IEEE Turkey Section

UBMK'22

Bildiriler Kitabı
Proceedings

7. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

7th International Conference on Computer Science and Engineering

14-15-16 Eylül (September) 2022 Diyarbakır - Türkiye

Morphological Annotation System in The Corpus of Internet Information Texts in Uzbek Language

Abdullayeva Oqıla Xolmo‘Mınovna
Department of Uzbek language and literature
Tashkent State University of Uzbek Language and Literature
Tashkent, Uzbekistan
abdullayevaoqila@gmail.com

***Abstract*—Morphological and syntactic annotation is one of the most important types of annotation at the current stage of development of text corpora because it is clearly applied in areas such as lexical and grammatical development. Morphological annotation of language units in Uzbek was done manually and semi-automatically. We offered special tags for the language corpus, if we identify the most common affixes of combinations speech parts and suffixes in words in Uzbek and analyze them with marked tags, it can serve as a foundation for the further development of language corpora. More than 70 special tags were selected for morphological annotation, and speech units were annotated in the corpus. Search and analysis results are displayed in 3 different ways in the user interface: 1) only by word search: in this case, any word is written in the search line. If the analyzed word is tagged in the corpus, it will appear with all its grammatical markers. 2) word+tag search: at this stage, the morphological or semantic feature of a word in a certain word group can be analyzed. For example, if a word with a specific morphological indicator is analyzed, a word is written in the search line, and a tag is selected from the list of morphological tags. At the time of choosing a morphological indicator, a word group must be specified. 3) search by tag only: this option is used when the researcher needs speech units belonging to a certain morphological index or a semantic group in a certain word group.**

Keywords— Corpus, lemma, token, annotation, tag, pop-up.

I. INTRODUCTION

One of the most important types of annotation in language corpora is lemmatization, the process of identifying and defining the basis (quote or dictionary) of each word in the corpus. Word lemmatization is a key level of annotation for all corpora, especially in student language corpora where spelling errors or word changes are significant. [1]. In this process, information about the basic form of the word (lemma) is added to the form used in the text. This simplifies the use of the case by allowing users to search for different affixes at the same time. Because of the complexity and variety of morphemes present in a language, the roots of words change so much that the search for a particular word in the corpus can become more complicated later. Lemmas are conditionally written in lowercase letters. In the corpus of electronic information texts of the Uzbek language, a lemma explanation is given to each unit of speech.

Another way in which research in corpus linguistics differs is whether linguistic analysis is encoded in the corpus data itself. Such encoding, called corpus annotation, can be achieved by editing the data or adding some analysis to it or storing the analysis separately, linking to the data [2]. If a large number of language samples in the corpus are tagged with linguistic symbols, then the corpus can easily use the function of automatic text tagging. Of course, since there are no automatic tags available in Uzbek like in English, the annotation is done manually. For example, we need to give a grammatical category to each word in the context and add a comment to the corpus to show parts of speech. The entered comment is saved, automatically interpreted each time a similar part of speech is encountered in context. For example, the word of saylovchilarning (elector’s) is commented, determined the signs as a <saylovchi (elector)> [morph. noun+ plural+ genitive case+affixed, semantically noun, nominative]. The annotation of the morphological corpus may seem a bit complicated, but in fact it is a process that has been practiced by linguists for many years through used manually. However, it is also a process of presenting the analysis of any data in a systematic and convenient way [3]. According to some scholars, the final product in the process of creating a language corpus is “annotation” [4]. The Uzbek language is considered an agglutinative language. The morphological form of the Uzbek language is very different from the English language. In English, word forms change with suffixes, or they don't change at all. For example, the verb go changes to went in the past tense except for the forms going, gone, will go. Or the verb fly changes from the form flying to the forms flew, flown in different tenses. The Uzbek language is an agglutinative language with rich morphological structures. In the Uzbek language, root words are formed by adding suffixes and prefixes. Some Uzbek words are formed by adding two or more suffixes to root words. In the morphological analysis of the Uzbek language, there are more than a hundred grammatical forms of one verb.

II. METHODS

We try to analyze the morphological and semantic annotation separately in the corpus. In fact, in the corpus, these interpretations must be made in their entirety at the same time. Morphological annotation is also called grammatical annotation in some studies[4]. It is sometimes referred to as a morphosyntactic annotation [5]. Because in the process of tapping a part of speech, its grammatical feature is taken into account in the joint analysis. We asked important questions before entering the morphological explanation in the corpus of Uzbek language Internet texts. 1) How do we break text into

signs or tokens? 2) How do we choose a set of phrases for tokens? 3) How do we choose which tag to match to which token? All of these are actually linguistic questions, and even if these questions are answered, completely unexpected units can be encountered in the tagging process. The question of whether a token can be interpreted automatically each time it is encountered after a morphological interpretation, and whether it can be corrected without adversely affecting the previous ones if a different tag selection occurs in the context, is now a practical and technical question. In the process of answering the questions, we created a list of tags. Tags represent speech parts. However, asking which phrase is most useful can lead to controversy and linguistic disagreement. The list is based on the ability to provide the most important grammatical explanations. The list contains abbreviations of Uzbek word groups and their grammatical forms in pointed briquettes.

1. **For the verb:**
- Auxilary verbs <Yor>;
- Transitive verbs <O‘tl>;
- Intransitive verbs <O‘ts>;
- Gerund verbs <Rav>;
- Participle verbs <Sif>;
- Infinitive verbs <Harn>;
- Active voice <Aniqn>;
- Reflexive voice <O‘zn>;
- Causative voice <Ortn>;
- Reciprocal voice <Birn>;
- Passive voice <Majn>;
- Positive form of verbs <Bo‘l>;
- Negative form of verbs <Bo‘lsiz>;
- Imperative form of verbs <Buymay>;
- Conditional mood <Shar>;
- Subjunctive mood <Maq>;
- General mood <Xab>;
- Past simple tense <O‘tz>;
- Present simple tense <Hozz>;
- Future simple tense <Kelz>;
- Verb phrase conjunction <KFSQ>;
- I person <1sh>;
- II person <2sh>;
- III person <3sh>.

2. For the noun:
- Nominative case <boshk>;
- Accusative case <Tushk>;
- Genitive case <Qark>;
- Dative case <Jo‘nk>;
- Locative case <O‘rink>;
- Ablative case <Chiqk>;
- I person form of possessive <1shE>;
- II person form of possessive <2shE>;
- III person form of possessive <3shE>.

3. For the adjective:
- Simple degree <Odd>;
- Comparative <Qiy>;
- Superlative <Ort>.

4. For the pronoun:
- Personal pronoun <Kish>;
- Interrogative pronoun <So‘r>;
- Demonstrative pronoun <Ko‘r>;
- Indefinite pronoun <Gum>;
- Universal pronoun <Belg>;
- Negative pronoun <Bsiz>;
- Reflexive pronoun <O‘z>.

5. **For the number:**
- Numeral <Sanoq>;
- Ordinal <Tar>.
-

For the other speech parts, we decided to take the whole: adverb <RAV>, imitations <TAQ>, predlogs <K>, conjunctions <B>, auxiliary words <Y>, exclamations <U>, modals <M>. The morphological annotation also explains other grammatical features of the parts of speech. For example, abbreviation of units <Qisq>, derivative word <Yas>, non-derivative word <TUB>, singular <Bir> or plural <Ko‘p> form and etc. Special tags were selected for each interpretation. The number of tags can be increased, and the number of tags in each language has increased over time. For example, for English, there are 61 tags in BNC C5, 132 tags in LOB, 197 tags in London-Lund corpus, and 270 tags in TOSCA-ICE. Of course, the number of tags also depends on the tagging corpus. We have tried to make the user understandable when choosing abbreviations for tags. Because for English, they prefer to use the Noun: Prop: Sing tag instead of the NP1 tag in the comments [4]. Abbreviative tags are determined for the noun – OT, for the verb – FL, for the adjective – SIF, for thr number – SON, for the adverb – RAV, for the pronoun – OL, for the imitative words – Taq, for the predlogs – K, for the conjunctions – B, for the auxiliary words – Y, for the exclamations – U, for the modals – M. In morphological annotation the OT1shE tag for I person possessive form of noun, FLAniqn tag for the active voice form of verb, OLBelg tag for designative form of pronoun are interpretated. To illustrate in sentence:

Shu_OLKo‘r tariqa_OT ayni_OLKo‘r kungacha_RAV mamlakatda_OTO‘rink koronavirus_OTQark infeksiyasiga_OTJo‘nk qarshi_K emlangan_FLMajnSif fuqarolar_OTKo‘p soni_SON3shE 458 555_RAQSONSanoq nafarga_SON yetdi_FLAniqnO‘tz ._TB (https://daryo.uz/2021/04/21/ozbekistonda-koronavirusga-qarshi-emlanganlar-soni-450-ming-kishidan-oshdi/).

This sentence is interpreted only morphologically, the sentence in the context is interpreted in a horizontal format. Some experts prefer to provide this analysis in a vertical format, which does not affect the quality of interpretation. There are no clear standard guidelines for tag size, but you can increase or decrease the size of the tags. We tried to get their core when choosing tags. Although there is no clear standard for the selection of tags, it depends on the purpose of each case being created. In the annotation of parts of speech, the morphological annotation also includes the definition of lemmas of language units. The annotation process is a step-by-step process. The algorithm of the process of performing operations in the case is explained by means of a diagram.

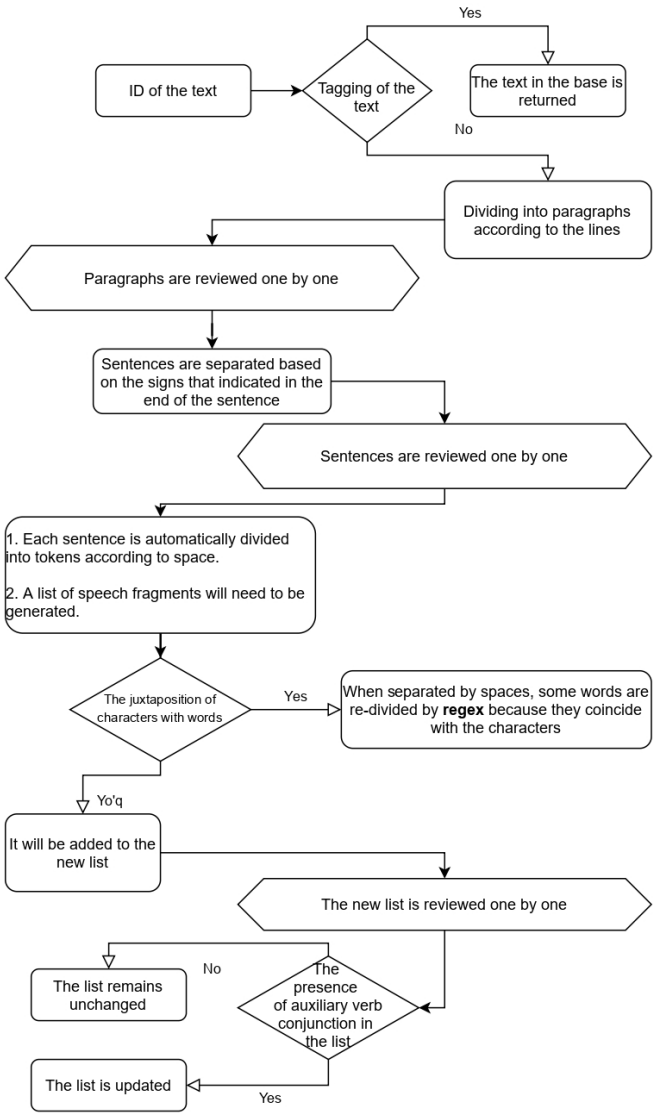

Diagram 1. Stages of the process of performing actions in the corpus

In this diagram, the last case ends with a list of words. After this process, the lemma of the language unit is identified and interpreted using special morphological tags. The Uzbek language uses a special programming language to encode tags in the corpus of information texts.

III. RESULTS

Morphological interpretation is a common and well-developed algorithm for annotating parts of speech. The most common types of morphological annotations in Uzbek texts are:

Noun:

- **OTBir**: *noun, singular*;
- **OTBirQark**: *noun, singular, genitive case*;
- **OTBirTushk**: *noun, singular, accusative case*;
- **OTO‘rink:** *noun, locative case;*
- **OTBir3shEQark: noun, singular,** *III person of possessive form, genitive case;*
- **OTKo‘p: noun, plural;**
- **OTKo‘p3shEQark: noun, plural,** *III person of possessive form, genitive case;*
- **OTKo‘pO‘rink:** *noun, plural, locative case.*

Adjective:

- **SIFOddYAS:** *adjective, simple, derivative;*
- **SIFQiy:** *adjective, comparative;*
- **SIFOrt:** *adjective, superlative.*

Verb:

- **FLO‘tz1Sh:** *verb, past simple, I person;*
- **FLMajnHozz:** *verb, passive voice, present tense;*
- **FLHozz3Sh:** *verb, present tense, III person;*
- **FLOrtnO‘tz:** *verb, causative voice, past simple;*
- **FLKFSQO‘tz:** *verb, verb phrase conjunction, past simple;*
- **FLBuym:** *verb, imperarative mood form;*
- **FLShar:** *verb, conditional mood form.*

Number:

- **SONSanoq:** *number, numeral;*
- **SONTar:** *number, son, ordinal.*

The logical coincidence of the above tags is shown, the number and type of encounters can be further extended. There is also the encounter and tagging of other units in the language. For example, auxiliary words or punctuation. Punctuation, such as commas and periods, is useful for automatic interpretation. This is because punctuation marks that are tagged once will be tagged automatically at subsequent places. We can see the morphological annotation of units in several sentences in horizontal format:

<*Islohot*_OTBir *istovchi*_SIFYas *mamlakatlar*_OTKo‘p *uchun*_K *tayyor*_SIF *tajribalar*_OTKo‘p *bor*_FL ._TB> <*Lekin*_B *bu*_OLKo‘r *bilimlar*_OTKo‘pYas *to‘g‘ri*_SIFOdd *qo‘llanmasa*_FLMajnBo‘lsizShar ,_TB *islohotlarga*_OTKo‘pJo‘nk *qarshi*_SIF *bo‘luvchilar*_OTKo‘pYas *turli*_SIF *bahonalar*_OTKo‘p bilan_B *islohotlarni*_OTKo‘pTushk *noma’lum*_SIFYas *muddatga*_RAVJo‘nk *ortga*_RAV *surib+yuborishi*_KFSQ, *vaqt*_RAV *boy+berilishi*_QFLMajn, *yana*_SIF *bir*_SONSanoq *avlod*_OT *umri*_OT3ShE *zoye+ketishi*_QFL *mumkin*_ AUXVerb ._TB> (https://kun.uz/uz/news/2021/05/03/islohot-formulasi-tuzalishga-yeng-shimargan-mamlakat-uchun-birinchi-navbatda-nimalar-muhim)

In the corpus of Uzbek-language information texts, when researchers perform an analysis of speech units through the user interface in general, the context window in which the analyzed speech unit is involved appears. For example,

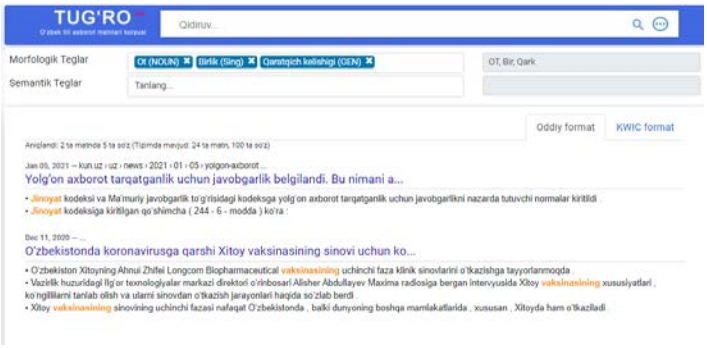

Fig. 1. User interface in the corpus

In th Figure 1, we can see that the user has analyzed the search for tags from the case. Contexts that contain parts of speech that are explained by the requested tag are given in plain format. As the sample case was searched during the initial operation, the search results showed that the system contained 24 texts and 100 tagged words. Of these, 5 words "vaksinasining" were identified in 2 texts. If the text in the case is linguistically annotated, information about any unit of speech is provided. If the researcher clicks on the word "vaksinasining" with the mouse tool, the linguistic annotation of the speech unit will appear in a special form – in the programming language – in the "pop-up".

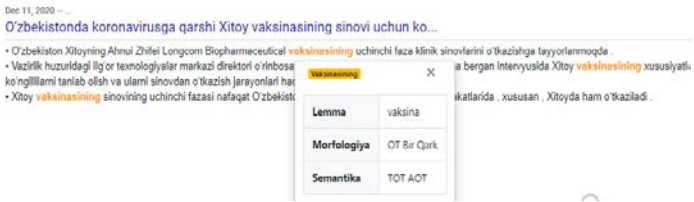

Fig. 2. A linguistic annotation depicted in a pop-up on the corpus

The Figure 2 is a pop-up analysis of the search results. By the uploaded text, the speech units are divided into automatic tokens, and the word "vaksinasining" is given as a single token. As a result of the analysis, the lemma, morphological and semantic annotation of the speech unit is given. So this unit is explained as follows: lemma: vaksina, <morf OT+Bir+Qark sem TOT AOT>.

DISCUSSION

The usefulness of large corpora depends on how easy it is to get information from them. Often, in order to obtain information from the corpus, it is necessary to add some information to the unit being searched. For example, homonymous phrases may belong to different word groups, and in order for corpus users to use them, this information must be provided in the corpus with appropriate comments. In general, morphological interpretation is a multi-step complex process in which unexplained text is lemmatized and grammatical explanations are added through coding. In the process of morphological tagging of speech units in the context of Uzbek language texts, it was difficult to categorize word-forms, to tackle the form of a compound verb and verb phrases, which at the same time formed a complex form of a sentence. For example, in the sentense of "Yangi qonun joriy qilingunga qadar amaldagi qonun o'z kuchida qoladi" the word combination "joriy qilingunga qadar" enacted with semantically "joriy qilinguncha", but our rules do not have a morphological index the "-gunga" form or when the phrases ".... qayd etilishicha, ... aytishicha" are encountered at the beginning of a sentence, the syntactic analysis is analyzed as an introductory word, while the morphological analysis is modalized because contextual analysis is performed in the corpus. Uzbek, which is an agglutinative language, differs from other languages in that the combination of suffixes in speech units is complex and rich, and the process of morphological analysis of the language is unique. For example, there are five word forms of the verb "sing" in BNC, there are "sing, singing, sang, sung, sings". In Uzbek, the verb of "sing" corresponded the verb of "ashula aytmoq" or "kuylamoq", there are about a hundred forms of this verb "kuylayman, kuylaysan, kuylaydi, kuyladi, kuylamoqchi, kuylagan, kuyladilar……", this line can be continued. Many of the words in the text may have morphologically ambiguous interpretations. This is due to the language's long history of change and rich morphology. Detection of morphological changes is required for many applications, such as pronunciation programs, search engines, and machine translation [6]. We also analyzed the results of the national corpus of the Turkish language directly in the construction of the corpus of informational texts of the Uzbek language and the designation of tags. The identification of more than 58 tags for parts of speech and affixes was studied in the national corpus of the Turkish language. We also tried to go in this direction.

CONCLUSION

In general, annotating parts of speech, defining a clear pattern, is important in translating natural language into computer language, in machine translation. One of the important processes in the creation of the corpus of information texts on the Uzbek language is the development of the principles of morphological analysis and morphological processing. Based on these principles, morphological and semantic annotation is carried out at the initial stage of the formation of the corpus of informational texts of the Uzbek language. The morphological annotation in the language corpus and the way the result is expressed in the corpus may be different. There are no mandatory annotation principles in the case. But the purpose of morphological annotation is to focus on problems and solutions in the annotation of languages where the morphological structure of the language is difficult and to show the researcher the correct result.

REFERENCES

[1] S. Brunni, L. Lehto, J. Jantunen, V. Airaksinen; How to annotate morphologically rich learner language. Principles, problems and solutions. Learner Corpus Research: LCR2013 Conference Proceedings BeLLS 2015 Vol. 6, ISBN 978-82-998587-7-9, p.133–152.

[2] T. McEnery, A. Hardie; Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press, 2012. p.312.

[3] Sinclair J., Svartvik J. The automatic analysis of corpora. // Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991. p.379–397.

[4] R. Garside, G. Leech, T. McEnery. Corpus annotation. – Routledge, 1997. p.292.
[5] G. Leech, A. Wilson; Recommendations for the morphosyntactic annotation of corpora, EAGLES Document EAG-TCWG-MAC/R., 1994. www.ilc.cnr.it/EAGLES/browse.html
[6] A. Itai, E. Segal. A Corpus Based Morphological Analyzer for Unvocalized Modern Hebrew. https://www.cs.cmu.edu/
[7] O.X. Abdullayeva; Linguistic annotation in language corpora and its principles. // National corpus of the Uzbek language: problems and tasks. – Tashkent, 2022, may. Vol. 1, –p. 131-136.