***Мотыгуллина Зухра Айвазовна** –*
кандидат филологических наук,
доцент
Казанский федеральный университет
E-mail: zuhra711@yandex.ru

***Motygullina Zukhra Aivazovna** –*
Doctor of Philology,
Ass.Professor
Kazan Federal University
E-mail: zuhra711@yandex.ru

***Сайдашева Эльмира Аглиулловна** –*
кандидат филологических наук,
доцент
Казанский федеральный университет
E-mail: esaidasheva@rambler.ru

***Saidasheva Elmira Agliullovna** –*
Doctor of Philology,
Ass.Professor
Kazan Federal University
E-mail: esaidasheva@rambler.ru

Поступила в редакцию 10.01.2019.
Принята к публикации 15.02.2019.

УДК 81

ПРИНЦИПЫ РЕЗЮМИРОВАНИЯ БОЛЬШИХ ОБЪЕМОВ ДАННЫХ: ОСОБЕННОСТИ ОТБОРА ЯЗЫКОВЫХ СРЕДСТВ

Л.Р. Сакаева, К.Р. Родригес
charlyr98@gmail.com

Казанский (Приволжский) федеральный университет, г. Казань, Россия
Университет информационных наук, г. Гавана, Куба

Аннотация. Естественный язык является основным способом человека для общения и понимания явлений и процессов. Ускоренное внедрение информационных и коммуникационных технологий во все сферы жизни общества способствует формированию больших объемов данных, содержащих слова, коды, символы и цифры. Однако первоначально лишь небольшая часть этих данных может быть понята и использована людьми. Существуют вычислительные методы, которые, используя методы мягких вычислений и естественный язык, строят резюме в виде предложений, описывающих большие объемы данных. Целью данной работы является краткое описание метода, используемого при построении лингвистических резюме, с акцентом на использовании естественного языка для предоставления важной и полной информации о данных. Во-первых, представлен общий обзор лингвистического резюмирования. Затем анализируются два примера применения этого метода к различным базам данных. В обоих случаях получаются короткие предложения на естественном языке, содержащие суммированную информацию о данных. Во втором примере бенефициары показали индекс удовлетворенности по логической схеме Ядова на уровне 0,7.

Ключевые слова: большие объемы данных, естественный язык, лингвистическое резюмирование, слово.

Для цитирования: Сакаева Л.Р., Родригес К.Р. Принципы резюмирования больших объемов данных: особенности отбора языковых средств. *Казанский лингвистический журнал*. 2019; 2 (1): 62–75. (на англ.)

PRINCIPLES OF SUMMARIZATION OF LARGE AMOUNTS OF DATA: FEATURES SELECTION OF LANGUAGE MEANS

L.R. Sakaeva, C.R. Rodríguez

charlyr98@gmail.com

Kazan Federal University, Kazan, Russia

University of Informatics Sciences, Havana, Cuba

Abstract. Natural language is the main resource of people to communicate and understand phenomena and processes. The accelerated introduction of information and communication technologies in all fields of society promotes the generation of large data volumes which contains words, codes, symbols, and numbers. However, originally only a small part of this data can be understood and used by people. There are computational approaches that, using soft computing techniques and natural language, construct summaries in form of sentences which describe large volumes of data. The purpose of this paper is to briefly describe a method used in linguistic summaries construction, emphasizing the use of natural language to offer relevant and comprehensive information about data. First, a background on the linguistic summarization of data is presented. Then two examples of the application of this technique to different databases are analyzed. In both cases, short sentences in natural language are obtained that offer summarized information about the data. In the second example, the beneficiaries showed a satisfaction index of 0.7 according to Iadov's logical framework.

Keywords: large data volumes, linguistic data summarization, natural language, word.

For citation: Sakaeva L.R., Rodríguez C.R. The use of words in linguistic summarization of large data volumes. *Kazan Linguistic Journal*. 2019; 2 (1): 62–75.

People communicate essentially using oral and written language. This could be interpreted in general terms as people communicate preferably using words. But not only with words do people communicate. In most of daily life processes (in economics, politics, science, health, education, etc.) they are also used acronyms, codes, symbols, and many numerical data.

The accelerated introduction of information and communication technologies in all fields of contemporary society has led to the generation of large volumes of data. These data are stored in the technological infrastructures of the organizations that generate them or in those of the global consortiums of telecommunications services (AT & T, Google, Sprint, etc.). In 2007, several consulting companies and software evaluators released a list of the largest databases to date. The biggest database among non-telecommunication organizations is the database of the World Climate Data Center with 220 terabytes (TB). In the field of telecommunications companies, the first place is for AT & T with 323 TB. One TB equals 1024 gigabytes (GB).

At the organizational level, these data volumes are very invaluable assets for their owners. However, only a small part of this data is originally able to be understood and used by people. Some of its reasons are the following:

- its volume, i.e., the number of records and attributes of those records,
- its type that can be numerical, ordinal, nominal,
- the existence of missing values, i.e., some records have missing values of some of their attributes,
- the ambiguity, this is presented in several ways, for example, to use indistinctly the words: tall, large, slender, to indicate that a person is high.

Some researchers name this situation as "rich data, poor information" phenomenon. This condition is a loss of opportunity for organizations because the correct interpretation of these data is an ability that affects the decision-making process and selection strategies.

There are computational approaches that, using soft computing techniques and natural language, construct summaries in form of sentences which describe large volumes of data. These approaches are based on the fact that words are the main resource of people to communicate and understand phenomena and processes. It is an active research topic in the scientific community. It has been developed and derived in many different areas such as Linguistic Data Summarization and Linguistic Data Description. A recent it be merged with big data. A new study

[1, p.165] returned the following results (see Table 1) about the presence of the topic "Linguistic Data Summarization" in scientific publications during the last five years.

Table 1

Presence of the topic "Linguistic Data Summarization" in scientific publications during the last five years

"Linguistic Data Summarization"	IEEE	Thomson Routers	Google Scholar	Semantic Scholar	SCOPUS	Library Genesis	Microsoft Academic
Last 5 years	62	45	22	7	5	2	1

The objective of this paper is to briefly describe an approach to construct linguistic summaries from databases, emphasizing the use of natural language to offer relevant and comprehensive information about the data.

Linguistic data summarization (background)

Initially, it is necessary to present two definitions that will help to understand this technique.

Knowledge Discovery in Databases (KDD) is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable structure in data [2, p.33].

The idea of linguistic summaries is to enable the summarization of a collection of data using linguistic terms and thereby making the information provided more comprehensible. Furthermore, the focus of these summaries should involve terms and ideas that are relevant to the goals and objectives of the recipient of the summary [3, p.273].

Linguistic Data Summarization (LDS) is considered a KDD approach to extract patterns from information stored in databases. LDS allows capturing and briefly describe the trends and characteristics that appear in a data set. Its use is especially feasible because it provides summaries that are not as concise as the average, median, variance, etc. and allows the processing of numerical data or not.

One of the most extensive approach to LDS is based on fuzzy logic. Fuzzy logic allows to manage the uncertainty about the degree of membership of a value to more than one class. These classes are associated with a linguistic term and are modeled by fuzzy sets. A fuzzy set has a membership function in a specified interval. The use of several fuzzy sets to model the behavior of an attribute is called linguistic variable (see Fig 1, it is linguistic variable “age”).

A linguistic summary like this approach can have the following structure:

QBR’s are S

Where:

S is the summarizer (it is a linguistic expression that determines a subset of objects within the database, is the focus of the summary, the task is to determine how many objects that fit the rest of the sentence satisfy him).

Q is a quantifier (a linguistic term that indicates data range -the quantity- that satisfy S).

R is the family of objects that are analyzed (the records in database).

B is a qualifier (a linguistic expression that determines a subset of objects within the database, is optional, can use as many as needed show details).

For example:

“Most of well-paid workers are adults”

The main characteristic of this formulation is that the *summarizer* and *quantifier* are expressed in linguistic terms. Using fuzzy sets these linguistic terms are provided with a formal semantic. The qualifiers can be nominal or represented by fuzzy sets. One advantage of the use of linguistic summaries is that they provide statements about the dataset in terms that a very easy for people to comprehend. These summaries have a final component that is the degree of validity of the proposition. It takes values between [0,1]. But due to its computational character we will not abound in its calculation form.

The process of constructing linguistic summaries consists of several steps that include from the selection of the data to be processed to the interpretation of the results. Although there are different methodologies for this, a common framework can be identified. Next, the activities of this framework are presented without discussing their computational peculiarities.

Selection of the data set: As its name suggests, it consists in establishing data which will be worked. This decision will depend on the specific interests that motivate the analysis and the real availability of data.

Pre-processing of data: It consists of making a first treatment of the completeness and ambiguity of the data. Techniques can be used to try with incorrect and missing data.

Transformation and reduction of data: The data is modified to make it more useful. This activity may include the change of continuous data into discrete data using fuzzy sets. Here the words begin to participate since the linguistic terms associated with each fuzzy set must be properly defined. Also, the data can be integrated, restructured, etc. Some techniques can use to determine the most relevant attributes and thus reduce the size of the database.

Data Mining (Linguistic Data Summarization): At this point, the data is ready to be mined. Therefore, the linguistic expressions that will be used for the summarizers, the quantifiers, and the qualifiers must be defined.

Examples of these elements can be:

- Summarizes and qualifiers: for attribute age (“child”, “adolescent”, “young”, “adult”).
- Quantifiers: “around 5” (relative quantifier), “most” (absolute quantifier).

After these elements are defined, the computational algorithms that analyze the entire database and build all the possible summaries are developed. In this phase, an initial value of validity of each summary is also calculated.

In *Transformation* and *Data Mining* steps, the use of words has a very important role. Because it is here where all components of linguistics summaries are determined.

Interpretation / evaluation of extracted knowledge: Finally, other quality measures of the summaries are calculated (the degree of imprecision, the degree of coverage, the degree of appropriateness, length of the summary). With them, the users according to their specific interests determine the definitive validity of each summary. Once the summaries are evaluated, they can be used for decision making.

Application examples

Now we will see two applications of this technique to summarize two different databases. The first example was taken from the work of some of the most prestigious authors on the subject of Linguistic summarization. The second example was taken from a research report developed by the author in 2017.

Example 1: Sales in a computer store (taken from [4, pp.298-300])

A database of sales of a computer equipment and accessories store is used, whose basic structure is the Table 2. Initially, the authors discover interesting relationships between commission and types of goods sold (see Table 3). Then they found the relationships between the groups of products and the date of sales (see Table 4). In the second case, the summaries represent less obvious relationships than those found before.

Table 2

Basic structure of the database example 1.

Attribute name	Attribute type	Description
Date	Date	Date of sale
Time	Time	Time of sale transaction
Name	Text	Name of the product
Amount (number)	Numeric	Number of products sold in the transaction
Price	Numeric	Unit price
Commission	Numeric	Commission (in %) on sale
Value	Numeric	Value = amount (number) x price; of the product
Discount	Numeric	Discount (in %) for transaction
Group	Text	Product group to which the product belongs
Transaction value	Numeric	Value of the whole transaction
Total sale to customer	Numeric	Total value of sales to the customer in fiscal year
Purchasing frequency	Numeric	Number of purchases by customer in fiscal year
Town	Text	Town where the customer lives

Table 3

Relations between the group of products and commission

Summary
About 1/2 of sales of network elements is with a high commission
About 1/2 of sales of computers is with a medium commission
Much sales of accessories are with a high commission
Much sales of components are with a low commission

Продолжение Table 3
About 1/2 of sales of software is with a low commission
About 1/2 of sales of computers is with a low commission
A few sales of components are without commission
A few sales of computers are with a high commission
Very few sales of printers are with a high commission

Example 2: Criminal process in a procuracy's office [5, p.10]

There is a criminal process database that are handled in a procurator's office. Specifically, data from 387 ordinary processes and 193 testified complaints are used. The structure of the database is shown in Table 5.

Table 4

Relations between the groups of products and times of sale

Summary
About 1/3 of sales of computers is by the end of year
About 1/2 of sales in autumn is of accessories
About 1/3 of sales of network elements is in the beginning of year
Very few sales of network elements are by the end of year
Very few sales of software are in the beginning of year
About 1/2 of sales in the beginning of year is of accessories
About 1/3 of sales in the summer is of accessories
About 1/3 of sales of peripherals is in the spring period
About 1/3 of sales of software is by the end of year
About 1/3 of sales of network elements is in the spring period
About 1/3 of sales in the summer period is of components
Very few sales of network elements are in the autumn period
A few sales of software are in the summer period

Table 5

Basic structure of the database example 2

Attribute name	Attribute type	Description
complaint number	Character varying	The number that uniquely identifies the complaint
military procuracy	Boolean	Specifies if the process is attended by the military prosecutor's office
date of event	Date	Date on which the event occurred
event time	Time without	Time at which the event occurred
police station	Character varying	Police station where the complaint originated
file number	Character varying	The number that uniquely identifies the file
start date	Date	Start date of the process

Продолжение Table 5

closing date	Date	Date of conclusion of the process
it is prioritized	Boolean	Specifies whether the process is addressed with priority or not
instruction body	Character varying	Instruction body that instructs the process
id of case	Numeric	The number that uniquely identifies the case
id of person	Numeric	The number that uniquely identifies the person
year of birth	Date	Year of birth of the person involved in the process
Race	Character varying	The skin color of the person involved in the process
Sex	Character varying	Gender of the person involved in the process
marital status	Character varying	Marital status of the person involved in the process
id of process	Numeric	The number that uniquely identifies the process
Procuracy	Character varying	Name of the Office of the Procuracy that deals with the process
process type	Character varying	Denomination of type of process (Ordinary, etc.)
procurator name	Character varying	Name of the procurator who attends the process
kind of person	Character varying	Designation of the role that the person has in the process

Database have 21 attributes. Nine of them were discarded, and 11 were kept to build the summaries. Four attributes were transformed to increase their usefulness.

- The attribute *year_of_birth*: it is a date of the form year-month-day. This attribute becomes the numeric attribute *age*, which was then discretized.
- The attributes *start_date* and *closing_date*: they were merged to create the numerical attribute *process_execution_time*, which was then discretized.
- The attribute *date_of_event*: it is a date of the form year-month-day. This attribute becomes the nominal attribute *occurrence_period*.

Then the attributes: *age* (Fig. 1), *process_execution_time* (Fig. 2) and *event_time* (Fig. 3) were discretized. Trapezoidal fuzzy sets were used for this process.

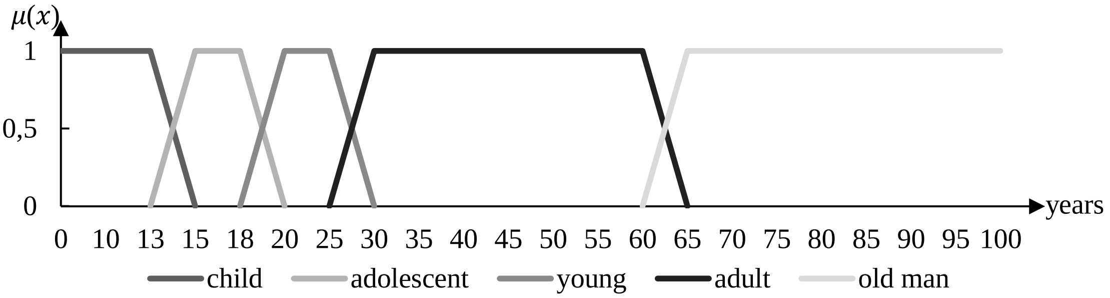

Fig 1. Linguistic variable “age”

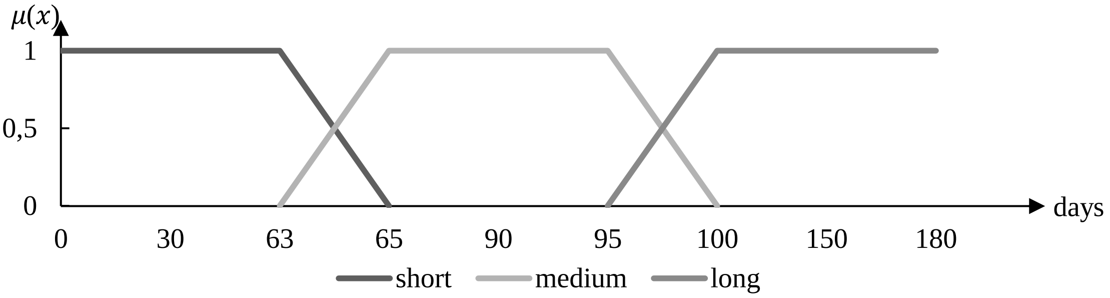

Fig 2. Linguistic variable “process_execution_time”

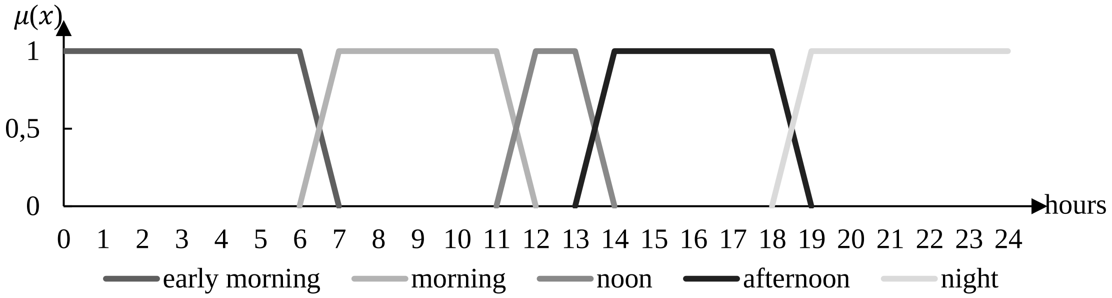

Fig 3. Linguistic variable “event_time”

The following summarizers and qualifiers were defined. Note that all of them are represented by words or phrases.

Summarizers:

- process execution time = {“short”, “medium”, “long”}
- age = {child, adolescent, young, adult, old man}
- hours = {“early morning”, “morning”, “noon”, “afternoon”, “night”}
- type of process = {“ordinary”, “testified complaint”}
- occurrence period = {“1st quarter”, “2nd quarter”, “3rd quarter”}

Qualifiers:

- race – take the values: “black”, “white” or “mestizo”.
- sex – take the values: male or female.

- marital status – take the values: “single”, “married”, “divorced” or “widowed”.
- is prioritized – take the values: “true” if the case is prioritized, “false” otherwise.
- military procuracy – takes the values: “true” if the process is carried out by the Military Procuracy's Office, “false” otherwise.

Finally, fuzzy (Fig. A) and non-fuzzy (Fig. B) quantifiers were defined.

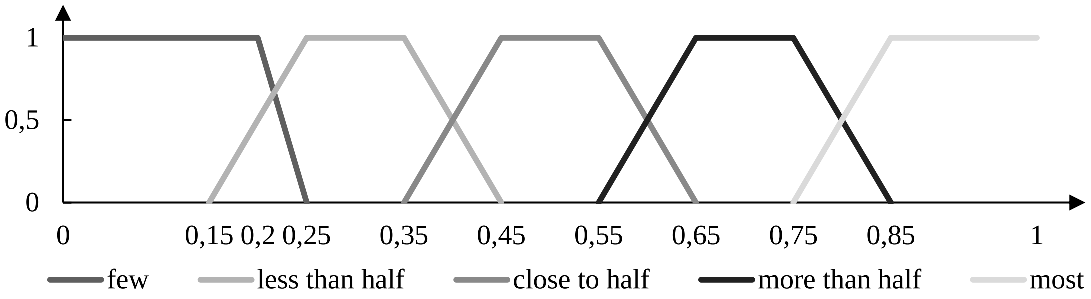

Fig 4. Fuzzy quantifiers

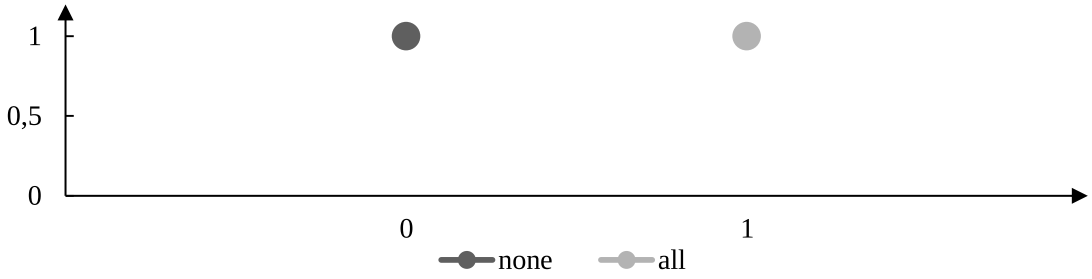

Fig 5. Non-fuzzy quantifiers

Table 6 shows some of the summaries obtained. In them can be observed the relationships between several of the attributes of the database. In this case, the workers of the prosecutor's office were asked about the usefulness of the linguistic summaries obtained. A modification of Iadov's technique to assess satisfaction [6, p.45] was used. As a result, a result of 0.7 was obtained, which represents satisfaction according to the value scale of this technique. They expressed that summary information could be used for different purposes, e.g., to develop social prevention actions, to evaluate the performance and to manage the work of the organization.

Table 6

Some summaries obtained

Summaries
More than half of processes of testified complaint that not carried out by the Military Prosecutor's Office occur during the 1st quarter year.
About half of processes of testified complaint that not carried out by the Military Prosecutor's Office occur in the afternoon.
All prioritized Ordinary processes are short duration.
Most of the prioritized Ordinary processes involve adults.
Less than half of single people are involved in process of testified complaint.

As the main conclusions we present three aspects:

- The processing of large volumes data is a challenge and an opportunity for process analysis and decision making.
- Linguistic data summarization allows processing large data volumes and showing the summarized information in natural language.
- Tasks of linguistic data summarization require the participation of a multidisciplinary team integrated by specialists in application area, computer engineers, linguists, etc.

Литература

1. *Кузьмина Н.В.* Методы исследования педагогической деятельности. Ленинград: Изд-во ЛГУ, 1970.

2. *Pérez I., Santos O., García R., Piñero P., Ramírez E*. Discovering linguistic summaries for help in project management // Cuban Journal of Computer Science. Vol. 12. Uciencia, 2018. Pp. 163–175.

3. *Fayyad U.* Knowledge Discovery in Databases: An Overview // Relational Data Mining. Berlin, Heidelberg: Springer, 2001. Pp. 28–47.

4. *Yager R.R., Yager R.L.* Using linguistic summaries and concepts for understanding large data // Engineering Applications of Artificial Intelligence. No 56. 2016. Pp. 273-280.

5. *Kacprzyk J., Zadrożny S.* Queries with Fuzzy Linguistic Quantifiers for Data of Variable Quality Using Some Extended OWA Operators //Advances in Intelligent Systems and Computing. Berlin: Springer. Vol 400. 2016. Pp. 295–305.

6. *Rodríguez C.R.* Construction of linguistic summaries of data from criminal processes. Research report (unpublished). Computers and Law Lab at University of Informatics Sciences. Havana, Cuba, 2017. 15 p.

References

1. Pérez I., Santos O., García R., Piñero P., Ramírez E. (2018). *Discovering linguistic summaries for help in project management* // Cuban Journal of Computer Science. Vol. 12. Uciencia. Pp. 163–175. (In English)

2. Fayyad, U. (2001). *Knowledge Discovery in Databases: An Overview* // Relational Data Mining. Berlin, Heidelberg: Springer. Pp. 28–47. (In English)

3. Yager, R.R., Yager, R.L. (2016). *Using linguistic summaries and concepts for understanding large data* // Engineering Applications of Artificial Intelligence. No 56. Pp. 273–280. (In English)

4. Kacprzyk, J., Zadrożny, S. (2016). *Queries with Fuzzy Linguistic Quantifiers for Data of Variable Quality Using Some Extended OWA Operators* // Advances in Intelligent Systems and Computing. Vol 400. Cham: Springer. Pp. 295–305. (In English)

5. Rodríguez, C.R. (2017). *Construction of linguistic summaries of data from criminal processes*. Research report (unpublished). Computers and Law Lab at University of Informatics Sciences. Havana, Cuba, 15 p. (In English)

6. Kuzmina, N.V. (1970). *Metody issledovania pedagogicheskoi diiatelnosti* [Methods of pedagogical activity research]. Leningrad: Izd-vo LGU. (In Russian)

Авторы публикации

Сакаева Лилия Радиковна *– доктор филологических наук, профессор*
Казанский федеральный университет
E-mail: liliyasakaeva@rambler.ru

Authors of the publication

Liliya Radikovna Sakaeva *– Doctor in Philological Sciences, Professor*
Kazan Federal University
E-mail: liliyasakaeva@rambler.ru

Родригес Родригес Карлос Рафаэль *–*
магистр в области управления ИТ-проектами, доцент
Университет информационных наук,
Гавана, Куба
E-mail: charlyr98@gmail.com

Carlos Rafael Rodríguez Rodríguez *–*
Master's Degree in IT projects management,
Associated Professor
University of Informatics Sciences
Havana, Cuba
E-mail: charlyr98@gmail.com

Поступила в редакцию 21.01.2019.
Принята к публикации 20.02.2019.

УДК 811

СЛОВОСЛОЖЕНИЕ, КОНВЕРСИЯ И АББРЕВИАЦИЯ КАК СПОСОБЫ ОБРАЗОВАНИЯ ХИМИЧЕСКОЙ ТЕРМИНОЛОГИИ В АНГЛИЙСКОМ И ТАТАРСКОМ ЯЗЫКАХ

В.Н. Хисамова, Л.М. Ибатулина
hisamovaven@yandex.ru, lucide@list.ru

Казанский федеральный университет, г. Казань, Россия

Аннотация. Терминология занимает важное место в науке. Без владения терминологией невозможно быть специалистом ни в одной из научных сфер жизни. Быстрое развитие науки ведет к появлению новых терминов. Таким образом, систематизация и стандартизация химической терминологии является актуальным вопросом на сегодняшний день. В данной статье рассматриваются способы словообразования в английской и татарской химической терминологии на примере словосложения, конверсии и аббревиации. Выбор данных способов терминообразования обусловлен их высокой продуктивностью. Цель настоящего исследования состоит в выявлении особенностей данных способов словообразования в химической терминосистеме двух генетически неродственных и типологически разноструктурных языков и определении их языковой репрезентации в академических текстах. Рассмотрены и представлены общие, характерные для исследуемых в данной статье языков, модели словосложения. Изучение вопроса о использовании конверсии показывает различия в моделях, являющихся продуктивными для каждого из языков, представленных в настоящей работе. При изложении материала о таком способе, как аббревиация, мы взяли во внимание только те модели, которые являются общими для английского и татарского языков.

Ключевые слова: словосложение, конверсия, аббревиация, термины, химическая терминология, английский язык, татарский язык.