HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

THEORETICAL AND PRACTICAL MODELING OF THE ANNOTATION OF LEXICAL UNITS IN THE ALISHER NAVOI CORPUS

Nargiza Gulomova
Navoi innovation university
nargizagulomova099@gmail.com
ORCID: 0000-0002-7716-1799
doi:https://doi.org/10.5281/zenodo.23801768

Abstract. Semantic analysis of natural language data is considered important in solving many problems in a number of directions, such as natural language processing (Natural Language Processing – NLP), natural language technologies (Human Language Technology - HLT). In particular, lexical units are semantically tagged to identify such phenomena as homonymy, polysemy, syncretism in the language corpora. It is precisely for this feature that corpora are used in semantic research. This article discusses the importance of tagging for language corpora, the standard tagging process, semantic tagging tools, the role of annotation, and two different approaches to semantic tagging for Alisher Navoi's corpus. Furthermore, the article examines issues related to the theoretical and practical modeling of the annotation of lexical units within a corpus based on the works of Alisher Navoi. It outlines the methodological principles for digitizing classical literary heritage, performing linguistic processing, and developing a system for semantic-grammatical tagging. The study substantiates the scientific and practical significance of models for encoding and structuring lexical units to facilitate the automated analysis of texts. The findings serve as an essential theoretical and methodological foundation for Uzbek computational linguistics, corpus linguistics, and the in-depth study of the vocabulary found in the works of Alisher Navoi.

Keywords: semantic annotation, theoretical and practical modeling, corpus linguistics, semantic tagging.

Enter. The term tag is actively used in world corpus linguistics, and its importance for the corpus is incomparable. It is known that texts are the basis of the corpus, and texts are composed of various lexical units. Natural Language Processing (NLP) takes into account all linguistic and non-linguistic elements of language as they occur in text. A tag is a key component of the global web language HTML (Hyper Text Markup Language) and is a special word enclosed in square brackets (<...>). In HTML, tags are not visible to the user. HTML and XML are implementations of the more complex language SGML (Standard Generalized Markup Language). HTML is used to display data, while XML (eXtensible Markup Language) is used to store and organize data. Thus, SGML, HTML and XML are document markup languages, and the tag is their main element. Unlike them, the tags in the linguistic corpus specifically mark text elements morphologically, syntactically, semantically, stylistically, prosodicly, anaphorically and stylistically [1]. Part of Speech tagging (POST) was developed for the LOB corpus at Lancaster University in the 70s of the XX century.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

The purpose of tagging text units is to encode text, that is, to explain text units to digital technology. As a result, it becomes possible to manage lexical and even non-lexical units. Language corpus differs from other types of text array systems and makes it important, and also increases the necessity and effectiveness of the results for language learning, educational process, and scientific research [2,3,6].

First, let's talk about the terms tag, tagging, annotation, linguistic annotation. Linguistic and extralinguistic information should be given to the text to perform various linguistic tasks/actions. For this, it is necessary to give a special explanation to the components of the existing text (for example, information about the word group, its meaningful expression) [17, 18]. This annotation is called a text annotation or linguistic annotation, and special symbols are called tags (Russian: razmetka, English: tag). Interpretation of text units is called annotation or linguistic interpretation (annotation), and symbolization of each linguistic unit is called tagging (rasmetka, tagging, mark-up).

For example, a noun group is labeled as “Ot” or “N”, an adjective group as “Sif” or “Adj”, a verb group as “F” or “V”. In the text, the word form “my book” belonging to the noun group <Ot, n-b nom., birl., I sh., p. eg.q.> is annotated, i.e. linguistically interpreted. The simplest example of tagging and annotation is general tagging and general annotation of phrases. It might look like this:

I am proud of my son.

Explain

I am (pronoun) proud of (verb) my (pronoun) son (subyekt)

Tags are service marks in text that contain information about the text itself.

For tagging corpus, several universities have specifically developed a system that describes which parameters of texts should be tagged. This system uses XML and is called the Text Encoding Initiative Guidelines (TEI Guidelines) [http://www.teic.org/Guidelines/index.xml]. 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 corpus development projects (including the British National Corpus) try to follow TEI recommendations in one way or another [19].

Traditionally, in the language of digital technology, tags are in pairs, i.e. opening and closing, in angle brackets. For example, <a> is an opening tag, is a closing tag (exactness in the term tag in information technology and natural language processing shows the importance of linguistic ontologies). The closing tag indicates the end of the message given in the opening tag.

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

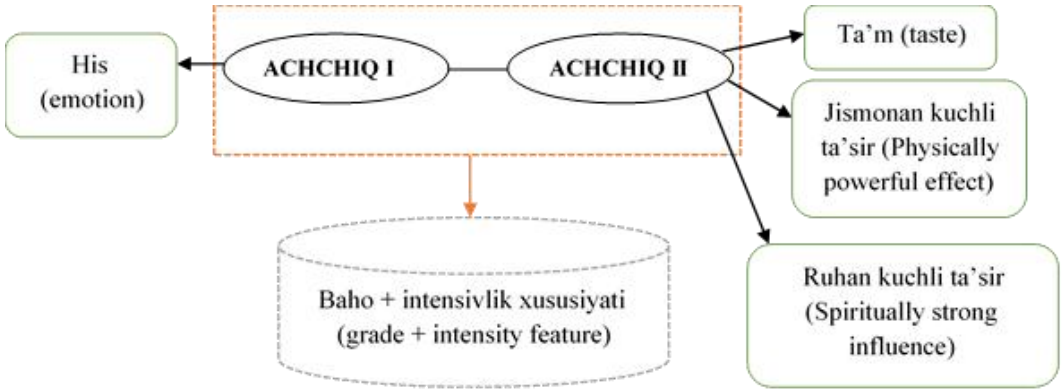
As it is clear, it is indicated that the lexical unit at the beginning of the sentence is the pronoun <pron>Men</pron>.

Default semantic tagging. In semantic tagging, lexical-semantic groups of lexical units are defined using special tags. A semantic tag is a special sign of a semantic group belonging to a lexical unit in the text, and semantic tagging is a special designation of semantic groups or categories. Semantic annotation (rasmetka) - specifying the meaning of a word or phrase in the texts of language corpora, indicating that it belongs to a certain semantic category or an even smaller semantic group (lexical-semantic group, semantic field and group). symbol, annotation set [5,8]. Semantic tags of the

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

corpus are the specification of word meaning(s), that is, their specific features, the creation of a set of comments related to homonymy, synonymy, antonymy, word categorization, determination of belonging to the semantic field, derivational characteristics , covers symbols such as fire meaning. Semantic tagging uses codes consisting of letters, abbreviations, numbers and words, or numbers only, although there is no single specific form. Coding and decoding is the process of transferring information into a language that can be understood by a machine (computer) and, conversely, from a language that can be understood by a computer into a language that can be understood by a person [16].

As it is known, in semantic tagging, the lexical unit belongs to a certain semantic category, and based on this, its semantic comments (lexical-semantic group, semantic field, etc.) are shown to the user. In this case, in semantic tagging, faceted classification is carried out according to the word's belonging to several semantic fields (chart 1).



Drawing 1. Belonging to the semantic field of the homonym "bitter" in the form of a faceted classification

Homonymous and polysemantic words found in texts are included in several semantic groups. For example, the word "ключ" in Russian is also a polysemous word, I. 1) thing: ключ от дверей – the key of the door; 2) technical equipment: гаечный ключ; 3) term: musical key. II. the key is the spring. The homonym given by the number I represents 3 meanings, and the homonym II is applied to the place where the water comes out and has now acquired the homonym. The semantic group they belong to can be understood from their meaning in the text.

In the National Corpus of the Russian Language [https://ruscorpora.ru/ access time: 2024.22.01] (NKRYA), several types of symbols are attached to one word form based on the principles of semantic tagging. Ambiguous words are assigned several alternative sets of semantic properties. Corpus acts as a source of automatic-semantic classification, providing various linguistic information. For example, when the word rodina was searched by the corpus, 4973 concordances were given out of 2917 texts belonging to 1730 authors, and the number of concordances with other words of this lexeme was 36449 out of 11931 texts [https://ruscorpora.ru/explore?req=родина]. It is noteworthy that the semantic bases of the lexeme rodina, additional semantic information, its synonyms and similar expressions are also shown in the corpus (table 1).

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Table 1. Grammatical, syntactic, semantic and additional comments (annotations) given to the lexeme rodina in the national corpus of the Russian language

Lemma	homeland
Grammar	noun, animate, feminine, nominative, singular
Basic semantics	subject names, adjectival, space and place
Additional semantics	subject names, adjectival, space and place
Similar words	fatherland (0.754806) – the land where the umbilical mine was spilled, fatherland (0.735216) – country, foreign land (0.62975) – a foreigner, fellow countryman (0.466587) – compatriot, ashes (0.463531) – culba
Syntactic properties of the word	subject (subject), subordinate clause, verb clause, there are dependent
Add. signs	capital, genderred, numred

All existing texts in the national corpus of the Russian language are semantically tagged according to their lexical-semantic properties. Words and phrases related to one semantic group are attached with special codes. For example:

– indicators of ethnic name - t:hum:etn;

– units meaning relatives - t:hum:kin;

– words denoting the content of plants - special ID codes such as t:plant are semantically tagged.

In general, a semantic tag combines words and phrases belonging to the same semantic group. Words belonging to the same semantic group are identified by one symbol. In order for the user or researcher to achieve efficiency in the search, the semantic field is displayed in the search filter, and the process is based on semantic tags.

Source analysis. Semantic tagging and annotation of texts was initially performed on the basis of dictionaries [10]. Modern semantic tagging systems use semantic networks or lexical classifications rather than dictionaries, among which WordNet [12, 7] is the most popular. For the first time, 234.136 word forms were semantically tagged based on Brown's corpus WordNet. 186.575 of them were multi-meaning [15]. The SemCor corpus created by Princeton University has 700.000 word forms, and 200.000 of them were manually tagged based on WordNet 1.6 version, and later, lexical units were automatically semantically tagged based on WordNet 1.7-2.0 versions.

Ideological semantic classification of lexical-semantic tagging of the texts of the national corpus of the Russian language [9], semantic filter of the corpus [20] have been thoroughly studied. In semantic tagging, at first, codes were used, then they were simplified to be understandable to the user, and semantic groups were presented [21].

Semantic tagging process in authorship corpus of Alisher Navoi. Semantic tagging in Alisher Navoi's corpus has a special basis, that is, essence. It is known that the language of the 15th century is different from the current Uzbek language, and there are many lexical units that are out of use or

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

whose formal expression has changed for today's reader. We call them explanatory words [4, 5]. It is unreasonable to show the semantic group or semantic field of such words without clarifying them to the user. Therefore, in our project, in order to expand the audience of those who read Alisher Navoi's works and to serve as an auxiliary tool in understanding the content of ghazals, in this corpus, at first, attention was paid to providing a contextual understanding of the words. Therefore, the process of semantic tagging for Alisher Navoi's authorship corpus has become the following approach:

In the semantic tagging of spiritual groups of words, the approach of marking the semantic group of lexical units with special tags and opening the semantic explanation of words with ID codes is taken into account. Since the main focus in the semantic tagging of Alisher Navoi's authorship corpus is on descriptive words, work is being done on the first approach.

Bringing Alisher Navoi closer to the people and turning his priceless spiritual heritage into the heart of our youth is not a one-day job. In the continuing education system, the teaching of Navoi's life and works based on the student's age is considered the most appropriate way to achieve effective education. The level of complexity of the words used in the text of ghazals, special codes were also used to determine the age of the audience for each ghazal, taking into account the scope, the idea expressed in the ghazal, and entering it into the database. In Alisher Navoi MK, the age of the audience is given in the metadata provided for each ghazal, from 15 to 18 years old (15+,16+,17+,18+) and separated by special codes when entering the corpus base: 15+1, 16+2, 17+3, 18+4. Determining the audience age of ghazals by means of special codes not only makes it convenient for the educational system, but it is also convenient for the general public to learn and memorize ghazals by selecting them accordingly [8].

In 2023, the practical project "Creation of Alisher Navoi's Authorship Corpus" was supported by the Innovative Development Agency under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan, and the members of the project were Alisher Navoi's "Badoye' ul-vasat" continues the work he started on the basis of his office. The main goal of this project is to create a semantically tagged database of 2600 ghazals in Alisher Navoi's "Khazayin ul-maoniy" collection ("Garayib us-sigar", "Navodir ush-shabab" and "Favoyid ul-kibar") as a result of creation and formation of Alisher Navoi's authorship corpus with the help of these databases, the audience of those who read and study Alisher Navoi's work from the 15th century sources, and those who understand Navoi is to be expanded.

As a result of semantic tagging of 650 ghazals in the "Badoye ul-vasat" department of Alisher Navoi's "Khazayin ul-maoni" kulliyati and carrying out research [13], the following results were achieved:

1. Linguistic and technical foundations of creating a corpus of works belonging to the pen of Alisher Navoi, a phenomenon of Turkic peoples, have been developed.

2. The words (phrases) in Alisher Navoi's works that are difficult for today's readers to understand are explained.

3. Semantic tagging of 22,859 explanatory words in 650 ghazals consisting of 10002 verses.

4. A copyright certificate was obtained from the Intellectual Property Agency for the semantically tagged base of ghazals belonging to the "Badoye' ul-vasat" department.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

5. Alisher Navoi's authorship corpus was created and a certificate was obtained.

6. In the corpus, a base of comments on proverbs, expressions, famous historical figures and place names is also formed (figure 1).

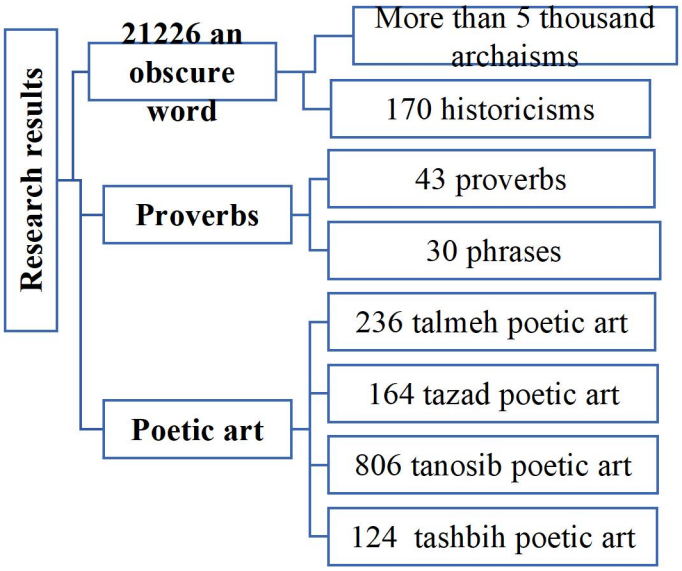


Figure 1. The amount of contextual semantic tags in 650 ghazals from Alisher Navoi's book "Badoye' ul-vasat".

7. A prototype of this project was prepared and presented to users, and a video instruction on use was also prepared (https://youtu.be/744o55Tqx9E).

8. The design of the body is developed with a national color.

Currently, within the framework of the project, in order to enrich the corpus base with other works of Alisher Navoi, 1950 ghazals in the remaining three divans of his collection "Khazayin ul-maoni" are being semantically tagged. As a result, the semantic tags of specific nouns in ghazals formed the following 4 groups (table 2):

Table 2. Semantic tags of proper nouns in Alisher Navoi authorship corpus

Semantic tags	Semantic field name
t:persn	Names (Jomiy, Yusuf, Nuh, Sulaymon, Muso, Iso, Maryam,)
t:topon	City (Misr, Eron, Batho, Buxoro, Bobil, Bojun – Xitoy, Bistom, Gulpoyagon – Erondagi tuman, Gurja)
t:divine understanding	Divine concepts (Ka'ba, Olloh, Rab, Bul Bashar – Odam Ato, Buroq, Iso, Muso, Sulaymon, Yusuf, Xizr, Dovud, Jamshid, Kavsar,
t:mountain	Mountains (Mano)

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

In the corpus, also 39 cognate groups of nouns,11of nouns denoting part (meronym), 24 abstract nouns, 17 adjectives, 6 numerals, 5 verbs, and 6 groups of lexical-semantic tags of the adverb series were formed(2-figure).

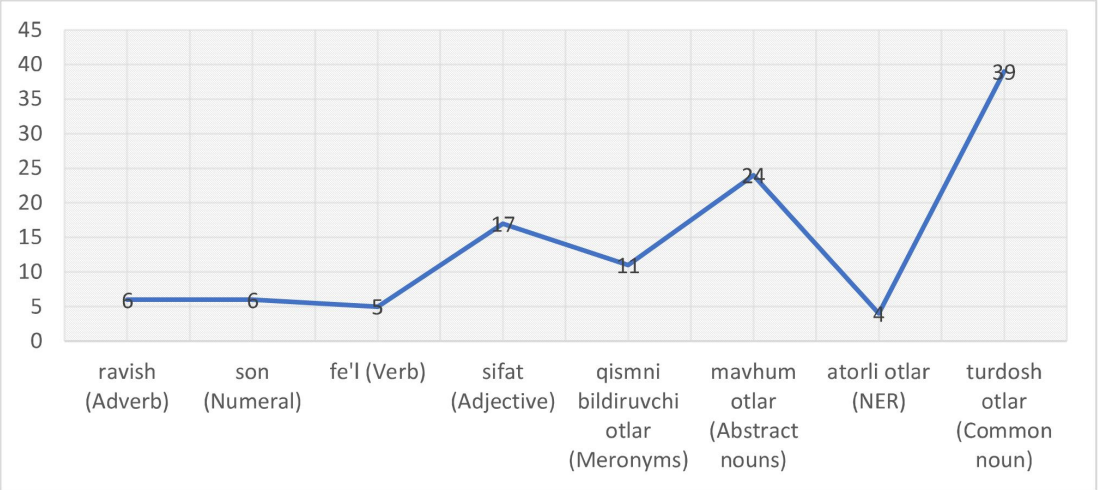


Figure 2. Statistics of lexical-semantic groups in Alisher Navoi's authorship corpus.

The scientific contribution of this article is that in the semantic tagging of the author's corpora of classical artists like Alisher Navoi, not only standard tagging (approach 2), but also semantic tagging of the contextual understanding of words that are incomprehensible to the modern user (approach 1) is also justified. In the second approach, the semantic field or semantic groups of a word are determined, as a result of which a semantic filter is opened in the search and effective search results are obtained, the possibility of semantic analysis of ghazals increases; and on the basis of the first approach, the contextual concept of the word is defined by explanation. As a result, the competence of reading and understanding Alisher Navoi's ghazals is formed. These two approaches provide many conveniences in the study and research of medieval sources. Particularly, it expands the audience of those who read and understand the texts.

Conclusion. Semantic tag is an indicator that shows whether lexical units belong to a specific semantic field or a semantic group in the corpus. In this case, carrying out various studies on lexical units allows to reveal the semantic scope of lexical units in the educational process, to determine homonymy and polysemous words from the context [6]. In Alisher Navoi's authorship corpus, the primary task is to interactively display the contextual semantics of words by explaining them to the user, and thus to expand the readership of Navoi's work, so firstly, the concept of explanatory words in ghazals and what they mean in verse is coded in the form of semantic tag ID was entered into the database and this process formed the essence of semantic tagging in the corpus. In this paper, we have explored a custom tagging approach for the Alisher Navoi authorship corpus in addition to standard tagging. From experience, it became clear that it is more important for the user to know the specific meaning of a word in the context of Alisher Navoi's works than to determine its semantic field. Therefore, the </contextual concept> tag has been customized in Alisher Navoi's authorship corpus. The lexicon of the era in which Navoiy lived differs from contemporary vocabulary in many respects.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Specifically, to ensure that archaisms, historicisms, and terms that have undergone semantic narrowing—which often require explanation—remain intelligible and accessible within his *ghazals*, it is necessary to apply semantic tagging and conduct lexical and morphological analysis. Within a corpus, categorizing polysemous words into semantic groups—based on the specific meaning each word conveys in a given line of verse—can help users expand their own vocabulary. When developing an author-specific corpus, it is essential to provide a comprehensive account of the semantic characteristics of the included texts and to organize the lexemes employed by the author into lexical groups under specific tags. By classifying the vocabulary found in his *ghazals* into lexical-semantic groups, one can gain a precise, thorough, and rapid understanding of the poet's individual style and his mastery of word usage.

References:

1. Abjalova M. Korpus lingvistikasi. [Matn]: uslubiy qo'llanma / M.A.Abjalova. – Toshkent: Bookmany print, 2022. – 102 b.
2. Abjalova M. Tagging and Annotation of Corpus Units. // International Journal of Language Learning and Applied LinguisticsISSN: 2835-1924 Volume 2 | No 12 | Dec -2023. – PP. 103-107. https://inter-publishing.com/index.php/IJLLAL/article/view/3228/2733
3. 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.
4. 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
5. 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
6. 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
7. Abjalova M.A. O'zbek tili ontologiyasi: yaratish texnologiyasi va konsepsiyasi. [Matn] : monografiya / M.A. Abjalova. –Toshkent: Nodirabegim, 2021. – 215 b. ISBN 978-9943-7804-5-3
8. Abjalova M.A., G'ulomova N.S. Alisher Navoiy mualliflik korpusi va uning imkoniyatlari. // Kompyuter lingvistikasi: muammolar, yechim va istiqbollar. Xalqaro ilmiy-amaliy konferensiya to'plami. Elektron nashr / ebook. – Toshkent: ToshDO'TAU, 22.04.2022. – B. 89-93.
9. Atkins S.Tools for computer-aided corpus lexicography: the Hector project // Acta linguistica Hungarica41, 1993.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

10. Daniel Jurafsky and James Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, volume 2. 02 2008. – 988 p.
11. Dolan W, Vanderwende L., Richardson S. Polysemy in a broad-coverage natural languigeprocessing system //Ravin Y., Leacock C. (eds.) Polysemy: Theoretical and Computational Approaches. N.-Y.: Oxford University Press, 2002.
12. Fellbaum C.,Grabowski J., Landes S. Performance and confidence in a semantic annotation task // Fellbaum C. (ed.)WordNet:An Electronic Lexical Database. Cambridge (Mass.): The MIT Press, 1998.
13. G‘ulomova N. Alisher Navoiy mualliflik korpusi va uning semantik teglari bazasini yaratish (“Badoye’ ul-vasat” devoni asosida): Fil.f.b.f.d. ilm.daraj. (PhD) olish uchun yozil. diss. – Toshkent, 2023. – 190 b.
14. https://ruscorpora.ru/. Murojaat vaqti: 2024.22.01
15. Miller G., Leacock C., Randee T., Bunker R. A semantic concordance // Proceedings of the 3rd DARPA Workshop on Human Language Technology, Plainsboro, New Jersey, 1993. P. 303–308.
16. Rahimov A. Kompyuter lingvistikasi asoslari. – Toshkent: Akademnashr, 2011. – B.19.
17. Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language. CEUR Workshop Proceedings, Vol. 3315. – pp. 93-98.
18. Tukeyev, U.A. Computational morphology of Turkic languages: textbook. Almaty: Qazaq university, 2023. – 161 p.
19. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск: ИГЛУ, 2011. – С.48-49.
20. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.
21. Шеманаева О.Ю., Кустова Г.И., Ляшевская О.Н., Рахилина Е.В. Семантические фильтры для разрешения многозначности в национальном корпусе русского языка: прилагательные // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007». М.: 2007. – С. 582-587.