLT, syntax will not be cover by the research.

4. MATERIAL AND METHODS

The model of lexical-grammar and lexical-semantic information formalization based on the method of position-digital encoding for ASP NLT proposed an approach to developing a semantic model is based on the following conceptual provisions:

− incoming NLT is linked text (that is, the discourse);

− the cohesion of discourse is provided by the graphemes by means of text (the ratio between the headers and the content of paragraphs of text, etc.), linguistic tools (grammar approvals, more anaphoric links and more) and extra-linguistic (temporary, causal relations, etc.);

− all of these tools is the tool the encoding of knowledge about the world (domain).

Based on this, a semantic analysis of the incoming NLT is distributed. This is caused by the development of the semantic code separate lexical-semantic, which have extra linguistic load semantics. The input unit of the semantic information content of an encoding method is the notion. The concept of the text can be presented as a separate word and phrase. The proposed method is based on the position-digital encoding semantic information. In this way, each analyzing concept gets its digital code that contains information about the semantic characteristics inherent in a single word in any context, and define the semantic characteristics of semantic compatibility in the text. Offers the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

following semantic characteristics, which in our view full enough able to describe as their own so the contextual semantics of the concepts: process, object, State, time, space, person, number, and quality. The necessity of introduction of semantic characteristics of "entity", due to it is contextual semantics (or rather pragmatics). So, for example, when the automatic development of texts of the political orientation of importance is not just a grammar category of the "creature" and "semantic" person ", but this characterization, officer (or first officer). This will depend on which degree of urgency (the importance of) will be provided with the appropriate text. The sign "+" marked semantic characteristics that are defined for the corresponding lexical-semantic classes.

In the proposed approach to encoding semantic information code, lexical-semantic class separates from the respective semantic descriptions of the sign "*". Semantic characteristics have a fixed position (Yes, as they are recorded in table 1) and take the value on the interval [0; 9]

Thus, we have 9-position numeric code that uniquely describes the semantics of the corresponding concepts. In the case of multivalency of the notion through the sign of "/", introduced an alternative semantic code, this ambiguity can manifest at the level of lexical-semantic classes and semantic characteristics. This method of encoding allows you to get a unique semantic key for each concept. In the process of automatic processing of semantic ambiguity is eliminated as by grammar parsing sentences (by matching the lexical-grammar and lexical-semantic classes), and at the stage of pragmatic analysis of the text.

As mentioned, every semantic characteristic (see columns 2-10 table 1) can take values from 0 to 9. For each semantic characteristics constructed its own code table. It should be noted that the values 0 and 9 are fixed for all code tables of semantic characteristics. 0 means that the characteristic is absent for the corresponding lexical-semantic class (for example, the lexical-semantic class 9 – "locality" of table 1 contains no semantic characteristic of "person" and respectively in the seventh position will be 0). 9 means that this characterization is uncertain, that is not essential for the shortcut combination of concepts. The process of automatic analysis processing algorithm takes 9 as such which is consistent with any value in the appropriate position.

Table1: The codes of lexico-semantic classes and their semantic characteristics

Code LSC	Lexico-semantic class	Semantic characteristics									
		process	object	state	time	space	property	person	number	quality	examples
1	Name		+					+	+		*Alexander*
2	Middle Name		+				+	+	+		*Petrovich*
3	Family name		+				+	+	+		*Ivanov*
4	Nickname		+				+	+	+		*Kashtanka*
5	Historical event	+			+	+			+		World War II
….	Other 7 objects type here, simplified for the article.										….
12	Water name		+			+			+		*Baikal*
13	The name of planets, stars		+			+			+		*Mars*
14	The name of the parts of the world		+			+			+		*Far East,West*
15	The name of the continents		+			+			+		*Africa*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Code LSC	Lexico-semantic class	Semantic characteristics									
		process	object	state	time	space	property	person	number	quality	examples
16	The name of the organizations		+			+	+		+	+	*Armed Forces*
17	The name of the institution		+			+	+		+		*Verkhovna Rada*
18	The name of the publications		+			+			+	+	*"Today"*
19	The name of months, days		+		+				+		*Wednesday, January*
20	Share									+	*If*
21	Modal	+			+					+	*you need*
22	Article		+						+		*the, an, a*
23	Preposition		+		+	+					*when, with*
24	Conjunction						+		+		*also, but*
25	Gerund	+	+	+	+	+	+		+		*working*
26	Adverb	+		+	+	+	+			+	today, slowly
27	Noun	+	+	+	+	+	+	+	+	+	run, bone, anger
28	Adjective			+	+	+	+		+	+	priority
29	Numeral						+		+		second, thirty
30	Verb	+			+		+		+	+	to run,
31	Participle	+		+	+		+		+	+	detailed
32	Adverbial participle	+		+	+		+				determining

Note that the column Name "semantic specifications" of the table.1 is the default setting and allows the researcher to or complement the back cells or form replace the name under their own applied tasks of analyzing the text. This allows you to customize the data in the system as a new subject area, and a new analysis NLT intact software [11].

The method of encoding of lexical-grammar information to implement automatic morphological analysis and synthesis. The proposed method is based on the position-digital encoding of grammar information in the dictionary article. In this way, each used previously wordform gets its code that contains information about a part of a language, and the specific grammar meaning. [12, 13] So it's a feature often helps identify and underline the functionality of language, which, as a rule, remained out of focus. Regarding this and highlight parts of the language, but rather a lexical-grammar class. A list of grammar classes adopted in work classification consists of 31 (but the list is not closed and can replenish) the selected group (lexical-grammar classes-LGC). Of course, separate the traditional parts of speech: noun, verb, adjective, numeral, preposition, conjunction, etc. Define: article, gerund, the word as a foreign origin. The entered changes concerning classes:

verbs, in a separate group select the verb last time, not time, infinitive, pointing to the features of the control verb bidding way;

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

verbs in separate classes select verbs last time, former time, infinitive, pointing to the features of the control case, the verb bidding way;

in separate classes highlighted verb forms is a participle and adverbial participle;

number (apart as the numeral-noun, numeral-adjective); for pronouns applies only as the character of their values, and the specificity of wordform and functioning in the language. Respectively, are considered pronouns that have adjectives the type of conjugation (some, any, no), pronouns-nouns (whoever), the personal pronouns (he, she); for the proposition added grammar category control case. To meet the demands of multilingualism and the creation of universal components introduced such lexical-grammar classes as the gerund and to build appropriate codes for the English language. [14] in table 2 are the lexical-grammar classes taken in a system of automatic morphological analysis. Here the sign "+" marked grammar category that is defined for the corresponding lexical-grammar classes.

Table 2: Codes of lexical-grammar classes and their grammar categories

Code LSC	Lexical-semantic class	Grammar categories								
		Type	Number	Case	Person	View	Condition	Degree	Time	Being
1	Noun	+	+	+						+
2	Adjective	+	+	+						
3	Numeral-noun	+		+						
4	Numeral-adjective	+	+	+						
5	Pronoun- personal	+	+	+	+					+
6	Pronoun-noun		+	+						+
7	Pronoun-adjective	+	+	+	+					
8	Verb, past tense	+	+	+		+	+			+
9	Verb, current tense		+	+	+	+	+		+	+
10	Infinitive			+		+				+
11	Verb imperative		+	+	+	+				+
12	Participle	+	+	+		+	+		+	+
13	Adverbial participle			+		+				+
14	Adverb							+		
15	Comparative adjective							+		
16	Possessive adjective	+	+	+						
17	Short adjective	+	+							
18	Brief communion	+	+	+					+	+
19	Comparative adverb							+		

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Code LSC	Lexical-semantic class	Grammar categories								
		Type	Number	Case	Person	View	Condition	Degree	Time	Being
20	Share									
21	Modal		+		+				+	+
22	Article	+	+							
23	Excuse			+						
24	Union									
25	Gerund									
26	Union-noun			+						+
27	Union adjective	+	+	+						
28	Verb auxiliary		+		+				+	+
29	Verb ending to “сь”, “ся” (in Cyrillic languages)	+	+	+	+	+	+		+	+
30	Verb impersonal					+			+	
31	Words of foreign origin									

At works [13] describes the values that can take these grammar categories that are defined for the corresponding lexical-grammar classes.

Formation of lexical-grammar code for word. Each wordforms be put in line with the code, which is formed on the basis of the cast. So, for example, for the word "linguistic" will this code: 2 * 921000000/2 * 924000000/.

The first source (the symbol *) goes to indicate lexical-grammar class, and then – grammar characteristics (race, number, case ...). As we have already noted, each grammar characteristics has a specific set of parameters that are marked with a number from 0 to 9. Even if wordform has a few morphological values (present homonym) – that all possible combinations are set via the symbol/.

Immediately noticeable features of the proposed method of encoding of the element: the possibility of access to grammar information, availability of case in the verb (to control the lexical-semantics information on the stage of parsing), the possibility of defining the Word as foreign words – define what this word disgusting the input language. Additionally, worldform a dictionary of official words that should not be on analysis when building a dictionary quasi-inflections and must contain the necessary grammar information to Lexical-grammar classes as a cardinal number-noun, Cardinal-adjective, conjunction, the auxiliary verb, pronoun-adjective, preposition, article. Lemmatization and paradigmatical dictionary constructed using quasi-inflections and similar lexical-grammar encoding. Consider the principles of these components of the linguistic database. Let's start with the dictionary quasi-inflections – consider the snippet from the dictionary and describe briefly the content of the process of its construction. For example, suppose we have this typed and marked dictionary word forms (see table 3).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Fragment of the dictionary (for Ukrainian language)

Wordform (*for Ukrainian language*)	**A set of grammar codes**
справа	1*211000002\
трава	1*211000002\
протрава	1*211000002\
октава	1*211000002\
контроктава	1*211000002\
застава	1*211000002\
підстава	1*211000002\
вистава	1*211000002\
постава	1*211000002\
лева	1*112000001\1*114000001\

The process of building a dictionary of quasi-inflections is covered: from the end of words in alphabetical order. Quasi-inflections, in this case, acts as a collection of letters from the end of the word. The process begins with the identification of the most commonly used code (dialing codes) for one last letter, for example, the letter "quasi-inflections". The future is determined by the most used code for quasi-inflections "ва" if it coincides with the code for "a", then such quasi-inflections is not paid to the dictionary, otherwise the end contributed to the dictionary and the process continues. If at some point determined, more short quasi-inflections corresponds to the same set of codes and other, which includes such, it is larger than the size of the quasi-inflections is removed from the dictionary. This passage is done on all letters of the alphabet. At the output, we get dictionary quasi-inflections in the form of the quasi-inflections and the corresponding predefined sets of codes. Of course, this process of quasi-inflections generating can take place recursively [13].

The AWP EXPERT and PARADIGM as components of linguistic software for automatic semantic analysis in ASP NLT.

System AWP EXPERT [21] allows you to enter and store semantic information that is intended for automated formation of encyclopedic knowledge about the world. For each attached to an object, a linguist identifies a semantic class, corresponding to which you may specify additional characteristics, such as "person", "number", etc. The choice of the following translated to numeric form is the following and is a representation of the data in the AWP. The system allows you to enter information for English, Ukrainian and Russian languages. Determined such semantic classes: space – this includes geographical names (names of countries, capitals, continents, oceans, seas, rivers, etc); [19] the time – here are the English names of the months of the year, days of the week, because in English texts, they will always be spelled with capital letters; the names of these proper names are for the correct determination of the kind of translation from English into Russian or Ukrainian; positions in the text presented as proper names, their separation is necessary for the proper translation, it is true and proper names-institutions; the units – which also includes spatial, temporal, and others.

The result of the linguist in the AWP EXPERT, fig. 1, is a dictionary of proper names, common for the input language and dictionary of proper names in a particular subject area for input languages [20].

Because many words that are handled by the linguist while working in the AWP EXPERT, may have the same digital positional semantic codes, proposed storage scheme provides the advantage that such codes are not duplicated and do not create additional data that increase the size of the database and pose with them, and are only a link to the corresponding word codes.

Actually, at the stage of processing of the words that you want to add to the system by entering the relevant semantic information

Journal of Theoretical and Applied Information Technology
15th March 2017. Vol.95. No 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

linguist chooses proposed semantic category from the list, reflecting the natural representation of information, and the system itself puts into compliance with numeric code – this allows you to significantly reduce the number of errors and speed up the introduction of the information because a linguist working with natural objects, not numeric codes.

Feature AWP EXPERT is that input data can be as relevant translation NLT dictionaries.

The developed software product support English, Russian and Ukrainian languages. For each language, you create a separate database, which is involved in accordance with the language of the text. The database regardless of the languages has single unified semantic parameterization as reflecting the same snippets of knowledge about the world.

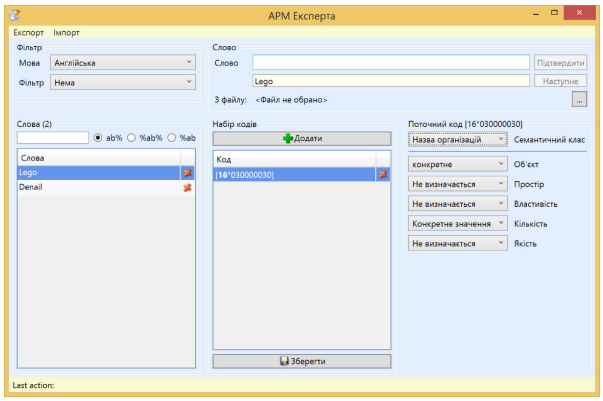

Fig. 1. AWP EXPERT general view (Ukrainian UI)

Formation of dictionary units comes in two modes: directly after the text file and the manual introduction of lexers operator. The contents of the text file can contain general background information (such as a dictionary of names, a list of units of measure, a dictionary of place names, etc.), the following files reflect generally accepted knowledge about the world and, as a rule, are not accompanied by explanatory context. The system AWP PARADIGM allows you to enter and edit the morphological information about the words and also provides the functionality for advanced features for building the components of linguistic support. In the system are presented in the form of a matrix of lexico-grammar forms and list of grammar categories, which introduces a single format to enter data for Russian, Ukrainian and English languages. Classification put in a basis of the automatic morphological analysis, focused on the fact that the results serve as initial data for automatic syntax, lexical and semantic analyses of several languages, fig, 2.

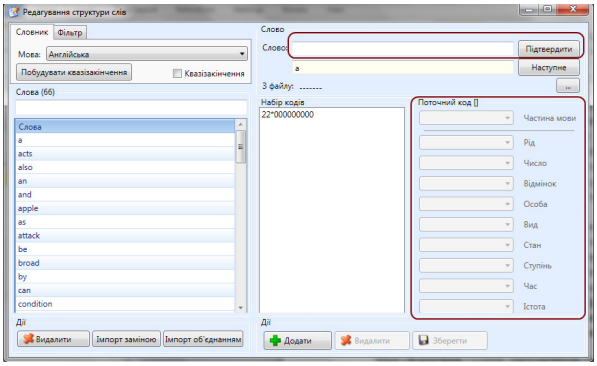

Fig. 2. General view of the AWP PARADIGMA (Ukrainian UI)

List of grammar classes adopted in work classification consists of 31 selected groups (lexical-grammar classes). Of course, separated the traditional parts of speech: noun, verb, adjective, numeral, preposition, conjunction, etc. Define article, gerund.

Interface programs AWP EXPERT and PARADIGM is developed using Microsoft technology stack, including using Net Framework version 3.5 and WPF using a template (pattern) design of Model-View-ViewModel (MVVM). The system AWP PARADIGM is an interface for the introduction of morphological information for objects that are added to the system. Well AWP PARADIGM has the function of building components of linguistic support and quasi-inflections dictionaries of words, and a number of additional features that help you lingware high-quality work with data and receive a variety of characteristics of input.

When you type the information for a particular word linguist does not work with digital encoding, and with the interface elements that reflect the appropriate lexical-grammar category, as well as allow you to choose only the morphological traits, which were fixed in advance for a specific lexical-grammar category.

5. TEST RESULTS THE AWP

The proposed approach in the implementation of the software allowed us to realize the following theoretical fundamentals – the study of the laws of language, i.e. the resulting dictionary quasi-inflections not under the subject field

However, the volume of the dictionary of quasi-inflections depends on the accuracy of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

filling the position-digital code, which in turn depends on the specific applied problems of processing of natural language text.

Depending on the applied problems on the interval [0.9] are determined by the value of grammar categories that are needed for solving the problem.

For the problem of the choice of automated stacking keyword list (concepts) in the information and search engines is enough to select the base set (in the AWP Paradigma) only the lexical-grammar classes that define the subject area and it is nouns and adjectives. For a given applied tasks of the value of the grammar category of "essential′ is not important (last position), so instead of a specific value can be set to 0. The volume of the dictionary of quasi-inflections given the official parts of speech and unchanged words will make up no more than 3 thousand articles.

For native NLT "essential" is not only important, but the English translation need to Supplement even more. So, when translated into English is not only the creature, man or not, because it depends on what pronoun can be used in the context of (she, he, it), that is, it is advisable to enter the value "3" is the creature man, on the interval [0,9] are reserved positions. In addition, the task of machine translation includes all stages of the text analysis including morphological, syntactic and semantic analysis of the text. It is therefore advisable to include in the code the value of grammar case for verbs that allow the parsing stage to automatically determine the group of the predicate and to reduce syntactic homonymies.

These features greatly affect the volume of the dictionary quasi-inflections, because the more the variability of the code, the greater the volume of the dictionary quasi-inflections, fig. 3.

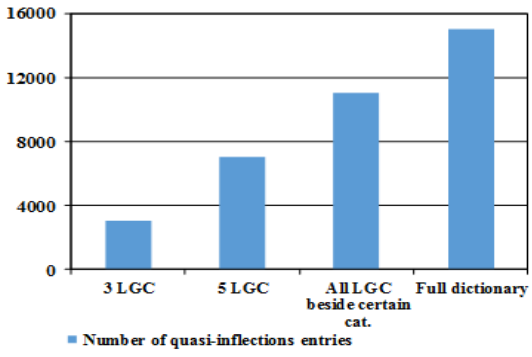

Fig. 3. The entries number of the dictionary quasi-inflections depending on the chosen lexical-grammar classes

An important criterion is that the resulting dictionary cases size is much smaller than the dictionaries for morphological analysis are obtained by other methods of coding that fig. 4 illustrate.

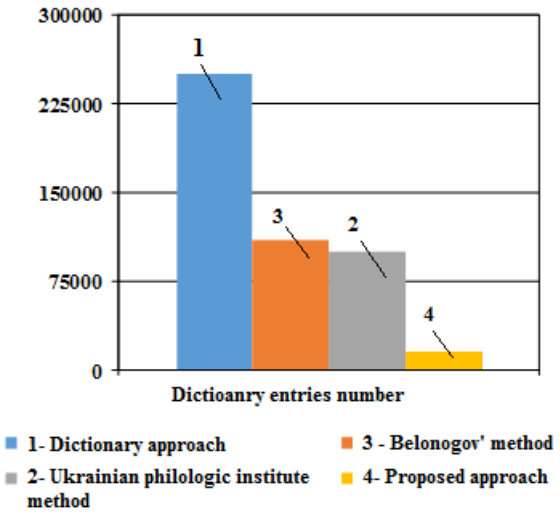

Fig. 4. Compare the number of articles in the dictionaries of linguistic databases of morphemic level

As an important criterion for evaluation is the convenience of lingware – the ability to enter data directly from a text file and the convenience of filtering help to significantly speed up the process and in practice and achieve the speed input is approximately thousand entries a day. While working and checking cover dictionary quasi-inflections language laws must, of course, build vocabulary, in fact, the procedure of building takes up to 10 minutes, allowing you to carry it every day during a set of a dictionary databases.

Once the test texts specialist will confirm that you have reached the required accuracy determination of morphological information – it can be argued that the functional completeness and for processing in industrial bulk no need to rebuild the dictionary quasi-inflections.

6. DISCUSSION

Obtained scientific results in practical terms brought to implementation in the form of AWP PARADIGM and AWP EXPERT. Developed software systems were investigated experimentally in Military institute department of Kyiv National University after Taras Shevchenko, and used in the preparation of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

methodological materials for training in the Institute of philology of Kyiv National University. It should be noted that if you download any NLT with the set of subjects to the AWP EXPERT, automatically receive only words capitalized, abbreviations, contractions, you are writing detected at the stage of premorphemics analysis (for example a million, km/h, in, rtbr) and the word suspicious on reduction or other classes that are not transferred to the stage of morphemic analysis. Automatic detection of "suspicious" words in the text is achieved by the fact that the AWP EXPERT combined with AWP PARADIGMA, and the words of the text originally validated based on all word forms language that is in the database. Directions for further research in this area is expanding and improving as the developed position-digital morphological code and models of semantics in order to expand the universal approach to most of the inflectional languages. Open issues are about covering non-inflectional languages, also applying current model of semantics to existent linguistic ontologies could be challenging. Practical development is located within the application of the results obtained in the systems of automation of processing NLT-search engines, NLT systems and other.

7. CONCLUSIONS

1. Improved method for position-digital encoding of lexical-grammar information in the direction of expansion of lexical-grammar classes (in particular, the introduction infinitive, article, etc.) and accessory grammar information (in particular, the introduction of the category of case for verbs, the category essential of nouns, etc.). This allowed within a unified system of encoding formally introduce lexical units in English, Ukrainian and Russian and languages (with the ability to spread to other languages). The inclusion in the code of the lexical-semantic valence verb allows more precision to conduct automatic parsing essential.

2. For the first time developed a procedure of automatically generation of quasi-inflections dictionary as a linguistic software components of automatic morphological analysis, which differs from the known fact that dictionary quasi-inflections reflects the patterns of wordforms in the language of the ending chain letters. The advantage of this dictionary is that it by 10 times smaller than the known; it depends on the language of the subject area and is able to provide automatic morphological analysis of new words in the text with precision to the word. 3. For the first method and lingware to encode semantic information and realization of automatic semantic analysis of lexical units, which unlike existing has extended the classification of semantic classes and cover semantic features are universal as regards input languages, and relative to the subject field (time, space, quantity, quality, etc.).

REFERENCES:

[1] I. Boguslavsky, et al. "Plausible Expectations-Based Inference for Semantic Analysis", *Proceedings on the International Conference on Artificial Intelligence (ICAI). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, 2016.

[2] A. Dobrov, et al. “Morphosyntactic Analyzer for the Tibetan Language: Aspects of Structural Ambiguity”, *International Conference on Text, Speech, and Dialogue. Springer International Publishing*, Vol. 9924, 2016, pp. 215–222.

[3] Yu. M. Shcherbina, V. A. Vysotskaya, T. V. Shestakevich. Metody i razdely matematicheskoy lingvistiki v strukture kursa «Kompyuternaya lingvistika», *Kompyuternye tekhnologii: nauchno-metodicheskiy zhurnal,* Vol. 117(104), 2009. pp. 203–208

[4] Yang, Lim Chan. “Lexical-semantic analysis of English adjectives with suffixes-ic/-ical”, 2016.

[5] Ma, Meng, et al. “OntoEvent: An Ontology-Based Event Description Language for Semantic Complex Event Processing”, *International Conference on Web-Age Information Management*, *Springer International Publishing*, 2015, pp. 448–451.

[6] B. Croft, J. Lafferty, eds. “Language modeling for information retrieval”, *Springer Science & Business Media*, Vol. 13, 2013, pp. 57–71.

[7] P. Cimiano, C. Unger, and J. McCrae. "Ontology-Based Interpretation of Natural Language." *http://www.aclweb.org/anthology/J/J15/J15-2006.pdf 2015.*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[8] I.V. Zamarueva, A.A. Shipnivska. "Morfemnaya obrabotka tekstov v sistemakh mashinnogo perevoda", *Vestnik Kievskogo natsionalnogo universiteta imeni Tarasa Shevchenko. Voenno-spetsialnye nauki*, Vol. 20, 2008, pp. 61–63.

[9] T.A. Gryaznukhina. Morfologicheskiy analiz nauchnogo teksta na EVM, *Naukova dumka*, 1989, p. 264.

[10] V.V. Balabin, I.V. Zamarueva, S.V. Lenkov, I.V. Pampukha. "Tekhnologicheskie aspekty realizatsii avtomatizirovannykh sistem mashinnogo perevoda", *Sb. nauk. pr. VIKNU*, № 26, 2010, pp. 55– 64.

[11] D. Vitkutė-Adžgauskienė, et al. "NLP infrastructure for the Lithuanian language", Proceedings *of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 2539–2542.

[12] A.Yu. Nikolaevskiy, I.V. Pampushki, L.A. Litvinenko, L.A. Ryaba, L.V. Solodeeva. "Postroenie i ispolzovanie slovarya kvazizakinchen", *Sbornik nauchnykh trudov Voennogo instituta KNU im. Tarasa Shevchenko*, № 14, 2008, pp. 154–158.

[13] I. Zamaruieva, O. Nikolaievskyi. "Procedure of Forming Dictionary of Quasi-Inflections Basing on Wordform Dictionary", *The Advanced Science*, Vol. 2013, Iss. 8, 2013, pp. 61–65.

[14] C. Green. "Approaches to English Clause Grammar", *Patterns and Development in the English Clause System. Springer Singapore*, 2017. pp. 1–16.

[15] S. Toldova, O. Lyashevskaya, A. Bonch-Osmolovskaya, & M. Ionov. "Evaluation for morphologically rich language: Russian nlp", *In Proceedings on the International Conference on Artificial Intelligence (ICAI)*, 2015, pp. 300–306.

[16] L. Berlanda, A. Fraser. "POS and Morphological Tagging & Lemmatizing", 2016. http://www.cis.uni-muenchen.de/~fraser/morphology_2016/folien_27_06_16_LB.pdf

[17] H. Furstenau, M. Lapata, "Semisupervised semantic role labeling via structural alignment", *Computational Linguistics*, No. 38(1), 2012, pp. 135–171.

[18] M. Tetsuro, K. Kuramitsu. "A declarative extension of parsing expression grammars for recognizing most programming languages" *Journal of Information Processing,* N. 24.2 2016, pp. 256–264.

[19] A. Lüdeling, J. Ritz, M. Stede, and A. Zeldes. "Corpus linguistics and information structure research", *The Oxford Handbook of Information Structure*. 2016, pp. 13–21.

[20] Wendy G. L., H. R. Martin. "Strategies for natural language processing", *Psychology Press*, 2014, p. 198.

[21] A. Yu. Nikolaevskiy. "Osobennosti formirovaniya komponentov znaniy ob okruzhayushchem mire v ARM EKSPERT", *Voennoe obrazovanie i nauka: nastoyashchee i budushchee: VII Mezhdunarodnaya nauchno-prakticheskoy. konferen*., 2012, pp. 152.