Solid State Technology
Volume: 63 Issue: 4
Publication Year: 2020

Prospects For The Creation Of The Uzbek Language Corpus

Sayfullaeva Rano Raufovna

Professor, Doctor of Philology of

Tashkent State Uzbek Language and

Literature University

Shirinova Raima Hakimovna

Doctor of Philology, Professor

Gaybullaeva Nafisa Izzatullaevna

Doctor of Philosophy in Philology(PhD)

Bukhara State University

mexrishka82@mail.ru

Abstract - The lexical system and structure of Russian linguistics have been sufficiently studied, and the services of Uzbek scholars in the comprehensive study of the lexicon of the Uzbek language are invaluable. With their fundamental research, they laid a vast foundation for the systematic study of the Uzbek lexicon; they also started this method in the first stage; Uzbek language laid the foundation for systemic lexicology, and the research of a group of scholars is directly devoted to the study of lexicon from the point of view of systems linguistics.

Key words: The lexical system, linguistics, systematic study, systemic lexicology, lexical semantic field (LSF), lexical semantic groups (LSG).

I.Introduction

Separate interdisciplinary research based on the theory of lexical semantic field (LSF) began, mainly in the 70-80s of the twentieth century. This stage is characterized by the study of lexicon as a whole system consisting of certain semantic groups (LSG), the relationships of certain semantic elements.

The LSF itself is a relatively wider area covering several LSGs. The synonymous series of a word is based on the general base of the semaphore in the coverage of LSG or LSF, as well as the system of meaning(s) of each lexical unit in the group. Sometimes even words that do not belong to a synonymous series have generalizing semaphores that can be combined into a single LSG, a synonymous series within the LSG; and a few LSGs make up the LSF. There are several types of relationships in LSG: commonality, contradiction, intersection. Even LSGs themselves can be micro- and macroLSGs that are part of each other (e.g., movement> forward movement> human movement). In any case, words to LSG are combined on the basis of a paradigmatic attitude (opposition).

II.Literature review

Language is a system of elements that form a whole. The meaning of each element that makes it up stems from the presence of other elements that make up the system at the same time. Various manifestations of the system-structural approach to language can be found in the linguistic concepts put forward by such prominent scholars of world linguistics as V.Gumboldt, A.Shleyxer, G.Shteyntal, V.Matezius, V.Skalichka, N.S.Trubetskoy, L.Elmslev, X.Uldal, L.Blumfield, E.Sepir, B.Uorf, Harris, A. Martins, L. Weisgerber, Baudouin de Courtenay, F. Fortunatov.

An even larger grouping of words is the thematic group (TG); this group is a collection of words from different word groups based on subject commonality. This set is based on a variety of: paradigmatic and syntagmatic relationships. For example, “sports” TG (football, scoring, stadium, fan) or “trade” TG (trading, bargaining, market, shop, seller, buyer, sale, sell). TG includes various LSGs. For example, a commercial building LSFi (shop, kiosk, boutique, supermarket), synonyms (to own, buy), antonyms (expensive - cheap), hyponyms (store - gastronomy) are included in the "trade" TG.

III.Analysis

By language owners, these units can be used mixed in speech. In the semantic tagging of language corpora, the boundaries of TG, LSG, and LSF must be clear, and this hierarchy must be correctly located in the search boxes. Otherwise, the internal division of groups will confuse the user. The clarity of the semantic search, the intelligibility of the search windows is achieved by the correct separation of the semantic field and the lexical-semantic groups. When tapping corpus units, we clearly define the boundaries of TG, LSG, and LSF. In deriving TG and LSF, we take logical, semantic symbols as a basis: it is appropriate to follow the principle that “TG represents the linguistic landscape of the world, its fragments, while LSF represents aspects of meaning and relationships”. Two approaches to the study of vocabulary: semantic (from word to concept) and onomasiological (from concept to word) define the main directions of word work in the language corpus, word learning, linguistic operations. A thematic group, lexical-semantic group, semantic field-based approach to the word allows the corpus to move from concept to word.

The principle of "from concept to word" has been followed in most of the corporations in the world with existing semantic annotations (markings, tags). In our view, the semantic tagging of corpus material is extremely important in two ways: from word to concept, from concept to word. In the corpus, units are semantically searched LSG> TG> LSF hierarchically. This is the "concept-to-word" principle of research. The first window of the search displays the LSFs; smaller LSFs are placed as they enter each LSF; each of which consists of TG and TGs of LSGs. The semantic tagging of the case material and the placement of the LSFs in the search box are done in several steps.

Lexical units exist in language by forming different paradigms based on different semantic relationships. Synonymic, antonymic, graduonymic, partonymic, hyponymic relations are such types of linguistic relations. The reasons why language units are grouped together according to the laws of dialectics, form separate lexical-semantic groups, and differ from other members within the group on the basis of certain characteristics have been extensively studied. In the sources, the types of relations between members of the paradigm in the semantic field are divided into groups such as synonymous, hyper-hyponymic, holo-meronymic, graduonymic, functionalimic, contradictory relationship. Based on the approach to lexicon based on the theory of semantic field, it is possible to go from word to concept through the lexical-semantic relationship between words, while in the corpus the word is understood. That is, such an approach results in a different view of semantic character-based search in the language corpus.

It is known that the language corps is a necessary tool for linguodidactical, linguistic and other social spheres. Corpus research has already proven that the corpus is a large database in word learning. The linguistic corpus differs from the explanatory dictionary, thesaurus, and hyperlug in that the whole meaning of the word is interpreted side-by-side with words that enter into a lexical-semantic relationship. In a glossary, only the meaning of the word is explained; although the thesaurus and hyperlug are aimed at revealing a relatively larger aspect of the word, the simplicity of the search differs from that of the corpus by its limited capacity. The large-volume language corpus is, in a sense, distinguished from the above-mentioned sources of information by the ability to interpret a word, to express it in conjunction with the language.

Adherence to the principle of "word to concept" in the semantic tagging of the corpus unit leads to a comprehensive expansion of word interpretation in the corpus. This allows the corpus user to use a semantic character search in the semantic character search window.

It should be noted as a shortcoming that the majority of language corporations in the global network do not interpret semantic search as broadly as an annotated dictionary. The fact that the corpus refers to the context in which the word is used does not indicate all of the semantic sign that the word can be linked to another unit. The user must find complete, coherent information about the word they are looking for in one place. If in the national corpus the meaning of all words in the language is explained, in the author's corpus the words belonging to the same author's lexicon, their meaning arising in the lexicon of the same author are given.

Another aspect of semantic tagging is the principle of "from word to concept" - the interpretation of word meanings, lexemes that enter into a lexical-semantic relationship: synonyms,

antonyms, hyponym / hyperonym, graduonymy, holo / meronymy. The system of semantic tags includes word, lemma, dictionary meanings, existing synonyms, graduonyms, hyponym / hyperonyms, holo / meronyms, antonyms. It is appropriate to include in the semantic tag such symbols as category, meaning, synonym, antonym, slot, hierarchy, hierarchy, origin, homonym, style, paronymy, variant, period of use.

It can be said that in the semantic tagging of language units for the corpus, the conclusion obtained so far as a result of the study of lexical-semantic relations in linguistics, theoretical material, dictionaries serve as a linguistic supply base. Electronicization of the text of existing dictionaries in the Uzbek language is the first stage of linking the dictionary with the corpus. Systematicity in the lexicon is not as obvious as at other levels of language. Inventory of vocabulary in terms of quality and quantity, the possibility of extensive research has expanded. The main source for semantic tagging of language units is the explanatory dictionary of the Uzbek language.

IV.Discussion

The dictionary does not mention any of the meanings of the headline, but its meanings, which are now in common use and understandable to many. While the keyword has one meaning, the dictionary article explains that meaning; if plural, each meaning is marked separately, in a certain order, with a dark black Arabic numeral. In the word-to-concept approach, it is important that the meanings of words are determined in the order in the dictionary. However, in the search from the concept to the principle, the meaning of the polysemous word should be attached to the corresponding LSG, TG and LSF, because the polysemous word belongs to one LSG with one meaning and belongs to another LSG / LSF with the other meaning.

The number of illustrative examples in the dictionary article will be limited; in contrast to the corpus: a searched word has the ability to refer to the context of thousands of examples. Linking an existing dictionary of antonyms and synonyms to the database of the corpus, and setting up a hyperlink from these dictionaries when searching, gives good results. As a dictionary reflecting the hyper-hyponymic, holo-meronymic relationship has not yet been created in the Uzbek language, there is no lexicographic material for semantic tagging of the word on the basis of these parameters. Semantic tagging is done independently, manually, based on research in this area.

A semantic tag is a character, a set of comments that specifies a meaning, indicating that a word or phrase in a language corpus belongs to a particular semantic category or smaller semantic group (LSG, semantic field, and gang). The semantic tags of the corpus include the specification of the meaning(s) of the word, the formation of a set of explanations related to the homonymy, synonymy of the word, categorization of the word, its thematic group, LSG, definition of semantic field, derivational characteristics, noun meaning. Observations of the principles of semantic tagging of existing national corpus showed that in the corpus three types of characters (category, lexical-semantic characteristic, derivational description) are attached to a word form. Lexical-semantic tags (1) taxonomy (LSG belonging to a lexeme) - a corresponding tag for noun, adjective, verb, form word groups; (2) mereology (a sign referring to "whole-piece", a sign belonging to "element-group /

group") - a tag belonging to the subject and non-subject units; (3) topology (topological position of the object being represented) - a tag belonging to the object names; (4) causation - a tag belonging to verbs; (5) Evaluation - subject and non-subject units are grouped according to their respective units of quality and form.

Semantic tagging uses letter, number, or number-only codes, even if there is no single standard form. The first letter or number represents the general semantic meaning, the next character represents a small semantic group that further specifies the meaning of the word. A semantic tag is not only a word, but also combines many compounds into a semantic group, in which case compounds that express the same meaning in different combinations are encoded by a single character. Semantic tagging is a work tool that expands the ability to perform linguistic operations through the corpus: the presence of a semantically annotated unit in the corpus makes it possible to respond not only to a particular word but also to a user-requested construct search.

The semantic mark is a continuation and extension of the morphological mark; it includes a system of three groups of characters: (1) a unit formed by word-forming indicators; (2) a sign belonging to a lexical-semantic feature: a noun, a noun, an original and a relative adjective; own and portable meaning; (3) private-semantic character: thematic (taxonomic) group, reflecting the semantic field. The semantic tagging of the case is divided into two main types based on the combination of programmer and linguist competencies:

1) terminological marking - the stage of naming a concept in the text;

2) provide a tag indicating the inter-unit relationship.

This involves determining the place of each unit in the semantic field, its relationship with another unit. If the first generation of the language corpus was a collection of electronic texts, then a tool with a query-responsive interface was later formed into linguistic, extralinguistic tagged corpora. Linguistically labeled corporations were initially only morphological, then morpho-syntactic, and in recent years have undergone a stage of development, such as having a perfect view of the linguistic mark - morphological, syntactic and semantic mark. As a result of the study of the NKRYa system of semantic tags, it was found that each tag consists of 3 to 7 comments, these tags represent the unit in a whole-piece relationship, semantic field, lexical-semantic group, and gang; we have observed that the center has its own name, function, depending on its location in the circle. A.A. Kretov divides tags into constants, operator-classifiers. For example, Sp t: constr building and structure (house, attic, bridge). In this system of tags, Sp is a constant and t: constr is a classifier. The first comment is the main, and the comment after the semicolon is the meaning in the context. Marking is done in both word order and alphabetical order.

The first approach is used to illuminate the lexical-grammatical feature of a unit, to indicate to which category it belongs. If the lexical unit has a single meaning regardless of the context, a constant sign is placed, and if the meaning changes depending on the context, a variable sign is placed.

The semantic marking system has a number of special features in addition to the features of the general linguistic marking system. A similar aspect to other linguistic markup systems is that the markup is based on linguistic support or database. A distinctive feature of semantic markup is the presence of two types of annotation: facet and pedigree method.

In practice, one or both of these two methods are used in combination, depending on the nature of the unit. While the facet method requires sequential interpretation of the unit, the tree method requires the identification of the group, gang, area to which the unit belongs. In the first method it goes from word to meaning, in the second method it goes from meaning to word. This indicates that the word -> meaning, meaning -> word principle applies, both in the expert's marking process and in the search that is displayed on the interface. In corpus linguistics, in the process of semantic tagging of an ambiguous unit, general principles should be developed, free from controversial views.

Not all semantics of a word appear as a result of a corpus search engine because the search result is based on a sequence of characters. If there is no semantic comment database in the language corpus, the result appears only on the basis of the form. The result of the semantic tagging should be such that the semaphores appear in sequence. This can be achieved only by adding a dictionary of the Uzbek language to the database of linguistic support of the language corps. The absence of such a dictionary in the Uzbek language increases the number of dictionaries in Uzbek lexicography, which should not be delayed. After all, the creation of the Uzbek language corpus is directly proportional to the development of Uzbek lexicography.

The language corpus reflects the speech event of language units: how the word is used in the language is observed. Search result example, context; but the semaphores are not displayed. In our view, in the semantic tagging of a polysemous word, it is necessary to distinguish between its general and terminological meaning. As a result of the search, the general meaning of the word should appear in the interface, then the terminological meaning. Formulas and models, which are generalized on the basis of the dictionary of aggregation, are useful.

V.Conclusion

1. The approach from word to concept and from concept to word in the lexicon defines the main directions of linguistic action on the word in the language corpus, and in the corpus the lemma is systematized on the basis of LSG, LSF. However, when covering a word in a synonymous series, LSG or LSF, it is based on the common base of the semaphore, the system of meaning(s) of each lexical unit in the group, where the whole lexical structure of the language is combined into semantic fields. Linguistic support is also required for the semantic tagging of noun units, as the set of semantic tags consists of a tag denoting affiliation to a word, lemma, microLSG, macroLSG, a tag denoting affiliation to TG, LSF.

2. An explanatory dictionary differs from a thesaurus in that the whole corpus of language is interpreted side by side with words that enter into a lexical-semantic relationship, in which only the

meaning of the word is interpreted in the explanatory dictionary; the thesaurus differs from the corpus in that the simplicity of the search is limited, while the aim is to uncover a relatively larger aspect of the word. Corporations refer to the context in which a word is used, not all of the semantic characters as the word demonstrates the ability to relate to another unit, since the user must find complete, coherent information about the word they are looking for in one place.

3. There is no need to refer to another word in the language body (as in the dictionary), because artificial intelligence has a large memory, paper is not wasted, there is always the ability to expand the volume as desired, the size is not difficult for the user and millions of words in just a few megabytes located. If the number of illustrative examples in a dictionary article is limited, it will be possible to refer to the context of thousands of examples of a single word searched in the corpus.

4. A semantic mark, a set of comments denoting a word / combination in the language corpus, indicating that it belongs to a certain semantic category, group, LSG, semantic field definition, derivation characteristic, noun meaning. Lexical-semantic tags are grouped by taxonomy, mereology, topology, causation, evaluation fields, as the semantic markup system consists of category, lexical-semantic characteristics and derivational description.

5. The semantic marking system has special features in addition to the features of the general linguistic marking system. While, like other linguistic marking systems, marking is based on linguistic supply or base, the specific aspect of semantic marking is that there are two types of annotation: facet and pedigree method, in practice one or both of these two methods are used in combination, depending on the nature of the unit. While the facet method requires sequential interpretation of the unit, the tree method requires the identification of the group, gang, area to which the unit belongs.

6. Units that reflect the lexicon of a particular language as a necessary tool for the semantic layout of the body: 1) dictionary, 2) a semantic dictionary that can fully interpret the lexicon of the language, 3) a linguistic module for the implementation of semantic marking - a set of rules, 4) semantic marking system, 5) additional software tool: a filter is distinguished that can distinguish between ambiguity and homonymy. Morphological, lexical homonymy in the process of semantic tagging; universal vocabulary that is part of a compound word (compound term); a word that does not exist in the dictionary; fragment; it is necessary to develop specific principles of marking the literal-symbolic construction, as these units will have a separate character in each language.

7. Based on the semantic tag of the corpus units' semantic field, gang, belonging to the group, the group is identified by the operator tag belonging to the gang, the constant tag belonging to the field. Indeed, defining a field boundary is important in semantic tagging. The fact that a polysemous word can only be an element of one / more fields requires solving certain problems in tapping a word / lemma, because the domain to which a polysemous word belongs is determined only on the basis of a semantic filter. Indeed, formulas showing the coherence of words in the language corpus will be limited.

8. A language product that is a cyber product differs from a paper dictionary in that it has the ability to display all the meanings of a lexical unit, but paper dictionaries are the main raw material

for the language body. As a result of semantic tagging it is necessary to achieve a sequence of semantics, which can be achieved by adding a dictionary of the Uzbek language to the database of the linguistic support of the language corpus. The absence of such a dictionary in the Uzbek language adds to the number of types of dictionaries that need to be created in Uzbek lexicography.

9. In the corpus linguistics of the world there are methods of tapping homonymous units, the elimination of homonymy in the process of automatic reading of the text - based on 1) grammatical norms and 2) statistics. Homonymy should be defined separately for literary speech and colloquial speech, because the dictionary mainly describes lexemes belonging to the modern Uzbek language, and includes a limited number of words specific to colloquial speech. However, homonyms used in literary speech and colloquial speech in the language corpus are used in the same way that homonymous forms not reflected in the dictionary can be observed in the language corpus. The most necessary tool for distinguishing homonymy in the text is a morpho-semantic filter. Creating a special program that differentiates homonyms - creating a morpho-semantic filter - regulates the search for homonyms in the language corpus.

10. Fixed tags can be thought of as the name of a semantic field: secondary tags are operators / classifiers that indicate belonging to LMGs and serve to define the meaning of a word; the constant tag represents belonging to the lexical-semantic field. Experiments show that separated LSFs, LMGs can be operator / classifier tags belonging to constants and constants, although it is natural that in subsequent studies, the sequence of these tags will be enriched, and the sequence of constants and operators / classifiers will be improved.

References:

1. Ўзбек тили лексикологияси (Муалифлар жамоаси). Тошкент: Фан, 1981.
2. Jamolxonov H. Hozirgi o'zbek adabiy tili. – Toshkent: "O`zbekiston milliy ensiklopediyasi" Davlat ilmiy nashriyoti, 2013.
3. Раҳматуллаев Ш. Ўзбек тили омонимларининг изоҳли луғати. Тошкент: Ўқитувчи, 1984. – 215 б.
4. Раҳматуллаев Ш. Ўзбек тили омонимларининг изоҳли луғати. Тошкент: Ўқитувчи, 1984. – 215 б.
5. Пинхасов Я.Д. Ҳозирги ўзбек адабий тили лексикология ва фразеология. Тошкент: Ўкитувчи, 1969.
6. Ўзбек тилининг изоҳли луғати. 5 томлик. Тошкент: "Ўзбекистон миллий энциклопедияси" Давлат илмий нашриёти, 2006. II том.
7. http://tech.yandex.ru/mystem – MyStem морфологик таҳлил дастури сайти
8. Рысаков С.В. Методы борьбы с омонимией. http://samag.ru/archive/article/3059.
9. Абжалова М.А. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (расмий ва илмий услубдаги матнлар таҳрири дастури учун). Филол. фан. бўйича фалсафа докт. дисс... – Фарғона, 2019. – 164 б.

10. Кобрицов Б.П. Модели многозначности русской предметной лексики: глобальные и локальные правила разрешения омонимии. Автореф… канд. филол. наук. Москва: РГГУ, 2004.;
11. Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов (Электрон ресурс). http://www.dialog-21.ru/media/2444/zelenkov_segalovich.pdf;
12. Кобрицов Б.П. Методы снятия семантической неоднозначности. НТИ, Сер.2, Вып. 3, 2004.;
13. Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка
14. (Электрон ресурс). http://elar.urfu.ru/bitstream/10995/1388/1/IMAT_2005_03.pdf.; Hearst M.A. Noun homograph disambiguation using local context in large text corpora // Proceedings of the 7th conference, UW Centre for the New OED & Text Research Using Corpora, 1991.;
15. Yarowsky D. Word sense disambiguation using statistical models of Roget's categories trained on large corpora // Proceedings of the 14th International Conference on Computational Linguistics, COLING'92, 23- 28 August, Nantes, France, 1992. – P. 454-460.
16. Ўзбек тилининг изоҳли луғати. Икки томлик. М., 1981.
17. Ўзбек тилининг изоҳли луғати. 5 томлик. Тошкент: “Ўзбекистон миллий энциклопедияси” Давлат илмий нашриёти, 2000-2006.
18. Менглиев Б. Лисоний тизим яхлитлиги ва унда сатҳлараро муносабатлар. Филология фанлари доктори диссертацияси. – Бухоро, 2001.