**Afruz M. Gurbanova**
Institute of Information Technology of ANAS, Baku, Azerbaijan
afruz1961@gmail.com
DOI: 10.25045/jpit.v09.i2.10

# ANALYSIS OF AUTOMATION METHODS OF THE TERM CREATION ACTIVITY

*The article deals with the issue of terminology, which is a consequence of conscious influence of society on the development of language. It is highlighted that vocabulary of any language is updated due to both internal and external resources, and a life cycle of a word depends on its usage. The principles of harmonization of terminology are defined. The process of determinologization of concepts is studied and the reasons for this process are determined. The role of special and common words of the lexicon in the determinologization process is noted. The methods of calculating the frequency and weight coefficients of terms in texts are studied and analyzed.*

***Keywords:*** *language policy, terminology creation, harmonization of terms, determinologization, weight factor of terms.*

## Introduction

The lexical composition of the language is very sensitive to the changes in society. The volume of the linguistic vocabulary enhances its functional capabilities. Vocabulary of language is renewed not only through the generation of new features, but also at the expense of internal resources, for example, inclusion of dialect words in literary language, giving different meanings to the words in the spoken language, or bringing back previously-used outdated lexicon, and at the expense of international words [1].

Regardless of the generation methods, the word must be adopted and welcomed by the language carriers in order to provide "citizenship right", in other words, the word must be regularly used in a particular context. Thus, the term must be compatible with the selection criteria: grammatical features of the term, its accuracy, transparency, laconicism, reliability, pronunciation and ease of use in speech, and the opportunities for creating its derivatives and so forth.

An interesting point is that foreign words are “safer” in a language. This is due to the fact that the native speaker "does not trust" the words in own language and believes that they do not represent a certain event, object, and so forth. This, in turn, may be a sign of low level of language competence of the language carrier.

Language policy issues include the regulation of terminology in language. In this regard, governments establish terminological commissions.

After the Azerbaijani language was given a state status in 1995, a new stage in the terminology work, which is one of the fields of linguistics, has begun [2]. A large number of terminological dictionaries have been developed in different fields. Some of the terms are based on internal resources, a part of which are loan words. Loan words are not challenging as they should not exceed certain standards. The terminological development of the language should be adjusted in this regard, and the process of term creation should not be stopped, which is implemented at the expense of internal resources. When determining the position of the foreign word in the native language, its position in the global linguistic environment should be taken into account. The availability of a certain number of words in in many other languages makes it even easier to be learnt.

As in other languages, the most convenient way to create and enrich the terminology in Azerbaijani is the creation of terminology and terminological word-combinations based on the internal capabilities of the language. Thus, when creating the terms for the expression of new concepts in different fields of science and technology, it is necessary to review the vocabulary of the language and to use its internal resources effectively. From this point of view, intra-linguistic opportunities, that is lexical, morphological and syntactical term creation have always dominated

in the enrichment of the terminology of the Azerbaijani language, which has a long-term development history [3].

New notions are generated due to the changes and innovations in the society have to be called using both internal and external resources.

**Principles of terminology harmonization**

Traditionally, terms are selected from the texts, systematized and published as a book after the examination by scientists and specialists (terminologists). Accessibility of information flood through the Internet has made this job even more difficult.

On the other hand, the work of term-building has been significantly simplified compared to the previous periods. Note that [4] offers a conceptual model of the Azerbaijani terminology information system and analyzes the possibilities of the system. When creating a new term, each scientist and specialist can refer to the Terminology Registry of the National Terminology Information System (NTIS) to find out whether this word has been previously registered as a term. If the searched term is in the same registry, then there is no need to adopt a new term. Otherwise, this term is searched and analyzed in the integrated NTISs of other countries, and more efficient decision is made, in other words, international harmonization is realized.

In modern era, strengthening international cooperation in the field of science, culture and economy requires the development of harmonization implementations of terminology of more developed national languages. The development of the principles of terminology harmonization is considered to be an important part of the tasks of the international co-operation of terminologists.

The scheduled internationalization of the terms, which are an integral part of the harmonization, should be fulfilled, in other words, precise conformity between the terms must be defined, and the meaning of the terms in various languages should be selected out of the synonyms of the internationalized terms, which are close to each other in terms of definition and form, and approved.

Terminology harmonization consists of the following steps:

1. Systematic comparison of national terminology and terminology systems;
2. Building a classification scheme, taking into account all the concepts represented in the compared national terminology;
3. Developing an agreement on the identification and determination of the unambiguous understanding of the equivalent national terms;
4. Internationalization.

The harmonization of national and internationalized notions and the terminology systems they are represented in is aimed at the development of a single technical standardization language in certain languages. The harmonization of terminology systems considers two factors:

- linguistic (related to the features of the language the term is referred to);
- extralinguistic (related to the subject field and the theory the term is described in).

Extralinguistic factors of the international harmonization of terminology systems are as follows:

- Integration of knowledge;
- internationalization of science and technology;
- theoretical and methodological unification of science and technology;
- distinctive to the modern era of the world civilization.

Linguistic factors of the international harmonization of terminology systems include:

- formation of languages for specific goals;
- internationalization processes of the terms that are the lexical unit of a language for specific purposes.

The principles of international harmonization of terminology systems are possible in the following cases:

- It is feasible and appropriate if two or more countries develop the same knowledge or field

of activity;
- It is feasible if the fields of activity in these countries are based on the same or similar theories with the system of same or closely related concepts.

The application of the terminology systems results in:
- international terminology dictionaries and standards;
- multilingual information-search thesaurus.

The technical committee of the International Organization for Standardization (ISO/TC37/SC1) developed the standard ISO 860: 2007 Harmonization of concepts and terms [5]. This standard specifies a methodological approach to the harmonization of concepts, conceptual systems, definitions and terms. It is applied to the development of the terminology harmonized in national or international level in bilingual or multilingual context.

**Determinologization of concepts**

It should be noted that it is possible to obtain supplementary information on terminology (etymology, definition, comments, etc.) via the Internet. Modern technologies allow presenting the terms not only as a text but also though other multimedia descriptions (photo, audio, video).

The accessibility of information and its broadcasting through social media creates great opportunities for the propagation of knowledge. Therefore, relatively rarely used concepts (terms) are also rapidly determinologized (as opposed to the terminologization).

As a result of the "stimulation of information and communication processes" during the scientific-technical revolution, a serious increase of terms in various fields of knowledge and their inclusion in the general lexicon was predicted. At that time, this process was called "terminological explosion". In the present globalization era, a "specific terminological expanse" occurs, which is defined as a tendency of the "intellectualization of the lexicon" by some linguists referring to the increase of the communicative role of terminology [6].

In the modern world, a special lexicon is often used "on a daily basis in a single semantic area" and subject to permanent determinologization.

The definition of the concept of determinologization has not been resolved so far. Determinologization is the loss of meaning of terms being excluded from a terminology system and converted into a common word. Determinologization in the developed national language is one of the main ways of enriching the nominative system of the language, and represents the interaction of specific and general lexicon. Determinologization represents the semantic and functional variation of the linguistic substrate (sub-layer) of the term. Functional determinologization is the loss of terminological function of the term. Whereas the semantic determinologization occurs when the term loses its terminological essence and its new semantics emerges.

The general lexicon is constantly enriched with terms, which is explained by the relevance of the certain concepts in Azerbaijani language, as in other languages. Determinologization, as an intellectualization method of communicative processes, provides an understanding of the recognition and enrichment of information shape of the world.

The determinologization process can be analyzed and evaluated automatically by determining the frequency of the terms in the public domain. It should be noted that the terminologization and determinologization indicators of the concepts depend on the usage areal of the language, i.e., the population of the country.

Terminologization is the process of converting an ordinary and daily used word into a term, while the inverse process is the inclusion of a term in commonly used word. This process was analyzed by Ingrid Meyer and Kristen Macintos in 2000 in their studies on terminology and was defined as determinologization [7].

They described two categories of determinologization:

1. The term maintains and preserves its correct meaning, however is not used by the field experts. In such cases, the thing may become popular, and its concept is understood by

people well enough. For example, as the medical terms are broadly used, and everyone is aware of them.

2. The word now describes completely different concept.

Terminology is a dynamic and constantly changing system. Thus, in this system, the lifecycle of terms, concepts and definitions ends. They go out of use, and are abolished as outdated and unscientific.

As in other languages, in the Azerbaijani language, several periods (such as scientific and technical revolution) of acquisition of lexical units from terminological systems were distinguished. Nowadays, a new era – an era of computing technology and the Internet has led to the creation of a specialized computer sub-language containing computer jargon and technical terms and representing a fairly rich terminology system. A part of the computer sub-language is dynamically represented in the mass media and on the Internet. As a result, the terms begin to be widely used.

Owing to the increased number of computer and Internet users, the terminological lexicon is available not only for professionals, but also for ordinary users. However, ordinary users only have an approximate knowledge of the meaning of those terms.

Accordingly, the popularization of the Internet and computer technology in the modern human life leads to a close relationship between daily communication and professional communicators: the new terminology and concepts change not only our speech, but also our thoughts and ideas and complement them. Overall, complex systems are designed not only for the transmission of general and specific knowledge about the world and the human being, but also for the formation of special aesthetic constructions for long-term prospects [8].

As mentioned above, many terms are currently determinologized or refer to new meanings. Their semantics are expanded or limited: they are used literary. The tendency of expanding the semantic structure of terms appears when they shift from one sphere to another. In this case, the initial meaning and value can be thrown or maintained. In literature, such words are classified as neologisms - previously known but obtaining a new usage.

The process of determinologization causes the loss of scientific accuracy of the term, and expands the scope of its implementation. It should be noted that the rapid development of science and technology, and the representation of research processes and results in the mass media lead to the determinologization process. According to many researchers, this process is the only way for individual terms to shift from general lexicon to specific one.

M.Fomina thinks that the main reasons for researchers to focus on determinologization are [9]:

- public awareness (information provision);
- rapid development of science and technology;
- increased level of education;
- clarity and transparency of the general meaning of the word.

Thus, the determinologization process can be presented as follows:

- a term is primarily used by terminologists in a special context;
- a term begins to be used by non-professionals, however in a particular context;
- the use of a term comes from a specific context.

It should be noted that the first stage is characteristic only for special scientific discourses. However, the second and third stages demonstrate many features of the determinologization process in the scientific-popular discourse.

The hidden sign of terms allows differentiating them from other language units and distinguishing most of the terms. Common concepts are defined by this hidden signs of terms. Several types of general concepts are available; consequently different types of terms may appear.

At each stage of the development of human knowledge and in each century, a number of scientific and technical concepts are generated. They are also linked to general concepts of philosophy, general systems theory, cybernetics, informatics and other methodological sciences. Some of these concepts can be used in various fields of knowledge as general scientific concepts.

It should be noted that general scientific and technical concepts are of common content and therefore applicable in various fields [10].

Thus, determinologization process, which is one of the methods of intellectualization of communicative processes, helps to understand the reality, to develop cognition and enrich the information description of the world.

**Methods of calculating weight coefficient of a term**

The frequency of separate terms represents the level of science and the development of the public structure at a certain stage.

The query, which specifies the user's need for information, consists of separate terms. When the search algorithm is implemented, queried terms are compared, and their proximity is determined, in other words, their relevance is identified.

The greater the weight of the term in the document, the more relevant the document is and the higher position the document occupies in the list of search results.

Thus, the key issue of the term measuring in search engine is required to provide the user with the relevant document.

A simple and more popular search model - Bull model uses a binary number system to measure the terms [11]. This method is implemented at the selection stage of indexed terms and results in the queried term to be equal to 1, while the other terms equal to 0. Thus, all the queried terms are considered to be of the same meaning.

However, this model has some drawbacks. Thus, the use of single weights results in many unsettled documents as a response to the user's request, leading to significant difficulties in revealing the search results. Selection of relevant documents from this set becomes really challenging. Solution to this problem is to assign different weight to the terms. The queried terms included in the same document may have different weights. Moreover, the weight of the same term may vary in different documents.

In addition to this method, the following methods are also used to measure the weight of the terms:

- statistical frequency model;
- probability model;
- semantic analysis model.

One of the simplest ways to evaluate the importance of term for any document is the use of a static frequency method - $TF-IDF$ ($TF$ *– term frequency,* $IDF$ *– inverse document frequency*). The statistical frequency model of term weight is closely linked to the indexing frequency method [11]. The principle is that if the word is frequently encountered in any document and if it is rarely found in the other set of documents, it may be of great significance for the document. The weight of the word is proportional to the number of words used in the document, and in the other set of documents it is inversely proportional to the frequency of that word. $TF-IDF$ is often used for text analysis and information search.

$TF$ - term frequency means how frequently the term is in the document. The importance of the word $t_i$ in a document is evaluated. Logically, assume that the term can be found in long documents relatively more rather than in shorter documents; therefore, the exact number is not searched. The relative number is applied when the necessary term is used in the text. In other words,

$$tf(t,d)=\frac{n_t}{\sum_k n_k} \tag{1}$$

where, $n_t$ denotes the number of the term $t_i$ included in the document $d_i$, $n_k$ -the total number of words included in the document $d_i$. It turns out that any word that is more frequently found in the document may be regarded as an important term, which precisely represents the content of this document.

The first form of the weight of the term was first given by Hans Peter Lun (1957). The weight of the term found in the document is proportional to its frequency [12].

When *TF* is counted , all terms are considered to be equivalent to each other for their importance. However, practice shows that the most frequently used words in the document are not always important term for this document. It is known that serving words, connections, pronouns, etc. are more commonly found, however, practically they do not affect the content of the text.

According to the analysis of Zipf's law of the content of the documents, if any word in the document has a high frequency, but is not a term, it mostly has to be found in many other documents of the corpus with high frequency [13]. This feature can be used in the process of selecting important terms for documents. Therefore, at the second stage, all documents of the data corpus are analyzed and the weight coefficient of the terms in that document is determined by taking into account the characteristics of occurrence in other documents of the corpus. This quantity is defined as a logarithm of the ratio of the total number of documents to the number of documents where the term is found, and called the inverse document frequency.

It should be noted that the statistical interpretation of the terminological specificity named as *inverse document frequency (IDF)* was defined by *Karen Spärck Jones* (*1972*) [14]. *IDF* measures the meaning of the term directly and is signed as follows:

$$idf(t, D) = \log \frac{|D|}{|\{d_i \in D / t \in d_i\}|} \tag{2}$$

$|D|$ denotes the number of documents in the corpus; $\{d_i \in D / t \in d_i\}$- the number of documents in the corpus $D$, where the term *t* is found ( $n_t \neq 0$ ).

The quantities of the term frequency in the document and inverse document frequency can be combined within the Single Frequency Indexing model.

$$w_{ij} = tf(t, d) \cdot idf(t, D) \tag{3}$$

Placing the expressions (1) and (2) in the formula (3), and taking into account the characteristics of all other documents in the information space, the following formula is obtained to calculate the weight coefficient $w_{ij}$ of the term $t_i$ in the document $d_i$:

$$w_{ij} = \frac{n_t}{\sum_k n_k} \cdot \log \frac{|D|}{|\{d_i \in D / t \in d_i\}|} \tag{4}$$

The formula (4) shows that the higher the frequency of the term $t_i$ in the document $d_i$, and the less the number of documents, where this term is found, the greater the weight of the term $t_i$ in the document $d_i$. In other words, if the document $d_i$ is a document concentrating the term $t_i$, then the term $t_i$ is of importance for this document.

Note that *TF – IDF* is a quite universal metric for measuring the importance of the term. Several formulas are based on this method. They differ for the use of coefficients, normalization, and logarithmic scale. One of the most popular formulas is the formula *BM* 25 [15].

A probability model is developed for assessing the weight of terms for building the compatibility between real information needs and terms. The probability model is based on the precise estimation of the probability that the given document is relevant to the request [16].

A certain term can express completely different definitions. Accordingly, the existence of this or that term in some documents does not indicate the relevance of the document to the request. The problems described are solved through the latent semantic indexing (LSI) [17]. The essence of this approach is that each document has an open and latent semantic structure. Analysis of such structure (LSI) allows any document to be described for the presence or absence of a certain term

or from the point of its semantic meaning. For example, a document can be sufficiently described with a term which is not included in its structure, and vice versa, i.e., some terms do not represent the essence of the document, and their coincidence with the requested terms does not mean the relevance of the document.

Thus, as a result of the latent factor counting, the power of the term can be corrected and the search of the document may be more adequate to its content. LSI is mathematically carried out through the singular separation of the matrix, one of the methods of the linear algebra [18]. Modern algorithms also use probability theory apparatus.

Note that certain work has been carried out on the automation of terminology building in Azerbaijan.

In the 80's of the last century, the Institute of Cybernetics of the Academy of Sciences of Ukraine jointly with the Institute of Cybernetics of the Academy of Sciences of Azerbaijan (now the Institute of Controlled Systems of ANAS) developed the automated analysis of terminology lexicon (AASTL) [19]. AASTL performs the classification of lexicon by particular parameters, formation of some lexical groups, calculation of system parameters of words and dictionaries. The data generated by the system is used for theoretical studies of lexicon and for correction of terminological dictionaries.

Within the framework of the NTIS concept in Azerbaijan, the National Terminology Web Portal [20] has been developed, which is an important step towards the automation of terminology building.

## Conclusion

The following conclusions were made in this research:

- Terminologization and determinologization, analysis of the terms harmonization principles and weight evaluation showed that the terminology itself contains dynamic events. Therefore, specialists in this field shall focus on innovation in society, taking into account the need to support the functional development of the language. Hence, the name of this innovation in the society should be identified in mother tongue;
- There is a need to develop a national standard on understanding and harmonization of terminology in Azerbaijan, which will lead to the organized internationalization of terms and the exact compatibility between them;
- It was revealed that the popularization rate of the field can be determined by the dynamics of the frequency of terms in any field of knowledge;
- The statistical model used to determine the importance of the term is one of the commonly used algorithms for keywords extract. Some newly developed methods use this model as part of the calculation process. Practically all the corpus-based methods depend on the weight function. This function balances some features of a term or expression in a particular text with similar rules in the whole corpus;
- As a result of the collection of terminology dictionaries in a single information system in Azerbaijan, the effectiveness of terminology will increase.

## References

1. Garipova F.H. Practical aspects of language policy: terminological work // Herald of VEHU, No. 3 (41), 2009, pp. 34-39.
2. Constitution of the Republic of Azerbaijan, Article 21. State language, 12 November 1995. www.azerbaijan.az/portal/General/Constitution/doc/constitution_a.pdf
3. Sadigova S. The term creativity in the modern Azerbaijani literary language, Baku: "Elm", 2010, 244 p.
4. Alguliyev R.M., Gurbanova A.M. Conceptual bases of building of terminological information system in Azerbaijan // Problems of information society, Baku, 2011, /No1, pp.3-8.

5. ISO 860: 2007 (en), Terminology work - Harmonization of concepts and terms. www.iso.org/obp/ui/#iso:std:iso:860:ed-3:v1:en
6. Bichkova O.N. Determinologization as a functional and communicative transformation of the term // Scientific Almanac: Language. Text. Discourse, Stavropol: Publishing House SSPI, 2011, Issue 9, pp.458-462.
7. Meyer I., Mackintosh K. L'étirement du sensinologique: aperçu du phénomène de la déterminologisation // Le Sens en Terminologie, 2000, pp.198-216.
8. Glazirina A.I., Antonova A.V. Determinologization in the Russian computer sublanguage: data from the component analysis // IIS UrGPU, Ekaterinburg, Russia, 2013, pp.20-26
9. Fomina M.I. The modern Russian language. Lexicology, 4th ed., Rev., Moscow: Higher School, 2001, 415 p.
10. Zubkova A.A. The term as a special lexical unit (on the material of the subject area "Logistics") / Actual questions of philological science of the XXI century: proceedings of the V Intern. sci. Conf. Young Scientists, Ekaterinburg, Ural Federal University, 2016, pp.107-112.
11. Salton J. Dynamic Library and Information Systems, Moscow, Publishing House Mir, 1979, 559 p.
12. Luhn Hans Peter. A Statistical Approach to the Mechanized Encoding and Searching of Literary Information // IBM Journal of Research and Development, 1957, vol. 1, no.4, pp. 309-317.
13. Zipf G.K. Human behavior and the principle of least effort: an introduction to human ecology. Cambridge, Addison-Wesley Press, 1949, 573 p.
14. Karen S. J. A Statistical Interpretation of Term Specificity and its Application in Retrieval // Journal of Documentation, 1972, vol.28, no.1, pp.11-21.
15. Zaragoza H., Craswell N., Taylor M., Saria S., Robertson S. Microsoft Cambridge at TREC-13: Web and HARD tracks / The Thirteenth Text Retrieval Conference (TREC-2004). Gaithersburg, Maryland, 2004, NIST Special Publication, pp. 1-7.
16. Sparck J.K. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments // Information Processing and Management, 2000, No. 36 (6), pp. 809-840.
17. Nekrestyanov I.S., Panteleeva N. Text search systems for the Web // Programming, 2002, No4, pp.33-57.
18. Indexing by Latent Semantic Analysis / S. Deerwester [and others] // Journal of the American Society for Information Science, 1990, No. 41 (6), pp.391-407.
19. Mammadova M.G., Skorokhodko E.F. Automated analysis system for terminological vocabulary, Moscow, VNIITI, Scientific and Technical Information, 1981, Series 2, No. 1, pp.14-18.
20. Terminology Commission under the Cabinet of Ministers of the Republic of Azerbaijan. www.terminology.az