**KYRGYZ-TURKISH MANAS UNIVERSITY**

**THE GRADUATE SCHOOL OF SOCIAL SCIENCES**

**SIMULTANEOUS TRANSLATION DEPARTMENT**

# VERB SENSE DISAMBIGUATION(VSD) IN THE KYRGYZ CORPUS AND THE PROBLEMS OF THEIR MORPHOLOGICAL TAGGING

**MASTER'S THESIS**

**Aizat KADYRBEKOVA**

**Thesis Advisor: Assoc. Prof. Aida Kasieva**

**BISHKEK 2023**

## DECLARATION OF AUTHORSHIP

The research under the thesis paper entitled ***"Verb Sense Disambiguation (VSD) in the Kyrgyz Corpus and the Problems of Morphological Tagging"*** was undertaken according to the criteria for the Master's degree at the Kyrgyz-Turkish "Manas" University. All sources cited in this thesis paper have been used ethically and in compliance with academic norms. This thesis paper's data came from a variety of sources, all of which were used in accordance with ethical and scientific standards. All references and citations for sources are included. As a result, I thus affirm that this thesis paper is entirely my own work and that I am fully aware that any rules violations would result in disciplinary action in accordance with University legislation.

**Aizat Kadyrbekova**

## ЭТИКАЛЫК ТАЛАПТАР

**"Кыргыз корпусундагы этиштердин кош маанилүүлүгүн жоюу(VSD) жана аларды энтектөө маселелери"** аттуу диссертациялык иш боюнча илимий-изилдөө иши Кыргыз- Түрк "Манас" университетинин Коомдук илимдер институтунун "Магистрдик диссертация жазуу эрежеленине" ылайык жүргүзүлдү. Диссертацияда колдонулган маалыматтар, эмгектер жана документтер академиялык жана этикалык эрежелерди эске алуу менен пайдаланылды. Ар кандай булактардан алынган маалыматтардын булактары, адабияттары илимий жана моралдык эрежелерге ылайык жазылып берилди. Диссертацияда колдонулган бардык адабияттарга жана булактарга шилтеме жасалды. Эгерде эрежелердин бузулгандыгы аныкталса, мага жана бул илимий ишке каршы козголо турчу укуктук жоопкерчиликти тартууга даярмын.

**Айзат Кадырбекова**

## ПРОТОКОЛ

Кадырбекова Айзат даярдаган “Кыргыз тилинин корпусундагы этиштердин кош маанилүүлүгүн жоюу(VSD) жана аларды морфологиялык энтектөө маселелери” аттуу диссертациясы төмөндө көрсөтүлгөн жюри тарабынан бир добуштан / көпчүлүк добуш менен Кыргыз-Түрк “Манас” университети, Коомдук илимдер институтунун Лингвистика (Котормо жана котормо таануу) багытында магистрдик диссертациясы кабыл алынды.

Диссертацияны коргогон күнү: .....

Жюри тарабынан ***бир добуштан / көпчүлүк добуш*** менен кабыл алынган диссертация Институттун башкаруу кеңешинин ..... номердүү жана ..... күнкү чечими бекитилди.

.....Коомдук илимдер Институнун мүдүрү

## TEZ ONAYI

Ayzat KADIRBEKOVA tarafınan hazırlanan ***"Kırgız Derleminde Fiil Anlamının Belirsizliğini Giderme (VSD) ve onların Morfolojik Etiketleme Sorunları"*** adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ / OY ÇOKLUĞU ile Kırgızistan-Türkşye "Manas" Üniversitesi, Sosyal Bilimler Enstitüsü Mütercim-Tercümanlık Anabilim Dalında yüksek lisans tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: ……..

Jüri tarafından ***oy birliğiyle/ oy çokluğuyla*** kabul edilen bu tez Enstitü Yönetim Kurulu.....

Sayılı ve ..... tarihi kararıyla onaylanmıştır.

.....

Sosyal Bilimler Enstitüsü Müdürü

## ABSTRACT

| | |
|---|---|
| **Researcher** | : Ayzat Kadyrbekova |
| **University** | : Kyrgyz-Turkish Manas University |
| **Department** | : Simultaneous Translation |
| **Type of study** | : Master thesis |
| **Number of pages** | :XX+142 |
| **Date of graduation** | |
| **Research supervisor(s)** | : Assoc. Prof. Aida Kasieva |

**VERB SENSE DISAMBIGUATION(VSD) IN THE KYRGYZ CORPUS AND THE PROBLEMS OF THEIR MORHPLOGOCAL TAGGING**

This thesis paper 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.

This study attempts to provide the first survey for verb sense disambiguation (VSD) of verbs in the Kyrgyz language, applying newly created Kyrgyz Corpus (2019-04-18) (named *the Kyrgyz Corpus)* powered by Corpus Query Processor (CQP) on the basis of the University of Saarland in Germany. Withdrawal of Kyrgyz verbs' morphological tagging are carried out applying CQP and syntactic analysis are done based on Universal Dependency (UD) platforms.

As a result, we believe that the materials of this paper, will advance the representation of the Kyrgyz Corpus, and contribute to the establishment of corpus linguistics as a new branch of linguistics and facilitate the distribution of it within Kyrgyzstan. In addition, we hope that it will also benefit to Kyrgyz Natural Language Processing field.

The outcomes of the research can be applied in teaching, studying and conducting linguistic researches in Corpus and Computer Linguistics, Semantics along with syntax

and morphology, Machine Translation. Moreover, it will benefit greatly for teachers who conduct syntax, grammar and morphology classes.

**Key words:** The Kyrgyz Corpus, Corpus linguistics, Natural Language Processing, Word Sense Disambiguation (WSD), Verb Sense Disambiguation, Kyrgyz verbs, Universal Dependency, Syntactic Parsing, POS tagging.

## КЫСКАЧА МАЗМУНУ


| | |
|---|---|
| **Даярдаган** | : Айзат Кадырбекова |
| **Университет** | : Кыргыз-Түрк «Манас» университети |
| **Багыты** | : Лингвистика (котормо жана котормо таануу) |
| **Иштин сыпаты** | : Магистрдик диссертация |
| **Беттердин саны** | : XX+142 |
| **Бүтүрүү датасы** | : |
| **Илимий жетекчи** | : ф. и. к., Аида Касиева |


**Кыргыз корпусундагы этиштердин кош маанилүүлүгүн жоюу (VSD) жана аларды морфологиялык энтектөө маселелери**


Бул магистрдик иш кыргыз тилиндеги этиштердин көп кездешүүчү кош маанилүүлүгүн жана анын түрлөрүн (VSD-Verb Sense Disambiguation) корпуска багытталган изилдөө маселелерин жана корпустагы кош маанилүүлүктү жоюу процессин автоматташтыруу мүмкүнчүлүктөрүн карайт. Прогрессивдүү чыпкалоо жана чийки маалыматтарды иштетүүнүн өркүндөтүлгөн ыкмалары токтоосуз өсүп жаткан маалымат агымын кармап туруу үчүн колдонулушу керек. Натыйжада, табигый тилди иштетүү тармагында түшүнүксүз жана кош маанилери бар сөздөрдүн санын азайтуу/жоюу (Word Sense Disambiguation (WSD) процесси деп аталат) — бул маанилүү кадамдардын бири. Бул илимий иште биз сөздөрдүн кош маанилүүлүгүн жоюу (WSD) ыкмаларын сунуштайбыз. Биздин эмгегибиз кыргыз тилиндеги этиштерге жана алардын кош маанилүүлүгүн жоюуга (VSD-Verb Sense Disambiguation) багытталып, табигый тилди иштетүүнүн (ТТИ) теориялык негиздеринин бири катары каралып, изилденет. Бул жагынан алып карасак, бирден-бир шарт - морфологиялык аннотацияланган корпусту колдонуу. Демек, жогоруда айтылган маселеге жана анын ыкмаларына баа берүү үчүн жаңы түзүлгөн кыргыз корпусу колдонулду.

Бул изилдөө Германиядагы Саарланд университетинин Corpus Query Processor (CQP) платформасы тарабынан колдоого алынган жаңы түзүлгөн кыргыз корпусундагы (2019-04-18) (Кыргыз корпусу деп аталган) этиштердин маанисин чечмелөө/кош маанилүүлүгүн жоюу (VSD) боюнча алгачкы илимий эмгек болуп саналат. Кыргыз тилиндеги этиштердин морфологиялык энтектелиши CQP платформасын колдонуу менен ишке ашырылды жана ал эми универсалдуу багыныңкылык (UD) платформасынын негизинде синтаксистик талдоо жүргүзүлдү.

Жыйынтыктап айтканда, бул эмгектеги материалдар кыргыз корпусунун мындан ары өнүгүшүнө жана корпустук тил илими Кыргызстандын илимий

чөйрөсүндө тил илиминин жаңы тармагы катары түптөлүшүнө салым кошуп, Кыргызстандын ичинде жайылышына шарт түзөт деп ишенебиз. Мындан тышкары, кыргыз тилин компьютерде иштетүү тармагына да пайда алып келет деген үмүттөбүз.

Изилдөөнүн натыйжалары корпус жана компьютердик лингвистика, семантика менен бирге синтаксис жана морфология, машиналык котормо тармактарын окутууда, изилдөөдө жана ушул эле тармактарды лингвистикалык изилдөөлөрдү жүргүзүүдө колдонулушу мүмкүн. Айрыкча, синтаксис, грамматика жана морфология сабактарын окутуп жаткан мугалимдерге чоң пайда алып келет.

**Түйүндүү сөздөр:** кыргыз тилинин корпусу, корпустук тил илими, табигый тилди иштетүү, сөздөрдүн кош маанилүүлүгүн (WSD), этиштердин кош маанилүүлүгүн жоюу, кыргыз тилиндеги этиштер, универсалдуу багыныңкылык, синтаксистик парсинг, сөз түркүмдүк энтектөө (аннотациялоо).

# ÖZ


| | |
|---|---|
| **Hazırlayan** | : Ayzat Kadırbekova |
| **Üniversite** | : Kırgızistan-Türkiye Manas Üniversitesi |
| **Anabilim Dalı** | : Mütercim-Tercümanlık |
| **Tezin Niteliği** | : Yüksek Lisans |
| **Sayfa Sayısı** | : XX+142 |
| **Mezuniyet Tarihi** | : |
| **Tez Danışmanı** | : Doç. Dr. Aida Kasieva |


## KIRGIZ DERLEMİNDE FİİL ANLAMININ BELİRSİZLİĞI GIDERME (VSD) VE ONLARIN MORFOLOJİK ETİKETLEME SORUNLARI


Bu tez çalışması, Kırgız dilinde en sık karşılaşılan fiil belirsizliği türlerinin (VSD-Verb Sense Disambiguation) derlem odaklı incelenmesi ve derlemdeki belirsizlik giderme sürecinin otomasyonu için olasılıklar konularını ele almaktadır. Artan bilgi akışına ayak uydurmak için aşamalı filtreleme ve gelişmiş ham veri işleme teknikleri kullanılmalıdır. Sonuç olarak, Kelime Anlamının Belirsizliğini Giderme (WSD) süreci olarak da bilinen, anlamları belirsiz olan kelime oluşumlarını ortadan kaldırmak bu önemli adımlardan biridir. Bu çalışmada, NLP sisteminin teorik arka planı için örneklerden biri olarak Kırgız dilindeki fiillerle (VSD-Verb Sense Disambiguation) sınırlı olan WSD yaklaşımlarını sunuyoruz. Bu konudaki tek ön koşul, morfolojik olarak açıklanmış bir derlemin kullanılmasıdır. Sonuç olarak, yeni oluşturulan Kırgızca derlem yukarıda bahsedilen konuyu ve yöntemlerini değerlendirmek için kullanılmıştır.

Bu çalışma, Almanya'daki Saarland Üniversitesi temelinde Corpus Query Processor (CQP) tarafından desteklenen yeni oluşturulan Kırgız Derlemini (2019-04-18) (Kırgız Derlemi olarak adlandırılmıştır) uygulayarak Kırgız dilindeki Fiillerin Anlamının Belirsizliğini Gidermek (VSD) için ilk araştırmayı sağlamaya çalışmaktadır. Kırgızca fiillerin morfolojik etiketlemesi CQP uygulanarak gerçekleştirilmekte ve sözdizimsel analiz Evrensel Bağımlılık (UD) platformlarına dayalı olarak yapılmaktadır.

Sonuç olarak, bu makalenin materyallerinin Kırgız Derleminin temsilini geliştireceğine ve derlem dilbiliminin yeni bir dilbilim dalı olarak kurulmasına katkıda bulunacağına ve Kırgızistan'da dağıtımını kolaylaştıracağına inanıyoruz. Ayrıca Kırgızca Doğal Dil İşleme alanına da fayda sağlamasını temenni ediyoruz.

Araştırmanın sonuçları, Derlem ve Bilgisayar Dilbilimi, sözdizimi ve morfoloji ile birlikte anlambilim, Makine Çevirisi alanlarında öğretim, eğitim ve dilbilimsel araştırmalarda uygulanabilir. Ayrıca, sözdizimi, dilbilgisi ve morfoloji derslerini yürüten öğretmenler için de büyük fayda sağlayacaktır.

**Anahtar kelimeler:** Kırgızca Derlemi, Derlem Dilbilimi, Doğal Dil İşleme, Kelime Anlamının Belirsizliğini Giderme (WSD), Fiil Anlamınının Belirsizliğini Giderme (VSD), Kırgızcadaki fiiller, Evrensel Bağımlılık, Sözdizimsel Ayrıştırma, POS etiketleme.

## АБСТРАКТ

| | |
|---|---|
| **Исследователь** | : Айзат Кадырбекова |
| **Университет** | : Кыргызско-Турецкий университет «Манас» |
| **Направление** | : Лингвистика (перевод и переводоведение) |
| **Вид исследования** | : Магистерская диссертация |
| **Количество страниц** | : XX+142 |
| **Дата защиты** | : |
| **Научный руководитель** | : Кан. фил. наук, Аида Касиева |

**Снятие неоднозначности глаголов (VSD) в Кыргызском корпусе и проблемы их морфологической разметки**

В данной магистрской диссертации рассматриваются вопросы корпусно-ориентированного исследования наиболее частотных типов неоднозначности глаголов (VSD-Verb Sense Disambiguation) в кыргызском языке и возможности автоматизации процесса дизамбигуации в корпусе. Чтобы не отставать от растущего потока информации, необходимо использовать прогрессивную фильтрацию и передовые методы обработки исходных данных. В результате, устранение вхождений слов с неясными и однозначными значениями - также известное как процесс разграничения смысла слов (WSD) - является одним из этих важнейших шагов. В данной работе мы предлагаем подходы к WSD, которые в нашем случае ограничены глаголами (VSD-Verb Sense Disambiguation) в кыргызском языке, выступая в качестве одного из примеров для теоретической базы системы NLP. Единственным условием в этом отношении является использование морфологически аннотированного корпуса. Следовательно, для оценки вышеупомянутой проблемы и ее методов был использован новосозданный корпус кыргызского языка.

В данном исследовании предпринята попытка провести первое исследование по деамбигуации смысла глаголов (ДСГ) в кыргызском языке, используя недавно созданный кыргызский корпус (2019-04-18) (названный Кыргызским корпусом) на основе Corpus Query Processor (CQP) на базе Саарского университета в Германии. Вывод морфологических тегов кыргызских глаголов осуществляется с помощью CQP, а синтаксический анализ проводится на основе платформ Universal Dependency (UD).

В результате, мы считаем, что материалы данной работы, продвинут представление корпуса кыргызского языка, а также внесут вклад в становление корпусной лингвистики как новой отрасли языкознания и будут способствовать ее

распространению в Кыргызстане. Кроме того, мы надеемся, что это работа также принесет пользу области обработки кыргызского естественного языка.

Результаты исследования могут быть использованы в преподавании, изучении и проведении лингвистических исследований в области корпусной и компьютерной лингвистики, семантики, а также синтаксиса и морфологии, машинного перевода. Кроме того, они принесут большую пользу преподавателям, ведущим занятия по синтаксису, грамматике и морфологии.

**Ключевые слова:** Корпус кыргызского языка, корпусная лингвистика, обработка естественного языка, снятие неоднозначности смысла слов (WSD), снятие неоднозначности смысла глаголов (VSD), глаголы в кыргызском языке, универсальная зависимость (УЗ), синтаксический разбор, POS-теги.

## PREFACE

With the advance of information revolution, information technology (IT) 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 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 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 Natural Language Processing (NLP) techniques are essential in this evolution, making them vital in success of information systems.

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 in terms of Kyrgyz Corpus and Universal Dependency platforms. This specific issue is commonly referred to as ***Word Sense Disambiguation (WSD)***. Word Sense Disambiguation (WSD) which comprises itself a subbranch ***Verb Sense Disambiguation (VSD)*** which has been our major study focus, 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 (NLP), since their presence would otherwise impair the effectiveness of the systems that have been created. Thus, this work is devoted to the thorough study of verbs in the Kyrgyz Language and their grammatical peculiarities along with their morphological and syntactic annotation taking into account verb peculiarities of the Kyrgyz language. To accomplish these set goals, the Newly-created Kyrgyz Corpus (for morphological analysis) web and Universal Dependency (for syntactic dependency treebanks) platform has been successfully implemented.

First and foremost, I would like to express my deepest gratitude to my research supervisor **Assoc. Prof. Aida Kasieva** for her excellent guidance, tolerance and valuable advice throughout the whole process. This paper could not have been completed without my supervisor's assistance.

Additionally, I would like to thank professor Elke Teich, the director of English linguistics and Translation Studies at the University of the Saarlandes (Germany). I would like to express my deep gratitude to Jörg Knappen, a computer linguist and researcher from the Universität des Saarlandes in Germany, for handling the technical parts of creating the Kyrgyz Corpus and for continuing to work on it up to now.

While conducting the research on this topic, a great support and valuable advice was provided by *Associate Professor Gulnura Jumalieva* from Kyrgyz-Turkish Manas University, who gave hand in solving the arising issues on different stages of the writing

process of this thesis paper. Also, I would like to thank Seide Musazhanova, for her assistance in Unisersal Dependency platform.

Being given an opportunity, I would like to offer my sincere gratitude to academic staff for providing its students with an excellent education and for their devotedness in teaching. And I would like to express my appreciation to Kyrgyz-Turkish Manas University for providing its students with an excellent education, possibilities, and conditions, which in turn greatly influences the students' personalities and educational backgrounds.

Last but not least, my special thanks go to my family, friends, for their moral encouragement and motivation.

**Бишкек 2023** **Айзат Кадырбекова**

## TABLE OF CONTENT

**CHAPTER 3**

**RESOLUTION OF AMBIGUITY IN NATURAL LANGUAGE PROCESSING (NLP). VERB SENSE DISAMBIGUATION**



**CHAPTER 4**

**VERB SENSE DISAMBIGUATION IN THE KYRGYZ LANGUAGE (ON THE BASIS OF THE NEWLY-CREATED KYRGYZ CORPUS)**

## TABLE OF ABBREVIATIONS

| | |
|---|---|
| CL | Corpus Linguistics |
| WSD | Word Sense Disambiguation |
| VSD | Verb Sense Disambiguation |
| NLP | Natural Language Processing |
| AI | Artificial Intelligence |
| DL | Deep Learning |
| ML | Machine Learning |
| MT | Machine Translation |
| CQP | Corpus Query Processor |
| POS | Parts of Speech |
| UD | Universal Dependency |

## LIST OF TABLES

## LIST OF FIGURES

# INTRODUCTION

Communication and information systems now deal with a tremendous volume of raw data that is constantly growing as the information revolution continues. As a result, it calls for access to data in a highly natural way, combined with established data presentation supported by formats that are readily available to and usable by a variety of users. Corpus research and contemporary linguistics, including online linguistics and computational linguistics, are more interconnected and thorough than ever. It is now possible to study languages at all levels thanks to a variety of NLP tools and linguistic resources. For example, the phonetics, morphology, syntax, semantics, and pragmatics of a single language may be studied using one or more linguistic corpora. Similar to how other fields such as sociolinguistics, psycholinguistics, neurolinguistics, theoretical/applied linguistics, cognitive linguistics, geographical linguistics, and others are impacted by language, language itself transcends strictly linguistic limits. This evolution makes language technologies built on Natural Language Processing (NLP) techniques crucial to the 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.

In this thesis 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). Word Sense Disambiguation (WSD) which comprises itself a subbranch (Verb Sense Disambiguation) has been a study focus, 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.

## Motivation for research

Natural language processing (NLP) is used in a wide range of modern applications, including automatic speech translation, automatic summarization, search engines that use semantic/topic search rather than word matching. But ambiguity is a problem that all of these applications have to cope with. The ability of a given expression to be understood in multiple ways is referred to as ambiguity. Ambiguity appears during many phases of the processing of a text or a sentence in NLP. Word forms that fall under the purview of this assignment may be unclear since, for instance, punctuation may signify more than just the conclusion of a sentence.

**Topicality of the research** is that this study attempts to use Corpus-based approach for conducting morphological analysis of verbs in the Kyrgyz language. Nowadays there are no restrictions on the amount of materials for making analysis due to the fact that corpora comprise millions of words, or even more, in our case the Kyrgyz corpus consists of more than 2 million words It is considered as an empirical science and it can provide valuable data for doing research. To make language learning, teaching and linguistic study more effective and quicker, linguists created various types of linguistic corpora which include naturally-occurring collections of written and spoken materials. Corpora can show results for syntactic and semantic tagging of words, phrases, sentences, even grammar, and word frequency and density with just one click in seconds. Thus, these opportunities offered by corpus linguistics motivated us to choose this method as a means of carrying out analysis.

In this paper we introduce the UD Annotatrix annotation tool for manual annotation of languages in Universal Dependencies. To study syntactic peculiarities and to show dependency of verbs in the Kyrgyz language we implemented Universal Dependency which is a platform used to consistently annotate the grammar of various human languages, including the parts of speech, morphological characteristics, and syntactic dependencies.

**The objectives of the research** are to provide theoretical background on Corpus Linguistics, Natural Language Processing along with Artificial Intelligence as long as they are considered to be the foundations of WSD, to investigate theoretical background of WSD and its history. We have also considered Kyrgyz Language and Verbs in Kyrgyz language as our main objective. To demonstrate and evaluate our research of

morphological and syntactic tagging in the sense resolution process of Kyrgyz verbs we applied Corpus-oriented approach and UD Annotatrix annotation tool. The following tasks had to be accomplished in order to achieve these set goals:

1) To elicit Kyrgyz verbs from the Kyrgyz Corpus Query Processor (CQP);
2) To adapt Kyrgyz verbs into classification that Abduvaliev has given in his book:
    - Morphological and syntactic analysis of simple verbs retrieved from Kyrgyz Corpus query processor;
    - Morphological and syntactic analysis of compound verbs taken from the Kyrgyz Corpus query processor;
    - Morphological and syntactic analysis of main verbs selected out from the Kyrgyz Corpus query processor;
    - Morphological and syntactic analysis of auxiliary verbs found in the Kyrgyz Corpus query processor;
    - Morphological and syntactic analysis of types of compound verbs retrieved from the Kyrgyz Corpus query processor;
    - Morphological and syntactic analysis of classification of notional verbs found in the Kyrgyz Corpus query processor;
    - Morphological and syntactic analysis of action verbs found in the Kyrgyz Corpus query processor;
    - Morphological and syntactic analysis of state verbs were chosen from the Kyrgyz Corpus query processor;
    - Morphological and syntactic analysis of modifying verbs and verbs of sense found in the Kyrgyz Corpus query processor;
3) To single out achieved result and evaluate, make an analysis of them according to each type of Kyrgyz verbs.

**The subject of the research** is Verb Sense Disambiguation in the Kyrgyz language and their morphological and syntactic tagging.

**The object of the research** is UD Annotatrix annotation tool and Kyrgyz Corpus query processor (CQP).

**The scientific novelty of the research:** To the best knowledge of the researcher there has not been written any research on Verb Sense Disambiguation in Kyrgyz linguistics especially in terms of Corpus analysis and Universal Dependencies. Thus, this study can be considered as the first work that is devoted to the study of Verb Sense Disambiguation theory in general and to the Kyrgyz verbs in particular. The study suggests a new theoretical approach according to which Kyrgyz verbs' ambiguity can be resolved, researched and analyzed. Kyrgyz verbs' examples are extracted from the Kyrgyz Corpus, are thoroughly examined and analyzed. And the verbs that are have been retrieved from the Kyrgyz corpus are manually annotated in UD Annotatrix annotation tool and results are downloaded, examined and analyzed. This approach can also be considered as another part of the novelty of the present dissertation work. Consequently, the attempt to apply corpus-based/corpus-driven approach, and UD tools to the study of the Kyrgyz verbs has been taken.

**Research methodology:**

- Corpus-based approach is implemented to retrieve Kyrgyz verbs along with their morphological tagging.
- Corpus-driven approach is used to analyze and show the frequency list, concordances of verbs in the Kyrgyz language;
- Quantitative method is used to demonstrate statistical data regarding frequency of verbs, concordances for certain type of verbs and words as well;
- Qualitative method is used to describe, explain and compare ambiguous meanings of verbs in the Kyrgyz language;
- Comparative method is used to compare, to translate and reveal similarities in verbs from the Kyrgyz language into English and their equivalents in English;
- Contrastive method is used in revealing differences while translating verbs' examples from the Kyrgyz language into English;
- Selective method is used to demonstrate Kyrgyz verbs by extracting them from the body of the Kyrgyz literary works that are provided in the Kyrgyz corpus;

- Descriptive method is used to depict theoretical background of Kyrgyz verbs, Kyrgyz language, Corpus Linguistics and Verb Sense Disambiguation in Natural Language Processing and Artificial Intelligence;

**Theoretical and practical importance of the research is** that theoretical background descriptions of the research can be applied in teaching, studying and conducting linguistic researches in Kyrgyz Natural language Processing, the Kyrgyz Corpus, Corpus and Computer Linguistics, Machine Translation. Especially, it will benefit greatly to teachers who conduct syntax and morphology classes. Moreover, we hope that this work will serve as a basis for further semantic studies of verbs, especially, compound verb structures and peculiarities which has not been studied thoroughly yet.

From the scope of practical importance, it is worth noting that CQP (Corpus query processor) is a unique method for language processing, which presupposes managing with a vast volume of computer-available data to conduct linguistic analysis and obtain accurate and exact statistical data regarding frequency of verbs in the Kyrgyz language. The analysis is carried out with the application of the newly created Kyrgyz Corpus and UD Annotatrix annotation tool. Until quite recent time linguists could only examine the limited amount of texts and manually retrieve necessary examples from them for their research and it was really time consuming and laborious. However, today thanks to emergence of Corpus Linguistics, it has become an easy task doing a research using computer readable huge amounts of texts. Thus, the results and outcomes of this work are also considered as helpful in a number of subfields of Linguistics such as: Kyrgyz Natural Language Processing, Kyrgyz Word Sense Disambiguation, Corpus Linguistics, Semantics and Syntax, Morphology and finally Translation studies, etc.

**Materials of the research:** In the process of writing the present dissertation work a large volume of literature sources in the field of Natural language Processing and Artificial Intelligence, Corpus Linguistics, and Word/Verb Sense Disambiguation research. One of the most challenging tasks in the discipline of natural language processing research is WSD. In this area, research was first conducted 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 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. 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. 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).

The structure of the thesis consists of four chapters:

*The first chapter* deals with theoretical background of Corpus linguistics, their classification and divisions, which are based on special categorization. The compilation and development of Kyrgyz corpus were discussed thoroughly. This chapter also introduces with a newly-created Kyrgyz Corpus CQP (2019-04-18).

*The second chapter* comprises theoretical and practical knowledge regarding Artificial Intelligence and Natural language processing as a basis of Word Sense Disambiguation.

*The third chapter* focuses on the ambiguity and its types, classification and what ambiguity is not. In this chapter, we have defined and investigated the Word/ Verb Sense Disambiguation Process itself.

*The last fourth* chapter deals with the practical resolution process of Verb Sense Disambiguation in two ways: Kyrgyz Corpus (morphological analysis) and Universal Dependency (syntactic analysis) methods.

# CHAPTER 1

# A THEORETICAL OVERVIEW OF CORPUS LINGUISTICS AND THE KYRGYZ CORPUS

*"You shall know a word by the company it keeps".*

*John Rupert Firth, Studies in Linguistic Analysis, 1957*

## 1.1 History of Creation of Linguistic Corpora

*Pre-electronic or Digital Age.* The history of creation of corpora dates back to the 1940s when the pioneer work in digital world of study, *The Index Tomisticus*, was produced. It was an electronic concordance of over 10.6 million words, created by friar Roberto Busa about the writings of Thomas Aquinas, was the first project where elements of machine processing of texts were applied (Busa, 1980). This Concordance was created over the course of 34 years. For the convenience of working with the concordance were used punch cards which are stiff pieces of paper that contain digital data expressed by the presence or absence of holes at prescribed positions. And Busa decided to present only the lemma, or headword, as the key word in the concordance with only the lemma, or headword, with all its word forms. Consequently, he carried out lemmatization of texts, which took place in two stages: combining all word forms with inflections under one lemma and attaching a code with an appropriate part of speech for each lemma and its word form. The lemmatization was carried out using the dictionary *The Lexicon Electronicum Latinum*, which was compiled by Busa and ten priests over a period of two years. The electronic dictionary was a table of lemma data, on the basis of which the computer carried out the lemmatization of the texts. This method of working with an electronic dictionary or list later determined much of the principle of electronic text processing. In 1973, the first volume of *the Index Tomisticus* was published, in the 1970s more than 40 volumes of *the Index Tomisticus* were published along with alphabetical indexes, word frequency tables, etc.

The last corpus of before the start of electronic era was the mixed corpus of spoken and written language by Querke, *The Survey of English Usage (SEU),* developed at the University of Durham in England (Quirk, 1972). Svartvik argues that in 1960 the term *"corpus*" was hardly ever used, and at the conference scholars argued at length about the

plural form of the word *"corpus". (corpuses, corpora, or even corpi*) (2007, p. 15). This corpus is considered to be the most well-structured and systematic corpus of the time. In the corpus, the oral and written forms of speech were represented by texts of various genres, with both formal and informal communication serving as sources. The corpus consisted of 200 text fragments, each with a volume of 5,000 word uses. This corpus symbolized the transition from the pre-electronic to the electronic era.

Consequently, all necessary prerequisites and preparation were done for the transition to the new electronic or digital age. The first concordances were created, which were thought to be synonymous with dictionaries and indexes. They were of great importance for the further development of corpus linguistics. Later in the composition of the article of the concordance, to indicate the word that searched for, the place of its usage, the context of use in the recorded language units became obligatory. In addition, a system of illustrating the context in the concordance "*key word in context"* was developed. At that time there was not any unified principles or rules for creating the corpora and compiling concordances. The scope and sources of such corpora and concordances also varied greatly: a corpus could be texts of sacred books (translations of the Bible, works of theologians) as well as individual works of fiction. From the modern point of view, these kinds of texts are not corpora, but archives or collections of individual texts. Moreover, the term *"corpus"* itself was also absent.

*The Electronic or Digital Age (1960s to present).* S. Johansson argues that, despite the works already published in the 1960s, Busa and the emergence of the first electronic corpus, scholars started to show active interest in the late 1970s (2008, p. 39). According to Johansson, real corpus linguistics was born in the 1970s, with the creation of the first laboratories and centers in which linguists and programmers began to cooperate on general problems of linguistics. Computational linguistics centers aimed at collecting, storing, and processing corpus texts were opened in Italy, the United States, England, Germany, Canada, France, Sweden, and Norway. By the mid-1970s, the first databases for electronic corpus storage and distribution were established: The Oxford Text Archive (OTA, 1976) and the International Computer Archive of Modern English (ICAME, 1977).

*First-generation corpora.* In the early 1960s, electronic enclosures first appeared. The first electronic corpus was the so-called "Brown corpus". The Brown corpus, named after The Brown University, Rhode Island, USA. Its name officially included the term "corpus". A group of scientists led by G. Kučera and N. Francis worked on the corpus from 1961 to 1964 (Francis & Kucera, 1998). R. Quirk, P. O'Connor, and J. Carroll also participated in the creation of this corpus. Ph. B. Gove, the editor of the third edition of Webster's Dictionary, took part in this project as well. The Brown corpus was a corpus of written American English and contained one million word uses from 500 texts published in 1961 alone. The corpus contained the following fifteen genres of written American-English: newspaper articles, scientific works, advertisements, hobby books, religious literature, biography, essays, fiction (detectives, adventures and westerns, popular science fiction, love stories, feuilletons). The texts in Brown Corpus were inscribed on a punch card, which indicated an information about the location of the text, its title, and the number of lines in the text.

Later, in 1968 Philip Bagley first coined the term *"metadata"* and used in his book named *"Extension of Programming Concepts"* to refer to all textual data in a corpus (Hoang, 2014, p. 195). From the middle of the 1960s the first concordancer programs appeared based on KWIC: the COCOA (COunt and COncordance Generation Atlas, 1967), and Collocations (CLOC, CoLOCation, 1978). When they were created, machine processing of texts was accompanied by manual markup (tagging), i.e., "attaching" a code (or tag) to each unit of text (word), and a unit with metadata about it (Baker et al, 2006, p. 154). Automatic text markup first started to be used when B. Green and J. Rubin created the automated text tagging program called TAGGIT in 1971. This tagset program was tested on the Brown corpus. TAGGIT marked up the text with 86 tags, highlighting the parts of speech are commonly divided into open classes (nouns, verbs, adjectives, and adverbs) and closed classes (pronouns, prepositions, conjunctions, articles/determiners, and interjections), punctuation marks, and individual morphemes in the text. The only malfunction of the program was that it did not consider homonymy, and 23% of the words in the corpus were marked with several tags simultaneously (McEnery & Hardie, 2012).

In 1978, Ellegard manually marked up the Brown corpus in terms of syntactic parsing. This project was carried out in three stages: clause structures in sentences, constituent structures of clauses, and word classes of individual words. After several years of revisions and corrections, Brown corpus's syntactic parsing was completed in 1979. Green and Rubin published all the data on the TAGGIT morphological analyzer, thus other scientists could refine, rework and make improvements to it (Johansson, 2008, p. 46) . Scholars consider the end of the 1970s to be the time of *official recognition* of the term "*corpus linguistics".*

In the 1980s, the TAGGIT program continued to be refined and improved, and in 1983 a group of scientists at Lancaster University, led by grammarian G. Leach and programmer R. Garside, tested and implemented an updated version of the morphological analyzer called CLAWS (the Constituent Likelihood Automatic Word-tagging System) (McEnery & Hardie, 2012).

The Brownian corpus has become the standard and sample for other corpus compilations, both in its volume and in the range of writing styles and genres represented in it. With the publication of the Brownian corpus in the mid-1970s, similar corpora began to appear, first in Britain and then in other countries. For example, in 1976 The Lancaster-Oslo-Bergen corpus (LOB) (1961-1978) was published as a joint corpus of the universities of Lancaster, Oslo, and Bergen (Leech, Johansson, Garside, & Hofland, 2008). In the early 1990s, similar corpora with a volume of at least one million words, consisting of 500 texts of fifteen different genres of writing, began to be created. Each text that is included had to contain at least 2,000 uses of the word. For example, The Australian Corpus of English, (ACE, 1986), the Wellington Corpus of New Zealand English, The Wellington Written English, (WWE, 1986), the Freiburg and Brown Universities American English Speech Corpus, The Freiburg-Brown Corpus, (FROWN, 1991-1992), The Freiburg London-Oslo (F-LOB, 1991-1992), The Kolhapur corpus of Indian written English, The Kolhapur corpus Indian English (1978). All these corpora were collectively called as the corpora of Brown Family. The only difference between these corpora were that they contained texts of one of the variants of written English:

American, British, Australian, New Zealand, Indian etc. (Baker et al, 2006). All corpora mentioned above contained only collections of written texts.

*Corpora of Spoken Language.* The spoken language (oral speech) corpus came much later than the written corpora that are listed above. The first glances started to be published in the 1990s. The London-Lund (LLC) corpus was constructed between 1975 and 1990 by Svartvick, Quirk, Greenbaum and Hofland based on two projects: the SEU corpus (1959-1989) and the Speaking English Corpus (SEC, 1975). The LLC corpus consists of 100 transcribed texts of spoken monologue and dialogues of 5,000 words each. Recordings dialogic speech is taken from conversations between friends and colleagues, regular conversations, and telephone conversations. Monological speech is represented by spontaneous comments, stories, and narratives as well as prepared speech, not read from the paper (Xiao, 2008, pp. 408-409). In addition to the grammatical tagging, the texts in the corpus are tagged at the prosodic level, i.e., they contain information about tone units, the beginning of sound (onset), the core place (word, syntagma), the direction of nuclear tones (rising, falling, even, rising-down), pitch, pause (short and long), stress (ordinary and dedicated). Texts from the SEU project has detailed prosodic markings: indications of different volume and tempo levels (fast, intermittent, mannerly-stretched, etc.), modifications in voice quality (pitch, rhythm, tension, etc.), additional characteristics (whisper, wheeze, etc.).

In 1984, Texas Instruments compiled a database of spoken English American speech *TI-DIGITS.* This corpus comprises speech that was created and collected at Texas Instruments, Inc. (TI) for the purpose of developing and testing algorithms for speaker-independent recognition of linked digit sequences. 326 speakers participated (111 men, 114 women, 50 boys, and 51 girls) who each say 77digit sequences (Leonard & Doddington, 1982). Each speaker group is divided into two subsets: test and training (Lamel & Cole, 1997).

In 1990, the TIMIT Acoustic-Phonetic Continuous Speech Corpus was created for acoustic-phonetic research, development, and evaluation of automatic speech recognition systems by Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments(TI). The corpus contains texts in eight major dialects of

spoken English American speech by 630 speakers (70% male and 30% female), who read aloud ten sentences each. To test speech recognition systems, the TIMIT corpus includes three types of texts: dialectal (1,260 sentences), phonetically rich (compact), i.e., covering the entire phonemic composition and individual combinations of phonemes that present some recognition difficulty (3,150 sentences), and phonetically diverse tests with repetition of each phoneme in a different context (1,890 sentences). For the third part of the TIMIT corpus, the texts were taken from the Brown corpus, for example, the dialogues of the theatrical productions of the time (Garofolo, et al., 1950).

The Resource Management (RM) corpus (1988) for testing of conjoint speech recognition systems includes more than 25,000 utterances from over 160 respondents speaking various regional dialects of American variant English. The corpus includes two sub-corpora: RM1 and RM2. The RM1 subcorpus consists of three parts. The Speaker dependent training part consists of 12 speakers, each of whom reads aloud 600 "training" sentences in two dialects and ten "rapid adaptation" sentences. The total volume of this subcorpus is 7,344 sentences. The Speaker independent sub-corpus contains 3,360 sentences read aloud by 80 persons in two dialects, and 40 sentences each taken from the main RM corpus. The RM2 subcorpus is an expanded version of the RM1 Speaker-dependent selection. The subset contains 10,508 sentences read out by two men and two women (2,652 sentences each) (Price, Fisher, Bernstein, & Pallett, 1993).

The Air Travel Information Service corpus (ATIS,1990) was designed to study spontaneous speech and speech synthesis. The corpus is also divided into a training part and a test part. ATIS contains texts of people talking to an auto-answering machine like "I would like a ticket to...", "I want to fly to Boston from New York next week". This corpus was later used to create dialogue systems that could answer questions like "Does Air Canada fly from Chicago to California?"( (Tur & Renato, 2011).

These corpora showed and proved that it is possible to train machines for automatic speech recognition and gave a rise to new terms: *tokenization and lemmatization* (dividing conjoint speech into separate words), *segmentation* (dividing conjoint speech into sentences and syntagma), parser (syntactic analyzer), *normalization* (bringing to the

normal phonetic form words spoken with different individual features of the speaker) based on the time alignment of the phrase (time alignment) and etc.

Looking through all types of corpora that are mentioned above, Kennedy came to conclusion that all corpora of particular genres of texts from different historical periods, texts of speech of particular professional communities, age groups or regional dialects are examples of first-generation corpora, because their aim is to study the speech as a particular and distinct form of language, rather than the language as a whole in all its diversity (Kennedy, 1998). Thus, according to his classification, we can argue that the first-generation corpora are special corpora as long as they mainly represent individual genres of oral speech.

In the 1960s and 1990s, corpus requirements and its principle gradually took shape: The involvement of texts of written speech with a total volume of up to one million uses of words became mandatory. However, the texts involved were mostly from the most common genres of written speech and each text fragment (not full text) contained about 2,000 uses of a word. Moreover, corpora of that time did not contain complete texts of written speech, but fragments with a fixed volume of words. The 1970s and 1980s, were the years of prosperity of corpus linguistics: centers and laboratories for processing texts into electronic(machine-readable) form emerged. Electronic processing of texts has brought to scientists the problem of the accuracy of electronic word processing, and tagging. After a number of experiments, they realized the value of manual markup.

Several speech corpora (some are mentioned above) were also compiled for military purposes to develop systems for the recognition and synthesis of live sounding speech. While tagging words in a spoken corpus the scientists had to pay close attention to transcription and tagging(markup). During this period, the meanings of such terms as "*corpus," "corpus linguistics," "markup," "meta-markup," "concordancer," and "morphological and syntactic tagging"* were defined. Consequently, the terms "*tokenization", "tokens", "lemmatization", "lemma", "segmentation", "normalization", "time alignment"* have appeared and defined in this field of study of language.

*Second-generation corpora, megacorpora.* In the early 1980s, a text markup (tagging) language or meta-language called *SGML (*Standard Generalized Markup Language) was developed. It is a set of tags that standardizes the markup of texts (Baker, Andrew, & McEnery, 2006, p. 149). This format remained the reference format until 2007, when it was replaced by the simplified *XML* (eXtensible Markup Language) format with a more unified and rigorous form of markup to prevent duplication of markup, as was the case with SGML (McEnery & Hardie, 2012, pp. 76-77).

In the 1990s, scholars at Lancaster University developed a number of programs for the following levels of markup: anaphoric referential markup (1992), prosodic markup (1993), semantic markup (1993), (2004), artistic-stylistic markup (1996 and 2004), pragmatic markup (2003), and speaker error markup (1999, 2003) (McEnery & Hardie, 2012, pp. 78, 83, 29).

T. McEnery and A. Hardie claim that the 1990s were the era of second-generation concordance programs which became more comfortable and effective. Second-generation concordancers ran on the IBM platform, so they could be used on personal computers that supported the IBM operating system. Second-generation concordancers, such as Micro-OCP (1988), Longman Mini-Concordancer (1989), Kaye concordancer (1990), etc., were also based on KWIC methodology and performed the following functions: an alphabetical list of concordances with a right and left contextual word environment, a corpus word list, elementary descriptive statistics such as word counts, type-token ratio (word-to-word ratio). However, insufficiency of a unified format, character representation standards, and markings has had a negative impact on the power and performance of second-generation concordancers. (McEnery & Hardie, 2012, p. 40).

In 1987, at a conference at Vassar College in Poughkeepsie, New York, the Text-to-Code Initiative Community was founded. In New York, the Text Encoding Initiative (TEI) community was founded and raised the problem of developing common standards for corpus composition, transcription, and markup (Bernard, 2018). The emergence of a large number of corpora created on the basis of different types of texts led to the need to create a unified set of rules that would contain all the rules for collecting, transcribing, and annotating texts of both oral and written forms of a language. In addition, issues of

ethics and copyright also emerged. Thus, if in the 1970s the use of hidden microphones for recording speech and giving personal names and addresses was considered acceptable, by the 1990s the use of such methods became unethical (Kennedy, 1998, pp. 76-78) (McEnery & Hardie, 2012, pp. 60-69). Thus, the TEI documents (Text Encoding Initiative Principles) became such a set of rules for ethics and copyright (Baker, Andrew, & McEnery, 2006, p. 157).

In 1991, the nonprofit company "Unicode Consortium" developed the Unicode character encoding standard for ASCII (American Standard Code for Information Interchange), designed for all types of written languages of the world, as well as for encoding non-printing characters (transcriptions, mathematical formulas, etc.). Currently, UTF-8 is the most common system for coding texts changing them into machine-readable form (McEnery & Hardie, 2012, pp. 37-38).

After several lively discussions on marking up or tagging of a corpus, in 1993, G. Leech published maxims for compiling meta-texts, i.e., metatexts, or biographical data about texts, with complete extra-linguistic information. According to G. Leech, meta-labeling should meet established requirements and include the following information about the criteria and sources for text selection and markup (tagging/labelling system):

1) the possibility of access to the original version of the material;
2) separate storage for metatext from the main text;
3) listing all set of used markup principles in a separate document;
4) availability of information about the authors of markup and main characteristics of markup (manual / automated by software, etc.);
5) understanding markup as an author's interpretation, its relativity;
6) obligatory presentation of the fullest possible information about the text in the markup, based on generally accepted linguistic principles;
7) inadmissibility of recognizing any markup as a standard but just a reference (Leech G. N., 1993).

From the late 1990s to the 2000s, concordancers WordSmith 0.4 (1996), MonoConc (2000), AntConc (2005) were developed and introduced. The difference of these

programs was the ability to process a large volume of texts of any script, and to perform complex statistical analysis (McEnery & Hardie, 2012, p. 35). Moreover, these and other concordance programs are characterized by their high functionality: one program can quickly compose a list of keywords, concordances, perform frequency analysis and collocation analysis. Definitely, it was a major breakthrough for this period of time.

In consequence, with the advance of technology, since the early 1990s, technical capabilities gave a big opportunity to scholars to compile, develop and construct large-volume corpora. The corpora of that time aimed at covering a wide range of language forms manifested in both written and oral speech, and representing the full diversity of the language. It became possible to automatically tag oral corpora at the prosodic, phonetic, morphological, lexical, syntactic, and discourse levels.

G. Kennedy (1998), Baker, Hardie, McEnery (2012, p. 35) call the corpora developed since the late 1980s till the end of 1990s *megacorpora,* because their volume has approached 100 million word uses. Such corpora traditionally include The Longman Corpus Network (BCN, 1991), The Bank of English, (BoE, 1993), The British National Corpus, (BNC, 1994), The American National Corpus, (ANC, 2008).

It is worth to note that one of the most ambitious projects developed in the late 1980s was the Collins Birmingham University International Language Database, or COBUILD corpus. The corpus was created by a group of scholars led by Sinclair. The project uses the so-called Birmingham Collection of Texts, which includes 20-million-word uses of written and spoken texts. The main corpus contains 7.3 million words, and the so-called "reserve corpus" contains 13 million words. The corpus contains 75 per cent of the written texts, and 25 per cent of the spoken texts published from the 1960s to 1982. The written speech is represented mainly by prose fiction texts. Twenty percent of the corpus is American English (Sinclair, 1987). According to Johansson, the COBUILD project was a breakthrough for its time for a number of reasons: 1) the size of the corpus exceeded 20 million words, 2) the sources were full texts, not short extracts, 3) it was the most representative due to inclusion of texts of oral and written speech of different genres. COBUILD was the most comprehensive corpus of its time and served as a base for formation of the Collins COBUILD Dictionary of English (1987).

After completion of the COBUILD project in 1991, Sinclair began to write about the enlargement of the volume of corpus (1991). In the 1990s, the scholar announced a project to expand the COBUILD corpus and compile on its basis a corpus called *The Bank of English (BoE).* This new project's purpose was to create a "dynamic corpus" named *The Bank of English (BoE)* which would include several hundred million words, which would be continuously updated with new texts of English spoken and written language. This kind of corpus was also called "monitor corpus", because it was expected to help monitor the changes that are taking place in the language currently (Baker, Andrew, & McEnery, 2006, p. 47) (McEnery & Hardie, 2012, pp. 65-116). Like COBUILD, the BoE corpus consists of 75% written texts and 25% spoken texts, with 70% being texts of the British variant of English, 20% of the American variant, and 10% of other national variants of English. By 1997 the English Language Bank corpus had 300 million word uses. For the first time, the truly dynamic corpus was created, the corpus was updated with new texts being added to the corpus every year. G. Kennedy writes that this kind of corpus presented scholars with new tasks in textual processing: up to 2.5 million words were added each month from newspapers (Baker, Andrew, & McEnery, 2006, p. 47). Although the developers were not yet completely convinced of the advisability of using monitor corpora, the COBUILD and BoE corpora formed the new standard or new principle in corpus formation: *balance and representativeness*. According to P. Baker (2006, p. 18), balance as a principle of corpus formation can only be realized in large reference corpora in which both oral and written forms of high, formal, and low registers should be represented. Claiming about the need for a representative corpus, Biber also writes that since a language is a system of different genres or styles, a reference corpus should include all styles and genres of speech, as well as territorial accents and dialects (territorial and regional dialects, sociolects etc.), formal and informal language. Besides, Biber argues that a language should be represented from a historical perspective, i.e., include the texts of all historical epochs too (Biber, Conrad, & Reppen, 1998, p. 12) (Biber, 1993, pp. 243-257). Meeting all principles and requirements about corpus compilation of this epoch, Word Banks Online was considered to be the most representative corpus containing 259.4 million uses of British English (41.4 million uses of spoken language)

and 189.4 million uses of American English (33.1 million uses of spoken language) (1997). This corpus almost represented English language all in its diversity and variety.

Another mega corpus, the formation of which was begun in the late 1980s by a group led by Della Summers, is the Longman Corpus Network. This corpus network is now a commercial database that consists of five main corpus: 1) The Longman Corpus of Learners' English (10 million word uses); 2) The Longman Corpus of Written American English(100 million word uses); 3) The Longman Spoken American Corpus (5 million word uses); 4) The Longman / Lancaster English Language Corpus (a joint corpus of written and spoken English, 30 million word uses) and 5)The Spoken British Corpus (10 million words) (1980). According to G. Kennedy, despite the fact that each of the parts of the Longman Corpus Network was assembled for a specific purpose, when they are combined into one the corpus has become a powerful tool, representing a large variety of texts of different genres of spoken language. Later, the spoken type of the corpus was used to create dictionaries and textbooks on communicative English grammar. Afterwards, it also served as a base for the spoken part of the British National Corpus.

The British National Corpus (BNC) was compiled between 1991 to 1995s at Oxford and Lancaster Universities. The main aim of creating this corpus was to construct a balanced and representative corpus of spoken and written English for academic, lexicographic, and commercial purposes. The corpus of 100 million words includes 10% of oral transcripts and 90% of written texts from the second half of the 20$^{th}$ century. 75% of written texts are texts of informative genre: scientific articles and monographs, political, business, cultural (music, theater) and secular news, religious and philosophical texts, articles from magazines about sports and housekeeping. 25% of the corpus - works of fiction. The balanced oral part of the corpus is divided into two: "contextual" and "demographic" texts. "The contextual" ("the context-governed texts") of the oral English sub-corpus contains texts of various genres and styles: Scientific informative style (lectures, news, classroom discussions, scientific advices and lessons); business (trade shows, meeting with trade unions, medical, legal and professional advice, interviews); public (sermons, political speech, council meetings, parliamentary readings, court hearings); leisure (sports commentary, after-dinner talks, club meetings, radio listener calls). The demographic

texts of the subcorpus presents speech recordings of regional dialects (southern, central, and northern) and different accents in the English language. The texts in the corpus were tagged using the automated tagging program CLAWS5 Tagset developed at Lancaster University. And the markup of the texts have been marked up using the SGML system according to the TEI ( (Lamel L. C., 1997) (Garside & Leech, 1996), (The official website of the *British National Corpus* URL: http://www.natcorp.ox.ac.uk ).

Another major corpus was The International Corpus of English (ICE), developed at University College London under the supervision of S. Greenbaum in 1996. The aim of the project was to collect texts of regional variants of English. The subcorpora include oral and written texts of regional English variants of Britain (ICE-GB), East Africa, India, New Zealand, Singapore, Canada, Hong Kong, Jamaica, the Philippines, the United States, Cameroon, Fiji, Ireland, Kenya, Malta, Malaysia, Pakistan, Sierra Leone, Sri Lanka, and Trinidad and Tobago. In total the corpus contained 60% of written texts and 40% of spoken texts. The morphological tagging is done with the help of CLAWS7 Tagset, the semantic one is based on the UCREL Semantic Analysis System (USAS) (Greenbaum, 1991) (Geenbaum, 2021).

Consequently, second-generation corpora contained at least one hundred million words, whose aim is to represent the written and spoken language in all its diversity. Typically, these are corpora available online, assembled and tagged according to TEI requirements. National corpora became monitor corpora and were compiled based on the principles of representative selection of texts and according to the rules that belongs to only second-generation corpora. In the 1990s, the British National Corpus was used as a new corpus model, and TEI became the standard for corpus compilation, which recommended the SGML markup language. The period from 1987 to 2004 saw the development of corpus regulations and rules for corpus collection, meta-tagging, and automated text markup software were developed.

*Third-generation corpora, or gigacorpora.* The beginning of the 2010s was marked by the emergence of great technical capabilities. Technical changes had great impact on corpus linguistics by providing with tremendous technical possibilities. Thus, a corpus manager and text analysis softwares, concordancers like BNCweb (2009), CQPweb

(2012), SketchEngine (2013), Wmatrix (2013) were developed. The creators of these softwares strived to solve the following problems in corpus linguistics: limited power of personal computers, incompatibility of personal computer operating systems, and legal restrictions on the distribution of corpora. To solve the legal issues, tthey simplified the procedure for obtaining access, the corpora switched to online versions, which increased the speed of processing requests and expanded the number of users. Direct access became available through a web browser equipped with an online search engine. The corpora operated online and allowed for users to make a contrastive analysis of a small private corpus like BNC corpus or texts from the Internet. M. Davies calls the concordancers that are developed in this period of time, hybrid corpora, because their interface became a kind of common field for creating corpora and performing frequency analysis on the morphemic, lexical syntactic and phrasal levels (2015).

After the 2000s, the trend of increasing the volume of corpora persisted. Mauranen (2013), Kuebler and Zinsmeister (2015, p. 10) characterize the corpora that are created in this era by the motto *"the bigger the corpus, the better".* And L. Flowerdew is the first to refer to this period as the age of the gigacorpus generation. (2004). Because at this time, a number of new corpora (COCA, Google Books Ngram) (see below) arose, amounting billions of words A huge number of corpora gave the opportunity to researchers to undertake larger-scale frequency studies, gain a high proportion of outcomes. and to investigate collocations of three, four, or more words.

Moreover, many linguists started to study collocations consisting of three, four, or more words easier. The linguists, Biber (2006) and K. Hyland (2008) call collocations "lexical bundles" in their works. In this type of "lexical bundles" one word can be variable. For example, in collocations of five words: *in the beginning of the, in the end of the, in the form of the,* the third word is variable, we can change with the words like *kind, type or format* etc. Subsequently, in the corpus linguistics a new notion called *n-grams* showed up, where bigrams are collocations consisting of two words, trigrams are collocations consisting of three words, and n-grams are collocations consisting of *n-words* (2015). Nowadays, to investigate collocations has become feasible thanks to the creation of large gigacorpora (Google Ngram, Google Books, COCA, etc.).

In 2008, The Corpus of Contemporary American English (COCA) was published, with a total volume of approximately 400 million words currently in use. The corpus includes both spoken and written texts. Written speech is collected from genres as fiction: short stories and plays from literary magazines, children's literature, first chapters of books published since 1990, and film scripts (113 million words); popular magazines from Time, Cosmopolitan, Men's Health, Good Housekeeping, Fortune, Christian Century, Sports (118 million uses); newspaper articles from 10 newspapers all over America: USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle (114 million uses); scientific articles from journals in different scientific fields (112 million uses) (The official website of *Corpus of Contemporary American English(COCA* available at URL: https://corpus.byu.edu/coca/). In the COCA corpus, the volume of spoken texts is 118 million words. This sub-corpus contains transcripts, video and audio recordings of a wide range of radio and television programs: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Today Show (NBC), 60 Minutes (CBS), Hannity and Colmes (Fox). The COCA corpus is dynamic, adding 20 million words a year. POS-tagging of texts is done with the CLAWS program. The corpus is accompanied by the WordAndPhrase concordance program's platform. Currently, the amount of the corpus reached to 520 million words.

The Google Books Ngram Viewer corpus of digitized texts, which includes the texts of over one billion e-books written between 1500 and 2008, was released in 2009. The Google N-gram Corpus surpassed 200 billion words in 2011 (Google Ngram Viewer, 2022). The Google Books Corpus was updated in 2014, containing 155 billion usage of written American English conversation and 34 billion uses of written British English discourse. In addition to English texts, the Google Books corpus includes works in six other languages: Spanish, French, Russian, German, Italian, and Hebrew, but in considerably lower quantities.

The next goes the Global Web-based corpus of English (GloWbE) (2013) containing 1.9 billion words. The corpus aims to represent as many regional variants of English around the world as possible. This corpus comprises the texts of web pages and websites of 20 regional variants of English (Davies, 2013).

The News on the Web (NOW) corpus (2016) currently exceeds 5.7 billion words. The corpus contains English-language texts from "2012 to yesterday". Every day the corpus is updated with texts for 4-5 million words. Every night, from 22:00 to 1:00, the texts are uploaded to the corpus: the program HTTrack reads the URLs from Google News and uploads to the corpus 9-10 thousand texts, and then, using the program JusText, the repeating and template texts are removed. Texts are tagged and lemmatized using CLAWS 7, and then the tagged-ready texts are added to the main body of the corpus. The site of the corpus also can, for example, track the most popular word of the day or of the year (Davies, 2018).

Unfortunately, the emergence of mega- and giga-corpus has shown that large reference corpora are unsuitable for studying the speech of particular professions or genres. Large corpora, despite their enormous size, contain mostly texts of the most common genres of oral and written speech (Laurence, 2013) (Davies, 2015) (Maurenan, 2013) (Sandra Kuebler., 2015) (Flowerdew, 2004). In the late 1990s and early 2000s it was proved that the principles of representativeness of special corpuses are observed in case of much smaller volumes, because the frequency of both terms and neutral words remains stable and unsteady (Sinclair, 2005) (Tognini-Bonelli, 2001).

Nevertheless, this era is distinguished by the fusion of corpus linguistics methods with the World Wide Web: automatic downloads of texts from the Internet, as in the case of the NOW and GloWbE corpora, treatment of the World Wide Web as a corpus (the Google Books corpus), and the incorporation of the tools themselves into the World Wide Web (Sketch Engine, BNC web). At this point, discussions regarding n-grams have gotten more significant. Furthermore, it became feasible to follow the evolution of a certain term's use on big data sets, such as the change in the form and meaning of the word through time in written (Google Books) or spoken language (COCA, NOW, GloWbE). Moreover, the emergence of corpora with large amount of texts has not diminished the relevance of the question of the need and representativeness of small corpora of professional speech.

### 1.2 Formation of Corpus linguistics as a Field of Linguistics

The terms "*corpus"* and *"corpus linguistics"* are not new in the world of scientific community of the twenty-first century. It is claimed that these two terms are defined as some of the major components of scientific and technological progress, since corpus linguistics is closely connected with the technology and is a part of the applied and computational linguistics as well. The history of corpus linguistics dates back to the middle of the XX century, or to be more precise, to 1961 when the first Brown American English Corpus was created at Brown University in the USA. The authors of this corpus, which consisted of one million words (500 texts of 2,00- words each), were H. Kučera and W. N. Francis (Zakharov & Bogdanova, 2013, p. 11). It was definitely distinct from most other linguistic topics, since it was not explicitly concerned with the study of any definite field of language. Accordingly, this linguistic news drew the attention and deep interest of the scientific community of the time, and it sparkled a number of lively debates and discussions in addition to the public response. And there was also a lot of negative comments from linguists at first. As T. McEnery and E. Wilson (2012, p. 25) point out in their book called "Corpus Linguistics", one of such eager critics who rejected the new trend at the time was Avram Noam Chomsky. Chomsky (Andor, 2004) expressively opposes the type of research evidence that corpora entail.

However, later on, researchers in linguistics started to use corpus data in their works, even to the degree that in the twenty-first century. Preliminary to corpus linguistics, much of the empirical studies into language has been done based on a manual analysis of a few texts. At that time, linguists had been constrained by a small number of texts that human could collect, handle, manage, and analyze effectively. Emergence of corpus linguistics, in the last two decades especially, made a boost, and a significant turn-around to the empirical study of language. Eventually, it started to play an important role in linguistic world entirely, and an increasing number of linguists did not imagine their works without corpus data. Linguists began to judge corpus linguistics not on the basis of a theory or philosophical argument about it, but rather on the huge amounts of results that corpus had produced. Adopting such a corpus-based approach, researchers have used the previous advocacy of empiricist views by British linguist J.R. Firth, who popularized the phrase "*You shall know a word by the company it keeps*,". (1957, p. 11)However, American structuralists (the "post-Bloomfieldians"), particularly Zellig Harris, exhibit an empiricist

corpus-based methodology in their work in a way that is probably even more obvious. For instance, he claims that the corpus-based approach (Harris, 1951) is an attempt to identify automatic methods for discovering a language's structure. Moreover, corpora turned out to be fruitful in a range of areas of linguistics (discourse analysis, language learning, semantics, pragmatics, sociolinguistics, theoretical linguistics, as a source of data for language description (lexicography, grammar), providing researchers with huge and diverse insights about their interest of work. Nowadays, it is hard to find a field of linguistics where a corpus-based approach has not been implemented fruitfully. Houston ( (2002, p. 1), notes that it is not an exaggeration to state that corpus linguistics has *"revolutionized language study"*. Due to corpus linguistics, linguists can now not only explore texts with millions of words with relative ease, but they are also aware of the fascinating insights that can be derived from the application of corpus methods to textual analyses: insights that would have been missed in a human-only analysis.

### 1.3 What is Corpus Linguistics?

Prior to defining a corpus linguistics, it would be accurate to determine the word *"corpus"* itself. The word *"corpus"* comes from Latin which means *"body"* and plural form is *corpora.* Wikipedia (2022) defines corpus as "large and structured collections of texts (now usually in electronic form) that are used for statistical analysis and hypothesis testing, case verification, or justification of language rules for specific domains". The most renowned corpus linguists T. McEnery and A. Wilson define a corpus as follows: "*a corpus is a collection of language extracts which are chosen according to specific linguistic criteria to be used as a language model"* (2001, pp. 2-3).

According to Baker ( (2010, p. 93) "corpus linguistics is a growing branch of linguistics that comprises the investigation of (typically) very large collections of electronically stored texts with the use of computer software". As a result, we may infer that corpus linguistics is a large "body"' of texts recorded electronically in a computer in order to do specific inquiry or research about language usage. Another linguist Sinclair (1991, p. 2) defines it as "a collection of naturally-occurring language text (only authentic texts), selected to characterize a state of variety of a language". Corpus linguistics, due to its authenticity, is regarded to be one of the most reliable sources for researchers that can

be implemented in a variety range of disciplines starting from language learning and teaching to a language usage like natural language processing (NLP) in computer science. McEnery and Hardie ( (2010, p. 1) claim that corpus linguistics deals with "a group of machine-readable texts that is considered to be appropriate as a foundation for investigating a specific set of research concerns."Corpus linguistics is therefore a methodology or an approach implemented to gain qualitative and quantitative data analysis. İts primary resource is a corpus(authentic texts in eletronic form and annotated). In a nutshell, corpus linguistics can be represented as a set of methods, procedures, and resources dealing with empirical data in linguistics.

Briefly, it is a collection of texts (a body of systematically gathered texts and transcribed speeches to represent a particular use of a language that is used for linguistic analysis), which is done first by uploading it onto corpus software, and then we can apply specific query tools (engines) with the help of computer software like finding frequency counts or concordance lists etc. and then derive results. It is obvious from the given definitions above that the development and usage of corpus linguistics has been closely connected to computers. Moreover, it is crucial to understand that our usage of corpus data is not restricted as long as technology develops day by day.

### 1.4 Benefits of Using Corpora in Linguistics

The term *"corpus"* also includes a textual and linguistic data management system, most recently referred to as *a corpus manager*. This is a specialized search system, which includes software tools for searching data in a corpus, obtaining statistical information and providing the user with results in a convenient form.

A search in the corpus allows you to build *a concordance* for any word - a list of all the uses of a given word in the context with references to the source. The corpora can be used to obtain a variety of inquiries and statistical data about real language usage and speech units as well. In particular, on the basis of corpora a researcher can acquire data about the frequency of word forms, lexemes and grammatical categories, trace the changes in the frequency and context in different periods of time and dynamics of changes in a language within a certain period of time, obtain data about the occurrence of lexical

units, etc. A representative corpus of linguistic data within a set of a certain period allows us to study the dynamics of the processes of changing the lexical composition of the language, to analyze the lexical and grammatical characteristics in different genres and by different authors.

Moreover, the corpus is intended to serve as a source and tool for multidimensional *lexicographic work* on a variety of historical and contemporary dictionaries. For instance, Samuel Johnson's "A Dictionary of the English Language", sometimes known as Johnson's Dictionary, was published on April 15, 1755. It is one of the most significant dictionaries in English language history. In June 1746, a group of London booksellers, who were dissatisfied with the dictionaries of the time, proposed Johnson to prepare a dictionary. It required seven years to complete the first version of the dictionary. He accomplished it all by himself, with just clerical aid of James Murray whose job was to reproduce the illustrative passages that Johnson had highlighted in the books. Throughout his life, Johnson published multiple updated versions and it took almost a half of century (Wikipedia contributors, 2022). Nowadays with the help of corpus data, lexicographers are able to accomplish their task in a short period of time and save a heaps of time. So, this is one of the advantages of corpus linguistics that can bring to linguistics.

Traditional school *grammars* and *textbooks* are often illustrated with artificially produced or edited examples of language use. In the future, they will be of little help to students who sooner or later have to deal with real language environment or real communication in their life. In this respect, corpora as sources of empirical and real (authentic) data play an important role in *linguistic pedagogy*. In language teaching, corpora provide a source for stimulating students' interest and motivate them to engage in independent study of authentic language use. An important application of corpus data is *Computer-Assisted Language Learning (CALL) Technology*, where corpus-based software is used to support interactive learning activities performed by students with the aid of computers (Potapova, 2005, pp. 3-4). Nineteenth-century grammarians illustrated their statements with examples taken from the works of recognized authors. For example, H. Paul in his German Grammar used the works of the German "classics" to illustrate each of his statements - in phonology, morphology, and syntax (Auer & Murray, 2015).

Today, grammarians also use a corpus approach, but the corpus now includes not only the classics, but also other types of texts from different variants, registers (genres), etc. And consequently, this allows the language to be described more adequately and sufficiently. In particular, nowadays there is a growing interest in oral grammar of a language.

In the early studies, the individual researchers were restricted to the small number of texts that they could acquire, manage, and analyze properly. Throughout the whole research process, they needed to go to libraries and get card index of a book and find that book from the shelves according to *card index* that were given by librarians. After they had to read a whole book in order to get necessary information and write them down, etc. Thus, it was time-consuming and by necessity a linguist was constrained to a few numbers of books. In the last two decades, due to corpus linguistics analysts can now not only explore texts with millions of words in a short period of time, sometimes even in seconds relatively easy. Furthermore, they are also aware of the huge number of results that can be derived from the application of corpus methods to textual analyses: results that would have been missed in a human-only analysis (Ngula, 2018, pp. 1-2). Nowadays, many linguists use the corpus as an "example bank," i.e., trying to find empirical support for the hypotheses, principles, and rules that they are working on. Examples, of course, can be retrieved at random, but the corpus linguistic approach provides a representative and balanced linguistic material, as well as a search tool that usually enables researcher to obtain a significant outcome from a particular corpus.

Corpus linguistics is sometimes referred to as "a bundle of methods from different areas of linguistic inquiry" (Lüdeling & Kytö, 2008). As a method of linguistic analysis, corpus linguistics is also related to contrastive research, which aims to establish the facts of similarity and difference between languages, dialects or language variants in the course of their comparative study, even it is possible to make diachronic research in a set linguistic interest of study.

The basic but primary philosophical concept underlying behind corpus linguistics can be considered in the following two ways:

a) a cognitive urge to understand how people use language in their daily communication activities;
b) if it is feasible to construct intelligent systems that can efficiently communicate with humans. With this motivation in mind, computer scientists and linguists have collaborated to create a language corpus that can be used to design intelligent systems (e.g., machine translation systems, language processing systems, speech understanding and recognition systems, text analysis and understanding systems, computer aided instruction systems, and for the benefit of the language community as a whole.

Finally, whereas a field of linguistics such as syntax, semantics and sociolinguistics aim to describe or evaluate one specific language periphery or language use, corpus linguistics is a broader concept, a methodology that can be applied almost to all aspects of language.

### 1.5 Classification of Corpora

Due to that corpus linguistics is a new concept in linguistics, it is not yet developed a clear understanding of what should include a corpus and how it should be categorized. Categorization criteria are can be both external and internal according to nature of corpus linguistics. External factors such as participants, occasion, social situation, communicative function of a language. Internal criteria are concerned with the language patterns inside a corpus. Taking all of these factors into account, we suggest the following classification of corpora according to categories such as text genre, data nature, text type, design purpose, and application nature (Dash, 2010).

**a) Genre of text**

- *Written Corpus*: A written corpus, by virtue of its genre contains only language data collected from various written, printed, published and electronic sources.
- *Speech Corpus:* A speech corpus (e.g., Wellington Corpus of Spoken New Zealand English) contains all formal and informal discussions, debates, previously made talks, analysis, casual and normal life talks, dialogues,

monologues, various types of conversation, on line dictations, instant public addressing, etc. There is no limitation to media involvement in such texts.

- *Spoken corpus* (e.g., London-Lund Corpus of Spoken English), a technical extension of speech corpus, contains texts of spoken language. In such corpus, speech is represented in written form without change except transcription. It is annotated using a special form of phonetic annotation/tagging.

**b) Nature of Data**

- *General Corpus*: A general corpus (for example, the British National Corpus) is a collection of general texts from various disciplines, genres, subject fields, and registers. Given the nature of its shape and function, the number of text collections is limited. That is, the number of text kinds, as well as the quantity of words and phrases, are restricted. It has the potential to evolve over time and to include new data as fresh texts become available. It is quite huge in size, rich in diversity, extensive in representation, and has a very broad range of applications.
- *Special Corpus:* A special corpus (for example, the CHILDES Database) is created from texts taken from a general corpus for specific variation of language, dialect, and subject, with an emphasis on unique and specific aspects of the issue under inquiry. Its size and content vary depending on the purpose. Because it has a significant number of texts which shows unusual features of language layers, that does not represent to the description of a language wholly. Its source is untrustworthy since it collects data from people who are not acting normally. Special corpus is not balanced (except within the limits of its intended function) and provides an inaccurate and indirect information of language segments when applied for other purposes. It differs in principles, because it features and is limited to one or more types of normal, authentic language variety. Because of the non-representative nature of the language concerned, corpus of language of children, non-native speakers, users of dialects, and particular social areas of communication (e.g., auction, medical discussions, gambling, court process, etc.) are defined as special corpus. Its key benefit is that the texts are chosen in such a manner that the phenomena of interest appear more frequently in it than in a balanced corpus.

- *A sublanguage corpus*, as the name is suggesting, is one that contains only one text variant or one stratum of a certain language. Because of homogeneity of its structure and specialized lexicon, the amount of data required to demonstrate normal authentic language is not large.
- *Sample corpus*: A sample corpus (Zurich Corpus of English Newspapers) is a type of special corpus composed of texts chosen with great care and thoroughly analyzed. Once a sample corpus is created, it is not added to or modified in any manner (Sinclair, 1991, p. 24) since any modification will disturb its composition and affect study requirements. Samples are few in number and of constant size in comparison to texts. As a result, they do not qualify as texts.
- *Literary corpus*: A special category of sample corpus is literary corpus. There are many kinds of literary subcorpora since the literature is very diverse in nature. Classification criteria considered for such corpus include author, genre (e.g., odes, short stories, fictions, etc.), period (e.g., 15$^{th}$ century, 18$^{th}$ century, etc.), group (e.g., Romantic poets, Augustan prose writers, Victorian novelists, etc.), theme (e.g., revolutionary writings, family narration, industrialization, etc.) and other valuable parameters.
- *Monitor corpus* (e.g., Bank of English): A monitor corpus is a growing and developing due to non-finite collection of texts with the potential for continual data induction and augmentation to reflect changes in language. The constant expansion of the corpus reflects changes in language while maintaining the relative weight of its components as indicated by criteria that are stable. Year after year, the same composition schema is used. The monitor corpus is built using texts of spoken or written in a particular year (Sinclair, 1991, p. 21). From monitor corpus we find new words, track a variation in usage, observe change in meaning, establish a long-term norm of frequency distribution, and derive a wide range of lexical information.

**c) Type of Text**

- *Monolingual corpus:* It (e.g., ISI Bengali Corpus) contains representative texts of a single language representing its use in a particular period or in multiple periods. It contains both written and spoken text samples. A monolingual corpus is the

most frequent type of corpus. It contains texts in one language only. The corpus is usually tagged for parts of speech and is used by a wide range of users for various tasks from highly practical ones, e.g. checking the correct usage of a word or looking up the most natural word combinations and definitions, scientific use, e.g. inquiring frequent patterns or new trends in a language.

- *Bilingual corpus:* A bilingual corpus (e.g., TDIL Bengali-Oriya Corpus) is created by combining corpora from two related or unrelated languages. If these languages are genetically or typologically similar, they form a parallel corpus (explained further below), in which texts are aligned according to certain predetermined characteristics. Size, content, and field may differ between corpora, which is not permitted in the case of parallel corpora.
- *Multilingual corpus:* Multilingual corpus (e.g., Crater Corpus) contains representative collections from more than two languages. Generally, here as well as in bilingual corpus, similar text categories and identical sampling procedures are followed despite the fact that texts belong to different languages.

**d) Purpose of Design**

- *Unannotated corpus*: It (e.g., TDIL Corpus) represents a simple raw state of plain texts without any additional linguistic or non-linguistic information.
- *Annotated corpus*: It (for example, British National Corpus) comprises tags and codes attached by designers, linguists and computer programmer to record extra information (analytical markings, parts-of-speech tagging, grammatical category information, and so on) into texts. Annotated corpus, as opposed to unannotated corpus, is better suited for delivering relevant information used in a variety of language technology activities such as morphological processing, sentence parsing, information retrieval, *word sense disambiguation*, machine translation, and so on.

**e) Nature of Application**

- *Translation Corpora:* Translation corpora are collections of original source language (L1) materials and their translations into target language(L2). These corpora often maintain the meaning and function of words and phrases across languages, and so provide a perfect foundation for comparing the equivalents of

certain meanings in two distinct languages under identical conditions. Furthermore, they enable the discovery of all cross-linguistic variations, i.e., alternate usage of certain meanings and concepts. As a result, translation corpora provide more fruitful resources for cross-linguistic data analysis, comparison analysis as well as rule development required for translation.

- *Aligned corpus:* It is (e.g., The Canadian Hansard Corpus) a kind of bilingual corpus where texts in one language and their translations into other language are aligned, sentence by sentence, phrase by phrase, or even word by word.
- *Parallel corpus:* A parallel corpus (for example, the Chemnitz German-English Corpus) comprises texts as well as translations in each of the languages involved, allowing for the double-checking of translation equivalents. Texts in one language are matched with their translations in another: sentence by sentence, phrase by phrase, or even word by word. Reciprocal parallel corpora are sometimes developed, with corpora comprising actual texts and translations in each of the languages concerned.
- *Reference corpus* (for example, the Bank of English) is designed to provide comprehensive and representative information about a language. It is vast enough to include all significant language variations and characteristic vocabulary, allowing it to be used for constructing grammars, dictionaries, thesauruses, and other resources. It is built on important factors agreed upon by the language community. It comprises both spoken and written language, as well as formal and informal registers expressing distinct social and situational registers. It is used as a 'benchmark' for lexicons, general tool performance, and language technology applications. With the expanding importance of internal criteria, reference corpus is being used to quantify special corpus deviation.
- *Comparable corpus*: A collection of 'similar' texts in more than one language or variety (e.g., Corpus of European Union). It comprises works in many languages that are not the same in topic, genre, or register. These are used to compare various languages. It follows the same composition pattern, however there is no consensus on the nature of resemblance because there are few comparable corpora. It is

essential for comparing languages and creating bilingual and multilingual lexicons and dictionaries.

- *Opportunistic corpus:* An opportunistic corpus is collection of electronic texts that may be accessed, transformed, and utilized for free or at a low cost; nevertheless, it is frequently unfinished and incomplete. As a result, consumers are left to fill in the blanks for themselves. They have a place in circumstances where size and corpus access are not an issue. The opportunistic corpus is a virtual corpus in the sense and it is a source of the real corpus (from the opportunistic corpus) which is created based on the demands of a certain project. Monitor corpus is commonly referred to as an opportunistic corpus.
- *Learner corpora:* Learner corpora, which can be broadly defined as electronic collections of texts produced by language learners, have been used to perform two distinct but related functions: they can contribute to Second Language Acquisition theory by providing a better description of interlanguage (i.e. transitional language produced by second or foreign language learners) and a better understanding of the factors that influence it; and they can be used as a tool for pedagogic purposes (Dash, 2010).

There are some other types of corpora exist such as *closed corpus, synchronic corpus, historical corpus (diachronic corpus), dialect corpus, idiolect corpus, and sociolect corpus,* and others. As a result, the categorization criteria described here is not absolute and final. It is open to re-categorization as well as sub-classification based on several other characteristics.

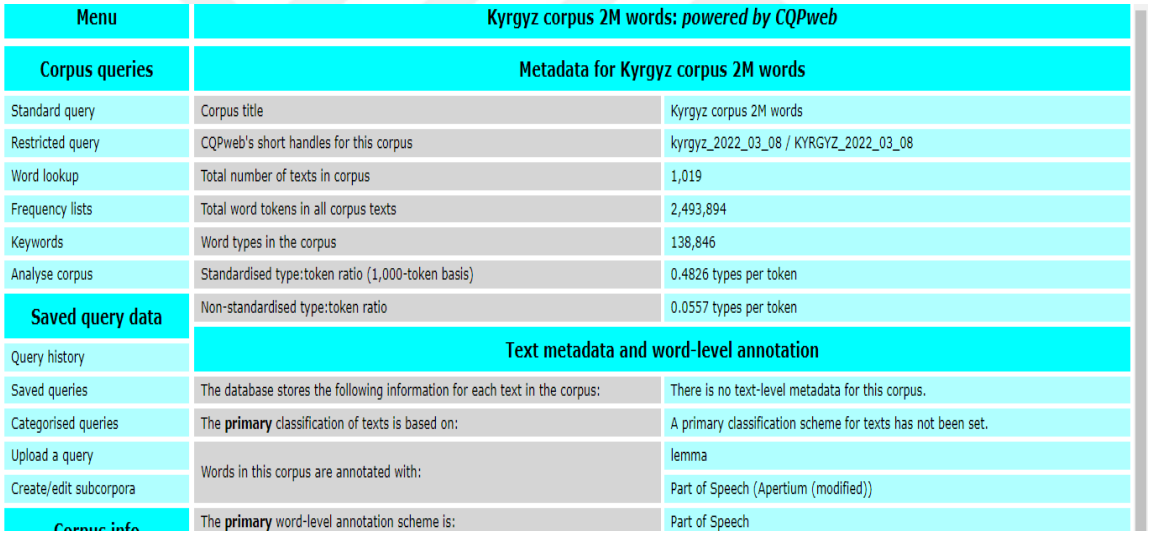
### 1.6 The compilation and development of the Kyrgyz Corpus

According to our prior investigation, we can claim that this is the first corpus, containing literary Kyrgyz texts with part-of-speech (POS) tagging. The Kyrgyz corpus was compiled in April 2019 within the framework of the DAAD exchange program as a result of collaboration of the Universität des Saarlandes and the Kyrgyz-Turkish "Manas" University (2019). At the moment the Kyrgyz Corpus comprises 2,493,894 words from texts of mostly literary genres (epics, novels, stories, fairy-tales, etc. in the poetic form and prose) and mass media (Newspaper "Erkin-too"). All texts were compiled from the

website bizdin.kg (texts comprising about 1 243161 words) and from “Erkin-Too” newspaper (texts including 1 000 000 words) using the special copyright permission from both entities given for creating Kyrgyz Corpus and processing Kyrgyz language through computer.

The corpus is annotated with part-of-speech tags and provided extralinguistic and per-text meta-data, and made available under a free license from CLARIN-D. (Kasieva, Knappen, Fischer, & Teich, 2019).

The corpus title is: The Kyrgyz Corpus (2019-04-18): powered by CQPweb (Corpus Query-Processor). Every document contained in the corpus is stored in a plain text format in the UTF-8 encoding. In the Menu section there is a subsection Corpus Info that comprises Corpus Metadata.



| Menu | Kyrgyz corpus 2M words: *powered by CQPweb* | |
|---|---|---|
| **Corpus queries** | **Metadata for Kyrgyz corpus 2M words** | |
| Standard query | Corpus title | Kyrgyz corpus 2M words |
| Restricted query | CQPweb's short handles for this corpus | kyrgyz_2022_03_08 / KYRGYZ_2022_03_08 |
| Word lookup | Total number of texts in corpus | 1,019 |
| Frequency lists | Total word tokens in all corpus texts | 2,493,894 |
| Keywords | Word types in the corpus | 138,846 |
| Analyse corpus | Standardised type:token ratio (1,000-token basis) | 0.4826 types per token |
| **Saved query data** | Non-standardised type:token ratio | 0.0557 types per token |
| Query history | **Text metadata and word-level annotation** | |
| Saved queries | The database stores the following information for each text in the corpus: | There is no text-level metadata for this corpus. |
| Categorised queries | The **primary** classification of texts is based on: | A primary classification scheme for texts has not been set. |
| Upload a query | Words in this corpus are annotated with: | lemma |
| Create/edit subcorpora | | Part of Speech (Apertium (modified)) |
| **Corpus info** | The **primary** word-level annotation scheme is: | Part of Speech |

Figure 1.1 The home-page of the Kyrgyz Corpus

The Kyrgyz Corpus is available at https://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418/index.php?thisQ=corpusMetadata&uT=y

The Kyrgyz Corpus has a web-based corpus management system, called CQP-Corpus Query-Processor (Hardie, 2012) that helps navigate the data and retrieve necessary information. To solve many different linguistic problems, a necessary condition is that corpus texts contain linguistic and metalinguistic information - markup or annotation corresponding to different levels of linguistic description - phonetic, morphological, syntactic, semantic and others. This kind of annotation become even more complicated

due to the agglutinative nature of the Kyrgyz language. Turkic Lexicon Apertium (Washington, Ipasov, & Tyers, 2012) (Washington & Tyers, 2018)(Washington & Tyers, 2018; Washington, 2012), an open-source machine translation platform has been selected as the most appropriate toolkit for POS tagging and parsing issues of the corpus. (Kasieva & Kadyrbekova, 2021, бет. 208-210).

*Table 1. 1 The POS tagset description of the Kyrgyz Corpus*

| # | Description | Apertium | LPs |
|---|---|---|---|
| | **Noun:** | | |
| 1 | *noun* | n | sg/pl; |
| 2 | *proper* | np | Case: nom/acc/ dat /gen /abl /loc; Possession: px1sg/px2sg/px3sg; px1pl/px2pl/px3pl; px3sp |
| | **Verb:** | | |
| 3 | *standard verb* | v | tv/iv; actv/pasv; |
| 4 | *modal verb/word* | vbmod | Tense: pres; past; fti; cni (Conditional); fts |
| 5 | *auxiliary* | vaux/cop | (future subjunctive); Mood: imp/ind/itg; neg; |
| | **Adjective:** | | pst (positive)/ comp (comparative)/sup |
| 6 | *adjective* | adj | (superlative) |
| | **Adverb:** | | |
| 7 | *adverb* | adv | |
| | **Pronoun:** | | |
| 8 | *pronoun* | prn | Person: p1/p2/p3/impers |
| 9 | *personal pronoun* | pers | |
| 10 | *indefinite pronoun* | ind | |
| 11 | *interrogative* | itg | |
| 12 | *pronoun* | | |
| 13 | *demonstrative* | dem | |
| 14 | *pronoun* | | |
| | *possessive pronoun* | pos | |
| 15 | *reflexive pronoun* | ref | |

| | | | |
|---|---|---|---|
| | **Auxiliary noun:** | | |
| 16 | *postposition* | post | |
| 17 | *past partciple* | pp | |
| | **Numeral:** | | |
| 18 | *cardinal* | num | |
| | *ordinal* | ord | |
| 19 | **Interjection:** | | |
| 20 | *interjection* | ij | |
| | **Conjunction:** | | |
| 21 | *coordinating conj.* | cnjcoo | |
| 22 | *subordinating conj.* | cnjadv | |
| | present participle | | |
| | gerund | pprs | |
| | | ger | |

The above Table 1 depicts the adaptation of the Kyrgyz parts of speech into Apertium tagset symbols and offers a detailed explanation of the developed tagset, which corresponds to Apertium symbols. The table under consideration offers a set of tags organized into 9 main parts of speech. Each POS tag is also associated with a set of linguistic features. As a consequence, 22 fundamental tags were created using Apertium symbols and taking into consideration the nuances of Kyrgyz language. However, choosing the correct Apertium symbol for various Kyrgyz parts of speech proved to be challenging.

Table 1.2 Linguistic Property assimilated into the POS tagset design

| # | Linguistic property | Code |
|---|---|---|
| 1 | Number | N |
| 2 | Possessiveness | Px (1,2,3) sg/pl |
| 3 | Person | P |

| | | |
|---|---|---|
| 4 | Case | nom, acc, dat, gen, abl, loc |
| 5 | Negation | Neg |
| 6 | Tense | pres, past, fti |
| 7 | Mood | imp, ind, itg, fts (future subjunctive), cni (conditional) |
| 8 | Voice | Actv /pass |

Table 2 contains main linguistic properties defined in the Kyrgyz grammar adapted for the codes in Apertium: includes necessary details relating to the features and peculiarities of the Kyrgyz language grammar.

The process of tagging in the Kyrgyz language ("splitting up" or "tokenization" prior to the POS tagging of each word into a token according to the context) is very complex due to the fact that not always and not in all cases the use of one tag will be appropriate. This is also due to the fact that the unification process has not yet been done, so there is a lot of work to be done both on working out the POS annotation and increasing its levels, as well as increasing the number of words in the corpus of the Kyrgyz language. Another problem we encountered during the morphological parsing is the problem of compound word combinations, verbs and, especially, phraseological phrases (Касиева & Сатыбекова, 2020, бет. 3-4).

It is worth noting that the issue of Part-of-Speech tagging of word forms in the Kyrgyz language earlier was studied in the article by T. Sadykov et.al (2018, бет. 90-94). Later Kochkonbaeva (2019)also wrote dissertation paper entitled "Development of morphological analyzer models and algorithms for Kyrgyz language". In this paper, a formal model of the morphological structure of the Kyrgyz language has been developed,

which allows the implementation of morphological analysis algorithms and word normalization.

Approximately 4.4 million people around the world speak Kyrgyz, which is the official language in Kyrgyzstan. It belongs to the Turkic language family and has a rich agglutinative morphology with word structures formed by productive affixations of derivational and inflectional suffixes to root words. A simple example of a Kyrgyz word formation is:

***(сиздер) камсыздандырылгандардансыздарбы (kamsyzdandyrylgandardansyzdarby)?***

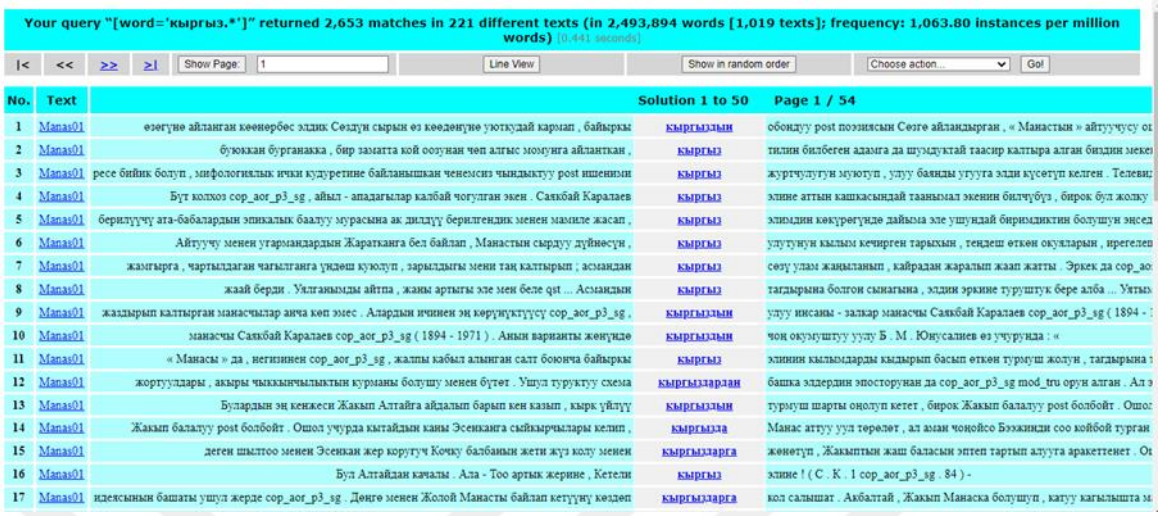
Which can be broken down into morphemes as follows:

***Камсыздандыр <v> <caus> + ыл <v> <pass> + ган <gpr_perf> + лар <pl> + дан <n> <abl> + сыздар <p2pl> + бы <quest>***

***And in the Kyrgyz language it sounds like: Камсыздандыр <эт.><арк.м> + ыл <эт.> <туюк м.> +ган <атоочт.> + лар <көп с.> +сыздар <2ж.көп м.> + бы <сур.м>.***

This verb can be translated into English as ***"Are you from those who have been insured?"***. For more details of the Kyrgyz language's grammar and word formation (morphology) one can refer to a number of books (Abduvaliev & Sadykov, 2008) (Davletov, 1980) (Abduvaliev, 2015).

Kyrgyz is currently a language with limited resources in terms of corpus linguistics; there are also other available Kyrgyz web corpuses, but without annotations. For this reason, the current corpus is the first Kyrgyz corpus to contain a morphological annotation that has been manually annotated beforehand. And the following is the sample search results for the word ***"Kyrgyz"*** in the corpus of the Kyrgyz language:



Your query "[word='кыргыз.*']" returned 2,653 matches in 221 different texts (in 2,493,894 words [1,019 texts]; frequency: 1,063.80 instances per million words) [0.441 seconds]

|< << >> >| Show Page: 1 Line View Show in random order Choose action... Go!

Solution 1 to 50 Page 1 / 54

| No. | Text | | | |
|---|---|---|---|---|
| 1 | Manas01 | өзөгүнө айланган көөнөрбөс элдик Сөздүн сырын өз көөдөнүнө уюткудай кармап , байыркы | кыргыздын | обондуу post поэзиясын Сөзгө айландырган , « Манастын » айтуучусу ог |
| 2 | Manas01 | буюккан бурганакка , бир заматта кой оозунан чеп алгыс момунга айланткан , | кыргыз | тилин билбеген адамга да шумдуктай таасир калтыра алган биздин мекен |
| 3 | Manas01 | ресе бийик болуп , мифологиялык ички кудуретине байланышкан ченемсиз чындыктуу post ишеними | кыргыз | жүргүлүгүн муютуп , улуу баяңды угууга элди күсөгүп келген . Телеви |
| 4 | Manas01 | Бүт колхоз cop_aor_p3_sg , айыл - ападагылар калбай чогулган экен . Саякбай Каралаев | кыргыз | элине аттын кашкасындай таанымал экенин билчүбүз , бирок бул жолку |
| 5 | Manas01 | берилүүчү ата-бабалардын эпикалык баалуу мурасына ак дилдүү берилгендик менен мамиле жасап , | кыргыз | элимдин көкүрөгүндө дайыма эле ушундай биримдиктин болушун эңсеп |
| 6 | Manas01 | Айтуучу менен угармандардын Жаратканга бел байлап , Манастын сырдуу дүйнөсүн , | кыргыз | улутунун кылым кечирген тарыхын , тең деш өткөн окуяларын , ирегелеп |
| 7 | Manas01 | жамгырга , чартылдаган чагылганга үңдөш күүлүп , зарылдыгы мени таң калтырып ; асмандан | кыргыз | сөзү улам жаңыланып , кайрадан жаралып жаап жатты . Эркек да cop_ao |
| 8 | Manas01 | жаай берди . Уялганымды айтпа , жаны артыгы эле мен беле qst ... Асмандын | кыргыз | тагдырына болгон сынагына , элдин эркине туруштук бере алба ... Ушы |
| 9 | Manas01 | жаздырып калтырган манасчылар анча көп эмес . Алардын ичинен эң көрүнүктүүсү cop_aor_p3_sg , | кыргыздын | улуу инсаны - залкар манасчы Саякбай Каралаев cop_aor_p3_sg ( 1894 - |
| 10 | Manas01 | манасчы Саякбай Каралаев cop_aor_p3_sg ( 1894 - 1971 ) . Аның варианты жөнүндө | кыргыздын | чоң окумуштуу уулу Б . М . Юнусалиев өз учурунда : « |
| 11 | Manas01 | « Манасы » да , негизинен cop_aor_p3_sg , жалпы кабыл алынган салт боюнча байыркы | кыргыз | элинин кылымдарды кыдырып басып өткөн турмуш жолун , тагдырына |
| 12 | Manas01 | жортуулдары , акыры чыккынчылыктын курманы болушу менен бүтөт . Ушул туруктуу схема | кыргыздардан | башка элдердин эпосторунан да cop_aor_p3_sg mod_tru орун алган . Ал |
| 13 | Manas01 | Булардын эң кенжеси Жакып Алтайга айдалып барып кен казып , кырк үйлүү | кыргызын | турмуш шарты оңолуп кетет , бирок Жакып балалуу post болбойт . Ошо |
| 14 | Manas01 | Жакып балалуу post болбойт . Ошол учурда кытайдын каны Эсенканга сыйкырчылары келип , | кыргызда | Манас аттуу уул төрөлөт , ал аман чоюйсо Бээжинди соо койбой турган |
| 15 | Manas01 | деген шылтоо менен Эсенкан жер корутуч Кочку балбанын жети жүз колу менен | кыргыздарга | жөнөтүп , Жакыптын жаш баласын эптеп тартып алууга аракеттенет . Ог |
| 16 | Manas01 | Бул Алтайдан качалы . Ала - Тоо артык жерине , Кетели | кыргыз | элине ! ( С . К . 1 cop_aor_p3_sg . 84 ) - |
| 17 | Manas01 | идеясынын башаты ушул жерде cop_aor_p3_sg . Дөңгө менен Жолой Манасты байлап кетүүнү көздөп | кыргыздарга | кол салышат . Акбалтай , Жакып Манаска болушуп , катуу кагылышта м |

*Figure 1. 2 Search results for the word "Kyrgyz" in the Kyrgyz Corpus*

*https://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418/*

At this stage, the issue of part-of-speech tagging is considered incomplete; there are still several points of controversy due to the linguistic differences between English and Kyrgyz. To be more precise, this is usually caused by the presence or absence of certain syntactic categories in Kyrgyz that cannot be found in the structure of English. This corpus is focused on creating a national corpus of the Kyrgyz language, as well as to become a starting point for creating an English-Kyrgyz/Kyrgyz-English parallel corpus. Due to the present trends of the method, it is anticipated that the Kyrgyz corpus linguistics will develop and become more advanced, provide corpus analysis tools, increase its volume considerably. And consequently, it will be used in a wider range of applications, particularly for many other languages. It appears that this multidimensional method to language study has an even brighter future.

**Deduction on Chapter 1**

Thus, a corpus is an electronically represented, usually tagged for analysis for linguistic purposes, provided with a relatively easy-to-use search engine, a representative array of texts representing as many "variants" of the language as possible. There is no doubt that corpus linguistics has significantly advanced the field of linguistics. Nearly every branch of linguistics, including lexis, grammar, discourse, pragmatics, sociolinguistics, stylistics, register linguistics, and many more, has now confirmed to the accuracy of this theory. Its results are also accurate, insightful, and objective. In the period

of the birth of corpus linguistics, there were no issues of computerization in this area, and "researchers pointed out the possibility of neglecting linguistic variation, i.e., territorial, social, professional, age, age, gender, individual and similar language differentiation" (Plungyan, 2006, pp. 76-77). Even theoretical linguists, who in the past would not have engaged with corpora, now see fascinating ways of this methodology may enhance their work (McEnery & Hardie, 2012). But today, by ignoring it, we deliberately limit ourselves to different frameworks when studying texts of a particular language, which calls into doubt the objectivity of this kind of research. With the advent of electronic corpora, the variety of forms of language existence has become more visible, and the possibilities of language data research have expanded. Modern linguistic corpora contain hundreds of millions of words uses, and the fact that with the help of an electronic corpus the results of sample word use can be obtained in seconds makes the task of linguists much easier. The presented typology of corpora, without claiming to be comprehensive, shows us the existing diversity of corpora of texts and allows us to be oriented in it for further scientific research.

# CHAPTER 2

# ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE PROCESSING (NLP) IN LINGUISTICS AS A FOUNDATION FOR VERB SENSE DISAMBIGUATION

## 2.1 The Beginning and Development of Artificial Intelligence (AI)

*At some stage therefore we should have to expect the machines to take control.*

Alan Turing, Intelligent Machinery, A Heretical Theory, 1951

"You cannot make a machine to think for you". It is common knowledge that is typically assumed to be routine. It will be the purpose of this chapter to question it. The majority of equipment created for industrial use is designed to do a single, particularly specific task accurately and quickly. It frequently repeats the same set of actions without ever changing them. Many people see the fact that there is genuine equipment accessible to be strong evidence in favor of the cliché mentioned above. A mathematical logician would not accept this argument because it has been demonstrated that there are theoretically potential computers that could perform tasks that are extremely similar to thinking. The replication of human intelligence functions by machines, particularly computer systems, is known as artificial intelligence (AI).

In order to be informed about the history of artificial intelligence, it is necessary to go back to previous dates in Milat. It is known that numerous ideas for humanoid robots were implemented throughout the Ancient Greek era. Even there are the myths of Mechanical men in Ancient Greek and Egyptian Mythology. Daedelus, who is thought to have governed the mythology of the wind, is one example of someone who attempted to build artificial humans. The goal of defining philosophers' systems of human mind has begun to be observed in history through the development of modern artificial intelligence. The year 1884 is crucial for artificial intelligence. On this day in history, Charles Babbage began working on a mechanical device that will display intelligent behavior. These experiments, however, convinced him that he would not be able to build a computer that would behave as intelligently as a human being, and he decided to put his work on hold.

Claude Shannon proposed that computers could play chess in 1950. Until the early 1960s, artificial intelligence research was carried out slowly (Mijwel, 2015, pp. 2-4).

The emergence of artificial intelligence officially in history dates back to 1956. At Dartmouth College, artificial intelligence was first discussed in a conference session in 1956. In his book "Stormed Search for Artificial Intelligence," Marvin Minsky predicted *that "Within a generation, the artificial intelligence modeling issue will be resolved"* (Minsky, 1956)". During this time, the first applications of artificial intelligence were released. In 1950 Claude Shannon's "Programming a Computer for Playing Chess" is the first published article on developing a chess-playing computer program (Press, 2016). The programs that are built at that time took their base on chess and logic theorems. The fact that the programs written during this time could be separated from the geometric structures utilized in intelligence tests gave a rise to the theory that intelligent computers could be built (Specialists, 2004).

The history of AI research has its theoretical foundations in the *Turing machine* (Turing, 1937; 1996), an idealized representation of a computing device that is capable of carrying out any specified set of instruction. The major work on AI was Computing Machines and Intelligence was published by Alan Turing. Turing, who gained fame during World War II for cracking the Nazi ENIGMA code, proposes this paper to address the question *of "Can machines think?"* and introduces *the Imitation Game* which later become known as *Turing Test* to determine whether a computer can exhibit the same intelligence (or the outcomes of the same intelligence) as a human. Since then, people have argued over the Turing test's usefulness (Turing, 1950). In the beginning of 1951, Marvin Minsky and Dean Edmunds built the first artificial *neural network* called *SNARC (Stochastic Neural Analog Reinforcement Calculator)* using 3000 vacuum tubes to simulate a network of 40 neurons. Later researchers (Rosenblatt, 1957; Minsky, 1960) (Quillian, 1969) tried to create computational models of mental processes based on this work. In the late 1950s, Isaac Asimov published his *Three Law of Robotics.* In order to develop the necessary competence to compete with a world champion, IBM's Arthur Samuel created the first game-playing program for checkers (draughts) between 1952 and 1962. The checkers player's impressive performance was a result of Samuel's *machine*

*learning programs*. Five years later, namely, in 1956 Allen Newell, Cliff Shaw, and Herbert Simon's *Logic Theorist* provided the original proof for theory of Artificial Intelligence. The Logic Theorist was a computer program developed by the Research and Development (RAND) Corporation to imitate human problem-solving abilities. It was presented at the *Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI),* which was organized by John McCarthy and Marvin Minsky. It is widely regarded as the first artificial intelligence program. McCarthy, who invented the phrase *"artificial intelligence"* at the very conference, brought together leading experts from a variety of fields for an open-ended discussion on the topic during this historic conference. This workshop conference, which took place in July and August 1956, is generally considered as the official birthdate of the new artificial intelligence field. Everyone whole-heartedly started to collaborate with the sentiment that AI was achievable. The significance of this event cannot be undermined, because it catalyzed the next twenty years of AI research. At that time high-level computer languages such as *FORTRAN, LISP, or COBOL* were invented (McGuire et al, 2006).

AI flourished between 1957 and 1974. Computers improved in speed, affordability, and accessibility while being able to store more data. Additionally, machine learning algorithms developed, and individuals became more skilled at determining which method to use for a given situation. Early experiments in problem solving and spoken language interpretation, such as *Joseph Weizenbaum's ELIZA* and *the General Problem Solver by Newell and Simon*, both looked promising. These accomplishments convinced governmental organizations like the Defense Advanced Research Projects Agency (DARPA) to support AI research at a number of institutions, in addition to the advocacy of top researchers namely the DSRPAI attendees (Dartmouth Summer Research Project on Artificial Intelligence). The government was particularly interested in a device that could process enormous volumes of data while also transcribing and translating spoken language. At this period, Researchers had underestimated the issue of *"word-sense disambiguation (WSD)"*. Identifying which sense of a word was used in a sentence was an open questionable issue in linguistics. A machine had to have some understanding of the subject matter of the statement in order to translate it. Expectations were very high, and optimism was also very strong. "*From three to eight years we will have a machine*

*which will have the general intellect of an average human being*," said Marvin Minsky to Life Magazine in 1970. Although there was a fundamental proof of concept, there was still much work to be done before *natural language processing, analytical thinking, and self-recognition* could be accomplished (Rockwell, 2018).

Two factors helped ignite AI in the 1980s: an increase in funding and the growth of the algorithmic toolbox. The *"deep learning"* methods first used by John Hopfield and David Rumelhart allowed computers to learn from experience. On the other hand, Edward Feigenbaum developed expert systems that imitated a human expert's *decision-making process*. When this was mastered for nearly all situations, the program could teach non-experts. It would ask an expert in a subject how to react in a certain situation. The employment of expert systems was widespread. As part of their *Fifth Generation Computer Project (FGCP)*, the Japanese government significantly supported expert systems and other AI-related projects from 1982 to 1990.

Ironically, AI prospered in the absence of government support and media hype. Many of the historic objectives of artificial intelligence have been accomplished by the 1990s and 2000s. Grandmaster and current *global chess champion Gary Kasparo*v lost to *IBM's Deep Blue*, a chess-playing computer program, in 1997. In this widely reported game, the current world chess champion lost to a computer for the first time, and it marked a significant advancement toward the development of artificially *intelligent decision-making programs*. The same year, Windows was updated to include *Dragon Systems' speech recognition software*. This was yet another excellent step in the right direction for the project of spoken language interpretation. Kismet, a robot created by Cynthia Breazeal in 2000 that could understand and show emotions, proved that even human emotion was fair game. Geoffrey Hinton published "Learning Multiple Layers of Representation" in 2006, which provided an overview of the concepts that led to *"multilayer neural networks that included best connections and training them to generate sensory data rather than to classify it,"* i.e., the new approaches to *deep learning* (contributors, 2022) (Naqvi, 2019) (Kelley, 2022).

A large percentage of the (original) focus of artificial intelligence research focuses what might be called linguistic intelligence and borrows heavily from an experimental

approach to psychology. As a result, this field of study was divided into two major subfields. The first was *artificial intelligence*, which aimed to create intelligent machines. The second was *computational psychology*, which sought to create digital simulations of human thought (Rescorla, 2019).

### 2.2 What is Artificial Intelligence

Artificial intelligence (AI) refers to the perception, synthesis, and inference of information produced by computers in contrast to the intelligence exhibited by humans and animals (2022). The goal of AI is to simulate human intellect using computers. The Oxford English Dictionary of Oxford University Press (2022) defines artificial intelligence as:

"the theory and development of computer systems able to carry out tasks which normally require humanly intelligence, such as visual perception, speech recognition, decision-making, and translation among languages". Taking into account the definitions given above, we can conclude that Artificial intelligence (AI) is the capacity of a digital computer or robot operated by a special computer program to carry out actions frequently performed by intelligent beings, namely, by human or animals. The term artificial intelligence (AI) was first coined by John McCarthy in 1956 when he held the first academic conference called Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI) and he became one of the founders of this field. One of his papers he defines artificial intelligence as (McCarthy, 2007, p. 2) "the art and science of creating intelligent devices, particularly clever computer programs. Although it is related to the job of utilizing computers to comprehend human intellect, AI should not be limited to techniques that can be observed by biological means". This term is closely associated with the effort to create artificial intelligence (AI) systems that possess human-like cognitive abilities like the capacity for reasoning, meaning-finding, generalization, and experience-based learning. It has been proven that computers can be programmed to perform extremely complicated tasks—like, for example, finding proofs for mathematical theorems or playing chess with remarkable proficiency ever since the development of the digital computer in the 1940s. Nevertheless, despite ongoing improvements in computer processing speed and memory space, there are currently no programs that can match

human adaptability across a larger range of activities or those involving a substantial amount of background knowledge. On the other hand, certain programs are capable of doing specialized jobs at levels comparable to those of experts and professionals who are human in a variety of applications, including voice or handwriting recognition, computer search engines, and medical diagnosis. (Copeland, 2022). And then what is Intelligence itself? Intelligence or state of being intelligent, most of the time is ascribed to human being's behavior. Even the most complex insect behavior is never interpreted as a sign of intelligence, while the most basic human behavior is attributed to intelligence. What is the distinction? Take the digging wasp, Sphex ichneumoneus, as one example. When the female wasp brings food back to her burrow, she first places it on the threshold, looks inside for intruders, and only then, if everything is well, brings her food inside. If the food is moved a few inches from the burrow entrance while the wasp is inside, the true nature of her innate behavior is revealed: upon her exit, she will repeat the same process every time the food is moved. Intelligence, which Sphex obviously lacks, is the ability to adapt to new situations in the environment, must be included to it (Tirri & Nokelainen, 2011).

Psychologists (Sternberg, 2012) typically don't define human intelligence in terms of a single characteristic but rather a composite of several different skills. The five components of intelligence—*learning, reasoning, problem solving, perception, and language use*—have received the majority of attention in AI research.

### 2.3 The Role of Artificial Intelligence in Linguistics

The study of artificial intelligence (AI) has been strongly connected to linguistics from its beginning. Chomsky's (1957) theory of universal grammar offered one of the first gateways for cross-disciplinary AI research. Additionally, the development of a computer intelligence capable of producing natural speech has been a key emphasis of AI research and development. It shouldn't be surprising that mathematicians and computer scientists dealing with artificial intelligence (AI) have been interested in linguistics from the very beginning. Initially, the relationship was one-sided. Theoretical or actual uses of AI were not particularly useful to linguists (Rosenberg, 1975). It is time to reconsider what this field of study means for linguists and language teachers since that research is

moving ahead quickly and AI has already seen broad and far-reaching practical applications in all areas of our life, particularly in linguistics.

Artificial intelligence (AI) lies at the cusp of man and machine. Thus, the ultimate goal of AI is to imitate human intellect using computers. And it makes use of the machine intelligence by applying ideas from human intellect to produce the required "human like intelligence" in machines. Consequently, we can claim that the foundation of artificial intelligence lies in Natural Language Processing (NLP), a science that combines machine language and human or natural language into practical, value-adding algorithms. Thus, a basis for AI testing can be created by having a deep understanding of computer languages, data, and everyday human languages.

### 2.4 Natural Language processing (NLP)

ELIZA is an early example of natural language processing computer software developed by Joseph Weizenbaum at the MIT (Massachusetts Institute of Technology) Artificial Intelligence Laboratory between 1964 and 1966. Eliza is a mock Rogerian psychotherapist. "Scripts" written originally in MAD-Slip (is a list processing computer language invented by Joseph Weizenbaum in 1960s) offered instructions on how to communicate, allowing ELIZA to process user inputs and engage in dialogue while adhering to the script's rules and guidelines. The most renowned script, DOCTOR, imitated a Rogerian psychotherapist (specifically, Carl Rogers, who was well known psychotherapist for merely repeating back to patients what they had just said) and employed scripted rules to answer to user inputs with non-directional questions. As a result, ELIZA was one of the earliest chatterbots and one of the first programs to take the Turing Test (The imitation game, introduced by Alan Turing in 1950, measures a machine's ability to exhibit intelligent behavior that is similar to or impossible to differentiate from human behavior) (contributors, 2022). Consequently, Eliza was developed to demonstrate and to check the effectiveness of human and machine communication by simulating conversation using a "pattern matching" and replacement approach. The program's overall approach is pretty straightforward; the text is read and examined for the presence of *keywords*. If such a term is identified, the phrase is converted using a keyword-related rule. If not a content-free remark, or specific conditions are

observed, a prior transformation process is done. The text that has been processed, and computed is then printed out for further analysis. As Weizenbaum points out (1976, pp. 8-9), this is one of the few dialogue genres in which listeners may pretend they know nothing about the world. Eliza's imitation of human dialogue was astonishingly successful: many individuals who engaged with *ELIZA* began to feel that it truly understood them and their issues, and many people remained to believe in *ELIZA's* skills even after the program's function was described to them.

Thus, current conversational agents, chatbots can do much more than entertain; they can answer questions of clients, book flights, and find restaurants, etc., all of which need a far more sophisticated comprehension of the user's purpose. Nonetheless, the basic pattern-based algorithms used to power *ELIZA*, other software programs and chatbots play an important role in natural language processing.

Natural Language processing (NLP) emerged in the 1950s as an intersection of artificial intelligence (AI) and linguistics. Natural language processing (NLP) (Nadkarni et al, 2011, pp. 544-545) is a subfield of linguistics, computer science, and artificial intelligence concerned with computer and human language interactions, specifically how to train computers to process and evaluate huge volumes of natural language data. As long as it is a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone. Nevertheless, we present the following definition; "Natural language processing (NLP) is the computational study of linguistic data, most frequently in the form of textual data like papers or publications." (Verspoor & Cohen, 2013). Natural language processing tries to design a representation of the text that adds structure to unstructured natural language by employing linguistic understanding and insights of linguistics. This structure might be syntactic in type, capturing the grammatical links between the text's parts, or more semantic in nature, capturing the meaning expressed by the text. The objective is to create a computer that can "understand" the human language, and contents of papers, including the contextual complexities of the language contained within them. The system can then extract information and insights from the papers, as well as categorize and arrange the documents themselves according to user's needs.

## 2.5 Natural Language Processing (NLP) in Linguistics

We must first consider words as the building blocks of language before discussing natural language processing (NLP). Moreover, words don't just emerge. Any specific text we investigate is made up of one or more words that were created by one or more specific speakers or writers in a specific dialect of a specific language, at a specific time, in a specific location, and for a specific purpose. The language is arguably the most significant aspect of word variety. Additionally, NLP algorithms work best when used in a wide range of languages. The online Ethnologue database states that there are currently *7157 languages* spoken around the world. (Eberhard et al, 2022). These human languages are incredibly diverse and complex and unique in their own ways. We have countless ways to express ourselves verbally and in writing. There are many different languages and dialects and regional accents, and each language has its own collection of terminology, grammar and syntax rules, and colloquial words. When we write, we frequently misspell, shorten, sometimes abbreviate words or omit punctuation. But we talk, mumble, stutter, and use words from other languages. Moreover, a language is dynamic and has a tendency to use variations depending on social developments(changes) and experiences that it has been exposed. And NLP contends with the help of an empiricist approach, namely, by choosing a proper general language model and then applying statistical, pattern recognition, and machine learning (ML) techniques to a significant quantity of language use, we can learn the complex and expansive structure of language. As challenging as it may sound, NLP uses machine learning capabilities to recognize and learn all of the above-mentioned aspects of human language. Computers can read text, listen to speech, analyze conversations, assess sentiments, and identify key points thanks to machine learning's ability to extract data and this is the integration of AI and deep learning (DL). Despite the fact that *deep learning (DL)*, supervised learning, and *other machine learning applications* are used more frequently to imitate human language, it has been observed that these technologies struggle to comprehend the semantic structure of human language. NLP, on the other hand, has shown that it is capable of understanding and analyzing the deep nature of human language systems (Kassaye, 2022). In general, NLP activities split

language into smaller, more basic parts, attempt to comprehend links between the parts, and investigate how the parts combine to form meaning.

Since it may be used for language understanding, translation, and invention, NLP has practically endless applications. Chatbots, which can understand questions submitted to them by clients in normal language, are a very real example of this. These chatbots can determine the purpose and significance of a customer's request and generate spontaneous responses based on the available data. Despite the fact that they are now only implemented as a first line of defense, it shows how deep learning and NLP have extremely real-world applications. Below are some examples of NLP's more typical applications in linguistics (Team, 2022), (Rosenberg, 1975, pp. 380-388), (Donges, 2022), (Paris et al, 1991).

1. *Language Translation*

   It must be obvious saying that translating speech and writing into another language is a very difficult task. Each language has its own distinct word patterns and grammatical structures. Word for word translation of writings or speech frequently fails because it might alter the underlying style and meaning. Natural language Processing (NLP) allows for the translation of words and phrases into other languages while maintaining the original meaning. These days, Google Translate is driven by Google Neural Machine Translation, which uses machine learning and natural language processing algorithms to recognize various linguistic patterns. Additionally, for more precise specialized translation, machine translation systems are trained to comprehend terms relating to a particular field, such as law, finance, or medical.

2. *Grammar Checking*

   Whenever we use a computer to type something, we occasionally misspell, mistype, or even forget to include a word in an email or report. We are notified when we made a mistake by red or blue underlines thanks to one of NLP systems' components. Errors in spelling and grammar are found and highlighted by automatic grammar checking. Grammarly, which uses NLP to provide spelling and grammar checking, is one particularly well-known example of this.

3. *Part-of-Speech tagging*

Another important element of NLP is part-of-speech tagging, which labels each word in a text with the correct part of speech (noun, verb, adjective, or adverb). It is helpful for identifying correlations between words and particular linguistic patterns. Given that most words might contain multiple parts of speech, this task is more difficult than it first appears. For instance, depending on the context, "rain" can be both a noun and a verb. Here are a few instances of typical part-of-speech tagging methods:

*Rule-Based Method*: If a word, like *"station"* or *"worker,"* ends in *"ion"* or *"er,"* it must be a noun. If the word ends in *"ed"* or *"ing,"* then it should be an adjective.

*Stochastic Method*: It generates POS tags based on how frequent and frequently a specific tag sequence happened.

Understanding the appropriate grammar structure improves comprehension of sentence meaning and connotation.

4. *Automatic text condensing and summarization*

Automatic text condensing and summarizing techniques condense a text's size to produce a shorter version. They maintain important details and eliminate some words or phrases that are either meaningless or do not include details necessary for comprehension of the text. In order to summarize massive amounts of digital text and produce summaries and synopses for indexes, research databases, or busy readers who don't have time to read the complete text, text summarizing uses NLP techniques. The finest text summary software uses natural language generation (NLG) and semantic reasoning to provide summaries relevant context and conclusions. When making news digests or news bulletins and coming up with headlines, this use of natural language processing is helpful.

5. *Syntactic and Semantic Analysis*

The two main methods for interpreting natural language are syntactic analysis (syntax) and semantic analysis (semantic). Language is a collection of valid sentences, but what exactly qualifies as a sentence? Semantics or syntax?

Semantics refers to the meaning being communicated, whereas syntax refers to the grammatical structure of the text. However, a sentence that is semantically accurate may not always be syntactically correct. For instance, the sentence *"cows flow supremely"* is grammatically correct (subject – verb – adverb), but it makes no sense.

Another example of syntactic analysis is parsing. A sentence is parsed when a computer formally breaks it down into its component parts. This process creates a parse tree, which may be used to further process and comprehend the syntactic relationships between the sentence's constituent parts. The parse tree for the phrase "*The thief robbed the flat"* may be found below. The three separate content types that the statement conveys are described:

1. Parts of speech

   *N=noun*

   *V=verb*

   *Det=determiner*

2. Phrases

   *Noun Phrases: "the thief", "the apartment";*

   *Verb Phrases: "robbed the apartment";*

   *Sentence: "The thief robbed the apartment".*

3. Relationships

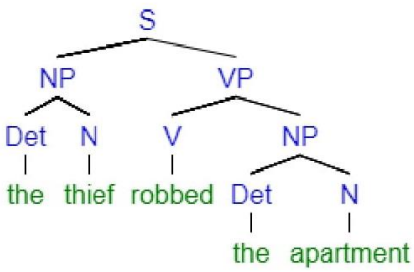



Figure 1. 3 Syntactic and Semantic Analysis

*6. Stemming*

Stemming is a pre-processing and efficiency method used in natural language processing that derives from morphology and information retrieval.

Stemming, in its simplest form, is the reduction of words to their word stem. The part of a word that remains after all affixes have been removed is known as the *"stem"*. For instance, *"touch"* is the root of the word *"touched"*. Additionally, *"touch"* is the root of *"touching,"* and so forth.

You might be wondering why we even need the stem. The stem is necessary since we'll come across different word variations that actually have the same stem and

signify the same thing. For instance, let's take a look at the following two sentences:

a) *I was taking a ride in the car.*
b) *I was riding in the car.*

The meaning of the word used in these two statements is the same in both.

Imagine all of the English terms in the lexicon now having each of their many fixes added to them. It would take a sizable database filled with several terms that, in reality, have the same meaning to store them all. By concentrating solely on a word's stem, this is resolved. *The Porter stemming algorithm* (Willett, 2006), developed in 1979, is a widely used stemming algorithm.

7. *Text Segmentation*

   Text segmentation is the process of breaking down text into understandable parts such as words, phrases, various subjects, the underlying intent, and more in natural language processing. Most of the time, the text is divided into its individual words, which, depending on the language, can be a challenging operation. Again, this is a result of how complicated human language is. For instance, using spaces to separate words in English often works well—apart from when terms like *"icebox"* should be combined instead of being separated by a space. The issue is that it is occasionally misspelled as *"ice-box."*

8. *Named Entity Recognition*

   The goal of named entity recognition (NER) is to identify the things in a text that may be located and put into one of several predefined categories. These categories might include everything from the names of people, companies, organizations and places to monetary values and percentages.

   For instance:

   Before NER: *Martin bought 300 shares of SAP in 2016.*

   After NER*: [Martin]Person bought 300 shares of [SAP]Organization in [2016] Time*.

9. *Relationship Extraction*

Relationship extraction explores the semantic relationships between the listed entities in the NER (Named Entity Recognition). This could entail learning who is married to whom, that a person works for a particular company, and so on. A machine learning model can be trained for each sort of relationship and this issue can also be rephrased as a classification problem.

10. *Sentiment Analysis*

    The goal of sentiment analysis is to automatically identify the sentiment present in text. Sentiment is an opinion communicated through language, either positively or negatively. The automatic determination of whether an online review (of a book, movie, or consumer goods) is favorable or unfavorable to the object being reviewed is one of the common uses of sentiment analysis. Today, businesses, marketers, and political analysts frequently use sentiment analysis as a tool in their toolbox for social media analysis. The study of sentiment analysis derives information from the context of positive and negative words in texts as well as from the linguistic structure of the texts.

11. ***Corpus Analysis***

    It is an approach to linguistic analysis that seeks to identify language usage patterns, such as grammatical or lexical patterns, that are crucial to a particular genre or type of text, and which can be a valuable resource for dialectology, sociolinguistics, and other related fields of linguistics. Corpora are collections of 'real-life' language samples that have been collected systematically or randomly and stored electronically. Understanding corpus and document structure through output statistics is important for activities like selecting samples wisely, getting data ready for additional models, and planning modeling strategies in research writing.

12. ***Word Sense Disambiguation (WSD)***

    Word sense disambiguation is the act of choosing a word's meaning from among its possible meanings using semantic analysis to discover which word makes the most sense in the context at hand. Word sense disambiguation, for instance, clarifies the difference between the meanings of the verbs *"make"* and *"make the grade"* (achieve) and *"make a bed"* (place).

13. *Speech Recognition*

    The process of accurately translating voice data into text is known as speech recognition, commonly referred to as speech-to-text. Any application that responds to voice commands or questions need to use speech recognition. The way individuals speak—quickly, slurring words together, with varied emphasis and intonation, in various dialects, and frequently using improper grammar—makes speech recognition particularly difficult.

14. *Natural Language Generation (NLG)*

    Natural language generation is the process of converting structured data into human language; it is frequently referred to as the opposite of voice recognition or speech-to-text. *NLG, or natural language generation*, is an artificial intelligence-driven software process that creates natural written or spoken language from both structured and unstructured data. It is beneficial for computers to communicate with users in human language that they can understand, as opposed to how a computer might.

15. *Virtual Agents and Chatbots:*

    Virtual agents like *Apple's Siri* and *Amazon's Alexa* recognize patterns in voice input using speech recognition, and they answer with the right action or a helpful remark using natural language generation. The same magic is worked by chatbots in response to text input. The greatest of these also capture contextual cues from human queries over time and employ them to offer even better options or responses. The next improvement for these apps is question answering, or the capacity to answer our questions—whether anticipated or not—in their own words in a way that is pertinent and beneficial.

16. *Spam Detection*

    One might not think of NLP as a solution for spam detection, yet the most effective spam detection algorithms search emails for language that frequently denotes spam or phishing. The overuse of financial phrases, recognizable poor grammar, aggressive language, improper urgency, misspelled company names, and other factors are examples of these indications. One of the few NLP issues

that experts believe to be “largely solved” is spam detection (although you could disagree that your email experience doesn’t match this).

**Deduction on Chapter 2**

Despite gaining popularity a decade ago, NLP is not a new field of study. Since the 1950s, it has attracted the interest of numerous scholars. In the upcoming years, it will have even more opportunities as it has developed into a necessary tool (Mah, Skalna, & Muzam, 2022, pp. 4-5). Although there are many real-world uses for natural language processing, teaching machines to understand natural language and produce original text is a real challenge. The majority of human languages follow a set of rules, but they also frequently deviate from these rules or make exceptions to them. Additionally, there may be significance in omission, a secondary context that alters the text’s entire meaning, and intentional ambiguity. Because of everything said above, teaching machines to understand natural language directly is extremely challenging and time-consuming process for humans. Instead, using incredibly big datasets and fast processors, *Deep Learning* (DL) gives machines the ability to extract rules and meaning from text on their own. As a result, *Deep Learning (DL), Artificial Intelligence (AI) and Natural Language Processing (NLP)* have a wide range of useful applications, such as chatbots, translation tools, and text production, etc.

# CHAPTER 3

# RESOLUTION OF AMBIGUITY IN NATURAL LANGUAGE PROCESSING (NLP). VERB SENSE DISAMBIGUATION

*"For a large class of cases-though not for all-in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language."*

*(*Wittgenstein, *Philosophical Investigations, 1968, 943)*

## 3.1 Ambiguity in a Language. Why NLP is difficult?

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 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 language, for instance. Similarly, language transcends purely linguistic boundaries, touching other disciplines such as sociolinguistics, psycholinguistics, neurolinguistics, theoretical/applied linguistics, cognitive linguistics, geographical linguistics, and others. In this respect, language technologies based on Natural Language Processing (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. These issues are related to the way how statistical NLP handles semantics. Majority of early researches in statistical natural language processing has focused on lower levels of grammatical processing, and some have doubted whether this natural language processing statistical

approaches can ever deal with meaning. But defining "meaning" is the fundamental challenge in providing an answer to this issue. Therefore, resolving *ambiguity* is one of the key goals while creating any NLP system. It is a universally recognized and demonstratable fact that many of the phrases in the languages are ambiguous. They can be interpreted in two or more different ways (Lyons, 1977). As a result, each kind of uncertainty or *ambiguity* of words necessitates a unique resolution process (Agirre & Rigau, 1996). First, let's consider what is ambiguity at all. A word, term, notation, sign symbol, phrase, sentence, or any other form used for communication is said to be *ambiguous* if it can be understood in more than one way. This is pronounced as /ˌæmbɪˈgjuːəti/. However, *ambiguity* is context-dependent, meaning of the same word, phrase, or even a whole sentence may be ambiguous in one context and pretty obvious (unambiguous) in another (Khamidi, 2009). A Linguist David Crystal (1988, p. 15) defines the word "*ambiguity"* in the following way: "*A word or sentence that expresses more than one meaning is referred to as being ambiguous, and this concept is related to language usage*." Hartmann and Stork offer a different definition *of ambiguity*, stating that it is a construction that provides for several interpretations. In the phrase *"Patent medicines are sold by frightening people*," for example, it is unclear whether the *meaning* is "*Patent medicines are sold by people who are frightening*" or "*Patent medicines are sold by people who induce fear into people*". Philosophically, this takes us on the threshold of Wittgenstein's (1968) position, which holds that *a word's meaning is determined by the context in which it is used*, (*A use theory of MEANING*) - see the quote at the beginning of the chapter. John Rupert Firth (1957) also maintaining the same opinion proposed his own theory of "context of situation." He highlights the context-dependent aspect of meaning, Firth believed that language should not be investigated as a mental system. Instead, he asserted that language represented a series of events that speakers uttered—an activity that one acquires via doing things - in the positivist and behaviorist ways. Any statement made by someone, in his opinion, must be interpreted in the context of the surrounding circumstance s (situations). He became well-known for this approach and his theory of "context of situation" became one of the central concepts in linguistics. Therefore, Firth advises breaking down meaning into a number of component functions like *phonetic, lexical, morphological, syntactic, and situational.*

Thus, semantic study becomes the place where the phonetician, grammarian, and lexicographer cooperate and integrate their work. Consequently, progress in the study of meaning is only possible with the amalgamation of phonetics, morphology, and syntax. Firth stresses how meaning is addressed by descriptive or structural linguistics at all levels of analysis and throughout the descriptive range. We first accept language occurrences as integral in experience, whole, repetitive, and interconnected, and then we apply theoretical schemata and make claims in terms of structures and systems at various levels of analysis and carry out disambiguation task within a set linguistic inquiry (Firth, 1935).

Ambiguity is typically a quality shared by signs in a language or, more generally, a system of signs that have various *(legitimate)* interpretations. The word *"legitimate"* is used to acknowledge the fact that many signs, in theory, can sustain any more than one interpretation. The term *"ambiguity"* is mentioned a lot in common language; frequently, *underspecificity* alone qualifies as ambiguity.

However, theorists have realized that it is useful to distinguish the phenomenon of ambiguity from other phenomena in many fields (e.g., *underspecification, vagueness, context sensitivity*). Philosophers are attracted by ambiguity for a variety of reasons, some of which we will examine here:

a) *Ambiguity* brings to light some of the distinctions between formal and natural languages and places demands on the use of the former to represent the latter (Stokhof, 2007).

b) Due to potential equivocation, *ambiguity* might negatively affect our capacity to judge the validity of arguments in natural language.

c) By opposing to easy categorization and interpretation, *ambiguity* in art can intentionally (or accidentally) boost the interest in a piece of art.

d) *Ambiguity* in the statement of the law can undermine their applicability and our ability to obey them.

e) Finally, the ability to resolve *ambiguity (disambiguation)* is a key component of cognitive comprehension and interpretation. We can gain understanding of

mind and interpretation by researching ambiguity and how we deal with it in real-world situations (Sennet, 2021).

Philosophers have been interested with ambiguity for a very, very long time. In *Aristotle's Sophistical Refutations*, fallacies were investigated in relation to this topic. At the end, these variations of fallacies connected to ambiguity and amphiboly[1] writing are identified by Aristotle (1984).

These ambiguities and amphibolies are available in three different varieties:

a) When the name or one of the expressions has clearly more than one meaning...

b) When we use them in this way out of habit;

c) When words that have only one meaning when used alone have multiple meanings when combined, such as *"knowing letters."* For each term, *"knowing"* and *"letters,"* there may be just one meaning; yet, when used together, they might signify either that the letters themselves possess knowledge or that someone else does.

Ambiguity also intrigued the Stoics (Atherton, 1993). Chrysippus once asserted that every phrase in a language is ambiguous; nevertheless, by this he meant that a single person could interpret a word delivered to him in a variety of different ways. The question of whether the language in which we think might contain ambiguous terms attracted attention of philosophers who were interested in the relationship between language and thought, particularly those who advocated a language of thought. For instance, Ockham was willing to accept ambiguity in mental sentences of a language of thought but not in those language's mental terms (Spade, 1996, p. 101). Later, in a well-known footnote, Frege discussed the role of a sense in natural language, writing (1948, p. 210):

[1] Amphiboly arises when a statement's language allows multiple possible interpretations. For instance: "The governor says, "Save soap and waste paper." So, soap is more valuable than paper (Schagrin & Rescher, 2021)."

*As long as the reference remains the same, such variations in meaning are permitted, but they must be avoided in the theoretical framework of a demonstrative science and shouldn't occur in a perfect language.*

Despite his passing, Frege's disdain for ambiguity continues. In order to clarify potentially ambiguous sentences, we generally employ formal languages. For instance, brackets being a paradigm example of a disambiguating device. In the long run, ambiguity is the ability to have more than one meaning or to be comprehended in more than one manner. Because natural languages are ambiguous, computers cannot grasp language the same way that humans can. The field of natural language processing (NLP) focuses on the creation of computer models for various language tasks.

### 3.2 What (Linguistic) Ambiguity Is Not?

Language philosophers and linguists use the term "ambiguity" to describe a phenomenon that is more particular than the existence of numerous acceptable interpretations. Distinguishing ambiguity from these related phenomena can be a challenging and confusing process. Below, we'll discuss testing for ambiguity, but for now, let's attempt to distinguish ambiguity by isolating it from other common situations that ambiguity is frequently conflated with (Sennet, 2021).

#### 3.2.1 Vagueness

Although it is notoriously (and paradoxically) difficult to define vagueness, it appears to result from a lack of clarity in the meaning or reference of a term or phrase. There are words that are ambiguous but not (apparently) vague, such as *"bat", (a)he wooden baseball bat*, b) *the bird*) which is clear that it is an ambiguous not vague. "*Is bald*" seems to be vague and requires more information to get the ultimate meaning (who is bald?), but not ambiguous. In order to adopt linguistic practice guidelines digitally, it is crucial to eliminate ambiguity and vagueness. Understanding the qualities of ambiguity and vagueness is necessary for successful resolution, but these concepts have not still been distinguished, categorized, or described in the context of language.

It's interesting to note that certain viewpoints on ambiguous language perceive ambiguity as at least equivalent to vagueness. While supervaluationism[2] views ambiguous phrases as conveying several distinct semantic values, Braun and Sider (2007) treats ambiguous sentences as expressing numerous distinct ideas. However, the fundamental idea of numerous expressions seems to differ from paradigmatic ambiguity, where two meanings of a term or phrase are unquestionably appropriate methods to make the term more explicit, rather than where several meanings are. One may even argue that these viewpoints treat ambiguity as a form of polysemy.

### 3.2.2 Context Sensitivity

Context sensitivity is the (potential) variation in content brought on by changes in the utterance's context alone, without changes in the word usage. The meaning of the phrase "*I am hungry*" varies depending on the speaker, because *"I"* is context-sensitive and changes its reference depending on who says it. The word *"I"* is not particularly ambiguous, on the other hand; rather, the puzzle around context-sensitive phrases has been how they might have a single meaning while referring to multiple things. The word "bank" is ambiguous and not obviously context-sensitive. Of course, contextual awareness can assist clarify an unclear phrase. However, ambiguity is a quality of the terms' meanings and is not characterized by how they interact with (extra-linguistic) context.

### 3.2.3 Under-specification and Generality

I have sisters living in Kingston, New York, and Toronto. If I only tell you that I'm going to see one of my sisters, you won't know which one I'll be seeing. If you're trying to figure out where I'm heading, this can be difficult. This, however, is not the result of the ambiguity in the phrase "one of my sisters". Its intent is obvious. The sentence is "sense-general"; it just underspecifies which sister I am going to see. In general, generality and underdetermination can leave a wide range of choices accessible without creating any ambiguity. One more terminological point: What we refer to as *"sense generality" is* often treated as *vagueness* in the cognitive linguistics literature (Dunbar,

[2] Supervaluationism is a semantics which deals with irreferential singular terms and vagueness (2022)

2001) a single lexeme with a single meaning that is nonspecific with respect to certain properties.

Considering how frequently the extension of a univocal term can split up into two or more distinct salient categories, it is simple to confuse sense generality for ambiguity. The sentence 'I ordered filet mignon' doesn't specify whether or not the filet was to be given to me cooked or raw. If the waiter presents the filet uncooked, you will undoubtedly be upset and say, "That's not what I meant," but not in the butcher shop. It might be challenging to determine when a difference in extension corresponds to a discrepancy in the term's meaning. However, we shouldn't abandon the distinction because it can be challenging to discern things apart in some instances.

### 3.2.4 Sense and Reference Transfer

Transference of sense or reference is one complex phenomenon (Nunberg G. D., 1978; 1995). You probably manage to refer to the car rather than yourself when you remark, "*I am parked on G St*." The phrase *"I am traditionally allowed a final meal"* said by a prisoner also has nothing to do with him or her (there are no traditions regarding him). The mechanisms of reference transfer are unclear, and there is some debate regarding how transferred terms interact with the syntax.

Naturally, sentences might possess several of these characteristics at once. "*My uncle asks if I am parked where the bank begins"* is sense-general, ambiguous, context-sensitive, vague and it contains reference-transfer. In spite of this, it is essential to differentiate between these features since the semantic treatment we give each one can vary greatly, testing for them can necessitate very specialized considerations, and their sources can vary greatly from phenomenon to phenomenon.

## 3.3 Types of Ambiguity

There are different sources and types of ambiguities. Linguistic theories have identified the following main types of ambiguity (Anjali & Babu Anto, 2014, pp. 2-4):

### 3.3.1 Lexical Ambiguity

As we have discussed earlier, it has been determined that something is ambiguous if there are two or more possible interpretations. If a ambiguity is detected in a single word, it is called as

*lexical ambiguity*. Lexical ambiguity occurs when the provided context is not enough to differentiate between two meanings of a single word. There are many instances of lexical ambiguity (Igiri, 2017, p. 8); in fact, practically every word has many meanings. Consider the word *"ambiguity"* itself. It can indicate uncertainty to what you mean; the desire to convey several meanings; the probability that one or both of two meanings were intended; and the reality that a statement has various definitions.

According to Fromkin (2003, p. 122), lexical ambiguity occurs when at least one word in a phrase has more than one meaning. *"This will make you smart,"* for example. Due to the term smart's dual meanings of *"intelligent"* and *"burning feeling,"* it is confusing. A word can be ambiguous with respect to its syntactic structure. E.g.: The word "silver" can be used as *a noun, an adjective,* or *a verb.*

a. She bagged two *silver* medals.

b. She made a *silver* speech.

c. His worries had *silvered* his hair.

Lexical ambiguity can be resolved by lexical category disambiguation i.e., parts-of-speech (POS) tagging. As majority of words may belong to more than one lexical category. P*art-of-speech tagging i*s the process of assigning a part-of-speech or lexical category such as a noun, verb, pronoun, preposition, adverb, adjective etc. to each word in a sentence.

### 3.3.2 Lexical Semantic Ambiguity

The type of lexical ambiguity, which occurs when a single word is associated with multiple senses. E.g.*: bank, pen, fast, bat, cricket etc.* For example, take a look at the following sentences:

The *tank* (container) was full of water.

I saw a military *tank* (vehicle).

Despite the fact that both phrases contain the word tank, which belongs to the grammatical category noun, their meanings are different. Using *Word Sense Disambiguation (WSD*) techniques, lexical semantic ambiguity is resolved. WSD aspires to automatically assign the meaning of the word in the context in a computational manner.

### 3.3.3 Syntactic Ambiguity

This form of ambiguity is also called structural or grammatical ambiguity. It occurs in the sentence because the sentence structure leads to two or more possible meanings. Loebner (2013) claims that independently of lexical ambiguities, the syntactic structure of a sentence may be ambiguous. There are two kinds of syntactic ambiguity*: Scope Ambiguity* and *Attachment Ambiguity.*

**3.3.2.1 Scope Ambiguity.** Scope ambiguity involves operators and quantifiers. Consider these two examples:

*Old men and women were taken to safe locations.*

The scope of the adjective (i.e., the amount of text it qualifies) is ambiguous, that is, whether the structure (*old men and women) or ((old men) and women).*

The scope of quantifiers is often not clear, and consequently, creates ambiguity.

*Every man loves a woman.*

The interpretations can be, *for every man there is a woman* and also it can be there is *one particular woman who is loved by every man.*

#### 3.3.2.2 Attachment Ambiguity

If a constituent can fit in more than one position in a parse tree, then the sentence has attachment ambiguity. *Attachment ambiguity* emerges uncertainty about which part of a sentence to attach a phrase or clause to (Jurafsky & Martin, 2019, p. 233). Let's consider the following example: *I saw the man with the telescope.*

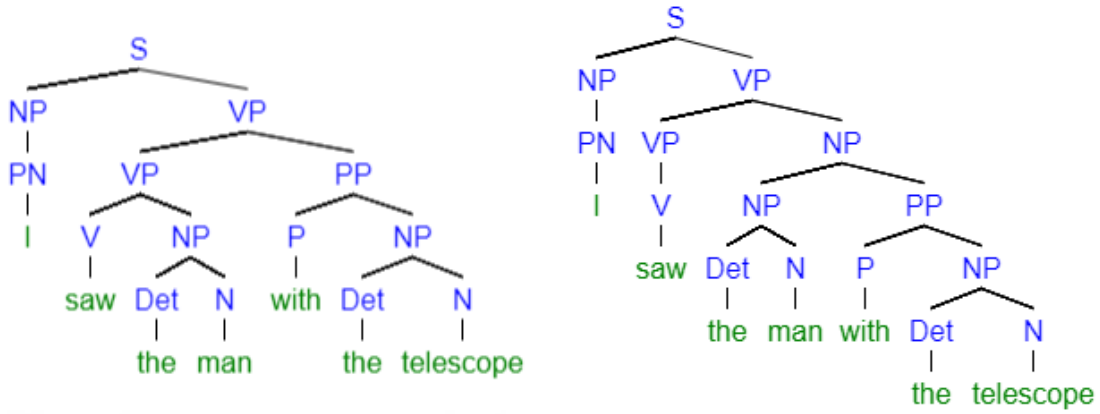



Figure 3. 4 Parse trees for an ambiguous sentence

This sentence is ambiguous, because it is unclear whether the man saw a girl who was carrying a telescope, or he saw a girl using or with help of his telescope. The meaning is dependent on whether the preposition *'with'* is *attached* to the I or the man.

The first phrase structure tree represents – The speaker used a telescope to see the man. The key element is the position of the PP directly under the VP. Though the PP is under VP, it is not selected by the verb because it is not a complement. The tree selects an NP only. In the sentence, the PP has an adverbial function and modifies the verb. In other meaning: "The boy saw a man using (the PP) - with the help of the telescope". In the second tree, The PP (with the telescope) occurs under the direct object NP, where it modifies the noun man, consequently, in this case it means that the speaker saw the man who was holding or carrying a telescope. The complement of the verb see is the entire NP - the man with the telescope. The PP in the first structure is generated by the rule: VP V NP PP. In the second structure the PP is generated by the rule: NP Det N PP.

### 3.3.3 Semantic Ambiguity

This happens when the words themselves have ambiguous meanings. There are two ways to read the sentence, even after the syntax and the meanings of the individual words have been defined. Look at this sentence:

*Seema loves her mother and Sriya does too.* Here the two interpretations are possible: Sriya loves Seema's mother or Sriya likes her own mother.

When a sentence contains an ambiguous word or phrase, semantic ambiguity occurs. Semantic ambiguities are caused by the fact that, in general, a computer cannot identify what is logical from what is not.

### 3.3.4 Discourse Ambiguity

Discourse level processing of ambiguity requires a common reality or knowledge base, and interpretation is done in this context. There are two distinctions of this ambiguity:

**3.3.4.1 Anaphoric ambiguity:** The entity that have previously been introduced into the discourse are called anaphora. Anaphora in linguistics is about referring backwards (or an entity in another context) in a text. Let us see one sentence; *"London had snow yesterday. It fell to a depth of a meter."* In this sentence, how do we relate the pronoun *"it"* with the previous sentence? We have three *antecedents* namely *"London", "snow"* and *"yesterday".* We can relate the anaphor to either *"London", or "snow", or "yesterday".* It would be able to get the correct meaning if we relate the anaphor to the *antecedent "snow".*

Anaphors are words that relate to other words in the same or different phrases but have little or no meaning on their own. Anaphoric ambiguity refers to a case where an anaphor has more than one potential reference in the same or another sentence within one contextual text.

**3.3.4.2 Pragmatic Ambiguity:** Pragmatic ambiguity describes a scenario in which a sentence might have several meanings depending on its context. One of the most challenging NLP tasks. Processing user intention, sentiment, belief world, modals, etc.—all extremely difficult tasks. Consider the example: *I love you too.*

This can be interpreted as following:

1. I love you (just like you love me)
2. I love you (just like someone else does)
3. I love you (and I love someone else too)

4. I love you (as well as I like you)

Pragmatic ambiguity emerges when a statement is undefined and the context lacks the details necessary to make it clear. The absence of information necessitates further inference.

To conclude, ambiguity appears at all NLP levels. Solving this type of ambiguity is a very difficult task, especially at the higher levels of NLP. In order to interpret the meaning of a word, phrase, or sentence, complementary contextual information is necessary. At higher levels, pragmatic and global knowledge are also required. Making a world model for activities requiring disambiguation is difficult. The development of disambiguation techniques requires linguistic resources and lexical tools. When it comes to the use of these strategies, resourceless languages lag behind resourceful languages. Automatic resolution of all these ambiguities has a number of long-standing issues, but once more, we can think positively about the development of comprehensive disambiguation techniques that address all the ambiguities because they are crucial to the proper operation of NLP applications like machine translation, information retrieval, and question answering, among others.

### 3.4 A Brief History of Research on Word Sense Disambiguation (WSD)

One of the first problems in computational linguistics, namely, in Natural Language Processing (NLP) Word Sense Disambiguation (WSD) was first formulated as a separate computer work during the early years of machine translation in the 1940s. The issue was first presented in a computational setting by Warren Weaver in his renowned 1949 memorandum on translation. Early researchers had a clear understanding of the importance and complexity of WSD. In fact, Bar-Hillel (1964) utilized the above instance to support his claim that WSD could not be resolved by an "electronic computer" due to the requirement to model all available knowledge.

One of the most challenging tasks in the discipline of natural language processing research is WSD. In this area, research was first conducted (Agirre & Edmonds, 2007) 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 (Agirre & Edmonds, 2007)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. (1955) 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 (Masterman, 1957).

WSD systems were generally rule-based and hand-coded in the 1970s when they were developed as a subtask of semantic interpretation systems in the field of artificial intelligence, but this made them prone to a knowledge acquisition bottleneck. 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. (Alot Ranjan & Diganta, 2015) 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.

When the statistical revolution sailed through computational linguistics in the 1990s, WSD emerged as a paradigm problem to which supervised machine learning approaches might be applied. 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 (1998); the availability of the online lexicon WordNet (Seo et al, 2004), (Canas et al, 2003) and the introduction of statistical approaches. Because information was both programmatically available and hierarchically arranged into word senses termed

synsets, WordNet (Miller, 1991) 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. (1991) introduced corpus-based Word Sense Disambiguation for the first time in 1991.

Since supervised techniques' accuracy peaked in the 2000s, focus has switched to coarser-grained senses, domain adaptability, semi-supervised and unsupervised corpus-based systems, combinations of diverse methods, and the revival of knowledge-based systems through graph-based techniques. The best performance is still retained for supervised systems.

### 3.5 Word Sense Disambiguation (WSD)

There are many terms that indicate meanings in various situations in all of the major languages used today. A method to determine a word's precise meaning in a given situation is called Word Sense Disambiguation (WSD) (Cucerzan et al, 2002; Nameh et al, 2011; Xiaojie et al, 2009). For instance, the word *"bank"* in English can mean numerous things, such *as "financial institution," "riverside," "reservoir,"* etc. These words with many meanings are referred to as ambiguous words, and the procedure for determining an ambiguous word's precise meaning in a given context is known as word sense disambiguation. Normal people have the innate ability to distinguish between the various meanings of a word in the given context, but computers only function according to the instructions. As a result, the system is provided with various rules to carry out a specific duty. 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.

Here is a famous example, let us determine the sense of pen in the following passage (Bar-Hillel, 1964):

*Little John was looking for his toy box. Finally, he found it. The box was in the **pen**. John was very happy.*

*WordNet* lists five senses for the word *pen:*

1) pen — a writing implement with a point from which ink flows.
2) pen — an enclosure for confining livestock.
3) playpen, pen — a portable enclosure in which babies may be left to play.
4) penitentiary, pen — a correctional institution for those convicted of major crimes.
5) pen — female swan.

The level of accuracy that WSD systems accomplish across a range of word types and ambiguities is consistent thanks to the steady advancement of research in this area. There has been extensive research on a wide range of methodologies, including dictionary-based approaches that draw on the knowledge stored in lexical resources, supervised machine learning approaches that train a classifier for each unique word on a corpus of manually sense-annotated examples, and completely unsupervised approaches that group word occurrences to infer word senses. The most effective algorithms among these to emphasize are supervised learning techniques (Edmonds & Agirre, 2008).

### 3.5.1 Word (Verb) Sense Disambiguation, Approaches and Methods

Verb Sense Disambiguation 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)" (Yarowsky, 2000). 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.

There are two main approaches to WSD – *deep approaches* and *shallow approaches*.

*Deep approaches* imply having access to an extensive body of global knowledge. These approaches are typically not seen as being very effective in actual practice, mostly because, outside of extremely specific disciplines, such a body of knowledge does not exist in a computer-readable version (2022). It can be challenging to distinguish between knowledge that is linguistic or general knowledge due to the long heritage in computational linguistics of exploring such techniques in terms of coded information. Margaret Masterman and her colleagues at the Cambridge Language Research Unit in England made the initial attempt in the 1950s. This project used a punched-card Roget's Thesaurus and its numbered "heads" as data, serving as an indicator of subjects, and it searched the text for repetitions using a set intersection algorithm. Although it wasn't very effective, it had important connections to later work, particularly Yarowsky's machine learning optimization of a thesaurus method in the 1990s (contributors, 2022).

*Shallow approaches* focus more on the words around the text than on the text itself. Through the use of a training corpus of words with their word senses identified, the computer may automatically generate these rules. Due to the computer's limited understanding of the outside world, this strategy, while theoretically less powerful than deep approaches, but produces superior outcomes in practice.

There are four conventional methods to WSD:

- ✓ **Dictionary- and knowledge-based methods:** These rely primarily on dictionaries, thesauri, and lexical knowledge bases, avoid using any corpus evidence.
- ✓ **Supervised methods:** These employ sense-annotated corpora as a training resource.

- ✓ **Semi-supervised or Minimally supervised methods:** Usage of a secondary source of information, such as a word-aligned bilingual corpus or a short-annotated corpus used as seed data in a bootstrapping process.
- ✓ **Unsupervised methods:** These forsake (nearly entirely) external data in favor of working directly with unannotated raw corpora. Word sense discrimination is another name for these methods.

Nearly all of these methods operate by selecting a window of n content words surrounding each word in the corpus that needs to be disambiguated and statistically evaluating those n words. Naïve Bayes classifiers and decision trees are two simple methods that are used to train and then disambiguate. Support vector machines and other kernel-based techniques have demonstrated greater performance in supervised learning in recent studies. The research community has also given graph-based techniques a lot of attention, and they presently attain performance that is very close to the **state of the art.**

**Deduction on Chapter 3**

In a sense, WSD research has come back to the starting point, going back to empirical techniques and corpus-based analyses that are typical of some of the problem's initial attempts to be solved. Researchers in the 1990s have undoubtedly improved on prior findings with access to significant resources and improved statistical methods, but it appears that we may have reached the upper bound of what is possible within the current paradigm. Due to this, it is now more important than ever to evaluate the current state of WSD and think about potential future research areas. By placing WSD in the context of the past 50 years of research on the subject, this chapter attempts to give that theoretical background, at least in part. 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. Moreover, it is relatively new field or focus of study in Kyrgyz language circumstance. Of course, one of the reasons why WSD is challenging is that it is inherently difficult to determine or even define word sense, and this problem is likely to be solved anytime 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 chapter is to provide with a substantial basis to number of researchers working in various branches of computational linguistics, NLP and AI who want to learn more about WSD. As WSD contributes to numerous applications as we have listed above and interest in it has grown recently. Although WSD is "an intermediate problem," it is challenging and possibly hard to evaluate in general. By incorporating WSD methods into more extensive applications, we can potentially inform and improve future work.

# CHAPTER 4

# VERB SENSE DISAMBIGUATION IN THE KYRGYZ LANGUAGE (ON THE BASIS OF THE NEWLY-CREATED KYRGYZ CORPUS)

## 4.1 Overview on Kyrgyz Language

The Turkic language, ***Kyrgyz*** (written "кыргыз тили", pronounced [qırgız tili], in English: /ˈkɪərgɪzˌ kərˈgiːz/), also known as Kirghiz or "Kirgiz," is spoken in Kyrgyzstan, China, Tajikistan, and Uzbekistan. Its categorization within Turkic language family is still unclear; it seems to alternately belong to the Kypchak (Northwestern) and South Siberian (Northeastern) branches. The southern dialects of Altay are the Turkic variations that are phonetically and phonologically closest to Kyrgyz, despite the fact that Kyrgyz exhibits substantial similarities to Kazakh that these varieties do not seen, particularly in its Talas dialects. There are numerous similarities among southern Kyrgyz varieties and Uzbek that other dialects lack. Kyrgyz, Kazakh, and Altay languages bear strong resemblances between each other. (Washington et al, 2012, pp. 1-2).

Kyrgyz is mostly spoken in Kyrgyzstan, where it is the official national language. Majority of Kyrgyz speakers live in Kyrgyzstan, where the language is recognized as the official national tongue. The large number of people in Kyrgyzstan are Kyrgyz speakers who are also fluent in Russian and/or Uzbek. Outside of Kyrgyzstan, there are additional major Kyrgyz-speaking communities, most notably in China (where the Kyrgyz are an officially acknowledged minority-Kizilsu Kyrgyz Autonomous Prefecture in Xinjiang Province), and Tajikistan (Gorno-Badakhshan Autonomous Region) and in some regions of Uzbekistan. Afghanistan and Pakistan are home to speakers of the Pamiri Kyrgyz dialect of the Kyrgyz language. The number of speakers is currently estimated to be more than 6 million. There is a highly strong correlation between these ethnic groups and their linguistic proficiency, even though not all ethnic Kyrgyz are proficient speakers of the language and at the same time not all proficient speakers are ethnic Kyrgyz. Kyrgyz is also spoken by many ethnic Kyrgyz groups through the former of Soviet Union Regions, Afghanistan, Turkey, northern Pakistan, and Russia (Kara, 2003).

Initially, Kyrgyz was written in Göktürk script (Кызласов, 1994), gradually, its alphabet was replaced by the Perso-Arabic alphabet (in use until 1928 in the USSR, still in use in China). Between 1928 and 1940 a Latin-script alphabet, the Uniform Turkic Alphabet, was widely applied. In 1940, Soviet authorities changed the Latin script into the Cyrillic alphabet for all Turkic countries. When Kyrgyzstan became independent following the Soviet Union's collapse in 1991, there were suggestions to adopt the Latin alphabet and it became popular. Although Latin Alphabet has not been implemented, it continues to be discussed occasionally (Altynbaev, 2019).

### 4.2 Verbs in the Kyrgyz Language

Verbs (этиштер) are words that show an action (кыймыл-аракет) (***ырда-sing),*** state of being (ал-абалды) ***(бол-to be, become)*** or mental activity ***(сүй-love).*** Verbs answers questions like ***What are (you) doing? What did (you) do? What will (you) do?*** The term ***"этиш"*** (lit.: "verb") comes from the word ***"эт-"*** meaning ***"work (иште-)", "do (кыл-)" or "make (жаса-)".*** It was actively used in ancient times, and but nowadays, in modern Kyrgyz, it is almost never used independently, and it is only used in the system of compound verb forms such as ***"кабыл эт-", "сабыр эт-", "былк эт-", "солк эт-"***. Sometimes it is been used along with the loanwords from other languages in compound verb structure: ***"звонить эт-", "оформить эт-",*** etc.

Among the parts of speech of the Kyrgyz language, the verb word groups play an important role and has a special place. There are certain peculiarities of verbs in Kyrgyz Language in general:

- Comparing to other parts of speech, verbs make up the majority of vocabulary(lexicon) in the language.
- Kyrgyz verbs rarely transfer to another parts of speech. Consequently, they form a stable lexical-grammatical category in the language. Cases of borrowing or adopting verbs from other languages is not common in Kyrgyz language.

- The grammatical structure of verbs is complicated. Thus, its grammatical categories cause various difficulty for analysis.

Verbs vary according to the categories of *mood (ыңгай), voice (мамиле), person/possessiveness (жак), tense (чак), and number (сан),* and have forms expressing positive and negative meanings (оң жана терс маанилер). The systematic structure of verbs also is divided into *transitive or intransitive with respect to the subject/object(субъект/объект)* of the action in a sentence. (these will be discussed separately below).

After all, the spatial movement of animate and inanimate substances in nature and the time frame of that movement are unlimited. Therefore, verbs mean different types of actions, state of beings, actions that are related to mental processes in our mind. Moreover, these actions take place or occur in various time units and diverse periods.

We believe that the issue of the infinitive form of verbs not only in the Kyrgyz language, but also in Turkic languages in general has not been clarified yet. In most Turkic languages, including Kyrgyz, as the infinitive form, imperative mood stems representing the second person singular are accepted. For example, words such as ***бас-(bas-walk), же-(je-eat), сүйлө-(süylö-speak), төлө-(tölö-pay)*** are considered as verb bases (stems). In the modern Kyrgyz language, root verbs and derivative verbs (туунду) always come with the meanings of commanding and demanding. That is why it is used in the same way in the register of dictionaries.

On the other hand, in a group of Turkic languages, verb forms ending in ***-мак (-mak)*** are used as the infinitive form of verbs (for example: Uzb. *йигмок (yigmok), ялинмок (yalinmak;* Uyg. *кайтмак- (kaytmak);* Azerbaijani. *нурланмаг (nurlanmag)*; Turkish. *йазмак (yasmak), гелмек (gelmek))*. In Kyrgyz, the participle ***-мак*** serves as an indicator of a gerund (кыймыл атооч) like in the following sentence*: Кел демек бар, кет демек жок.* (Proverb).

These kind of inconsistencies among Turkic languages which occur due to their internal agglutinative nature causes controversy among researchers on this field. Techniques and tools which can be used to solve these issues and serve for modelling or

unifying Turkic languages remain open in Turkology as a matter of the future (Hakkani-Tür, Oflazer et al, 2002).

In this paper, we made a decision to write root form of verbs with hyphen at the end of each verb (***"отур-" otur-sit***) relying on Judahin's *Dictionary of Kyrgyz-Russian languages* (1965) and Junusaliev's book on *Lexicology of Kyrgyz language* (1959).

The verbs in the Kyrgyz language word differs from other parts of speech by the following features:

- In terms of meaning, it mainly indicates movement, action and a state of being.
- It has its own unique grammatical categories such as *mood (ыңгай), voice (мамиле), person/possessiveness (жак), tense (чак)*.
- It has its own suffixes (куранды мүчөлөр-сөз жасоочу мүчөлөр) when added to root of a word form a new verb. For example***:***
  - ***ла:*** *камчыла-, сүйлө-, ойло-;* ***-дан:*** *каардан-, ардан-, кубаттан-;*
  - ***а:*** *сына-, күчө-, сана-;* ***-ай:*** *азай-, көбөй-, чоңой-.*
- Syntactically, it mainly acts as a predicate of a sentence. For instance: Бирок алар өздөрүн асмандан жерге кулап түшкөндөй ***сезишти*** (Aitmatov, 1997)**.**

*Table 4.3 The Syntactic Role of the Verbs in the Kyrgyz language*

| Coordinating conjunction | Subject | Direct object | Adverbial modifier of manner | Predicate |
|---|---|---|---|---|
| ***Бирок*** | ***алар*** | ***өздөрүн*** | ***асмандан жерге кулап түшкөндөй*** | ***сезишти*** |

- It usually takes place at the end in the order of a sentence. It may also appear at the beginning of a sentence in some literary genres like poetry etc. when it is used for stylistic purposes.

### 4.2.1 Simple and Compound Verbs

In the Kyrgyz Language, verbs are divided into simple (жөнөкөй) and compound (татаал) verbs according to their structure.

Simple verbs consist of only one word and denote one lexical meaning: *бас-, тур-, сүйлө-, аткар-, учкаш-, отур-, кыймылда-, уйкусура-, бекин-, мыкчы-, ойгон-, жөлөн-* etc. Structurally, simple verbs are formed in two ways:

1. Verbs that are made up from root word (уңгу сөздөн): *ич-, айт-, бас-, ук- etc;*
2. Derivative verbs (туунду этиштер) which are constructed at the result of adding suffixes (куранды сөз жасоочу мүчөлөр) to root word: *камчы+ла-, сүй+лө-, этсир+е-, каар+дан-, кам+ын- etc.*

***1.*** *Хан макул болуп, жаш-карынын баарын* ***чакырды.***

Хан_n_nom макул_ij болуп_v_iv_prc_perf, _cm жаш_quio карынын_n_gen_sg баарын_prn_ind_px3sp_acc чакырды_v_tv_ifi_p3_sg ._sent

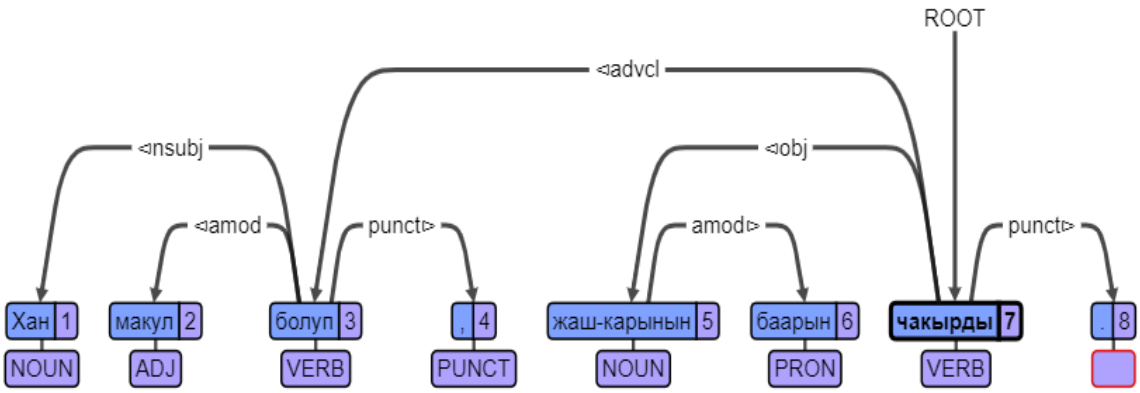



*Figure 4.1 Dependency parsing of the simple verb "чакырды"*

"<Хан>"

"хан" NOUN @nsubj #1->3

"<макул>"

"макул" ADJ @amod #2->3

"<болуп>"

"бол" VERB @advcl #3->7

"<,>"

"," PUNCT @punct #4->3

“<жаш-карынын>”

“жаш-кары” NOUN @obj #5->7

“<баарын>”

“баары” PRON @amod #6->5

“<чакырды>”

“чакыр” VERB @root #7->0

“<.>”

“.” @punct #8->7

**2.** *Коён эсин жыйып, кантип кутулуштун аргасын* ***издеди (The Story by Aldarkoso).***

Коён_n_nom эсин_n_px3sp_acc жыйып_v_tv_prc_perf ,_cm кантип_adv кутулуштун_v_tv_ger_pres_gen аргасын_n_px3sp_acc издеди_v_tv_ifi_p3_sg ._sent

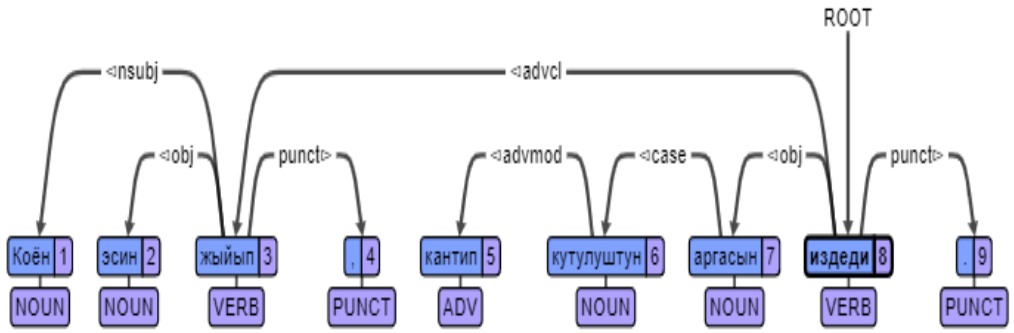



Figure 4. 2 Dependency Parsing of the Simple Verb “издеди”

”<Коён>”

“коён” NOUN @nsubj #1->3

“<эсин>”

“эс” NOUN @obj #2->3

“<жыйып>”

“жый” VERB @advcl #3->8

“<,>”

“,” PUNCT @punct #4->3

“<кантип>”

“кантип” ADV @advmod #5->6

“<кутулуштун>”

“кутулуш” NOUN @case #6->7

“<аргасын>”

“арга” NOUN @obj #7->8

“<издеди>”

“изде” VERB @root #8->0

“<.>”

“.” PUNCT @punct #9-

The simple verbs that are given above “***чакырды-invited”, “издеди-searched”*** are root verbs. They are made up from one word which denotes one lexical meaning. And grammatical category also provided by the same verbs: чакыр-***ды***, изде-***ди*** are affixes of the past tense and first person singular.

On the other hand, compound verbs consist of more than one word but represent one lexical meaning: *тосуп чык-, ойлоп тап-, чаап ташта-, коштоп жүр-, баш бак-, кол кой-, алдан май-, тарс эт-, тартып бара жат- etc.*

In a sentence, ***compound verbs*** play the same role as simple verbs do, i.e. despite the fact that they are composed of more than one word serve as one part of a sentence, namely, *predicate. Let us consider the following examples:*

1. *Кечке жуук Арсен Саманчин* ***чыдап олтура албады***. (Aitmatov, 1997).

Кечке_n_dat жуук_adv Арсен_np_ant_m_nom Саманчин_np_ant_m_nom чыдап_v_tv_prc_perf олтура_v_iv_prc_impf албады_v_iv_. _sent

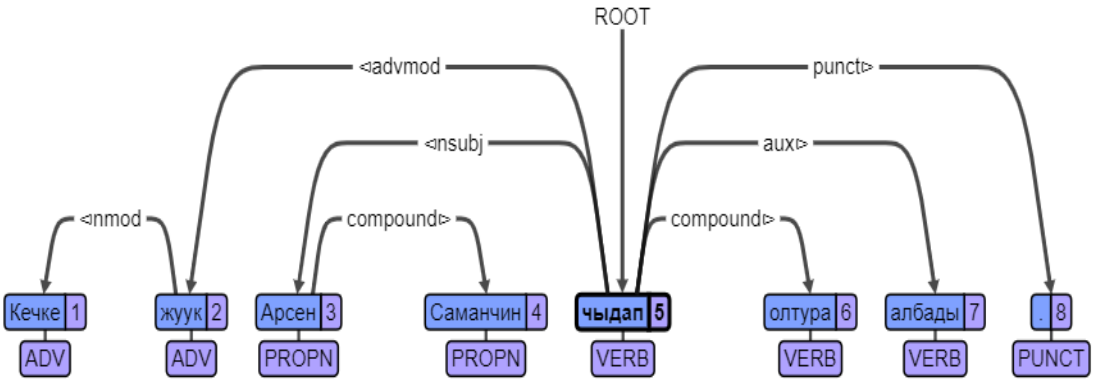



*Figure 4. 3 Dependency parsing of the compound verb "чыдап олтура албады"*

"<Кечке>"

"кеч" ADV @nmod #1->2

"<жуук>"

"жуук" ADV @advmod #2->5

"<Арсен>"

"Арсен" PROPN @nsubj #3->5

"<Саманчин>"

"Саманчин" PROPN @compound #4->3

"<чыдап>"

"чыда" VERB @root #5->0

"<олтура>"

"олтур" VERB @compound #6->5

"<албады>"

"ал" VERB @aux #7->5

"<.>"

"." PUNCT @punct #8-

2. *Сахнага чыккан акындарды эл дуулдата кол чаап , кызуу* ***коштоп турду.*** (Osorov, 2021).

Сахнага_n_dat чыккан_v_iv_gpr_past_subst_nom акындарды_n_pl_acc эл_n_nom дуулдата_unknown кол_n_nom чаап_v_tv_prc_perf ,_cm кызуу_unknown коштоп_unknown турду_v_iv_ifi_p3_sg ._sent

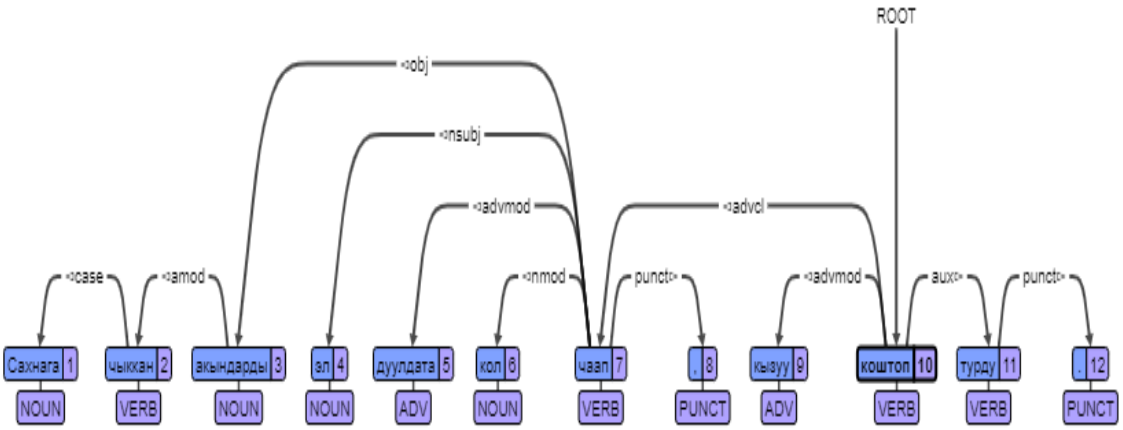



Figure 4. 4 Dependency parsing of the compound verb “коштоп турду”

”<Сахнага>”

“сахна” NOUN @case #1->2

“<чыккан>”

“чык” VERB @amod #2->3

“<акындарды>”

“акын” NOUN @obj #3->7

“<эл>”

“эл” NOUN @nsubj #4->7

“<дуулдата>”

“дуулдат” ADV @advmod #5->7

“<кол>”

“кол” NOUN @nmod #6->7

“<чаап>”

“чаап” VERB @advcl #7->10

“<,>”

“,” PUNCT @punct #8->7

“<кызуу>”

“кызуу” ADV @advmod #9->10

“<коштоп>”

“кошто” VERB @root #10->0

“<турду>”

“тур” VERB @aux #11->10

“<.>”

“.” PUNCT @punct #12->11

Word order of compound verbs in a sentence also the same as simple verbs, they always come at the end of a sentence and main verb pair (notional verb-негизги маани

берүүчү этиш) of a compound verb structure precedes auxiliary verb (жардамчы этиш). In these compound verb pairs ***“чыдап олтура албады”, “коштоп турду”, “чыдап”*** and ***“ коштоп”*** are the main notional verbs**,** auxiliary verbs are “***олтура албады” “ турду”.***

However, the nature and structure of compound verbs have not yet been fully investigated in both Kyrgyz linguistics and Turkology. There are different opinions about this issue among scientists (Хидирова & Авязова, 2008).

### 4.2.2 Main (notional) and Auxiliary Verbs

In compound verbs which consist of more than one pair of verbs and notional meaning is always attributed to the first pair, i.e. to the main verb (негизги этиш). The second pair denotes no lexical meaning on its own, it just takes a supportive role for main verb in a sentence, that is why it is called auxiliary verb (жардамчы/көмөкчү этиш). Auxiliary (helping) verbs bring clarity and some addition to the meaning of the main verbs, and help to express various grammatical meanings like tense, number etc.

Thus, we came to definition that *a word in the compound verb system that fully preserves the original/notional meaning is called the main verb,* and *the word whose verb meaning is weakened or sometimes completely lost and supports and adds the meaning to the main verb is called an auxiliary verb.*

Turn your attention to the following examples of auxiliary verb annotation:

1. *Ал кымыз деп ууну* ***куюп берди*** (Er Toshtuk, 1996)**.**

Ал_prn_pers_p3_sg_nom кымыз_n_nom деп_v_tv_prc_perf ууну_n_acc куюп_v_tv_prc_perf берди_vaux_ifi_p3_sg ._sent

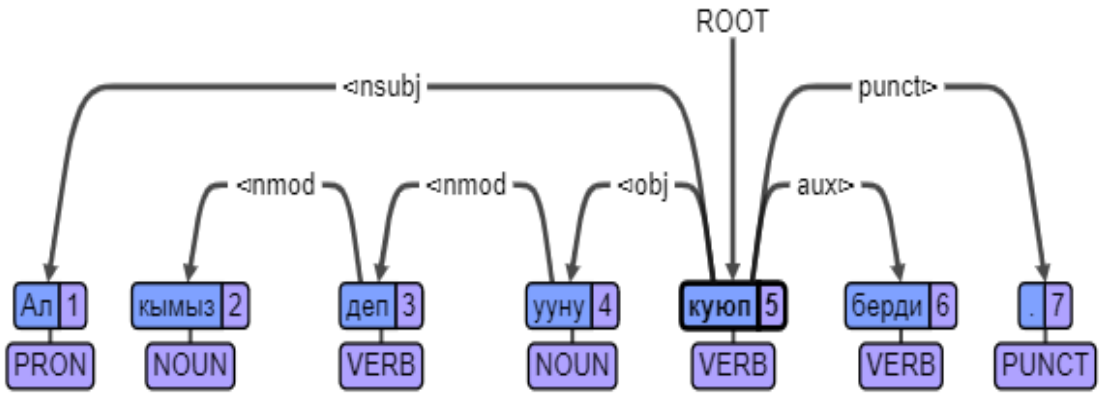



Figure 4.5 Dependency parsing of the main verb “куюп” and the auxiliary verb “берди”

”<Ал>”

“Ал” PRON @nsubj #1->5

“<кымыз>”

“кымыз” NOUN @nmod #2->3

“<деп>”

“де” VERB @nmod #3->4

“<ууну>”

“уу” NOUN @obj #4->5

“<куюп>”

“куй” VERB @root #5->0

“<берди>”

“бер” VERB @aux #6->5

“<.>”

“.” PUNCT @punct #7->5

In this compound verb “***куюп берди”***, “***куюп”*** is the main verb which denotes main idea of action. The second verb “***берди”*** is the auxiliary verb that does not have any meaning just expressing the main verb’s number and tense categories (third person singular, past tense).

2. Сарыбайдын тишин бир сермеп ***сууруп салды*** (Сынган кылыч).

Сарыбайдын_np_ant_m_gen тишин_n_px3sp_acc бир_num сермеп_adv сууруп_unknown салды_v_tv_ifi_p3_sg ._sent

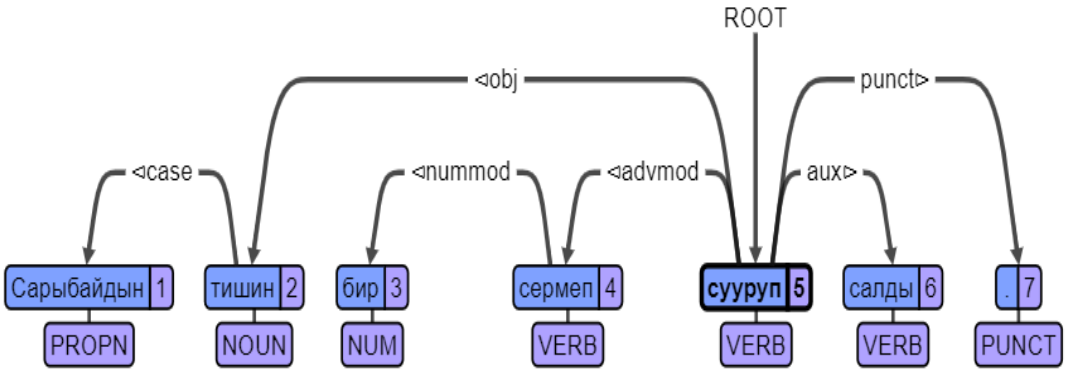



Figure 4.6 Dependency parsing of the main verb “сууруп”, and the auxiliary verb “салды”

”<Сарыбайдын>”

“Сарыбай” PROPN @case #1->2

“<тишин>”

“тиш” NOUN @obj #2->5

“<бир>”

“бир” NUM @nummod #3->4

“<сермеп>”

“серме” VERB @advmod #4->5

“<сууруп>”

“суура” VERB @root #5->0

“<салды>”

“сал” VERB @aux #6->5

“<.>”

“.” PUNCT @punct #7->5

In this compound verb pairs “***сууруп салды”***, “***сууруп”*** is the main verb which denotes main idea of action. The second verb “***салды”*** is the auxiliary verb that does not have any meaning just expressing the main verb’s number and tense categories (third person singular, past tense).

### 4.2.3 Types of Compound Verbs

Compound verbs, in the Kyrgyz language are classified into three depending on the nature of the words that they comprise.

1. ***Compound verbs with verbal pairs*** *(чакчыл түгөйлүү татаал этиштер).* Compound verbs whose constituents are only verbs are called ***compound verbs with verbal pairs.*** The first couplet of such verbs is always in the present tense and indicates the main action. And the second pair adds to its grammatical meaning and becomes an auxiliary verb. For instance:

   ***1.*** Абил аны колу менен ***көрсөтө берди.***

   Абил_np_ant_m_nom аны_prn_pers_p3_sg_acc колу_n_px3sp_acc-ind менен_post көрсөтө_v_tv_prc_impf берди_vaux_ifi_p3_sg ._sent

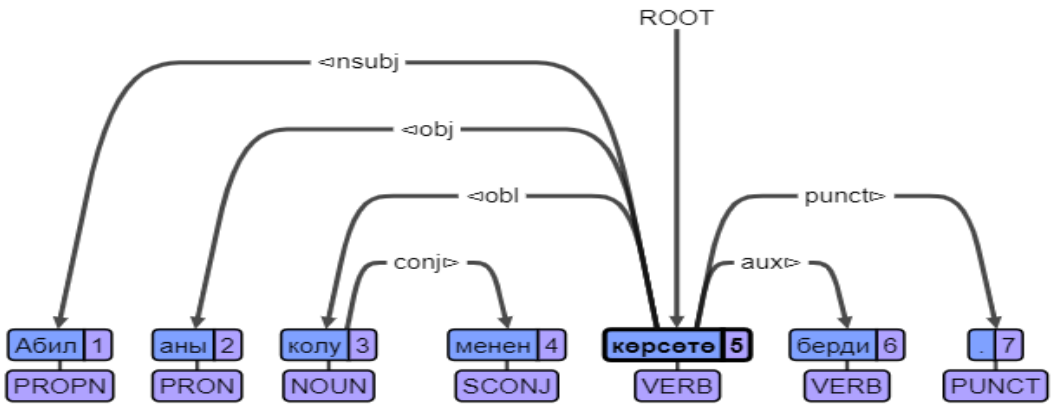



Figure 4.7 Dependency parsing of the compound verbal pairs “көрсөтө берди”

”<Абил>”

“Абил” PROPN @nsubj #1->5

“<аны>”

“ал” PRON @obj #2->5

“<колу>”

“кол” NOUN @obl #3->5

“<менен>”

“менен” SCONJ @conj #4->3

“<көрсөтө>”

“көрсөт” VERB @root #5->0

“<берди>”

“бер” VERB @aux #6->5

“<.>”

“.” PUNCT @punct #7->

***2.*** *Капыстан бетме - бет чыга түшкөн өлүм кызыл жүздүү жигитти* ***апкаарытып таштады*** (Kasymbekov, 1998)**.**

Капыстан_n_nom бетме-бет_n_nom чыга_v_iv_prc_impf түшкөн_v_iv_gpr_past өлүм_n_nom кызыл_adj жүздүү_n post жигитти_n_acc апкаарытып_v_tv_p3_sg таштады_v_tv_ifi_p3_sg ._sent

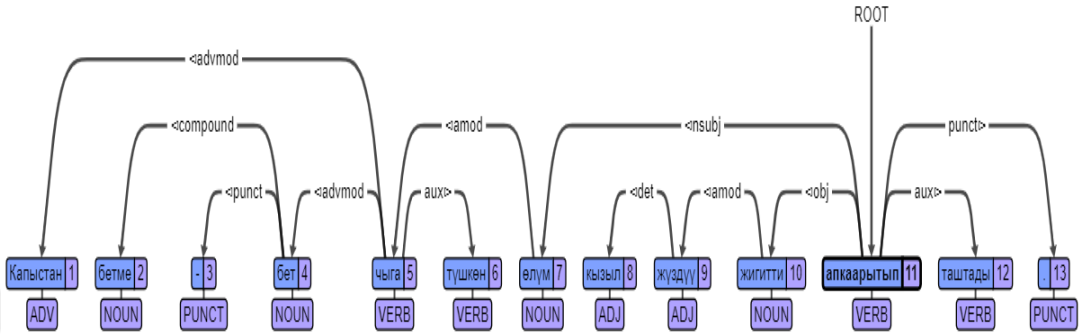



Figure 4.8 Dependency parsing of the compound verbal pairs ***“апкаарытып таштады”***

»<Капыстан>»

«Капыстан» ADV @advmod #1->5

«<бетме>»

“бетме” NOUN @compound #2->4

“<->”

“-“ PUNCT @punct #3->4

“<бет>”

“бет” NOUN @advmod #4->5

“<чыга>”

“чыга” VERB @amod #5->7

“<түшкөн>”

“түшкөн” VERB @aux #6->5

“<өлүм>”

“өлүм” NOUN @nsubj #7->11

“<кызыл>”

“кызыл” ADJ @det #8->9

“<жүздүү>”

“жүздүү” ADJ @amod #9->10

“<жигитти>”

“жигитти” NOUN @obj #10->11

“<апкаарытып>”

“апкаарытып” VERB @root #11->0

“<таштады>”

“таштады” VERB @aux #12->11

“<.>”

“.” PUNCT @punct #13->11

***Compound verbs with verbal pairs*** are those whose constituents are only ***verbs: "көрсөтө берди", "апкаарытып ташгады"***. Such verbs always have a first couplet that expresses the main action in the present tense as in the ***"көрсөтө"*** and ***"апкаарытып"***. Additionally, the second pair gains grammatical significance *(past tense and first person singular)* and changes into an auxiliary verb like ***"берди"***, ***"ташгады".***

There are compound verbs which constitute more than *three verb pairs*.

1. *Ибрагим Хайал шамдын түбүнө жетип, үлбүрөп турган билигин* ***карап туруп калды*** (Broken Sword).

Ибрагим_np_ant_m_nom Хайал_np_ant_m_nom шамдын_n_gen түбүнө_n_px3sp_dat жетип_v_tv_prc_perf үлбүрөп_adv турган_v_iv_past_p3_sg билигин_n_acc карап_v_tv_prc_perf туруп_vaux_prc_perf калды_v_iv_ifi_p3_sg ._sent

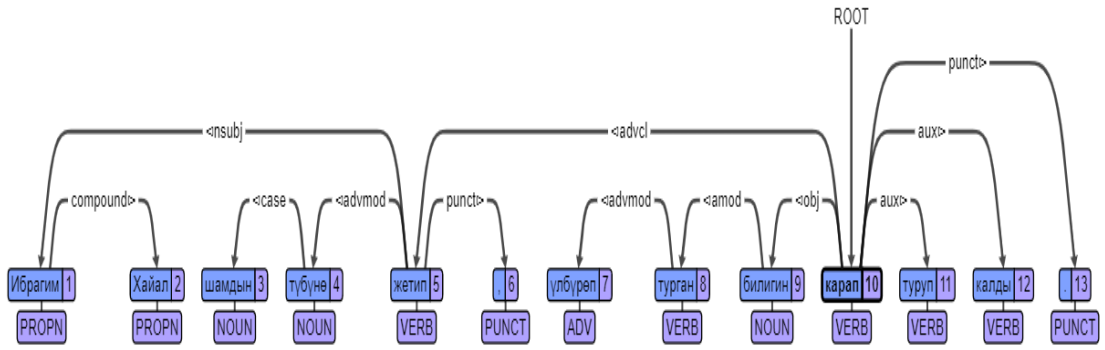



*Figure 4.9 Dependency parsing of the compound three verbal pairs* ***"карап туруп калды"***

"<Ибрагим>"

"Ибрагим" PROPN @nsubj #1->5

"<Хайал>"

"Хайал" PROPN @compound #2->1

"<шамдын>"

"шам" NOUN @case #3->4

"<түбүнө>"

"түп" NOUN @advmod #4->5

"<жетип>"

"жет" VERB @advcl #5->10

"<,>"

"," PUNCT @punct #6->5

“<үлбүрөп>”

“үлбүрө” ADV @advmod #7->8

“<турган>”

“тур” VERB @amod #8->9

“<билигин>”

“билик” NOUN @obj #9->10

“<карап>”

“кара” VERB @root #10->0

“<туруп>”

“тур” VERB @aux #11->10

“<калды>”

“кал” VERB @aux #12->10

“<.>”

“.” PUNCT @punct #13->10

In this example, grammatical category is shown by the last pair “***калды”.*** The left two pairs ***“карап туруп”*** are formed with help of derivational suffixes that make up a verb from verbal root (чакчыл формада түзүлөт). Mostly, in these kinds of pairs like in this particular case, the first verb ***(карап)*** expresses the main meaning, the second one ***(туруп)*** shows the continuity of an action.

2. *Аксакалдар колдорун көкүрөктөрүнө алышып, үн көтөрө салам* ***айтып калып жатышты*** *(Сынган кылыч).*

Аксакалдар_n_pl_nom колдорун_n_pl_px3sp_acc көкүрөктөрүнө_n_pl_px3sp_dat алышып_v_tv_coop_prc_perf ,_cm үн_n_nom көтөрө_v_tv_prc_impf салам_vaux_aor_p1_sg айтып_v_tv_prc_perf калып_vaux_prc_perf жатышты_vaux_ifi_p3_pl ._sent

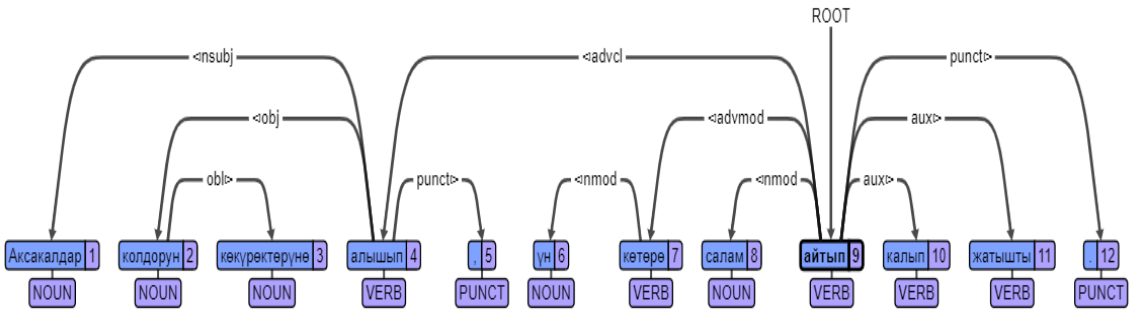



Figure 4.10 Dependency parsing of the compound three verbal pairs ***“айтып калып жатышты”***

”<Аксакалдар>”

“Аксакал” NOUN @nsubj #1->4

“<колдорун>”

“кол” NOUN @obj #2->4

“<көкүрөктөрүнө>”

“көкүрөк” NOUN @obl #3->2

“<алышып>”

“ал” VERB @advcl #4->9

“<,>”

“,” PUNCT @punct #5->4

“<үн>”

“үн” NOUN @nmod #6->7

“<көтөрө>”

“көтөр” VERB @advmod #7->9

“<салам>”

“салам” NOUN @nmod #8->9

“<айтып>”

“айт” VERB @root #9->0

“<калып>”

“кал” VERB @aux #10->9

“<жатышты>”

“жат” VERB @aux #11->9

“<.>”

“.” PUNCT @punct #12->9

In this instance, the final pair of these verbs “***жатышты***” indicates the grammatical category *(past tense, third person plural, cooperative mood).* The derivational suffixes that build a verb from its verbal root (чакчыл формада түзүлөт) are used to produce the left two pairs, “***салам айтып калып”*** The first verb ***(салам айтып)*** typically provides the major meaning in these types of pairs, while the second verb ***(калып)*** illustrates the continuity of an activity.

3. ***Compound Verbs with Nominal Pairs*** *(Атооч түгөйлүү татаал этиштер).* In these verb pairs, the first pairs are composed of nouns. For example: *ашык бол-, киши бол-, кол кой-, бир кой-, баш бак etc*. Compound verbs with nominal pairs can be divide distributed into 5 groups nouns depending on *the case* of the paired nouns.

    a) Compound verbs with nouns in nominative case: *киши бол-, арача бол-, ашык бол-, көз сал-, эн сал-, бир сал-, тил ал-, байыр ал-, ат*

*кой-, каршы чык-, убада бер-, ал жет-, күч жет-, казан ас-, байге сай- etc.*

*b)* Compound verbs with nouns in dative case: *ишке сал-, калыпка сал-, жөнгө сал-, добушка сал-, кайгыга сал-, башка сал-, эпке кел-, уятка кал-, четке как-, добушка кой-, изине түш-, жоопко тарт- etc.*

*c)* Compound verbs with nouns in accusative case: *ачууну жаз-, жарпты жаз-, башты жаз-, жоопту бер-, сабатсыздыкты жой-, экини жой-, карызды жой- etc.*

d) Compound verbs with nouns in locative case: *бейпилде жат-, өкүттө кал- etc.*

*e)* Compound verbs with nouns in ablative case: *өтөсүнөн чык-, үстүнөн чык-, анттан тай-, тилден кал-, жандан кеч- etc.*

1. Жогортон уруксатсыз силерди өткөрүп жиберсек ***жоопко тартылып калабыз*** (Kyrgyn).

sent_cm жогортон_adv уруксатсыз_n post силерди_n_pl_acc өткөрүп_v_tv_caus_prc_perf жиберсек_v_tv_prc_cond_p1_pl жоопко_n_dat тартылып_v_tv_pass_prc_perf калабыз_vaux_aor_p1_pl!_sent

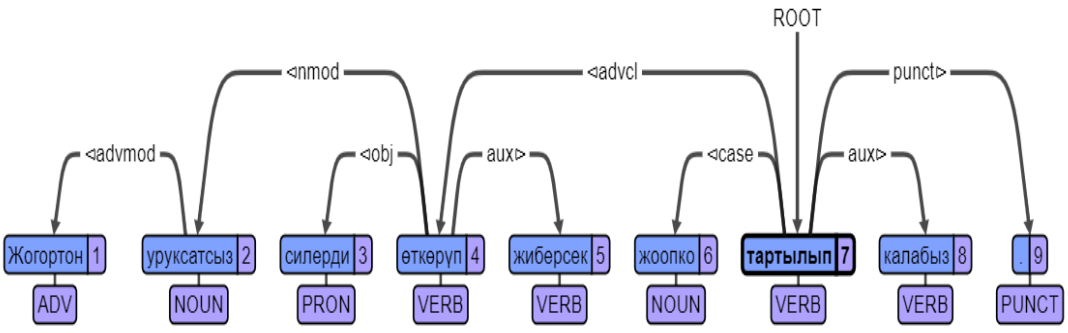



*Figure 4.11 Dependency parsing of the nominal verb pairs "жоопко тартылып калабыз"*

"<Жогортон>" "жогору" ADV @advmod #1->2

“<уруксатсыз>”

“уруксат” NOUN @nmod #2->4

“<силерди>”

“силер” PRON @obj #3->4

“<өткөрүп>”

“өткөр” VERB @advcl #4->7

“<жиберсек>”

“жибер” VERB @aux #5->4

“<жоопко>”

“жооп” NOUN @case #6->7

“<тартылып>”

“тарт” VERB @root #7->0

“<калабыз>”

“кал” VERB @aux #8->7

“<.>”

“.” PUNCT @punct #9->7

In these nominal verb pairs “***жоопко тартылып калабыз***”, the first pair is composed of noun - ***“жооп”*** in dative case - ***“жоопко”.***

2. Арсен ***тилден калды*** (Тоолор кулаганда).

   Арсен_np_ant_m_nom тилден_n_abl калды_v_iv_ifi_p3_sg ._sent

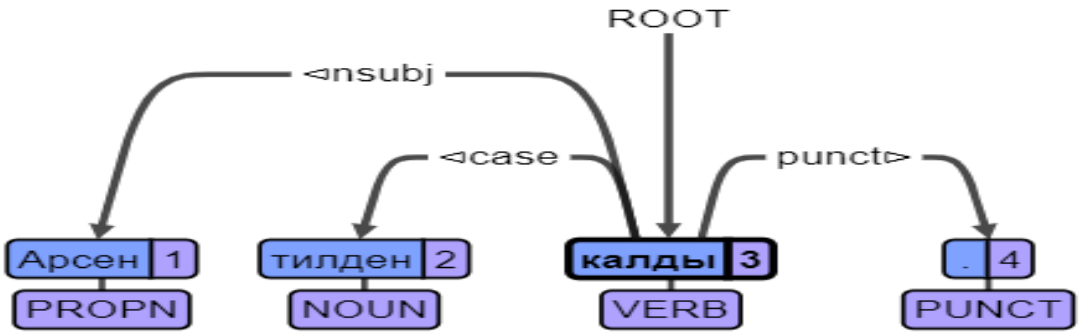



Figure 4.12 Dependency parsing of the nominal verb pairs “тилден калды”

“<Арсен>”

“Арсен” PROPN @nsubj #1->3

“<тилден>”

“тил” NOUN @case #2->3

“<калды>”

“кал” VERB @root #3->0

“<.>”

“.” PUNCT __ @punct #4->3

In these *nominal verb pairs* ***“милден калды”***, the first pair is composed of noun - ***“мил”*** in ablative case - ***“милден”.***

4. ***Compound Verbs with Ideophone Pairs.*** There are also compound verbs with the first pair of verbs are either ideophones (тууранды сөздөр) or figurative (imaginary) words (элестүү сөздөр): *тырп эт-, күп эт-, үрп эт-, жарк эт-, заңк эт-, жарк-журк эт-, солк эт-, мыңк эт-, тарс эт-, чырм эт-, болк эт-, ар эт-, быш эт-, быш де-, кыш де-, чү де-, кош де-, кың де-, күңк де-, кылт эт-, чү кой-, дыр кой-, жылт кой-, чөк түш-, бүк түш- etc.*

1. *Кесилген чачты көрүп, Акбалбандын жүрөгү* ***болк этти*** *(Кел-кел).*
   Кесилген_v_tv_pass_gpr_past чачты_n_acc көрүп_v_tv_prc_perf ,_cm Акбалбандын_n_pn_ant_sg_gen жүрөгү_n_px3sp_nom болк_unknown этти_v_iv_ifi_p3_sg ._sent

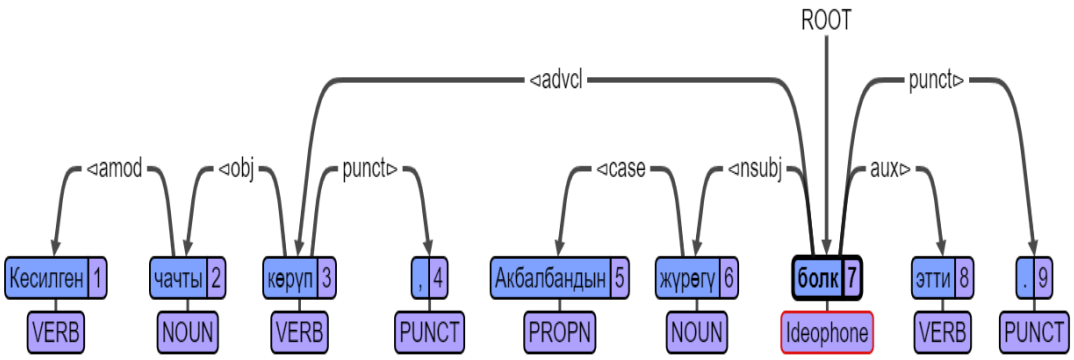



*Figure 4.13 Dependency parsing of the ideophone verb pairs “болк этти”*

”<Кесилген>”

“Кес” VERB @amod #1->2

“<чачты>”

“чач” NOUN @obj #2->3

“<көрүп>”

“көр” VERB @advcl #3->7

“<,>”

“,” PUNCT @punct #4->3

“<Акбалбандын>”

“Акбалбан” PROPN @case #5->6

“<жүрөгү>”

“жүрөк” NOUN @nsubj #6->7

“<болк>”

“болк” Ideophone @root #7->0

“<этти>”

“эт” VERB @aux #8->7

“<.>”

“.” PUNCT @punct #9->7

**2.** Эртең менен караса, айлана жарк - журк этет (Алтын шакек).

Эртең менен_adv караса_v_tv_prc_cond_p3_sg, cm айлана_v_iv_prc_impf жарк_ -_guio журк_unknown этет_v_iv_aor_p3_sg. Sent

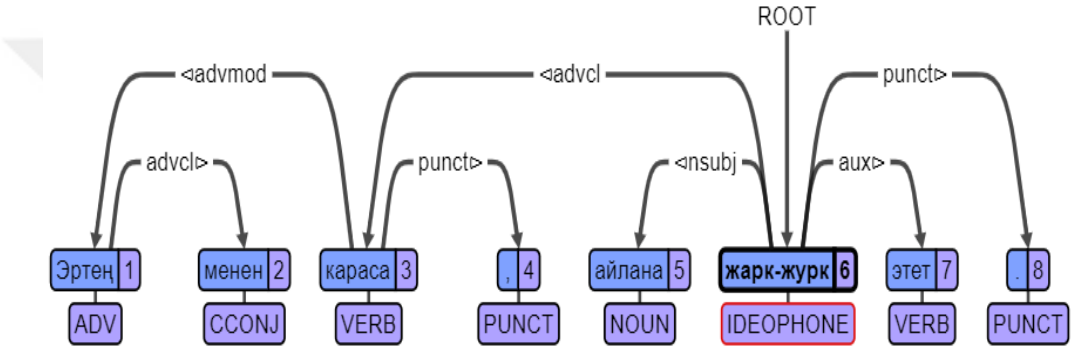



Figure 4.18 Dependency parsing of the ***ideophone verb pairs “жарк-журк этет”***

”<Эртең>”

“Эртең” ADV @advmod #1->3

“<менен>”

“менен” CCONJ @advcl #2->1

“<караса>”

“кара” VERB @advcl #3->6

“<,>”

“,” PUNCT @punct #4->3

“<айлана>”

“айлана” NOUN @nsubj #5->6

“<жарк-журк>”

“жарк-журк” IDEOPHONE @root #6->0

“<этет>”

“эт” VERB @aux #7->6

“<.>”

“.” PUNCT @punct #8->6

In these cases, the first notional verbs constitute from ideophones like ***“жарк-журк”, “болк”,*** the rest pairs ***“этет”, “этти”*** are auxiliary verbs which depict the grammatical categories (*present tense, third person singular, in the second, past tense, third person plural*).

## 4.3 Notional or Semantic Classification of Verbs

Despite the fact that verbs generally convey action and state of being, they are not actually the same in terms of general grammatical meaning. Some of them clearly express actions, movements while others do not. Thus, the verbs in the Kyrgyz language are internally divided into several notional groups as follows:

1. action verbs (кыймыл этиштер),
2. state verbs (абал эиштер),
3. modifying verbs or verbs of change of state (өзгөрүм этиштер),
4. verbs of sensation (сезим этиштер)

### 4.3.1 Action Verbs (Кыймыл этиштер)

Action verbs obviously express action. They are further subdivided according to the relationship between the subject performing the action and the object involved in the action:

a. Verbs that indicate the subject's movement/action, direction to something: *жүгүр-, чурка-, бас-, жорт-, чап-, ур-, кайт-, чык-, көтөрүл-, калкы-, серүүндө- etc.* Such verbs cause the subject of the sentence to move, to act, that is, the doer/subject that performs the action itself moves or does some kind of action.

1. Караан бери ***жүгүрдү*** (Сынган кылыч).

Караан_n_nom_sg бери_adv жүгүрдү_v_iv_ifi_p3_sg :_sent

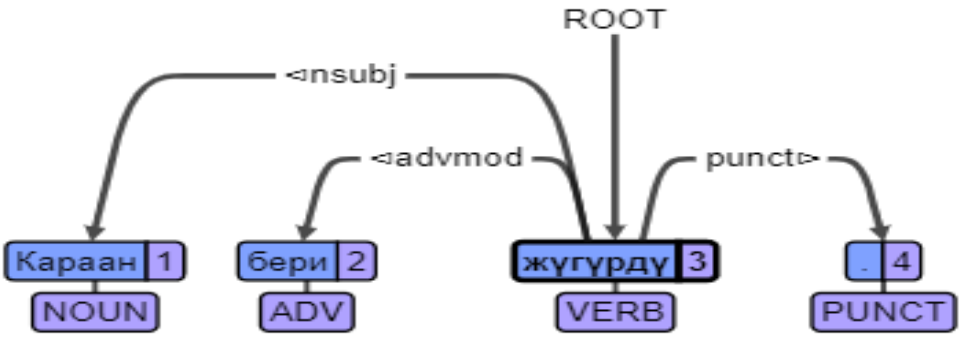



Figure 4.19 Dependency Parsing of the action verb ***“жүгүрдү”***

«<Караан>»

«Караан» NOUN @nsubj #1->3

«<бери>»

«бери» ADV @advmod #2->3

«<жүгүрдү>»

«жүгүр» VERB @root #3->0

“<.>”

“.” PUNCT @punct #4->3

b. Verbs that cause changes in their object: *каз-, курут-, майла-, самында-, көтөр-, алып бер-, алып кел- etc.* Such kind of verbs cause change in the object of a sentence to what it is directed.

Согушка жараамдуулугун бир кыйла ***көтөрдү (***Сынган кылыч).

Согушка_v_tv_ger_pres_dat жарамдуулугун_n_sg_acc бир кыйла_adj_advl көтөрдү_v_tv_ifi_p3_sg ._sent

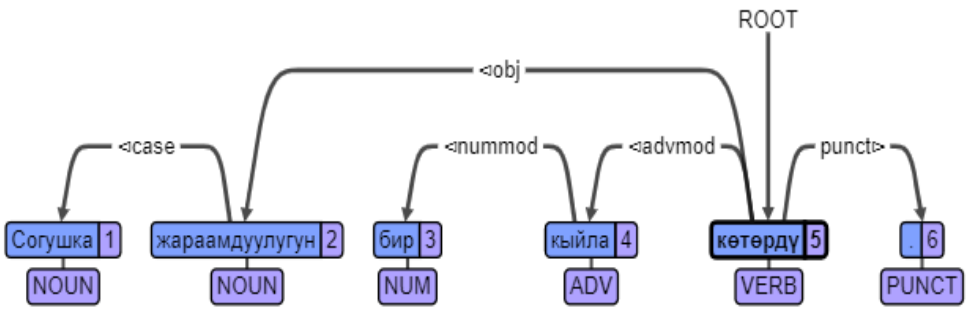



Figure 4.20 Dependency Parsing of the action verb ***“көтөрдү”***

”<Согушка>”

“Согуш” NOUN @case #1->2

“<жараамдуулугун>”

“жараамдуулук” NOUN @obj #2->5

“<бир>”

“бир” NUM @nummod #3->4

“<кыйла>”

“кыйла” ADV @advmod #4->5

“<көтөрдү>”

“көтөр” VERB @root #5->0

“<.>”

“.” PUNCT @punct #6->5

### 4.3.2 State Verbs (Ал-абал этиштер)

This kind of verbs do not apparently express action rather they indicate a state of being of subject in a sentence: *отур-, олтур-, жат-, тур-, бол- etc.* To get deep understanding, let us consider these two sentences:

1. *Айзаада жанында* ***отурат.***

Айзаада_propn_nom_ant_sg жанында_n_px3sp_loc отурат_v_iv_aor_p3_sg ._sent

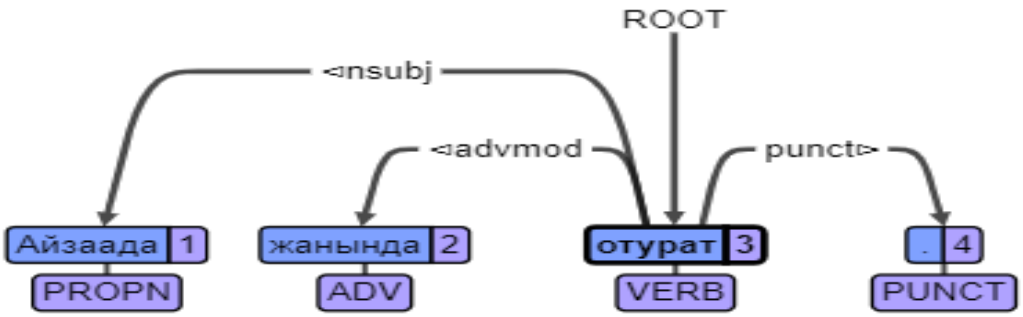



Figure 4.21 Dependency Parsing of the state verb ***"отурат"***

"<Айзаада>"

"Айзаада" PROPN @nsubj #1->3

"<жанында>"

"жанында" ADV @advmod #2->3

"<отурат>"

"отур" VERB @root #3->0

"<.>"

"." PUNCT @punct #4->3

2. *Чокусунда тартылган түндүк окшоп кичинекей көл* ***жатат*** (Kasymbekov, 1998).

Чокусунда_n_px3sp_loc тартылган_adj_subst_nom түндүк_adj окшоп_v_iv_prc_perf кичинекей_adj көл_n_nom жатат_v_iv_aor_p3_sg ._sent

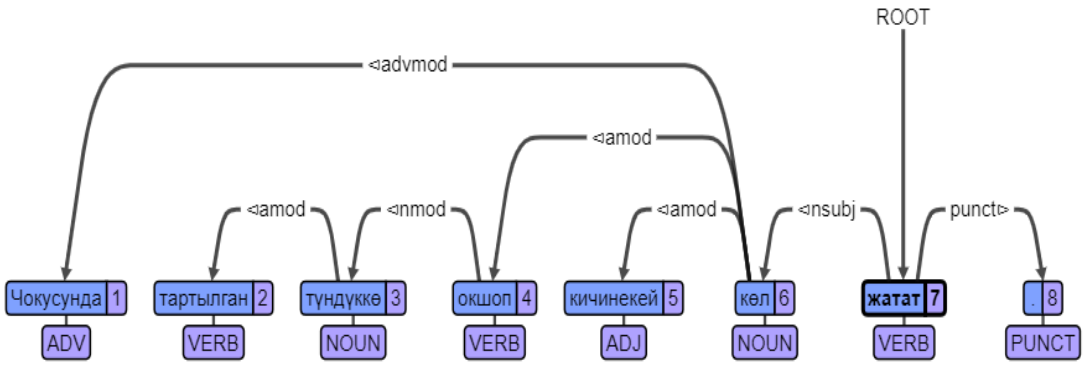



Figure 4.22 Dependency Parsing of the state verb ***"жатат"***

»<Чокусунда>»

«Чоку» ADV @advmod #1->6

«<тартылган>»

«тарт» VERB @amod #2->3

«<түндүккө>»

«түндүк» NOUN @nmod #3->4

«<окшоп>»

«окшо» VERB @amod #4->6

«<кичинекей>»

«кичинекей» ADJ @amod #5->6

«<көл>»

«көл» NOUN @nsubj #6->7

«<жатат>»

«жат» VERB @root #7->0

«<.>»

«.» PUNCT @punct #8->7

As we can see from the examples, the verbs "***отурат", "жатат"*** are indicating the state of being of the subjects "Айзаада", "көл" rather than an action. We usually infer a meaning of state verbs from the context (the whole sentence).

*State verbs* sometimes attain *compound form* structure:

1.Ал эми Саякбай Каралаев доордун айтуучусу ***болуп калды***.

Ал эми_cnjadv Саякбай_np_ant_m_nom Каралаев_np_cog_m_nom доордун_n_gen айтуучусу_v_tv_gpr_pot_subst_px3sp_nom болуп_v_iv_prc_perf калды_v_iv_ifi_p3_sg ._sent

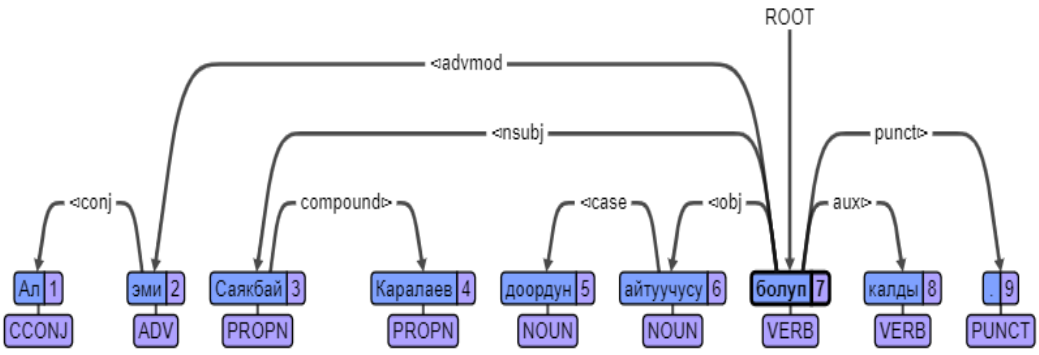



Figure 4.23 Dependency Parsing of the compound state verb ***"болуп калды"***

"<Ал>" "Ал" CCONJ @conj #1->2

“<эми>”

“эми” ADV @advmod #2->7

“<Саякбай>”

“Саякбай” PROPN @nsubj #3->7

“<Каралаев>”

“Каралаев” PROPN @compound #4->3

“<доордун>”

“доор” NOUN @case #5->6

“<айтуучусу>”

“айт” NOUN @obj #6->7

“<болуп>”

“бол” VERB @root #7->0

“<калды>”

“кал” VERB @aux #8->7

“<.>”

“.” PUNCT @punct #9->7

2. Нөшөрлөп төккөн калың жаан манасчынын күчүнө күч кошуп, шериктеш ***болуп турду!*** (Karalaev, 2010)**.**

Нөшөрлөп_v_tv_prc_perf төккөн_v_iv_past_p3_sg калың_adj жаан_n_nom манасчынын_n_gen күчүнө_n_px3sp_dat күч_n_nom кошуп_v_tv_prc_perf ,_cm шериктеш_n_nom болуп_v_iv_prc_perf турду_v_iv_ifi_p3_sg !_sent

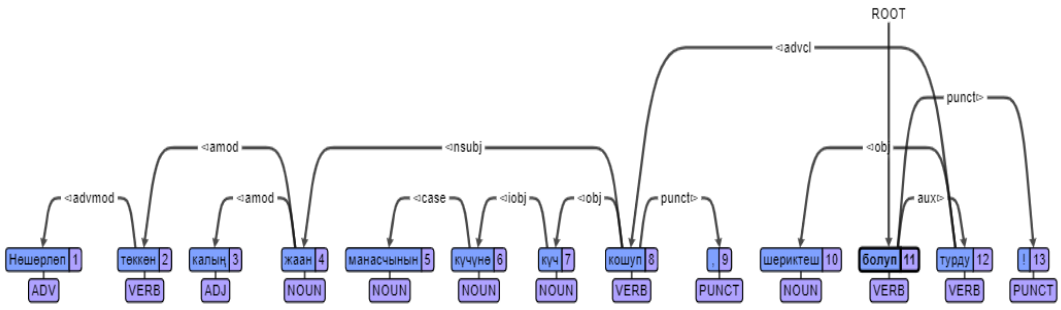



Figure 4.24 Dependency Parsing of the compound sate verb ***“болуп турду”***

”<Нөшөрлөп>”

“Нөшөрлө” ADV @advmod #1->2

“<төккөн>”

“төк” VERB @amod #2->4

“<калың>”

“калың” ADJ @amod #3->4

“<жаан>” “<,>”

“жаан” NOUN @nsubj #4->8 “,” PUNCT @punct #9->8

“<манасчынын>” “<шериктеш>”

“манас” NOUN @case #5->6 “шерик” NOUN @obj #10->12

“<күчүнө>” “<болуп>”

“күч” NOUN @iobj #6->7 “бол” VERB @root #11->0

“<күч>” “<турду>”

“күч” NOUN @obj #7->8 “тур” VERB @aux #12->11

“<кошуп>” “<!>”

“кош” VERB @advcl #8->12 “!” PUNCT @punct #13->1

Th ese examples illustrates that the compound state verbs “***болуп турду”, “болуп калды”*** are indicating the state of being of the subjects “Саякбай Каралаев”, “жаан” rather than an action. We usually infer a meaning of state verbs from the context (the whole sentence).

### 4.3.3 Modifying Verbs or Verbs of Change of State (Өзгөрүм этиштер)

Verbs of change of State or Modifying verbs indicate that the subject or object has undergone some change in quantity or quality: *агар-, жаша-, түлө-, той-, чанай-, кампай-, ичиркен-, семир-, арыкта- etc.*

***E.g.*** *Арык жүрүп* ***семирдим,***

*Ачка жүрүп* ***тоюндум,***

*Жардылыктан* ***байыдым,***

*Жалгыздыктан* ***көбөйдүм*** (Karalaev, 2010)**.**

Арык_adj жүрүп_v_iv_prc_perf семирдим_v_iv_ifi_p1_sg ,_cm
Ачка_adj жүрүп_v_iv_prc_perf тоюндум_ v_iv_ifi_p1_sg ,_cm

Жардылыктан_n_acc байыдым_v_iv_ifi_p1_sg ,_cm

Жалгыздыктан_n_acc көбөйдүм_v_iv_ifi_p1_sg ._sent

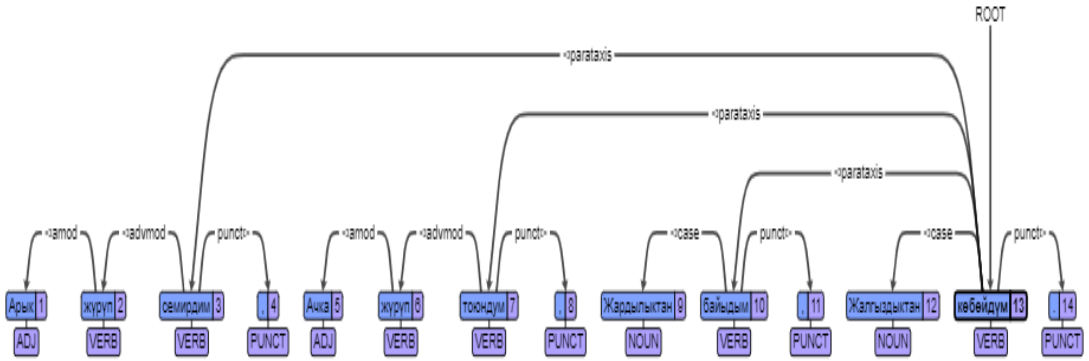



Figure 4.25 Dependency parsing of *the modifying verbs* ***“семирдим, тоюндум, байыдым, көбөйдүм”***

”<Арык>”

“Арык” ADJ @amod #1->2

“<жүрүп>”

“жүр” VERB @advmod #2->3

“<семирдим>”

“семир” VERB @parataxis #3->13

“<,>”

“,” PUNCT @punct #4->3

“<Ачка>”

“Ачка” ADJ @amod #5->6

“<жүрүп>”

“жүр” VERB @advmod #6->7

“<тоюндум>”

“той” VERB @parataxis #7->13

“<,>”

“,” PUNCT @punct #8->7

“<Жардылыктан>”

“Жарды” NOUN @case #9->10

“<байыдым>”

“байы” VERB @parataxis #10->13

“<,>”

“,” PUNCT @punct #11->10

“<Жалгыздыктан>”

“Жалгыз” NOUN @case #12->13

“<көбөйдүм>”

“көбөй” VERB @root #13->0

“<.>” “.” PUNCT @punct #14->13

From these examples, action is not very noticeable. As it is seen above, the verbs **“семирдим”, “тоюндум”, “байыдым”, “көбөйдүм”** do not express the action of the subject, but they denote rather the qualitative change. We notice the action of qualitative change gradually. Of course, both quantitative and qualitative changes in matter are the result of certain kind of action. Therefore, we call such words as verbs of quantitative and qualitative change.

### 4.3.4 Verbs of Sense (Сезим этиштер)

Sense verbs do not move cause action in the subject or object of a sentence, but only convey actions that is going in their mind or mental activity: *эсте-, түшүн-, ойло-, ук-, тыңша-, бил-, сүй-, сез-, көр-, эшит-, байка-, баамда- etc*. In such verbs, there is no action performed or about to be performed, but in the lexical meaning of these words, there is some inferred action, mental action, although it is not a clear action as we understand it.

For instance:

1. *Кыйды неме мунун көңүлүн эмне өйүп турганын **баамдады.***

Кыйды_adj неме_prn_nom_ мунун_prn_dem_gen көңүлүн_n_px3sp_acc
эмне_cnjcoo өйүп_ v_iv_prc_perf турганын_v_iv_ger_past_px3sp
баамдады_v_iv_ifi_p3_sg :_sent

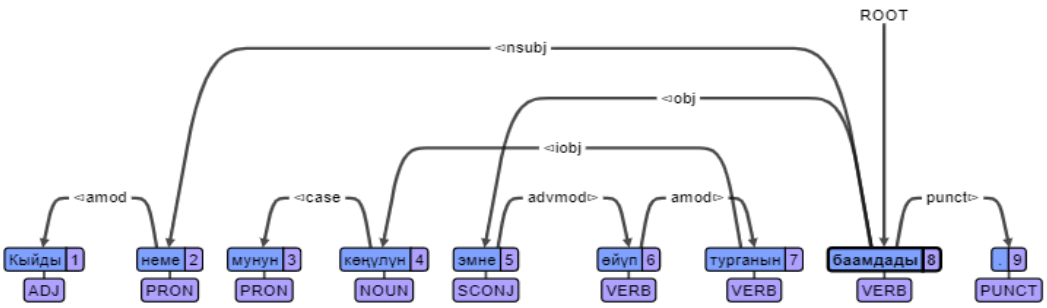



Figure 4.26 Dependency parsing of *the sense verb* **“баамдады”**

”<Кыйды>” “Кыйды” ADJ @amod #1->2

“<неме>”

“неме” PRON @nsubj #2->8

“<мунун>”

“бул” PRON @case #3->4

“<көңүлүн>”

“көңүл” NOUN @iobj #4->7

“<эмне>”

“эмне” SCONJ @obj #5->8

“<өйүп>”

“өйү” VERB @advmod #6->5

“<турганын>”

“тур” VERB @amod #7->6

“<баамдады>”

“баамда” VERB @root #8->0

“<.>”

“.” PUNCT @punct #9->8

2. *Ал Нүзүптүн ар жигитке бирден от жагууга буйрук кылганын* ***эстеди.***

Ал_prn_pers_p3_sg_nom Нүзүптүн_pn_ant_nom_sg_acc ар_det_ind жигитке_n_dat бирден_num_subst_abl от_n_nom жагууга_v_tv_ger_dat буйрук_n_nom кылганын_v_tv_ger_past_px3sp_acc эстеди_v_tv_ifi_p3_sg ._sent

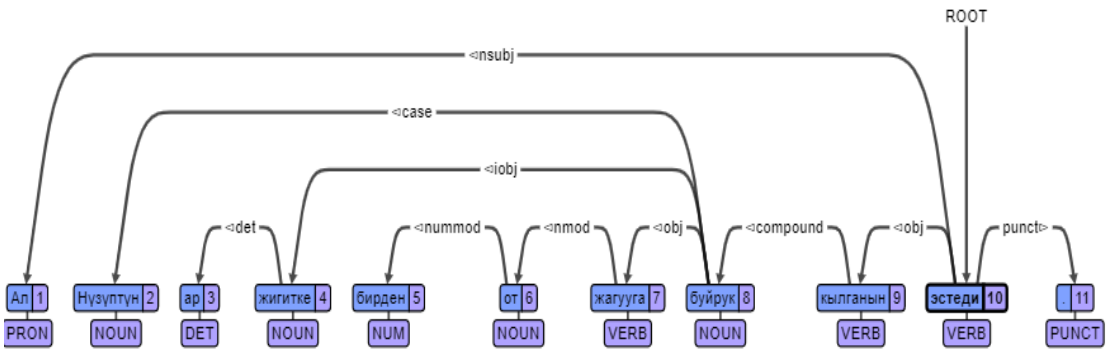



Figure 4.27 Dependency parsing of *the sense verb* **“эстеди”**

”<Ал>”

“Ал” PRON @nsubj #1->10

“<Нүзүптүн>”

“Нүзүп” NOUN @case #2->8

“<ар>”

“ар” DET @det #3->4

“<жигитке>”

“жигит” NOUN @iobj #4->8

“<бирден>”

“бир” NUM @nummod #5->6

“<от>”

“от” NOUN @nmod #6->7

“<жагууга>”

“жак” VERB @obj #7->8

“<буйрук>”

“буйрук” NOUN @compound #8->9

“<кылганын>”

“кыл” VERB @obj #9->10

“<эстеди>”

“эсте” VERB @root #10->0

“<.>”

“.” PUNCT @punct #11->10

The verbs **“баамдады”, “эстеди”** do not involve any action, but both of them potentially contain only the result of the action in the subject’s mind and mental activity.

**Deduction on Chapter 4**

The third chapter of the thesis deals with the morphological and syntactic analysis of verbs in the Kyrgyz language. In this chapter, we attempt to perform resolution process of *Verb Sense Disambiguation.* In order to study the Kyrgyz verbs, we applied the Kyrgyz Corpus and UD annotatrix annotation tool successfully. Examples which have been elicited form the Kyrgyz literary works that are compiled in the Kyrgyz corpus are examined, analyzed, disambiguated and then evaluated morphologically and syntactically.

As a result, we can claim that an important part of the Kyrgyz language’s vocabulary belongs to verb groups. Due to the fact that most of the verbs in the Kyrgyz language are ambiguous they directly or indirectly determine the syntactic valency of words in the category, especially nouns, and consequently they determine what kind of sentence is about to be structured. So, verbs in the Kyrgyz language form the core of the grammatical system.

## CONCLUSION

The importance of studying verbs in the Kyrgyz language is caused by the reason that, on the one hand, they reflect human's intention or action of what he/she speaks about, his/her psychological state or just a state of being and emotion and, on the other hand Kyrgyz verbs cause some difficulty due to their ambiguity. One and the same verb can have more than one meaning. Thus, this present paper aimed at revealing the nature of Kyrgyz verbs and solving the Kyrgyz verb's sense disambiguation and studying their lexico-grammatical and syntactic features in the framework of the Corpus linguistics and Universal Dependencies.

The general overview of the thesis embraces three different branches of linguistics, namely, *Corpus linguistics* which comprises itself The Newly-Created Kyrgyz Corpus, *Artificial Intelligence and Natural Language Processing* as a foundation for *Verb Sense Disambiguation.*

The *first chapter* of the thesis introduces with a general overview of Corpus Linguistics. This chapter provides with a broad information on Corpus linguistics' emergence and history, and its establishment as a separate field of linguistics, and classification according to certain types of criteria. The importance of using corpora in modern linguistics and its methods are discussed and some suggestions were made regarding usage of them in scientific works. At this stage, we mention the compilation and development process of Newly-created Kyrgyz Corpus.

The *second chapter* of the thesis discusses *Artificial Intelligence* as a basis of *Natural Language Processing* which is in turn a foundation for Word Sense Disambiguation. The emergence of Artificial Intelligence, the role of Artificial Intelligence in Linguistics was described in details. Also, *Natural Language Processing*, and its applications in linguistics were thoroughly explained.

The *third chapter* of the thesis deals with the concept of *"ambiguity"* in a language which cause difficulties to Natural Language Processing. And other concepts which are misused or misunderstood with the word *"ambiguity"* (which is not ambiguity) and its types was taken under our detail consideration. As our main objective we provided with

a broad theoretical background *of Word Sense Disambiguation and its subfield Verb Sense Disambiguation,* and their history in scientific world and importance in linguistics along with their methods and approaches.

The last, *fourth chapter* presents the morphological and syntactic analysis which have been carried out *in the Kyrgyz Corpus* and *Universal Dependency platform*. Results of analysis of Kyrgyz verbs were given in detail with the figures of syntactic parse trees and the results have been evaluated. In this section of our whole thesis paper, the practical resolution process of the Kyrgyz Verbs Sense Disambiguation was performed.

In order to achieve the goals and aims set in this paper, the following tasks have been carried out:

1. 32 sentences which contained Kyrgyz verbs have been elicited from the Kyrgyz Corpus Query Processor (CQP) and translated into English language;
2. The retrieved Kyrgyz verbs have been adapted into classification that Abduvaliev has given in his book:
    - Morphological and syntactic analysis of *simple verbs* have been carried out using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;
    - Morphological and syntactic analysis of *compound verbs* have been carried out using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;
    - Morphological and syntactic analysis of *main verbs* have been made using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;
    - Morphological and syntactic analysis of *auxiliary verbs* have been carried out using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;
    - Morphological and syntactic analysis of types of *types of compound verbs* have been made using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;

- Morphological and syntactic analysis of classification of *notional verbs* have been done using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;
- Morphological and syntactic analysis of *action verbs* have been carried out using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;
- Morphological and syntactic analysis of *state verbs* have been undertaken using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;
- Morphological and syntactic analysis of *modifying verbs and verbs of sense* have been made using Kyrgyz Corpus query processor and UD Annotatrix annotation tool;

3. In total, 32 sentences have been retrieved and singled out from the Kyrgyz Corpus from different literary genres. All these verbs have been morphologically tagged and syntactically analyzed in UD Annotatrix annotation tool. The syntactic tree parse of each analyzed verb has been given in figure format. Achieved results have been evaluated, made an analysis of them according to each type of verbs that have been chosen.

As this study is one of the first works carried out in the scope of Corpus and UD which are used for resolution of ambiguities in the Kyrgyz verbs, the study suggests new approaches, i.e. application of corpus-based and Universal Dependency tool for scientific research on different linguistic issues.

## ДИССЕРТАЦИЯНЫН КЫСКАЧА МАЗМУНУ

### Кыргыз корпусундагы этиштердин кош маанилүүлүгүн жобу ажана аларды энтектөө маселелери

Маалымат жана коммуникация технологияларынын өнүгүшү менен алардын системалары эбегейсиз чоң көлөмдөгү чийки иштетиле элек токтоосуз көбөйүп жаткан маалымат менен иштешип жатышат. Ошол маалыматтарды ар түрдүү адамдар жана ар кандай максатта колдоно алышы үчүн ал маалыматтар табигый бир негизде коомчулукка тартууланышы керек. Ушул багытты көздөгөн корпустук тил илими жан ага негизделген изилдөөлөр, онлайн лингвистика, интернет лингвистика ж.б. ушул сыяктуу компьютердик тил илимдери бири-бири менен тыгыз байланышып, бирдикте иш алып барышууда. Ошентип азыркы заманбап тил илиминде тилди ар түрдүү деңгээлден, өңүттөн алып, табигый тилди иштетүү процессинен өткөрсөк болот. Мисалга алсак, фонетика, морфология, синтаксис, семантика, прагматика ж.б. тил илимдерин бир же бир нече тилдик корпусту жана анын куралдарын колдонуп, терең изилдөө ишин жүргүзүүгө болот. Табигый тилди иштетүүнүн негизинде курулуп жаткан шул сыяктуу эволюциялар/куралдар маалыматтарды туура иштетип, колдонуучуга түшүнүктүү кылып жеткирүүдө зор роль ойнойт.

Табигый тилди иштетүү системалары жогоруда кеп кылынган тапшырмаларды толук жана туура аткаруу үчүн кадимки биз колдонгон табигый тилдин табиятын терең аңдап-билип, түшүнүшү керек. Тилди иштетип жатканда табигый тилдеги көп маанилүүлүктөр кыйынчылыктарды туудурат. Тилде көп маанилүүлүктөр ар түрдүү деңгээлде болушу мүмкүн: фонологиялык, морфологиялык, синтаксистик, семантикалык жана прагматикалык ж.б. Ошондуктан, алгач тилдеги көп маанилүүлүктөрдү жоюу табигый тилди иштетүүдө эң маанилүү тапшырмалардын бири.

Бул илимий иш кандайдыр бир контекст аркылуу сөздөрдүн маанисин чечмелөө менен лексикалык көп маанилүүлүктү жоюу маселелерин карайт. Бул маселе англис тилинде Word Sense Disambiguation (WSD)(сөздөрдүн кош

маанилүүлүгүн жоюу) деп аталат. Биздин илимий иш сөз түркүмдөрүнүн ичинен этиштердин кош маанилүүлүгүн жоюуга арналган. Анткени, дүйнөдөгү тилдердин көпчүлүгүндө көп маанилүү сөздөр кездешет. Дегеле табигый тилди иштетүү куралдарын кызматка киргизүүдө кандайдыр бир контекстке таянып, көп маанилүү сөздөрдүн санын азайтуу/маанилерин жоюу абдан маанилүү иш. Бул сыяктуу сөздөрдүн бар болушу системалардын натыйжалуу иштешине тоскоолдук жаратат.

**Бул илимий иштин актуалдуулугу:** бул илимий иштин актуалдуулуктарынын бири кыргыз тилиндеги этиштердин көп маанилүүлүгүн жоюу үчүн жаңы түзүлгөн кыргыз тилинин корпусунун негизинде ар бир этишке морфологиялык энтектер берилип, морфологиялык анализдин жасалышында болуп саналат. Азыркы күндө анализ кылууда керектелүүчү материалдардын санында жана көлөмүндө чек жок, анткени тилдик корпуста бир нече миллиондогон айрым учурларда миллиарддаган (мисалы, BNC, COCA деп аталган англис тилинин корпустары) сөздөрдү өз ичине камтыйт. Мисалга алсак, кыргыз тилинин жаңы түзүлгөн улуттук корпусунда 2 миллиондон ашык сөз морфологиялык деңгээлде энтектелип киргизилген. Натыйжада, корпустук ыкманы колдонуп жүргүзүлгөн илимий иштер эмпирикалык маалыматтар жана көп өлчөмдө натыйжалар менен байыйт. Корпустук ыкманы колдонуп, бир нече секунд ичинде компьютердин баскычын бир жолу чыкылдатуу менен сөздөрдүн, фразалардын, сүйлөмдөрдүн татаал сөз жана сүйлөмдөрдүн семантикалык, синтаксистик, морфологиялык, ж.б. деңгээлдердеги энтектерин, жыштыктарын, кездешүү тыгыздыгын алууга болот жана ошону менен бирге убактыңыз да үнөмдөлөт. Ушул себептер корпустук тил илимин ишибизде негизги ыкма катары тандап алууга түрткү болду.

Ошондой эле, бул илимий иште Универсалдуу багыныңкылык (англ.: Universal Dependency) долбоорунун UD Annotatrix annotation tool деп аталган энтектөө/аннотациялоо куралы менен тааныштырабыз. Бул курал этиштердин синтаксистик мүнөзүн, түзүлүшүн жана сүйлөм ичиндеги башка сөздөргө болгон көз карандуулугун чагылдыруу үчүн колдонулду. Бул UD Annotatrix annotation куралы ар түрдүү тилдердин грамматикасы, сөз түркүмдөрү, морфологиялык

мүнөздөмөсү жана синтаксистик көз карандылыгы көрсөтүлгөн, тынымсыз жаңылатылып туруучу атайын платформа болуп саналат.

**Бул илимий иштин максаты:** Сөздөрдүн (биздин иште этиштер менен чектелген) көп маанилүүлүгүн жоюунун (WSD) негизи болгон корпустук тил илими, жасалма интеллект, табигый тилди иштетүү тармактары боюнча теориялык маалымат топтоо; Сөздөрдүн (биздин иште этиштер менен чектелген) көп маанилүүлүгүн жоюунун (WSD) теориясын жана тарыхын, тил илиминдеги колдонулушун изилдөө; бул илимий иштин негизги максаты катары кыргыз тили жана андагы этиштер тууралуу баяндоо; сөздөрдүн (биздин иште этиштер менен чектелген) көп маанилүүлүгүн жоюу (WSD) процессин ишке ашыруудагы кыргыз тилиндеги этиштердин морфологиялык жана синтаксистик энтектелишин көрсөтүү жана ага анализ илимий анализ жүргүзүп, жыйынтыктарына илимий баа берүү үчүн корпустук ыкмалар жана UD Annotatrix аннотация куралы колдонуу. Бул илимий иштин жогоруда белгиленген максаттарына жетүү үчүн төмөндөгү алдыга коюлган милдеттерди жүзөгө ашыруу керектелет:

1) Кыргыз корпусундагы тексттерден этиштерди тандап алып чыгуу

2) Табылган этиштерди Абдувалиев “Азыркы кыргыз тилинин морфологиясы” китебинде жазган этиштердин классификациясы боюнча бөлүштүрүү

    - Кыргыз тилинин корпусунан алынган жөнөкөй этиштерге морфологиялык жана синтаксистик анализ жасоо;
    - Кыргыз тилинин корпусунан алынган татаал этиштерге морфологиялык жана синтаксистик анализ жасоо;
    - Кыргыз тилинин корпусунан алынган негизги этиштерге морфологиялык жана синтаксистик анализ жасоо;
    - Кыргыз тилинин корпусунан алынган жардамчы этиштерге морфологиялык жана синтаксистик анализ жасоо;

- Кыргыз тилинин корпусунан алынган татаал этиштин түрлөрүнө морфологиялык жана синтаксистик анализ жасоо;
- Кыргыз тилинин корпусунан алынган этиштердин маанилий топторуна/классификациясына морфологиялык жана синтаксистик анализ жасоо;
- Кыргыз тилинин корпусунан алынган кыймыл-аракет этиштерине морфологиялык жана синтаксистик анализ жасоо;
- Кыргыз тилинин корпусунан алынган ал-абалды билдирген этиштерге морфологиялык жана синтаксистик анализ жасоо;
- Кыргыз тилинин корпусунан алынган өзгөрүм жана сезим-туюм этиштерине морфологиялык жана синтаксистик анализ жасоо;

3) Ар бир кыргыз тилиндеги этиштерге алардын мүнөзү боюнча манилерин аныктап, ар бирин талдап, анализ жүргүзүү, анализдерди жыйынтыктап, натыйжаларына илимий баа берүү

**Изилдөөнүн объектиси:** Кыргыз тилиндеги этиштердин көп маанилүүлүгүн жоюу, алардын морфологиялык жана синтаксистик энтектелишин/аннотациясын берүү.

**Изилдөө ишинин предмети**: UD Annotatrix аннотациялоо/энтектөө куралы жана кыргыз тилинин корпусу.

**Изилдөөнүн илимий жаңылыгы:** бул тармакта жүргүзүлгөн изилдөөлөргө таянып, кыргыз тил илиминде корпустук жана UD Annotatrix аннотациялоо/энтектөө ыкмаларын колдонуу аркылуу этиштердин көп маанилүүлүгүн жоюу боюнча азырынча илимий-иштер аткарыла элек. Ошондуктан, бул илимий иш бул өңүттөн алып караганда алгачкы эмгек десек жаңылышпайбыз. Бул иште этиштердин көп маанилүүлүгү боюнча теория менен бирге практикалык изилдөөлөр да сунушталды. Кыргыз тилинен алынган

этиштердин көп маанилүүлүгүн жоюуга, аларды тереңдеп анализдөөгө жана маанилерин ийгиликтүү чечмелеп берүүгө далалат кылынды. Бул илимий иште, дегеле кыргыз тил илиминде UD Annotatrix аннотациялоо/энтектөө куралынын эң биринчилерден болуп колдонулушу да өзүнчө бир жаңылык тартуулап жатат.

**Изилдөөнүн илимий методикалык негиздери:** бул илимий иште төмөндөгү ыкмалар колдонулду:

Кыргыз тилиндеги этиштердин морфологиялык энтектери менен бирге алуу үчүн кыргыз тилинин корпусу, атап айтканда, *корпустук негизделген ыкма* (Corpus-based approach)колдонулду.

*Корпуска багытталган ыкма* (Corpus-driven approach ) этиштердин жыштыгын көрсөтүү үчүн колдонулду.

*Сандык (квантитативдик) ыкма* бул иште камтылган этиштер боюнча статистикалык маалыматтарды берүү үчүн колдонулду.

*Сапаттык (квалитативдик) ыкма* кыргыз тилиндеги этиштердин көп маанилүүлүгүн анализдөө, сыпаттап берүү, жана түшүндүрүү үчүн колдонулду.

*Контрасттык ыкма (*Comparative method) этиштерди англис тилине которуудагы кээ бир өзгөчөлүктөрдү аныктоодо колдонулду.

*Тандап алуу ыкмасы* кыргыз тилинин корпусундагы ар түрдүү жанрдагы адабияттардын арасынан этиштерди тандап алууда колдонулду.

*Сыпаттоо ыкмасы* кыргыз тили, жана андагы этиштердин, корпустук тил илиминин, этиштердин көп маанилүүлүгүн жоюу, табигый тилди иштетүү жана жасалма интеллект боюнча теориялык изилдөөлөрдү берүүдө колдонулду.

**Иштин теориялык жана практикалык баалуулугу:** бул иштин изилдөөнүн негизинде топтолгон теориялык маалыматтары машина котормо, корпустук жана комьютердик тил илимдери, кыргыз тилинин корпусунун өнүктүрүү, жасалма интеллект жана кыргыз тилин табигый тилди иштетүү процессинен өткөрүү

жаатында изилдөө иштерин жүргүзүүдө, аларды окутуп-үйрөтүүдө чоң бир база болуп бере алат. Мындан сырткары синтаксис жана морфология сабактарын окуткан мугалимдерге да жардамчы болуп бере алат. Буга кошумча, бул иш дегеле тилдеги этиштердин семантикасын изилдөөдө, өзгөчө азырынча толук изилденип бүтө элек, түркология илиминде дагы деле кызуу талаш-тартыштарды жаратып келген татаал этиштерге, алардын тутумуна, түзүлүшүнө, классификациясына жана жалпы эле табиятына анализ жасоого да жардам берет деген үмүттөбүз.

Ал эми практикалык баалуулугу тилди иштетүүнүн өзгөчө бир ыкмасы болгон, компьютердик эбегейсиз чоң көлөмдөгү маалымат камытыгандыгынан улам практикалык иштерде так жана туура маалымат алууга көздөлгөн корпустук ыкманын колдонулушу десек болот. Бул иштин анализи, практикалык бөлүгү кыргыз тилинин корпусу жана UD Annotatrix annotation куралдары колдонулуп ишке ашырылды. Корпустук тил илими тил илиминде кеңири колдонула элек же дегеле пайда боло электе изилдөө ишин аткаруу процессине аябай көп убакыт коротулуп, көп эмгек жумшоого туура келчү. Физикалык мүмкүнчүлүктөрдөн улам аз гана көлөмдөгү адабиятарды окуп изилдеп чыгууга мүмкүн эле. Ал эми азыркы учурда, корпустук тил илиминин негизинде компьютердик форматтагы чоң көлөмдөгү маалыматтар (тилдик корпустар миллиондогон сөздөрдү, айрым корпустар миллиарддаган сөздөрдү ичине камтыйт) менен иштешүү кыйла жеңилдеди. Ошондуктан, бул иштин жыйынтыгында келип-чыккан натыйжаларды төмөндөгү багыттарда колдонууга мүмкүн: кыргыз тилин табигый тилди иштетүү процессинен өткөрүүдө, корпустук тил илимин изилдөөдө, семантика,синтаксис, морфология жана кыргыз тилиндеги этиштердин көп маанилүүлүгүн жоюуда.

**Изилдөөнүн материалдары:** бул ишти жазып баштоодон мурун жана жазып жаткан убакта табигый тилди иштетүү, жсалма интеллект жана корпустук тил илими, жана сөздөрдүн  (анын ичинде этиштердин) көп маанилүүлүгүн жоюу боюнча бир нече адабияттарга көз чаптырдык. Табигый тилди иштетүүдө эң кыйын аткарылышы керек болгон жана аткарылып жаткан тапшырмалардын бири бул тилдеги сөздөрдүн маанилкрин чечмелөө аркылуу көп маанилүүлүгүн жоюу болуп саналат. Бул багытта алгачкылардан болуп, 1940-жылдарынын аягында Зиф (Zipf)

өзүнүн “Маанинин мыйзамы” (“Law of Meaning” ) деген эмгегин сунуштаган. Анын ткориясы боюнча, азыраак колдонулган сөздөр менен эң ко колнулган сөздөрдүн арасында маанилик айырмачылыктар бар, тактап айтканда, азыраак колдонулган сөздөргө караганда көбүрөөк колдонулган сөздөрдүн маанилери көп болот. Кийинчерээк бул маанилик байланыш теориясы British National Corpus тарабынан тастыкталган. Ал эми Каплан 1950-жылы бүтүндөй контексттик сүйлөмдөрдүн маанилери менен анын тутумундагы эки эле сөздүн ар тараптуу маанилери менен салыштырып караса болоорун сунуштаган. Андан соң 1957-жылы, Мастерман Рожеттин эл аралык Тезарусунда сөздүн туура жана так маанисин аныктоо үчүн категориялардагы баш аталыштарды кантип колдонсо болоорун түшүндүрүп, эмгек жазып чыккан. Сөздөрдүн көп маанилеринин ичинен туура маанини аныктоо үчүн тандап алуу менен лексика-семантикага чектелген ыкмаларды биргелештирип колдонууга болоорун баса белгилейт. 1979-жылдары Риджер жана Смол “сөз адистери” ("word experts.") түшүнүгүн киргизген. Масштабдуу көлөмдөгү лексикалык булактардын жана тилдик корпустардын пайда болушу менен, 1980-жылдары сөздүн кош маанилүүлүгүн жоюу тармагы чоң өзгөрүүгө дуушар болуп, салмактуу алдыга өнүгүү кадамын таштады. Жыйынтыгында, изилдөөчүлөр автоматтык түрдө маалымат алуу куралдары менен бирге кол аркылуу жасалып алынуучу куралдарды биргелештирип, массалык түрдө колдоно башташты. Кийинчерээк 1986-жылы Леск сөздөрдүн арасындагы жалпы окшош маанилерге (сөздүктөгү берилген аныктамалар) негизделген өзүнүн алгоритмин иштеп чыккан. Бул алгоритмде көп маанилүү сөздөрдөгү сөздүн сунуштала турчу мааниси ал сөздүн жалпысынан эң көп кездешкен маанисине (maximum number of overlaps) барабар. Бул ишинде Леск сөздөрдүн аныктамаларын алуу үчүн Oxford Advanced Learner’s Dictionary of Current English (OALD)сөздүгүн колдонгон. Бул алгоритм сөздүн маанисин сөздүктөгү аныктамаларга негиздеп иштеп чыгуу ыкмасынын пайда болуп, түптөлүшүнө негиз салган. Ал эми Гусри Азыркы заманбап англис тилинин сөздүгүн (Longman Dictionary of Contemporary English (LDOCE)) колдонуп, коддордун (the subject codes) негизинде сөздөрдүн маанилерин чечмелөөгө далалат кылган. Жалпысынан айтканда, сөздөрдүн анын ичинде этиштердин коп маанилүүлүгүн жоюу боюнча урунттуу жана маанилүү

материалдарды баяндап өттүк. Дагы кененирээк маалымат алуу үчүн бул магистрдик иштин ички баптарына көз чаптырсаңыз болот.

Бул илимий иштин дагы бир манилүү материалы кыргыз тилиндеги этиштер жана алардын өзгөчөлүктөрүн камтыган теориялык маалымат болуп саналат. Бул тууралуу маалымат берүү үчүн Абдувалиев агайдын “Азыркы кыргыз тилинини морфологиясы” аттуу китебине таянып, изилдөө жүргүздүк.

*Заттардын кыймыл-аракетин, ал-абалын, акыл-ойдогу аракеттерди билдирген сөздөрдү этиш сөздөр деп айтабыз. Этиш сөздөр эмне кылып жатат? Эмне кылды? Эмне кылат?* деген суроолорго жооп берет. Этиш деген сөз өзү этимологиясы боюнча *эт-* деген “аткар”, “иште” деген маанини билдирген сөздөн келип чыккан. Бул сөз эскирген сөздөрдүн катарына кирет. Азыркы учурда дээрлик олдонуудан чыгып, кээ бир учурларда гана татаал этиштердин тутумунда гана колдонулуп келет. Мисалы: “кабыл эт-”, “сабыр эт-” ж.б. Кээде башка тилден алынган сөздөр менен бирге оозеки кепте “звонить эт-”, “оформить эт-“ деп колдонулуп келет.

- Кыргыз тилиндеги башка сөз түркүмдөрүнөн айырмаланып, этиш сөздөр чоң рольду ойнойт жана өзгөчө мааниге ээ. Кыргыз тилиндеги этиштердин жалпысынан төмөндөгүдөй өзгөчөлүктөрү бар:
- Башка сөз түркүмдөрүнө караганда этиштер тилдин сөз казынасынын көпчүлүгүн бөлүгүн ээлейт.
- Кыргыз тилиндеги этиштердин башка сөз түркүмдөрүнө өзгөрүшү абдан сейрек кездешет жана оңой эле өзгөрө бербейт. Алар тилде туруктуу лексикалык грамматикалык категорияга ээ. Башка тилдерден этиш сөздөрдүн кыргыз тилинде колдонулушу чанда гана учурларда байкалат.
- Этиштердин грамматикалык категориясы абдан татаал процесс, оңой менен түшүнүп болбойт. Ошондуктан, этиш сөздөр анализ кылууда бир топ кыйынчылыктарды туудурат.

Этиштер ыңгай, чак, сан, мамиле, жана жак категориялары боюнча өзгөрөт. Ошондой эле этиштер оң жана терс формада да болушу мүмкүн. Түзүлүшү боюнча

этиштер сүйлөмдүн объект жана субъектисине карай өтмө жана өтпөс этиштер болуп бөлүнөт.

Жыйынтыктап айтканда, жандуу жана жансыз заттардын мейкиндиктеги кыймылы, иш -аракет болуп өткөн мезгил жана убакыт ченемдери да ар түрдүү жана чексиз. Ушул себептен улам, этиштер да кыймыл- аракеттин, ал-абалдын, аң-сезимдеги кыймылдардын ар түрдүүлүгүн чагылдырып турат. Демек кыргыз тилиндеги этиштер да түзүлүшүнө карай жөнөкөй жана татаал, ал эми татаал этиштер тутумундагы этиштер негизги (маани берүүчү) этиштер жана жардамчы этиштер болуп бөлүнөт. Ал эми татал этиштер өз кезегинде этиш түгөйлүү, атооч түгөйлүү жана тууранды сөз түгөйлүү болуп 3кө бөлүнөт.

Этиш сөздөр жалпы алып караганда, кыймыл аракеттик жана ал-абалдык маанини билдиргени менен кээ бир учурларда кыймыл-аракет байкалса, кээде дээрлик байкала бербейт. Ушул себептен улам да кыргыз тилиндеги этиштер маанилик жагынан кыймыл, абал жана сезим этиштер деген топторго ажырайт.

Ушул классификациялардын негизинде кыргыз тилинен алынган этиштерге жүргүзүлгөн синтаксистик жана морфологиялык анализдердин айрымдарына токтолуп өтсөк:

1. *Хан макул болуп, жаш-карынын баарын* ***чакырды.***

   Хан_n_nom макул_ij болуп_v_iv_prc_perf, _cm жаш_quio карынын_n_gen_sg баарын_prn_ind_px3sp_acc чакырды_v_tv_ifi_p3_sg ._sent

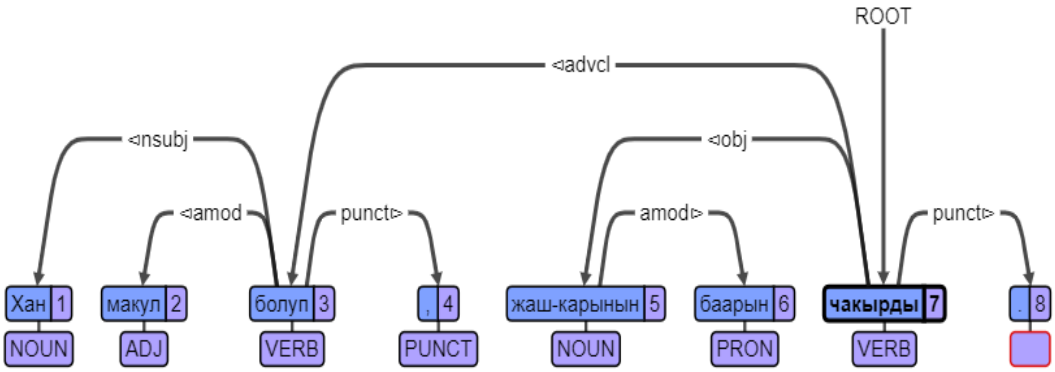



*Figure 4.14 Dependency parsing of the simple verb "чакырды"*

"<Хан>"

"хан" NOUN @nsubj #1->3

"<макул>"

"макул" ADJ @amod #2->3

"<болуп>"

"бол" VERB @advcl #3->7

"<,>"

"," PUNCT @punct #4->3

"<жаш-карынын>"

"жаш-кары" NOUN @obj #5->7

"<баарын>"

"баары" PRON @amod #6->5

"<чакырды>"

"чакыр" VERB @root #7->0

"<.>"

"." @punct #8->7

*2.Коён эсин жыйып, кантип кутулуштун аргасын* ***издеди (The Story by Aldarkoso).***

Коён_n_nom эсин_n_px3sp_acc жыйып_v_tv_prc_perf ,_cm кантип_adv кутулуштун_v_tv_ger_pres_gen аргасын_n_px3sp_acc издеди_v_tv_ifi_p3_sg ._sent

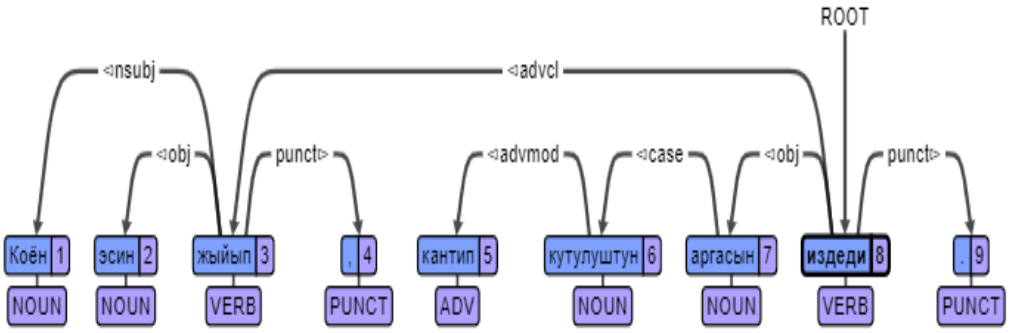



Figure 4. 15 Dependency Parsing of the Simple Verb “издеди”

”<Коён>”

“коён” NOUN @nsubj #1->3

“<эсин>”

“эс” NOUN @obj #2->3

“<жыйып>”

“жый” VERB @advcl #3->8

“<,>”

“,” PUNCT @punct #4->3

“<кантип>”

“кантип” ADV @advmod #5->6

“<кутулуштун>”

“кутулуш” NOUN @case #6->7

“<аргасын>”

“арга” NOUN @obj #7->8

“<издеди>”

“изде” VERB @root #8->0

“<.>”

“.” PUNCT @punct #

Жогорудагы мисалда берилген “***чакырды-invited”, “издеди-searched”*** этиштери бир гана сөздөн туруп, бир гана маанини билдиргендиктен жөнөкөй этиштер болуп саналат. Грамматикалык категориясы боюнча бир ган уңгудан туруп, мүчөсү

чакыр-***ды***, изде-***ди*** биринчи жактын жекелик түрүн көрсөтүп, кыймыл-аракет өткөн чакта ишке ашкандыгын белгилеп турат.

Ал эми төмөндө татаал этиштер берилген сүйлөмдөрдүн анализдери:

3. *Кечке жуук Арсен Саманчин* ***чыдап олтура албады***. (Aitmatov, 1997).

Кечке_n_dat жуук_adv Арсен_np_ant_m_nom Саманчин_np_ant_m_nom чыдап_v_tv_prc_perf олтура_v_iv_prc_impf албады_v_iv_. _sent

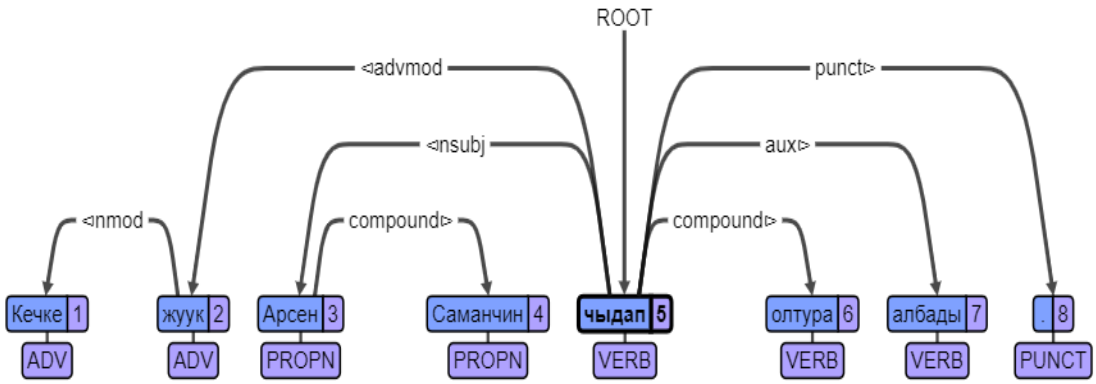



*Figure 4. 16 Dependency parsing of the compound verb "чыдап олтура албады"*

"<Кечке>"

"кеч" ADV @nmod #1->2

"<жуук>"

"жуук" ADV @advmod #2->5

"<Арсен>"

"Арсен" PROPN @nsubj #3->5

"<Саманчин>"

"Саманчин" PROPN @compound #4->3

"<чыдап>"

"чыда" VERB @root #5->0

"<олтура>"

"олтур" VERB @compound #6->5

"<албады>"

"ал" VERB @aux #7->5

"<.>"

"." PUNCT @punct #8-

4. *Сахнага чыккан акындарды эл дуулдата кол чаап , кызуу* ***коштоп турду.*** (Osorov, 2021).

Сахнага_n_dat чыккан_v_iv_gpr_past_subst_nom акындарды_n_pl_acc эл_n_nom дуулдата_unknown кол_n_nom чаап_v_tv_prc_perf ,_cm кызуу_unknown коштоп_unknown турду_v_iv_ifi_p3_sg ._sent

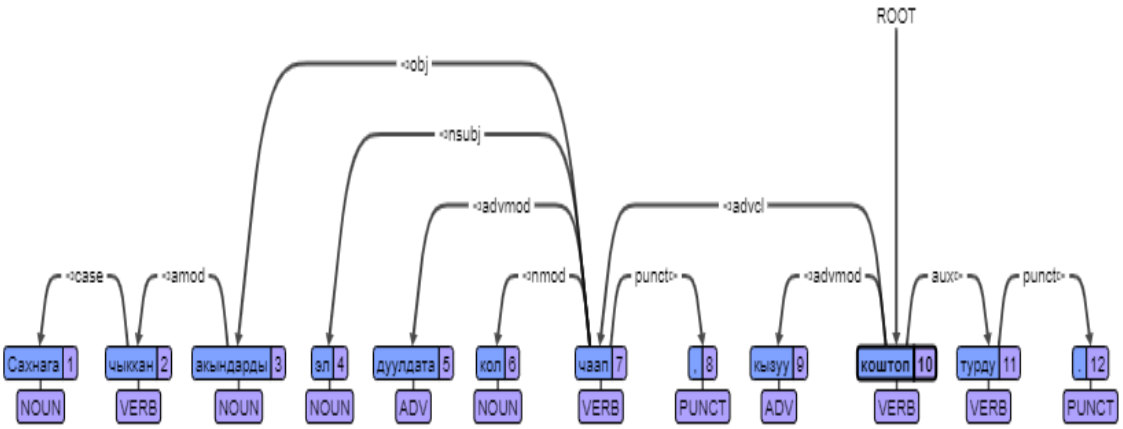



Figure 4. 17 Dependency parsing of the compound verb “коштоп турду”

”<Сахнага>”

“сахна” NOUN @case #1->2

“<чыккан>”

“чык” VERB @amod #2->3

“<акындарды>”

“акын” NOUN @obj #3->7

“<эл>”

“эл” NOUN @nsubj #4->7

“<дуулдата>”

“дуулдат” ADV @advmod #5->7

“<кол>”

“кол” NOUN @nmod #6->7

“<чаап>”

“чаап” VERB @advcl #7->10

“<,>”

“,” PUNCT @punct #8->7

“<кызуу>”

“кызуу” ADV @advmod #9->10

“<коштоп>”

“кошто” VERB @root #10->0

“<турду>”

“тур” VERB @aux #11->10

“<.>” “.” PUNCT @punct #12->1

Сүйлөм тизмегинде татаал этиштер да түзүлүшүнө карабастан жөнөкөй этиштердейэле функцияны аткарат. Татаал этиштерде негизги этиш биринчи, ал эми жардамчы этиш андан кийин ага жанаша орун алат. ***“чыдап олтура албады”, “коштоп турду” татаал этиштеринде “чыдап”*** and ***“ коштоп”*** негизги этиштер, ал эми жардамчы этиштер “***олтура албады” “ турду”*** болуп эсептелет.

5. Караан бери ***жүгүрдү*** (Сынган кылыч).

   Караан_n_nom_sg бери_adv жүгүрдү_v_iv_ifi_p3_sg :_sent

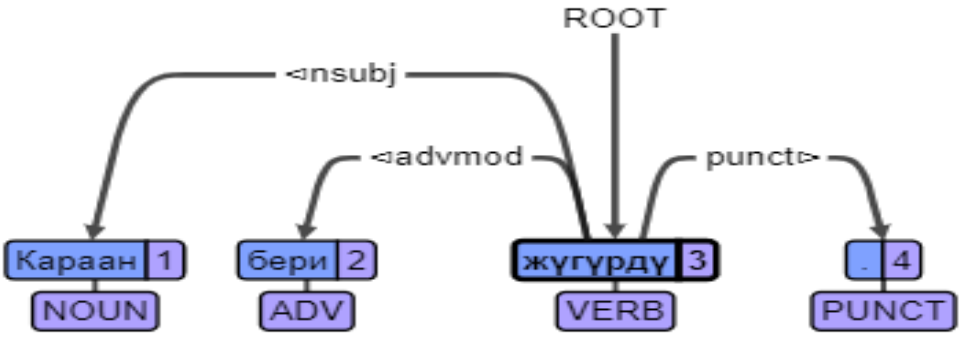



Figure 4.19 Dependency Parsing of the action verb ***“жүгүрдү”***

«<Караан>»
«Караан» NOUN @nsubj #1->3
«<бери>»
«бери» ADV @advmod #2->3
«<жүгүрдү>»
«жүгүр» VERB @root #3->0

“<.>”

“.” PUNCT @punct #4->3

6. Согушка жараамдуулугун бир кыйла ***көтөрдү (***Сынган кылыч).

Согушка_v_tv_ger_pres_dat жарамдуулугун_n_sg_acc бир кыйла_adj_advl көтөрдү_v_tv_ifi_p3_sg ._sent

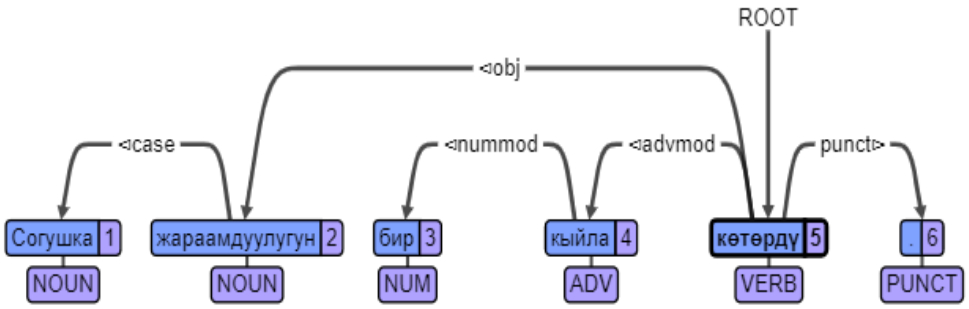



Figure 4.20 Dependency Parsing of the action verb ***"көтөрдү"***

"<Согушка>"

"Согуш" NOUN @case #1->2

"<жараамдуулугун>"

"жараамдуулук" NOUN @obj #2->5

"<бир>"

"бир" NUM @nummod #3->4

"<кыйла>"

"кыйла" ADV @advmod #4->5

"<көтөрдү>"

"көтөр" VERB @root #5->0

"<.>"

"." PUNCT @punct #6->5

Бул мисалдардагы "көтөрдү жана жүгүрдү" делген эки этиш кыймыл этиштерге кирет, анткени экөө тең кыймылды аткарып жаткан субъектинин, кыймыл багытталган объектинин да ал-абалынын өзгөрүшүнө себеп болуп, алардын

кыймылга келишин шарттап жатат. Ал эми түзүлүшү боюнча экөө тең жөнөкөй этиштер катарына кирет.

**Бул илимий иштин түзүлүшү:**

Бул илимий иш негизги *төрт баптан* турат:

*Биринчи бапта* корпустук тил илими, анын тарых-таржымалы, анын классификациясы жана атайын бир критерийлердин негизинде бөлүнүшү, тил илиминдеги корпустар, жана ошондой эле жаңы түзүлгөн кыргыз тилинин корпусунун ((2019-04-18) түзүү жана өнүктүрүү этаптары боюнча кененирээк сөз кылынды.

*Экинчи бапта* болсо сөздөрдүн көп маанилүүлүгүн жоюунун фундаменти болгон жасалма интеллект жана табигый тилди иштетүү тармактары боюнча кенен теориялык изилдөөлөрдүн жыйынтыктары, алардын тил илиминдеги колдонулушунун мисалдары сунушталды.

*Үчүнчү бап* көп маанилүүлүк деген эмне жана алардын түрлөрүн, классификациясын, дегеле көп маанилүүлүктү жоюу процессинин жүрүшүн жана ыкмаларын изилдейт.

Ал эми акыркы *төртүнчү бапта* кыргыз тилиндеги этиштердин кош маанилүүлүгүн жоюу процесси эки жол (морфологиялык жана синтаксистик энтектөө) менен ишке ашты.

**Илимий иштин корутундусу:** бул илимий иште коюлган максаттарга жетүү үчүн төмөндөгү милдеттер аткарылды:

1. Кыргыз корпусунан этиштер (сүйлөм менен бирге) тандалып алынды жана ал сүйлөмдөр англис тилине которулуп берилди.;
2. Тандалып алынган этиштер Абдувалиев өзүнүн китебинде берген классификация боюнча бөлүштүрүлдү:
    - Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *жөнөкөй этиштерге* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *татаал этиштерге* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *негизги этиштерге* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *жардамчы этиштерге* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *татаал этиштердин түрлөрүнө* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *маани берүүчү этиштерге/маанилик топтомдоруна* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *кыймыл-аракет этиштерине* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *ал-абал билдирген этиштерге* морфологиялык жана синтаксистик талдоо жүргүзүлдү;
- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *сезим-туюм этиштерине* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

- Kyrgyz Corpus query processor жана UD Annotatrix annotation куралдарын колдонуп, *өзгөрүм этиштерине* морфологиялык жана синтаксистик талдоо жүргүзүлдү;

3. Жалпысынан кыргыз тилинин корпусунан ар түрдүү жанрдагы тексттерден 32 сүйлөм тандалып алынып, алардын ар бири морфологиялык жана синтаксистик энтектөөлөрү менен кошо камтылды. Ар бир анализделген этиштин синтаксистик дарак форматы сүрөт түрүндө иллюстрацияланып берилди. Ар бир тандалып алынган этиш өзүнүн түрүнө жараша анализделип, жетишкен натыйжалар жана жыйынтытардын ар бирине илимий баа берилди.

Бул иштин жыйынтыгында дагы деле аткарылчу иштер көп экенин баса белгилеп, этиштердин көп маанилүүлүгүн жоюуда ж.б. илимий тилдик изилдөөлөрдө бул иште колдонулган корпустук жана универсалдуу багыныңкылык ыкмаларын башка изилдөөчүлөргө да сунуштайбыз.