UDC 81'33
SRSTI 16.21.51
DOI 10.48371/PHILS.2022.64.1.001

SEMANTIC MARKUP IS ONE OF THE COMPONENTS OF THE NATIONAL LANGUAGE CORPUS

*Amirbekova A.B.[1], Talgatkyzy G.[2] , Urakova L.[3]
*[1]Candidate of Philological Sciences, Institute of Linguistics named after A. Baitursynov, Almaty, Kazakhstan
[2] PhD student, al-Farabi Kazakh National University, Almaty, Kazakhstan
[3] PHD Akdeniz University, Antalya, Turkey
*[1]e-mail: Marghan01@mail.ru
[2] e-mail: gulalyzhan@mail.ru
[3]e-mail: urakovayanc@gmail.com

Abstract. The article describes the principles of semantic markup in the National Corpus of the Kazakh language. The purpose of the article is to consider and develop a system of semantic tags ready for use in the language corpus. This approach is based on the semantic classification of vocabulary and is universal and applicable to any language. The practical significance of dictionary and text corpus markup is to improve the quality of search and expand user capabilities. The scientific significance of the article is determined by the fact that the markup and semantic classification should be focused on any programming paradigm. We have chosen a functional paradigm. The main results of the article are, firstly, the semantic marking of national corpora significantly improves the quality of the search and expands the user's capabilities when requesting linguistic information; secondly, the semantic information about each token in which an entry is made is presented as a set of semantic markups or tags and is usually reflected in the semantic classification of the language's vocabulary. Conclusions are drawn about further possibilities of using corpus data for modern studies of lexical and grammatical semantics.

The publication was made within the framework of a scientific project on No. BR11765619 on the topic " Development of the national corpus of the kazakh language as information-innovation state language base: research and training internet resource", supported by the Ministry of Education and Science of the Republic of Kazakhstan.

Keywords: semantic tags, semantic classification, markup, vocabulary, subgroup, subcorpus, national corpus.

Introduction

Recently, corpus linguistics has been rapidly developing, which is undoubtedly due, on the one hand, to the emergence of "breakthrough" developments in computer technology and, on the other hand, the rapid development of the Internet and Internet technologies, the very existence of which can be defined as an impetus to the progress of information and communication

systems in general. Studies conducted on the material of texts corpus in a particular language have made it possible to correct the existing descriptions of languages. For example, on the material of the National Corpus of the Kazakh language, new dictionaries, grammars have been created. Research is being carried out on the problems of general and applied linguistics and particular issues. Among special programs for natural language processing, automatic markup programs occupy a special place.

The development of new types of markup and the constant replenishment of language corpora with new materials are important for a simple user, since the existence of parallel electronic resources with different approaches, different markups create favorable opportunities for a deeper and more versatile study of the structure of the language.

In developing the tagging system for semantic annotation, we relied on the markup for the Russian National Corpus [1, p. 381], the Bashkir language corpora [2, p. 23], the Chinese language corpora [3, p. 24], the multimedia colloquial Turkish language corpora [4, p.9], and also on works devoted to taxonomic groups [5, p. 13]. The reference point in solving many issues that arise when creating the Kazakh language corpus is, first of all, the National Corpus of the Russian Language, the first Russian experience in developing an information and reference system with semantic markup. The developers of the corpus themselves, in turn, were based on the Lexicography system, which, of course, was redesigned taking into account the features of the corpus [6 , pp.156–158].

Description of materials and methods

What is a semantic markup for? Let's answer this seemingly simple question in the words of S. Gindin: “There are two approaches to the description of the language - “from form to meaning” and “from meaning to form”. In the first case, the descriptor's task is to characterize as accurately and completely as possible all the existing meanings and ways of using a certain linguistic element. In the second approach, on the contrary, all the ways of embodying some content, some semantic complex, available in a given language, are listed and described. Both of these approaches complement each other and are equally necessary for linguistics and language teaching. And the “user”, the addressee of linguistic products, whether they are grammars and dictionaries of his native language or the language that he is still learning, needs descriptions of both types. If you come across an unfamiliar word or an unknown construction, you reach for a dictionary or list of grammatical elements, which will list the meanings of the stranger who puzzled you. But when you speak or write, how often do you find yourself in the position of a mute: what you want to say, you seem to know or at least feel, but you can’t find the right word or grammatical form” [7, p. 56].

Indeed, semantic markup is very important for solving the problems of studying vocabulary, in particular, the problems of word compatibility and its` syntactic. There are two approaches to creating a semantic annotation: faceted and tree classification. The latter is implemented in the Russian Semantic Dictionary [8, p. 21] in the form of a tree, where each label is a derivative of the main one. When

conducting a faceted classification, the researcher starts with the idea that several independent classifications can be used simultaneously for various reasons. One or another lexical unit, which may have many features, has the potential for a non-intersection of these features. In the analysis, both classifications were used, since, in our opinion, it is rather difficult to choose one type of material classification. According to A. A. Kretov: "Lexico-thematic markup: the allocation of taxonomic groups, mereology, topology, if it was possible to determine belonging to a particular themetaxon" [9, p.126]. The traditional approach to linguistic markup is implemented within the framework of the TEI (Text Encoding Initiative) [10, p.6] project and the XCES (Corpus Encoding Standard for XML) [11] standard. The advantages of markup in the TEI standard are its completeness, based on a carefully developed theory of the structure of text and document, ease of personalization and adaptation to a specific material due to the modular organization and specification mechanism, independence from a specific platform or software product [3, p.24].

Another approach to the organization of structural-semantic markup, proposed in 2004 by C. Tantek, is based on practical experience in working with a networked electronic environment and proposes to solve the problem of introducing new elements into a hypertext markup language based on microformats. In 2011, the creators of the largest search engines united in the Schema project. org - an initiative to develop a unified schema for semantic markup based on the effective structuring of the supplied information resources and their semantic markup with microformats.

Microformats are entities on top of HTML that can be used to describe any information on Web pages. The microformat specification is a way to mark up content to define specific types of information such as testimonials, information about a person, or an event. The standard is a set of classes that describe all kinds of entities and their properties. Now there are several hundred of them.

Meta-marking is effective in studying the conditions for the existence of a language, identifying relationships in it, and studying individual subsets of a language. XML (Extensible Markup Language) has become the standard for semantic description, bringing metadata to the electronic information environment that describes the structure, properties, and semantics of information resources.

Of particular importance for automatic text, analysis is linguistic markup itself, which consists in assigning special labels to texts (their components), which provide the ability to automatically identify texts according to various parameters, to carry out their syntactic and semantic analyzes.

A. L. Sharandin, who develops the principles of "lexical grammar", points out the special significance of the defectiveness of the paradigm for this type of linguistic description. Moreover, we will quote, “in the ontological aspect, the semantic defectiveness is not the actual defectiveness. The uniqueness of the sign content represented by the lexical meaning lies in the fact that the word tends to have only its own set of grammatical forms” [12, p. 25].

The premise of this work is the study, the results of which are presented in the article [13,p.349]. As well as the development of grammatical word profiles [14, p.721].

Modern systems of semantic markup use binding not to dictionaries, but to semantic networks or lexical classifications, among which the most popular is WordNet (http://wordnet.princeton.edu/), which uses the division into values from the Oxford Advanced Learners Dictionary (OALD).

The first subcorpus of the Brown Corps was marked on its basis [Miller et al. 1993], containing 234,136 marked-up word usages, of which 186,575 are polysemantic.

Then came the LEXAS system [Ng & Lee 1996], which manually labeled 192,800 usages of the two hundred most common nouns and verbs. The SemCor corpus [Fellbaum et al. 1998], created at Princeton University, contained 700,000 words, 200,000 of which (full words) were manually marked up according to WordNet 1.6 values, and subsequently automatically recoded into WordNet 1.7.-2.0.

Lexical classifications, originating from ontologies, are less sensitive to semantic nuances. They distinguish between two meanings of a word only if one of them belongs to class X and the other to class Y. Such, in particular, is the classification of lexical units used in the FrameNet project (http://framenet.icsi.berkeley.edu/), the classification of the SenseLearner system developed in Lancaster [Scott Songlin Piao et al. 2005], as well as taxonomies developed for the corpora of the Russian language - the corps of the Laboratory of General and Computer Lexicology and Lexicography of the Faculty of Philology of Moscow State University, the Syntax Corpus.

Results

Markup (tagging, annotation) consists in attributing special tags (tag, tags) to texts and their components: external, extralinguistic, structural, and linguistic proper, describing lexical, grammatical, and other characteristics of text elements.

The set of these metadata largely determines the opportunities provided by corpora to researchers. When choosing these data, it is necessary to be guided by the objectives of the study and the needs of linguists, as well as the possibilities for introducing certain additional features into the text. Linguistic markup types include:

- morphological markup. Morphological marks include not only a feature of a

part of speech but also features of grammatical categories characteristic of a given part of speech. This is the main type of markup: firstly, most large corpora are just morphologically marked up corpora, secondly, morphological analysis is considered as the basis for further forms of analysis

- syntactic and semantic, and, thirdly, advances in computer morphology make it possible to automatically mark up large hulls;

- syntactic markup resulting from parsing. This type of markup describes syntactic links between lexical units and various syntactic constructions.
- semantic markup. Although there is no unified semantic theory for semantics, most often semantic tags denote the semantic categories to which a given word or phrase belongs, and narrower subcategories specify its meaning;
- anaphoric markup. Fixes referential links. prosodic markup. In prosodic corpora, marks are used to describe stress and intonation. In colloquial speech corpora, prosodic marking is often accompanied by so-called discourse marking, which serves to indicate pauses, repetitions, reservations, etc.

There are other types of markup.

The technological process of creating a case can be represented as the following steps or steps.

1. Definition of the list of sources.

2. Digitization of texts (conversion to computer form). It should be said that as far as the task of entering texts into a computer was difficult and time-consuming before, today this problem is solved quite easily, at least concerning modern texts and modern orthography.

This ease is based on advances in optical input (scanning) and recognition of textual information and the global computerization of modern life, including in areas related to the processing of textual information.

Texts in electronic form for creating corpora can be obtained in a variety of ways - manual entry, scanning, author's copies, gifts and exchange, the Internet, original layouts provided to compilers of corpora by publishers.

3. Text preprocessing. At this stage, all texts obtained from different sources undergo philological verification and correction. A bibliographic and extralinguistic description of the text is also being prepared.

4. Converting and graphemic analysis. Some texts also go through one or more stages of pre-machine processing, during which various kinds of re-coding (if required), removal or transformation of non-text elements (figures, tables), removal of hyphens, “hard-line endings” from the text, ensuring uniform writing dashes and so on.

As a rule, these operations are performed automatically. Usually, at the same stage, the segmentation of the text into its structural components is carried out.

5. Text markup. Text markup consists in attributing additional information (metadata) to texts and their components. The meta description of corpus texts includes both meaningful data elements (bibliographic data, features that characterize the genre and style features of the text, information about the author) and formal ones (filename, encoding parameters, markup language version, performers of work stages). This data is usually entered manually. Structural markup of a document (selection of paragraphs, sentences, words) and proper linguistic markup are usually carried out automatically.

6. At the next stage, the results of automatic markup are corrected: error correction and disambiguation (manually or semi-automatically).

7. The final stage is the conversion of marked-up texts into the structure of a specialized linguistic information retrieval system (corpus manager), which provides fast multi-aspect search and statistical processing.

8. And finally, providing access to the body. The package may be available within the display class, may be distributed on CD-ROM, and may be available in WAN mode. Different categories of users may be granted different rights and different options.

Of course, in each specific case, the composition and number of procedures may differ from those listed above, and the actual technology may turn out to be much more complicated.

Semantic markup allows the user to create meaningful queries when searching for examples of the use of a word in a certain meaning, and for the developers of the Corpus to create semantic filters for automatic disambiguation.

Within the lexico-semantic group, subclasses are distinguished, of which the largest (about a hundred units in each subclass) and the most interesting for analysis are nouns and verbs.

The principles of the semantic description of lexemes in the dictionary were described in detail in the publications of Kobritsov, Lyashevskaya, Kustova, we will only say that each meaning of a word is given by a set of semantic labels indicating that a lexeme belongs to one or another lexical class, for example:

Museum

1) "objective name", "spatial object" (go to the museum);

2) "objective name", "collection" (artist's museum);

3) "objective name", "organization" (national museum).

This semantic markup program transfers into the text sets of features that describe all the meanings of a word; the task of subsequent filters is to select the correct one and remove the rest.

Many polysemic words do not require disambiguation, for example, institute

1) A higher educational institution;

2) Research institution;

3) In pre-revolutionary Russia: a closed (with a boarding school) women's secondary educational institution for the children of the nobility. All three values are described in the same way: "subject name", "organization" [15, p.137].

Discussion

The basis of further research will be three theses with varying degrees of scientific strength and novelty.

The first of them is that the lexical semantics of a word contains the possibilities of its speech use. This hypothesis is the most general and therefore of little interest. It is accepted by most linguists and is essentially a truism. Because different groups of nouns manifest themselves differently in syntax, and this depends primarily on their meaning.

The second thesis: the lexical semantics of a word determines the features of the implementation of its morphological paradigm. Accordingly, the lexico-semantic group of words should be characterized by some generality of the paradigm (dominants are distinguished here based on average values). This idea is very relevant for modern grammar.

The third thesis: the set of word usages of a lexeme (the speech paradigm of a word) correlates with the intuitive structuring of lexico-semantic groups in the mind of a native speaker.

In particular, it can be assumed that the core of such a group is characterized by the greatest correspondence to the "dominant" grammatical profile. This hypothesis is the strongest; it needs special proof and verification. In particular, it requires an appeal to the data of linguistic statistics.

As a result, the problem of continuous semantic markup of a very large amount of texts is solved in the language corpus, which can be performed only in automatic mode.

One of the techniques in the fight against polysemy, which generates noise when searching by semantic features, is the optimization of the initial semantic dictionary, namely, the establishment of a hierarchy of values and, if necessary, their renumbering.

The additional criterion of the semantic query "search only by the first meaning of the word" will ensure the issuance of the most probable value. Thus, using the order of word values in the markup is a simple and fairly effective tool for improving the adequacy of the output.

Conclusion

The use of semantic markup is one of the latest trends in corpus linguistics. Semantic markup is one of the newest, most powerful, yet definitely, the least used search engine optimization techniques. Once you understand the concept and techniques of semantic markup, you can greatly improve the position of a language corpus and even your site in search engine results.

Semantic markup is code that you place in a corpus or on your website to help search engines get more informative corpus results for SERP users.

Semantic markup tells search engines what words stand for, i.e. their meaning (semantics). With semantic markup, corpus content is indexed and returned in search results in a different way.

Usually, when creating a semantic classification for a dictionary or thesaurus, the lexicon is divided into topics called semantic classes, and, if necessary, subgroups are created in each of the classes.

These subgroups are also tagged, and the tagging system can be applied to a dictionary or thesaurus. If the token does not match any existing semantic subgroup in the class, you can either create a new subgroup and a new semantic tag for it or use the semantic tag of the class to which the token belongs.

The problem is to create an axiomatic basis for such a classification, i.e., a minimum set of semantic features through which other semantic features can be

defined. The fact is that no one can think of such a classification of the language that will be universal and will not be devoid of some semantic classes or subgroups.

Indeed, text corpora allow researchers to access a large amount of illustrative material and statistical information. For the researcher within the framework of corpus linguistics, there are great opportunities for studying authentic material.

Currently, the Corpus has implemented a search system for lexical-semantic manifestations, based on partial semantic markup of texts. With this definition, most of the words in the text upon detection of one or more semantic and word-forming features, for example, 'person', 'substance', 'space', 'speed', 'movement', 'possession', 'property of a person', 'diminutive ', 'verbal name' etc. Facet classification is used, in which one word can fall into several classes. At the first stage, the search was carried out in terms of the appearance of features in the dictionary. Text markup is carried out automatically using the *Qazkorpus.kz* program (authored by A. Baitursynov Institute of Linguistics) in accordance with the semantic dictionary of the Corpus. Since manual processing of semantically marked texts is very laborious, semantic homonymy in the Corpus is not removed: several alternative sets of semantic features are assigned to polysemantic words.

There are three groups of tags assigned to words to reflect lexical and semantic information:

1. Class (a name, a reflexive pronoun, etc.)
2. Lexical and semantic features (a lexeme's thematic class, indications of causality or assessment, etc.)
3. Derivational features (a diminutive, an adjectival adverb, etc.)

The set of semantic and lexical parameters is different for different parts of speech. Moreover, nouns are divided into three subclasses (concrete nouns, abstract nouns, and proper names), each with its own hierarchy of tags.

Lexical and semantic tags are grouped as follows:

1. Taxonomy (a lexeme's thematic class) – for nouns, verbs, adjectives and adverbs.
2. Mereology (“part – whole” and “element – aggregate” relationships) – for concrete and abstract nouns
3. Topology – for concrete names
4. Causation – for verbs
5. Auxiliary status – for verbs
6. Evaluation – for abstract and concrete nouns, adjectives and adverbs

The semantic markup is based on the classification system of Kazakh vocabulary, adopted in the database "15 volume explanatory dictionary of the Kazakh language", which has been developed since 2012. Since then the dictionary was essentially expanded, several new semantic classes and the derivational parameters were added for the needs of the Corpus.

REFERENCES

[1] Kustova G.I. Mental predicates in metatext constructions of the 2nd person. Based on the materials of the international conference "Dialogue-2018": « Computational Linguistics and Intelligent Technologies». Issue, 17 (24), M., 2018, pp. 380 - 390. [In Rus.]

[2] Sirazitdinov Z.A. Corpora of the Bashkir language: development principles // Oriental Studies, 2016, 23.[In Rus.]

[3] Corpus Chinese. (LIVAC Synchronous Corpus). Access mode: http://www.rcl. cityu.edu.hk/livac (accessed 01/17/2022), pp. 11– 26.

[4] The Spoken Turkish Corpus (a multimedia corpus of spoken Turkish). Access mode: http://stc.org.tr (date of access: 06/17/2013), pp. 6 - 39.

[5] Kretov A. A. Theoretical problems of lexico-semantic typology (on the example of Caucasian languages). Vestnik VSU, series: linguistics and intercultural communication, No. 1 (2020), pp. 6-13. [In Rus.]

[6] Kustova G.I. Vocabulary Markup in the National Corpus of the Russian Language: Principles, Problems, Prospects . National Corpus of the Russian Language: 2003–2005. Results and prospects. M.: Indrik, 2005. Pp. 155–174. [In Rus.]

[7] Gindin S. Semantic dictionaries – map of the language world. Russian language, No. 5, 2001. Pp. 12-19[In Rus.]

[8] Russian semantic dictionary. IV. Explanatory dictionary, systematized by classes of words and meanings. Rep. Ed. N.Yu. Shvedova. M .: Azbukovnik, 2007. 924 p. [In Rus.]

[9] Kretov A.A. Corpus of Parallel Texts: Architecture and Possibilities of Use . National Corpus of the Russian Language: 2003—2005. Moscow: Indrik, 2005. [In Rus.]

[10] Text Encoding Initiative [Electronic resource]. Access mode: http://www. tei-c. org .Access date: 03/15/2014. P. 12.

[11] Corpus Encoding Standard for XML [Electronic resource]. Access mode: http://xml. cover pages. org/xces. Html. Date of access: 03/15/2014.

[12] Sharandin A. L. Methodology of lexical grammar . Interaction of vocabulary and grammar in Russian: problems, results, and prospects. Tambov, 2009, pp. 13-28. [In Rus.]

[13] Norman B. Yu., Mukhin M. Yu. On grammatical profiles of lexico-grammatical groups: search for internal links between lexical and grammatical semantics. SLAVIA: Casopis pro slovanskou filologii, 2015. Roc. 84, ses. 3. Pp. 348-359.

[14] Janda L. A., Lyashevskaya O. Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian . Cognitive Linguistics, 2011, Vol. 22, No 4. Pp. 719-763. [In Rus.]

[15] Kukanova V.V. Principles of semantic markup of the National Corpus of the Kalmyk language //Oriental Studies, 2014. 137 p. [In Rus.]

СЕМАНТИКАЛЫҚ БЕЛГІЛЕУ ҰЛТТЫҚ ТІЛ КОРПУСЫНЫҢ ҚҰРАМДАС БӨЛІКТЕРІНІҢ БІРІ РЕТІНДЕ

*Әмірбекова А.Б.[1], Талғатқызы Г.[2], Уракова Л.[3]

*[1]Филология ғылымдарының кандидаты, А.Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан

[2]Докторант, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан

[3]PhD, Акдениз университеті, Анталия, Түркия,

*[1]e-mail: Margan01@mail.ru

[2] e-mail: gulalyzhan@mail.ru

[3]e-mail: urakovayanc@gmail.com

Аңдатпа. Мақалада қазақ тілінің Ұлттық корпусындағы семантикалық таңбалау принциптері сипатталған. Мақаланың мақсаты – тілдік корпуста қолдануға дайын семантикалық тегтер жүйесін қарастыру және дамыту. Бұл тәсіл сөздіктің семантикалық жіктелуіне негізделген және әмбебап және кез келген тілге жарамды. Сөздіктер мен мәтіндер корпусын белгілеудің практикалық мәні іздеу сапасын жақсарту және

пайдаланушылардың мүмкіндіктерін кеңейту болып табылады. Мақаланың ғылыми маңыздылығы белгілеу және семантикалық классификация кез келген бағдарламалау парадигмасына бағытталуы керек екендігімен анықталады. Біз функционалдық парадигманы таңдадық. Мақаланың негізгі нәтижелері, біріншіден, ұлттық корпустың семантикалық таңбалануы, бұл іздеу сапасын айтарлықтай жақсартады және лингвистикалық ақпаратты сұрау кезінде пайдаланушының мүмкіндіктерін кеңейтеді; екіншіден, жазба жасалған әрбір лексема туралы семантикалық ақпарат семантикалық белгілер немесе тегтер жиынтығы ретінде беріледі және әдетте тіл сөздігінің семантикалық классификациясында көрініс табады. Лексико-грамматикалық семантиканың заманауи зерттеулері үшін корпус деректерін пайдаланудың одан әрі мүмкіндіктері туралы қорытындылар жасалады. **Мақала Қазақстан Республикасы Білім және ғылым министрлігінің қолдауымен «Мемлекеттік тілдің ақпараттық-инновациялық базасы ретіндегі қазақ тілінің ұлттық корпусын әзірлеу; ғылыми-зерттеу және оқыту интернет-ресурс» тақырыбындағы № BR 11765619 ғылыми жобасының аясында жасалды.**

Тірек сөздер: семантикалық тегтер, семантикалық классификация, белгілеу, лексика, топша, субкорпус, ұлттық корпус.

СЕМАНТИЧЕСКАЯ РАЗМЕТКА КАК ОДИН ИЗ КОМПОНЕНТОВ НАЦИОНАЛЬНОГО ЯЗЫКОВОГО КОРПУСА

*Амирбекова А.Б.[1], Талгаткызы Г.[2], Уракова Л.[3]
*[1]Кандидат филологических наук, Институт языкознания имени А.Байтурсынулы, Алматы, Казахстан
[2]Докторант КазНУ им. Аль-Фараби, Алматы, Казахстан
[3] PhD, Университет «Акдениз», Анталия ,Турция
*[1]e-mail: Marghan01@mail.ru
[2] e-mail: gulalyzhan@mail.ru
[3]e-mail: urakovayanc@gmail.com

Аннотация. В статье описаны принципы семантической разметки в Национальном корпусе казахского языка. Цель статьи – рассмотреть и разработать систему семантических тегов, готовую к использованию в языковом корпусе. Этот подход основан на семантической классификации лексики и является универсальным и применимым к любому языку. Практическое значение разметки словарей и корпусов текстов заключается в повышении качества поиска и расширении возможностей пользователей. Научная значимость статьи определяется тем, что разметка и семантическая классификация должны быть ориентированы на любую парадигму программирования. Мы выбрали функциональную парадигму. Основными результатами статьи являются, во-первых, семантическая маркировка национальных корпусов, значительно повышающая качество поиска и расширяющая возможности пользователя при запросе лингвистической информации; во-вторых, семантическая информация о каждой лексеме, в которой сделана запись; она представлена в виде набора семантических разметок или тегов и обычно отражается в семантической классификации словаря языка. Делаются выводы о дальнейших возможностях использования корпусных данных для современных исследований лексико-грамматической семантики. **Публикация выполнена в рамках научного проекта №BR 11765619 на тему «Разработка национального корпуса казахского языка как информационно-инновационной базы госязыка: научно-исследовательский и обучающий интернет-ресурс»,** поддержанного Министерством образования и науки Республики Казахстан.

Ключевые слова: семантические теги, семантическая классификация, разметка, лексика, подгруппа, подкорпус, национальный корпус.

Статья поступила 15.02.2022