REPUBLIC OF AZERBAIJAN

On the rights of the manuscript

ABSTRACT

of the dissertation for the degree of Doctor of Philosophy

THE CORPUSES OF THE NATIONAL LANGUAGES AND PRINCIPLES OF THEIR FORMATION (On the basis of the materials of British and American English)

Speciality: 5704.01 – The theory of language
Science field: Philology – linguistics

Applicant: **Sevinj Salim Mammadzadeh**

Baku – 2021

The work was performed at the department of the "Languages and Pedagogy" at Odlar Yurdu University.

Scientific supervisor: prof. Doctor of Philological Sciences
Mesmekhanum Yusif Gaziyeva

Official opponents: Doctor of Philological Sciences, assoc prof.
Kamila Abdulla Valiyeva

Assoc. prof. Doctor of Philosophy
Javanshir Khankishi Muradov

Doctor of Philosophy
Ramila Misirkhan Faradjova

Dissertation council – ED 1.06 of the Supreme Attestation Commission under the President of the Republic of Azerbaijan operating at the Institute of Linguistics named after Nasimi, Azerbaijan National Academy of Sciences.

Chairman of the Dissertation council: academician
Mohsun Zellabdin Naghisoylu

Scientific secretary of the Dissertation council:
Doctor of Philosophy on Philology, assoc.prof.
Sevinj Yusif Mammadova

Chairman of the scientific seminar:
Doctor of Philological Sciences, assoc.prof.
Gulsum Israfil Huseynova

INTRODUCTION

The actuality and the usage rate of the research work. At present the rapid progress of computational technology, computerization of printing and publishing, appearance of internet, acceleration of information exchange, collection of a huge amount of language material stored in electronic media require use of these sources. Application of language material, especially words in the linguistic means (dictionaries) by preservation of such files caused formation of corpus linguistics.

Corpus linguistics that is a new branch of computational linguistics has a dual character. On the one hand, formation of corpuses, determination of their sub-corpora, supply of corpus texts with the necessary annotations, construction of relations between corpuses; on the other hand, implementation of the linguistic researches by using corpuses, determination of supporting role of corpuses in solution to the concrete problem. In recent years significant work has been done to research the different languages including English using the mathematical and statistical methods, to study the quantitative indicators of the language in the synchronous and diachronic plans, to compile frequency dictionaries, to construct the systems of automatically searching linguistic information. However, these researches, applied studies have not been carried out on the basis of the voluminous materials. As materials of the corpuses are large and wide enough it is possible to verify accuracy of the results obtained, to carry out the modern researches that will give the representative results.

Corpus linguistics is considered to be a branch of computational linguistics. This new linguistic direction formed in the light of compiling the frequency dictionaries develops in connection with such problems as machine translation, application of the mathematical and statistical methods in linguistics, and finally, formation of systems of natural language processing. The further development of corpus linguistics does not exclude the possibility of consolidation of all issues of computational linguistics.

Use of corpuses is one of the characteristic features of the modern linguistics. The different linguistic information and material form the corpuses are applied in solving problems in broad terms.

Revealing of some linguistic factors, confirmation of their regularity and coincidence require research and analysis of voluminous language material. Solution of this task is possible only at the level of corpus linguistics. Corpus linguistics is used both in application, teaching language and in its research. All these factors confirm the urgency of the research connected with corpus linguistics.

The object and the subject of the research. The object of the research is corpus linguistics as a whole. The relationship between the other directions of corpus linguistics and computational linguistics are also highlighted at the level of the research object. The subject of the dissertation deals with the purpose and tasks of creation of corpuses, peculiarities of corpuses, social and dialectal variation of the language, creation of new corpuses using the experience.

The aim and the tasks of the research. The main purpose of the research is construction of corpuses, determination of their sub-corpora, study of applicability of corpuses in solving linguistic issues. For this purpose, the following tasks are to be solved:

- determination the factors that gave impetus to the evolution and formation of corpus linguistics;
- study of the reasons of formation of the first corpuses;
- description of the main linguistic corpuses, clarification of their characteristic features;
- study of the peculiarities of formation of the different corpuses through determination of their classification criteria;
- analysis of the construction of corpus grouping the language material by genres, dialect and speech distinction;
- study of the principles of construction of corpuses on the standard language, dialect, language variant, slang;
- description of the similarities and differences between corpuses on the standard English, variants and dialects of English;

- the comparative analysis of the variants of English on Brown corpus, LOB and FLOB;

- study of the main issues of annotation of corpuses;

- clarification of the place of corpus linguistics in computational linguistics;

- determination of sphere of compiling corpus and frequency dictionaries, relations and connections between corpus and machine translation, corpus and mathematical and statistic researches.

The methods of the research. The descriptive, historical-comparative and statistic-distributive methods are used in the dissertation.

The basic provisions giving to the defence:

1. Creation of linguistic corpuses is a constantly improving process based on the existing experience.

2. Brown corpus and corpuses created on the different variants, dialects of English confirm the possibility of creation of corpuses on the dialects of the Azerbaijani language and also the Turkic languages.

3. The optimal information search in the corpuses depends on marking of the texts included in the corpus.

A correct determination of marking increases possibilities of using the corpus.

4. Expansion of the scope of sub-corpora of linguistic corpuses is possible.

Inclusion of the special terminological sub-corpora in the corpuses gives the material for determination of the peculiarities of enrichment of the terminological fund. The terminological corpus of English can be applied in the compiling electronic terminological dictionaries.

5. Computational linguistics and corpus linguistics are interrelated directions. All issues solved by computational linguistics can be accomplished within the corpus.

6. A linguistic corpus is to be an open system and its constant enrichment is to be meant.

7. The development of the modern computational technologies

increases the perspectives of corpus linguistics, and increase in volume of linguistic base is possible.

The scientific novelty of the research. For the first time, the history of corpus linguistics, the first corpuses and their difference from the modern corpuses, application of corpus in computational linguistics are researched comprehensively.

Transition from the applied issues to the theoretical issues of linguistics and computational linguistics that are the basis of corpus linguistics are traced, the dual nature of the corpus is revealed.

For the first time, the problem of study of relationship between the fields of computational linguistics is put forward and compiling dictionaries, machine translation and corpus linguistics are comparatively researched in this direction. The main scientific novelty of the research is the study of social and regional dialects of British and American English and the formation of their corpora, their application to the field of computer linguistics and machine translation, as well as the problems during the application and usage for the first time in the field of Azerbaijani linguistics.

The theoretical and the practical importance of the research. The main theoretical issues of the research include the problem of annotation of a corpus. The studies prove the direct link between marking and the highlighted problem. Generalization of experience of corpus makes it possible to consider the theoretical problems of creation of the modern national corpuses from a new angle. One of the aspects of the theoretical significance of the dissertation is connected with the analysis of the experience of Brown, LOB, FLOB and other corpuses on the basis of the material of the Azerbaijani language.

The obtained results can be used in formation of the corpus of the national language, in creation of sub - corpora of this corpus. In addition, the dissertation can be used at the seminars and special courses on computational linguistics. The results of the research can also be applied in marking the text.

The approbation and the applying of the work. There are papers on the theme of dissertation published at the following conferences: the international conferences “The actual problems of cognitive and applied linguistics” (October 20-21, 2016) and “IV International scientific conference of young researchers” (April 29-30, 2016); the Republican scientific conferences: “XX Republican scientific conference of doctoral students and young researchers” (May 24-25, 2016) and “The Republican scientific conference dedicated to the 90th anniversary of the First Turkological Congress” December 28, 2016). 9 articles and 3 theses have been published on the theme of the dissertation.

The name of the organization where the dissertation has been accomplished. The dissertation has been accomplished at the Department of the “Languages and Pedagogy” at Odlar Yurdu University.

The volume of the structural sections of dissertation separately and the general volume with the sign. The dissertation consists of introduction, three chapters, conclusion and list of references and the list of abbreviations. The introduction consists of 6 pages, the first chapter - 40 pages, the second chapter - 54 pages, the third chapter -33 pages, the conclusion- 2 pages, the list of references – 11 pages and the list of abbreviations- 1 page. The dissertation consists of 148 pages, 38 509 words and 246, 314 signs.

THE BASIC CONTENT OF THE WORK

The actuality of content and the usage rate of the research work is based, the object and the subject of research, the aim and the tasks, the method and ways of the research, the basic provisions giving to the defence are defined, the information about the scientific novelty of the research, the theoretical and practical importance of the work, the approbation and the applying of the work, the name of the organization where the dissertation has been accomplished, the volume of the structural sections of dissertation separately and the general volume with the sign is given in the part of **“Introduction”** of the dissertation.

The first chapter of the dissertation is **“Corpus linguistics and its creation”.** The first subchapter of this chapter contains the excursus into the history of corpus linguistics. Being a part of computational linguistics corpus linguistics is a scientific field aimed at study and creation of language corpuses (text corpuses) by means of computational technology. Corpus linguistics which is the most modern branch of applied linguistics has very ancient historical roots. These roots are connected with the collection and systematization of the language material.

Both compiling card indexes and their storage cause the different difficulties and problems. For example, a number of researchers’ work is used in order to choose and collect the words related to the history of language from the written monuments, record the examples of their use in the cards. It takes years. It is necessary to arrange the words in alphabetical order and type them. At the same time there are such problems as making special filing cabinets, expanding repositories with the increasing number of filing cabinets etc. Use of electronic machines created new opportunities and conditions for the solution of the problems of the development of computational technologies.

In 1960 a decision was taken to construct the first linguistic corpus in electronic media. The standard corpus of the modern American English was constructed in the USA Brown University from 1961 to 1964 inclusively. The first computationalized corpus – the Brown Corpus – contained about 500 texts from the American newspapers, journals and books. The authors of the corpus W.Francis and H.Kučera used a huge amount of materials and collections. For the first time creation of the corpus of English in the USA depends on both the development of technology in the USA and correct substantiation of the idea of corpus reasons, purposes of its constructions and the worldwide interest to this language used in the country. Creation of Brown corpus greatly helped linguistis’ and non-linguistis’ researches in the field of English. It was a great impetus to the development of linguistics and the English language.

The second computational corpus was prepared by Lancaster and Oslo universities together with the Bergen scientific centre. This corpus was compiled in 1970-1978 and its name is abbreviation from the initial letters of the names of cities where these universities and scientific centre are situated: LOB (Lancaster – Oslo-Bergen). The structure of LOB linguistic corpus is similar to the structure of Brown corpus. Corpus includes one million word forms.

Both Brown and LOB corpuses include one million word forms that were substantiated by coverage of all word forms in the language. It should be noted that Brown corpus includes the material of American English and LOB Corpus includes the material of British English. Both corpuses are still urgent and useful. They are researched in detail and the subtlety of English is studied through these corpuses. Shortly thereafter, production of high-speed computers with voluminous memory began. In addition, the work on entering texts into the computational through scanners has been facilitated. This work used to be carried out with the keyboard for a long time. Entering texts printed before through scanners and OCR program (Optical Character Recognition) into the computational memory accelerated collection of texts. As a result, it was possible to enter the texts consisted of billions of words into the corpus. There were more than 600 computational corpuses in 1990. British National Corpus (BNC) and Corpus of Contemporary American English (COCA) take an important place among the modern corpuses of the English language. The corpus provides the opportunity to use the materials connected with the variants of English during the research of a certain sphere of English. BNC and COCA are essential to resolve the difficulties in this direction.

Corpus linguistics began forming as a branch in the first half of 1990. "*Corpus linguistics was in its prime*". This expression was noted in J.Svartvik's article dedicated to corpus linguistics in 1992.[1]

[1] Svartvik, J. Directions in Corpus Linguistics / J. Svartvik. – 1991. – p. 116.

"*Corpus linguistics is closely linked to computational linguistics; it also uses it and at the same time enriches it"*[2].

Though at present linguistics do not have complete information about the corpus they are aware of significance of the corpus. "*A corpus language is an electronic collection of texts of any language"*.[3] This thought can be explained as follows: a person reads a text in order to get some information and use it. A linguist begins to divide the text into parts, and such separated texts take places in the corpus. At the present time literature and materials that used to be collected for years can be collected in a short time. Today the time is spent on study of material, not on its collection. Corpus greatly facilitates hard work. The role of corpus is indispensable for linguistics. The quantity and quality of the material received from the corpus is much higher in comparison with the period before corpuses. Using more than 10 000 examples during the research of the language is higher than using 10 examples. The corpus of this language is used in order to find these 10000 examples. Thus, appearance of corpus linguistics in the 1980 s, preparation of a number of corpuses confirmed the urgency and importance of corpus linguistics in the end.

The second subchapter of this chapter considers the main corpuses of the English language. British National Corpus (BNC) is considered to be one of the biggest corpuses. It is a 100 million word collection. The corpus was created in Oxford University with the participation of Lancaster University and the British library.

90% of the BNC is samples of written corpus use. These samples were extracted from regional and national newspapers, published research journals or periodicals from various academic fields, fiction and non-fiction books, other published material, and unpublished material such as leaflets, brochures, letters, essays written by students of differing academic levels, speeches, scripts,

[2] Захаров, В.П. Корпусная лингвистика / В.П. Захаров, А.С.Злыгостева. – Санкт-Петербург: – – 2013. – с. 13.

[3] Кулагина, О.С. Исследования по машинному переводу / О.С. Кулагина. – Москва: Наука, – 1979. – с. 14.

and many other types of texts. *The corpus also includes the different styles and is endowed with the themes. This corpus consists of only the texts in modern English. There are also loan words used in the British English in the corpus.*[4]

The texts presented in BNC were chosen and collected by three criteria: time, space and type of creation. All texts cover the same period from the standpoint of time.

The examples from the texts created after 1975 are included here. The examples from the fiction account 25% of the total amount. 75% of the written texts were taken from the field of science, culture, sociology and politics.

The corpus of the modern American English is considered to be the biggest corpus that combines the different genres and free access. This corpus covers mainly five genres: oral literature, fiction, the famous newspapers, journals and academic publications. The material of the oral genre combines 85 million words collected from about 150 TV channels and radio broadcasts. The literary genre consists of 81 word forms. These also include short tales and plays, film scripts. Samples from the journals represent 86 million word-forms. The material on various subjects was chosen from 100 journals for the corpus: news, sports, religion and other themes. Newspapers (81-million word forms) covered sports, news, financial and economic sections of ten USA newspapers. Samples from the academic publications (81 mln word forms) combine examples from about 100 scientific journals.[5]

The third subchapter of this chapter is dedicated to the classification of corpuses. Corpus is divided into non-electronic and electronic types of corpuses according to the format criterion. Non-electronic corpuses are related to the previous periods. At present only electronic corpuses are used. The second criterion applied in the classification of corpuses is the peculiarity of access of texts to the corpus. The fragments of the texts of the same volume were collected

[4] Захаров, В.П. Göstərilən əsəri, – s. 65.

[5] Косякова, М. Корпус современного американского английского: [элэктр. ресурс] / презентация – 2014. - http://www.slideshare.net/, – с. 3-6.

from the different texts in the Brown corpus, so this corpus is a selective text. Corpuses can be with full text.

In this case a full text is included in the corpus. The whole text of the written monument is included in its corpus. Actually, inclusion of the whole text in the corpus depends on the tasks of the research.

Being static and updated corpuses are divided into two groups. In general, the range of classification criteria of corpuses is wide enough. These criteria are connected with the purposes of corpuses. With the development of computational linguistics the new forms of use of corpuses are found. The future perspectives and types of corpus linguistics are also based on the annotation of corpus materials. Morphological, syntactic, semantic, terminological marking confirms the necessity of registration of at least four corpuses.

There is a sufficient number of the different types of corpuses. V.P.Zakharov distinguishes *the different types of corpuses on purposes and indicators.*[6] It should be added that his classification cannot be accepted completely, it can be accepted only conditionally. Thus corpuses are multi-purpose and specialized according to their purposes. Multi-purpose corpuses collect texts of the different genres, and specialized corpuses cover only one genre or one group of genres. Corpuses of texts can be grouped on genres: literary, folklore, dramatic, publicistic etc. Marking criterion divides a corpus into the coded (marked) and non-coded (unmarked) groups. It can be called by another name, for example: indexed and unindexed. Indexed corpus contains the words and sentences with tags (morphological, syntactic, semantic etc.).

One of the important criterions for users is its usability. It is possible in corpuses with free access through on-line. There must be right to use in historical corpuses. Closed corpuses are intended for other purposes, public use is forbidden. Corpuses are divided into monolingual, bilingual and multilingual ones according to the criterion of text volume.

[6] Захаров, В.П. Göstərilən əsəri, – s. 16-17.

The variants and dialects of the language are opposed in the monolingual corpus. For example, bilingual and monolingual corpuses as variants of English can be divided into two main types:

1) corpuses that demonstrate numerous original texts and their translation into one or several other language;

2) corpuses that combine texts covering the same field regardless of writing in one or several languages.

Such corpuses are mostly used by translators. Both types of corpuses are widely used in translation, machine translation, compiling terminological dictionaries and also in comparative research of languages.

The second chapter of the dissertation “The problems of compiling corpuses of regional and social dialects” deals with the role of the social factors in formation of the standard language corpuses.

Influencing the other languages English shares words and the world gives impetus to the development of English. *One of the main indicators of spread of English is appearance of new lexemes, changing semantic structure of the vocabulary.*[7]

Nevertheless, use of English as an official language in the different countries causes appearance of new words in them, and consequently, vocabularies of the language variants differ from one another. In that case the variant of the used standard English and common standard language are formed in each country. Thus, there is a need to form a corpus of any standard language and conduct comparative research in order to reveal the similarities and differences between these languages. The main purpose of the Broun corpus was to compile frequency dictionary on the basis of the standard variant of American English and determine the lexicographical corpus of this standard language.

Later LOB, FLOB and other corpuses were formed on the basis of the certain language. Such formation of corpuses undoubtedly

[7] Ярцева, В.Н. Развитие национального литературного английского языка / В.Н. Ярцева. – Москва: – 1969. – с. 57.

requires clarification of the issues related to the standard language, dialect, variant language, slang.

In I.V.Arnold's opinion, *the standard English is an official language of Great Britain, the literary English language spoken in schools, universities, used in radio and television, by educated people. Its vocabulary differs from the dialectal words. The local dialects are variants of English that have no literary forms and they are used and understood on the quite different territories.*[8]

English is an international language in the world. It has become means of international communication and received the title of "the global language" in the past ten years. It is an official language of Great Britain; most British understand this language and speak it. However a part of the population in the North and centre of Wales speak Welsh, the population in the West of Scotland speaks Scots; Irish is spoken in North Ireland. English spoken in the fourth sides of Great Britain has its peculiarities and distinguishing features. The attitude to the regional and social variants in Great Britain depends on the accent and dialects between the North and South. The standard English was formed on the basis of the South dialect.

Essentially it means dialects of London and its surrounding areas. "*The dialects of the northern regions are considered to be more difficult and backward in comparison with the southern dialects; it is accepted as the low-level people's language*".[9]

British English is traditionally named the standard English by its lexical and grammatical peculiarities.

The privileged segments of the population living in London mainly speak the standard English. Cockney English is the accent or dialect of English traditionally spoken by working-class Londoners. It is commonly associated with the East End of London. It is mixed with the Saxon dialect from the phonetic standpoint. Cockney is considered to be the cleanest social local dialect of England.

[8] Arnold, I.V. The English Word / I.V. Arnold. – Moscow: – 1966. – p. 262.

[9] Матюшенков, В.С. Словарь английского сленга / В.С. Матюшенков. – Москва: – 2002. – с. 214.

Slangs are preserved in some dialects as a result of the language development. A number of lexemes that occurred in the earlier periods of the language formation are still used in some dialects. Slangs have certain grammatical and phonetic peculiarities of the dialects to which they relate. It is impossible to speak only slangs in any language. Slang can be used together with the other words and expressions in the literary language. According to the results of study many words that are neutral from the stylistic standpoint (mainly in English) were formed due to metaphors and other stylistic figures. *Additionally, a neutral word used in the speech was subjected to the semantic changes several times in the early stages of the language development. The word that used to be a metaphor later was understood as a homonym.*[10]

The main vocabulary of slang in the English language is remarkable for their distant past. Most modern slangs existed long before the formation of the literary language. *Like the other layers of language slang is constantly developing and enriched with the words with new meanings.*[11]

It should be noted that the elements which entered English from the regional dialects of America are unstable. As a rule after a while they fall into disuse. Only 40 % of them are related to the dictionary of Australian slang. The other 35% are Cockney words and 25% are Americanisms.[12]

The geographical-local variants of English have come a long way of historical development. The England, Wales, Scotland and Ireland variants of English are used in the United Kingdom. Furthermore, as English is widespread in the other places of the world there are quite a lot of its variants.

[10] Маковский, М.М. Системность и а системность в языке / М.М. Маковский. – Москва: – 1980. – с. 124-162.

[11] Маковский, М.М. Английские социальные диалекты / М.М. Маковский, – Москва: – 1982. – с. 23.

[12] Baker, S. The Australian Language / S. Baker. – Sydney, London: – 1936. – p. 288.

The second and third subchapters of this chapter of the dissertation deal with the analysis of the corpuses of Great Britain and American English.

English spoken in four countries of Great Britain (England, Scotland, Wales, and Ireland) has its specific peculiarities. English of each country is quite different.

It was assumed for a long time that RP (Received Pronunciation) is a social variant and patois of the privileged segments of the English population.

The word "received" in the XIX century was accepted as a notion of the literary language. It was mostly understood as the language of the aristocracy. Later this notion was propagated as "the King's English".

English spoken in the USA is called American English. In I.V. Arnold's opinion, as *American English is a regional variant it cannot be called a dialect. This variant is derived from the standard English, there is American National standard.*[13]

American English like America itself has very interesting history of the development. Three and half centuries period is reflected in the vocabulary of American English. Such words and terms as "*Blue laws, sunbonnet, law cabin, forty-niner, cash house, motel, boby-sitter"* give information about the past and present of America. *American English does not remain within the country, quite a lot words have passed into the other languages (Ok, telephone).*[14] Indian words, through the Spanish and Portuguese languages, the Aztec languages influenced the language of England before English took roots in America. The names of some Indian tribes now are the onomastic units in the USA, for example: "Iowa, Kansas, Michigan" etc. Some of them took roots in the European languages, too. The words that came to Europe from New World had rich exotic peculiarities. This was due to the mixing of the language used in America with the languages of the neighboring countries, for

[13] Arnold, I.V. The English Word / I.V. Arnold. – Moscow: – 1966. – p. 265-266.

[14] Thomas, P. Words and ways of American English / P. Thomas. – New-York: Random House, – 1952. – p. 4.

example: *potato, tomato, chocolate, cocoa, cannibal, maize, savannah* etc. Though these words are considered to be usual they were new words for the English and Europeans before. For example, when the word "barbeque" appeared in British English it had been already used and known in America. One of the variants spoken in many places and heard in the most famous music is "Black English" (primarily spoken by most black people in the United State. At one time this variant was not interesting to linguistics. However at present there is great interest in this dialect. This dialect differs from the standard English and white people's English. In Mc.Devis's opinion, *there is no factor that differs the Black from the white people, but their pronunciation rules make it possible to determine their racial origin.*[15] One of the North American writers emphasized some words used by the Black living in the southern states in his work in 1888 and gave their explanation, for example: *buccra* "a white man", *brottus* "a small present" etc.

The American linguist W.A.Stewart made numerous attempts to study differences between "Black English" and "standard English": Such words as *goeber* "hazelnut", *juke* "a small box" etc. entered the vocabulary of English. The word "Goeber" is of African origin. "Guya" means "groundnut" in Hausa language (Western African language).[16]

U-huh "yes" and other words are understood both by the White and Black in the USA. But despite this, the people from Great Britain neither understand nor accept this word.[17] Actually, Black English is still controversial. Some scientists call it a language, some call it a slang, and some call it a dialect. The USA media repeatedly noted that Black English is an unnecessary dialect; this dialect must be corrected. In fact, Black English is a variant of American English that differs from the literary English language.

[15] McDavid, R. Dialects: British and American standard and nonstandard- in linguistics / R. McDavid. – New-York. – 1969.

[16] Chamberlain, A.F. Negro Dialect / A.F. Chamberlain. – 1988. – p.23.

[17] Steward, W. Difference between Black English and standart English / Steward W. – 1996. – p. 102.

The third chapter of the dissertation is called “Application of the corpuses of the national languages”. The first subchapter of this chapter deals with the research of relationship between the directions of computational linguistics. The different linguistic researches widely use computational linguistics in the modern world. One of the greatest problems of computational linguistics is structural, grammatical and semantic text processing, or rather OCR (Optical Character Recognition) regardless of written or oral speech.

Modern computational linguistics is rapidly developing and drawing attention with its great scientific and applied achievements. Systems of Automatic translation from all languages into the other ones were created. Translation from a number of languages into one another is perfect. Machine translation systems created in such countries as Russia, the USA, Japan are considered satisfactory today. Computational linguistics has succeeded in compiling monolingual and bilingual dictionaries.

Computational linguistics as a branch of linguistics is divided into two parts: 1) theoretical and; 2) applied computational linguistics. Relationship between the language and computational takes a special place in the theoretical sphere. It is understood as transformation of the natural language into computational. The natural language is used by computational again. In general, theoretical computational linguistics studies the peculiarities of the natural language. From this standpoint, theoretical computational linguistics is of great importance. Researches on computational linguistics are very useful for the practical purposes of the language formation.

One of the main issues of applied computational linguistics is natural language processing and creation of artificial speech. Applied computational linguistics is closely linked to informatics, mathematics, logic, cognitive psychology. The first Talking Machines imitated human voice. Modern attempt to create electronic speech is called speech synthesis.

There are problems connected with the flexibility of speech synthesis systems. For example, speech production is a more complex process than pronunciation. Such factors as intonation, pause etc. are taken into account. At present many problems of automation of texts in most cases can be solved in computational. Many standard programs were drafted in order to get statistical indicator in the machine.

Significant progress has been made lately in the areas of scientific and technical texts, official document processing, machine translation, automation of preparing reports. Preparing translation systems, automatic dictionaries for the different languages necessitated creation of a new field in linguistics. Thus a new scientific field appeared at the interface of linguistics and some exact sciences in the middle of the XX century. At first it was called "mathematical linguistics", "computing linguistics" and other names. *At last a new scientific field "computational linguistics" is known as a new field of linguistics in the world science and Russian linguistics, and this term has already gained right to citizenship in the world science.*[18]

In the general, the issue of relationship between computational linguistics and corpus linguistics is often put forward. Researchers sometimes separate them, sometimes include corpus linguistics in computational linguistics. Actually the term "computational linguistics" is in itself a subject of the dispute. It used to be called "mathematical linguistics", "computing linguistics". Now it is called "computational linguistics", and the main reason is use of computational in this sphere. At first the notion "computing linguistics" was used because it was directly related to calculating of language units. The need to calculate was undoubtedly caused by finding frequency of word use. Long before computers a researcher had to read a text several times in order to solve such a task and register the required word, then the registered words were calculated and their quantitative indicators were determined.

[18] Марчук, Ю. Н. Основы компьютерной лингвистики / Ю. Н. Марчук. Изд.-во МГУ, – 2000. – с. 38.

Careful work in the systems of electronic machines, transformation into the talking note, provision of their visibility etc. make linguistics to cooperate with other scientific fields. This collaboration enables them to set up letters of computational keyboard so that their order would correspond to their frequency of use. *"Modern computational keyboard causes problems because of the order of symbols that are not based on the research of the Azerbaijani language".*[19]

Many problems noted by F.Veysalli have been solved at the level of the researches for the Azerbaijani language. However, the results of these researches have not been sufficiently used. Typing in Azerbaijani on the computational requires arrangement of letters-graphemes of the Azerbaijani alphabet on keyboard on the basis of the quantitative characteristic of the language.

Automatic translation as achievement of computational linguistics has become reality and has been given for use of people. At present the programs of translation from many languages into English, and vice versa have already taken their places on the internet. Posting of automatic translations programs on the different internet sites is still continuing. The most complicated problems of translation from the Indo-European languages and vice versa have been already elaborated. The influence of the national language corpuses on the electronic lexicography is considered in the second subchapter of this chapter.

Electronic dictionaries were compiled long before the other dictionaries in the machine translation system.

There is now a wide choice of dictionaries in the software market. Only one dictionary could be added to the first electronic dictionaries. Now numerous dictionaries can be included in electronic dictionaries. Previously expansion of electronic dictionary by users was impossible, but now it is possible in the modern LINGVO 4.6. and other versions. *There are two translation regimes in the electronic dictionaries according to the translation regime:*

[19] Veysəlli, F. Dilçiliyin əsasları / F.Veysəlli. – Bakı: Mütərcim, – 2013. – s. 56.

automated and interactive regimes. There is word-for-word, or literal translation in the first regime. The lexical base of such dictionaries is very weak. Unfortunately, such dictionaries do not help translators and do not influence their work.[20] In the second regime a user presses the foreign word seen on the monitor, finds the translation function from the open window and translates the needed word. This is an effective regime for translators.

There are a number of advantages of computational dictionaries; appeal to dictionary is much quicker and simplier, at the same time appealing to several dictionaries one can compare the meanings of the word. Periodic updating occurs, new meanings are added and new words, language innovations take their places in these dictionaries in the short term.

“Dictionary.com”, “Dictionaries”, “Online dictionaries” are the most widespread and widely used computational dictionaries in English. Each of them consists of more the 200 dictionaries. “Your dictionary” consists of the dictionaries in 240 languages and more than 30 books. Such dictionaries contain information about the belonging of the search word to the concrete part of speech and its synonym, antonym and homonym. These dictionaries cover European and at the same time oriental languages.

The idea of creation of corpuses was initially connected with the frequency dictionaries compiling. Compiling such dictionaries essentially depended on the problem statement. “Mathematical statement of problem” means numerical data of the materials, i.e. codification, solution of their parts. Certainly solution of the problem requires choice of the optimal variants of methods.

Algorithm of the problem must be formed on the basis of the chosen method. Frequency dictionaries that take an important place in Azerbaijani linguistics were compiled through mathematical-statistic methods. Some difficulties arise during compiling this dictionary. Firstly, it is necessary to choose such texts that sufficiently reflect the norms of the literary language. It is not so easy

[20] www.xreferat.com/31/4352-1mashinniy -perevod.html

at all. Secondly, a more complicated problem is to choose total volume of texts.

The main source for the frequency dictionary of the Azerbaijani language was newspapers. 100 thousand words were chosen from the texts.

The third subchapter of the chapter is dedicated to the issue of machine translation and linguistic corpus. Machine translation is a process of translation written or oral texts from one language into the other one by means of computational. Study of foreign languages is necessary not only for travelling and hosting but also for watching famous Hollywood films, reading writing on the foreign production and web pages.

We come to the conclusion that we have to use such a foreign language even within our country.

There can be two approaches to machine translation: deductive and inductive. The first method is based on the use of the model "text-meaning-text". In this case it is possible to get the high-level machine translation making the best use of language semantics. The general scheme can be imagined as transition of the morphological, syntactic and semantic levels to the text semantics. Thus the basis of this system is transition from the text to its semantics. This transition takes place due to vocabularies of incoming and outgoing languages. Surely, grammar rules play an important role, too. The second approach considers the text as a diversified system.

The words used in the different dialects impede the perfection of machine transition. For example, a word used in a region of America can have quite a different meaning in Britain. A translator who knows these dialects gives the meaning that corresponds to the context during the translation, but computational expresses it only by the word arranged in the system. For example, *kitchen* means "a room or part of a room used for cooking and food preparation in English. In American English this word means "hair that resides at the nape of the neck", for example: *"Sheila should comb her kitchen* *because it is looking kinda rough"*[21]. It is likely that machine translation will be wrong and this sentence will not be translated correctly. But human translation will be correct taking into account the context.

V.Ingve writes: *"The works in the area of machine translation are in front of the semantic barrier. We begin to understand that the only method in machine translation is the following: machine must "perceive" what it translates"*.[22] In our opinion, this is very correct explanation. The mind of computational cannot be like human mind. Perception of translation by computational is a very complex process. When translating any text a person uses own artistic, literary knowledge, worldview. A person looks for the word according to the context, finds it' and adds more artistry to the translated text.

Semiautomatic translation as a branch of machine translation is also much known. Computers are not completely used in semiautomatic translation. They are used only as auxiliary means. At that moment a translator uses special dictionaries compiled with the aid of machines. A translator does not waste his time on the search for a word in the dictionary and finds the meaning of all unknown words in the short time.

The last subchapter of the third chapter deals with the issues of annotation and marking in the corpus. Corpus linguistics is a branch of applied linguistics that studies the general principles of formation and use of linguistic corpuses. The main sources of the language material are texts. Corpus linguistics differs from translational linguistics. Corpus linguistics studies speech. Traditional grammar focuses on the research of the language. Corpus linguistics researches speech that comes from its material. A corpus consists of a databank of natural texts compiled from writing and a transcription of recorded speech. Corpus consists of real texts and they are speech products that are result of the communication process of the different types.

[21] Moody, S. The Diversity of English in America // Popular linguistics, - 2011. – vol 1, - issue 2, - p. 48-50.

[22] Ингве, В. Значение исследовании в области машинного перевода // НТН, – сер. 2, № 7, – 965. – с. 44.

Opportunities for analysis, segmentation and segment analysis are wide in corpus linguistics because its object is a finished text, and the units forming this text are revealed. The word forms of the texts are studied in the corpus linguistics. The structure and mathematical purpose of the corpus allows for determining the inner circle of the word form. Corpus gives the material for calculating the probability of a certain sequence of word forms. One of the advantages of corpus linguistics is the following: those who research the quantitative peculiarities of the language may directly use the corpus material. Use of corpus gets rid of such labor – intensive work as choice, collection and loading of material. It should be noted that available electronic libraries can be used with some purposes. There are search systems in many user programs (for example, Word). Some problems can be solved by using these systems. It is possible to determine frequency dictionary of writer's language, total volume of frequency of words used by the writer.

One of the important problems solved in the process of formation of corpus in corpus linguistics is provision of transition from marked text to unmarked text. The term *annotation* is used in corpus terminology of English. This term entered Russian, too. As Azerbaijani corpus linguistics is new the appropriate terminological base has not been formed yet. Nevertheless, such variants of terms as *annotasiya, annotasiyalama, nişanlama, markerləmə* can be used. Annotation or marking means linguistic information included in corpus.

Corpus is not completed only by choice of texts, determination of contexts and inclusion them in the corpus. In this case corpus loses its significance and becomes meaningless collection in comparison with electronical library. One of the important decisive factors of corpus is annotation of included texts. At first annotation covered only linguistic information. With the development of corpus linguistics it became necessary to provide texts with additional information. Corpus included information about its creation, genre of text, author, date of writing, name of the work from which the text is chosen, precise information about edition, page number. They are not linguistic data. If linguistic information is connected with syntactic structure it is related to a sentence, but if linguistic information is connected with a lexeme and grammatical peculiarities it is related to a word.

Solution to the mentioned problems poses certain difficulties. When user directly accesses information block about the author he gets information about this author's works that are to be used and directly accessing text corpus from here he acquires corpuses. In the other case, user chooses context from corpus and then gets information about its belonging, date of writing. The author of corpus faces such tasks as different directions of access information about printed version of the context. So placing of reference to marking, annotation materials in the corpus, the different views and approaches to them are taken into account in corpus linguistics.

"Conclusion" of the dissertation contains the generalization of the main results and findings and findings of the research:

1. Corpus linguistics was formed in order to collect and store enormous amount of language material, to solve the different linguistic problems using this material.

2. When creating the national corpus of the language it is important to take into consideration selection criteria of the collected material. When using the corpus text there must be opportunity to analyze the words, word-combinations, grammatical categories used in the text.

3. A range of classification criteria of corpuses is wide enough. These criteria are connected with the purposes of corpuses. With the development of computational linguistics the new forms of use of corpuses are found. The future perspectives and types of corpus linguistics are also based on the annotation of corpus material. Morphological, syntactic, semantic, terminological marking confirms the necessity to note at least four corpuses.

4. In corpus linguistics, standard language, social dialects, slang and language variants form corpus. Recently, English is considered a global language and the English of each country is quite different. Dialects of Great Britain outnumber the dialects of America. From

this point of view, one of the main issues in creating the corpus of the language is the importance of the text fragments included in it to cover all dialect forms of the language. Feasibility of creation of the different corpuses on the bases of language variants, dialects, patois, slang is confirmed in translation of such available corpuses as Brown, LOB, FLOB.

5. Corpus linguistics differs from traditional linguistics. Corpus linguistics studies speech. In traditional linguistics, the main focus is on the study of language. The study of speech by corpus linguistics is based on its material. These materials are written, printed, and transcribed examples of oral speech that have been incorporated into electronic media. The texts that make up the body are real existing texts, and they are speech products that are the result of different types of communication processes.

6. The concept of annotation or marking refers to the linguistic information included in the case. The corpus is not completed only by selecting the texts, defining the contexts in the order of random numbers and entering them into the corpus. As corpus linguistics developed, it became clear that texts needed to be supplemented with additional information.

The following thesis and articles relating to the subject of the dissertation have been published:

1. Korpus dilçiliyində məqsədi və əsas istiqamətləri. // – Bakı: Filologiya məsələləri, – 2015. №1, – s. 258-261

2. Korpus dilçiliyi kompüter dilçiliyinin əsası kimi // – Bakı: Filologiya məsələləri, – 2015. №1, – s. 103-108.

3. American variety of English language // – Toronto (Canada): International Journal of English Linguistics, – 2015. – Vol 5, – p.159-163.

4. Sosial and Regional Variations of English Language // - USA: Global Journals of Human-Social Sciences, – 2015. – Vol 15, – s. 13-15.

5. Different dialects in corpus linguistics. (Based on bnc and coca.) // – Odlar Yurdu Universitetinin Elmi və Pedaqoji Xəbərləri, –

Bakı: – № 41 – 2015. – s. 152-158

6. Scientific Interest to American English // – Gənc tədqiqatçıların IV beynəlxalq elmi konfransı, filologiya elmləri, – Bakı: – 29-30 May, – 2016. – s. 1281-1282.

7. İngilis dilinin standart dil variativi // Koqnitiv və tətbiqi dilçiliyin aktual problemləri adlı beynəlxalq elmi konfrans, – Bakı: - 20 Oktyabr, – 2016. – s. 320-322.

8. Amerikan ingiliscəsinin “Black English” dialekti // Doktorantların və Gənc tədqiqatçıların XX Respublika elmi konfransının materialları, – Bakı: – 24-24 May, – 2016. – s. 182-184.

9. Müasir ingilis dilində slenqlər // Birinci Türkoloji Qurultayın 90 illik yubileyinə həsr olunmuş Respublika Elmi Konfransının materialları, – Bakı: – 28 Dekabr, – 2016. – s. 235-236.

10. Erkən amerikanizmlər və onların ingilis dilində təzahürü // – Bakı: ADU-nun Elmi xəbərləri, – Cild 2, №3, – 2017. – s. 49-52.

11. Qloballaşma və onun dünya dillərinə təsiri // – Bakı: Filologiya məsələləri, №4, – 2018. – s. 188-195.

12. Amerika ingiliscəsinin inkişaf yolu və Noah Vebster // – Bakı: Slavyan Universitetinin “Elmi əsərləri”, – 2019. – s.183-188.

13. Maşın tərcüməsi və linqvistik korpus // “Müasir təhsilin inkişafında beynəlmiləlləşmənin rolu” mövzusunda Beynəlxalq elmi-praktiki konfransın materialları, – Bakı: – 29 Aprel, – 2019. – s.49-51

14. Using national corpora in translation // “Вопросы современной науки: проблемы, тенденции и перспективы”, международная научная конференция, – Выпуск 9, – Москва: – 13 Сентябрь, – 2019. – с. 44-46.

15. Theoretical basis of machine translation // Филологические социокультурные вопросы науки и образования» IV международная научно-практическая очнозаочная конференция, – Краснодар: 25 Октябрь, – 2019. – с. 486-492.

The defence will be held on 20 April in 2021 at 14^{00} at the meeting of the Dissertation council – ED 1.06 of Supreme Attestation Commission under the President of the Republic of Azerbaijan operating at Azerbaijan National Academy of Sciences, the Institute of Linguistics named after Nasimi.

Address: Baku, AZ 1143, The avenue H.Javid 115, V floor, ANAS, The Institute of Linguistics named after Nasimi.

Dissertation is accessible at the Library of Azerbaijan National Academy of Sciences, The Institute of Linguistics named after Nasimi.

Electronic versions of dissertation and its abstract are available on the official website of the Institute of Linguistics named after Nasimi, Azerbaijan National Academy of Sciences.

Abstract was sent to the required addresses on 19 March in "2021"

Signed for print: 15.03.2021
Paper format: 60x84 16\$^{1}$
Volume: 38,358
Number of hard copies: 20