International Journal of Language Learning and Applied Linguistics

ISSN: 2835-1924
Volume 2 | No 8 | Aug -2023

.

Problems of Semantic Tagging of Terms Related to Astronomy in the Corpus

[1]Bakhtiyarova Fotimakhon

[1]Teacher, Alisher Navo'i Tashkent State University of the Uzbek Language and Literature,
fbahtiyarova83@gmail.com

Annotation: The article presents a practical approach to solving the semantic labeling of terminology related to astronomy for academic buildings of the Uzbek language, creating their data warehouse, and selecting lexicographic products. The methods of semantic labeling of terms related to the field of astronomy in the corpus and the problems arising in this process were analyzed, and a practical solution was recommended to them. The Uzbek language curriculum shows tasks that are important to complete at first when it comes to terms related to astronomy, as well as astronomical aspects of the complexity of labeling iconic terms are based on the example of terms related to one- and multi-valued astronomy by their semantic structure, i.e. these terms have the properties of one- and multi-meaning, while methodological recommendations were given to solve such problematic aspects.

Key words: markup; semantic markup; corpus linguistics; semantic filter; lemma; grammar; corpus markup; semantic touch; term system; sensory terms in the corpus; term system of terms; unambiguous unit; multivalued unit.

Introduction. Semantic markup is the process of attaching semantic tags or symbols to a text or other data type to identify and describe its semantic structure. This allows computer systems to better understand the content and context of data. Within the semantic mark of the text (word, sentence, paragraph) is assigned a mark indicating its semantic role, type or relationship with other elements of the text.

Semantic markup is useful in many fields, including natural language processing, data retrieval, and its analysis. It helps systems automatically retrieve information, classify texts, search for meaning, analyze links between different elements of a text, and much more.

Main part. Semantic markup can also be used to create ontologies. Ontologies make it possible to structure and organize information, as well as establish connections between different concepts and entities. The use of semantic markup can significantly increase the efficiency and accuracy of textual data processing, as well as make them more convenient for automatic processing and analysis.

The article "Zadachi i prinsipi semanticheskoy razmetki leksiki v NKRYa" ("Tasks and principles of semantic markup of vocabulary in the National Corpus of the Russian language) written in cooperation with E.V. Rakhilina and a group of specialists provides valuable information about the tasks and principles of semantic markup in the NCRL.

The article "Lingvisticheskie annotirovannie korpusa russkogo yazika (obzor obshedostupnix resursov)" (Linguistically Annotated Corpora of the Russian Language) by T.I. Reznikova, M.V.

Kopotev is devoted to the linguistic marking of the National Corpus of the Russian language (NCRL)[1], their possibility and uniqueness[2]. In the corpus of the National Russian language, semantic marking is carried out automatically: in the text, one or more semantic and formative symbols are recorded in most lemmas. In this, a detailed classification includes not only nouns, but also non-scientific lexicon, adjectives, verbs and adverbs. It should be noted that one lemma can refer to several classes at once. At the same time, all semantic symbols are automatically transferred to the corpus, and lexical homonyms cannot be combined into a single lemma. Currently, work is being carried out on the creation of semantic filters of the corpus and its application to the corpus markup. They allow automatic lexical homonymy filtering in given contexts or constructions[3].

Among the special programs for natural language processing, automatic parsing takes a special place. It is not enough to have a set of texts to solve different linguistic problems. It is also required to have a variety of additional linguistic and extralinguistic information that is clearly indicated in the texts. Using like-Brown corpus material, it is easy to determine the frequency of words – their regular use in specific contexts. However, it will be the frequency of tokens (word forms). To determine the frequency of lexemes, each word should be assigned a lemma. In order to count frequencies in the context of grammatical categories, they must also be labeled appropriately. If large-scale labeling is done manually, it takes too long, so researchers have developed automatic labeling methods in the case[4].

The problem. Tagging (tagging, annotation) of corpora is a time-consuming process, especially considering the size of modern corpora. For some types of tagging, in particular, anaphoric, and prosodic, it is still very difficult to create automatic systems, and most of the work is done manually, for morphological and syntactic analysis, there are various software tools called taggers and parsers, respectively. As a result of the work of automatic morphological analysis programs (taggers), each lexical unit is characterized by grammatical features, including a part of speech, a lemma and a set of grammars (for example, gender, number, case, animate/inanimate, transitive, etc.). As a result of the work of automatic syntactic analysis programs, syntactic relations between words and phrases are established, and syntactic units are given appropriate properties (sentence type, syntactic function of the phrase, etc.).

However, automatic natural language parsing is not error-free and is polysemous – it usually has several parsing options for a single lexical unit (words, phrases, sentences). In this case, we can talk about grammatical homonymy. In general, disambiguation (morphological, syntactic) is one of the most important and difficult tasks of computer linguistics. Both automatic and manual methods are used to eliminate ambiguity in the creation of corpora. The new generation of corpora contains hundreds of millions of words, so the principles of developing systems that minimize human intervention are advanced. Automatic resolution of morphological or syntactic ambiguity, as a rule, is based on the use of high-level information (syntactic, semantic) using statistical methods[5].

Currently, a lexical-semantic search system based on partial semantic tags of corpus texts is being implemented. Most words in the text with this markup have one or more semantic and word-formation features, such as "face", "fashion", "space", "speed", "movement", "possession", "human property features" etc. During classification, a single word can belong to several classes. At the first stage, the search is carried out according to the features available in the dictionary.

Text marking is performed automatically using the Semmarkup program (by A.E. Polyakov) according to the semantic dictionary of the corpus. Because manual processing of semantically marked texts is very laborious, semantic homonymy is not removed in the corpus: several alternative sets of semantic features belong to polysemous words.

[1] http://www.ruscorpora.ru/
[2] Резникова Т.И., Копотев М.В. Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) // http://ruscorpora.ru/sbornik2005/04reznikova.pdf
[3] Гулямова Ш. Ўзбек тили семантик анализаторининг лингвистик асослари: Филол.фан.док-ри дисс... – Фарғона, 2022. – Б. 78. – 240 б.
[4] Копотев М. Введение в корпусную лингвистику – Прага, 2014.
[5] https://ozlib.com/1010208/literatura/ponyatie_razmetki

Semantic designation is based on the classification system of the Russian dictionary adopted in the database "Lexikograf" developed in 1992 under the leadership of E.V. Paducheva and E.V. Rakhilina[6]. For the needs of the corpus, the vocabulary was significantly increased, the content was expanded and the structure of semantic classes was improved, word-forming features were added.

Typing the words of the semantic dictionary is based on the morphological dictionary of the DIALING system (total volume of about 120 thousand words). This is an extension of "Grammaticheskiy slovar russkogo yazyka" ("Grammatic dictionary of the Russian language") created by A.A. Zaliznyak[7]. The current version of the semantic dictionary includes the words of important word groups of speech: nouns, adjectives, numbers, pronouns, verbs and adverbs[8].

The problems of lexical-semantic tagging of corpus units in world linguistics were solved in the research of scientists such as E.V.Paducheva, E.V.Rakhilina, A.V.Sannikov[9].

The issue of semantic tagging of Uzbek corpus materials was partially reviewed in Sh.Khamroeva's dissertation[10]. In this study, the importance of semantic group, gang and field classification in semantic tagging of text was studied, and semantic tag models were developed based on the classification of semantic fields of the Uzbek language.

In A. Eshmuminov's research, the concept of semantic markup, tools and methods, the importance of semantic markup for the corpus and its user, as well as its linguistic bases: possibilities and criteria are given in detail.[11]

In D. Akhmedova's research, the models of semantic tagging of noun units and the linguistic support of lexical-semantic tagging of Uzbek language units for the corpus were developed[12].

Significant work is being done on Uzbek computer linguistics. The first sub-corpus of the national corpus of the Uzbek language - the educational corpus - has been created and is being improved[13].

In Uzbek linguistics, there are studies of M. Mahmudov, M. Ayimbetov, S. Karimov on computer linguistics, and lexicographic processing of text[14].

Sh. Hamroeva studied the issue of creating a morpho analyzer of the Uzbek language and its linguistic support as a research object, while Sh. Gulyamova studied the linguistic foundations of the semantic analyzer of the Uzbek language[15].

---

[6] Semantic dictionary viewed as a lexical data base. // COLING-1992. печ. Nantes, 1992. 0,5 Е.В. Падучева, М.В. Филипенко

[7] Зализняк А.А. Грамматический Словарь Русского Языка. Москва, изд-во «Русский язык» - 1980.

[8] https://ruscorpora.ru/page/instruction-semantic/

[9] Апресян Ю.Д., Иомдин Л.Л., Санников А.В., Сизов В.Г. Семантическая разметка в глубоко аннотированном корпусе русского языка // Труды международной конференции “Корпусная лингвистика – 2004”. – Санкт-Петербург: Издательство Санкт-Петербургского университета, 2004. – С.41-54.

[10] Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Фил. фан. бўйича фалсафа доктори (PhD)…дис. – Бухоро, 2018. – 250 б.

[11] Эшмуминов А. Ўзбек тили миллий корпусининг синоним сўзлар базаси. Монография. – Термиз, 2021. – 118 б.

[12] Ахмедова Д.Б. Атов бирликларини ўзбек тили корпуслари учун лексик-семантик теглашнинг лингвистик асос ва моделлари: Филол. фан. Бўйича фал. доктори (PhD) диссертацияси. − Бухоро, 2020. − 247 б.

[13] http://uzschoolcorpara.uz

[14] Махмудов М.А., Пиотровская А.А., Садыков Т. Система машинного анализа и синтеза тюркской словоформы // Переработка текста методами инженерной лингвистики. – Минск, 1982.; Муҳаммедов С.А. Статистический анализ лексико-морфологической структуры узбекских газетных текстов: Автореф. дисс… канд. фил. наук. – Ташкент, 1980.; Бабанаров А. Разработка принципов построения словарного обеспечения турецко-русского машинного перевода: Автореф. дисс… канд. фил. наук. – Л., 1981.; Айымбетов М.К. Опыт лингвостатистического анализа лексики и морфологии каракалпакского публицистического текста: Автореф. дисс… канд. фил. наук. – Ташкент, 1987.; Каримов С., Қаршиев А., Исроилова Г. Абдулла Қаҳҳор асарлари тилининг луғати. Алфавитли луғат. Частотали луғат. Терс луғат. – Тошкент, 2007.; Ризаев С. Ўзбек тилининг лингвостатистик тадқиқи: Фил. фан. док. дис…автореф. – Тошкент, 2008.; Пўлатов А. Компьютер лингвистикаси. – Тошкент: Akademnashr, 2011.; Норов А. Компьютер лингвистикаси асослари. – Қарши, 2017. – 136 б.; Жуманазарова Г.У. Фозил Йўлдош ўғли достонлари тилининг лингвопоэтикаси: Фил. фан. док. дис…автореф. – Тошкент, 2017.; Ўринбоева Д.Б. Ўзбек фольклори матнларининг лингвостатистик тадқиқи. – Тошкент: Фан, 2010.; Ўринбаева Д. Халқ оғзаки ижоди: жанрий-лисоний ва лингвостатистик тадқиқ муаммолари: фил. фан. док. дисс. автореф. – Самарқанд, 2019.− 74 б.

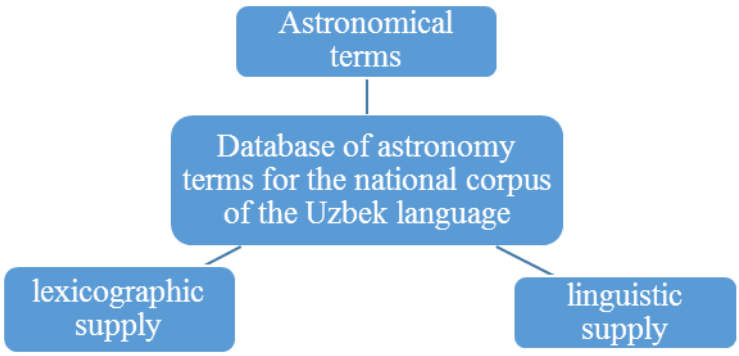

Figure 3.1. The structure of the database of astronomy terms in the national corpus of the Uzbek language

The solution. The process of semantic tagging of astronomical terms of the Uzbek language first requires the creation of a database of astronomical terms. First of all, the structure and content of the terms should be defined. For the educational corpus of the Uzbek language, the structure of the database of terms related to astronomy should be as follows:

The database dictionary of astronomy terms in the national corpus of the Uzbek language is formed on the basis of (the 2-volume and 5-volume) explanatory dictionary[16] of the Uzbek language.

Thus, it was observed that there are a number of problems in the process of semantic tagging of Uzbek astronomical terms. Many words in the Uzbek language are polysemantic and homonymous, which requires the development of perfect filters in the semantic classification system and linguistic models based on them. This issue was resolved in Sh. Gulyamova's doctoral dissertation. Based on her recommendation to eliminate ambiguous, homonymous and polyfunctional words, semantic tagging is appropriate.[17]

To do this, first of all, each of them is separately classified, polysemy, homonymy and polyfunctionality between words are distinguished by a filter and a linguistic model created through it. Therefore, the study of the issue of semantic tagging of astronomy terms in the Uzbek language as a separate study provides a perfect solution to the problematic issues related to the terminology of the field in the educational corpus.

The list of used literature:

1. Apresyan Yu.D., Iomdin L.L., Sannikov A.V., Sizov V.G. Semanticheskaya razmetka v gluboko annotirovannom korpuse russkogo yazyka // Trudy mezhdunarodnoj konferencii “Korpusnaya lingvistika – 2004”. – Sankt-Peterburg: Izdatelstvo Sankt-Peterburgskogo universiteta, 2004. – P.41-54.

[15] Ҳамроева Ш. Ўзбек тили морфологик анализаторининг лингвистик таъминоти. Филология фан. д-ри. дисс. авти. – Фарғона, 2021. – 76 б.; Гулямова Ш. Ўзбек тили семантик анализаторининг лингвистик асослари: Филол.фан.доктори (DSc) диссертацияси. – Фарғона, 2022. – 281 б.

[16] Ўзбек тилининг изоҳли луғати. 2 жилдли. – Москва, 1981; Ўзбек тилининг изоҳли луғати. 5 жилдли. – Тошкент: Ўзбекистон миллий энциклопедияси, 2006. 1-жилд. – 680 б.; Ўзбек тилининг изоҳли луғати. 5 жилдли. – Тошкент: “Ўзбекистон миллий энциклопедияси” Давлат илмий нашриёти, 2006. 2-жилд. – 671 б.; Ўзбек тилининг изоҳли луғати. 5 жилдли. – Тошкент: “Ўзбекистон миллий энциклопедияси” Давлат илмий нашриёти, 2006. 3-жилд. – 688 б.; Ўзбек тилининг изоҳли луғати. 5 жилдли. – Тошкент: “Ўзбекистон миллий энциклопедияси” Давлат илмий нашриёти, 2007. 4-жилд. – 608 б.; Ўзбек тилининг изоҳли луғати. 5 жилдли. – Тошкент: “Ўзбекистон миллий энциклопедияси” Давлат илмий нашриёти, 2007. 5-жилд. – 592 б.

[17] Гулямова Ш. Ўзбек тили семантик анализаторининг лингвистик асослари: Филол.фан.доктори (DSc) диссертацияси. – Фарғона, 2022. – 281 б.

2. Ahmedova D.B. Atov birliklarini ўzbek tili korpuslari uchun leksik-semantik teglashning lingvistik asos va modellari: Filol. fan. Bujicha fal. doktori (PhD) dissertaciyasi. − Buhoro, 2020. − 247 p.
3. Ajymbetov M.K. Opyt lingvostatisticheskogo analiza leksiki i morfologii karakalpakskogo publicisticheskogo teksta: Avtoref. diss… kand. fil. nauk. – Tashkent, 1987.
4. Gulyamova Sh. Uzbek tili semantik analizatorining lingvistik asoslari: Filol.fan.dok-ri diss... – Fargona, 2022. – B. 78. – 240 b.
5. Zaliznyak A.A. Grammaticheskij Slovar Russkogo Yazyka. Moskva, izd-vo «Russkij yazyk» - 1980.
6. Eshmuminov A. Uzbek tili millij korpusining sinonim sўzlar bazasi. Monografiya. – Termiz, 2021. – 118 b.
7. Karimov S., Қarshiev A., Isroilova G. Abdulla Қahhor asarlari tilining lugati. Alfavitli luғat. Chastotali luғat. Ters luғat. – Toshkent, 2007.
8. Mahmudov M.A., Piotrovskaya A.A., Sadykov T. Sistema mashinnogo analiza i sinteza tyurkskoj slovoformy // Pererabotka teksta metodami inzhenernoj lingvistiki. – Minsk, 1982.
9. Muhammedov S.A. Statisticheskij analiz leksiko-morfologicheskoj struktury uzbekskih gazetnyh tekstov: Avtoref. diss… kand. fil. nauk. – Tashkent, 1980.; Babanarov A. Razrabotka principov postroeniya slovarnogo obespecheniya turecko-russkogo mashinnogo perevoda: Avtoref. diss… kand. fil. nauk. – L., 1981.;
10. Norov A. Kompyuter lingvistikasi asoslari. – Қarshi, 2017. – 136 b.; Zhumanazarova G.U. Fozil Jўldosh ўғli dostonlari tilining lingvopoetikasi: Fil. fan. dok. dis…avtoref. – Toshkent, 2017.; Ўrinboeva D.B. Ўzbek folklori matnlarining lingvostatistik tadқiқi. – Toshkent: Fan, 2010.
11. Reznikova T.I., Kopotev M.V. Lingvisticheski annotirovannye korpusa russkogo yazyka (obzor obshedostupnyh resursov) // http://ruscorpora.ru/sbornik2005/04reznikova.pdf
12. Rizaev S. Uzbek tilining lingvostatistik tadқiқi: Fil. fan. dok. dis…avtoref. – Toshkent, 2008.; Pulatov A. Kompyuter lingvistikasi. – Toshkent: Akademnashr, 2011.;
13. Urinbaeva D. Halk ogzaki izhodi: zhanrij-lisonij va lingvostatistik tadқiқ muammolari: fil. fan. dok. diss. avtoref. – Samarқand, 2019.− 74 b.
14. Hamroeva Sh. Uzbek tili morfologik analizatorining lingvistik taminoti. Filologiya fan. d-ri. diss. av-ti. – Fargona, 2021. – 76 b.; Gulyamova Sh. Ўzbek tili semantik analizatorining lingvistik asoslari: Filol.fan.doktori (DSc) dissertaciyasi. – Farғona, 2022. – 281 b.
15. Hamroeva Sh. Uzbek tili mualliflik korpusini tuzishning lingvistik asoslari: Fil. fan. bўjicha falsafa doktori (PhD)…dis. – Buhoro, 2018. – 250 b.
16. Uzbek tilining izoҳli luғati. 2 zhildli. – Moskva, 1981; Ўzbek tilining izoҳli luғati. 5 zhildli. – Toshkent: Uzbekiston millij enciklopediyasi, 2006. 1-zhild. – 680 b.;
17. Uzbek tilining izohli luғati. 5 zhildli. – Toshkent: “Uzbekiston millij enciklopediyasi” Davlat ilmij nashriyoti, 2006. 2-zhild. – 671 b.;
18. Semantic dictionary viewed as a lexical data base. // COLING-1992. pech. Nantes, 1992. 0,5 E.V. Paducheva, M.V. Filipenko
19. https://ruscorpora.ru/page/instruction-semantic/
20. http://www.ruscorpora.ru/
21. http://uzschoolcorpara.uz