Technology for creating digital explanatory dictionaries

Anush Melikyan[1,*]

[1]Don State Technical University, 344003, Gagarin Sq., 1, Rostov-on-Don, Russia

Abstract. The work considers a new dictionary technology adapted to the digital environment. A network-based knowledge database is developed as a modern way of organizing explanatory-translation dictionaries. The work describes the organization of multilingual explanatory-translation dictionaries by ISMA technology and produces a system of expert organization scheme to support the creation of explanatory dictionaries based on the network knowledge database.

1 Some lexicography problems that arise in a digital environment

In recent years, with the spread of the Internet, numerous dictionaries containing the Armenian language have appeared in it. But such dictionaries are mostly copies of paper dictionaries, which are usually bilingual with different professional orientations [1]. The large number of such dictionaries is not easy to use; in addition, the ease of their juxtaposition on the Internet already reveals uncertainties [2-5]. For example (Fig. 1) when we switch from language A to language B (using A → B transition vocabulary) we get $B_1,\ldots, B_n$ translations for any A_0 current word. However, when trying to get the translation of the word B_i given in the list back to language A with B → A dictionary, then the derived word in the $A_1,\ldots, A_n$ vocabulary the original word A_0 often is not found. Such disagreements become more evident when we use three dictionaries: A → B, B → C, C → A. In order to solve this problem and not to repeat the mistake of the authors of other dictionaries, this work proposes a new technology for the creation of dictionaries and introduces new principles for organizing them which will minimize such ambiguities. The suggested united dictionary is also suitable for integrating the words of different fields into one common vocabulary.

[*]Corresponding author: anushmelikyan@mail.ru

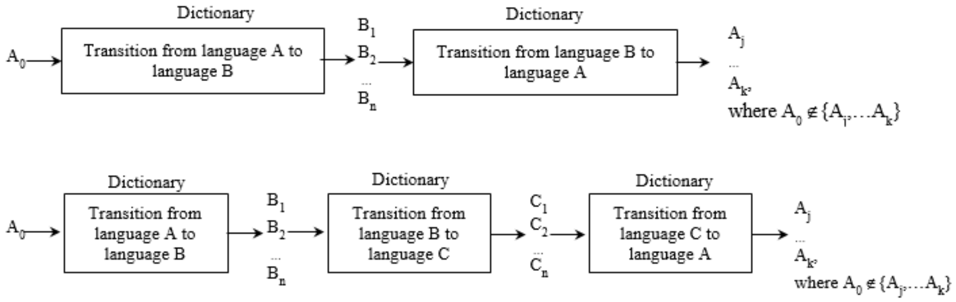

Dictionary
Transition from language A to language B
Dictionary
Transition from language B to language A
A_0
B_1
B_2
...
B_n
A_j
...
A_k,
where $A_0 \notin \{A_j,\ldots A_k\}$
Dictionary
Transition from language A to language B
Dictionary
Transition from language B to language C
Dictionary
Transition from language C to language A
C_1
C_2
...
C_n

Fig. 1. Switching from language A to language B

2 On the new dictionary forming technology adapted to digital environment

2.1 Direct and inverse data storage modes

In this work, any dictionary is recommended to be viewed as a knowledge base [6] where information objects (IO) with their characteristics are stored. In direct storage, each information object in the database is represented by one segment of the information carrier, and the characteristics of that object by a specific field in that segment (Fig. 2).

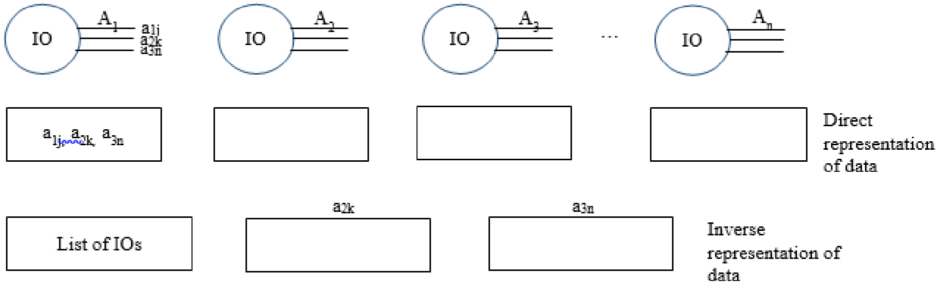


Fig. 2. Representing information objects in direct and inverse storage

During inverse storage, each characteristic is represented by a single note, which contains the list of numbers of information objects with this feature (Fig. 2).

2.2 Direct and inverse forms of digital dictionaries

In the case of organizing digital dictionaries, IO is the word of the language being translated, and the characteristic is the explanation, that expresses some of its meaning. In the case of digitalization of the translation dictionaries, the word expressing the same meaning in the translated language is attached to this meaning. Nowadays digitized classical dictionaries can be viewed as directly organized database [7-8].

This work suggests to organize the dictionary as an inverted knowledge base, that is, its organizational unit is the meaning of the word, any concept of the subject world (hereinafter, concept), and all the words of the translated language are attached to this concept (Fig. 2) are the relevant meaning of the concept [9]. Since the subject world is common to both translation and translated languages, that is to say, concepts defined in one language must be present in the multitude of concepts defined in the second language, then any concept of the knowledge base can be added to its meaning in the list of synonyms of the words spoken in the second language [10].

Now let us consider the process of creating such a dictionary.

Since the decisive element here is the concept, it is defined from the outset by its position and relations in the subject world. The next step is to add to the concept the words that express its meaning in language A, and then the words that express its meaning in language B. That is, when translating the words of A language into B, their mediating concept is activated (Fig. 3).

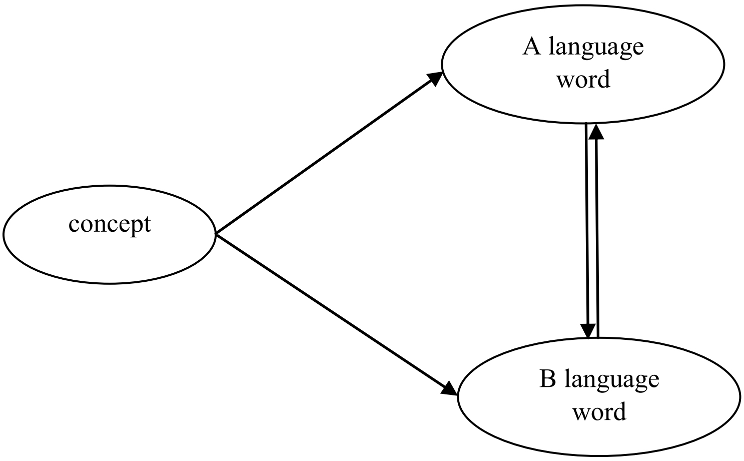


Fig. 3. Activation of mediating concept

In the classical dictionaries, both in the process of formation and operation of the dictionary, the above-mentioned concept is not clearly defined, and this gives rise to errors in bilingual dictionaries, which have already been mentioned at the beginning of this article. Experience shows that the presence of the concept in both knowledge base formation and word translation contributes to the reduction of A → B transition errors. So, the creation of a dictionary starts with the definition of concepts and the creation of their knowledge base. These concepts have some relationship with each other. Concepts with their relationships can be viewed as a conceptual model of the subject world, which is called computer network of knowledge base in informatics.

3 A number of techniques to present words and their meanings in new technology dictionaries in the digital environment

3.1 Network knowledge base as a modern way of organizing explanatory-translation dictionaries

Thus, based on the description mentioned above, the proposed electronic dictionary organized by the ISMA technology is an electronic knowledge base of organizational concepts representing the concepts of the real world [11]. The concepts are identified with their corresponding real-world names, which are presented in the words of the languages provided in the electronic dictionary. One of these languages is considered to be the main one. In our system, Armenian was chosen as the main language, as in the 14 languages included in the dictionary, in our opinion, Armenian had the least degree of multiple-meanings. The concept of interconnections plays an important role in the identification of concepts. They are given during the development of the electronic dictionary. The concepts with their connections can already be viewed as knowledge network base. The formation of this database is done by the linguist who is to act as a real-world expert. He has to analyze real-world objects, form their general-class, and input the organizational unit corresponding to that class - the "concept". He is to identify that unit of the knowledge base, give it its name in the languages used in the dictionary, and integrate their relationship with other units of the knowledge base. Creating interpretation connections, the linguist acts as a knowledge engineer for the given concept, selects the type of neighbor-to-neighbor relationship code and then the associated neighbor concept with the main language words of the dictionary. Physically the concepts are represented by digital tables stored in the external memory of the computer, and the links between them are shown in rows in that table.

3.2 Proposed composition of inter-conceptual relations of the concepts in the knowledge database representing the meanings of words.

When designing the knowledge network database at the base of the dictionary, we offer a set of links between the concepts of the network, which we have divided into two groups: 1` links that represent the semantic relationships between the concepts of the network; 2` links that represent the syntactic relationships of the names that represent the concepts of the network.

Table 1. Links between the concepts of the network: the meanings of words

	Contact Relationship Name	Explanation of the relation to the “concept” K_i with the given “concept” K_0
1	Class-Type	K_0 represents the class of concepts that comprises K_i
2	Part- totality	K_0 –represents the “concept” which is the component of th given K_i “concept”
3	Sphere	K_0 represents the sphere of subject world which includes the given K_i “concept”
4	Period of time	K_0 represents the period of time, during which the given K_i “concept survives and operates
5	Cause	K_0 –represents the cause of the given K_i “concept”
6	Consequence	K_0 –represents the consequence of the given K_i “concept”
7	Source	K_0 represents the material-body from which the material-body represented by the given K_i “concept” is given
8	Result	K_0 - represents the material-body derived from the material-body represented by the given K_i “concept”
9	On	K_0 represents the “concept” (object) on which the given K_i “concept” is located
10	In	K_0 represents the “concept” (object) within which the given K_i “concept” is located

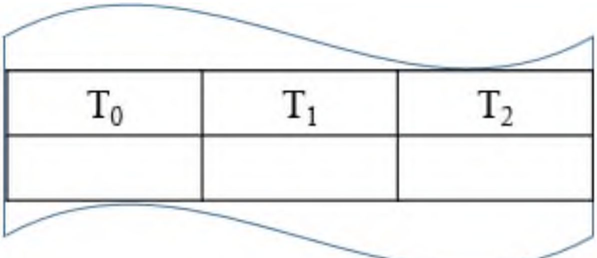
3.3 Proposed composition of inter-conceptual relations of the concepts of the knowledge database representing the syntax of words.

Table 2. Links between the concepts of the network: the syntax of words

	The name of the syntactic relation	**The explanation of the syntactic relation to K_0 "concept" in the given K_i "concept"**
1	Predicate (subject)	K_0 represents the verb for which the subject is the "concept" represented by the given K_i
2	Direct object	K_0 represents the verb for which the direct object is the "concept" represented by the given K_i
3	Indirect object	K_0 represents the verb for which the indirect object is the "concept" represented by the given K_i
4	Metaphoric indirect object	K_0 represents the verb for which the metaphoric indirect object is the "concept" represented by the given K_i and which is presented in the text with dative case
5	Partitive object	K_0 represents the verb for which the partitive object is the "concept" represented by the given K_i
6	Metaphoric Partitive object	K_0 represents the verb for which the metaphoric partitive object is the "concept" represented by the given K_i and which is presented in the text with ablative case
7	Attitude object	K_0 represents the verb for which the attitude object is the "concept" represented by the given K_i
8	Object of means	K_0 represents the verb for which the object of means is the "concept" represented by the given K_i
9	Attribute- determined	K_0 represents the object belonging to the given K_i "concept". In the text K_i is presented by genitive case
10	Adjective (attribute)	K_0 represents the noun "concept" for which as attribute can be the given K_i "concept"

4. The organization of multilingual explanatory-translation dictionaries with ISMA technology

It has already been mentioned that the main organizational unit of the described electronic dictionary - the concept - is stored in the form of an electronic table. As there are no principle limitations on their lines in such tables, for multilingual dictionaries multiplicative tables can be arranged, in the lines of which the name of the given concept can be written in a particular language (Fig 4).



In the knowledge base, select the "concept" table that contains the introductory word T_0 – row type,
T_1 – code of language
T_2 – the list of synonyms of the concept in T1 language

Fig. 4. Arrangement of multiplicative tables

The number of concepts included in the dictionary developed by the ISMA technology at http://translator.am/am/index.html# address as of May 6, 2020 is 109,000, and the number of words in the tables by language is as follows:

Table 3. Numbers of words (of a set of languages)

Languages	**Number of words**
Eastern Armenian	154 000
Western Armenian	153 000
Old Armenian (Grabar)	20 000
Hamshen dialect	11 000
English	160 000
French	31 000
German	21 000
Russian	157 000
Turkish	55 000
Zazaki	3 700
Kurdish	19 000
Latin	12 200
Talysh	55 000
Lezghian	23 200

It is easy to notice that the structure of the electronic dictionary developed by ISMA technology differs substantially from that of classical dictionaries and is not suitable as a dictionary for direct operation. For this purpose, the software running the glossary provides a software module, which develops an electronic database upon request, and the information displayed on the screen coincides with the classical structure dictionaries. And the fact that the output of the dictionary is formed by software makes the output format of the dictionary flexible. Thus: The name of the language being translated is given to the entry of the dictionary and the word requiring translation in that language, as well as the names of all the languages that require the translation of that word. That is to say, outputs of the given word are translated with their synonyms in several languages at the same time. This is accomplished by the existence of a software module operating with the algorithm given in Fig. 5.

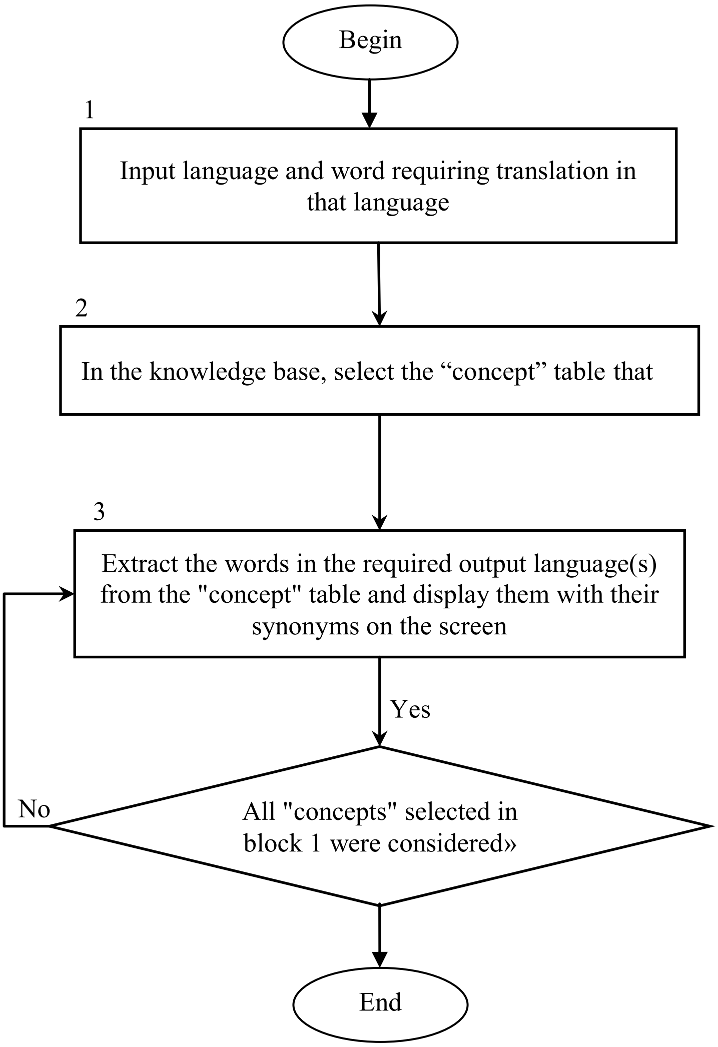
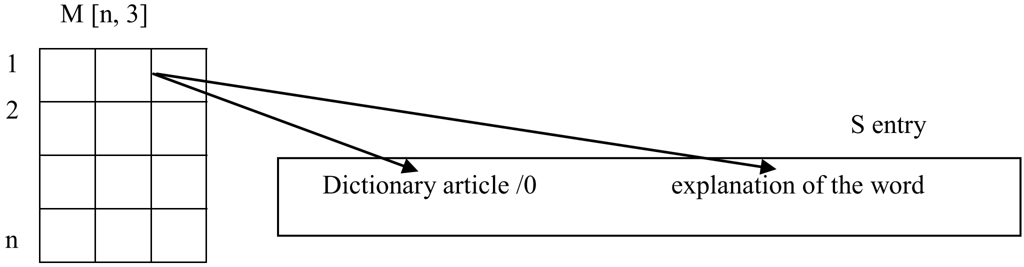

Begin
1
Input language and word requiring translation in that language
2
In the knowledge base, select the “concept” table that
3
Extract the words in the required output language(s) from the "concept" table and display them with their synonyms on the screen
Yes
No
All "concepts" selected in block 1 were considered»
End

Fig. 5. The algorithm of translating a given word with synonyms

5 An expert system for the creation of explanatory dictionaries based on the network knowledge base

In the process of formation of the dictionaries of the given class, the main task is to select the composition of the concepts [12]. Below we suggest an expert system method and scheme which will allow to somehow formalize the process of selecting these concepts. In this work, we proceed from thesis that the basic information needed for concept development is available in the classical explanatory dictionaries. These dictionaries can be viewed as a list of dictionary articles [13]. At present, such dictionaries are digitized and available for software development, as a result of which the digitized dictionary can be presented by a two-dimensional array M and and with S-symbol entry adjacent to it (Fig. 6).



n-is the number of dictionary articles in the explanatory dictionary

Fig. 6. Presentation of the digitized dictionary by two-dimensional array M and adjacent S-symbol entry

M (i, 1) contains a reference to the index of the S entry from which the entry of the i dictionary entry begins. M (i, 2) is the code of the part of speech of the word represented in M (i, 1). And M (i, 3) contains a reference to the index of the S entry from which the text description entry of the given dictionary article begins. If the word is multi-meaning, it corresponds to several lines of M (n, 3) matrices where the values of M (i, 1) coincide. In our opinion the above-mentioned expert system (ES) should be able to rearrange the M (n, 3) matrix lines in such a way that those groups remind the linguist, who creates the explanatory dictionary, of the concept corresponding to that group. In order to insert the matrix line into a specific group, it is necessary to determine the semantic similarity of the text description introduced by the field M (i, 3) with the descriptions of other words in that group. If the meaning represented by the field M of the k1 group M (k1, 3) is too close or coincides with the meaning of another k2 field M (k2, 3), then we can interpreted that we should include a concept in the knowledge base, the name of which will have 2 synonyms represented by the words M (k1.1), M (k2.1). If the meanings presented in the third columns of the group are close but do not match, this suggests that the knowledge base should have concepts that belong to the same subcategory. It will allow the linguist to structure the semantic network of the KB (knowledge base). To evaluate the proximity of the dictionary articles above, it is recommended to use the well-known criteria for proximity assessment of texts used in information search modes of electronic library systems [14]. The latter are represented by matrix R [i, j] element of which is $0 < r_{ij} < 1$, and i and j are the numbers of the corresponding texts. That is, R [i, j] indicates the semantic relationship between the texts. If R [i, j] = 0, then these texts are very close to each other. If R [i, j] = 1, they are very far apart in meaning.

6 The scheme of the organization of the expert system work:

Thus, the problem is to search for groups of words close to each other with the help of R [i, j] matrix. We suggest to divide the n-element keyword groups into two groups at the beginning of the work so that the average value of the intergroup words (P_1) is greater than the links of the words between groups (P_2), then get a new matrix on account of the rearrangement of lines and columns, where the 1st group words will be represented by lines from 1 to m, and from m to n the second group words. Based on the grouping mentioned above we will split R [i, j] (n × n) symmetric matrices into two R1 and R2. In the future, we will apply the approach of dividing the matrix into two matrices R1 and R2 separately. Such a process will continue as long as $P_1/P_2 < \varepsilon$. We recommend using the well-known method of taxonomy theory for dividing the matrices above into two [15], which we will realize with the following steps:

1. We search for the minimum element of matrix R [i, j] (n × n) and then place it in the symmetric zero matrix R'[i, j] (n × n), at the same time we assign all elements to the j-column of the r [i, j] matrix i max value to remove them from further observations.
2. We look for the minimum $R_{i1,\ j1}$ element in the matrix R [i, j], where j1 is one of the nonzero elements of the current moment R '[i, j]. That is, we look for elements of the R '[i, j] matrix from the nearest R [i, j] matrix. We repeat this step until all the elements of the R [i, j] matrix have been considered, that is, until all of the elements have been assigned max value. It is easy to notice that if the R '[i, j] matrix rows are viewed as points in the space and its elements as links between those points, then the image graph in space will have no cycles (Fig. 7).
3. Now we are looking for the element with the maximum value in the matrix R '[i, j] (n × n) and remove it from the view. Since the corresponding graph had no cycles, this process divides the graph into two autonomous subtitles, which allow grouping the lines corresponding to the nodes of those subtitles into a matrix R [i, j] divided by the given R [i, j] matrix into searching matrices R1 and R2.

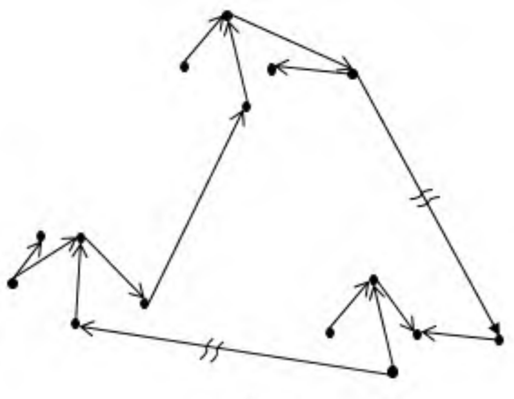


Fig. 7

Fig. 7. Absence of cycles in the image graph in space

4. We form a binary b-tree representing the matrices R1 and R2 obtained above. We assign the above-mentioned R1 and R2 to the terminal nodes of this tree (a hierarchical one-level tree with two subordinate branches). Based on the symbolic record S (described in the 5th section of this article), we form new, truncated, symbolic records S1 and S2, containing the names of dictionary entries corresponding to the rows of matrices R1 and R2, respectively.

5. We sequentially scan the terminal nodes of the current binary tree, if the rank of the matrices is more than two, then steps 3 and 4 of these algorithms are applied relative to it. As a result, the binary tree becomes multilevel and the split submatrices Ri and corresponding to them symbolic records Si are assigned to its nodes, alongside with this symbolic records Si, which contain one or two names of dictionary entries, are assigned to the terminal nodes of this tree.

The expert system contains a dialogue subsystem that sequentially scans all terminal nodes of the resulting multi-level binary tree and, based on the character strings assigned to them, suggests the linguist (knowledge base engineer) to use the content of these records as identifiers of new concepts in the knowledge base or as synonyms to the concepts that already exist. If the proposal of the expert system is accepted by the linguist, then the dialogue system removes the considered terminal nodes and switches to the higher nodes. Thus, as a result of the described dialogue, the linguist introduces into the knowledge vase a new terminal concept or a superior concept, which is associated with the existing class-subclass relationship. The above proposed method of creating and operating an expert system was tested on the example of creating a knowledge base system, which is currently working at Translator.am. The resulting linguistic system serves as an explanatory and translation dictionary for 14 languages (including Russian, English, Armenian, etc.).

7 Summary

The introduction of electronic versions of translation dictionaries and explanatory dictionaries made it possible to organize a comparison of the contents of different dictionaries. The article shows that this process allows one to identify erroneous or questionable solutions proposed by such dictionaries. To avoid such errors, a new technology for creating electronic dictionaries is proposed, which is based on the use of a knowledge base of a network structure. Semantic web nodes supported in such bases are represented by concepts, which are a mapping of abstract and tangible objects of the real world. It is obvious that the composition of such objects is the same for all users of these dictionaries, regardless of nationality. Therefore, while making the optimal choice of concepts, that is the composition of the nodes of the semantic network of the knowledge base, it is possible to assign to these nodes words in different languages that describe the concepts of these nodes. Then, to get a translation of a specific word, it is enough to programmatically activate those nodes to which the given word is assigned, and then, from the activated nodes, read words or phrases describing this node in other languages. During operation, the user gets the impression of using a conventional dictionary. The essence of the described approach is the choice of the composition of concepts and the relationships between them. It is proposed to automate this process using an expert system. At the initial stage of its work, from the digital mappings of explanatory dictionaries (see matrix M and symbolic notation S), it makes up a matrix of incidence (connections) between dictionary entries selected from explanatory dictionaries (see matrix R [i, j]). To determine the semantic relationship between them, it is proposed to use the previously developed methods of documentary search, with the help of which the semantic relationship between the texts of interpretations of individual dictionary entries is determined. Based on the approaches of the theory of taxonomy, this article proposes an algorithm for grouping dictionary entries. As a result, the obtained groups and subgroups indicate possible concepts of the knowledge base, since dictionary entries included in one group can identify the corresponding concept and act as synonyms for this particular concept. The article proposes an interactive (dialog) subsystem that, by processing the incidence matrices, obtains a kind of binary b-tree. Processing the dialogue system of this tree allows one to give the linguist suggestions on the composition of concepts and the structure of the knowledge base being created. Thus, using the proposals of the expert system, the linguist can act as an engineer for knowledge bases and form the semantic network of this base. The described approach was tested in the process of development of an explanatory and translation dictionary for 14 languages, the possibilities of which can be found at translator.am.

References

1. K. S. Toshtemirovich, American Journal of Applied Sciences, July, 126 - 130 (2020) DOI: 10.37547/tajas/Volume02Issue07-20
2. L. Liu, Lingua, **214**, 11-27 (2018) doi.org/10.1016/j.lingua.2018.08.001
3. J. L. Dong, L. Z., Y. H. Chen, W. C. Jiang, *Signal Processing: Image Communication*, **76**, 81 - 88 (2019) doi.org/10.1016/j.image.2019.04.006
4. T.-C. Liu, P.-H. Lin, Computers in Human Behavior, **27(1)**, 373–383 (2011) doi.org/10.1016/j.chb.2010.08.016
5. L. Zhuhadar, Computers in Human Behavior, **51(B)**, 1107-1115 (2015) doi.org/10.1016/j.chb.2015.03.021
6. Y. Zhanga, M. Ye, Y. Gan, W. Zhang, Knowledge-Based Systems, **193** (2020) doi.org/10.1016/j.knosys.2019.105444
7. M.-C. L'Homme, M. C. Cormier, International Journal of Lexicography, **27(4)**, (2014) DOI: 10.1093/ijl/ecu023
8. S. Rajan, K. R. Soumya, *Procedia Technology*, **25**, 272-279 (2016) doi.org/10.1016/j.protcy.2016.08.107
9. L. Zou, K. Pang, X. Song, N. Kang, X. Liu, Information Sciences, **524**, 165 – 183 (2020) doi.org/10.1016/j.ins.2020.03.002
10. N. Guan, D. Song, L. Liao, Knowledge-Based Systems, **164**, 38–44 (2019) doi.org/10.1016/j.knosys.2018.10.008
11. J. Han, X. Teng, X. Tang, X. Cai, H. Liang, Knowledge-Based Systems, **195**, 105701 (2020) doi.org/10.1016/j.knosys.2020.105701
12. S. Khamroeva, Theoretical & Applied Science, **87(07)**, 463–466 (2020) DOI: 10.15863/TAS.2020.07.87.88
13. G. Dima, *Procedia - Social and Behavioral Sciences*, **63**, 93-98 (2012), doi.org/10.1016/j.sbspro.2012.10.016
14. V. Chernenkiy, Y. Gapanyuk, V. Terekhov, G. Revunkov, Y. Kaganov, *Procedia - Computer Science*, **145**, 143-152 (2018) doi.org/10.1016/j.procs.2018.11.022
15. M. Jeon, *Emotions and Affect in Human Factors and Human-Computer Interaction, Chapter 1 - Emotions and Affect in Human Factors and Human–Computer Interaction: Taxonomy, Theories, Approaches, and Methods*, 3 – 26 (2017) doi.org/10.1016/B978-0-12-801851-4.00001-X