KYRGYZ-TURKISH MANAS UNIVERSITY
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
SIMULTANEOUS TRANSLATION DEPARTMENT

PARTS-OF-SPEECH TAGGING OF ADVERBS IN THE KYRGYZ CORPUS

(MASTER'S THESIS)

Aizhan SATYBEKOVA

BISHKEK 2023

ЭТИКАЛЫК ТАЛАПТАР

ПРОТОКОЛ

TEZ ONAYI

КЫСКАЧА МАЗМУНУ

Даярдаган	: Айжан Сатыбекова
Университет	: Кыргыз-Түрк “Манас” университети
Багыты	: Лингвистика (Котормо жана котормо таануу)
Иштин сыпаты	: Магистрдик диссертация
Беттердин саны	: xviii + 122
Коргоо датасы	: 02 / 02 / 2023
Илимий жетекчиси	: Ф.и.к., доц. Аида Касиева

КЫРГЫЗ ТИЛИНИН КОРПУСУНДАГЫ ТАКТООЧТОРДУН ЭНТЕКТЕЛИШИ

Бул магистрдик диссертация кыргыз тилиндеги тактоочтордун морфологиялык өзгөчөлүктөрүн корпустук лингвистика жана табигый тилди иштетүү багыттарында изилдейт. Анткени кыргыз тилиндеги тактоочтордун морфологиялык табияты татаал маселе болуп саналат. Кыргыз тилинде бир гана тактоочтор гана эмес, этиштин өзгөчө формасы болгон чакчылдар да табигый тил иштетүү жаатында камсыздалган сөз түркүмдү энтектөө инструменттери боюнча тактооч катары белгиленет. Кыргыз тил илиминде тактоочторду чакчылдардан ажырата турган касиеттер жана б.у.с. маалыматтар баяндалса да, тактоочтор камтыган көптөгөн маселелер изилденбей келет. Ошондуктан, бул эмгектин максаты кыргыз тилиндеги тактоочтордун табиятын табигый тил иштетүү нугунда карап чыгуу зарылчылыгы пайда болгон. Буга кошумча, кыргыз тили агглютинативдик тил болгондугуна байланыштуу, тактоочторду энтектөө процесси дагы да татаалдашат. Ошентип, бул диссертациянын алкагында биздин максат – 2019- жылы түзүлүп, 2022- жылы жаңыланган Кыргыз тилинин корпусундагы тактоочтордун морфологиялык энтектелиши жана ал үчүн колдонулган белгилер тууралуу кеңири маалымат берип, аларды корпустагы мисалдар менен далилдеп көрсөтүү болмокчу. Тактоочтордун дагы да терең талдоо үчүн, албетте, ал тактоочторду курчаган контекст да эске алынат жана алгач ал тактоочтун кайсы категорияга караганы, б.а. кайсы түрү экендигин аныктайбыз. Андан соң ага ылайыктуу энтекти тандап белгилейбиз. Тактоочтун контексттеги байланышын аныктоо учун, ал сүйлөмдүн синтаксистик багыныңкы байланыштары атайы түзүлгөн веб-тиркемеде аткарылат. Ошентип, сүйлөмдөгү ар бир сөздүн, анын ичинде тактоочтун да байланышы белгиленет жана бул процесстин жыйынтыгында тактоочтун мааниси жана функциясы ортого чыгат. Изилдөөнун максатына жетүү этабында, төмөндөгү методдор жана усулдар колдонулду:

1. Салыштырма анализ: кыргыз жана англис тилдериндеги тактоочтордун окшоштугун жана айырмачылыктарын аныктоо үчүн колдонулду;

2. Квалитативдик анализ: Кыргыз тилиндеги тактоочтордун мүнөздөмөсүн, алардын өзгөчөлүктөрүн жана классификацияланышын сыпаттоо үчүн колдонулду;

3. Квантитативдик анализ: Тактоочтордун категорияларга жана классификациясына жараша алардын кыргыз тилинин корпусунда кезиккендигин жана жыштыгын сандык түрдө көрсөтүү үчүн колдонулду;

4. Корпуска негизделген анализ: Диссертацияда көрсөтүлгөн тактоочтор кыргыз тилинин корпусунан алынып, алардын морфологиялык энтектелиши корпуста кандай көрсөтүлгөндүгү талкууланат;

5. Терең морфологиялык анализ жалаң гана тактоочтордун морфологиялык аспекттерине байланыштуу маселелерди иликтөөгө колдонулду;

6. Буларга кошумча, жогоруда берилген методдордун катарын бул маселелерди чечүүгө арналган атайы веб-тиркемелерди да көрсөтсөк болот. Мисалы, кыргыз тилинин корпусун энтектөөдө жана тактоочторго энтектерди иштеп чыгууда Turkic Lexicon Apertium Toolkit инструменттеринин колдонулушу;

7. Ал эми алардын ситаксистик байланышын багынынкы абалын контекстте көрсөтүү үчүн атайы сунушталган Annotatrix веб-тиркемеси колдонулду.

Натыйжада, бул диссертациянын материалдары кыргыз тилинин корпусунун репрезентативдүүлүгүн жогорулатууга жана аны менен бирге Кыргызстанда корпустук лингвистиканын жаңы тармак катары жаралышына, анын өнүгүшүнө өз салымын кошот деп ишенебиз.

Ачкыч сөздөр: тактооч, энтектөө, Кыргыз тилинин корпусу, аннотация, морфология, синтаксиситик парсерлер, морфология, Turkic Lexicon Apertium.

ÖZ

Yazar	: Aycan SATIBEKOVA
Üniversite	: Kırgızistan-Türkiye Manas Üniversitesi
Anabilim Dalı	: Mütercim-Tercümanlık
Tezin Niteliği	: Yüksek Lisans Tezi
Sayfa Sayısı	: xviii + 122
Savunma Tarihi	: 02 / 02 / 2023
Tez Danışmanı	: Doç. Dr. Aida KASİEVA

KIRGIZCA DERLEMDEKİ ZARFLARIN SÖZCÜK TÜRÜ OLARAK İŞARETLENMESİ

Bu yüksek lisans tezi, Kırgız dilinde zarfların morfolojik özelliklerini derleme dilbilimi çalışmalarında ve doğal dil işleme alanlarında incelemektedir. Çünkü Kırgız dilinde zarfların morfolojik yapısı karmaşık bir konudur. Kırgız dilinde sadece zarflar değil, fiilin özel bir şekli olan belirteç (çakçıl) da doğal dil işleme alanında sunulan kelime grubu etiketleme araçlarında zarf olarak adlandırılmaktadır.Kırgız dilbiliminde zarfları belirteçlerden ayıran özellikler ve bilgilerin sunulmasına rağmen, zarflarla ilgili birçok konu keşfedilmemiş durumdadır. Dolayısıyla bu çalışmanın amacı Kırgız dilindeki zarfların doğasını doğal dil işleme açısından incelemektir. Ayrıca Kırgız dilinin sondan eklemeli bir dil olması nedeniyle zarfların etiketleme işlemi daha da karmaşık hale gelmektedir.Böylece bu tez çerçevesinde amacımız, 2019 yılında oluşturulan ve 2022 yılında güncellenen Kırgız dili derlemindeki zarfların morfolojik sınıflandırılması ve bunun için kullanılan işaretler hakkında kapsamlı bilgi vermek ve bunları derlemden örneklerle göstermektir. Zarfların daha derinden bir analiz yapmak için zarfların bağlam içindeki anlamı da dikkate alınır ve önceden zarfın hangi kategoriye ait olduğu, yani hangi tür olduğunu belirleyeceğiz. Ondan sonra ona uygun etiketleme seçip işaretleyeceğiz. Zarfın bağlam içindeki ilişkisini belirlemek için cümlede sözdizimsel alt bağlantı (bagınıñkı baylanış) olarak oluşturulmuş bir web uygulamasında gerçekleştirilir.Böylece zarf dahil cümledeki her kelimenin ilişkisi belirlenir ve bu işlem sonucunda zarfın anlamı ve işlevi ortaya çıkar. Araştırmanın amacına ulaşma aşamasında aşağıdaki yöntem ve teknikler kullanılmıştır:

Karşılaştırmalı analiz: Kırgız ve İngiliz dillerindeki zarflar arasındaki benzerlikleri ve farklılıkları belirlemek için kullanılmıştır;

Niteliksel analiz: Kırgız dilinde zarfların özelliklerini ve sınıflandırılmasını tanımlamak için kullanılmıştır;

Niceliksel analiz: Kategorilere ve sınıflandırmaya bağlı olarak Kırgız dili derleminde zarfların oluşumunu ve sıklığını nicel olarak göstermek için kullanılmıştır;

Derlem tabanlı analiz: Tezde sunulan zarflar Kırgız dili derleminden alınmış ve morfolojik etiketleme derlemede nasıl olarak belirtildiği gösterilmiştir;

Zarfların morfolojik yönleriyle ilgili sorunları araştırmak için derinden morfolojik analiz kullanılmıştır;

Yukarıda belirtilen yöntemler arasında bu sorunları çözmek için tasarlanmış özel web uygulamaları da gösterebiliriz. Örneğin, Kırgız dili derleminin etiketlemesinde ve zarfların geliştirilmesinde Turkic Lexicon Apertium Toolkit uygulamasının kullanımı;

Sözdizimsel ilişkilerini bağlam içinde göstermek için özel olarak önerilen Annotatrix web uygulaması kullanılmıştır.

Sonuç olarak, bu tez Kırgız dili derleminin temsilciliğinin artmasına ve bununla birlikte Kırgızistan'da derleme dilbilimi yeni bir alan olarak oluşturulmasına ve geliştirilmesine katkıda bulunacağına inanıyoruz.

Anahtar kelimeler: Zarf, etiketleme, Kırgız dili derlemesi, öz, morfoloji, sözdizimsel ayrıştırıcılar,Turkic Lexicon Apertium

АННОТАЦИЯ

Выполнил	: Айжан Сатыбекова
Университет	: Кыргызско-Турецкий университет “Манас”
Специализация	: Лингвистика (Перевод и вереводоведение)
Вид работы	: Магистерская диссертация
Количество страниц	: xviii+ 122
Дата защиты	: 02 / 02 / 2023
Научный руководитель	: Доц. Др. Аида Касиева

ЧАСТЕРЕЧНАЯ РАЗМЕТКА НАРЕЧИЙ В КОРПУСЕ КЫРГЫЗСКОГО ЯЗЫКА

Данная магистерская работа изучает морфологические особенности наречий с точки зрения корпусной лингвистики и обработки естественного языка, так как морфологическая сущность наречий сама по себе является сложным вопросом. В кыргызском языке, не только наречия, но также деепричастия, являющиеся особой формой глагола также размечаются в качестве наречий в соответствии с разработанными инструментами частеречной разметки в рамках естественной обработки языка. Несмотря на тот факт, что критерии различия наречий от деепричастия а также другие особенности наречий уже были описаны в кыргызской лингвистике, еще остаются полностью не изученными другие вопросы касающиеся наречий. В связи с этим, появилась необходимость в рассмотрении сущности наречий в рамках обработки естественного языка что и явилось целью данной работы. В дополнение вышеперечисленному, агглютинативный характер кыргызского языка добавляет сложности процессу частеречной разметки наречий. Таким образом, целью работы в рамках данной диссертации является предоставление детальной информации о морфологической разметке наречий в корпусе кыргызского языка, созданный в 2019 и дополненный в 2022 году наряду с информацией о тэгах, использованные при частеречной разметке наречий. Данная информация о подкреплена примерами извлеченными из корпуса. Для более глубокого изучения наречий, также учитывается контекст использования наречия и определяется к какой категории будет относиться то или иное наречие. Далее, после определения его вида выбирается и определяется правильный тэг для частеречной разметки наречия. Для определения синтаксических зависимостей наречия в контексте, наречия анализируются в специально разработанном веб-приложении. Таким образом, определяется связь каждого слова в предложении, наречия в том числе, и выявляются значение и функция наречия в результате осуществления данного процесса. На этапе достижения цели данного исследования, были использованы следующие методы:

1. Сравнительный анализ: Был использован для определения схожестей и различий наречий в кыргызском и английском языках;

2. Квалитативный анализ: Был использован для описания характеристик, особенностей и классификации наречий в кыргызском языке;

3. Квантитативный анализ: Был использован для предоставления встречаемости наречий в зависимости от вида к которому наречие относится а также частотность встречаемости наречий в цифровых данных;

4. Анализ основанный на базе корпуса: Наречия, предоставленные в диссертации были извлечены из корпуса кыргызского языка и были рассмотрены прикрепленные морфологические тэги для гаречий в корпусе кыргызского языка;

5. Детальный морфологический анализ был использован только для изучения вопросов касающихся морфологических аспектов наречий;

6. В дополнение к вышеперечисленным методам, в качестве еще одного метода можно рассмотреть использование специального веб-приложения, разработанного в целях решения такого рода задач. К примеру, при аннотировании корпуса кыргызского языка и разработке новых тэгов для наречий кыргызского языка был применен инструментарий Turkic Lexicon Apertium Toolkit;

7. Что касаемо отображения синтаксических зависимостей, для этой цели был использовано веб-приложение Annotatrix.

В результате мы надеемся, что материалы данной диссертации будут способствовать повышению репрезентативности корпуса кыргызского языка, становлению корпусной лингвистики как новой отрасли в Кыргызстане а вместе с ним и ее развитию.

Ключевые слова: наречие, частеречная разметка, корпус кыргызского языка, аннотация, морфология, синтаксический парсеры, Turkic Lexicon Apertium

ABSTRACT

Researcher	: Aizhan SATYBEKOVA
University	: Kyrgyz-Turkish Manas University
Department	: Simultaneous Translation
Type of study	: Master thesis
Number of Pages	: xviii + 122
Date of defense	: 02 / 02 / 2023
Research supervisor(s)	: Assoc. Prof. Aida Kasieva

PARTS-OF-SPEECH TAGGING OF ADVERBS IN THE KYRGYZ CORPUS

This master's thesis explores the morphological features of adverbs in terms of corpus linguistics and natural language processing, as the morphological nature of adverbs is a complex issue in itself. In Kyrgyz, not only adverbs but also **"Chakchyls"** which are special forms of verbals (participle/perfect participle), are also marked up as adverbs according to the developed toolkit of parts-of-speech tagging within natural language processing. Despite the fact that the criteria for distinguishing adverbs from "chakchyls" (special verb forms) as well as other features of adverbs have already been described in Kyrgyz linguistics, other issues concerning adverbs have not yet been fully studied. As a consequence, there is a need to consider the nature of adverbs within the framework of natural language processing, which is the purpose of this paper. In this regard, there has appeared a need to consider the essence of adverbs in the framework of natural language processing, which is the purpose of this work. In addition to the above, the agglutinative nature of the Kyrgyz language adds complexity to the process of parts-of-speech tagging of adverbs. Thus, the purpose of the work within this dissertation is to provide detailed information about the morphological annotation of adverbs in the corpus of the Kyrgyz language, created in 2019 and supplemented in 2022, along with information about tags used in parts-of-speech tagging of adverbs. This information is supported by examples extracted from the Kyrgyz corpus. For thorough study of Kyrgyz adverbs, the context of the use of adverbs is also taken into account that determines which category this or that adverb belongs to. Further, after determining its type, the appropriate tag for the parts-of-speech tagging of adverb is selected and identified. To determine the syntactic dependency of adverbs in context, adverbs are analysed in a specially designed web-application. Thus, the relationship of each word in the sentence, including adverbs, is determined, and the meaning and function of the adverb have been revealed as a result of this process. At the stage of achieving the objective of this study, the following methods have been applied:

1. Comparative analysis: It was used to determine the similarities and differences of adverbs in the Kyrgyz and English languages;
2. Qualitative analysis: This was used to describe the characteristics, features and classification of adverbs in the Kyrgyz language;
3. Quantitative analysis: It was used to give the frequency of occurrence of adverbs according to the type to which the adverb refers as well as the frequency of use of adverbs in numerical data;
4. Corpus-based analysis: The adverbs presented in the dissertation were extracted from the Kyrgyz corpus and the attached morphological tags for adverbs in the corpus were considered
5. A detailed morphological analysis was used only to study questions concerning the morphological aspects of adverbs;
6. In addition to the above methods, as another method, we can consider the usage of a special web application designed to solve this kind of tasks. For example, when annotating the Kyrgyz corpus and proposing new tags for the adverbs of the Kyrgyz language, the Turkic Lexicon Apertium Toolkit was used;
7. As for the rendering of syntactic dependencies, the web-application 'Annotatrix' has been used.

As a result, the materials of this dissertation will contribute to increasing the representativeness of the corpus of the Kyrgyz language, the formation of corpus linguistics as a new branch in Kyrgyzstan.

Keywords: adverb, parts-of-speech tagging, corpus of the Kyrgyz language, annotation, morphology, syntactic parsers, Turkic Lexicon Apertium.

PREFACE

The XXI century can be characterized as the time of tremendous and rapid technology development, which has penetrated into every field of our life including language and linguistic study. Thus, today the technological innovations in Linguistics have also become the question of necessity, rather than priority. Corpus linguistics, which is considered as the comparatively new branch of Linguistics that originated in the 1960s, is tightly related to the computational linguistics, i.e., computational processing of the linguistic data on its every level starting from phonetic and ending with registers and discourse. This deals with even younger subbranch of Natural Language Processing (NLP), which is in turn is a subtype the Artificial Intelligence (AI).

Corpus linguistics has started its development in the middle of the past century and nowadays the world linguistic corpora such as (BNC), COCA, TNC have billions and billions of words in volume. This fact clearly reflects the level of the development of corpus linguistics among other world languages. Speaking in terms of the corpus linguistics among Turkic languages, the largest in volume is considered to be the BOUN Corpus containing more than 400 million of words and token, while other turkic languages are a bit humble in volume.

Due to the endless efforts, on the basis of the project agreed between the Kyrgyz-Turkish Manas University and Universität des Saarlandes (Germany), The Kyrgyz Corpus was created in 2019, containing 1 243 161 morphologically annotated words and has been expanded up to 2 millions in 2022. This dissertation is a part of both this project and computer processing of the Kyrgyz language that is intended to improve the quality and representativeness of the corpus by investigating annotation features of one of the complicated parts of speech in Kyrgyz, such as Adverbs and its types.

Thus, this work is devoted to the thorough study of adverbs and their grammatical peculiarities along with selecting their morphological annotation taking into account adverb features in Kyrgyz. Appropriate morphological annotation of adverbs in the corpus allows ensuring more accurate tagging taking into consideration their syntactic and semantic nature. Therefore, additional tags for different types of adverbs have been proposed in this work to identify their additional grammatical categories in order to allow computers to recognize the Kyrgyz language as correctly as possible.

First of all, I would like to express my sincere gratitude to my supervisor Aida Kasieva for her excellent guidance, generosity, and valuable advice. Without her, I would not have been able to complete this work. Also, I would like to express my gratitude to professor Elke Teich, the head of Translation studies and English linguistics at the Universität des Saarlandes (Germany). I express my gratitude to the computer

linguist and researcher Jörg Knappen, the person in charge for the technical side of building the Kyrgyz Corpus, from the Universität des Saarlandes (Germany) for the realization of such a great project and continuing contributing to it up today.

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 performing this thesis.

Being given an opportunity, I would express my deep gratitude to Kyrgyz-Turkish Manas University for ensuring its students with such a great in quality of education, opportunities and conditions, that in its own turn makes a huge contribution to the personality and educational background of its students.

Bishkek, 2023 Aizhan Satybekova

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The XXI century is tightly related to such processes as globalization and technological development all over the world in all spheres of routine activity. Nowadays we have crossed the threshold of the age of technology as it is of great importance. The issue of digitalization and full transfer to the technological solutions including textual material are among the problems arousing in the daily basis. Thus, helping computers to understand the Kyrgyz language is the priority task. This can be achieved through combination of such fields as computer engineering and artificial intelligence, which are parts of machine-learning process. Bearing in mind all these, this dissertation paper intends to make contribution in corpus-based study of adverbs in the Kyrgyz corpus and its Natural Language Processing by providing linguistic information about Kyrgyz adverbs, their morphological annotation and parts-of-speech tagging description.

Motivation for the research

This dissertation is a corpus-based research study that is intended to make contribution to the development of the Kyrgyz Corpus by providing relevant and evident description of adverbs in Kyrgyz. Taking into account grammatical peculiarities of adverbs in English and Kyrgyz, along with different standard symbols that are suggested by various recognized platforms, this study considers relevant annotation guideline for adverbs in the Kyrgyz corpus. Adverbs in the Kyrgyz language are studied in detail to accelerate the quality of the corpus and further corpus development. That process is of great importance since their morphological annotation is a baseline for the further analyses on the syntactic and semantic levels of the linguistic data of any corpus. The topic of the dissertation is motivated and encouraged to study particularly adverbs and in the practical part of the thesis, the possible variants of tagging adverbs are suggested to provide the Kyrgyz corpus

with the precise tagging information about adverbs in Kyrgyz. It is performed to show the syntactic relations of adverbs within a sentence and to observe whether adverb is tagged appropriately. The results of the thesis are directly related with the ability of computer programs to recognize the Kyrgyz language.

The topicality of the research is that adverbs have been studied on the basis of the corpus by applying corpus-based and corpus-driven approaches.

Living in the XXI century implies the mandatory knowledge of the very basics of using computers (not to mention programming language), with which human interacts tightly and intensely throughout the last two decades. Being in the condition of dynamically developing technological lifestyle, the knowledge of computer technologies is getting even more acutely necessary. In such a way, by providing morphological analyses for the Kyrgyz adverbs within the Kyrgyz Corpus, we are intending to contribute to Machine Learning of the Kyrgyz language. Appropriately annotated corpus on its all linguistic levels will form the base for integrating the huge machine-translation systems with the Kyrgyz language. The quality of machine-translation is directly dependent from the quality of annotation provided in the corpus. Thus, this issue is really topical nowadays as machine-translation applications have become extensively popular due to their wider use in every sphere of life.

The objective of the research is to study adverbs in the Kyrgyz corpus and their morphological annotation along with their peculiarities and features. Thus, it is a corpus-based study. As the nature of the Kyrgyz adverbs is complex, the aim of the dissertation is to ensure appropriate tagging for adverbs in Kyrgyz as sometimes “ambiguities” or some obstacles at the stage of part-of-speech annotation to work out proper variants for future annotation works in the Kyrgyz Corpus. To achieve the goal, the following objectives have been set:

a) To study the basic concepts applied in the Corpus Linguistics;
b) To consider the aspects of adverbs in the Kyrgyz language;

c) To provide the pre-processing of adverbs for the Kyrgyz corpus;
d) To provide quantitative (statistical) data on the adverbs and their annotation in the Kyrgyz Corpus;
e) To conduct syntactic analyses of adverbs using the tool of Universal Dependencies to reveal its position and relation with other words within the sentence;
f) To provide appropriate tags for adverbs, taking into account their additional aspects in the Kyrgyz corpus;
g) To compile a list of "tags" (obtained in the course of analyses) that can be suggested for the morphological annotation of adverbs for their processing in the Kyrgyz corpus

The subject of the research: adverbs extracted from the Kyrgyz Corpus.

The object of the research: morphological annotation of Kyrgyz adverbs and the issues related to their annotation in the corpus.

The scientific novelty of the research. The scientific novelty of the research is related to the fact that Kyrgyz adverbs are being investigated from the point of view of corpus linguistics and is closely connected with Natural language processing (NLP). The thesis is devoted to the comprehensive studies of Kyrgyz adverbs with the application of such new methodologies and approaches, as used in Corpus linguistics and Universal dependency platform by applying web-based platforms (Turkic Lexicon Apertium) and Universal Dependencies. The strength of the corpus approach lies in the combination of frequency data for all words in the corpus and the linguistic environment in which those words occur. This combination allows a detailed examination of typical usage patterns of lexis and grammar.

Corpus-based approach helped to retrieve the statistical data from the corpus regarding the adverb, allowing to conduct quantitative analysis while Universal

Dependency tool ensured the implementation of the scrupulous analysis of adverbs and their functions on the syntactic level.

Methodology applied in the research:

a) Quantitative method was applied to expose the statistical data regarding quantity of adverbs;
b) In order to present the frequency data, the corpus-based approach is used;
c) Selective method is applied to demonstrate the Kyrgyz adverbs and their complexity by extracting them from the body of the Kyrgyz Corpus;
d) Comparative analysis is used to describe, count and compare the features of adverbs within the Kyrgyz Corpus;
e) Descriptive method is applied to reveal the main concepts of Corpus Linguistics and the main features of adverbs;
f) The Universal Dependency analysis tool has been used for analyzing adverbs on the syntactic level.

The hypothesis of the research. The study of the adverbs of the Kyrgyz language within the frame of Corpus linguistics provides the appropriate morphological tagging through the syntactic dependency parsing.

Theoretical and practical significance of the thesis is the analysis of the adverbs extracted from the body of the Kyrgyz Corpus, study adverbs in the Kyrgyz language. On the basis of the analysis, it is intended to put forward proposals on fulfilling the annotation of existing adverbs in the corpus by suggesting new tags of the additional categories of adverbs for their morphological annotation.

The practical importance of the paper is to present new methodology for teaching foreign languages, such as Data-Driven Learning, where the focus is made on the study

of a language as data and students are the researcher who accomplishes the task in accordance with the guideline. The method includes data processing, using statistical methods, analysis of the results obtained.

CHAPTER I – CORPUS LINGUISTICS

1.1.Theoretical Background of Corpus Linguistics

The term "corpus linguistics" is not yet clearly defined. The history of corpus linguistics begins in his mid-twentieth century (1950s), although the term itself was not created at that time. The term “corpus” (“corpora” in plural) designates “body” from the Latin. Currently, this means a representative collection of texts of a particular language, dialect, or other subset of language used for linguistic analysis. More specifically, it refers to (a) (loosely) any body of text; (b) a body of machine-readable text (in general); and (c) (more accurate) A limited collection of machine-readable text chosen as representative of a language or language variety (Botley, 2000).Corpus encloses a huge collection of representative samples of texts covering different language variants used in different areas of language interaction. It is computer compatible, can be used for research and application, is representative of the source language, can be processed by humans and machines, has unlimited data, and is systematic in construction and representation. (Dash N. , 2008: 145-160p.).

However, up today there is no clear distinct definition for the corpus linguistics. As the corpus linguistics includes not only the linguistic data, but also can be used as a tool for conducting numerous linguistic studies in different aspects. Thus, the question may arise whether corpus linguistics is a ‘method’, ‘methodology’, ‘approach’, ‘theory’ or ‘paradigm’? Various linguists have various views regarding this term. Charlotte Taylor in her article namely “What is corpus linguistics? What the data says” enumerates the number of corpus theorists, who shared their opinions regarding this term (Taylor, 2008).

Taylor Charlotte, the senior lecturer in English language and Linguistics at the University of Sussex, enumerates the views of the number of corpus linguists regarding

the ‘corpus linguistics’. She devoted an article, namely “What is Corpus Linguistics? What the data says” to this topic. She says that the corpus linguistics could be considered as a tool, methodology, discipline, linguistic theory or in a combination of all of these. Supporters of the view of corpus linguistics as a “methodology” are considered to be McEnery, Tono, Xiao, Wilson and Meyer. They share following opinion: “corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning” (McEnery T. R., 2005). That view is also supported by Thompson and Hunston, claiming: “basically, corpus linguistics is a methodology that can be aligned to any theoretical approach to language”. While Stubbs (Stubbs, British traditions in text analysis: From Firth to SInclair., 1993) and Teubert are representing the corpus linguistics as a “theoretical approach to the study of language”. But we tend to consider corpus linguistics as the study of linguistic phenomena with the help of large collections of unified, annotated at different linguistic levels machine-readable texts, namely **corpora**. These are used in a wide range of purposes, like compilation of linguistic dictionaries, in the sphere of translation and translation studies, language learning, machine learning etc.

So, at present the term “linguistic corpus” can be defined as a large, machine-readable, unified, structured, tagged, philologically competent linguistic data set for specific linguistic tasks (Zakharov, 2013). The standardized representation of verbal material on a machine-readable medium allows the use of standard programs for its processing. So, there are following main criteria of a linguistic corpus according to which it is built and should meet the requirements:

1.1.1 Notable Features of Corpus.

- Quantity: It must be large and contain a large amount of data in verbal or written form. Size is actually the sum of the components that make up the body.
- Quality: All text should come from concrete spoken and written language.

The role of linguists here is very important. He has to make sure that the voice data comes from normal communication and not from experimental conditions or artificial environments.

- Representativeness: It should include examples from a wide range of texts. Maximizing the representation of linguistic diversity requires a balance across all domains of language use. Future analyzes developed on this will require validation and authentication of information from the corpus representing the language.
- Simplicity: It should enclose plain texts in plain format. In other words, we expect an unbroken string of characters (or words) unannotated on any of the linguistic levels within texts. Any kind of annotation containing various linguistic and non-linguistic information contrasts with simple plain text.
- Equality: The samples used in the corpus should be of unified size. However, this is debatable and not universally accepted. Sampling models can be significantly modified to make the corpus more representative and multidimensional.
- Retrievability: Data, information, examples, and references should be easily retrievable from corpus by the end-users. This pays attention to preserving techniques of language data in electronic format in computer. The present technology makes it possible to generate corpus in PC and preserve it in such way that we can easily retrieve data as and when required.
- Verifiability: The corpus should be open to all kinds of empirical validation. A corpus data might be used for any type of validation. This makes corpus linguistics a step ahead of intuitive approaches to language learning.
- Augmentation: It should be increased periodically. This makes the corpus "equivalent" and registers the linguistic changes that occur in language over time. Over time, with the addition of new linguistic data, the corpus gains a historical dimension for diachronic researches and for presenting linguistic cues to stop changes in life and society.
- Documentation: Complete component information should be stored

separately from the text itself. It is always recommended to separate document information from text and only include a minimal header that cites the document. For corpus management, this effectively separates plain text and annotations with little programming effort.

All of the above-mentioned criteria are of great importance especially on the very initial stages of compiling and building the linguistic corpus, consisting of procedures of Natural Language processing (NLP hereinafter), which provide the technical software of the linguistic corpora. Some steps performed within the NLP.

1.2.Brief history of Corpus Linguistics

The technologies used in corpus linguistics are much older than electronic computers: many of them have their roots in the tradition of the late XVIII and XIX centuries, when linguistics was first proclaimed a "real", or empirical science. The process of collecting texts in order to study the language in use cannot be characterized as the new one due to the fact that people began to compile the list of all the words in a certain text together with their contexts in the Middle Ages. (Wonner, 2022). One of the main influences on modern corpus linguistics came from comparative-historical linguistics. This is not surprising, since linguists engaged in historical research have always used texts or collections of texts as primary evidence. In general, the period of development of the corpus linguistics can be divided into 4 large time slots:1st – pre-digital starting XIII century and ending 1960s; 2nd - since 1960 till 1990, when appeared the first “digital corpora” of a little scale; 3rd period falls between 1990-2000 years, when the corpus linguistics starts to develop by leaps and bounds: and the last, 4th period starts since 2000 and up to now can be called as an era of “gigacorpora”. The era of pre-electronic enclosures began in the 13th century and ended by the early 1960s. The notion of "corpus" in its linguistic meaning emerged only towards the end of the

pre-electronic era, as it recognized a discrete religious or literary work or a collection of works by a single author for which a concordance was compiled manually and assembled mainly for theological, literary, and lexicographical research. (Solnyshkina M., 2020) Many pre-digital corpora were associated with the scriptures of various religions, and the most researched of these was the corpus of biblical texts. The Bible-based word lists with verse indication were called concordances (symphonies). A. Cruden calls a concordance a dictionary or index to the Bible in which all the words used in the "divinely inspired scriptures" are arranged alphabetically and the place in which the word is used is also indicated, so that one can easily find a verse with this word and compare the several meanings in which it is used (Solnyshkina M.I., 2020). All concordances of the pre-electronic era were different from those of today and represented some sort of indication of where a word or phrase was used. The first concordance was compiled in the thirteenth century by Friar Anthony of Padua for the fifth-century Latin version of the Vulgate Bible. This concordance was called "*Concordantiae Morales*". (Michael McCarthy, 2010)

After the publication of the first edition of Cruden's concordance in 1737, concordances to the works of great writers began to be compiled on the same principle. Thus, an important work for the development of corpus linguistics was the "Concordance to the works of W. Shakespeare in all editions" (1,787) by A. Beckett, since in addition to information about the place of use of a word (play, act and action), it contains an excerpt of the work in which this word was used was presented (Beckett, 1787).

The 18th and 19th century corpora were mostly dictionaries, and the period itself was characterized by the development of lexicography. As early as the late nineteenth and early twentieth centuries, corpora began to be created for the purpose of linguistic research or, more frequently, for practical purposes (e.g., to calculate the frequency of language units). At the turn of the XIX and XX centuries, several projects were organized to collect empirical material for lexicographic purposes. Based on them, the

Dictionary of the American Version of the English Language was compiled under the editorship of Noah Webster (Noah Webster's “An American English Dictionary”) (1828) and The Oxford English Dictionary (The Oxford English Dictionary, OED) (1884). To create the research base of the Oxford Dictionary, two thousand volunteer readers collected about five million citations totaling about 50 million word uses in order to illustrate the meanings and usage of 414,825 words in the dictionary. Based on the collected texts of the English dialect speech of J. Wright compiled a "Dictionary of English Dialects" The English Dialect Dictionary (1898-1905) (Kennedy, 1998). A turning point in the history of the development of concordances was the development of a methodology for using keywords in the system Keyword Out of Context (KWOC) or Keyword in the name (1856) by A. Crestadoro for the systematization of catalogs in the State Library of Manchester. In 1958, H.P. Lund refined this technique and introduced "keywords in context" into the computer selection technology (KWIC), according to which the keyword was located in the center, and the concordance lines could be placed to the left or right of the keyword, including the necessary context. (Korycinski, 1992). So, this can be nominated as the threshold between the pre-digital and digital eras in the framework of corpus linguistics.

The electronic concordance “*Index Thomisticus*” with a total volume of more than 10.6 million word uses, created by the monk R. Busa for the works of Thomas Aquinas, was the first work in which elements of machine text processing were applied (Busa, 1980). Concordance was created for five years: from 1962 to 1966. For the convenience of working with concordance and its brevity, R. Busa decided to present in it as a keyword only the lemma, or the title word, with all its word forms. To do this, he carried out the lemmatization of texts, which took place in two stages: combining all word forms with inflections under one lemma and attaching a code with the corresponding part of speech for each lemma and its word forms. Lemmatization was carried out on the basis of the Latin machine Lexicon Electronicum Latinum dictionary, which R. Busa and ten priests compiled for two years. The electronic dictionary was a

table with lemmas, on the basis of which the computer carried out the lemmatization of texts. This method of working on the basis of an electronic dictionary or list later largely determined the principle of electronic text processing. The first volume of the "Index Tomisticus" was published in 1973 and over 40 volumes of the "Index Tomisticus" with alphabetical indexes, word frequency tables, etc. were published in the 1970s (Svartvik, 2007). According to Leech (Leech, Corpora and theories of linguistic performance, 1992) it was in the 1950s, in the era of American structuralists such as Harris, Fries and Hill, among others, when the concept of collecting real data came into its own. While the work of early biblical and literary scholars provided the basic method of word search and indexing, structuralists were the forerunners of the corpus not only in the sense of data collection, but also about the commitment to bringing real language data into focus of the linguists. McEnery and Wilson (McEnery & Wilson, 1996) use the term "early corpus linguistics" to refer to all works on corpus linguistics performed before the appearance of Chomsky. This covers the period of time until the late 1950s, when structuralism was the main linguistic science. Well-known linguists of the structuralists tradition, such as Boas, Sepir, Bloomfield and Harris, use methods for analyzing language that can undoubtedly be called corpus. Though there were foundations and prerequisites for the emergence and the development of the corpus linguistics as it is nowadays, it has gone through a lot of stages mentioned above in addition with the criticism expressed mainly by Chomsky in the middle of the twentieth century, where the concepts of rationalists and empiricists have been clashed. Rationalist theory, in the case of linguistics, builds on the development of theory of mind and has cognitive plausibility as a fundamental goal. The goal is to develop a linguistic theory that not only mimics the external influences of human language processing, but actively claims to represent how language processing actually occurs. Empirical approaches to language, on the other hand, are dominated by observing naturally occurring data, usually through a corpus. In this case, it is possible to determine whether sentence x is a valid sentence in language y by searching a corpus of

language questions and gathering evidence for sentence grammaticality or to the contrary. (McEnery & Wilson, 1996, p. 5)

Thus, in the pre-electronic era, all the prerequisites for the transition to electronic era buildings were created. The first concordances were developed, which were understood as synonymous with dictionaries and pointers. The first concordances were of great importance for the further development of corpus linguistics, since in the composition of the concordance article, the indication of the desired word, the place of its use, the context of the use of fixed language units were considered mandatory. In addition, a system of context illustrations in the concordance "keyword in context" was developed. The buildings lacked a single principle of collecting texts, uniform rules for composing concordances. Their volume and sources also varied greatly: the corpus could be texts of sacred books (Bible translations, works of theologians), as well as individual works of fiction. From a modern point of view, such texts are not corpora, but archives or collections of individual texts. The term "corpus" itself was also missing.

The electronic era of the development of corpus linguistics starts from the 1960s up to nowadays. Johansson argues that, despite the works of R. Busa already published in the 1960s and the appearance of the first electronic corpus, scientists became actively interested in corpus linguistics only in the 1970s (Johansson S. , 2008). In his view, true corpus linguistics emerged in the 1970s with the creation of the first laboratories and centers where linguists and programmers began to work on general linguistic problems and ways of processing texts. Centers of computer linguistics aiming at corpus text collection, storage and processing have been established in Italy, The United States, England, Germany, Canada, France, Sweden and Norway. However, the very first electronic corpus is considered to be the **Brown Corpus of Standard American English**, that is the first of the modern, computer readable, general corpora. It was compiled by W.N. Francis and H. Kučera, Brown University (The Brown Corpus, 2022). This corpus consists of one million American English texts printed in 1961. The

corpus texts are selected from 15 different text categories, such as: newspaper articles, academic writings, advertisements, hobby books, religious literature, biography, essays, fiction (detectives, adventure and westerns, popular science fiction, love affairs, feuilletons) making the corpus an excellent standard reference. The texts in the "Brown Corpus" were printed on a punched card, which contained information about the location of the text, its name, as well as the number of lines in the text. Nowadays, this corpus is considered small and slightly outdated. However, the corpus is still used. Much of its usefulness lies in the fact that the Brown corpus layout has been copied by other corpus compilers.

In 1968, F. Bagley first introduced the term "metadata" to denote all data about texts in the corpus (Nguen T.H., 2014). Since the mid of 1960s, the first KWIC-based concordance programs have appeared, like: COCOA, COunt and COncordance Generation Atlas) (1967) и «Collocations» (CLOC, CoLOCation) (Stubbs, Notes on the History of Corpus Linguistics and Empirical Semantics , 2007). When they were created, machine text processing was accompanied by manual markup, i.e., "attaching" a code (or tag) to a text unit with information about it (Baker, 2006). Automatic text markup began to be talked about when, in 1971, B. Green and J. Rabin wrote a program for automated text markup TAGGIT, the first approbation of which was the markup of the Brown corpus. TAGGIT carried out markup using 86 tags that highlight significant and service words, punctuation marks and individual morphemes in the text. The program "did not remove homonymy", and 23% of the words in the corpus were marked with several tags at the same time (McEnery T., 2012). In 1978, Alvar Ellegard carried out a syntactic mark-up of part of Brown's corpus by hand. After several years of checks and corrections, the work on the partial part-of-speech tagging of the Brown Building was completed in 1979. B. Green and J. Rubin published all the data on the morphological analyzer TAGGIT so that other scientists could refine and improve it (Johansson S. , 2008, p. 46). Scientists consider the end of the 1970s to be the time of official recognition of the term "corpus linguistics" (Svartvik, 2007, p. 2). The Brown Corpus" has become the standard for compiling corpora

both in terms of volume and in terms of the range of styles and genres of written speech represented in it. With the publication of the "Brown Corpus" in the mid-1970s, similar cases began to appear, first in the UK, then in other countries. For example, in 1976, The joint corpus of the Universities of Lancaster, Oslo and Bergen (The Lancaster-Oslo-Bergen corpus (LOB) (1961-1978) was published (Lars Engwall, 2019).

The 1970s became decisive in the development of corpus linguistics: centers and laboratories for the development of electronic text processing tools appeared. The KWIC technique made it possible to systematize the form of concordance representation, later the first concordance programs appeared, such as COCOA (COunt and COncordance Generation Atlas) and CLOCK (CoLOCation). Electronic processing of corpora posed the scientists with the problem of the accuracy of electronic text processing, which gave good results only together with manual markup. So, this period can be called as “the time of corpora of the first generation”.

By the mid-1970s, with advances in technology and the consequent availability of sound recording, corpora began to form for a wider range of research purposes. During this period, the modern interpretation of the meanings of such terms as "corpus", "corpus linguistics", "markup", "meta-markup", "concordancer", "morphological analyzer" was fixed. When studying oral speech, the terms "tokenization", "tokens", "segmentation", "normalization", "time alignment" appeared (Solnyshkina M.I., 2020, p. 13). In the 1980s the morphological text analyzer CLAWS (Constitutional Likelihood Automatic Word-tagging System) was developed. At that time, the CLAWS system correctly marked up about 95% of the Brown corpus. It used a probabilistic approach. The system went through a number of changes before the final version of CLAWS 4 was received. The CLAWS 1 system was based on a hidden Markov model and was able to correctly mark up from 96% to 97% of any text. (Makhmutova, 2018). By these steps, corpus linguistics comes to start the development of the corpora of the second generation.

In the early 1980s, the text markup language, or the metalanguage SGML (Standard Generalized Markup Language), was developed, which is a set of tags that standardizes the markup of texts (Baker, 2006, p. 149). With the advent of numerous corpora based on

different types of texts, it became necessary to have a single set of rules containing all the rules for collecting, transcribing, and annotating texts of both oral and written discourse. In 1987, at a conference at Vassar College in Poughkeepsie, New York, the Text Encoding Initiative (TEI) community was founded, which posed the problem of developing uniform standards for the compilation, transcription and markup of corpora (Burnard, 2013). So, exactly this community has taken the responsibility for creating the set of rules for the compiling and annotating the corpora. However, they weren't the only one concerned with this issue. In 1991, the non-profit 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-printable characters (transcription, mathematical formulas, etc.) (Solnyshkina M.I., 2020, p. 144). Attempts to standardize corpus compilation were also made by the Expert Advisory Group on Language Engineering (EAGLES) (1993), which proposed its own standard for collecting and marking up texts in the Corpus Encoding Standard (CES) (1998), which was based first on the SGML markup language (1998), currently the XML markup language – XCES (2000) (Baker, 2006, p. 50).

One of the most ambitious projects developed in the late 1980s was the Collins Birmingham University International Language Database (the International Language Database at the University of Birmingham and Collins Company), or the COBUILD Corpus. The corpus was created by a group of scientists under the leadership of J. Sinclair. (The history of COBUILD, 2022). The project uses the so-called Birmingham Collection of Texts, which includes 20 million word uses of written and spoken texts. The main corpus had a volume of 7.3 million words and the so-called "reserve corpus" a volume of 13 million words. The corpus consists of 75% of written texts and 25% of texts of the spoken speech. The COBUILD corpus contains texts published in the period from the 1960s to 1982. Written speech is mainly represented by prose fiction texts. The corpus contains oral codified speech, in which only prevalent, non-specialized vocabulary is used. 75% of oral speech is the speech of men over 16, and 25% is the speech of women. 20% of the corpus consists of texts from the American version of English. According to S. Johansson, the

COBUILD project was a breakthrough for its time for several reasons: 1) the corpus exceeded 20 million words; 2) the sources were full texts rather than short fragments; 3) it was the most representative and included oral and written texts of different genres. COBUILD became the most comprehensive corpus of its time and formed the basis of The Collins COBUILD Dictionary of English (1987) (The history of COBUILD, 2022).

Another megacorp, the formation of which was started in the late 1980s by a group led by D. Summers, is a network of corpora of the Longman publishing house, The Longman Corpus Network. This Corpus Network is currently a commercial database consisting of five main corpora: 1) The Longman Corpus of Learners' English (containing 10 million of words); 2) The Longman Written American Corpus (100 million of words); 3) The Longman Spoken American Corpus (5 million of words); 4) The Longman/Lancaster English Language Corpus (30 million of words); 5) The Spoken British Corpus (10 million of words). Each part of The Longman Corpus Network has been assembled for a specific purpose; the combined corpus has become a powerful tool that captures a wide variety of texts from different genres of speech produced by native and non-native speakers of English. This type of corpus was used to create dictionaries and textbooks on communicative English grammar. (Solnyshkina M.I., 2020, бет 18).

Thus, second-generation corpora are corpora with a volume of at least a hundred million word in use, the aim of which is to represent the full diversity of written and spoken speech. The compilers sought to present as many genres and styles of oral and written speech of various segments of the population as possible. As a rule, these are corpora available online, assembled and placed according to TEI requirements. National corpora became monitor-based and were compiled on the basis of the principles of representativeness of the selection of texts and according to the rules, characteristic of second-generation corpora.

The third-generation corpora are belonging to the beginning of the XXI century, that can be marked by the following technological progress: the fourth-generation concordancers of the BNCweb (2009), CQPweb (2012), SketchEngine (2031) have been developed that are partly analogous to the concordancers of the previous editions.

Concordancers of the fourth generation were developed to solve the following problems: limited power of personal computers, incompatibility of personal computer operating systems and legal restrictions on the distribution of corpora. To solve the legal issues and simplify 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 tool (McEnery T., 2012). The tendency to increase the volume of corpora continued after the 2000s. A. Mauranen, S. Kubler and H. Zinsmeister (Kuebler, 2015) characterize this generation by the motto "the bigger the enclosure, the better", and L. Flowerdew is the first to call this epoch the era of the Gigacorpus generation1 (Flowerdew, 2004). At this time, a number of new corpora appeared (COCA, Google Books Ngram), the volume of which amounted to several billions of words. The large volume of the corpora made it possible to carry out frequency studies on a larger scale and to study collocations consisting of three, four or more words. Subsequently, these collocations were called n-grams, where bigrams are collocations consisting of two words, tri-grams are collocations consisting of three words, and n-grams are collocations consisting of n words (Rayson P. , 2015).

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 contains spoken and written texts. (Solnyshkina M.I., 2020). The COCA corpus is dynamic and is added by 20 million words each year. Parts-of-speech annotation of texts is done with the CLAWS program. The corpus is accompanied by the WordAndPhrase concordance program. This corpus will be used while studying the adverbs in the practical part of the thesis. Apart of this, there were created and fulfilled a great number of other corpora, reaching the index of billion of words within the corpora.

Thus, this period is characterized by the fusion of corpus linguistics methods with the World Wide Web: there were created programs that automatically download texts from the Internet. In addition, it became possible to track the development of the use of a word on large data sets, such as changes in the form and meaning of a word over time in written or spoken speech. The emergence of corpora with large amounts of texts has not

diminished the relevance of the question of the necessity and representativeness of small corpora of professional speech.

1.2. Classification of Corpora

The classification scheme I propose here is a reasonable range at the moment. It provides a rational way of classifying the corpus as much as possible using clearly delineated categories. Various taxonomy criteria are applied to the corpus, subcorpus and their related components. Linguistic criteria can be external and internal. External criteria are most often mapped onto the corpus from text typologies dealing with participants, opportunities, social settings, communicative functions of language, etc. Internal standards deal with repetition of speech patterns within language. Considering all these aspects, we roughly classify the corpus as follows: Genre of text, Nature of data, Type of text, Purpose of design, and Nature of application.

1.2.1. Text types Included into the Linguistic Corpora

Written Corpus: A written corpus (*e.g., Kyrgyz 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*) includes all formal and informal discussions, debates, previous conversations, spontaneous analysis, casual and regular conversations, dialogues, monologues, various types of conversations, instant public speaking, and others.

Spoken Corpus: A speech corpus (e.g., London-Lund Corpus of Spoken English) is a technical extension of the speech corpus and contains spoken text. In such corpora, speech is represented in written form without any alterations other than transcription. Annotation is implemented through some form of phonetic transcription.

1.2.2. Data Type within the Linguistic Corpora

General Corpus: General corpus (e.g., *British National Corpus*) consist of common texts belonging to different disciplines, genres, subject areas and registers. The number of text collections is limited due to the nature of their format and usefulness. This means that the number of text types and the number of words and phrases they contain are limited. It has the ability to grow over time, with new data being added as new text becomes available. It is very large, diverse, widely represented and has a wide range of uses.

Special Corpus: Special corpus (e.g., *CHILDES Database*) is designed from texts sampled in general corpus for specific variety of language, dialect and subject with emphasis on certain properties of the topic under investigation. It varies in size and composition according to purpose. Its high percentage of unusual features prevents it from adding to the description of a language. Since it collects information from people who don't act normally, its origin is not trustworthy. When used for other purposes than what it was intended for, the special corpus presents a distorted and "skewed" view of language segments because it is unbalanced (outside of the parameters of its intended use). Since it uses a variety of typical, genuine language, it is different in principle. Corpora of languages used byr children, non-native speakers, dialect users, and people from special areas of communication (e.g., auctions, doctor visits, gambling, court proceedings) are considered to be Special corpora. Its main benefit is that texts are chosen so that the phenomena one is looking for occur more frequently in it than in a balanced corpus. A corpus enriched in this way will be smaller than a balanced corpus that provides the same kind of data. (Sinclair, 1991).

Sublanguage corpus: It is made up of just one type of text of a certain language. It is on the opposite side of the linguistic spectrum of the reference corpus. Because of the homogeneity of its structure and specialized lexicon, only a small amount of data is needed to show typical good and closure properties.

Sample corpus: Sample corpus (e.g., *Zurich Corpus of English Newspapers*) is one of the categories of special corpora, consisting of samples containing a limited collection of carefully selected and extensively studied texts. A sample corpus is never expanded upon or altered after it has been created, (Sinclair, Corpus, Concordance, Collocation, 1991) because doing so would upset its balance and compromise the research requirements. Compared to texts, samples are scarce and always the same size. As a result, they are not considered texts.

Literary corpus: The literary corpus is a special category of sample corpus. Among the classification criteria taken into account for the creation of such a corpus there are author and genre (e. g. poetry, fiction, and short stories, among others), period (e.g. as well as the 15th and 18th centuries), group (e.g. authors of Victorian novels, Augustan prose, Romantic poets, etc.) themes (e. g. industrialization, family stories, revolutionary writings, etc.) as valued parameters, as well as other concerns.

Monitor corpus: The Monitor corpus (e.g. Bank of English) is a growing, non-exclusive collection of texts with an ever-expanding range of data that reflects language change. The constant growth of the corpus reflects linguistic change and does not affect the relative weights of the components defined by the parameters. The same composition scheme is adopted annually. The base of the monitor corpus refers to texts of spoken or written language over a period of one year (Sinclair, 1991).

With the help of the monitor corpus, we can discover new words, monitor usage variation, spot meaning changes, establish a long-term norm for frequency distribution, and glean a wealth of lexical data. Due to the availability of new data sources and the development of new procedures, the composition of a monitor corpus changes over time, making previously scarce materials more abundant.

1.2.3. Text Types in the Linguistic Corpora

Monolingual corpus: It includes representative texts from a single language that

show how it was used over the course of several different periods. It includes both written and spoken text samples as long as their interaction with one another and proposed research by the investigators is not impeded. (Botley, 2000).

Bilingual corpus: When corpora from two related or unrelated languages are combined, a bilingual corpus is created. If these languages are genetically or typologically related, they become parallel corpora, where texts are aligned in accordance with predetermined rules. It is not allowed for parallel corpora to have different sizes, contents, or fields.

Multilingual corpus: Multilingual corpus (e.g., *Crater Corpus*) contains a representative collection from two or more languages. In general, the same text categories and sampling methods as in bilingual corpora are used here, even if the texts belong to different languages.

1.2.4. Division According to the Annotation Criterion

Un-annotated corpus: It (e.g., *Kyrgyz Corpus in Sketch Engine*) represents a straightforward raw state of plain texts devoid of any additional linguistic or nonlinguistic data. Although the utility of the corpus is greatly increased by annotation, it is still extremely useful in language study.

Annotated corpus: It (e.g., *British National Corpus* contains tags and codes that corpus designers have added from outside to record additional information (analytical marks, parts-of-speech tags, grammatical category information, etc.) into texts. Annotated corpora, as opposed to unannotated corpora, are better suited to providing pertinent information useful for a variety of language technology tasks, such as morphological processing, sentence parsing, information retrieval, word sense disambiguation, machine translation, etc.

1.2.5. Purpose of Application of the Linguistic Corpora

Translation Corpora: The majority of original texts in the source language and their translations into the target language make up translation corpora. These corpora typically maintain the meaning and function of words and phrases across languages, making them the perfect foundation for comparing the realization of specific meanings in two different languages under the same circumstances. Additionally, they enable the discovery of all cross-linguistic variants, i. e. alternative interpretations of specific meanings and concepts. As a result, translation corpora offer more useful tools for developing the rules needed for translation as well as for analyzing cross-linguistic data (Aijmer, 2000).

Aligned corpus: It is a kind of bilingual corpus (for example, The Canadian Hansard Corpus), where text in one language and translations in another are matched sentence by sentence, phrase by phrase, or word by word.

Parallel corpus: Parallel corpus (e.g., *Chemnitz German-English Corpus*) contains texts and translations in each relevant language, allowing you to double-check equivalent translations. The text in one language and its translation into another language are aligned sentence by sentence, phrase by phrase, even word by word. A cross-parallel corpus may be designed to include a corpus containing canonical text and translations for each language.

Reference corpus: (e.g., *Bank of English*) It is designed to provide thorough knowledge about a language. It can be used to write grammars, dictionaries, thesauruses, and other materials because it is big enough to represent all pertinent varieties of language and defining vocabulary. It is put together using pertinent criteria that the linguistic community has agreed upon. It includes verbal and written language that reflects different social and situational registers as well as formal and informal speech. It serves as a "benchmark" for lexicons, the effectiveness of general tools, and language technology applications. Reference corpus is used to gauge the deviance of special corpus due to the

increasing influence of internal criteria. (Bouillon, 2001).

Comparable corpus: (e.g., *Corpus of European Union*) a collection of texts that are "similar" but are written in different languages or dialects. There are texts there that are not identical in content, genre, or register and are written in various languages. These are employed to compare various linguistic systems. There is disagreement regarding the nature of similarity despite the fact that it has the same composition pattern and there are few examples of comparable corpora. It is essential for linguistic comparison and the creation of bilingual and multilingual lexicons and dictionaries.

Opportunistic corpus: A low-cost collection of electronic texts that can be obtained, converted, and used for free or very cheaply is known as an opportunistic corpus. These texts are frequently unfinished and incomplete. Users must therefore complete any gaps on their own. The opportunistic corpus is a virtual corpus in that the needs of a specific project determine which actual corpus is chosen (from the opportunistic corpus). Generally speaking, the monitor corpus is an opportunistic corpus.

Additional specification types include *closed corpus, synchronic corpus, historical corpus, dialect corpus, idiolect corpus, sociolect corpus*, and others. As a result, the classification system presented here is not definitive and final. It is open to reclassification as well as sub-classification based on various criteria.

1.3. Technology of Building the Linguistic Corpus

There are many problems with corpus design, development, and management. First of all, there is the traditional question of representativeness and balance of the linguistic material underlying the dictionaries and grammars created on the basis of the corpus. Depending on the type of corpus and the intended use, different issues with corpus development and processing may arise. Concerns with the development of speech corpora are distinct from those with the development of text corpora. The creation of a speech

corpus involves a number of considerations, including the intended use, the choice of informants, the setting, the method of data collection, the size of the corpus, the problem of transcription, the type of data encoding, the management of data files, the editing of input data, the processing of texts, and the analysis of texts, etc. (Butler, 1992). The size of the corpus, text representation, choice of time period, choice of documents, and collection of text documents (e.g., magazines, books, and newspapers.), a technique for data sampling (e.g., arranging gathered materials in accordance with one's needs), entering data in a certain way (e.g., arbitrary, predictable, selective, etc.), ccorpus sanitation (e.g. correction of typographical errors, omission of foreign words, quotations, dialects, etc.) all of them are the significant criteria taken into account while compiling a linguistic corpus.

1.3.1. The Technological Stages of Creating a Linguistic Corpus:

1. Determining the list of sources;
2. Digitization of texts (conversion into computer form);
3. Text preprocessing. At this stage, all the texts received from different sources undergo philological verification and correction. Bibliographic and extralinguistic descriptions of the text are also prepared.
4. Conversion and graphematical analysis. Some texts also go through one or more stages of pre-processing, during which various kinds of transcoding (if necessary), removal or conversion of non-text elements (drawings, tables), removal of hyphenation, "hard line endings," ensuring uniform writing of dashes, etc. are done. As a rule, these operations are performed automatically. This is usually the time when the text is segmented into its structural components.
5. Markup text. Partitioning of the text is to attribute additional information (metadata) to texts and their components.
6. The next step is to correct the results of automatic markup: correct errors and

remove ambiguities (manually or semi automatically).

7. The final stage is conversion of marked-up texts into

8. The final stage is conversion of marked-up texts to the structure of the specialized linguistic information retrieval system (corpus manager), providing a quick multi-aspect search and statistical processing.

9. Finally, providing access to the corpus. The corpus can be accessed within the display classroom, can be distributed on

10. CD-ROM and can be accessed in a global network mode.

The goal of the corpus creators is to collect as many texts as possible belonging to the subset of the language for the study of which the corpus is being created. But the important thing is not only and not so much the quantity of linguistic material, but its proportionality. The most important concept of corpus linguistics is *representativeness*. Representativeness of the corpus (in relation to the problem area) is a requirement for the composition and volume of the corpus and its most important property. By representativeness is understood the ability of the corpus texts to reflect all the properties of the problem domain relevant to the type of linguistic research, in a certain proportion determined by the frequency of the phenomenon in the problem domain. However, the issue of size eventually loses significance in the context of text sample representation. A large corpus may not always accurately represent a language or variety in comparison to a smaller corpus that is properly balanced with texts from all languages. L. Flowerdew writes that representativeness must be considered more important than corpus volume, and for corpuses of written professional speech the volume can range from 20,000 to 250,000 words (Flowerdew, 2004)**.** It is always preferable to place emphasis on the quality of data or text samples rather than the quantity or amount of data in a corpus. In a corpus, data from every possible domain of language use should be fairly represented. In order to achieve proper text representation, the size and balance of the corpus as a whole must be balanced against the variety of source texts.

As we noted above, to solve various linguistic problems, it is necessary for texts and separate linguistic units within texts to contain additional linguistic and metalinguistic information - markup (annotation). Modern corpora, in addition to meta-labeling (reflecting various extra-linguistic information about the text, including its title, author, genre affiliation, etc.), contain markup corresponding to different levels of linguistic description - morphological, syntactic, phonetic, etc. The choice of these data should be guided by the purposes of the study and the needs of linguists, as well as by the possibilities of introducing those or other additional features into the text.

1.3.2. Corpus Processing

After a corpus is created, corpus processing becomes necessary. Systems, tools, techniques, and software must be developed in order to access the language data and extract pertinent data from the corpus. Corpus processing is essential for language technology projects as well as for traditional linguistic research and development.

There are many corpora processing techniques, including statistical analysis, concordance, lexical collocation, key-word search, local word grouping, lemmatization, morphological processing and generation, chunking, word processing, parts-of-speech tagging, annotation, parsing, etc. It has been noted that the results of corpus processing frequently defy common assumptions about a language's characteristics. For English, French, German, and other languages that are similar, there is a wide variety of corpora processing software available. Incorporating the nature of languages into the design of corpus processing tools for our own languages is essential.

1.3.3. Basic Natural Language Processing Procedures:

Tokenization, Lemmatization, Stemming and Parsing

Tokenization (or grapheme analysis), i.e., breaking down the stream of characters in natural language texts into separate meaningful units (tokens, word forms), is a necessary condition for further text processing. If languages had perfect punctuation, tokenization would be no problem - even the simplest program could divide a text into words, guided by spaces and punctuation marks. But in reality, languages do not have this kind of delimitation which makes the task of tokenization more difficult. For example, there are cases in English that cannot be unambiguously tokenized. Cf.: the line *"chap"* can be an abbreviated form of the word *"chapter"* or the word "*chap"*, which is located at the end of a sentence. The line *"Jan."* can be regarded as an abbreviated form of the word *"January"* or as a proper name placed at the end of a sentence. In the first case, the dot should be assigned to the same token, as the word, in the second case it should be a separate tag. It should be noted that many small but often important problems must be solved at the tokenization stage, and that many applications that process text often ignore difficult cases (e.g., handling hyphens, accounting for abbreviations and compound words, complex words, spacing, polysyllabic units, etc.) or handle them with special algorithms in the search procedure. (Greene, 1971) .

Lemmatization is another specific task related to morphological markup, i.e., the procedure of forming the original word form for word forms in a text. In many languages, a word can occur in several forms with different inflections. For example, the English verb *"walk"* may be represented by the following forms: *walk, walked, walks, walking*. The basic form, *"walk",* recorded in the dictionary, is called *"lemma".*

The process, somewhat different from lemmatization, is called ***stemming;*** it consists in finding the stem (base) of a word. The difference is that a stemmer does not produce a normal form, but it only tries to cut off the changeable part of a word, leaving

the base for further processing. Below are examples of stemming and lemmatization. The following sentence is given:

[The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs].

One of the most popular stemmers, Snowball Analyzer, produces the following stems:

[quick] [brown] [fox] [jump] [over] [lazy] [dog]. Lemmas of this sentence will be as follows: [the] [quick] [brown] [fox] [jump] [over] [the] [lazy] [dog]. Stemmers are usually simpler to implement and process data faster, but show lower accuracy. For example, the token better corresponds to the lemma good, but stemming is likely to generate stem bett. For a number of applications this may not be decisive, but in corpus linguistics, lemmatization usually takes place. Stemmers work quite well on texts of agglutinative languages, while for inflectional languages, particularly Slavic languages, their results are no longer as good. (Halteren, 1999).

Parsing involves matching a linear sequence of a language's lexemes (words, tokens) with that language's formal grammar. A dependency tree (syntax tree) or component tree is typically the end result. One of the most crucial areas of computational linguistics is the creation of automatic parsers (parsers) for large corpora. Along with various statistical programs that “train” on manually marked up parsers, many parsers use approaches based on contextual (linguistic) rules that model specific linguistic theories and grammatical and syntagmatic rules for constructing texts. The development of parsers is closely intertwined with the development of these theories. Because the most sentences are ambiguous in any theory based on rules (or list of restrictions), a strategy should be developed removal of ambiguity. Many disambiguation strategies rely on quantitative data - the frequency of a given structure in given corpus, sampling constraints for given lexical units that have been derived or extracted from corpus data, etc.

1.3.4. Corpus Annotation and its Types

To solve various linguistic problems, it is not enough just to have an array of texts. It also requires that the texts explicitly contain various kinds of additional linguistic and extralinguistic information. This is how corpus linguistics came up with the idea of a tagged corpus. Tagging (annotation) is the process of assigning specific tags to texts and their constituent parts. These tags are extralinguistic, external, and include information about the author, the text, and its genre and subject. Information about the author may also include the author's name, age, gender, and other details. There is also another term, namely "meta-data". The term "meta-data" refers to the coding of information that includes linguistic, structural (word form, chapter, paragraph, sentence, etc.), and structural (chapter) information. The annotation on the morphological level is called part-of-speech tagging (or POS tagging), and can be useful, for example, in distinguishing words which have the same spelling, but different meanings or pronunciation (Leech, Developing Linguistic Corpora: A Guide to Good Practice; Adding Linguistic Annotation, 2004). The set of these metadata largely determines the opportunities that corpora provide to researchers. The choice of this data should be guided by the purposes of the study and the needs of linguists, as well as by the possibilities of introducing certain additional features into the text. Among the linguistic types of annotation there are following division:

- **Morphological annotation,** also known as Parts-of-Speech tagging (PoS-tagging). The goal of part-of-speech annotation is to give each lexical unit in the text a code that identifies its part of speech. It improves the specificity of data retrieval from a corpus. In actuality, morphological tags encompass both grammatical categories specific to this part of speech as well as the attribute of part of speech. The majority of large corpora are simply morphologically annotated corpora; morphological analysis is seen as the foundation for further forms of analysis, such as syntactic and semantic; and at last, developments in computer morphology make it possible to automatically annotate large corpora. The annotation data are presented in some structured form and include: lemma,

attribute of part-of-speech, attributes of grammatical categories. In 1980, an annotated version of the Brown corpus appeared, in which word forms were lemmatized, their surface-syntactic functions marked, etc. (Zakharov, 2013, p. 34)

- **Syntactic annotation** – The annotation of corpora for syntactic analyses with parse trees followed part-of-speech tagging (Gries & Berez, 2017). On the basis of the results of the morphological analysis, a syntactic annotation, or parsing, is carried out. The syntactic relationships between lexical units and different syntactic constructions are described by this type of tagging (e.g. g. verb phrase, dependent clause, etc.). In a simple sense, syntactic annotation (or parsing) is a process by which we identify phrases used in a sentence, classify them, mark out their types, assign a specific phrasal identity to them, see how they are syntactically linked, and explore how they generate different layers of information in a sentence (Dash N. S., 2021). The term "treebank" refers to a parsed corpus and alludes to the tree diagrams that are employed during the parsing process. Rarely does corpus annotation contain a visual representation of the tree structure. Sets of labeled brackets are commonly used to represent identical information. Therefore, *Pearl sat on a chair* will appear in treebank in the manner described below:

[S[NP Pearl_NP1 NP][VP sat_VVD [PP on_II [NP a_AT1 chair_NN1 NP] PP] VP] S]

where constituents are denoted by opening and closing square brackets and phrase type e is annotated at the beginning and end, and morpho-syntactic information is attached to words by underscore characters. g. [S dot. S]. A full parsing scheme in this system aims to provide an in-depth analysis of sentence structure, whereas a skeleton parsing scheme typically uses a less finely distinguished set of syntactic constituent types and ignores the internal structure of some constituent types. Automatic parsing has a lower success rate than part-of-speech annotation, so post-editing by human analysts is most frequently used. Full manual parsing has the drawback of inconsistent editing or corpus parsing by the analyst(s) involved. More thorough guidelines are offered to combat this, but even then,

ambiguities may still arise because different interpretations are open to debate. (Boguraev, 1996).

Syntactic annotation is one of highly complex forms of text annotation. It is used on a written corpus after the words of a corpus are passed through the stages of morphological annotation. Mostly, syntactic annotation is carried out with the help of syntactic parsers. **Syntactic parsing** is the automatic analysis of syntactic structure of natural language, especially syntactic relations (in dependency grammar) and labelling spans of constituents (in constituency grammar) (Wikipedia, 2022).

Unlike the morphological annotation, ways of representing syntactic structure and syntactic relations are not so unified. There is a variety of syntactic theories and formalisms:

1. Dependency grammar;
2. Phrase structure grammar;
3. The structural grammar;
4. Traditional syntactic doctrines of sentence members;
5. Functional grammar;
6. Syntactic semantics interface, etc.

But in the practical part of the work, we will turn to the universal dependency structure, for analyzing the adverbs on the syntactic level and comprehending better their peculiarities.

- **Semantic annotation.** In semantic annotation, either the semantic properties of individual words in a text (basically, word sense annotation) or the semantic connections between individual words in a text are marked. Semantic annotation is applied to a piece of a text to annotate the most appropriate sense a word denotes in the context of its use (Löfberg, 2003). Semantic annotation is one of the most useful methods applied in the word sense disambiguation as well as in attaching

semantic domains to words based on their use in texts, the specification of the meaning of words, the resolution of homonyms and synonyms, the categorization of words, the allocation of thematic classes, signs of causativity, evaluative and derivational characteristics, etc. Regarding which semantic features should be annotated, there is no universally accepted standard.

One of the variants of semantic annotation:

1) category (proper name, reflexive pronoun, etc.);

2) lexical and semantic characteristics (thematic class lexemes, signs of causativity, estimates, etc.);

3) derivational characteristics ("diminutive", "subjective adverb", etc.).

The semantic annotation scheme takes the POS annotated corpus as input text

assigns specific semantic information to words to convey their general meaning field of words. Most often this task is performed using a single lexicon words and multi-word units where the main meaning of the word is predefined (Rayson P. S., 2008).

- **Anaphoric annotation:** All pronouns and noun phrases are co-indexed within a large framework of cohesion in the anaphoric annotation scheme. If such systems correctly defined the reference (the relationship between a word and a thing that is designated by this word, that is its referent), substitute pronouns would yield more accurate results. The study and testing of mechanisms like pronoun resolution use such an annotation scheme. Both machine translation and text comprehension depend on it.

- **Prosodic Annotation:** The patterns of intonation, stress, and speech pauses are identified through prosodic annotation. Since prosody is much more impressionistic than other linguistic levels, it is a more challenging type of annotation. A trained ear is needed for attentive listening. The elements which are subject to prosodic annotation are following: syllables, pauses, lengthening, pitch variation, accent variation, loudness, tonal

variation and silenceIn the Lancaster/IBM Spoken English Corpus, stressed syllables with and without independent pitch movement are identified using various symbols. All unstressed syllables are left unmarked, even if their pitch can be predicted from the tone marks of the accented syllables around them.

- **Discoursal Annotation**: A corpus is annotated at the textual and discursive levels and used for linguistic analysis in discoursal annotation. Despite its potential use in discourse analysis, this type of annotation has never been widely employed. This may be because linguistic categories depend on their context and their identification in texts is more contentious than other linguistic phenomena Some have added 16 discourse tags to the London-Lund Spoken Corpus in order to track trends in adolescent speech (Johansson S. &., 1996).

Needless to say, not all types of information annotated in text are necessary for all types of applications and uses of the corpus. Additional levels of annotation can be freely added to the corpus, even if it is already annotated with other levels of information. The final finding is that the higher the level of information annotated in the corpus, the greater the usefulness of the corpus across domains and disciplines. A corpus, which is annotated at different levels with different types of information, is indeed a treasure island that is ready to serve the needs of text users coming from different walks of life.

1.4. About the Kyrgyz Corpus

The Kyrgyz corpus has a web-based corpus management system to help you navigate the data and extract the necessary information. The first steps towards the creation of the Kyrgyz language corpus were made as part of the DAAD exchange program between Saarland University and Kyrgyz-Turkish "Manas" University under the direction of Prof. Elke Teich. In the initial stage, it consisted of 1,205,888 words in 84 written texts of mostly literary genres (epics, novels, stories, fairy-tales)

while in the august of 2022, due to the fruitful and hard work of two Universities, the project started enhancing.

The results of the longitudinal work indicate that the Kyrgyz POS tagging present interesting outcomes and challenges to be modeled with them. Taking into account the peculiarities and specifics of Kyrgyz POS, the Apertium toolkit was used in POS tagging of the Kyrgyz corpus to adjust morphological and syntactic features. As a result, a more detailed model for marking parts of speech for Kyrgyz sentences in the Kyrgyz corpus is proposed. In other words, this step served as a prerequisite for more accurate and detailed labeling, which allows to avoid morphological ambiguities in the words of the corpus, which can be caused by the phenomenon of agglutinative complexity of the Kyrgyz language.

The formation of a corpus is an urgent task for many modern communities, since not only codified languages, but all its implementations in all styles and genres, can correspond to the state status of a language. As Plungian**,** one of the creators of the National Corpus of the Russian language, points out, the corpus is an effective and useful tool, especially when the Corpus is large in volume and complete in coverage, **i.e.,** it is a National Corpus of Language (Plungian, 2003:483-484p.). A corpus is, in a first approximation, a collection of texts in a given language, presented in electronic form and provided with a scientific apparatus. The apparatus *"built"* into the corpus is called *"markup"* or "*annotation"*; the better the corpus, the completer and more perfect its annotation. As international experience shows, the creation of a corpus is a long, time-consuming process that requires the efforts of many centers and institutes, supported by government programs and information resources. Most of the major languages of the world have their own national corpus, varying in completeness and the level of scientific treatment of the texts.

At the moment a research project by language teachers is exploring the creation of corpus linguistics. The creation of the corpus will make it possible to solve many problems using information technology: to study the history of the Kyrgyz language, to teach and learn the Kyrgyz language on the basis of the virtual corpus, to conduct statistical monitoring of the functioning of lexical, grammatical and stylistic language means, to work on lexicographic support of the modern Kyrgyz language, its *standardization, codification, to create dictionaries, textbooks, reference books, to conduct statistical analysis* of various language units.

At the moment Kyrgyz Corpus contains 2,493,894 words including 1019 different texts of various genre with the morphological annotations. Not all the words out from this are annotated or even the annotated ones require improving and adding some new, correct annotation tags.

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

Figure 1.4.1. The Kyrgyz Corpus – metadata for the corpus

Deduction on Chapter 1

Thus, the corpus, which is a reflection of the language, represents, as well as the language itself, a dynamically developing system that presupposes all new criteria and approaches to the description and analysis of language material and the development of new methodological procedures. The corpus may contain detailed information about a particular language, but it is impossible to assemble a corpus covering the entire language, because it is impossible to collect all the samples of the use of this language, therefore, one should always assume that the corpus is just some finite collection of samples of the infinite universe of the language. It can be concluded that the corpus is a representative array of unedited texts presented in electronic form, usually marked up for analysis for linguistic purposes, provided by a relatively easy-to-use search system, representing the maximum number of language variants. If during the period of the origin of corpus linguistics, researchers pointed out that language variability could be neglected, then with the advent of electronic corpora, the variety of forms of language existence became more evident and the possibilities of studying language data expanded. The extensive typology of the corpora, created taking into account various criteria and their diversity allows both the linguist and the non-professional user to choose the one that corresponds to the goals and objectives of a specific and specific scientific research.

The Kyrgyz language, its peculiarity and uniqueness, its history and modernity, its future, the possibility of cardinal changes are all important and topical not only for linguists, but also for specialists in many fields: *culture, economy, history, politics*, etc., including linguistics. Even though right now we have only small and “raw” version of the Kyrgyz Corpus, little-by-little that corpus will also become a really representative, balanced and containing billions of words in its core.

CHAPTER II

THE PECULIARITIES OF ADVERBS WITHIN THE CORPUS LINGUISTICS

2.1. Adverbs in the Kyrgyz Language – Adverbs in the Kyrgyz Corpus

The initial data regarding the adverbs in the Kyrgyz language have almost emerged in the 30s and 40s of the XX century, when Batmanov devoted a whole chapter for the issue of adverbs in the Kyrgyz language. (Davletov S. , 1980). He has pointed out the obstacles related to the issue of partitioning the adverbs from other parts-of-speech in the Kyrgyz language as the role of adverb could be played by adjectives, nouns and some forms of verbs and shortly stopped at each of them. One more prominent work dedicated to the adverbs in the Kyrgyz language is considered to be the article of Kudaibergenov S., "Taktoochtor" issued in 1949. But this article was insufficient in terms of classifying adverbs as an independent part-of-speech and didn't reveal the characteristics of adverbs on the syntactic level.

The adverbs in the Kyrgyz language have been recognized as an independent part-of-speech since 1950s amidst the school grammar textbooks. Since then there were given several variants for the definition of an adverb in the Kyrgyz language, but all of them were in some way not accurate on sufficient level. The very initial step in considering the adverbs separately as a part-of-speech belongs to the prominent Kyrgyz linguist Kudaibergenov. However, another linguist Davletov S. has conducted specialized researches devoted to the adverbs as well as established the main distinctions of adverbs from other parts-od-speech and proved that adverbs can function in Kyrgyz language as an independent and separate part-of-speech (Abduvaliev, 2015). According to Davletov S., and Kudaibergenov S., adverbs are the words, which indicate the various states of an

action, mostly immutable on the morphological level, modifying mainly verbs, implementing the function of complements within the sentence (Davletov S. K., 1980).

According to its composition, adverbs are divided into two groups – inflectional and derivational. The first group includes itself the adverbs designating their function by their original form, without adding any suffixes and prefixes (i.e morphemes), for instance – *"бат" "bat"* (fast), *"жай" "jai"* (slow), *"төмөн" "tömön"* (down), *"ылдам" "yldam"* (rapid), *"жакшы" "jakshy"* (good), *"азыр" "azyr"* (now), *"зорго" "zorgo"* (hardly) *"эрте""erte"* (early) etc. While the derivational adverbs are formed with the help of adding one or more derivational morphemes, i.e., in the root of the adverb we can observe several derivational morphemes or two or more roots (Abduvaliev, 2015, p. 278). Exempli gratia: *"адам+**ча**" "adam+**cha**"* (like a human), *"кытай+**ча**" "kytai+**cha**"* (chinese), *"бала+**ча**" "bala+**cha**"* (like a child), *"чака+**лап**" "chaka+**lap**"* (by a bucket). In the Kyrgyz language, most of the adverbs are considered to be of the derivational origin due to the fact, that according to their external form they do coincide with another part-of-speech. For instance: *"капчап" "kaptap"*(by sacks), *"менчилеп" "menchilep"*(like me), "Динарачылап" "Dinarachylap" (like Dinara) etc. in these adverbs the affix "-ып" "-уp" is designating the gerund as well as the affix ***"-ый" "-уі" ("-ла""-la", "-й" "-i"***), e.g., *"учурлай" "uchurlai"* (in the moment), *"ылдамдай" "yldamdai"* (getting faster), "*жазды-жаздай" "jazdy-jazdai"* (lit. whole spring), *"жайды-жайлай" "jaidy-jailai"* (whole summer)is also showing the presence of the gerund form. This kind of words are functioning as adverbs within the context only (Akmataliev, 2015).

Being considered as an independent part-of-speech, it should be mentioned the main features of adverbs that distinct them form other part-of-speech. They are as follow:

1. The meaning of the adverbs comes to be abstract in the term of lexicon, modifies the situation (of place, time, reason, purpose, measure or the quality) of an action/movement. It can be demonstrated in this instance: "*Жакшы бала жакшы окуйт", "Jakshy bala jakshy okuit"* (A good boy studies well). In this sentence the first word

“жакшы” “jakshy” was relevant to the object of the sentence, when the second appearing of it modified the predicate of the sentence, indicating the quality of the action.

2. In terms of morphology, the adverbs in Kyrgyz do possess the affixes, applying to them only, as an example can be previously mentioned affixes **“-cha”** and **“-lap”**. Bearing in mind that the adverbs can appear to be adjective in some circumstances, the ability to join the affixes to its root can service as a distinction mark of an adverb from the adjective, for example: *“Жаңы заманда эл таалайы ачылды” “Jany zamanda el taalayi achyldy”* – “New times have brought prosperity to the people”, where the word *“жаңы” “jany”* is functioning as an adjective, while in the following case: *“Жаңы заманда эл жаңыча жашай баштады” “Jany zamanda el janycha jashai bashtady”* (At new times, people started to live in a new way) the word *“jany”* has added the affix *“-cha”* and turned into adverb, modifying the predicate *“жашай баштады” “jashai bashtady”* (started to live). It also should be mentioned that adverbs are not conjugated according to morphological categories, such as: cases, number, person, tense, i.e., they are unchangeable by these categories. (Davletov S. K., 1980, p. 195)

3. On the syntactic level, they go in conjunction with the verb, in most cases they do function as an attribute within a sentence. E.g. – *“Кубатбек алдына койгон максатын ишке ашырыш үчүн күндүр-түндүр иштеди”, “Kubatbek aldyna koigon maksatyn ishke ashyrysh üçün kündür-tündür ishtedi”* – “Kubatbek worked day and night to achieve his goal”. So, here the adverb *“күндүр-түндүр” “kündür-tündür”* (day and night) is modifying the verb *“иштеди” “ishtedi”* (worked), forming the verb phrase in the sentence structure and playing the role of attribute (Sartbayev, 1958).

2.1.1. Formation ways of the adverbs

Adverbs in the Kyrgyz language take place among the essential parts-of-speech, due to the fact that being entirely independent part-of-speech they played their own role in enriching the lexicon in the language and are the one from others who have wide range of usage in the speech. The basic of the adverbs is made up with the inflectional adverbs. But there are plenty enough of other adverbs, that have appeared by adding derivational affixes. The one of the founders of Kyrgyz linguistics, Turkologist, researcher of Manas, public figure Yunusaliev B.M. has shared his view regarding the origin of adverbs as follow: "Adverbs are considered to be transitive part-of-speech. They were formed from other part-of-speech to adverbs through lexical, morphological and syntactic means". He continues: "There are no any special affixes that could have helped to form adverbs. The category of adverbs, appeared morphologically are formed with the help of other different affixes belonging the other part-of-speech." (Karymshakova, 2009).

2.1.2. Morphological way of forming adverbs

Conventionally, adverbs in the Kyrgyz language can be formed in two ways – morphologically and syntactically. Morphologically formed adverbs are constructed by adding the derivational affix to the base. This kind of adverbs are internally subdivided according the part-of-speech they have been formed from.

1. Adverbs formed from the nouns. In the formation of the adverbs from the nouns following affixes are used:

a) **-cha;** This affix is common to the whole Turkic languages, not only to Kyrgyz. This affix is also considered to be the one that forms the adverb in other Turkic languages as well. With the help of this affix, newly formed adverbs designate the similarity of the action that is incorporated into the base of a word to the action of some other thing. For example: *"бүркүт**чө** шаңшуу" "bürküt**chö** shanshuu"* (to whistle in a high pitch as an eagle), "*Ит**че** шыйпаңдап", "it**che** shyipandap"* (to

wag its tail). If the affix ***-cha*** is attached to the word, that nominates a particular nation or people as well as coincide to the context of the words used in the speech, then this word will designate the meaning of "in the way of particular this or another people's" which was mentioned in the base of the attached word. E.g., "*казак**ча** ырдоо*" "*kazak**cha** yrdoo*" (singing in Kazakh), "*француз**ча** сүйлөө*" "*fransuz**cha** süilöö*" (speaking French) etc.

b) **-lap;** This affix actually one of the sophisticated as it can create disambiguates, especially in determining the part-of-speech of the particular unit. Sometimes it can serve as an verbal adverb, when the word ends with ***-lap***, can be changed according to the verbal categories and can normally function without any sense loss in wide usage even when the affix **-yp** is omitted. In these cases, the three phonemes in ***-lap*** is contemplated as two separate affixes, like: ***-la*** – the affix forming verb and ***-yp*** – the affix of verbal adverb. For comparison:

If the words with the affix ***-lap*** will not possess this feature, they will be considered as adverbs. In this case the affix ***-lap*** cannot be splitted in terms of word-formation and it functions one, unite affix that forms the adverb. It is not possible to omit the phoneme ***-p*** and conjugate it in accordance with other verbal categories. E.g., *"арабал**ап**" "arabal**ap**"* (with the help of trolley), *"айл**ап**" "ail**ap**"* (during months), *"килол**оп**" "kilol**op**"* (by kilograms) etc.

c) **-lai;** This affix is also one of the widely practiced within the process of adverbs formation. This morpheme is also combined with a verb forming affix ***-la*** + affix ***-i***

used for forming adverbial participles. In determining the part-of-speech of the words with this morpheme, the same principles are applicable as are for the morpheme **-lap**. When this affix is attached to the nouns of time, newly formed adverbs in most cases denote the possible continuity of the action, for instance: "*күз**дөй**"* "*küz**döi***" (up to the autumn), *"жаз**дай**" "jaz**dai**"* (up to the spring), *"жайлы-жай**лай**"* "*jaily-jay**lai**"* (whole summer) etc.

d) **-chalyk;** adverbs that are formed through adding this affix to the noun tend to designate the sign of the similarity or comparison to the root of the attached word (noun), e.g., "*Марат**чалык***", *"Marat**chalyk**" (as Marat, other examples are the same as they all are the proper nouns) "Алия**чылык**" "Aliya**chylyk**"* "*Жибек**чилик***" "*Jibek**chilik***", "*Адилет**челик***" "*Adilet**chelik***" etc.

2. Adverbs formed from the pronouns. In the formation of adverbs of this type there are following affixes participate:

a) **-cha;** This affix is attached mostly to the personal, demonstrative, and reflexive pronouns and shows the resemblance of one's action or condition to the action and condition of someone/thing that is implied by a particular pronoun, e.g., *"мен**че**" "men**che**"* (according to me), *"сен**че**" "sen**che**"* (as for you), *"тигилер**че**" "tigiler**che**"* (as for them), *"өзүңөр**чө**" "özünör**chö**"* (depending on the context, like: decide it by yourself).

b) **-chylap**; this kind of affix also attached to the personal, demonstrative, and reflexive pronouns, and creates adverb as a result: *"менчилеп" "men**chilep**"* (like me), *"сен**чилеп**" "sen**chilep**"* (like you), *"тигил**чилеп**" "tigil**chilep**"* (like them), *"өзүңөр**чүлөп**" "özünör**chülöp**"* (to be on your own).

c) **-chalyk;** adverbs formed with the help of this affix are acquiring the meaning of measurement and becoming synonymous to the adverbs that have been formed by the affix ***-cha:*** *"мен**челик**" "men**chelik**"(*as much as me (depending on

context)), *"сен**челик**" "sen**chelik**"*(as much as you), *"ошон**чолук**" "oshon**choluk**"* (as much as it is possible).

3. Adverbs formed from the adjectives. They are formed with the help of these:

a) **-lata;** This affix turns the adjective into adverb and modifies the action according to the sign implied in the base of the adjective. For instance: *"кичүү**лөтө**" "kichüü**lötö**"* (giving priority to the younger ones), *"улуу**лата**" "uluu**lata**"* (giving priority to the elder ones), *"сакалдуу**лата**" "sakalduu**lata**"* (the same as previous example). Nota Bene: adverbs that are formed in this way are less frequent in the Kyrgyz language than in other Turkic languages, e.g., in the Bashkir language with **-lata** are formed adverbs from the numerals.

b) ***-cha;*** This affix creates adverbs from simple or derivative adjectives which express the time, e.g., *"жаңы**ча**" "janу**cha**"* (in a new way), *"эски**че**" "eski**che**"* (in an old way) – there were used simple adjectives, while there: *"былтыркыча" "byltyrky**cha**"* (as in the last year) *"мурункуча" "murunku**cha**"* (as earlier) – derivative adjectives were used for the formation of adverbs.

c) **-syn;** a small number of adverbs are formed with the help of this affix. Basically, this affix denotes a period and is attached to the derivative adjectives with already added affix ***-gy,*** for instance: *"эртеңки**син**" "ertenki**sin**"* (something that will happen tomorrow), *"кечки**син**" "kechki**sin**"* (something that will happen in ethe evening), *"кышкы**сын**" "kyshky**syn**"* (something that will happen during winter), *"түнкү**сүн**" "tünkü**sün**"* (something that will happen in the night-time).

4. Adverbs formed from the numerals. Such adverbs are mostly done by adding the affix **-lap** to the cardinal numeral and demonstrates the feature of an action through the quantity of a numeral. For example: *"он**дон**" "on**dop**"* (by tens), *"отуз**дан**"*

*“otuz**dap**”*(by thirties), *“жүз**дөп**” “jüz**döp**”* (by hundreds), *“миң**деп**” “min**dep**”*(by thousands).

5. Adverbs formed from other adverbs. In the formation of this kind of adverbs the affixes **-cha** and in some cases **-yncha** are applied. These affixes are attached to some adverbs of time and forms another adverb, which designates the limitation of the time that is implied in the base of an adverb to which the affix is attached: e.g., *“бүг**үнчө**” “büg**ünchö**”* (for today), *“эрте**ңче**” “ert**enche**”* (tomorrow) *“азыр**ынча**” “azayr**yncha**”* (for now).

-lap; sometimes with the help of this affix another derivational adverb can be formed from the simple one. E.g., *“эртел**еп**” “ert**elep**”* (soon), *“тез**деп**” “tez**dep**”* (rapidly), *“аз-аз**дап**” “az-az**dap**”* (depending on context, like “step-by-step”, “little-by-little”).

There is also another type of adverbs that have been formed from the words by accepting case affixes, e.g.:

1. Adverbs that contain the suffixes of locative and ablative cases in their composition: *“мурдатан” “murdatan” (earlier), “илгертеден” “ilgerteden” (from days of old), “эртеден” “erteden” (earlier)* etc.;
2. Adverbs that contain the suffixes of ancient dative (***-ry***) and ablative cases in their composition: “арытан” “arytan” (), “беритен” “beriten” ();
3. Adverbs that contain the suffixes of ancient dative (***-ry***), locative and ablative cases in their composition: “арытадан” “arytadan” (), “беритеден” “beriteden”(), “ичкертеден” “ichkerteden” ();
4. Adverbs that contain the suffixes of ancient dative (***-gary***), and ablative cases in their composition: “илгертен” “ilgerten” (), “жогортон” “jogorton” ().

In such a way, the existence of all above-mentioned affixes is the reason for the enriching of the adverb’s lexicon from the one side, and they are playing the role of the

formal signs that distinct the adverb class from the other parts-of-speech withing the language.

2.1.3. Syntactic way of forming adverbs

Beside the morphological approach, there is another way as well to form the adverb, this is performed on the syntactic level. This kind of adverbs contain in its structure two or more words which coming together compiles one unity. Adverbs that have several stems in its structure that together making up one, unite meaning as well as function as one member of a sentence are called as adverbs formed syntactically. Adverbs of this type is divided internally into two groups – compound and complex adverbs.

a) Compound adverbs – in the Kyrgyz language they are called as "Kosh taktooch". Compound words that have the features of adverbs from the grammatical perspective are called compound adverbs. Their components are usually hyphenated, e.g., *"оозмо-ооз" "oozto-ooz" (*word of mouth*)*, *"көзмө-көз" "közmö-köz"* (eye-to-eye), *"өйдө-төмөн" "öidö-tömön"* (up and down) etc.

Among compound adverbs, there appears another little classification according to their formation structure. There are compound adverbs, that are formed through linking or the iteration of the components, for instance: *"бат-бат" "bat-bat"* (quickly), *"кайра-кайра" "kaira-kaira"* (again and again), *"ары-бери" "ary-beri"* (this way and that way).

Second type is the type of compound adverbs, where both of its components have the suffix of that or another case, e.g., *"кээ**дe**-кээ**дe**" "kee**de**-kee**de**"* (sometime), *"анда-санда" "anda-sanda"* (from time to time), *"күндө-күндө" "kündö-kündö"* (every day) etc.

Speaking on the third type of compound adverbs, the affix **-ma** is attached to the first component while the second component is remained as it is, for example: *"жекеме-жеке" "jekeme-jeke"* (privately), *"үймө-үй" "üimö-üi"* (pile of), *"оозмо-ооз" "oozmo-looz"* (word of mouth) etc.

b) Apart of the compound adverbs there is second type of syntactically formed adverbs, namely complex adverbs. Complex words that have the features of adverbs from the grammatical perspective are complex adverbs. Their components are written separately and no any punctuation is set between them, e.g., *"алда канча" "alda kancha"* (much more), *"бир кезде" "bir kezde"* (at one time), *"ар дайым" "ar dayim"* (always), *"эч жерде" "ech jerde"* (nowhere) *"эч убакта" "ech ubakta"* (nevermore), *"эч качан" "ech kachan"* (never) etc. In most cases, as the first component of the complex adverbs are used these words: "ar", "ech", "bir", "alda". This fact demonstrates adverb's proximity to pronouns.

In the process of language development, a number of complex adverbs underwent phonetic changes and became one integral linguistic unit. At the moment, these adverbs developed into "simple" word, precisely simple adverb, e.g., "бүгүн" "bügün" (today), "бүрсүгнү" "bürsügünü" (the day after tomorrow), "быйыл" "byiyil" (this year), "былтыр" "byltyr" (last year), "нечак" "nechak" (long time ago) etc. This category of adverbs may be considered as complex only from the historic perspective.

2.2. Lexical and semantical division of adverbs

While classifying the adverbs on lexical and semantical levels, not only the lexical meaning of an adverb should be taken into account, but also its relations with antonyms and synonyms as well as with the verbal predicate since these aspects will support in determining the peculiarities of a particular adverb. For instance: *"... ал илгери кеткен, сен кийин кетесиң" "...al ilgeri ketken, sen kiyin ketesin".* In this sentence, the adverbs

“илгери, кийин” “ilgeri, kiyin” can function as the adverb of place and answer the question *Where?* at the same time as an adverb of time and answer the question *When?* Therefore, it is not possible to say exactly which of the abovementioned contexts will be right within this sentence. So, if we substitute the predicates of the sentence to *“kelsin”* and *“kel”,* then the following sentence will appear: *“ал илгери келсин, сен кийин кел” “al ilgeri kelsin, sen kiyin kel”.* In this case the with the help of predicates these adverbs will designate the concept of time only. In order to consider the relations of adverbs with antonyms and synonyms, let’s have a look at the following sentences:

1. ***Жаңы*** *чыктык, жолоочулук бүткөн жок. /* ***Jany*** *chyktyk, joloochuluk bütkön jok.* / We **just** have got hit the road, the whole road ahead.
2. *Эки күн көз ачырбай кар бороилоп жаап,* ***жаңы гана*** *басылган. / Eki kün köz achyrbai kar boroilop jaap,* ***jany gana*** *basylgan.* / The blinding eyes blizzard that had been raging for two days, had **just** subsided.
3. ***Жаңы*** *портфелин оодара баштады. /* ***Jany*** *portfelin oodara bashtady*. / He began to rummage through the **new** briefcase.

In the first two sentences, there are two words ***“жаңы” “jany”***, ***“жаңы гана” “jany gana”*** (both translated as “just”) designating the time. It is not possible to consider them as an adjective since they are followed by other adverbs, like *emi, emi ele, azyr, azyr ele, emi gana, azyr gana* and are synonymous to each other as well as the adverb ***“жаңы” “jany”*** is conjugating the verb. Secondly, these words are not able to accept the suffix of the comparative degree ***“-ыраак” “-yraak”***. While the word “жаңы” “jany” in the third sentence is modifying the noun and even though it can attach to itself the affix of the comparative degree and, (*“жаңыраак портфель” “janyraak portfel”/ the newer briefcase*), it cannot be synonymous to the aforementioned adverbs from the lexical and semantical perspective. As well, the antonyms for the adverbs from the 1st and 2nd sentences are different from the one for the 3rd sentence, in such a way it is possible to identify the adverbs by accounting the synonyms and antonyms.

Taking into consideration all the features of the adverbs, they are divided into two large groups: Attributive adverb and adverbs that function as adverbial modifiers on the syntactic level.

2.2.1. Adverbs that function as adverbial modifiers

Adverbs of this type are demonstrating the time, place, reason or purpose of the action. They are subdivided within according to these criteria.

a) Adverbs of time

The adverbs of time designate the time in general and answers the questions like *"Качан?" "Kachan?"* (When?), *"Кайсы убакка чейин?" "Kaisy ubakka cheiyin?"* (Until which time?). Adverbs of time are the largest group among others and include following adverbs: *"анан" "anan"* (then), *"мурун" "murun"* (earlier), "илгери" "ilgeri" (at ancient times), *"кечке" "kechke"* (a whole day*), "күнү-түнү" "künü-tünü"* (day and night), *"жылдан-жылга" "jyldan-jylga"* (annually), *"эртели-кеч" "erteli-kech"* (sooner or later) and many others.
*"Багың ачылар күнүң **бүгүн**" "Bagyn achylar künün **bügün**"* (**Today** is the day when you find happiness), "***Илгери** мен мындан жаман күндөрдү өткөргөм." "**Ilgeri** men jash kezimde, myndan jaman kündördü bashyman ötkörgöm"* (Long ago, when I was younger, there were worse days).

b) Adverbs of place

Adverbs of place modify the place or direction of an action/movement, and answer questions – "Кайда?" "Kaida?" (Where?), "Кайдан?" "Kaidan?" (Where from?). E.g., *"төмөн" "tömön"*(down), *"өйдө" "öidö"* (up), *"жакын" "jakyn"* (near), *"артка" "artka"* (back), *"ылдый" "yldyi"* (low), *"алга" "alga"*(forward), *"алыста" "alysta"(far away), "туш-тушитан" "tush-tushtan"*(from different sides). *"Эки колу*

менен жер таянып, ***алга*** *жыгылды." "Eki kolu menen jer tayanyp* ***alga*** *jygyldy."* (She fell forward, leaning two hands on the ground.) "*Анын үнү толкунданып, алыска алда кайда угулду" "Anyn ünü tolkundanyp, alyska* ***alda kaida*** *uguldu"* (His voice was agitated and sounded far away).

c) Adverbs of reason and purpose

Adverbs of reason are used to indicate the reason for an action and answer questions such as: Why? What for? What is the reason? Following adverbs are the sample for this type of adverbs: *"айласыздан" "ailasyzdan"* (for particular reason), *"себептен улам" "sebepten ulam"*(due to), *"аргасыздан" "argasyzdan"*(because of the predicament), *"амалсыздан" "amalsyzdan"* (out of desperate need). ***Adverbs of purpose*** contain at the core the aim of an action/movement and designate one purpose by answering the questions: For what purpose? What for? This meaning is rendered through the words "*атайы" "atayi"*, *"атайын" "atayin"* – both translated as "special, especially".

In the Kyrgyz language, adverbs of reason and purpose are very small in number and tight from the usage perspective. Sometimes, they might acquire the meaning of description.

2.2.2. Attributive adverbs

Attributive adverbs are designating the number and qualities of an action/movement. According to this, they are divided into two groups: attributive adverbs of quality and attributive adverbs of number.

a) Attributive adverbs of quality are demonstrating the quality and the way of how the action/movement is performed and answer questions *How? In what way?* E.g., "*Бир нерсе кылуу* ***аста-аста****, талкалап салуу – бир паста" "Bir nerse*

kyluu ***asta-asta****, talkalap salyy – bir pasta"* (It is easier to ruin something than to create).

b) Attributive adverbs demonstrate the rapidness of an action, e.g., *"Ээси анча эмес, ээрчиме, кошоматка* ***бат*** *жыгылат" "Eesi ancha emes, eerchime, koshomatka* ***bat*** *jygylat"* (The one who has no will and his head is not on his shoulders, **quickly** becomes a victim of sycophants and flatterers);

c) Attributive adverbs designate the state of an action, characterize the action, e.g., *"Асан сыртта жүгүрүп, алка-шалка тердеди" "Asan syrtta jügürüp,* ***alka-shalka*** *terdedi"* (During the running outside, Asan sweat like a pig);

d) Attributive adverbs also demonstrate the state of reiteration or the degree of duration of an action, for instance: *"Жети кадам баскан соңб кайра отурушту" "Jeti kadam baskan son,* ***kaira*** *oturushtu"* (After walking another distance, they sat down **again**.);

e) Through the comparison of the one's action to the action of another person/thing, attributive adverbs designate the state of implementation of the action, e.g., "*Тиги киши кыргызча шар кетпегени менен, тилди дурус түшүнөт" "Tigi kishi kyrgyzcha shar ketpegeni menen tildi durus tüshünöt"* (Even though that man is incapable of speaking Kyrgyz immediately, he does comprehend well).

Attributive adverbs of measure designate the number, quantity or volume of an action and answer the questions How many? How much? Examples: *"Ал убакта тагдыр мени ошончолук каргады деп айтпас элем" "Al ubakta tagdyr meni oshoncholuk kargady dep aitpas elem"* (At the time I couldn't say that fate had cursed me to such an extent), *"Канчалык күтсө да Айдарбек келбеди" "Kanchalyk kütsö da Aidarbek kelbedi"* (Even though s/he waited for so long, Aidarbek didn't come), "*Менчелик аны эч ким жакшы көрбөйт" "Menchelik any ech kim jakshy körböit"* (Nobody do like him/her so much as I do).

2.3. Syntactic characteristics of adverbs

In order to give a syntactic description of an adverb, it is necessary to consider the following questions.

2.3.1. The relationship of adverbs with other words within a sentence

Adverb, like the rest parts-of-speech is combined with the other part-of-speech words and has dependent or independent meaning, but the degree of combination with another words is not the same. This peculiarity is closely related with the lexical and grammatical peculiarities of those words, with which the adverbs are combined together.

a) In the Kyrgyz language, the adverbs are sufficiently related with the verbs and demonstrate various features. E.g., *"Акылбектин баласы **эбак** жаткан" "Akylbektin balasy **ebak** jantkan"* (The son of Akylbek has **long since** gone to bed);
b) When the adverb is combined with the adjectives, it complements the sense of the adjective and modifies the adjective, e.g., *"өзү гвардеецче тыкан киши" "özü gvardeetsche tykan kishi"* (He is neat as the guardsman);
c) There are also cases, when the adverb is combined with other adverbs and in this case, they will have different correlations in terms of the sense, for example there can be combined adverbs of time and place, e.g., *"эртең эрте" "erten erte"* (tomorrow early), *"бүгүн кечинде" "bügün kechinde"* (tonight).

2.3.2. The syntactic function of adverbs

Adverbs also do function as a member of the sentence along with other parts-of-speech, but the degrees of usage the adverbs in terms of the different members of a

sentence is not the same. Mostly, adverbs function within the sentence as an adverbial modifier, like adverbial modifiers of time, place, manner, number, reason and purpose.

a) Adverbs can combine with nouns and function as adverbial modifier, e.g., *"... **Мурун** турмуш ушуга окшош кең беле?" "... **Murun** turmush ushuga okshosh ken bele?"* (Was ever life so prosperous before?);
b) When adverb is combined with a noun or an adjective, then adverb do function as attributive adverb as well, e.g., *"казакча бий" "kazakcha byi" (Kazakh dance), "жаңыча турмуш" "janycha turmush" (a new life). "өзгөчө ширин" "özgöchö shirin" (especially pretty);*
c) In rare cases, adverbs can function as predicates as well, for example: *"Токой ичи туңгуюк **жымжырт**" "Tokoi ichi tunguyuk **jymjyrt**"* (There was a deafening silence in the forest).

2.4. The place of adverbs in the structure of the sentence

Not all adverbs share the same peculiarities according to their place in the sentence, due to the fact that this feature is related to the meaning of an adverb as well as to the grammatical peculiarity of the word, with which the adverb is combined within a sentence (Sartbayev, 1958, p. 129). For instance, the adverbs of time and place are free in terms of their place in the sentence, because most of them are subordinate from the verbal predicates only, and therefore no matter where the adverb is located in the sentence, it will remain in the subordinate relation to the verbal predicate, for example: *"Бир бүркүт айланып жүргөн **анда**" "Bir bürküt ailanyp jürgön **anda**"* (An eagle was rounding there) and *"Бир бүркүт **анда** айланып жүргөн" " Bir bürküt **anda** ailanyp jürgön"* (There was rounding an eagle) in these examples the adverb *"anda"* is subordinating to the predicate *"ailanyp jürgön",* and due to this fact it is possible to change the word order o the adverbs of time and place.

Here were considered the main characteristics of the adverbs, that will be really valuable information for the processing the Kyrgyz language texts, containing adverbs and their tags, which in their turn accelarate in comprehending by the computers our native language. The next section wiil be devoted to the corpus-driven study of the adverbs and their statistical data will be presented.

2.5. Adverbs in the Kyrgyz Corpus – statistical data

Linguistics is a subject, which has a long relationship with statistics and mathematics. Mathematical linguistics, computational linguistics, corpus linguistics, applied linguistics, forensic linguistics, stylometric, etc. requires different statistical and quantitative results obtained from natural language corpus. Without adequate knowledge of statistical information about different properties of a language we can make mistakes in handling linguistic data as well as in observation. Invalid source specified.. Both quantitative and qualitative analysis can be performed on corpora. In quantitative analysis, we categorize various linguistic characteristics or features of a specific language, count them, and even create more intricate statistical models in an effort to explain what is observed. The most fundamental statistical measure, the frequency count, which simply counts the number of times something occurs in a corpus, can be useful in this situation**Invalid source specified.**. That tool is highly invaluable when conducting various researches in terms of teaching a language to foreigners, compiling dictionaries and for a great number of the linguistic studies. In the next section the term of frequency will be considered a bit closer.

2.5.1. What is frequency in the linguistic corpus?

Frequency (or relative frequency) is the ratio of the number of instances of a given object to the total number of instances of all objects in the set. Sometimes the number of instances of a given object is called the absolute frequency of that object. In linguistics and lexicostatistics, such objects are words, letters, and word combinations, in cryptography and informatics also combinations of letters and simply any symbols. A set can be understood as a text, a set of texts (corpus), or even a language. **Word counts** have existed since the Hellenistic period. Before current computers made such tasks significantly simpler, Thorndike & Lorge tallied 18,000,000 running words with the help of their colleagues in 1944 to produce the first comprehensive frequency list. In 1953, Harry Yosselson of Detroit published the first frequency dictionary of the Russian language after studying a collection of one million words. In the Soviet Union, frequency dictionaries later started to appear. A group led by Lidia Zasorina from Leningrad and Gorky - from St. Petersburg - published a dictionary with 400,000 words in Tallinn during the 1960s. Another dictionary based on one-million-word forms was published in 1977 by Petersburg and Nizhny Novgorod. We are currently in the era of frequency dictionaries, which are generated almost entirely automatically. The most well-known modern Russian frequency dictionary is the dictionary of Olga Lyashevskaya and Sergey Sharov, which was published in 2009 and is based on a corpus of texts with a volume of 92-million-word forms. This dictionary is freely available on the Internet, and anyone can use it.

A **frequency list** is a sorted list of words (word categories) and their frequency in computational linguistics, where the frequency is typically the number of occurrences in a particular corpus and rank is the position in the list. Frequency lists are a useful tool when building an electronic dictionary, which is a prerequisite for a wide range of applications in computational linguistics. Those lists are not intended to be given directly to students, but rather to serve as a guideline for teachers and textbook authors. The frequency of words is calculated using the formula:

$$Freq_x = \frac{Q_x}{Q_{all}},$$

where ***Freqx*** is the frequency of the word "x", ***Qx*** is the number of uses of the word "x", ***Qall*** is the total number of uses of the word. In most cases frequency is expressed as a percentage. In dictionaries the frequency of words can be reflected by the marks - commonly used, little-used, etc. The frequency of words and letters was the most important tool of cryptanalysis in the era before the widespread use of block ciphers. The terms frequency and rate (ciphers) should not be confused. If we know how many times in the corpus each word form/lemma occurs, we can build a frequency dictionary. This dictionary contains frequent and rare words in the given corpus. Frequency IPM - instances per million words are how words appear per million words.

The reasons for which the **frequency** dictionaries are required:

- Automatic text processing;
- Language Teaching;
- Teaching native speakers;
- Teaching a language as a foreign language;
- When learning a foreign language, the first thing you need to know is frequent words.

For example, in 1953, Michael West compiled the so-called “General Service List” for English, a list of common words, a list of commonly used words, in which he included about 2,000 of the most frequent English words, claiming that they cover over 90% of oral speech and 85-90% of written speech in English. That is, a foreigner who has learned these 2,000 words can understand a significant portion of texts quite easily. And this is, generally speaking, very convenient for a learner, for example, of the English language, because he understands that some words, he is likely to need, some he is likely not to need, and he can

distribute his efforts accordingly. For the Russian language, such a dictionary was also recently published, compiled by a team of authors headed by Sergey Sharov, one of the authors of the frequent dictionary of the Russian language. This, too, can come in handy when teaching Russian as a foreign language. Therefore, **word frequency** is known to have various effects. Higher word frequency has a good impact on memory, perhaps because the learner is exposed to more information (Laufer, 1997). High word frequency has a beneficial impact on lexical access; this is known as the word frequency effect.

Another task for which knowledge of word frequency is useful is automatic processing of natural language for different purposes. Take automatic spell checking as an example. For example, you can calculate the probability of this or that mistake, you can estimate the probability with which the user misses a letter, with what probability the user replaced "you are" with "your" or "professor" with "pro**ff**essor", because it is clear that it somehow depends on the arrangement of letters on the keyboard, for example. However, another thing that is important to consider is to think about what word the user wanted to write. This is where frequency comes into play again and we look at some corpus of texts for the frequency of those words.

One of the key elements in language processing is lexical frequency. It serves as a cornerstone for applied, corpus-based, and psycholinguistic research. Frequency counts from corpora are used by linguists and have become standard practice. Unfortunately, the argument that corpora might not always give an accurate picture of how frequently lexical items exist in a language is being raised. However, linguists all around the world are trying to solve these rising issues to make corpora useful for further data research.

The frequency dictionary provides information about which words occur **most often** in the corpus or sub-corpus that the user has selected. Let us look at the adverbs in the Kyrgyz language and their frequency. When we look up a definite word, in our case, an adverb in the query field, we get its occurrences per million words. We may express this as a percentage of the whole corpus; for instance, let us make a query of the word “азыр”.

Your query “азыр” returned 975 matches in 167 different texts (in 2,493,894 words [1,019 texts]; frequency: 390.955 (instances per million words). The percentage is just another way of looking at the count 390.955 context, to try to make sense of it relative to the totality of the written corpus.

2.5.2. Tables of frequencies of adverbs extracted from the Kyrgyz Corpus

For the further analysis to be made in the following chapter, the frequency tables of adverbs in the Kyrgyz Corpus were compiled up in accordance with their classification. This information facilitates the task of labeling the adverbs correct tags taking into account all the peculiarities of an adverb as an independent part-of-speech and creates a basis for the elimination all the tags attached wrong or fulfill the ones which weren’t labeled with any tag at all. This tables will also give us a comprehension what and in what way should be done the further efforts on the developing the Kyrgyz Corpus, particularly in terms of annotating the adverbs. So, there are following tables:

Table 2.5.2.1. Мезгил бышыктагыч тактоочтор – Adverbial modifiers (adverbs of time)

1. азыр	390.955
2. алда качан	8.421
3. анан	294.720
4. анда-санда	8.020
5. бая күнү	4.010
6. бир кезекте	2.406
7. бирде	124.304
8. бүгүн	143.952
9. бүрсүгүнү	3.208
10. быйыл	67.766
11. жаздын күнү	0.401
12. жайдын күнү	4.010

13. жылдан жылга	4.010
14. илгери	62.152
15. капкачан	6.416
16. кечке	37.692
17. кечээ	66.563
18. күздүн күнү	1.203
19. кыштын күнү	4.812
20. мурда күнү	0.401
21. мурун	124.304
22. эбак	7.619
23. эми	1394.413
24. эң мурун	3.609
25. эртели кеч	0.401
26. эртең	93.027
27. эртең жок	14.836
28. эчак	18.445

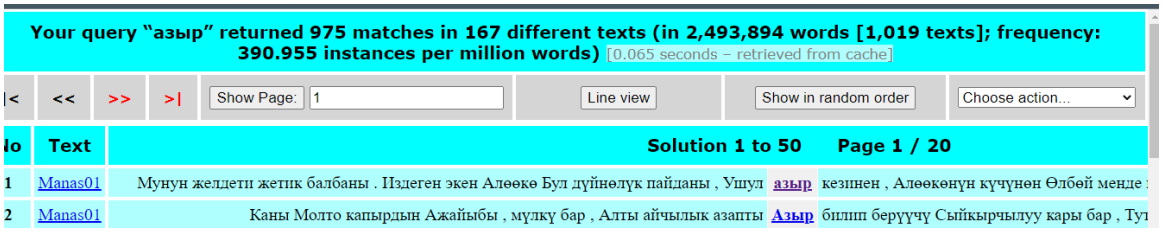


Figure 2.5.2.1. The screenshot of the adverb of adverbial modifier of time "**azyr**" with its frequency data extracted from the Kyrgyz Coprus

Table 2.5.2.2. Орун бышыктагыч тактоочтор – Adverbial modifiers (adverbs of place)

1. алга	30.875
2. алда кайдан	6.817
3. алдына	295.121
4. алыска	32.880
5. алыста	18.044
6. анда	332.813
7. арт жагынан	14.034
8. артка	33.682
9. жакын	318.378

10. жогору	123.101
11. илгери	62.152
12. кийин	791.533
13. мында	150.367
14. өйдө	147.961
15. төмөн	85.008
16. ылдый	66.563
17. ыраак	7.619

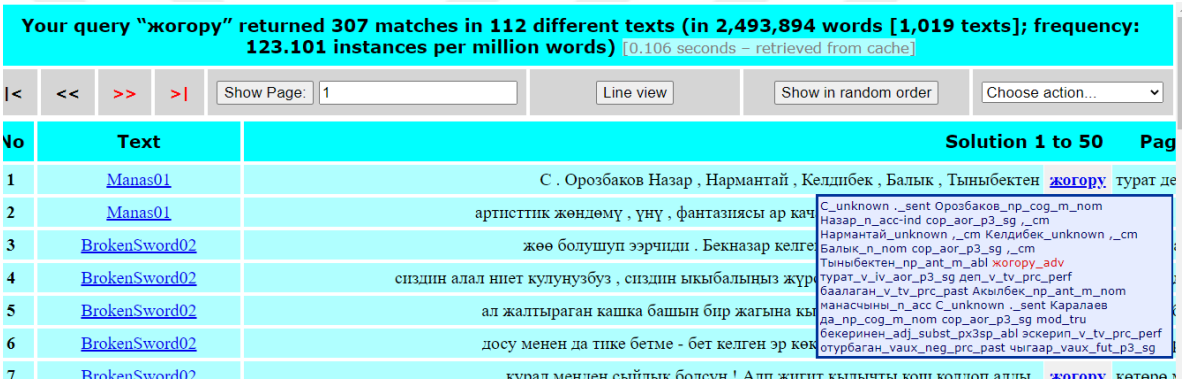


Figure 2.5.2.2. The screenshot of the adverb of adverbial modifier of place "**jogoru**" with its frequency data extracted from the Kyrgyz Coprus

Table 2.5.2.3. Себеп жана максат бышыктагыч тактоочтор – Adverbial modifiers (adverbs of reason and purpose)

1. айласыздан	2.807
2. амалсыздан	0.401
3. аргасыздан	5.614
4. атайлап	2.807
5. атайын	354.466
6. капилеттен	15.237
7. капыстан	24.059
8. эрксизден	2.005

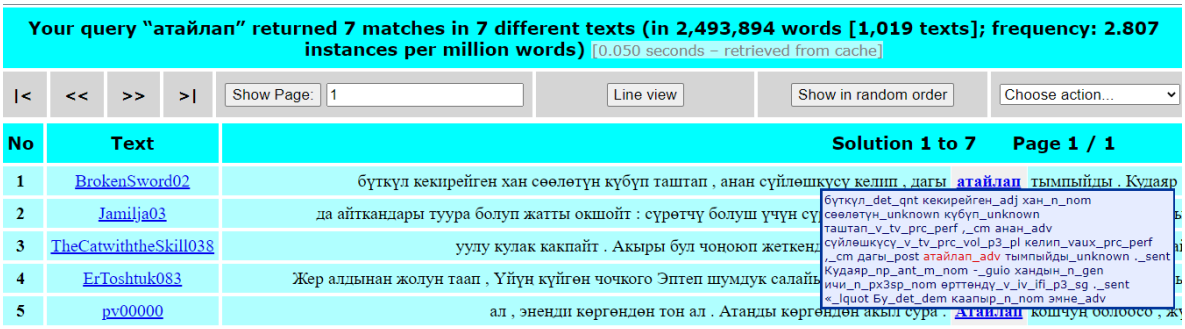


Figure 2.5.2.3. The screenshot of the adverb of adverbial modifier of purpose "***atailap***" with its frequency data extracted from the Kyrgyz Coprus

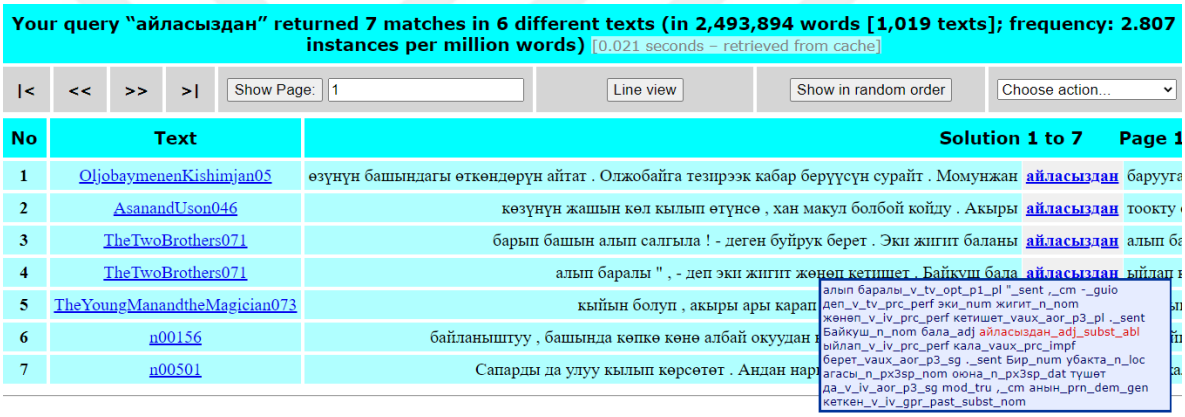


Figure 2.5.2.4. The screenshot of the adverb of adverbial modifier of reason "***ailasyzdan***" with its frequency data extracted from the Kyrgyz Coprus

Table 2.5.2.4. Сын-сыпат аныктагыч тактоочтор – Attributive adverbs (adverbs of quality)

1. акырын	393.762
2. аста	17.242
3. адамча	9.624
4. тымызын	79.795
5. тез	180.040
6. бат	62.553
7. дароо	116.284
8. ылдам	10.826
9. жай	225.751
10. бирге	287.502
11. бетме-бет	0.401
12. кайра	1,052.170

13. араң	125.507
14. зорго	4.812
15. жымжырт	22.455
16. жайбаракат	5.614
17. анча-мынча	5.614
18. кайта	103.453
19. үстүнө	178.035
20. бекер	69.770
21. улам	391.356
22. аябай	169.213
23. абдан	132.323
24. чапчаң	3.208

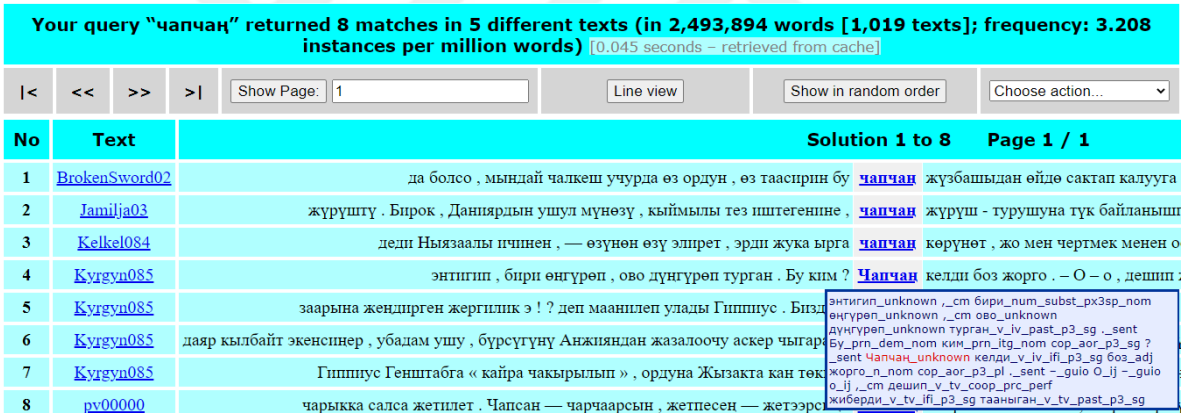
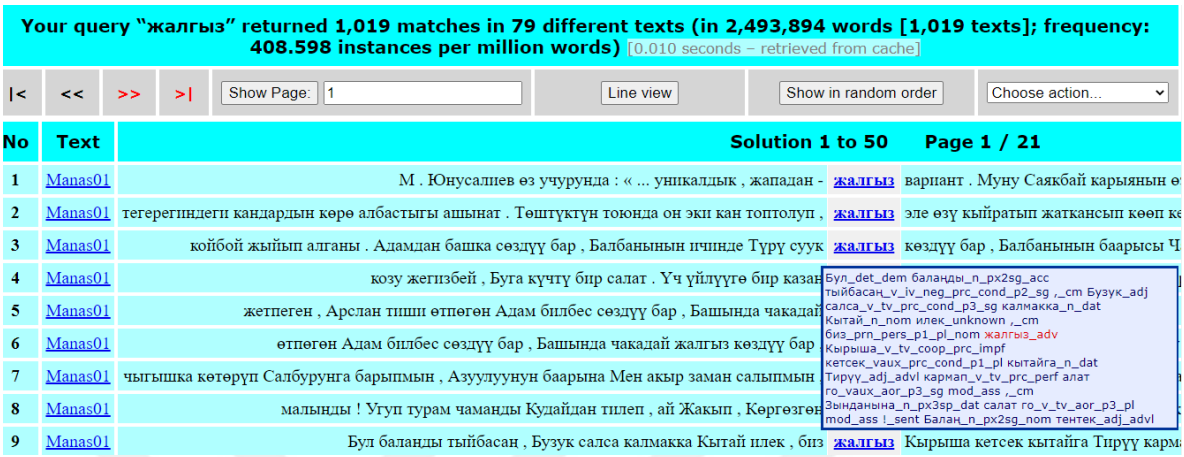


Figure 2.5.2.5. The screenshot of the adverb of attributive adverb of quality "***chapchan***" with its frequency data extracted from the Kyrgyz Coprus

Table 2.5.2.5. Сан-өлчөм аныктагыч тактоочтор – Attributive adverbs (adverbs of number)

1. көп	1,230.205
2. аз	281.087
3. көптөн	42.905
4. аздан	22.054
5. анча-мынча	5.614
6. далай	67.766
7. бирден	99.042
8. экиден	24.861
9. жалгыз	408.598



Your query "жалгыз" returned 1,019 matches in 79 different texts (in 2,493,894 words [1,019 texts]; frequency: 408.598 instances per million words) [0.010 seconds – retrieved from cache]

|< << >> >| Show Page: 1 Line view Show in random order Choose action...

No	Text	Solution 1 to 50 Page 1 / 21		
1	Manas01	М . Юнусалиев өз учурунда : « ... уникалдык , жападан -	жалгыз	вариант . Муну Саякбай карыянын ө
2	Manas01	тегерегиндеги кандардын көрө албастыгы ашынат . Төштүктүн тоюнда он эки кан топтолуп ,	жалгыз	эле өзү кыйратып жаткансып көөп ке
3	Manas01	койбой жыйып алганы . Адамдан башка сөздүү бар , Балбанынын ичинде Түрү сыяк	жалгыз	көздүү бар , Балбанынын баарысы Ч
4	Manas01	коzу жегизбей , Буга күчтү бир салат . Үч үйлүүгө бир казан		
5	Manas01	жетпеген , Арслан тиши өтпөгөн Адам билбес сөздүү бар , Башында чакадай		
6	Manas01	өтпөгөн Адам билбес сөздүү бар , Башында чакадай жалгыз көздүү бар		
7	Manas01	чыгышка көтөрүп Салбурунга барыпмын , Азуулуунун баарына Мен акыр заман салыпмын		
8	Manas01	малынды ! Угуп турам чаманды Кудайдан тилеп , ай Жакып , Көргөзгөн		
9	Manas01	Бул баланды тыйбасаң , Бузук салса калмакка Кытай илек , биз	жалгыз	Кырыша кетсек кытайга Тирүү карм

Бул_det_dem баланды_n_px2sg_acc
тыйбасаң_v_iv_neg_prc_cond_p2_sg ,_cm Бузук_adj
салса_v_tv_prc_cond_p3_sg калмакка_n_dat
Кытай_n_nom илек_unknown ,_cm
биз_prn_pers_p1_pl_nom жалгыз_adv
Кырыша_v_tv_coop_prc_impf
кетсек_vaux_prc_cond_p1_pl кытайга_n_dat
Тирүү_adj_advl карман_v_tv_prc_perf алат
го_vaux_aor_p3_sg mod_ass ,_cm
Зынданына_n_px3sp_dat салат го_v_tv_aor_p3_pl
mod_ass !_sent Балан_n_px2sg_nom тентек_adj_advl

Figure 2.5.2.6. The screenshot of the adverb of attributive adverb of measure "jalgyz" with its frequency data extracted from the Kyrgyz Coprus

All of these data are extracted from the Kyrgyz Corpus (Knappen, 2022) . After division of the adverbs according to their main type, the "pie" has been constructed, illustrating the quantitative ratio of these types within the Kyrgyz Corpus. Leaning on this data, it will be possible to count and take a decision on which type of adverbs a tagset of a newly suggested categories should be developed for in priority.

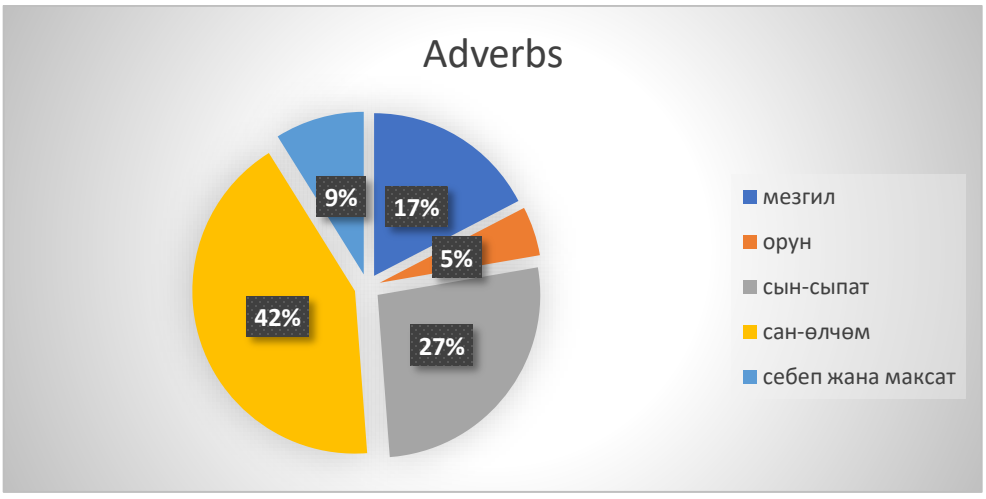


Figure 2.5.2.7. Pie of types of adverbs within the Kyrgyz Corpus

Diagram of the frequency of adverbs per one million words on average: мезгил – 99,6, орун – 28,066, себеп жана максат – 50,875, сын-сыпат – 151,96, сан-өлчөм – 242,46.

2.6. Universal Dependencies

With the development of automated text parsing and analysis tools, cross-linguistic research is becoming increasingly important, allowing us to detect similar linguistic phenomena in the context of several languages, translate from one language to another, analyze the semantic structure of multilingual texts, conduct comparative cross-linguistic research, form a linguistic typology based on several languages.

However, multilingual analysis in the field of grammatical parsing has long been obstructed by the fact that the parsing schemes for different languages were significantly different. This made it very difficult to perform comparative evaluations and cross-linguistic evaluations of experiments. Grammatical parsing of a sentence, which is the basis of almost any linguistic study, is characterized by comparing the morphological and syntactic features that determine the role of a lexical unit in a sentence, its dependence on other sentence members, and so forth. Each language has its own well-established own set of features. Thus, the lack of standardization of linguistic features creates difficulties in solving the following problems: comparing empirical studies of multiple languages; performing cross-lingual transfer of language constructions; evaluating cross-lingual learning; creating and maintaining multilingual systems; performing comparative cross-lingual studies; testing linguistic typology; conducting research toward building a universal cross-language text parser (Lyukina, 2012).

Therefore, contemporary developments in deep learning have made it now possible to create parsers based on neural networks that are trained on large amounts of texts of the target languages and are multilingual (Lyukina, 2012). Those parsers are now able to analyze sentences from different languages using one structure approach. One of the current online web-based platforms for sentence parsing is **Universal Dependencies**, often used as **UD**.

A framework called **Universal Dependencies** (UD) is used to consistently annotate the grammar of various human languages, including the parts of speech, morphological characteristics, and syntactic dependencies. Over 300 people have produced almost 200 treebanks in over 100 languages as part of the open community project known as UD. A **treebank** is a specific approach to analyzing sentences applying schemes and showing the relations of phrases in the sentence. A treebank is a parsed text corpus that annotates syntactic and semantic sentence structure in linguistics. In the early 1990s, the development of parsed corpora transformed computational linguistics, which derived from large-scale empirical data. Here we can see an example of a treebank (Figure 2):

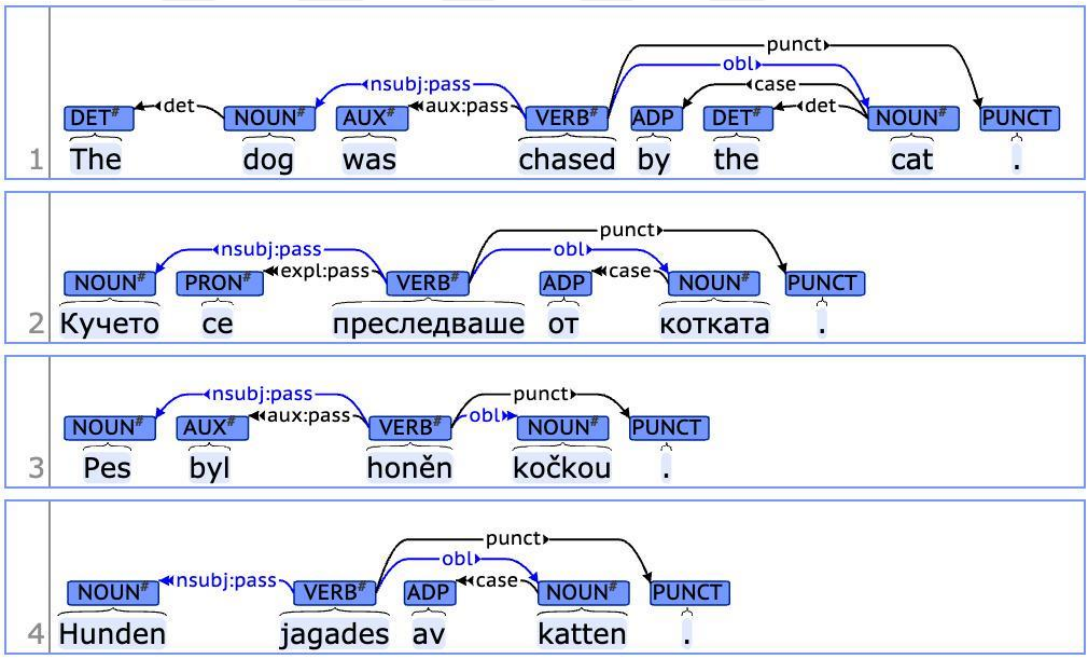


Figure 2.6.1. An example of treebanks in different languages

As we can observe above from the picture, all these sentences in various languages have the same grammar, and structurally can be explained in the same way. These sentences are annotated through unified UD tools (Hajic, Nivre , Gryuter, & Miyao, 2004).

The main objectives of the project are defining a cross-lingual consistent grammatical parsing scheme; supporting multilingual linguistic research; supplementing (not replacing) language-specific schemas, and annotating specific parsings selected from a universal set of conventions. The working hypothesis in UD is that languages have common parts of speech (nouns, verbs, adjectives, etc.) and common grammatical features (subject, predicate, etc.). Some languages may not have all these features; some may have features that are not universal. This approach has proven to be very successful, and currently, more than 60 languages are supported in UD.

However, the Kyrgyz language is not a part of this hugely important project and has not been analyzed yet. Currently, the Kyrgyz language is on the list of languages that are Possible Extensions for further development (see Figure 3.).

Possible Future Extensions

People have expressed interest in providing annotated data for the following languages but no data has been provided so far.

Amharic	1	-		Afro-Asiatic, Semitic
Ancient Greek	1	19K		IE, Greek
Archaic Irish	1	-		IE, Celtic
Assamese	1	-		IE, Indic
Bengali	2	-	W	IE, Indic
Bhojpuri	1	-		IE, Indic
Bororo	1	<1K		Bororoan
Cappadocian	1	-		IE, Greek
Cusco Quechua	1	-		Quechuan
Czech	1	1,198K		IE, Slavic
Danish	1	-		IE, Germanic
Dargwa	1	-		Nakh-Daghestanian, Lak-Dargwa
English	3	1,209K		IE, Germanic
French	1	-		IE, Romance
Frisian	1	-	W	IE, Germanic
Georgian	1	-		Kartvelian
Greek	1	23K		IE, Greek
Gujarati	1	-		IE, Indic
Hiligaynon	1	-		Austronesian, Central Philippine
Kabyle	1	23K		Afro-Asiatic, Berber
Kannada	1	-		Dravidian
Khoekhoe	1	-		Khoe-Kwadi
Kiga	1	-		Niger-Congo, Bantoid
Korean	2	-		Korean
Kyrgyz	1	-		Turkic, Northwestern
Ladino	1	-		IE, Romance
Laz	1	2K		Kartvelian
Macedonian	1	-		IE, Slavic
Magahi	2	7K	W	IE, Indic
Maghrebi Arabic French	1	-		Code switching
Mandyali	1	-		IE, Indic

Figure 2.6.2. A set of languages that are in the Possible Extensions list

Prior to Universal Dependencies, there were several attempts to create unified cross-linguistic parsing schemes, and UD is an integration of several of them. The basis of

UD is the unification of several universal approaches. Nevertheless, the main feature of UD is the application of treebank structures. Those treebank structures can be annotated via an online free website called UD Annotatrix. It is a client-side, browser-only, language-independent tool for editing dependency trees according to the guidelines established by the Universal Dependencies project. UD-Annotatrix supports uploading corpora to the browser to add, remove, and edit dependencies in a wide variety of formats (Washington , 2021).

The UD annotation scheme performs a syntactic analysis of sentences in terms of dependency grammar of dependencies. Each dependency is characterized by a syntactic function, which is shown by a label on the edge of the dependency. In the dependency grammar software, the UD annotation scheme is controversial. The main cause of disagreement concerns the analysis of function words. The UD prefers to subordinate function words to content words, a practice that is at odds with most work in the dependency grammar tradition.

UD has its own Universal Parts of Speech tags (Hajic, Nivre , Gryuter, & Miyao, 2004):

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Figure 2.6.3. The Parts of Speech tags of UD

As we may observe from the table, in Universal Dependencies adverbs are annotated as **ADV**. Adverbs are words that usually change the meaning of verbs in terms of

time, location, direction, or manner. They may also modify adjectives and other adverbs, as in extremely briefly or arguably incorrect.

In contrast to pronouns, which can be classed as interrogative, relative, demonstrative, etc., pronominal adverbs allude to situations in context rather than identifying them explicitly. Pronominal adverbs are similarly given the ADV part-of-speech tag, but they vary from other adverbs in some ways. Be aware that some words may be classified as numbers in some languages (such as Czech), but in this system of universal tagging, they are classified as adverbs. Particularly, multiplicative and adverbial ordinal numbers, such as "for the first time", function syntactically like adverbs and are designated as **ADV** (Hajic, Nivre , Gryuter, & Miyao, 2004).

It should be noted that some verb forms, such as transgressive or adverbial participles, share characteristics and functions with adverbs. They can be categorized as VERB or **ADV** depending on the language and context like in the Kyrgyz language.

2.6.1. Dependency Parsing of Kyrgyz Adverbs to Identify their Syntactic Function

To provide syntactic annotation of Kyrgyz adverbs, we refer to Universal Dependency web-based project, which is intended to suggest universally recognized symbols for identify their syntactic relation. To prove that, let us annotate several sentences in the Kyrgyz language through the platform and see how adverbs in Kyrgyz usually function. Further, we will have practical examples of ADV annotation, its function within the sentence and consider their various lexical/morphological peculiarities.

1. *Ал **азыр** ар кандай жыйынтыка даяр.*

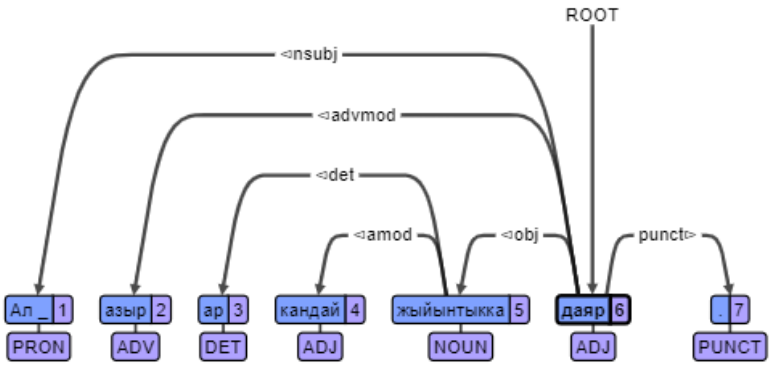


Figure 2.6.1.1. Dependency Parsing of adverbial modifier of time "***азыр***"

"<Ал _>"

"PRON" PRON @nsubj #1->6

"<азыр>"

"азыр" ADV @advmod #2->6

"<ар>"

"ар" DET @det #3->5

"<кандай>"

"кандай" ADJ @amod #4->5

"<жыйынтыкка>"

"жыйынтык" NOUN @obj #5->6

"<даяр>"

"даяр" ADJ @root #6->0

"<.>"

"." PUNCT @punct #7->6

An adverbial modifier is an (non-clausal) adverb or adverbial phrase that serves to modify a predicate or a modifier word. The term "**adverbial modifier**" refers to elements that operate like adverbs regardless of whether they are realized by adverbs, adpositional phrases, or nouns in specific morphological cases in various grammatical traditions. Adverbials that are realized as adverbs (**advmod**) are distinguished from those that are shown as nouns or adpositional phrases.

In this sentence, the adverb of time ***"azyr***" is combined with the subject (pronoun) of the sentence (that is expressed by the symbol <nsubj>) and modifies the predicate ***"dayar"*** expressed by the adjective. This relation is clearly reflected by these "#2->6" symbols, where the numbers 2 and 6 correspond to the order of the word within the sentence and these "->" symbols designate to which a first word is related to within a sentence, i.e., in this sample the adverb of time ***"azyr"*** depends from the predicate of the sentence ***"dayar"***, by which we can observe the first dependency of the adverb from the predicate within the syntax. For this case, the tag ***"advmod"*** is attached to the adverb on the syntactic level.

2. *Эл* ***эчак*** *көчүп кеткен.*

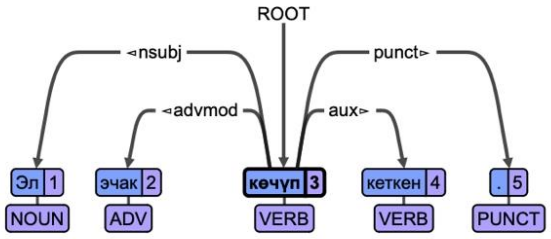


Figure 2.6.1.2. Dependency Parsing of adverbial modifier of time "***эчак***"

"<Эл>"

"эл " NOUN @nsubj #1->3

"<эчак>"

"эчак" ADV @advmod #2->3

"<көчүп>"

"коч" VERB @root #3->0

"<кеткен>"

"кет" VERB @aux #4->3

"<.>"

"." PUNCT @punct #5->3

In this example, the adverb of time ***"echak***" is combined with the subject (noun) of the sentence and modifies the predicate ***"köchüp"*** expressed by compound verb. This relation is clearly reflected by "#2->3" symbols, where the numbers 2 and 3 are attached to adverb and verb correspondingly. Here, the adverb of time ***"echak"*** depends from the predicate of the sentence ***"köchüp"***, by which we can again observe dependency of the adverb from the predicate within the syntax. The tag ***"advmod"*** is attached to the adverb.

3. *Токтогулдун ой-санаасы* ***алда кайда*** *кетти ошол тапта.*

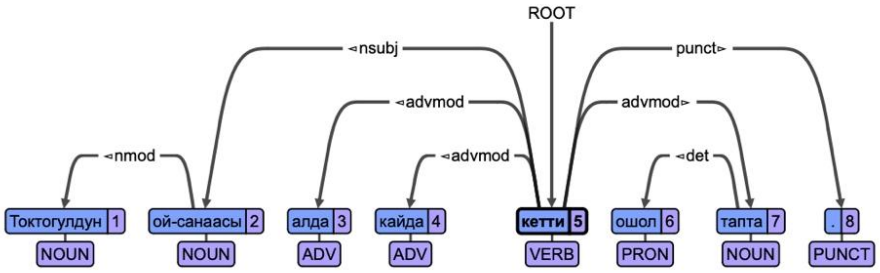


Figure 2.6.1.3. Dependency Parsing of adverbial modifier of place "***алда кайда***"

"<Токтогулдун>"

"Токтогул" NOUN @nmod #1->2

"<ой-санаасы>"

"ой-санаа" NOUN @nsubj #2->5

"<алда>"

"алда" ADV @advmod #3->5

"<кайда>"

"кайда" ADV @advmod #4->5

"<кетти>"

"кет" VERB @root #5->0

"<ошол>"

"ошол" PRON @det #6->7

"<тапта>"

"тап" NOUN @advmod #7->5

"<.>"

"." PUNCT @punct #8->5

Here, we can observe the complex adverb of place ***“alda***” and ***“kaida***” is combined with the subject (noun) of the sentence ***“oi-sanaasy”*** and modifies the predicate ***“ketti”*** expressed by the verb. As there is a complex adverb, two parts of it are numbered separately and considered as a separate unit of the sentence. This relation is clearly reflected by “#3->5” and “#4->5” symbols, where the numbers 3 and 4 are attached to the constituents of the adverb and 5 to the verb correspondingly. The complex adverb of place ***“alda” “kaida”*** depends from the predicate of the sentence ***“ketti”***, by which we can again observe dependency of the adverb from the predicate within the syntax. The 2 separate tags ***“advmod”*** is attached to the constituents of the complex adverb on the syntactic level.

4. *Эгер* ***бүгүн*** *Нүзүптүн башы кетсе,* ***эртең*** *сеники кетет.*

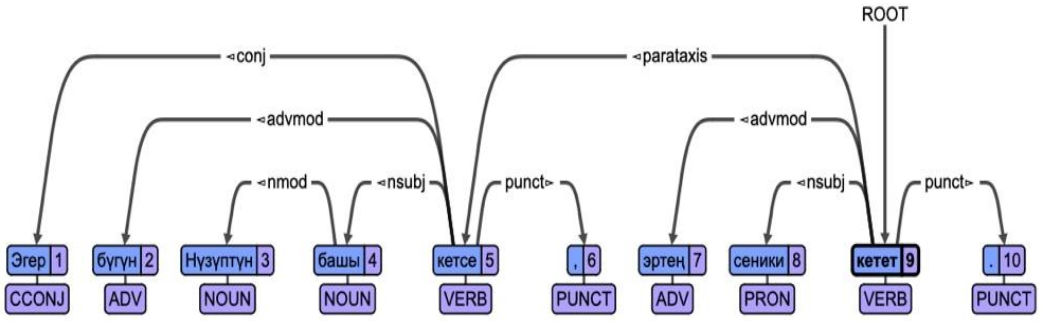


Figure 2.6.1.4. Dependency Parsing of adverbial modifier of time “***бүгүн***”, “***эртең***”

"<Эгер>"

"эгер" CCONJ @conj #1->5

"<бүгүн>"

"бүгүн" ADV @advmod #2->5

"<Нүзүптүн>"

"Нүзүп" NOUN @nmod #3->4

"<башы>"

"баш" NOUN @nsubj #4->5

"<кетсе>"

"кет" VERB @parataxis #5->9

"<,>"

"," PUNCT @punct #6->5

"<эртең>"

"эртең" ADV @advmod #7->9

"<сеники>"

"сен" PRON @nsubj #8->9

"<кетет>"

"кет" VERB @root #9->0

"<.>"

"." PUNCT @punct #10->9

There are two adverbs of time ***"bügün"*** and ***"erten"*** in the structure of a complex conditional sentence, that modify the predicates of the simple clauses ***"ketse"*** and ***"ketet"*** expressed by verbs. Their relation to the predicates is designated with the symbols **"**#2->5**"** and **"**#7->9**"** correspondingly. Being within the structure of the complex sentence, they

depend from various predicates, however the predicate #5 is dependent from the root of the sentence, from the predicate #9. Both adverbs are tagged as ***"advmod"*** according to their syntactic function.

5. *Исхак сол колун* ***алга*** *шилтеп таштады.*

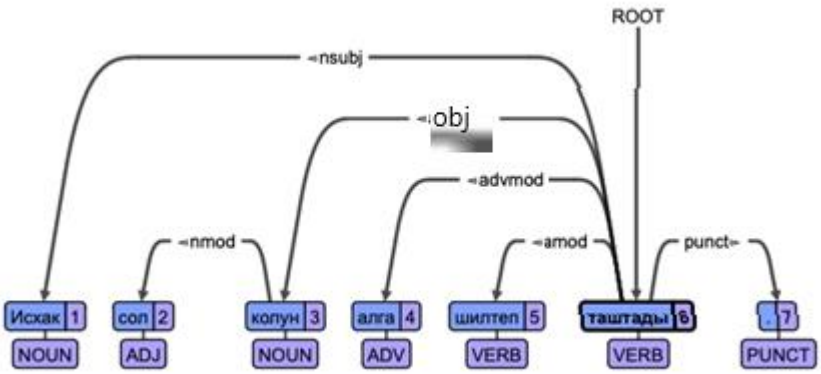


Figure 2.6.1.5. Dependency Parsing of adverbial modifier of place (direction) "***алга***"

"<Исхак>"

"Исхак" NOUN @nsubj #1->6

"<сол>"

"сол" ADJ @nmod #2->3

"<колун>"

"кол" NOUN @obj #3->6

"<алга>"

"алга" ADV @advmod #4->6

"<шилтеп>"

"шилте" VERB @amod #5->6

"<таштады>"

"ташта" VERB @root #6->0

"<.>"

"." PUNCT @punct #7->6

The adverb of place (direction in this case) ***“alga”*** as a rule modifies the part of the complex verbal predicate and the root of the sentence ***“tashtady”***. By this, that adverb is showing the direction of the action performed by the subject and conventionally combines with the verb of the sentence, that is expressed through this “#4->6” symbol. Thus, it is dependent from the predicate ***“tashtady”*** and ***“advmod”*** tag is attached to the adverb.

6. *Бүткүл суу келүүчү жабдуулары* ***эбак*** *эскирип бүткөн.*

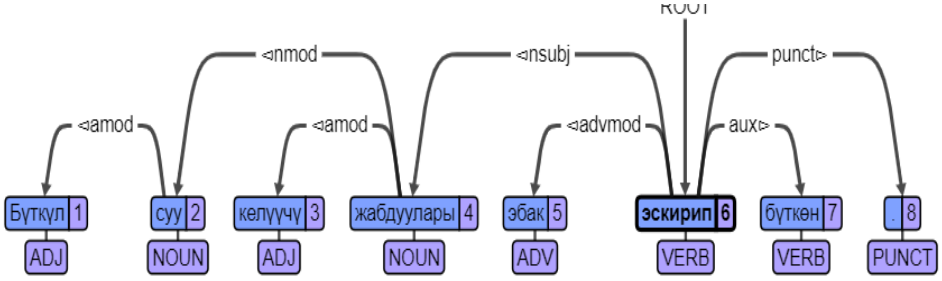


Figure 2.6.1.6. Dependency Parsing of adverbial modifier of time “***эбак***”

"<Бүткүл>"

"бүткүл" ADJ @amod #1->2

"<суу>"

"суу" NOUN @nmod #2->4

"<келүүчү>"

"кел" ADJ @amod #3->4

"<жабдуулары>"

"жабдуу" NOUN @nsubj #4->6

"<эбак>"

"эбак" ADV @advmod #5->6

"<эскирип>"

"эскир" VERB @root #6->0

"<бүткөн>"

"бүт" VERB @aux #7->6

"<.>"

"." PUNCT @punct #8->6

In this sample we see the adverb of time ***“ebak”*** modifying the predicate if a sentence, expressed through the complex verb and relating to the main verb of the predicate. That adverb is intensifying the time it was taken for the action that took place in the utterance. There is seen another similarity of adverbs to adjectives according to their degree categories. But this kind of thorough analysis will definitely support in determining the adverbs from the other parts of speech. This adverb is both combined and dependent from the verbal predicate of the sentence ***“eskirip”*** and has a tag ***“advmod”*** as well as the symbols “#5->6” displaying its relation to other parts of a sentence.

7. Эми *карынын көзү алыска кетти.*

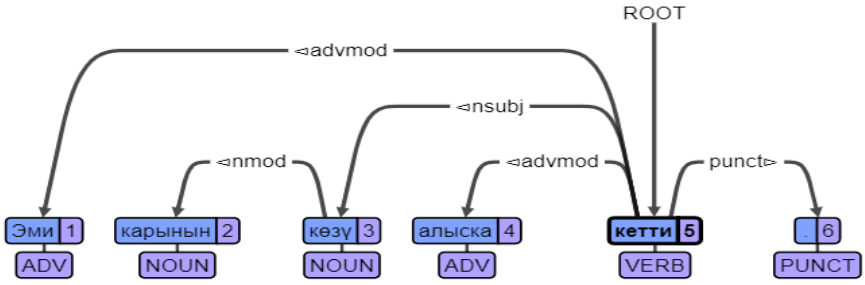


Figure 2.6.1.7. Dependency Parsing of adverbial modifier of time “***эми***”

"<Эми>"

"эми" ADV @advmod #1->5

"<карынын>"

"кары" NOUN @nmod #2->3

"<көзү>"

"көз" NOUN @nsubj #3->5

"<алыска>"

"алыс" ADV @advmod #4->5

"<кетти>"

"кет" VERB @root #5->0

"<.>"

Comment: This sentence contains two adverbs: adverb of time ***"emi"*** and the adverb of place (direction) ***"alyska".*** Both are dependent from the verbal predicate of the sentence ***"ketti".*** The second adverb of direction ***"alyska"*** is placed right before the verbal predicate and modifies it and is dependent from it. Regarding the adverb of time ***"emi"*** it is also modifying the verbal predicate designating the time of happening action. The relations of the adverbs with the predicate are shown through "#1->5" and "#4->5" symbols for ***"emi"*** and ***"alyska"*** correspondingly. Both of these adverbs have the tags ***"advmod"*** on the syntactic level.

8. *Тигине* ***алыста*** *кандай тоо кыркалары көрүнүп турат!*

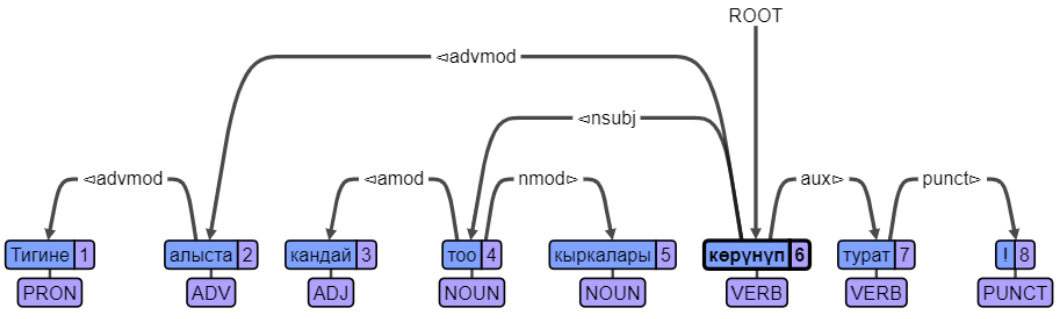


Figure 2.6.1.8. Dependency Parsing of adverbial modifier of place "***алыста***"

```
"<Тигине>"
	"тиги" PRON @advmod #1->2
"<алыста>"
	"алыс" ADV @advmod #2->6
"<кандай>"
	"кандай" ADJ @amod #3->4
"<тоо>"
	"тоо" NOUN @nsubj #4->6
"<кыркалары>"
	"кырка" NOUN @nmod #5->4
"<көрүнүп>"
	"көр" VERB @root #6->0
"<турат>"
	"тур" VERB @aux #7->6
"<!>"
	"." PUNCT @punct #8->7
```

The adverb of place (direction) ***"alysta"*** is the adverb formed with the help of accepting the case suffixes, here the locative case siffix ***-da*** has been attached to the root ***"alys"*** and formed the adverb ***"alysta".*** It is located in the beginning of the sentence and related to the

verbal predicate (expressed by the complex verb) in the passive voice ***“körünüp”,*** that can be viewed through these “#2->6” symbols**.** The dependency tool has assigned the ***“advmod”*** tag that is fully matches with its function.

9. *Атасы **араңдан зорго** макул болду*.

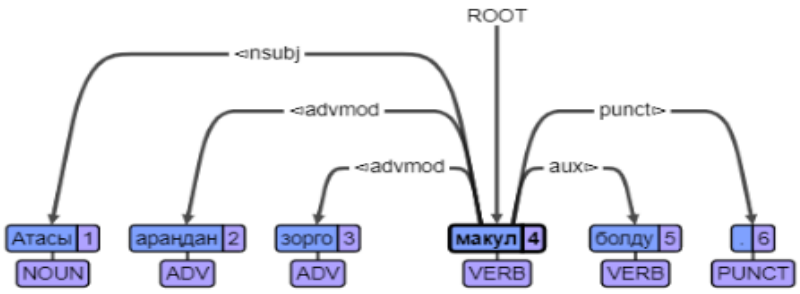


Figure 2.6.1.9. Dependency Parsing of attributive adverb of quality “***араңдан зорго***”

"<Атасы>"

"ата" NOUN @nsubj #1->4

"<араңдан>"

"араң" ADV @advmod #2->4

"<зорго>"

"зор" ADV @advmod #3->4

"<макул>"

"макул" VERB @root #4->0

"<болду>"

"бол" VERB @aux #5->4

"<.>"

"." PUNCT @punct #6->4

This sentence contains compound adverbs that have the suffixes of ablative – *"aran**dan**"* and dative *"zor**go**"* cases. They are attributive adverbs of quality as well as degree adverb, due to they are showing the degree of efforts that have been implemented for the action to be performed. They are modifying the verbal predicate and are dependent from it, that is clearly reflected by the symbols "#2->4", "#3->4". Syntactically, they are both tagged as ***"advmod"*** correspondingly.

10. *Ал өзүн **араң** токтотуп турду.*

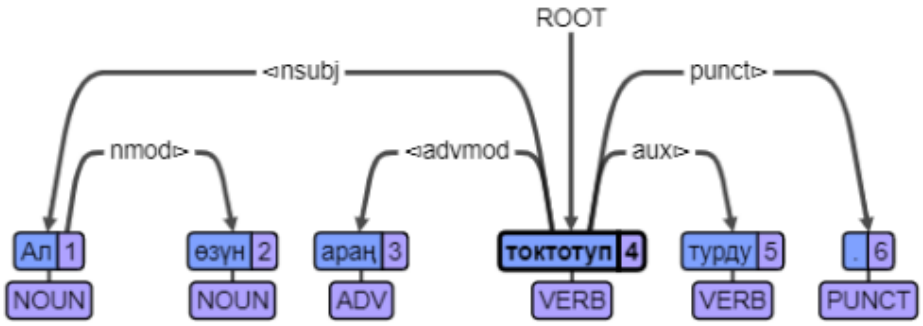


Figure 2.6.1.10. Dependency Parsing of attributive adverb of quality "***араң***"

"<Ал>"

"Ал" NOUN @nsubj #1->4

"<өзүн>"

"өзү" NOUN @nmod #2->1

"<араң>"

"араң" ADV @advmod #3->4

"<токтотуп>"

"токто" VERB @root #4->0

"<турду>"

"тур" VERB @aux #5->4

"<.>"

"." PUNCT @punct #6->4

In this sentence we can observe the inflectional, attributive adverb of quality ***"aran"***, as it specifying the way of how the action is applied by the subject of the sentence. It is located right before the verbal predicate (represented by complex verb) ***"toktotup"***, is combined with the predicate and depends from it as well. The combination is shown by the symbols "#3->4" while the dependency is marked by the tag ***"advmod"***.

11. *Эне кайсы тилде сүйлөсө, бала* ***эркизден*** *ошол тилде сүйлөйт*.

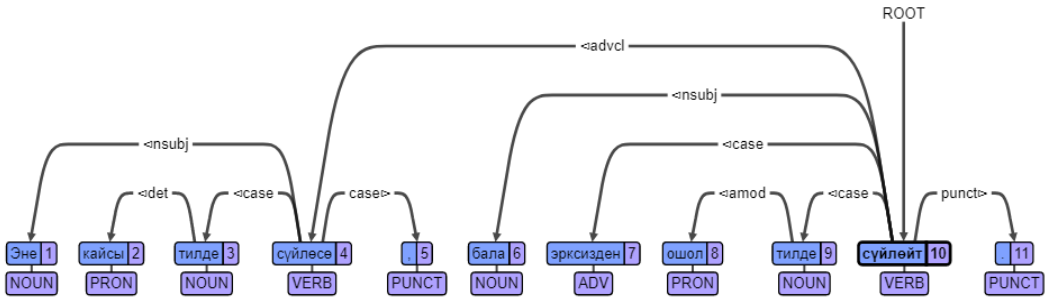


Figure 2.6.1.11. Dependency Parsing of adverbial modifier of reason "***эрксизден***"

"<Эне>"

"Эне" NOUN @nsubj #1->4

"<кайсы>"

"кайсы" PRON @det #2->3

"<тилде>"

"тил" NOUN @case #3->4

"<сүйлөсө>"

"сүйл" VERB @advcl #4->10

"<,>"

"," PUNCT @case #5->4

"<бала>"

"бала" NOUN @nsubj #6->10

"<эрксизден>"

"эрксизден" ADV @case #7->10

"<ошол>"

"ошол" PRON @amod #8->9

"<тилде>"

"тил" NOUN @case #9->10

"<сүйлөйт>"

"сүйл" VERB @root #10->0

"<.>"

"." PUNCT @punct #11->10

In this instance, we see the adverb of reason "***erksizden***". It has been formed by adding the negative form ***"-syz"*** and the suffix of the ablative case ***"-dan"*** to the root "***erk***", that is noun by itself. Here we see how adverbs can be formed morphologically, as a result of which it turned to be the adverb of reason. That's why the relation of this adverb to the root of the sentence was designated as ***"case"***, and not as adverbial modifier, even though morphologically it was designated correctly. That is the peculiarities of morphologically formed adverbs. So, this adverb explains the reason which caused the action in the independent clause within the conditional sentence. It is dependent from the predicate of the independent clause ***"süilöit"*** and therefore following symbols "#7->10" have been attached to the adverb on the syntactic level.

12. *Жер* ***алдындан атайлап*** *алып келген катыным.*

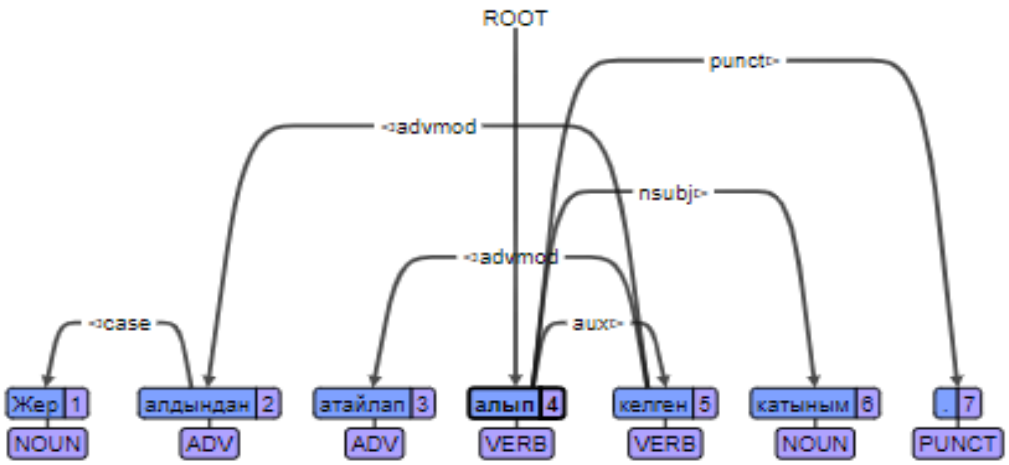


Figure 2.6.1.12. Dependency Parsing of adverbial modifier of place “***алдынан***” and purpose “***атайлап***”

"<Жер>"

"Жер" NOUN @case #1->2

"<алдындан>"

"алды" ADV @advmod #2->5

"<атайлап>"

"атайы" ADV @advmod #3->4

"<алып>"

"ал" VERB @root #4->0

"<келген>"

"кел" VERB @aux #5->4

"<катыным>"

"катын" NOUN @nsubj #6->4

"<.>"

"." PUNCT @punct #7->4

In this sentence we can see 2 adverbs, one is the adverb of place ***“aldynan”*** and second is adverb of purpose ***“atailap”*** and both are derivational adverbs. They do have the case affixes in their structure, e.g., the first adverb was formed by uniting the noun root and the ablative case affix “***aldy+nan***”, while the second was formed through joining together the adverb “***atayi***” and the affix “***-lap***” that helps to form derivational adverb from other adverbs. Regarding the syntax, the first adverb ***“aldynan”*** is combined with the noun ***“jer”*** and is in ablative case. This relation is shown through the symbols “” This adverb is related with the noun “***jer***” The adverb “***atailap***” specifies exactly the way and the purpose of action that was performed and is related to the main verb of the predicate (again expressed by the complex verb) within the sentence ***“alyp”***. This is seen in these “#3->4” symbols while syntactically it modifies the predicate thus has the tag “***advmod***”.

13. *Байкуш бала* ***айласыздан*** *ыйлап кала берет*.

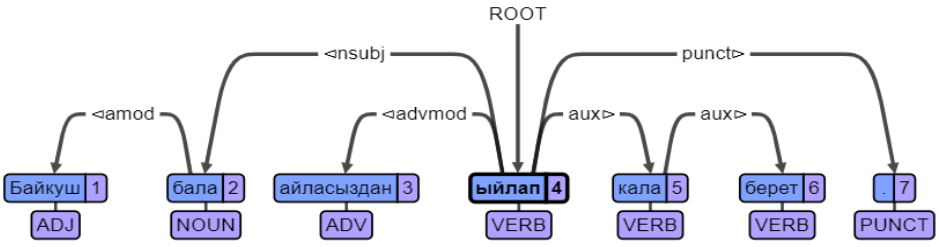


Figure 2.6.1.13. Dependency Parsing of adverbial modifier of reason “***айласыздан***”

"<Байкуш>"

"байкуш" ADJ 2 @amod #1->2

"<бала>"

"бала" NOUN 4 @nsubj #2->4

"<айласыздан>"

"айласыздан" ADV 4 @advmod #3->4

"<ыйлап>"

"ыйл" VERB @root #4->0

"<кала>"

"кал" VERB @aux #5->4

"<берет>"

"бер" VERB @aux #6->5

"<.>"

"." PUNCT @punct #7->4

The adverb of reason ***"ailasyzdan"*** is also formed by adding the negative form ***"-syz"*** and the suffix of the ablative case ***"-dan"*** to the root "***aila***", that is noun. Together, it becomes an adverb of reason. It is combined with the root of the sentence "***yilap***" (represented by verbal adverb) and modifies it, due to which it has got the tag ***"advmod".*** The relationship of the adverb and the root of the sentence is reflected through this "#3->4" symbols.

14. ***Чапчаң*** *келди боз жорго.*

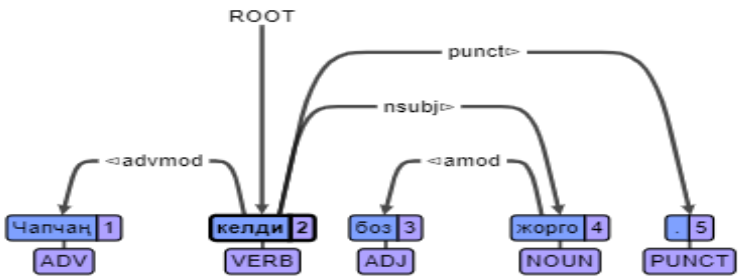


Figure 2.6.1.14. Dependency Parsing of attributive adverb of quality "***чапчаң***"

"<Чапчаң>"

"Чапчаң" ADV @advmod #1->2

"<келди>"

"кел" VERB @root #2->0

"<боз>"

"боз" ADJ @amod #3->4

"<жорго>"

"жорго" NOUN @nsubj #4->2

"<.>"

"." PUNCT @punct #5->2

The inflectional adverb ***"chapchan"*** is attributive adverb of quality. It describes the way of how fast the horse arrived the destination. It is modifying the verbal predicate of the sentence "***keldi***" thus has the tag "***advmod***" on the syntactic level. Speaking on the dependency, this adverb is related to the predicate, that can be seen by these "#1->2"

symbols. So, there is the conventional function of attributive adverb of quality modifying verbs.

15. Баскандан эринген, ***бир кезекте*** жүгүрөөр.

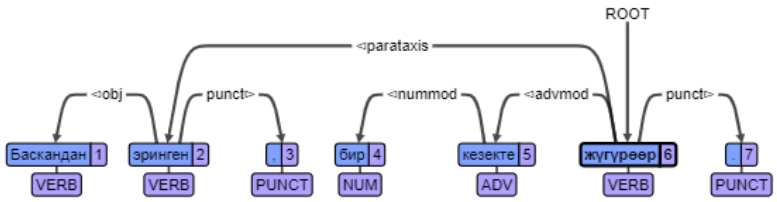


Figure 2.6.1.15. Dependency Parsing of adverbial modifier expressed by compound adverb "***бир кезекте***"

"<Баскандан>"

"Бас" VERB @obj #1->2

"<эринген>"

"эрин" VERB @parataxis #2->6

"<,>"

"," PUNCT @punct #3->2

"<бир>"

"бир" NUM @nummod #4->5

"<кезекте>"

"кезекте" ADV @advmod #5->6

"<жүгүрөөр>"

"жүгүр" VERB @root #6->0

"<.>"

"." PUNCT @punct #7->6

In this sentence we consider the complex adverb of time ***"bir kezekte"*** but was separately annotated on the morphological level as numeral for "***bir***" and adverb for "***kezekte***" though they bring together one lexical meaning. The adverb "***kezekte***" is combined with the predicate of the sentence "***jügüröör***" and modifies it, i.e., dependent from the predicate, that is seen by the tag "***advmod***". The relation of the adverb to the predicate is clear from these "#5->6" symbols. At the same time, the numeral "***bir***" is dependent from the adverb "***kezekte***" and expressed through "#4->5" symbols along with the tag ***"nummod"***, showing that numeral determines the adverb. This correlation shows that this complex adverb "***bir kezekte***" are one unit on the syntactic level.

2.7.The Table of Possible Tags for the Additional Categories of Adverbs

After considering the most of the types of adverbs on both morphological and syntactic levels, the task of working out the possible tags for the additional morphological categories of adverbs becomes clearer. This has great theoretical value from the perspective of the development both Corpus linguistics in Kyrgyzstan and machine-learning branch in particular. Of course, the proposed tagset is only little step at improving the process of teaching computers the natural languages, but accurately developed and attached tags will ensure the adequate perception of the natural language (Kyrgyz in our context) by computer. The more accurate and correct tags are developed, the better will be output of the language analysing process by the computer. In order to have these principles preserved, following criteria have been considered.

Tagset of other categories for the adverbs of the Kyrgyz language were scrutinized taking into account all the peculiarities discussed above. As the most basic and easy for the automatic-processing category was identified the division of adverbs according their composition, then the lexical and semantical classification of adverbs in the Kyrgyz language. Bearing in mind the peculiarities of the adverbs in Kyrgyz, the second level category after the PoS-tagging will be the division of adverbs according their composition, after which will follow lexical/semantical classification, including attributive adverbs and adverbs functioning as adverbial modifiers.

Classification of adverbs according to their composition and some suggestions for their detailed morphological annotation:

Table 2.7.1. Tags for the 2nd level category of adverbs

Category name	**Proposed tag**
Simple adverbs	<simp>
Compound adverbs	<comp>

(P.S. – simple adverbs can be further classified as inflectional or derivational depending on its type, but that requires much more detailed analysis).

Table 2.7.2. Tags for the 3rd level category of adverbs

Category name	**Proposed tag**
Adverbs of adverbial modifiers	<advmod>
Attributive adverbs	<attr_adv>

As these two large groups have been divided, each contain several important internal subdivisions so that they also do require separate tags, such as:

Table 2.7.3. Tags for the 3rd level category of adverbs

Category name	Proposed tag
1. For the adverbs of adverbial modifiers: a) Adverbs of time	<tm>
b) Adverbs of place	<plc>
c) Adverbs of reason	<rsn>
d) Adverbs of purpose	<purp>
2. For the attributive adverbs: a) Adverbs of quality	<qual>
e) Adverbs of number	<num>

Deduction on Chapter 2

So, we have looked through the adverbs as an indepedent part-of-speech and all the characteristic peculiarities regarding them. They can be formed both morphologically and syntactically, sometimes they may appear some ambuguities regarding their sense and shape. The issue of adverb has been raised only in the middle of the XX century, even though more than half of the centure has passed since then, the issue of adverbs in the Kyrgyz language has not been resolved completely yet. Starting the second half of the XX century such prominent liguists, as Davletov S., Kudaibergenov S. became the pioneers in considering the adverbs as an independent part-of-speech among other Kyrgyz linguists at that times. Over the decades they have been evolving and developing and even up to day their formation path is still in progress. For now adverbs in Kyrgyz have some distinct criterea according to which it is possible to distinct them from other part-of-speech, but surely that is not applicable entirely whole word class. Some cases are still generates argues and need to be thought over and considered thoroughly.

Bearing in mind this fact, this paper intends to make its own contribution in this direction and tries to give correct information regarding the adverbs and their morphological tags on the basis of corpus-based study of adverbs. The following section is going to be practical one where the syntactic relation of adverbs and their functions will be under the focus, which will maintain or reject some other peculiarities of adverbs described above.

We considered one of the syntactic parsers Universal Dependencies that helps to unify the grammars of world languages into one standardized structure, particularly focus was made on the adverbs of the Kyrgyz language and their syntactic annotation in Universal Dependency Platform with the help of unified tagsets. For annotating the adverbs, we were needed to make research on the unified tools in Universal Dependencies.

Furthermore, we studied the Treebank approach while annotating sentences that contained adverbs in the Kyrgyz language.

In this section, we tried to give the syntactic treebanks in order to investigate the functions and the features of adverbs in the Kyrgyz language. Through this analysis we came to the conclusion that the adverbs in the Kyrgyz language might play different roles as a member of a sentence, that was mentioned previously in the 2nd chapter. It was found out that the adverbs could be combined not only with the verbs, but with nouns as well preserving its function of modifying the verb/ predicate. Based on these examples, we have tried to show the diversity of the adverbs through Treebank structures. The possible variants of tagging the additional grammatical categories of adverbs were proposed for their implementation into the Kyrgyz Corpus built on the basis of the CQP Web.

We hope that the examples we have given above will help other researchers in their study of adverbs in the Kyrgyz language and promote further contribution into the development of the Corpus linguistics in the Kyrgyz Republic.

CONCLUSION

The thesis work revealed the nature of adverbs and their lexicogrammatical features in the framework of the Corpus linguistics. Corpus Linguistics is a multidimensional area. It is an area with a wide spectrum for encompassing all diversities of language use in all domains of linguistic interaction, communication, and comprehension. The introduction of corpus in language study and application has incorporated a new dimension to linguistics. In principle, *Corpus Linguistics* is an approach that aims at investigating language and all its properties by analyzing large collections of text samples. This approach has been used in a number of research areas for ages: from descriptive study of a language, to language education, to lexicography, etc. In general, it covers large amount of machine-readable data of actual language use that includes the collections of literary and non-literary text samples to reflect on both the synchronic and diachronic aspects of a language.

It was found out, that the uniqueness *corpus linguistics* lies in its way of using modern computer technology in collection of language data, methods used in processing language databases, techniques used in language data and information retrieval, and strategies used in application of these in all kinds' language-related research and development activities. The Kyrgyz Corpus has been developed on the basis of the project that was implemented in cooperation with the Department of Simultaneous Interpretation of Kyrgyz-Turkish Manas University and the University of Saarland. The project started in 2018 and is still going on under the continual efforts of academic staff of the departments of two universities.

So, the following objectives were implemented:

a) We studied and defined the basic concepts applied in the Corpus Linguistics: the notion of linguistic corpus, how a corpus is designed, what criteria and principles are applied in compiling a corpus as well as the broad classification of linguistic corpora;

b) We studied all the aspects of adverbs in the Kyrgyz language provided by well-known Kyrgyz linguists (Abduvaliev I., Kudaibergenov S., Davletov S.) from both

morphological, syntactic perspectives, including their structure, ways of formation of adverbs and lexical-semantical classification;

c) We have selected the adverbs in the Kyrgyz corpus, tagged them within the selected sentences and identified their additional aspects: adverbial modifiers of various types (time, place, reason, purpose), and attributive adverbs. It should be noted, that the Kyrgyz corpus lacks these categories for annotation of adverbs;

d) With the application of the Kyrgyz Corpus, we provided quantitative (statistical) data on various types of adverbs with their frequency of use in the Corpus as a result of which the adverbs of measure are the most prevailing among others;

e) Syntactic analyses of adverbs using the Universal Dependencies web-based syntactic parser (the platform for all languages that has a common unified scheme for syntactic analysis, which facilitates the cross-linguistic research) has been applied to reveal dependency relationship of adverbs within the considered sentence;

f) Having analysed all the selected adverbs of the Kyrgyz language, the missing aspects for adverbs are suggested for the introduction them into the Kyrgyz corpus;

g) A list of additional “tags” (obtained in the course of analyses) for the grammatical aspects have been offered to improve the representation of adverbs in the Kyrgyz Corpus that will be suggested for the future works dealing with Kyrgyz adverbs as well as in creating a guidebook for the Corpus use.

ДИССЕРТАЦИЯНЫН КЫСКАЧА МАЗМУНУ

Кыргыз тилинин корпусундагы тактоочтордун энтектелиши

Бул диссертация корпустун негизинде жүргүзүлүп, кыргыз тилиндеги тактоочтордун актуалдуу жана туура сүрөттөлүшүн берүү аркылуу кыргыз тилинин корпусунун өнүгүшүнө салым кошуу болуп саналат. Англис жана кыргыз тилдериндеги тактоочтордун грамматикалык өзгөчөлүктөрүн, ошондой эле ар кандай таанылган платформалар тарабынан сунушталган ар кандай стандарттык энтектөө символдордун эске алуу менен, бул изилдөө кыргыз тилдин корпусунда тактоочторду аннотациялоо боюнча тиешелүү сунуштарды карайт. Корпустун сапатын өнүктүрүү менен кыргыз тилинин корпусунун өнүгүүсүн алдыга илгерилөөдө тездетүү максатында кыргыз тилиндеги тактоочтор кеңири изилденет. Бул процесстин мааниси чоң, анткени алардын морфологиялык аннотациясы (энтектелиши) ар кандай корпустун лингвистикалык маалыматтарынын синтаксистик жана семантикалык деңгээлдеринде андан ары талдоо үчүн негиз болуп саналат. Диссертациянын темасын тандоодо даана тактоочтордун терең изилдениши түрткү болуп, диссертациянын практикалык бөлүгүндө кыргыз тилинин корпусундагы тактоочторду тагыраак маалымат менен камсыздандыруу максатында аларга грамматикалык аспекттери боюнча кошумча энтектердин варианттары сунушталууда. Бул маселе тактоочтордун сүйлөмдөгү башка сөздөр менен болгон синтаксистик байланыштарын көрсөтүү үчүн аткарылып, дагы да тактоочтордун морфологиялык дэңгээлде туура белгиленгенине көз чаптырмакчы. Диссертациянын жыйынтыгы компьютердик программалардын тактоочторду туура ажырата билиши менен түздөн-түз байланыштуу.

Изилдөөнүн актуалдуулугу тактоочтордун корпустун негизинде жана корпуска багытталган ыкмаларын колдонуу менен изилденгенинде жатат.

Изилдөөнүн максаты кыргыз тилинин корпусундагы тактоочторду жана алардын морфологиялык аннотациясын (энтектелишин), ошондой эле алардын башка грамматикалык өзгөчөлүктөрүн жана мүнөздөмөлөрүн изилдөө болмокчу. Ошондуктан, бул иш корпустун негизинде жүргүзүлгөн изилдөө болуп саналат. Кыргыз тилиндеги тактоочтордун табияты татаал болгондуктан, диссертациянын максаты кыргыз тилиндеги тактоочторго тийешелүү болгон кошумча морфологиялык энтектерди иштеп чыгып, сунуштоо, анткени кыргыз тилинин корпусун аннотациялоо боюнча келечектеги иштерди жүргзүүдө аннотациялоо этабында кандайдыр бир "түшүнүксүздүктөр" же кээ бир эки маанилүүлүктөрдү туудурган тоскоолдуктар пайда болот. Коюлган максатка жетүү үчүн төмөнкүдөй милдеттер коюлган:

1. Корпустук лингвистикада колдонулган түшүнүктөрдү окуп, изилдеп чыгуу;
2. Кыргыз тилиндеги тактоочтордун аспекттерин карап чыгуу;
3. Кыргыз тилинин корпусуна киргизилүүчү тактоочторду алдын-ала талдап, керектүү түргө келтирилип, табигый тил иштетүү процесстерине даярдоо;
4. Кыргыз тилинин корпусундагы тактоочторго жана алардын энтектелишине статистикалык (квантитативдик) маалымат менен камсыздандыруу;
5. Universal Dependencies (Универсалдуу Багыныңкылардын) платформанын инструментарийлерин колдонуу аркылуу сүйлөм тутумундагы башка сөздөр менен болгон байланыштарын аныктоо максатында синтиксистик талдоосун жүргүзүү;
6. Кыргыз тилинин корпусунда жайгашкан тактоочтордун кошумча аспекттерин эске алуу менен, грамматикалык жактан туура болгон энтектерди камсыздандыруу;
7. Практикалык бөлүгүндө аткарылган иштин натыйжасында сунуштоого иштелип чыккан тактоочтордун кошумча грамматикалык аспекттерин чагылдырган энтектерин тизмесин түзүп чыгуу.

Изилдөөнүн предмети: Кыргыз тилинин корпусунан алынган тактоочтор.

Изилдөөнүн объектиси: Кыргыз тилинин корпусундагы тактоочтордун морфологиялык энтектелиши жана корпустагы аннотациялоо процессине байланыштуу болгон көйгөйлөр.

Изилдөөнүн илимий жаңылыгы: Кыргыз тилиндеги тактоочторунун корпус лингвистикасынын көз карашынан изилденгендигине жана табигый тилди иштетүү менен тыгыз байланышкандыгына байланыштуу. Диссертация корпустук лингвистика, Универсалдуу Багыныңкылар (Universal Dependencies (UD) platform) менен Turkic-Lexicon Apertium веб-платформалары сымал жаңы методологияларды жана ыкмаларды колдонуу аркылуу кыргыз тилинин тактоочторун комплекстүү изилдөөгө арналган. Корпуска негизделген ыкмасынын артыкчылыгы корпустагы бардык сөздөр үчүн берилген жыштык маалыматтардын ал сөздөр кезиккен тилдик контексттери менен курчалагындыгынын айкалышында жатат. Корпуска негизделген ыкма корпустан такчтоочтордун статистикалык маалыматты ортого алып чыгууга мүмкүндүк берип, "Универсалдуу багыныңкы" платформасы болсо синтаксистик дэңгээлде тактоочторго жана алардын аткарган функцияларына кылдаттык менен талдоо жүргүзүүгө өбөлгө түзүп берген.

Изилдөөнүн максатына жетүү этабында, төмөндөгү методдор жана усулдар колдонулду:

1. Салыштырма анализ: кыргыз жана англис тилдериндеги тактоочтордун окшоштугун жана айырмачылыктарын аныктоо үчүн колдонулду;

2. Квалитативдик анализ: Кыргыз тилиндеги тактоочтордун мүнөздөмөсүн, алардын өзгөчөлүктөрүн жана классификацияланышын сыпаттоо үчүн колдонулду;

3. Квантитативдик анализ: Тактоочтордун категорияларга жана классификациясына жараша алардын кыргыз тилинин корпусунда кезиккендигин жана жыштыгын сандык түрдө көрсөтүү үчүн колдонулду;

4. Корпуска негизделген анализ: Диссертацияда көрсөтүлгөн тактоочтор кыргыз тилинин корпусунан алынып, алардын морфологиялык энтектелиши корпуста кандай көрсөтүлгөндүгү талкууланат;

5. Терең морфологиялык анализ жалаң гана тактоочтордун морфологиялык аспекттерине байланыштуу маселелерди иликтөөгө колдонулду;

6. Буларга кошумча, жогоруда берилген методдордун катарын бул маселелерди чечүүгө арналган атайы веб-тиркемелерди да көрсөтсөк болот. Мисалы, кыргыз тилинин корпусун энтектөөдө жана тактоочторго энтектерди иштеп чыгууда Turkic Lexicon Apertium Toolkit инструменттеринин колдонулушу;

7. Ал эми алардын ситаксистик байланышын багынынкы абалын контекстте көрсөтүү үчүн атайы сунушталган Annotatrix веб-тиркемеси колдонулду.

Изилдөө гипотезасы. Корпустук лингвистикынын алкагында кыргыз тилинин тактоочторунун изилделениши синтаксистик багыныңкы байланыштарын аныктоо аркылуу тактоочтоорго туура делген морфологиялык аннотацияны (энтектерди) камсыздайт.

Диссертациянын теориялык жана практикалык мааниси. Бул иштин теориялык баалуулугу жүргүзүлгөн анализдин негизинде морфологиялык деңгээлде иштелип чыгып, тактоочтун кошумча грамматикалык аспекттери боюнча энтектерин сунуштоо аркылуу кыргыз тилинин корпусундагы тактоочтордун учурда бар болгон аннотациясын (энтектелиши) толуктоо ниетинде камтылган. Ал эми иштин практикалык маанисинин маңызы болсо чет тилди окутуу методикасында Data-Driven Learning (көлөмдүү маалыматтарды изилдөө аркылуу үйрөнүү) аттуу ыкмасын

тартуулоодо жатат. Бул ыкмада тилдин берилиштер катары окутулушуна маани берилип, студенттер болсо мугалим берген тапшырмаларга ылайык ишин жүргүзгөн изилдөөчүлөрдүн кызматын аткарып калышат. Бул ыкмага берилиштерди иштетүү, статистикалык ыкмалардын колдонулушу жана жүргүзүлгөн иштин натыйжасында чогулган жыйынтыктардын үстүнөн талдоо жүргүзүлүшү кирет.

Биринчи бөлүмдө корпустук лингвитика тууралуу теориялык маалыматтар камсыздалмакчы. Корпуста тилдин өз ара аракеттенүүсүнүн ар кандай чөйрөлөрүндө колдонулган ар кандай тилдик варианттарды камтыган репрезентативдик текст үлгүлөрүнүн чоң жыйнагы камтылган. Ал компьютерге шайкеш келет, изилдөө жана колдонуу үчүн колдонулушу мүмкүн, баштапкы тилдин өкүлү, адамдар жана машиналар тарабынан иштетилиши мүмкүн, чексиз маалыматтарга ээ, курууда жана көрсөтүүдө системалуу. Корпус өз ичине тилдик өз ара аракеттенүүсүнүн ар кайсы тармактарында колдонулган тилдик варианттарын чагылдырган, демек реперезентативдүү тексттерди үзүндүлөрүнүн ири жыйнагын камтыйт. Ал компьютерлер менен шайкеш келип, изилдөөлөр үчүн колдонулуп, баштапкы тил үчүн репрезентативдүү боло билип, адамдар да компьютерелердин тарабынан дагы иштетилиши мүмкүн. Корпус өз ичинде сан жактан чектелбестен, түзүлүүдө системалык иүнөзгө ээ (Dash N. , 2008: 145-160p.).

Бирок бүгүнкү күнгө чейин корпустук лингвистика деген түшүнүккө толук кандуу так аныктама дээрлик жок. Корпус лингвистикасы лингвистикалык маалыматтарды гана камтыбастан, ар кандай аспектилерде көптөгөн лингвистикалык изилдөөлөрдү жүргүзүү үчүн инструмент катары колдонулушу мүмкүн. Ошентип, корпустук лингвистика кантип каралышы керек деген суроо ортого келип чыгууда. Ар кандай лингвисттер бул термин боюнча ар кандай пикирлерди айтышат, бирок жалпысынан корпустук лингвистика “метод", "методология", "ыкма", "теория" же "парадигма" деген

түшүнүктөрдүн ортосунда калып калган, себеби корпустук лингвистика тил илиминин бир багыты катары да каралышы мүмкүн, изидөөлөрдө колдонулган ыкма катары, тилдин ар кайсы тармактарга жараша ар кандайча колдонулган тилдик варианттарын чагылдырган теория, же өз алдынча илимий изилдөөнү талап кылган парадигма катары дагы каралышы мүмкүн болуп калган. Бирок биз корпустук лингвистика дегенде компьютерлер окуй ала турган, тилдик ар кайсы деңгээлде энтектелип, бир түргө келтирилген, көлөмү ири болгон тексттер аркылуу тилдик ар кандай кубулуштарын изилдөөнү камтыган түшүнүккө жакынырааkпыз.

Ал эми учурку тапта, “тилдик корпус” деген түшүнүккө ири көлөмдө болгон, унификацияланган, компьютер окуй ала турган түргө келтирилген, түзүмү ирээттелген, тилдин ар кайсы дэңгээлде энтектелип, филологиялык көз караштан туура делген жана атйын тилдик изилдөөлөрдү жүргүзүү үчүн иштелип чыккан тилдик маалыматтардын топтому деген аныктама берилип жатат (Zakharov, 2013). Демек, лингвистикалык корпусту түзүү үчүн түзүлүшүнүн максатына жараша аныкталган негизги критерийлери бар, ошолорго ылайык курулуп, коюлган талаптарга жооп бериши керек. Бул критерийлерге көлөм, сапат, репрезентативдүүлүк, ичинде камтылган тексттердин түзүмү боюнча болгон жөнөкөйлүгү, корпуста камтылган бирдиктердин (сүйлөмдөрдүн) көлөмү (узундугу) тең болгодугу, маалыматтын корпустан жеңил алынып чыгышы, ичиндеги маалыматтын тастыктала билүүсү, корпус өзү бара бара көлөмү жактан үзгүлтүксүз чоңоюшу менен корпустун ичичнде сакталган маалыматы тууралуу болгон маалыматтын документиштирилген түрдө болушу кирмекчи.

Жогоруда корпустардын түзүлүш принциптери тууралуу айтылып кетсе, тилдик корпусту түзүп жатканда, ага кирүүчү тексттер да өз алдынча табигый тил иштетүү процессинин бир катар баскычынан өтүп, анан гана корпуска кирет. Ал баскычтар төмөнкүдөй.

Токенизация (б.а. графемдик талдоо) – табигый тилде киргизилдген тексттерде символдордун узун бир сабын маниге ээ болгон бирдиктиреге (лексемаларга, сөз формаларга) ажыратуу процессин айтабы. Ар бир ошондой ажыратылган бирдикти “токен” деп аталат. Бул тексттин иштетилүүсүнүн алгачкы жана тексттин кийинки иштетилүүсүнө эң керектүү баскычы. **Лемматизация** бул морфлогиялык энтектөө процессине тыгыз байланышта болгон өзгөчө милдеттүү, тексттеги сөз формаларынын баштапкы формсына келтирүүчү процесс. Текстте болгон ар бир сөз бирдигинин уңгусун чыгарган процесс. Андан кийинки болуп өтүүчү процесс **стемминг** деп аталат. Анын маңызында сөздүн негизинин аныкталышы жатат. Лемматизациядан болгон айырмасы стеммингде сөздүн түпкү уңгусуна жетүүсүн көздөбөстөн, сөздүн өзгөрмө бөлүгүн ажыратып, мындан нары иштөө үчүн сөздүн өзгөрүлбөс негизини гана калтырат. Андан кийинки жүргүзүлгөң процесс **парсинг** деп аталат. Бул мурункуларга караганда алда кыйла татаал бир процесс, себеби изилденүүчү тилдин лексемаларын (сөздөрү) ошол тилдин формалдуу грамматикасы менен сызыктык (линейный) ырааттуулугу менен салыштырылмакчы. Бунун натыйжасында синтаксистик дарагы (сүйлөмдүн тутумундагы болгон багынынңкы байланыштардын түзүмү) түзүлүп калат. Бир сүйлөмдүн алкагында ичиндеги компоненттер бири бири менен кандай байланышта болгонун жана кандайча бири биринен көз каранды болгону көрүнүп калат.

Ар түрдүү лингвистикалык маселелелрди чечүү үчүн бир гана ири көлөмдө болгон тексттердин жыйындысына ээ болуу жетишсиз. Корпустун ичиндеги тексттер ар түрдүү лингвистикалык жана экстралингвитикалык маалыматка ээ болуусу шарт. Мындай муктаждыктын негизнде энтектөө деген процесс пайда болгон (“энтек” деген термин Б. Шаршембаев менен Т. Садыков биргеликте Лейпциг корпусунда жайгашкан кыргыз тилинин корпусу үчүн сунушталган энтектер боюнча жазган макалада биринчи жолу

кезигет дагы, англис тилиндеги “tag” деген сздүн ордуна колдонууга сунушталат). **Энтектөө процесси** (англис тилиндеги “tagging” деген сөз менен уңгулаш) деген текстке жана анын маанилүү бөлүктөрүнө атайын энтектерди (белгилерди) коюлуучу процесс. Бул энтектер өз ара экстралингвитикалык дэңгээлде берилиши мүмкүн дагы алардыгн ичине текстти жазган автор, текст тууралуу, анын жанры ж.б.у.с. маалыматты билдирген энтектер берилиши мүмкүн. **Морфологиялык энтектөө** ошондой эле сөз түркүмдү аныктоочу белгилерди коюу процесси деп да аталат. Бул процесстин негизги максаты тексттеги бар болгон ар бир лексемага/сөзгө анын кайсы сөз түркүмүнө тиешелүү экенине карай аларга ошол сөз түркүмүнө дал келген энтектерди белгилөө. Бул өзүнүн структурасына ээ, ал леммадан, сөз түркүмдү аныктоочу белгиден жана ал түркүмгө караштуу болгон грамматикалык аспекттерин көрсөткөн белгилерин камтыйт. Дүйнөлүк дэңгээлде алгач электрондук түрдө түзүлгөн англис тилинин америкалык вариантын чагылдырган Браун корпусундагы (Brown Corpus of American English) сөздөр алгач аннотацияланган варианты 1980-жылы пайда болгон. Анда сөздөрдүн баары уңгуларына дейре ажыратылып, ал сөздөрдүн синтаксистик функциялары үстүртөн белгиленип кеткен (Zakharov, 2013, p. 34).

Синтаксистик энтектөө – корпуста тексттер морфологиялык деңгээлдеги энтектелишинен кийин синтаксистик анализ дагы синтаксистик “даректардын” жардамы менен ишке ашырылат. Башкача айтканда, синтаксистик энтектөө деп сүйлөмдүн тутумунадгы орун алган сөз айкаштарын аныктоочу, атайын сөз айкаштык мүнөздөмөнү берүүчү, сүйлөм ичинде өз ара байланыштарын көрсөтүп, сүйлөмдүн ичиндеги сөз айкаштар өзүндө камтыган маалыматы менен бири бирине кандайча таасир тийгизгенин көрсөткөн процесс деп айтсак болот (Dash N. S., 2021). Синтаксистик талдоонун алкагында “трибанк” (“treebank” – “tree” – дарак, “bank” - контейнер) деген термин сүйлөмдөгү сөздөр, сөз айкаштар же татаал

сүйлөмдүн тутумундагы жөнөкөй сүйлөмдөрдүн өздөрүнүн бири бири менен байланышкан жолу боюнча талданып, ал өз ара ички байланыштар “даракка” окшогон синтаксистик анализдин схемасы аркылуу чагылдырып көрсөтүлөт.

Морфологиялык аннотациядан айырмаланып, синтаксистик түзүлүштүн чагылдыруу жолдору жана синтаксистик байланыштар унификацияланган түргө келтириле элек. Буга байланыштуу синтаксистик теориялар менен формализмдер бир нече түрлөрү бар, алар төмөндө берилет:

7. Багыныңкы грамматика;
8. Фразалык түзүм грамматикасы;
9. Структуралык грамматика;
10. Сүйлөм мүчөлөрүнүн салттуу синтаксистик доктриналары;
11. Функционалдык грамматика;
12. Синтаксистик семантика ж.б.

Семантикалык энектөө – энтектөөнүн бул түрүндө тексттин ичинде болгон сөздөрдүн семантикалык сапаттары (б.а. сөздүн маанисинин энтектелиши) же тексттин ичиндеги айрым турган сөздөрдүн ортосунда болгон семантикалык байланыштары белгиленет. Семантикалык энтек колдонулган контексттеги сөздүн эң ылайыктуу маанисин көрсөтүү үчүн тексттин бир бөлүгүнө берилет (Löfberg, 2003).

Кыргыз тилинини корпусу керектүү маалыматты издеп табууга жана маалыматты туура издөөгө өбөлгө түзгөн вебге негизделген башкаруу системасына ээ. Кыргыз тилинин корпусун түзүү багытында алгач кадамдары Кыргыз-Түрк Манас университет менен Заарланд университтин ортосундагы DAAD алмашуу программасынын алкагынын, Элке Тэйк профессордун жетекчилигинин алдында ишке жүзүнө ашырылган. Баштапкы этапта Кыргыз тилинин корпусу 1, 205, 888 сөздөн туруп, ар кайсы жанрдагы (дастандар, романдар, жомоктор, аңгемеоер ж.б.) 84 текстти камтыган. Ал эми бул эки

университет тарабынан жасалган үзгүлтүксүз, талыкпаган жана жемиштүү эмгектин аркасы менен 2022 жылы кыргыз тилинин корпусу кеңейип, учурда көлөмү 2, 493, 894 сөз түзүп туруп, 1019 ар түрдүү жанрдагы, морфологиялык деңгээлде энтектелген тексттерди камтыган.

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

Диссертациянын **экинчи бөлүмү** толугу менен кыргыз тилиндеги тактоочтордун терең, тилдин морфологиялык менен синтаксистик деңгээлинде тең изилденишине багытталган. Ошондой эле, кыргыз тилинин тактоочтордун өзгөчүлүктөү корпустук лингвистиканын алкагында дагы каралып чыкты. Бөлүмдүн башында тактоочтордун грамматикалык аспекттерине көңүл бурулуп, алардын түзүлүшүнө карай, уюшулуу жолдорун, лесика-семантикалык багыты боюнча бөлүнүшү жана синтаксисте тактоочтордун алган орду менен аткарган функциялары каралып чыкты. Тактоочтор кыргыз тил илиминде 1950-жылдырадын ортосунан гана өз алдынча сөз түркүм катары каралып баштаган. Тактоочторду өз алдынча сөз түркүмү катары кароо боюнча алгачкы кадамдары кыргыз тил илиминдеги

көрүнүктүү, кызыл тил илимпозчу Сарыбай Кудайбергенвго таандык. Ал эми кийинчирээк тактоочтор тууралуу бир өзүнчө китепти жазып, ар тараптуу изилдөөнү жүргүзүп, анын маңызын жетишээрлик түрдө ачып берген Самак Давлетовго таандык. Давлетов тактоочторду башка сөз түкүмдөрүнөн ажыратып чыгып, алрды башка сөз түркүмдөрүнөн айырмалап турган өзгөчөлүктөрүн бөлүп чыгып, кыргыз тилинде алар өз алдынча функциясын аткара аларын далилдеп берген (Абдувалиев, И. 2015). Кудайбергенов менен Давлетов тактоочко берген аныктамасына таянсак, алар кыймыл-аракеттин ар түрдүү ал абалын, сыпатын көрсөтүп, морфологиялык деңгээлде дээрлик өзгөрүлбөс мүнөзгө ээ болуп, учурлардын көбүндө этиштер менен айкашып, сүйлөмдө бышыктоочтук милдетти аткарган сөздөрдү тактооч деп беришкен (Давлетов С., 1980).

Түзүлүшүнө карай тактоочтор жөнөкөй жана татаал деп 2 топко бөлүнөт. Жөнөкөй тактоочтор өз алдынча дагы экиге бөлүнөт – тубаса жана туунду тактоочтор. Тубаса тактоочтор учурда турган абалында мындан нары дагы башка мүчөлөргө бөлүнбөгөн тактоочторду айтабыз. Мисалы, *“кеч”, “эрте”, “азыр”, “”чапчаң”* ж.б.у.с. Ал эми туунду тактоочтор болсо өз курамында сөз жасоочу мүчөнү камтып, же бир нече уңгусу бар болот. Мисалы, *“эртели-кеч”*, *“быйыл”*, *“бетме-бет”, “менчилеп”, “балача”* ж.б.у.с.

Ошондой эле, тактоочтордун морфологиялык жана синтаксистик жол аркылуу уюшулушу жана башка сөз түркүмдөрүнөн тактоочко айланган амалдары талкууланат. Бул маселеге экинчи бөлүмдүн орчундуу бөлүгү арналган, жана өзүнүн болгон бардык өзгөчөүлүктөрү тууралуу маалымат берилген. Өзгөчө бир нече сөз түркүмдөрдөн тактоочко кайсыл мүчөлөлдүн жалганышы аркылуу айланышын, ошону менен бирге кээ бир учурларла башка сөз түркүмдөрү менен окшош болуп кетери дагы белгиленет.

Лексика-семантикалык нуктан каралганда, кыргыз тилиндеги тактоочтор ири 2 топко бөлүндү, бышыктыгаыч тактоочтор жана аныктагыч тактоочтор.

Бышыктагыч такоочтор дээрлик этиштерди бышыктап, кыймыл-аракеттин болуп жатканын мезгилин, орунун, максатын жана себебин көрсөтүп турушат. Ал эми аныктагыч тактоочтордун тобуна киргендер кыйыл-аракеттин ишке кандай жолу менен ишке ашканын, анын сын-сыпатын жана сан-өлчөмүн билгизип турат.

Кийин, тактоочтордун синтаксистик деңгээлдеги аткарган функциясы, сүйлөмдө ээлеген орду, ичинде башка сөздөр менен айкалышуусу жана багыныңкы байланыштары терең изилденди. Сүйлөмдөрдө тактоочтор дээрлик көбүнчөдө этиштер менен айкалышып, кыймыл-аракеттин түрдүү ал абалын, сын сыпатын көрсөтүп, баяндоочту бышыктап тургандыгы аныкталды. Кээ бир учурларда сүйлөмдө башка функциясын аткарып калган учурлар да кезигет, сейрек болсо да. Синтаксисттик өңүттө кыргыз тилиндеги тактоочтор англис тилиндеги тактоочтордон салыштырмалуу жеңилирээк аныкталаары байкалды, себеби англис тилинде тактоочтук милдетти бир гана такооч эмес, тактоочтук сөз айкаштары, тактоочту тутумунда камтыган сүйлөмдүн жарымын түзгөн тизмектик дагы тактоочтук милдетти аткарып калышы мүмкүн. Бул өзгөчөлүктөрүн аныктоо максатында тактоочту камтыган сүйлөмдөрдү кыргыз тилинин корпусунан тандап алынып, 25 сүйлөмгө анализ жасалды, бирок практикалык бөлүмгө 15 тандалып алынып, аларга деталдуу түрдө морфологиялык жана синтакситик талдоо жүргүзүлгөн. Тактоочторго синтаксистик деңгээлде талдоо жүргүзүүдө Универсалдуу Багыныңкылыктын (UD-Universal Dependencies) долбоорунун алкагындагы иштелип чыккан UD Annotatrix платформасы колдонулду. Тактоочторго берилген энтектерди талдоодо кыргыз тилинин корпусу колдонулуп, тактоочтордун кошумча аспекттери боюнча сунуштоого энтектерди иштеп чыгууда Turkic Lexicon Apertium Toolkit инструментарии колдонулду. Тактоочтун грамматикалык аспекттери боюнча кошумча энтектерди иштеп чыгуу максатында тактоочтун лексика-семантикалык категорияларынын түрлөрү боюнча изилденип чыгып, кыргыз тилинин корпусунда кезиккен жыштыгы каралып чыгыв, лексика-

семантикалык бөлүнүшүнө карай таблицалар түзүлүп, алардын жыштык иаалыматтары көрсөтүлгөн.

Ал таблицалар дагы квантитативдик ыкманын жардамы менен талданып, кыргыз тилинин корпусунда бул категориядагы тактоочтор жалпы жонунан ээлеген көлөмү төмөнкүдөгү диаграммадан көрүүгө болот:

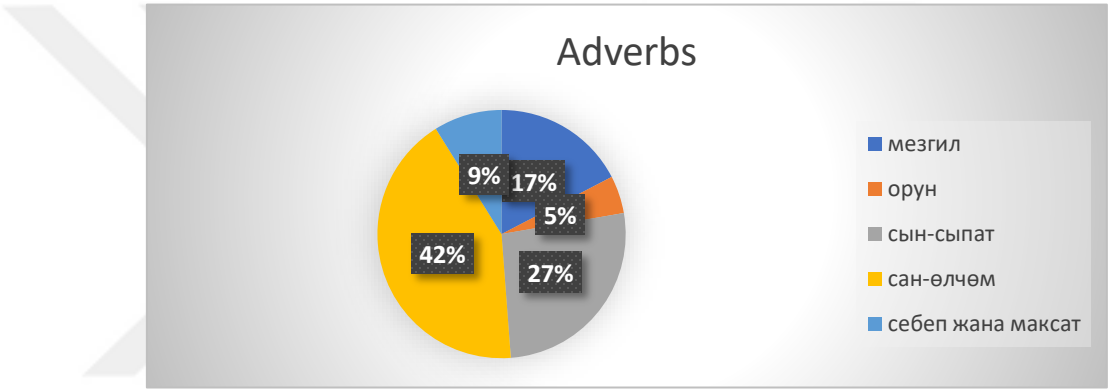
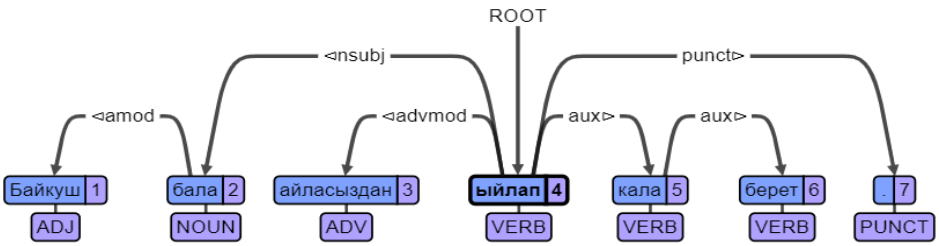


Diagram of the frequency of adverbs per one million words on average: мезгил – 99,6, орун – 28,066, себеп жана максат – 50,875, сын-сыпат – 151,96, сан-өлчөм – 242,46.

Бул статистикалык маалыматка таянуу менен, тактоочтордун грамматикалык аспекттери боюнча кошумча энтектерин кайсыларын алгач кошуу зарылчылыгын аныктоодо өбөлгө болуп бере алат.

Ал эми практикалык бөлүктө болсо тактоочтордун синтаксистик деңгээлде кандай байланышта болгону менен сүйлөм ичинде алган оруну жана аткарган кызматы каралып чыкты. Бул максатты аткаруу үчүн UD Annotatrix платформасы колдонулду. Бир эки мисал келтирели:

*Байкуш бала **айласыздан** ыйлап кала берет.*



"<Байкуш>"

"байкуш" ADJ 2 @amod #1->2

"<бала>"

"бала" NOUN 4 @nsubj #2->4

"<айласыздан>"

"айласыздан" ADV 4 @advmod #3->4

"<ыйлап>"

"ыйл" VERB @root #4->0

"<кала>"

"кал" VERB @aux #5->4

"<берет>"

"бер" VERB @aux #6->5

"<.>"

"." PUNCT @punct #7->4

Себеп бышыктагыч тактооч “***айласыздан***” туунду тактоочтордун тобуна киргендиги байкалып турат, анткени “***айла***” деген зат атоочтун уңгусуна терс формадагы сын атоочту жасай турган мүчө “***-сыз***” кошулуп, сын атоочту жаратты да, кийин чыгыш жөндөсүнүн мүчөсү “***-дан***” кошулуп, уңгусунда жөндөмөлөрдүн мүчөлөрү катып калуу жолу аркылуу пайдп болгон тактоочтордун катарына кирип, тактоочко айланды. Ал эми синтаксистик деңгээлде карай турган болсок, сүйлөмдөгү баяндоочту ролун аткарган татаал этиштин негизги этишин “***ыйлап***” бышыктап турат. Анын бышыктоо милдети “***advmod***” аркылуу көргөзүлсө, анын баяндооч менен байланышы “#3->4” символдору аркылуу чагыдырылган.

Чапчаң *келди боз жорго.*

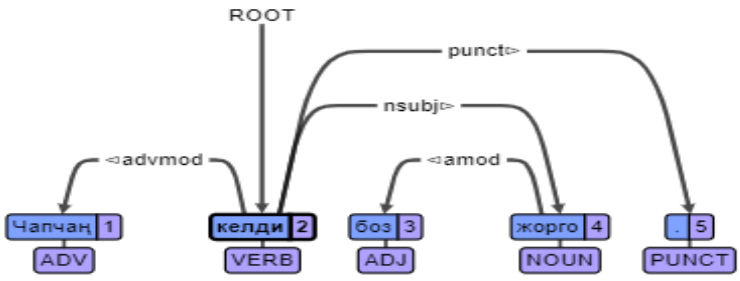


Figure 2.6.1.14. Dependency Parsing of attributive adverb of quality “чапчаң”

"<Чапчаң>"

"Чапчаң" ADV @advmod #1->2

"<келди>"

"кел" VERB @root #2->0

"<боз>"

"боз" ADJ @amod #3->4

"<жорго>"

"жорго" NOUN @nsubj #4->2

"<.>"

"." PUNCT @punct #5->2

Тубаса тактооч “***чапчаң***” аныктагыч тактоочтордун катарын кирип, кыймыл-аркеттин сапатын көрсөтүп турат, бул мисалда этиштин ылдамдыгын чагылдырып жатат. “***Чапчаң***” тактоочу сүйлөмдүн баяндоочун бышыктап тургандыктан, ага “***advmod***” энтеги берилип, этиш менен байланышы “#1->2” символдору аркылуу белгиленип турат.

Тактоочтордун көпчүлүк түрлөрү морфологиялык жана синтаксистик деңгээлде каралып чыккандан кийин, тактоочтордун кошумча морфологиялык категориялары үчүн мүмкүн болгон энтектерди иштеп чыгуу милдети түшүнүктүү болду. Кыргыз тилиндеги тактоочтор үчүн кошумча категориядагы энтектердин топтому жогоруда каралган бардык өзгөчөлүктөрдү эске алуу менен кылдат изилденген. Автоматтык түрдө иштеп чыгуу үчүн эң негизги жана жөнөкөй категория катары тактоочтордун курамы боюнча бөлүнүшү, андан кийин кыргыз тилиндеги тактоочтордун лексикалык-семантикалык классификациясы болуп каралган. Кыргыз тилиндеги тактоочтордун өзгөчөлүктөрүн эске алып, сөз түркүмүн белгилөөдөн кийинки экинчи деңгээлдеги категория тактоочторду курамы боюнча бөлүнүшү болмокчу. Андан кийин аныктагыч жана бышыктагыч тактоочторду камтыган лексикалык жана семантикалык классификация ирээттелмекчи.

Тактоочтордун түзүмү боюнча бөлүштүрүлүшү:

Tags for the 2nd level category of adverbs

Category name	Proposed tag
Simple adverbs	<simp>
Compound adverbs	<comp>

(N.B. – simple adverbs can be further classified as inflectional or derivational depending on its type, but that requires much more detailed analysis).

Тактоочтордун лексикалык жана семантикалык категориялары боюнча бөлүнүшү:

Tags for the 3rd level category of adverbs

Category name	Proposed tag
Adverbs of adverbial modifiers	<advmod>
Attributive adverbs	<attr_adv>

Бул эки чоң топ бир нече маанилүү ички бөлүмдөрдү камтыгандыктан, алар өзүнчө энтектерди талап кылат, мисалы:

Tags for the 3rd level category of adverbs

Category name	Proposed tag
3. For the adverbs of adverbial modifiers: f) Adverbs of time	<tm>
g) Adverbs of place	<plc>
h) Adverbs of reason	<rsn>
i) Adverbs of purpose	<pur>
4. For the attributive adverbs: b) Adverbs of quality	<qul>
j) Adverbs of number	<num>

Диссертациялык иште корпустук лингвистиканын алкагында тактоочтордун табияты жана алардын лексикалык-грамматикалык өзгөчөлүктөрү каралып чыккан. Корпустарды тил илимине жана практикалык колдонууга киргизилиши тил илимине жаңы багытты киргизди. Негизи, корпус лингвистикасы-бул тексттердин үлгүлөрүнүн чоң жыйнактарын талдоо аркылуу тилди жана анын бардык касиеттерин изилдөөгө багытталган ыкма. Жалпысынан алганда, корпуска машина окуй ала турган ири көлөмдөгү тилдин колдонулушу тууралуу маалыматты камтыган, бул маалыматтын арасында адабий жана адабий эмес тексттери тилдин синхроникалык дагы диахроникалык дагы аспекттерин чагылдыруу максатында киргизилген.

Кыргыз тилинин корпусу Кыргыз-Түрк Манас университетинин жана Заарланд университетинин синхрондук котормо бөлүмү менен биргеликте ишке ашырылган долбоордун негизинде иштелип чыккан. Долбоор 2018-жылы башталып, эки университеттин бөлүмдөрүнүн профессордук-окутуучулук курамынын тынымсыз аракетинин натыйжасында бүгүнкү күнгө чейин уланып келүүдө.

Натыйжада, бул диссертациянын материалдары кыргыз тилинин корпусунун илгери өнүгүүсүнө жана аны менен бирге Кыргызстанда корпустук лингвистиканын жаңы тармак катары жаралышына, анын өнүгүшүнө өз салымын кошот деп ишенебиз.