International Journal of Language Learning and Applied Linguistics

ISSN: 2835-1924
Volume 2 | No 12 | Dec -2023

.

Tagging and Annotation of Corpus Units

Manzura Abjalova Abdurashetovna

Doctor of Philology (DSc), Associate Professor of Tashkent State University of Uzbek Language and Literature. Tashkent, Uzbekistan

Abstract: Linguistic annotation is one of the features that distinguish linguistic corpora from text systems such as text archives, electronic encyclopedia and dictionary systems, and electronic libraries. Linguistic annotation of text units is created as a result of tagging of corpus units, and it is this possibility that increases the importance of language corpora in education, research and other processes. In many sources, the terms tag, tagging, annotation (linguistic annotation) are not distinguished. Therefore, the meaning of these terms has been explained in this article.

Key words: tag, tagging, annotation, linguistic comment, Corpus, metadata.

To perform a different linguistic task/action, the text must be processed with linguistic and extralinguistic additional information. For this purpose, it is necessary to provide a special explanation (for example, information about the vocabulary) to the components of the existing text. This linguistic information is called a text annotation, and special linguistic symbols are called a tag (Russian: razmetka, English: tag). Explaination of text units is called annotation or linguistic interpretation, and symbolization of each linguistic unit is called tagging (mark-up; Russian: razmechat). For example, a noun is tagged as "Ot" (in Uzbek) or "N" (in English), an adjective as "Sif" (in Uzbek) or "Adj" (in English), a verb as "F" (in Uzbek) or "V" (in English). In the text, the lexical form "*kitobim*" belonging to the noun group "*Ot, n-b nom., birl., I sh., b. eg.q.*" (in Uzbek) is annotated, i.e. linguistically explained. The simplest example of tagging and annotation is general tagging and general annotation of phrases. It might look like this:

1. *Samiya xalqaro tanlovga yaqinda boradi. (Samiya will soon go to the international competition.)*

Let's explain:

Samiya (ot) *xalqaro* (sifat) *tanlovga* (ot) *yaqinda* (ravish) *boradi* (fe'l).

Samiya(N) *will*(aux.V) *soon*(Adv) *go*(V) *to*(Prep) *the*(Det) *international*(Adj) *competition*(N)

2. *Men o'g'lim bilan faxrlanaman. (I am proud of my son.)*

Let's explain:

Men (olmosh) *o'g'lim* (ot) *bilan*(ko'makchi) *faxrlanaman* (fe'l).

I(Pron) am(aux.V) proud(V) of(Prep) my(Adj. Pron)) son(N).

Tag is symbol**, tagging** is symbolizing a lexical unit**, annotation** is a linguistic explanation of text units using tags.

A look at history. In the 80s, a standard for marking up electronic texts called SGML1 (Standard Generalized Markup Language) was adopted. It was developed in the typography industry, but

quickly spread to other industries. The goal of SGML is to make it possible to edit, analyze, and modify documents written in different word processors.

SGML introduced the concept of tags. Tags are service marks in the text that help to get information about the text unit. Each case can be assigned special tags, thereby creating dialects of the SGML language. SGML Markup Language is the "constructor" of languages. It is considered a very complex language and is rarely used. But on its basis, well-known markup languages such as HTML and XML were created.

For tagging textual data (corpus), several universities have specially developed a system that describes which parameters of texts should be tagged. This framework uses XML and is called the Text Encoding Initiative Guidelines (TEI Guidelines). It is a list of different features of text that can be encoded, tagged and indexed. For example, the system lists various corrections, quotations, abbreviations, proper nouns, initials, acronyms, foreign words, etc. in the text. Currently, almost all corpora projects (including the British National Corpus) attempt to follow the TEI recommendations in one way or another. [Кутузов А.Б.].

By convention, tags are enclosed in angle brackets in pairs, that is, in opening and closing positions. For example, <a> is the opening tag, is the closing tag. The closing tag indicates the end of the message given in the opening tag. We give an example of the idea with the above sentence:

<pron>Men</pron><N>o'g'lim</N><prep>bilan</prep> <V>faxrlanaman</V>.

<pron>I</pron> < aux.V> *am*</aux.V> <V>*proud*</V> <Prep>*of*</Prep> <Adj Pron>*my*</Adj Pron> <N>*son*</N>.

As it is clear, it is indicated that the lexical unit at the beginning of the sentence is the pronoun <pron>*Men*</pron>.

Yoki yana

<ds>Samimiyatni o'zingizga bezak qilib oling</ds> – deydilar onajonim.

The part of this sentence "*Samimiyatni o'zingizga bezak qilib oling"* (Make yourself an ornament of sincerity) is given in <ds> and </ds> tags, which means direct speech (ds).

The <pause> tag can be used oddly in spoken corpora. It doesn't matter if it's an opener or a closer. This means there is a stop at the tagged location.

Tags consist of short characters or symbols. For example, in Uzbek: *sifat – Sif, fe'l – F, ot – O, olmosh – Olm*; in English: adjective – Adj, verb – V, noun – N, adverb – Adv. Tags are not visible to the user. An Annotated text display program interprets tags according to specific rules and provides the user with text formatted according to these rules.

Automatic annotation / tagging. Tagging large cases is time-consuming and expensive. Therefore, in the 70s of the XX century, projects to make annotations using a computer began to appear. Then the TAGGIT program tagged 77% of the word categories of Brown's corpus. The remaining 23% were manually tagged for ten years. In the 1980s, the CLAWS (Constituent Likelihood Automatic Word-tagging System) system had a tagging rate of 95%. Probability theory is applied in it. Information about this was provided below. Today, systems for automatic tagging of word groups (morphological analysis, word-class tagging) and automatic tagging of sentence fragments (syntactic analysis, parsing) have been developed. These capabilities are also essential in Internet search vs. machine translation. Also, under the name "Automatic processing of texts" (http://www.aot.ru), Russian language processing using computer technologies has been created. A group of experts from the Faculty of Linguistics of the Russian State Humanities University developed the following module for Russian, German and English languages in this system [Abjalova, 2020]:

- ✓ graphematic (clarification of word boundaries);
- ✓ morphological (identification of Part of Speech and its categories);
- ✓ syntactic (marking of parts of a sentence);

- ✓ semantic (determining the semantic relationship in words).

In general, there are two types of text tagging: 1) writing **meta-data** and 2) linguistic annotation, i.e. attach tags.

1. *Meta-data (meta-information, meta-commentary, metalinguistic information, extra-linguistic information*) is the name of the source included in the corpus, the author, the time of creation, the place of the event, the style, the genre, from the point of view of the linguistics of the period in which the poet lived (characteristics of the language for a certain period) approach: in-depth study of the poet's style and skill in using artistic words (examples of folk art: proverbs, proverbs, riddles, various proverbs, wise words, phrases), highlighting the socio-educational significance of the genre, in the process of genre analysis, it is possible to determine which age group is suitable for the content.
2. In the *linguistic annotation* (a sequence of special linguistic tags (symbols)), the lexical units in the text are determined according to their linguistic characteristics: the grammatical meaning of the word form – part of speech of the word and its grammatical categories (verbs, nouns, adjectives, etc.), specific semantic characteristics information about the words symbols, archaism, historicism is included in the corpus in the form of tags (symbolic signs), as a result of which the educational and research value of the corpus increases, it also allows users to perform a special search in the corpus.

Types of linguistic interpretation:

1) **morphological** (part-of-speech tagging or POS-tagging) – tagging of word groups. For example, noun – Ot (N), verb – F (V), adjective – Sif (Adj);
2) **syntactic analysis or parsing** – description of syntactic relations between lexical units and various syntactic structures;
3) **semantic** – description according to the semantic categories to which a given word or phrase belongs and smaller categories that determine its meaning;
4) **explain anaphoric –** referent relations, for example, connection with pronouns;
5) **prosodic** – uses tags describing stress and intonation;
6) **discourse** – the text is specially interpreted to indicate pauses, repetitions, reminders, etc. in the corpus of oral speech;
7) **stylistic** – indicates the stylistic character of the lexical unit.

When implementing these linguistic annotations, it is desirable to observe the following basic principles:

- ✓ Theoretically neutral (traditional) annotation scheme – rather than creating confusion by using annotation schemes unique to each corpus, i.e. elements from the tagging specification of large linguistic corpora are taken as a basis for general use, the Uzbek language is the basis for the recognition of the hulls as modern hulls in global demand, and in its place such a hull serves as a standard hull. As soon as a theory of authorship is created without the use of a widely known annotation scheme, the user of the corpus is forced to delve deeper into the annotation system. Naturally, such an excessive search does not please the user.
- ✓ A generally accepted system of linguistic concepts - for linguistic corpora to be globally significant, it is appropriate to use international signs and symbols for labels. This is mainly considered important in language didactics and makes the process of language learning and teaching more convenient.
- ✓ Efficient input of parameters – human factor and semi-automated process are reliable in correctly tagging a very large number of linguistic units. This requires the productive work of a team of dedicated professionals.

- ✓ Adherence to international standards – Adherence to the TEI international standard of labeling is based on extensive experience.

Below, we will focus more on the morphological type of linguistic interpretation.

The basic unit of **morphological tagging** is a text form or token, which is understood as a string of characters and is usually equivalent to a simple word form. Such symbolic tagging is necessary for the operation of a computer program. The process of separating tokens from text is called tokenization. In some literature, it is also called graphematic analysis. It is worth noting that token is a term related not only to corpus linguistics, but also to other fields. A character from a probel (space) to a probel is a token [Копотев, 2003. 33-37]. Since the corpus object is the text, and the smallest unit is the word (word form), words and word forms are taken into account as tokens in corpus linguistics.

Along with tokenization in corpus, there is another important process. This step is important for processing the data included in the corpus. In this process called *lemmatization*, the initial form of a word is automatically determined, the initial form itself is called a *lemma*. Lemmatization is very important for inflected languages. The reason is that the base of the inflected word is restored in the process of lemmatization. For example: *copies → copy, bases → basis, oxen → ox; вижу → видить, иду → идти, пальчик → палец.*

It is known that in the public education institutions of almost all countries, the student reads a sentence from the elementary grades and identifies the noun, adjective, number, verb, adverb, and pronoun groups in it. In corpus linguistics, this is the category tag of the word. **Part-of-speech tagging** ((POS tagging or PoS tagging or POST), in Russian: частеречная разметка)) is a stage of automatic text processing, the task of which is to identify the groups of words (forms) used in the text and the grammatical is to determine the characteristics. With this task, POS-tagging is considered one of the first stages of automatic text analysis.

Word group tagging (POS-tagging, Part of Speech tagging) is important for annotating units in the corpus base. The need to define the word group in this way is connected with the fact that the computer does not distinguish homonyms and polysemantic words.

Such features and peculiarities of the created cases increase the possibility of working with them and the importance of the cases.

Tagging word groups and grammatical categories [Asiryan, A.K. 2017.] in corpus linguistics and to eliminate ambiguities in the classification of words, the word is not based only on its form in the dictionary, but on the basis of its expression in the text (sentence), its category tag and other words in the sentence (paragraph, phrase) it is important to consider the possibility of merger.

Identification of sentences members tags is a complex process. Because it is impossible distinguish all Uzbek words for 12 parts of speech. It is known that in Uzbek language there are 12 word groups (independent word groups: noun, verb, adjective, adverb, numeral, pronoun; auxiliary word groups: conjunction, postposition, particle; separate word groups: modal, interjection, imitative words).

A word can be polyfunctional depending on the state of its realization in the sentence structure and the semantic valence of the N-gramm words. [Abjalova, 2020: 73-77] E.g. The sentences classified according to the parts of speech the first word "*sick*" is answering to the question "who?" and it is noun, in the second one (it is answering the question What kind of?) it is an adjective "*It was brought the sick to the hospital*" and "*It was brought sick person to the hospital*" (how? answers the question) is a word in the function of the category of quality [Abjalova, Iskandarov, 2021]. Out of 11,000 borrowed words in the Uzbek Explanatory Dictionary, 66 such polyfunctional words were identified [Qurbonova M., Abjalova M. and others].

To tag word groups, it is not enough to enter a list of words and their groups in the linguistic database. The loss of consistency, as in the case of determining the word group above, or finding a group of polyfunctional, homonymous [Abjalova, 2020. 73-77] or polysemous words expressed in a sentence prompts even an expert linguist to think and search. Also, many words in the Uzbek

language have not been identified as belonging to a specific category. Taking into account such problems that exist in every natural language, several methods are used for tagging word groups.

In most cases, tagging of word groups is based on the following methods (algorithms): 1) method based on rules; 2) stochastic (or statistical) method. Each of these methods has its own characteristics, which are described in detail in our preliminary research [Abjalova, Iskandarov, 2021].

In conclusion, it should be said that tagging of corpus units is 1) accurate acquisition of statistical data in the corpus; 2) language learning and teaching using corpus; 3) identifying semes of lexical units in the corpus; 4) identifying homonymous units used in the context; 5) makes it possible to reveal the semantics of ambiguous and polyfunctional words. Therefore, the tagging of corpus units is important.

References:

1. Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. [Matn]: monografiya. – Toshkent, 2020. – 176 b. ISBN 978-9943-6939-0-6.
2. Abjalova M. Linguistic modules of the program of editing and analyzing texts in the Uzbek language (for the program of editing texts in official and scientific style): Doctor of Philosophy (PhD)… dis
3. Abjalova, M.A., Yuldashev A. 2021. Methods for Determining Homonyms In Homonymy And Linguistic Systems. ACADEMICIA: An International Multidisciplinary Research Journal. Vol. 11, Issue 2, February. Impact Factor: SJIF 2021 = 7.492 (https://saarj.com). ISSN: 2249-7137
4. Abjalova M., Iskandarov O. Methods of Tagging Part of Speech of Uzbek Language. // IEEE – UBMK – 2021: 6th International Conference on Computer Science and Engineering. 15-16-17 September 2021. Ankara – Turkey. DOI: 10.1109/UBMK52708.2021.9558900. – pp. 82-85. Impakt Factor 5.5
5. Abjalova M. Korpus lingvistikasi. [Matn]: uslubiy qo‘llanma / M.A. Abjalova. –Toshkent: Nodirabegim, 2022. – 110 b.
6. Abjalova M., Gulomova N. Author’s Corpus of Alisher Navoi and its Semantic Database. // IEEE – UBMK – 2022: 7th International Conference on Computer Science and Engineering. 24-26 September 2022. – Diyarbakir, Turkey. – pp. 182-187. Impakt Factor 5.5. DOI: 10.1109/UBMK55850.2022.9919546
7. Abjalova M., Gulomova N. ALISHER NAVOI AND THE THIRD RENAISSANCE PERIOD. // Procedia of Theoretical and Applied Sciences. Vol. 4 (2023). 28.02.2023. – pp. 111-115. ALISHER NAVOI AND THE THIRD RENAISSANCE PERIOD | Procedia of Theoretical and Applied Sciences
8. Abjalova M., E. Adalı and O. Iskandarov, "Educational Corpus of the Uzbek Language and its Opportunities," 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, 2023, pp. 590-594, doi: 10.1109/UBMK59864.2023.10286682.
9. Asiryan, A.K. 2017. Сравнение инструментов морфологической разметки. *Научный взгляд в будущее, 10.30888/2415-7538.2017-07-01-027.*
10. Qurbonova M., Abjalova M. va boshq. O‘zbek tili o‘zlashma so‘zlarining urg‘uli lug‘ati. [Matn]: o‘quv-uslubiy lug‘at / M.Qurbonova, M.Abjalova, N.Axmedova, R.To‘laboyeva. – Toshkent: Nodirabegim, 2021. – 988 b.
11. Копотев М. В., Мустайоки А. Принципы создания Хельсинского аннотированного корпуса русских текстов (ХАНКО) в сети интернет // Научно-техническая информация. Сер. 2: Информационные системы и процессы. № 6: Корпусная лингвистика в России. 2003. – С. 33-37.
12. Кутузов А.Б. Курс «Корпусная лингвистика». – 45 с. Лицензия: http://creativecommons.org/licenses/by-sa/3.0/