UDC 004.912:81
DOI https://doi.org/10.32782/tps2663-4880/2025.44.1.19

AUTOMATIC SEGMENTATION OF SENTENCES OF THE TEXTS IN THE AZERBAIJANI LANGUAGE

АЗАВТОМАТИЧНА СЕГМЕНТАЦІЯ РЕЧЕНЬ ТЕКСТІВ АЗЕРБАЙДЖАНСЬКОЮ МОВОЮ

Rahmanova N.,
orcid.org/0009-0008-1941-3020
Doctorant
Institute of Linguistics of National Academy of Sciences of Azerbaijan named after Nasimi

The article has been devoted to problems of volume-pragmatic segmentation of texts in the natural language. In the research work automatic segmentation of sentences and indentions of texts in the Azerbaijani language are considered. Automatic segmentation of texts into indentions and sentences is of great importance in modern linguistics, particularly in computer linguistics. Automatic segmentation of sentences in Azerbaijani texts is possible on the basis of markers that indicate the beginning and the end of a sentence. Along with the main marker indicating the beginning of a sentence, it becomes necessary to identify cases in which special or specific markers are present. The article reveals that in Azerbaijani texts certain specific phenomena sometimes occur at the beginning of a sentence. In particular, the use of an ellipsis at the beginning of a sentence is possible. Cases involving special markers indicating the end of a sentence are also observed. The article also includes research aimed at identifying such exceptional cases.

Automatic segmentation of texts in the Azerbaijani language into indentions and sentences is carried out as to the graphical signs, used in the text. Investigations show that punctuation rules of any language, its graphemes give possibilities to determine the markers, indicating the beginning and end of the indentions and sentences of the text in communicative speech. In the article signs and sign combinations being able to be used in the indentions of the texts in the Azerbaijani language, including the end of the sentences are determined and on the bases of them adequate algorithm was set up. The fact that this algorithm ended in representative result has been examined by the way of experiment. The consequences of the investigation make it possible to set up mutual algorithm of automatic segmentation of indentions and sentences of texts for the Turkish languages. In this context, it is necessary to verify whether specific symbols exist in the orthography of the relevant Turkic languages. If such symbols are present, corresponding modifications to the algorithm will be required. The results obtained can be used in the development of the national corpus of the Azerbaijani language as well as of Turkic languages.

Key words: text, literary text, paragraph, sentence, lexeme, word form, segmentation.

Статтю присвячено проблемам об'ємна-прагматичної сегментації текстів природної мови. У дослідженні розглядається автоматична сегментація речень і абзаців текстів азербайджанською мовою. Автоматична сегментація текстів на абзаци та речення має велике значення для сучасної лінгвістики, зокрема для комп'ютерної лінгвістики. Автоматична сегментація текстів азербайджанською мовою на абзаци та речення здійснюється на основі графічних знаків, що використовуються в тексті. Автоматична сегментація речень у текстах азербайджанською мовою можлива на основі маркерів, які позначають початок і кінець речення. Поряд з основним маркером, що вказує на початок речення, виникає необхідність визначення випадків наявності спеціальних або специфічних маркерів. У статті виявлено, що в азербайджанських текстах інколи спостерігаються окремі специфічні явища на початку речення. Зокрема, можливе використання трьох крапок на початку речення. Також фіксуються випадки наявності спеціальних маркерів, що позначають кінець речення. У статті проведено дослідження, спрямоване на виявлення таких виняткових випадків.

Дослідження показують, що правила пунктуації будь-якої мови, її графеми надають можливість визначати маркери, які позначають початок і кінець абзаців та речень тексту в комунікативному мовленні. У статті визначено знаки та поєднання знаків, які можуть використовуватися для виділення абзаців у текстах азербайджанською мовою, зокрема для позначення кінця речень, і на їх основі побудовано відповідний алгоритм. Репрезентативність результатів цього алгоритму перевірено експериментальним шляхом. Результати дослідження дають змогу створити універсальний алгоритм автоматичної сегментації абзаців і речень текстів для тюркських мов. У цьому контексті необхідно перевірити наявність специфічних знаків в орфографії відповідних тюркських мов. Якщо такі знаки існують, виникає потреба у внесенні відповідних змін до алгоритму. Отримані результати можуть бути використані під час створення національного корпусу азербайджанської мови, а також тюркських мов.

Ключові слова: текст, художній текст, абзац, речення, лексема, словоформа, сегментація.

Introduction. Modern linguistics and perspectives have actualized and solved directions of computer linguistics, working out language materials by means of computer, automatic translation formation of National corpses and its application for scientific-practical purposes, automatic segmentation of language units of different levels and issues of their attraction to investigation. Generally, analysis, including automatic analysis are carried out on the basis of language units. Issues linked both with anal-

ysis and synthesis demand the selection, collection and grouping of the language units. When hierarchic systematization is taken as a basis, attraction of units of still smaller level, for examples, phonemes, root words and affixal morphemes, lexemes including word-forms (root morphemes having accepted grammatical suffixes – root + suffix) is relatively simple and having the possibility of using the existing dictionaries such works can be realized manually by the investigator. Great number of sentences in the language material, differentiation of their structure, and number of words included into the composition of the construction, also existence of simple and complex sentences, for getting representative results, the required volume of collection of the sentence become more complicated. From this viewpoint the necessity of automatic segmentation of any text, especially a literary one in the sentence level is demanded.

A Few words on the Necessity of the Study of the Problem

The problem of automatic segmentation is not a new problem for computer linguistics. Different researches have been carried out on this problem [1, 2, 3, 4]. In the process of automatic analysis of word forms, the work of segmentation of these word forms into roots and suffixes, segmentation of word forms in the sentences, division of texts into sentences belong to the problem of automatic segmentation [5, 6]. Distinctions in language systems, in the word order of the sentences, in graphemics, sentence structure, in the rules of punctuation make the application of automatic algorithm of segmentation and its program insurance and in a number of cases make it even impossible. From this viewpoint study of the problem of automatic segmentation of texts in any language is important.

Aims and Objectives

The main aim of the research is setting up the automatic segmentation of texts in the Azerbaijani language. To reach the goal the following objectives are intended for the realization:

-to determine the volume-pragmatic parts of the text in the Azerbaijani language;

-to determine stable volume-pragmatic units of the texts in the Azerbaijani language;

-to determine graphic means of the borders of sentences in the sentence;

-to clarify the possibilities of graphic signs in the borders of the sentence which are used in other positions in the text.

Selection of material and determination of its volume-pragmatic units

As the aim of the research is the segmentation of the texts into sentences in the Azerbaijani language selection of texts are required. In the modern linguistics texts are determined differently, which depend on the acceptation of the differentiating signs of the texts. The main categorial signs of the text are its completeness, perfection and informativity and coordination among the units [7]. As the problem of research in visages automatization and use of computer, the informativity of the text, its completeness, semantic ties among the units of the text do not become necessary for the object of investigation and realization of this problem is more purposeful, if it is carried out on the literary texts, because a literary text is complete as to informativity, completeness and as to the ties among the units. Along with literary texts have their own peculiarities.

Texts are divided into volume-pragmatic and conceptual context parts Hierarchically the volume-pragmatic segmentation of the texts contain parts, chapters, paragraphs, special segmentation, and other parts as sentences. Our research, carried out on 50 different stories, show that in these stories always constantly the number of the marked valume-pragmatic units are two. All the stories attracted to the analysis have been established in a certain number of paragraphs. We mark here paragraph by the sign of $\mathbf{A}_1$ changing index. In this case the first indentions of the text will be $\mathbf{A}_1$, the second paragraph $\mathbf{A}_2$, but the last one $\mathbf{A}_n$. In separately-taken stories paragraph in $\mathbf{A}_1$ number is marked. In special cases the number of paragraphs may be equal to one. Each indentions is established in a certain number of sentences. If we sign the indentions with $\boldsymbol{S}_j$ changing index, the index of **j** too, may change from **1** till **M**. In a special case an indention may consist of only one sentence. For this the automatic segmentation of different stories are chosen as materials for investigation. Algorithm must be used as to the following systematization.

1. Automatic segmentation of the text into paragraphs: in this process the determined indentions are places on the **ABZ** massive, lined up as to $\boldsymbol{A}_1$ systematization.

2.$\boldsymbol{A}_1$ paragraphs are automatically segmented into sentences and the sentences of the paragraph being marked by $\boldsymbol{S}_j$ systematization are placed on the **SENT** massive.

Construction Methodology of Algorithms

Automatic segmentation of the Azerbaijani texts into paragraphs

For the automatic determination of paragraphs by machine there must be a marker pointing to the beginning and end of the indention symbol or a sign and this marker must possess its own code in the machine system. It is known in the word programme of the

machine paragraph, indicated by the sign ¶. From the **FIND** operator as a sign of indention ^**p** is used.

In the machine files instead of indention spaces between lines or breach of the lines are used. Putting them in simpler words in order to show the indention in the text, one line is left empty which is realized by pressing the knob "enter" twice on the claviature. In such cases at the beginning of indention placement of two systematic indention indicators ¶ ¶ or ^**p** ^**p** take place.

Text contains great number, consisting of graphemes of the alphabet of the language, punctuations and figures. In addition to this great number of the signs of indentions and spaces are added. We accept the text formally as the systematization of ak. In this case the first sign in the text shall be a1, the second a2 and so on.

Below the first indention of the story called "letter box" written by Jalil Mammadguluzadeh is introduced: As it is seen from the examples the indention consists of a text fragment between two ¶ signs.

¶ *Noyabr ayının 12-ci günü idi. Hava çox soyuq idi. Amma hələ qar əlaməti görsənmirdi. Həkim axırıncı dəfə xanın naxoş övrətini yoluxub, cavab verdi ki, dəxi naxoşun əhvalı yaxşıdır; belə ki, bir həftəyədək səfərə çıxmaq mümkün olar. Xan çox tələsirdi İrəvana getməyə; çünki xanı orada çox vacib işlər gözləyirdi. Və bir də ki, qorxurdu qar yağa və hava dəxi də soyuya və naxoş üçün yola çıxmaq qeyri-mümkün ola. Xan götürdü qələmi, irəvanlı dostu Cəfər ağaya bu məzmunda bir müxtəsər kağız yazdı:*[8, p. 86] ¶

Translation of the text:

¶ *It was the 12th day of November. The weather was very cold. But no sign of snow was visible. The doctor for the last time examined the ailing child of the Khan and said that the child was already good and it was possible for him to travel for the period of a week. The Khan was hurrying to make a trip to Iravan, because very important work awaited him there. Besides he was afraid that it might snow and the weather might become colder and for the ailing child to join him in the trip might have been impossible. The Khan took up his pen and wrote to Jafar agha a brief letter in this content:* ¶

In the very end of the text the sign of indention being marked, the general number of indentions in the text must be one unit less than the number of these signs. In the story, chosen for the research, the total number of indentions indicates that they are equal to 81. The automatic segmentation of the story into indentions is realized as to the following algorithm.

Given $\mathbf{A_i}$; $\mathbf{i} = 0$; $\boldsymbol{a_k}$; $\mathbf{k} = 0$;

1. In the text check up the sign $\mathbf{a_k}$
2. In the text $\mathbf{a_k}$ exists, pass to 3, not exists pass to 16
3. $\boldsymbol{i} = \boldsymbol{i} + 1$
4. $\mathbf{A_i}[\boldsymbol{a_k}]$
5. Check up $\mathbf{a_k} = ¶$. If yes, pass to 6, if not, pass to 16
6. $\boldsymbol{k} = \boldsymbol{k} + 1$
7. $\boldsymbol{A_i}[\mathbf{a_k a_{k+1}}]$; $\boldsymbol{A_i} \in \mathbf{ABZ}$
8. Check up sign $\mathbf{a_k}$ in the text
9. $\mathbf{a_k}$ exists in the text, pass to 10, if no, pass
10. Check up $\mathbf{a_k} = ¶$
11. Yes, pass to 3, no, to 12
12. $\boldsymbol{k} = \boldsymbol{k} + 1$
13. $\mathbf{A_i}[\mathbf{a_k a_{k+1} a_{k+2}}]$; $\mathbf{A_i} \in \mathbf{ABZ}$
14. Check up sign $\mathbf{a_k}$ in the text
15. $\mathbf{a_k}$ exists in the text, pass to 10, if not, pass to 16
16. **END**

Algorithm sets up the marked systematization of change of $\mathbf{A_i}$ index of the indentions, existing in the text, on the **ABZ** massive. The segmentation of the texts is carried out on the basis of elements on **ABZ** massive.

Automatic Segmentation of Texts into sentences in the Azerbaijaniani language

Segmentation of texts into sentences demands determination of signs, markers showing the beginning and end of the sentences. Each language has its own rules of punctuation and these rules are based on these punctuation marks. Great number of punctuation marks in different languages, we may say, that overlap. This system of punctuation marks include the following elements: dot **(.)**, commar **(,)**, interrogative mark **(?)**, exclamatory mark **(!)**, hypen **(–)**, semicollons **(;)**, colons **(:)**. Besides all this in the texts, different combinations of punctuation marks are used: three dots **(...)**, interrogative and exclamatory marks **(?!)**, exclamatory and interrogative mark **(!?)**, double exclamatory marks **(!!)**, three exclamatory marks **(!!!)**, an interrogative mark and two dots **(?..)**. In addition to this the usage a number of accepted punctuation marks are possible: an ordinary bracket **(·)**, transcription bracket **[·]**, big bracket **{·}**, apostrophe **(')**, a star **(*)**, a descending line **(\)**, sign of paragraph **(§)** etc. The usage of last marks in the texts are not productive and they are met in special texts. In the texts several abridgements are also used. For e.g. in the Azerbaijani language **(və s.) (etc.)**, və iliaxır **(və i.a.) (and so on)**, və bu kimi **(və b.k.) (and things like this)**, bizim eradan əvvəl **(b.e.e.) (before christ)**. In English too, such reductions take place, such as: **etc. = and so on; e.g.** (for example; **a.m.**

(time belonging to the first half of the day); **p.m.** (time belonging to the second half of the day); **p. = page; fig. = figure; vol. = volume; Dec. = December; Aug. = august** etc.

Terminal signs showing the final border of the sentence is a dot, sign of interrogation and exclamatory marks. If we carry out statistic investigation, with a small relative error, we may say that the end of 90% of the sentences in the texts end in these three signs. But of course it is natural that the end of the sentences are marked with other markers as well. For e.g. when the thought is uncompleted at the end of syntactic construction three dots are put. Sometimes, for the sake of strengthening emotionality and expressiveness, in order to express positive or negative attitude the author uses, interrogative marks or the combination of dots. It is possible to find examples to this in the literary texts. For e.g. *"Aylar keçmişdi, illər keçmişdi..."* [8, p. 84*]. "Kolxoz sədri dilləndi: – Yoldaş mütəxəssis maşını..." [s. 84]. "Bu da bir istəkdir, bir ricadır ki, belə eləmə. Qəlbin hara, ayağın ora... Kim nədirsə özüdür. Bu ağaclar kimi, sütunlar kimi, pillələr kimi. Yox pillələr hələ gözləyir..."* [8, p. 97].

Translation of the extract: *"Months passed, years passed…" [p. 84]. "The chairman of collective farm sounded: – Comrade machine of the specialist…" [p.84]. "This is a desire, a wish for you not to do it, where there is a will, there is a way… what is he, he is himself. Like these trees, columns, stair-cases. No, stair-cases are still waiting…"* [8, p. 97]. In these sentences there 4 examples, in which sentences or syntactic constructions end with three dots. The first and the fourth of the mentioned above examples in the Azerbaijani language are completed sentences. In the second and the third examples sentences are incomplete. Truly speaking, there are no predicates in these examples. In the second thought *(Comrade machine specialist)* is incomplete. In the third example in the Azerbaijani example *(Qəlbin hara, ayağın ora...)* elepsis is apparent. A sentence can be restored on the basis of context or associatively. Analysis of such cases, namely, evaluation of completion or incompletion of sentence construction is not carried out in the process of automatic segmentation, but automatic segmentation must segment syntactic constructions and sentences ending in three dots as well. Namely, the last sign of the sentence, or (…) three dots in the end of the sentences must be accepted as the marker of the end of the sentences.

"Buyur, sözün-sovun? Nədən ötrü, harandır?..." [8, p. 101] *(Your word, your questions please? Why, where?...)* In this sentence the end of the sentence is completed in unterrogative mark and two dots **(?..)**. The similar case is possible with an exlamatory mark and two dots **(!..)**. We come across the case of the terminal signs indicating the end of the sentence marking the sign of which is indicated with the sign of inverted commars within the bracket **(")**. In the Azerbaijani language we meet terminal signs to be marked both before and after the brackets. For e.g. *"Sabahınız xeyr, əzizim", "Yaxşıyam", "Könlün nə istəyir?"* [8, p. 132–133]. *("Good morning, my dear", "I am well", "What do you wish to have?")* As to the accepted norms terminal sign must be put after the inverted commas, but in fact putting the sign before the inverted commas is more correct because the speech of the author is introduced with the inverted commas and accordingly the punctuation marks belonging to this speech must be placed before the inverted commas. But here we observe the breach of norms and choice of such texts for segmentation, the mentioned above cases must be taken into consideration during the arrangement of algorithm.

Dot, which is one of and the main terminal signs is met within the sentence as well. Such cases are met during the usage of reductions and other accepted phrames. For e.g. If there is a proper noun or initials coming before them, a dot is put after the final letter of the expressing the initial. For e.g. *"Divardakı lövhədə "Direktor M.A.Kərimov" yazılmışdı"* [8, p. 141] *("On the plate on the wall "Director M.A. Karimov was written").*

All these cases are special or exceptional cases. But the terminal mark of dot if being accepted as a marker indicating the end of the sentence, automatic segmentation algorithm may end in a wrong result. Let's say that during the automatic segmentation the sentences *"Divardakı lövhədə "Direktor M.A.Kərimov" yazılmışdı"* can be splintered into the following segments: $\mathbf{S_1}$ = *"Divardakı lövhədə "Direktor* M. $\mathbf{S_2}$ = A. $\mathbf{S_3}$ *– Kərimov" yazılmışdı".*

For the lequidation of the mentioned errors, all the proper markers, signs indicating the end of the sentence must be determined and they must be taken into consideration during setting up of the algorithm. It is also worth mentioning that for the cases, not withstanding the general norms, it is possible to address to subalgorithm in the cases when such exceptional cases appear.

Let's now determine signs, markers indicating the beginning of sentences. In this respect the first sign many be indention. Usually a sentence in a text begins directly with an indention. Namely, algorithm shall start with the search of the sign of indention. Dialogical speech in may cases from the indention starts with the sign of a dash, which finds its solution with the search of the sign of indention. Nevertheless in the text, the speech of another person is introduced

after the sign of inverted commas after the colons. In this case the sign showing the start of the sentence shall be inverted commas. Within the indention the following sentence after the first sentence, after the terminal sign or possible sign for the specific case is characterized by the introduction of space (^**w**) or start with capital letter. The mentioned signs and sign systematizations become the starting markers of the sentences. Thus for the beginning of the sentence the sign of indention (^**p**), terminal sign + space, for e.g., · ^**w**; colons + space + inverted commas (**:^w“**) are differential markers. For the end of the sentence dot, sign of interrogation, exclamatory mark, interrogative mark + colons, exclamatory mark + colons, colons, three dots, dot + markers of inverted comas are chosen. During the establishment of algorithm different exceptional cases are also taken into consideration. By using the markers, indicating the beginning or end of the sentences, Let's set up the algorithm carrying out automatic segmentation of texts into sentences. We must say that before now the problem of segmentation of texts into indentions having found its solution, segmentation of texts into sentences shall be carried out on the **ABZ** massive. We mark the massive of sentences attained by the result of segmentation by **SENTS**, the proper sentence is marked by $\mathbf{S_j}$**, j [1, n]**.

Automatic segmentation algorithm of Azerbaijani text into sentences

Given: **Ai; i = 0; ak; k = 0; Sj; j = 0**

1. $\boldsymbol{i = i + 1}$**;** $\boldsymbol{k = k + 1}$ from the **ABZ** massive $\mathbf{A_i}$, from the **ABZ** massive $\mathbf{a_k}$ is taken
2. Check up the sign of $\mathbf{a_k}$ in the indention
3. If exists $\mathbf{a_k}$ in the indention, pass to 4, if not, pass to 16
4. **j = j + 1**
5. $\mathbf{S_j}$ exists, pass to 6, if not pass to 23
6. $\mathbf{S_j}$ [**ak**] $\mathbf{S_j}$ place $\mathbf{a_k}$ symbols systematically
7. Check up $\mathbf{a_k}$ = ¶ if there is pass to 8, if not, to 17
8. $\boldsymbol{k = k + 1}$
9. $\mathbf{S_j}$ $\mathbf{[a_k a_{k2}]}$; place $\mathbf{a_k}$ symbols systematically in $\mathbf{a_k}$ **Sj** ∈ **SENTS**
10. Check up sign $\mathbf{a_k}$ in the indention
11. If there exists $\mathbf{a_k}$ pass to 12, if not to 4
12. Check up $\mathbf{a_k}$ = If it is one of **[·], [?], [!], [!?], [?!], [!..], [...], [?..], [:], [”]** these signs, pass to 13, if not to 17
13. $\boldsymbol{k = k + 1}$
14. $\mathbf{S_j}$ $\mathbf{[a_k a_k a_k]}$; in $\mathbf{S_j}$ placement of $\mathbf{a_k}$ symbols systematically; $\mathbf{S_j}$ ∈ **SENTS**
15. Check up the sign $\mathbf{a_k}$ in the text
16. If in the text exists $\mathbf{a_k}$ pass to 17, if not, to 4
17. $\mathbf{a_k}$ = ^**w** yes, pass to 4, no to 18
18. $\boldsymbol{k = k + 1}$
19. $\mathbf{S_j}$ $\mathbf{[a_k a_k a_k a_k]}$ placement of $\mathbf{a_k}$ symbols in $\mathbf{S_j}$; $\mathbf{S_j}$ ∈ **SENTS**
20. Check up the sign $\mathbf{a_k}$ in the text
21. If in the texts exists $\mathbf{a_k}$ pass to 22, if not to 4
22. Check up if it is one of the signs $\mathbf{a_k}$ **= [·], [?], [!], [!?], [?!], [!..], [...], [?..], [:], [”]** pass to 13, if not, to 17
23. **END**

Experiment

The work of the established algorithm has been checked up as to 2 systematic indentions taken by the way of occasional choice from 5 literary texts. The number of checked up indentions is. 10 Automatic segmentation algorithm has segmented from to indentions, 2,5 indentions 3 sentences, in two indentions 5 sentences and in one indention in 6 sentences. In 10 indentions 35 sentences have been discovered. 10 sentences have started after the sign of indention, 1 sentence has started with three dots + space, two sentences with interrogative mark + space, one sentence with exclamatory mark + space, 21 sentences have started with dots + space mark. The marks in the end of the sentences composed of dots – thirty one times, one time three dots, 2 times interrogative marks, one time an exclamatory mark. Experiment affirmed the workability of the algorithm.

Discussion

New signs, different from the signs mentioned in the investigated part, showing the beginning and end of the sentences, in the experimental examination of the established algorithm have not been discovered. This can be associated with the few number of the indentions chosen for the experiment. In the texts of bigger volumes, automatic segmentation on the bases of algorithm established on the sentences, showing the beginning and end of the sentences may cause the appearance of new markers. When such cases are discovered, it will be necessary to carry out additional investigations for the determination of the beginning and end of the sentences. We think that qualification, improvement and specification of algorithms will be possible during the process of their concrete application.

Perspective Application of the investigation and its continuation

This research work on the segmentation of the texts in the Azerbaijani language into indentions and sentences can be applied in the languages of Turkic group. Algorithms of automatic segmentation of texts in other Turkic languages can be applied in the same order. At this time the graphics and the rules

of punctuation must be taken into consideration. Analysis shows that distinctions in the rules of punctuation though are few, but they exist among the group of Turkic languages for the indication of ordinal numeral in writing after the figure dot is put. The Azerbaijani sentence *"Kitabın 21-ci səhifəsini açdı" (He opened the 21st page of the book)* in the Turkic language the cardinal numeral "21-ci" (the 21st), shall be "21". In this case, in the middle of the sentence a dot is used and this does not show the end of the sentence. It shows that in each of the Turkic languages, taking adequate rules of punctuation into consideration, it is possible to set up mutual algorithms of the texts in Turkic languages for the segmentation into indentions and sentences. Such an algorithm in automatic translation of one Turkish language into the other and in the establishment of mutual corpses of Turkish languages can find its application.

Conclusion. It is possible to formulate texts in the Azerbaijani language for the determination of the borders of indentions and sentences on the basis of grapheme and rules of punctuation of the texts in the Azerbaijani language. Such a formulation creates ground for the determination of signs and sign combinations, used before the indentions and sentences. In the Azerbaijani text at the beginning of the sentence the sign of indention, inverted comas and dot + space are used. At the end of the sentence three terminal signs – a dot, interrogative mark and exclamatory marks are more productively used. Besides at the end of the sentences three dots, an exclamatory mark and two dots, an interrogative mark and two dots, an exclamatory mark + an interrogative mark, an interrogative mark + an exclamatory mark, sign of semi colons can be used.

Experimental examination of automatic segmentation algorithm established as to the use of signs at the beginning or at the end of the sentences confirms its working exactly. Algorithm can be improved in the procession of still greater missives.

REFERENCES:

1. Slocum J. A survey of machine translation: its history, current status, and future prospects. *Computational Linguistics*. Vol. 11, 1985. No. 1. P. 1–17. URL: https//aclanthology.org›J85-1001.pdf (date of application: 13.01.2026).

2. Marchuk Yu. N. Machine translation in the All-Union Translation Centre. *Overcoming the Language Barrier: Proceedings of the III European Congress on Information Systems and Networks* (Luxembourg, 3–6 May 1977). Munich : Verlag Dokumentation, 1977. P. 623–624.

3. Smith N. A. From words to corpora: recognizing translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP*). Philadelphia : Association for Computational Linguistics, 2002. P. 95–102.

4. Otero P. G. Extraction of translation equivalents from parallel corpora using sense-sensitive contexts. *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*. Budapest, 2005. P. 97–102. URL: https://www.researchgate.net/publication/228666382 (date of application:13.01.2026).

5. Mahmudov M. Türk dillərinin milli korpusu : monoqrafiya. Bakı : Elm və təhsil, 2018. 392 s.

6. Melamed I. D. Models of translational equivalence among words. *Computational Linguistics*. Vol. 26, No. 2. – 2000. P. 221–249.

7. Zanettin F. Translation-driven corpora: corpus resources for descriptive and applied translation studies. London; New York : Routledge, 2014. 252 p. – DOI: https://doi.org/10.4324/9781315759661

8. Azəri S. Seçilmiş əsərləri. C. 1. Bakı : Azərnəşr, 1989. 462 s.

Дата першого надходження рукопису до видання: 14.11.2025
Дата прийнятого до друку рукопису після рецензування: 12.12.2025
Дата публікації: 31.12.2025