DOI:10.26104/IVK.2022.45.557#

ИЗВЕСТИЯ ВУЗОВ КЫРГЫЗСТАНА, № 6, 2022

*Касиева А.А., Кадырбекова А.К.*

**КЫРГЫЗ ТИЛИНДЕГИ ЭТИШТЕРДИН КОШ МААНИЛҮҮЛҮГҮН ЖОЮУ**
**(жаңы түзүлгөн кыргыз тилинин корпусунун негизинде)**

*Касиева А.А., Кадырбекова А.К.*

**СНЯТИЕ НЕОДНОЗНАЧНОСТИ СМЫСЛА ГЛАГОЛОВ В КЫРГЫЗСКОМ ЯЗЫКЕ**
**(на основе новосозданного кыргызского корпуса)**

*A. Kasieva, A. Kadyrbekova*

**VERB SENSE DISAMBIGUATION IN THE KYRGYZ LANGUAGE**
**(on the basis of Newly-Created Kyrgyz Corpus)**

*УДК: 004.89/81.37:811.1*

*Сүйлөмдө колдонулган сөздүн контексттик мааниси тилдеги өндүрүлгөн продукциянын тактыгын жана маанилүүлүгүн аныктайт. Сөздөрдүн кош маанилүүлүгү маселеси ошол сөздөрдүн маанисин чечмелөө же жоюу аркылуу чечилет. Бул макала кыргыз корпусундагы этиштердин кош маанилүүлүгүн VSD-Verb Sense Disambiguation процесси аркылуу жоюу маселесин аркалайт жана аны автоматташтыруу мүмкүнчүлүктөрүн изилдейт. Күндөн-күнгө токтоосуз өсүп жаткан маалымат агымын жетиштүү түрдө иштетүү үчүн чийки маалыматтарды өркүндөтүлгөн ыкмалардан өткөрүп, аларды чыпкалоочу мыкты каражаттардын колдонулушун талап кылат. Демек ушул багыттан алып карасак, түшүнүксүз болгон кош маанилүү сөздөрдүн анык маанисин илгеп чыгарып, алардын кош маанилүүлүгүн жоюу Word Sense Disambiguation (WSD) процесси деп аталат. Бул процесс табигый тилди иштетүүдө маанилүү кадамдардын бири жана ал семантика тармагына караштуу болгону менен бирге, морфология жана синтаксиске да негизделет. Бул эмгекте, табигый тилди иштетүүдөгү (ТТИ) жалпы сөздөрдүн ичинен этиштин кош маанилүүлүгүн жоюу (VSD-Verb Sense Disambiguation) процессинин жүргүзүлүшүн көрсөтүп берүүгө далалат кылабыз. Буларды аткаруу үчүн бизде бир гана шарт болушу керек. Ал – тилдик корпус жана андагы сөздөрдүн сөз түркүмү боюнча энтектелиши. Демек, жогоруда айтылган маселелерди чечүүдө колдонулган ыкмаларга илимий баа берүү үчүн жаңы түзүлгөн кыргыз корпусундагы маалымат колдонулду.*

***Негизги сөздөр:*** *кыргыз тили, корпустук лингвистика, сөздөр, этиштер, көп маанилүүлүк, синтактикалык парсинг, морфологиялык энтектөө, табигый тил, тилди иштетүү.*

*Целью разотождествления смысла слова (WSD) является правильная идентификация значения слова в контексте. Во всех естественных языках присутствуют неоднозначные значения слов, которые часто трудно разрешить автоматически. Поэтому WSD считается важной проблемой в обработке естественного языка (NLP). В данной статье рассматриваются вопросы корпусно-ориентированного исследования наиболее частотных типов неоднозначности глаголов (VSD-Verb Sense Disambiguation) в кыргызском языке и возможности автоматизации процесса дисамбигуации в корпусе. Чтобы не отставать от растущего потока информации, необходимо использовать прогрессивную фильтрацию и передовые методы обработки исходных данных. В результате одним из таких важнейших этапов является устранение вхождений слов с неясными и неоднозначными значениями - также известный как процесс разграничения смысла слов (WSD). В данной работе мы предлагаем подходы к WSD, которые в нашем случае ограничены глаголами (VSD-Verb Sense Disambiguation) в кыргызском языке, который выступает в качестве одного из примеров для теоретической базы системы NLP. Единственным предварительным условием в этом отношении является наличие корпуса с частеречной разметкой. Соответственно, для оценки вышеупомянутой проблемы и ее методов был использован новосозданный корпус кыргызского языка.*

***Ключевые слова:*** *кыргызский язык, корпусная лингвистика, слова, глаголы, многозначность, синтаксический парсинг, морфологическая разметка, естественный язык, языковая обработка.*

*This article considers the issues of corpus-oriented study of the most frequent types of ambiguity of verbs (VSD – Verb Sense Disambiguation) in the Kyrgyz language and the possibilities for automation of the disambiguation process in the corpus. Progressive filtering and advanced raw data processing techniques must be used to keep up with the growing information flow. As a result, eliminating word occurrences with unclear-ambiguous meanings – also known as the Word Sense Disambiguation (WSD) process – is one of these crucial steps. In this work, we offer WSD approaches, that are, in our case, restricted to verbs (VSD – Verb Sense Disambiguation) in the Kyrgyz language, acting as one of examples for the NLP system's theoretical background. The only prerequisite in this regard is the usage of a morphologically annotated corpus. Consequently, the Newly-created Kyrgyz corpus has been used to evaluate the above-mentioned issue and its methods.*

***Key words:*** *the Kyrgyz language, Corpus linguistics, words, verbs, polysemy, syntactic parsing, morphological tagging, natural language, language processing.*

**Introduction.** With the advance of information revolution, telecommunications and information systems deal with a huge, constantly increasing massive volume of raw data. Accordingly, it requires the developed data presentation accompanied by formats that are available and can be used by a variety of users, along with access to data in a very natural way. More than ever, corpus research and modern linguistics (such as internet linguistics, computational linguistics, etc.) are becoming integrated and comprehensive.

With the help of various Natural Language Processing (NLP) programs and linguistic databases, it is now feasible to study languages at all levels. One or more linguistic corpora may be used to research phonetics, morphology, syntax, semantics, and pragmatics of a particular

DOI:10.26104/IVK.2022.45.557#



language, for instance. Similarly, language transcends purely linguistic boundaries, touching on other disciplines such as sociolinguistics, psycholinguistics, neurolinguistics, theoretical/applied linguistics, cognitive linguistics, geographical linguistics, and others. In this respect, language technologies based on NLP techniques are essential in this evolution, making them vital to success of information systems.

NLP systems need a deep understanding of language. A great difficulty in processing a language causes an ambiguity in natural language that occurs at all of its levels: phonological, morphological, syntactic, semantic, and pragmatic. Therefore, resolving ambiguity is one of the key goals while creating any NLP system. As a result, each kind of uncertainty or ambiguity of words necessitates a unique resolution process [1].

In this paper we consider the resolution of a particular type of lexical ambiguity, namely, the different senses a word which might have in a particular context. This specific issue is commonly referred to as Word Sense Disambiguation (WSD). WSD, which comprises Verb Sense Disambiguation (VSD) as its subbranch has been the focus of the present study since the majority of languages contain words that are ambiguous and have more than one meaning. The elimination of ambiguity of these words is a critical step in creating any tool for natural language processing, since their presence would otherwise impair the effectiveness of the systems that have been created.

**A Brief History of Research on Word Sense Disambiguation (WSD).** One of the most challenging tasks in the discipline of natural language processing research is WSD. In this area, research was first conducted [2] in the late 1940s when Zipf first put forth his “Law of Meaning” idea in 1949. According to this theory, the less frequent words and the more frequent words have a power-law connection. Compared to less frequent words, more frequent words have more senses.

Later, the British National Corpus received confirmation of the [2]. relationship. Kaplan discovered in 1950 that two words on each side of an ambiguous word in a context are comparable to the context's entire sentence [3]. Masterman first put forth his theory in 1957, explaining how to use the headers of the categories in Roget's International Thesaurus to determine the true meaning of a word [4]. In order to determine the precise meaning of an ambiguous word, Wilks created a model in 1975 called "preferred semantics," which combined selectional constraints and a frame-based lexical semantics. In 1979, Rieger and Small developed the concept of unique "word experts." Due to the availability of large-scale lexical resources and corpora in the 1980s, WSD research underwent a notable progress.

As a result, researchers began combining various automatic knowledge extraction tools along with manual handcrafting techniques. Later in 1986, Lesk introduced his algorithm based on overlaps between the *glosses (*Dictionary definitions) of the words in a sentence. In this algorithm, the preferred meaning of the ambiguous word is expressed by the maximum number of overlaps [5]. Lesk used the Oxford Advanced Learner’s Dictionary of Current English (OALD) to obtain the dictionary definitions. Later, this approach laid the basis for other Dictionary-based WSD works. In 1991, Guthrie employed the subject codes to disambiguate the exact sense using the Longman Dictionary of Contemporary English (LDOCE). Three significant advancements in the field of NLP research took place in the 1990s: the launch of Senseval [6], the availability of the online lexicon WordNet [7,8], and the introduction of statistical approaches. Because information was both programmatically available and hierarchically arranged into word senses termed synsets, WordNet [9] revolutionized this field of study. WordNet is now an important online sense inventory exploited in WSD research.

The sense classification issues are successfully solved using statistical and machine learning techniques. Modern approaches to WSD use supervised learning techniques that are trained on corpora that have been manually sense-tagged. Brown et al. [10] introduced corpus-based Word Sense Disambiguation for the first time in 1991.

**Verb Sense Disambiguation (VSD).** Word sense disambiguation (WSD), a challenge in natural language processing, is the process of figuring out which "sense" (meaning) of a word is activated by the use of the word in a certain context. WSD is a natural classification problem which categorizes an occurrence of the word in context into one or more of its sense classes given the term and its potential senses, as listed in a dictionary. The characteristics of the context, such as the words nearby, serve as the basis for classification.

Consequently, Verb Sense Disambiguation (VSD) is a sub-problem of the Word Sense Disambiguation (WSD) problem that tries to identify in which sense a polysemic verb is used in a given sentence. In his famous book entitled “Handbook of Natural Language Processing” David Yarowsky proposes the following definition for VSD: *“the process of examining verbs in a particular context and identifying precisely which sense of each verb is most appropriate is known as verb sense disambiguation* (VSD)” [11]. Up to that point, VSD did not receive much attention in the WSD research. Most WSD systems use largely collocation-based features to disambiguate verbs in the same way as nouns.

In this paper, we will investigate the role of VSD and describe its resolution process in Kyrgyz language using the newly-created Kyrgyz corpora [12]. Corpus annotation become even more complicated due to the agglutinative nature of the Kyrgyz language. Turkic Lexicon Apertium [13] [14], an open-source machine translation platform has been selected as the most appropriate toolkit for POS tagging and parsing issues of the corpus [15].

DOI:10.26104/IVK.2022.45.557#



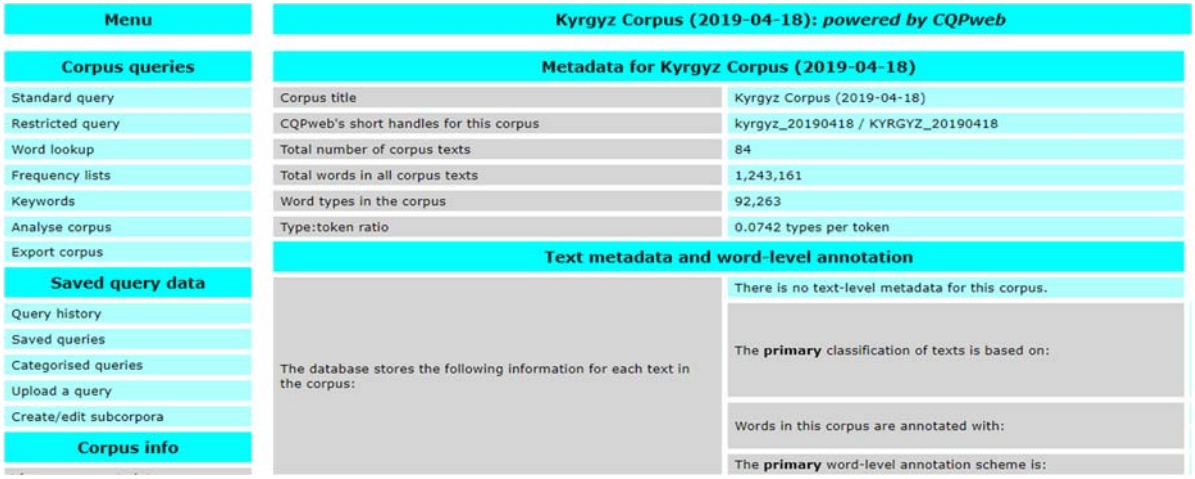


**Figure 1.** The home-page of the Kyrgyz corpus

***Source:*** The Kyrgyz Corpus is available at https://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418/i-ndex.php?thisQ=corpus-Metadata&uT=y

The sense of the context in which a word is used in a sentence determines how accurate the output produced will be. The ambiguous words are resolved using word sense disambiguation. In this paper, an attempt will be made to provide an overview of the ways that might be used to resolve ambiguity, particularly, verb sense disambiguation, and analysis of verbs in the Kyrgyz language that are retrieved from The Kyrgyz Corpus. The Kyrgyz language has a rich agglutinative morphology with word structures formed by productive affixations of derivational and inflectional suffixes to root words. Let us consider the following sentence:

*Капыстан бетме - бет чыга түшкөн өлүм кызыл жүздүү жигитти* ***апкаарытып таштады*** *(Lit.; The sudden death that came out face to face stunned the red-faced young man)***.**

1. The following is the Parts of Speech (POS) tagging (morphological tagging) of the considered sentence.

***Капыстан (Kapystan)***_n_nom ***бетме-бет (betme-bet)***_n_nom ***чыга (chyga)***_v_iv_prc_impf ***түшкөн (tüshkön)*** _v_iv_gpr_past ***өлүм (ölüm)*** _n_nom ***кызыл (kyzyl)***_adj ***жүздүү (jüzdüü)***_n post ***жигитти (jigitti)***_n_acc ***апкаарытып (apkaarytyp)*** _v_iv_p3_sg ***таштады (tashtady)***_v_tv_ifi_p3_sg._sent.

2. And this is the syntactic analysis of the same sentence. In order to study the structure of sentences and their tagging, we used Universal Dependencies (UD). Universal Dependencies (UD) is a platform for consistent annotation of grammar (parts of speech, morphological features and syntactic dependencies) in different languages [16]. We propose to review the syntactic features of example in Kyrgyz that was made on the UD Annotatrix website UD Annotatrix Annotation tool [17].

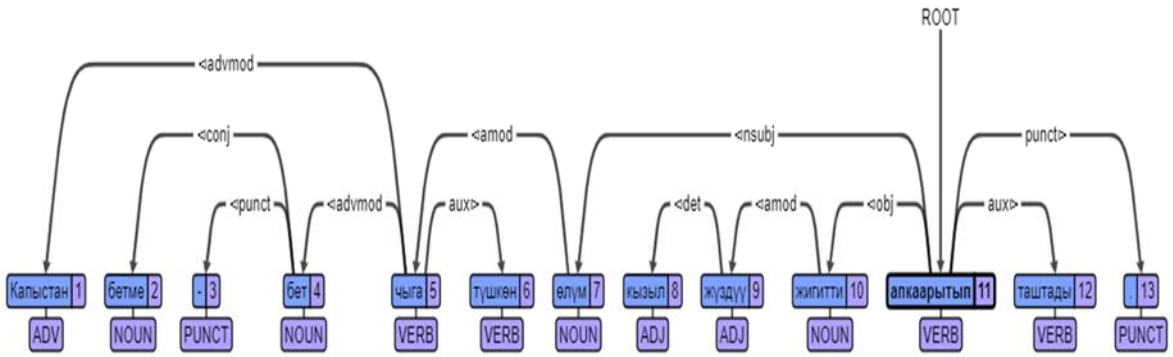


**Figure 2.** Syntactic Analysis via Universal Dependency Annotatrix Platform.

***Source:*** UD Annotatrix annotation tool is available at https://jonorthwash.github.io/udannotatrix/server/public/html/annotatrix.html#1

DOI:10.26104/IVK.2022.45.557#



3. Here is the one more analysis of the considered sentence. This treebank is made in a Javascript app for generating syntax trees called ***Syntax Tree Generator*** [18]. Looking at the analysis, we can observe that the verbs ***"апкаарытып (apkaarytyp)", "маштады (tashtady)"*** belong to verb group in the parts of speech and they always come together (there is the attachment between them) falling under the VP (Verb Phrase) branch. Consequently, we can claim that they are compound verb.

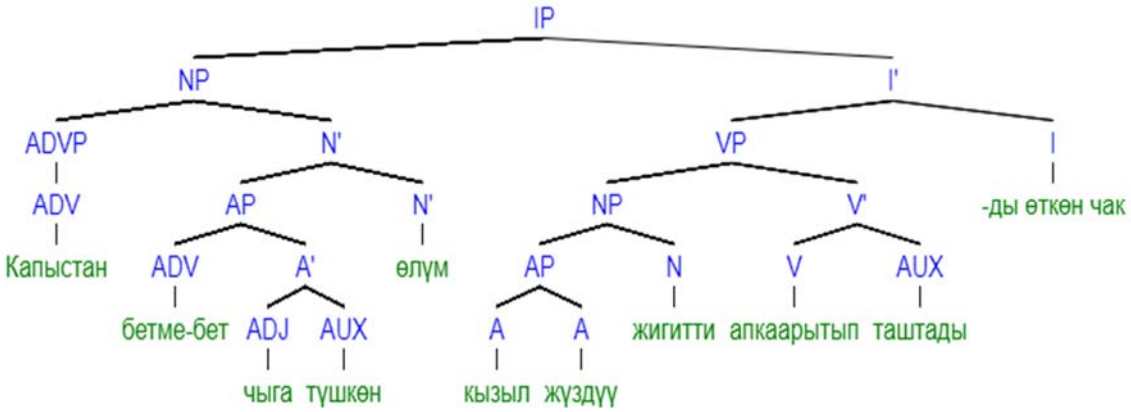


**Figure 3.** Syntactic Analysis via Syntax Tree Generator.

From these three analyses, we can infer that verbs ***"апкаарытып (apkaarytyp)", "маштады (tashtady)"*** are considered to be *compound verbs with verbal pairs (чакчыл түгөйлүү татаал этиштер).* Compound verbs which constitute from verbs only are called ***compound verbs with verbal pairs.*** The first couplet of such verbs is always in the present tense and indicates the main action as in our case the verb ***"апкаарытып"***. And the second pair ***"маштады"*** adds to main verb grammatical meaning showing the past tense and first person singular and becomes an auxiliary verb.

**Conclusion.** We have made an effort to cover the main areas of work and sketch the broad outlines of advancement in the field, even though we are aware that much more might be added to what is shown here, taking into account that it is a relatively new field in the study of the Kyrgyz language. Of course, one of the reasons why WSD is challenging is that it is inherently difficult to determine or even define the word sense, and this problem is likely to be solved any time soon. Even yet, it is evident that current WSD research would profit from taking a deeper look at lexical semantics and theories of meaning. The main goal of this paper is to provide a substantial basis to number of researchers working in various branches of computational linguistics, NLP and AI, who is interested in learning more about WSD. As WSD contributes to numerous researches, as we have listed above, an interest to it has grown recently. Although WSD is "an intermediate problem," it is challenging and possibly hard to evaluate. By incorporating WSD methods into more extensive applications, we can potentially improve future work.

***Acknowledgements.*** *I would like to thank Professor Elke Teich, the head of English Linguistics and Translation Studies at the Department of Linguistics and Language Technology of Saarland University (Germany) for creating and hosting the Kyrgyz Corpus. I would like to express my deep gratitude to Master of Sience Jörg Knappen for providing the technical aspect for creating and enhancing the Kyrgyz Corpus.*

*We are also grateful to Associate Professor Doctor of Philology Gulnura Dzhumalieva from the Kyrgyz-Turkish Manas University, Department of Simultaneous Translation for her support. Moreover, we would like to thank bachelor, graduate, and postgraduate students of Simultaneous Translation Department of the Kyrgyz-Turkish Manas University for annotating and validating the corpus data.*

**References:**

1. E. Agirre and G. Rigau, "Word Sense Disambiguation using Conceptual Density," in Proceedings of the 16th International Conference on COLING, Copenhagen, 1996.
2. E. Agirre and P. Edmonds, "Word Sense Disambiguation," Algorithms and application, vol. 33, p. 33.
3. A. Kaplan, "An experimental study of ambiguity and context," Mechanical translation, vol.2, no.2, pp. 39-46, November 1955.
4. M. Masterman, "The Thesaurus in Syntax and Semantics," Mechanical Translation, vol. 4, no. 1-2, pp. 35-43, 1957.
5. P. Alot Ranjan and S. Diganta, "Word Sense Disambiguation: a survey," International Journal of Control Theory and Computer Modeling (IJCTCM), vol. 5, no. 3, pp. 2-3, 2015, July.
6. R. Mihalcea, "SENSEVAL," University of North Texas, 19 October 2008. [Online]. Available: http://web.eecs. umich.edu /~mihalcea/senseval/.
7. H. Seo, H. Chung, H. Rim, S. Myaeng and S. Kim, "Unsupervised Word Sense Disambiguation using WordNet relatives," Computer Speech and Lanaguage, vol. 18, no. 3, pp. 253-273, 2004.

DOI:10.26104/IVK.2022.45.557#




8. A. J. Canas, a. Valerio, J. Lalinde-Pulido, M. Carvalho and M. Arguedas, "Using WordNet for Word Sense Disambiguation to Support Concept Map Construction," String Processing and Information Retrieval, vol. 2857, pp. 350-359, 2003.
9. G. Miller, "wordnet: An online lexical database," International Journal of Lexicography, vol. 3, no. 4, 1991.
10. P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R. L. Merger, "Word-Sense Disambiguation Using Statistical Methods," Association for Computational Linguistics, vol. 29th Annual Meeting of the Association for Computational Linguistics, p. 264–270, June 1991.
11. D. Yarowsky, Handbook of Natural Language Processing, vol. Chapter 26, Marcel Decker, 2000.
12. E. Teich, J. Knappen, S. Fischer and A. Kasieva., "The Kyrgyz Corpus," CQP Web, Saarland University, The Department of Language Science and Technology, April 2019. [Online]. Available: https://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418/index.php?thisQ=corpusMetadata&uT=y
13. J. N. Washington, M. Ipasov and F.M. Tyers, "A finite-state morphological transducer for Kyrgyz," LREC, pp. 934-940, 2012.
14. J. Washington and F. Tyers, "Turkic Lexicon," 14 June 2018. [Online]. Available: https://wiki.apertium.org/wiki/Turkic_-lexicon
15. A.A. Kasieva and A.K. Kadyrbekova, "Corpus Annotation Tools: Kyrgyz Language Corpus (Turkic Lexicon Apertium and PENN Treebank Tools), "Общество, язык и культура XXI века, no. 20, pp. 207-214, 2021.
16. "Universal Dependencies," 2014-2021. [Online]. Available: https://universaldependencies.org/.
17. F.M. Tyers, M. Sheyanova and J. N. Washington, "UD Annotatrix: An annotation tool for Universal Dependencies," Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pp. 10-17, 23-24 January 2018.
18. M. Shang, "Syntax Tree Generator," 28 October 2011. [Online]. Available: http://mshang.ca/syntree/.
19. Мамбеталиева А.М. Особенности и различия в выборе учебников для студентов і курса по специальности английского языка в обучении английский языка и литературы. Наука, новые технологии и инновации Кыргызстана. 2008. №. 3-4. - С. 149-150.