



**KYRGYZ-TURKISH MANAS UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
SIMULTANEOUS TRANSLATION DEPARTMENT**

**DEVELOPING OF A BANK OF SYNTACTIC
STRUCTURES OF THE KYRGYZ LANGUAGE WITHIN
THE FRAMES OF THE “UNIVERSAL DEPENDENCIES”
PROJECT**

SEIDE MUSAZHANOVA

MASTER’S THESIS

JUNE 2024

**KYRGYZ-TURKISH MANAS UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
SIMULTANEOUS TRANSLATION DEPARTMENT**

**DEVELOPING OF A BANK OF SYNTACTIC STRUCTURES OF
THE KYRGYZ LANGUAGE WITHIN THE FRAMES OF THE
UNIVERSAL DEPENDENCIES PROJECT**



MASTER'S THESIS

Seide MUSAZHANOVA

Thesis Advisor: Assoc. Prof. Aida KASİEVA

JUNE 2024

KABUL VE ONAY

..... danışmanlığında tarafından hazırlanan
“.....” adlı bu çalışma, jürimiz tarafından
Kırgızistan - Türkiye Manas Üniversitesi Lisansüstü Eğitim Enstitüsü Mütercim
Tercümanlık Anabilim Dalında Yüksek Lisans tezi olarak

OYBİRLİĞİ / OYÇOKLUĞU ile kabul edilmiştir.

...../...../.....

Jüri Unvanı,	Adı Soyadı	Kurumu	Onay	İmza
--------------	------------	--------	------	------

Başkan				
--------	--	--	--	--

Üye				
-----	--	--	--	--

(Danışman)				
------------	--	--	--	--

Üye				
-----	--	--	--	--

Üye				
-----	--	--	--	--

Üye				
-----	--	--	--	--

ONAY

Bu tezin kabulü Enstitü Yönetim Kurulunun tarih ve sayılı
kararı ile onaylanmıştır.

...../...../.....

Prof. Dr. İsmet ALTINTAŞ

Lisansüstü Eğitim Enstitüsü Müdürü

КАБЫЛ АЛУУ ЖАНА ЧЕЧИМ

.....жетекчилигинде.....

тарабынан даярдалган “.....”

аттуу бул диссертация комиссия тарабынан Кыргыз-Турк "Манас" университетинин Жогорку окуу жайдан кийинки билим берүү институту багытында Магистрдик диссертация катары БИР ДОБУШТАН / КӨПЧУЛУК ДОБУШ менен кабыл алынды.

...../...../.....

Комиссия	Наамы, аты – жөнү	Иштеген жери	Тастыктоо	Колу
Төрага				
Мүчө (Илимий жетекчи)				
Мүчө				
Мүчө				
Мүчө				

DECLARATION OF AUTHORSHIP

The research under the thesis paper entitled “*Developing of a Bank of Syntactic Structures of the Kyrgyz Language within the Frames of the Universal Dependencies project*” was undertaken according to the criteria for the Master’s degree at the Kyrgyz-Turkish Manas University. All sources cited in this thesis have been used ethically and in compliance with the academic norms. The data used in the thesis came from a variety of sources and were used in accordance with ethical and scientific standards. All references and citations for sources are included. Thus, I affirm that this thesis is entirely my work and that I am fully aware that any rules of violations would result in disciplinary action following University legislation.

Date: 17/05/2024

Seide Musazhanova

PREFACE

Language is the foundation of human communication, and understanding its syntactic structures is crucial for various natural language processing tasks. This dissertation aims to develop a comprehensive bank of syntactic structures for the Kyrgyz language, a language with an enriched history and unique linguistic features, as part of the "Universal Dependencies" project, focusing on understanding the syntactic structures of the Kyrgyz language.

The "Universal Dependencies" project offers a standardized framework for annotating language syntactic structures, allowing cross-linguistic comparison. This dissertation explores the Kyrgyz language's intricate syntactic patterns, contributing to the "Universal Dependencies" framework. The process involved meticulous data collection, annotation, and analysis, providing a deep understanding of the language's unique features and distinct typological characteristics.

This dissertation is significant because it progresses the field of computational linguistics and might have implications for text analysis, machine translation, and information retrieval, among other NLP applications. By making the language's syntactic structures more understandable, this effort aims to aid in the creation of more accurate and effective language processing systems for Kyrgyz.

I owe a debt of gratitude to Dr. Gulnura Zhumalieva and Assoc. Prof. Aida Kasieva, my research supervisor, for their unwavering support and direction during this research work. Without their solid encouragement, persistent trust in me, and maternal care, I could not have finished this task. Their knowledge, support, and helpful criticism have all been very helpful in crafting this dissertation. I am really grateful for the chance to study under such outstanding professors and academics. Their mentoring has enhanced my academic endeavors and given me a strong feeling of duty and commitment to our field.

I also extend my gratitude to Associate Professor of Linguistics at Swarthmore College Jonathan North Washington and Professor Elke Teich, the Head of English Linguistics and Translation Studies at the Department of Linguistics and Language Technology of Saarland University whose knowledge in Computational Linguistics have been fundamental in this study. We are deeply grateful for their patient mentorship, encouragement, expertise and thorough feedback that they gave.

I would like to also express my wholesome gratitude to the academic staff of the Translation Department of KTU Manas for their invaluable guidance and support throughout the journey of completing this dissertation paper. You have always been an inspiration to me with your persistent drive to quality and caring for the future generation of language professionals.

Throughout this process, I have been fortunate to have the needed support of my family. Their love, encouragement, and understanding have been the root of my success. I would like to express my deepest gratitude to my parents and siblings, whose infinite belief in me has been a constant source of inspiration, thank you for giving me the value of education and for always being my biggest support. Their belief in me has been my guiding light, and I dedicate this work to them.

I hope that this dissertation will serve as a vital source for further research in the field of computational linguistics and contribute to the broader goal of preserving and developing the Kyrgyz language.

Seide MUSAZHANOVA

TABLE OF CONTENTS

PREFACE.....	i
LIST OF ABBREVIATIONS	v
LIST OF TABLES	vi
LIST OF FIGURES	vii
ABSTRACT.....	1
ДИССЕРТАЦИЯНЫН МАЗМУНУ	3
GENİŞ ÖZET	22
АБСТРАКТ	37
INTRODUCTION	39
CHAPTER 1. THE GENERAL OUTLOOK OF THE THESIS PAPER AND INTRODUCTION TO UNIVERSAL DEPENDENCIES WEB PROJECT	41
1.1 Background and significance of the study	41
1.2 Research objectives and research questions	44
1.3 Overview of the structure of the thesis	47
Deduction to Chapter 1	50
CHAPTER 2. SYNTACTIC ANALYSIS AND DEPENDENCY PARSING OVERVIEW	51
2.1 Phrase Structure, Constituency, and Its Rules	51
2.2 Dependency Structure and Dependency Parsing Algorithm	56
2.3 Constituency-based and Dependency-based Treebanks	66
2.4 Applications of Treebanks in Computational Linguistics	75
Deduction on Chapter 2	79
CHAPTER 3. DEPENDENCY RELATIONS IN THE KYRGYZ LANGUAGE BASED ON UNIVERSAL DEPENDENCIES TAGSETS.....	81
3.1 Parts-of-Speech-tags and Tools Used in Universal Dependencies	81

3.2 Dependency parsing techniques and dependency relations in the Kyrgyz language	84
Deduction to Chapter 3	137
CONCLUSION	138
BIBLIOGRAPHY.....	140
CURRICULUM VITAE.....	144



LIST OF ABBREVIATIONS

UD	Universal Dependencies
POS	Parts-of-Speech
CL	Computational Linguistics (Corpus)
NLP	Natural Language Processing
NLU	Natural Language Understanding
TB	Treebank
NP	Noun Phrase
VP	Verb Phrase
SVO	Subject-Verb-Object
SOV	Subject-Object-Verb
GPSG	Generative Phrase Structure Grammar
HPSG	Head-Driven Phrase Structure Grammar
LFG	Lexical Functional Grammar

LIST OF TABLES

Table 3.1 A table representation of a simple sentence.....	82
Table 3.2 Core POS categories in UD	83



LIST OF FIGURES

Figure 3. 1. A picture of a dependency structure in a simple sentence.....	82
Figure 3. 2. 37 universal syntactic relations used in UD	83
Figure 3. 3. Example of acl relation in the sentence	85
Figure 3. 4. Example of advcl relation in the sentence.	86
Figure 3. 5. Example 2 of advcl relation in the sentence.	87
Figure 3. 6. Example of advmod in a sentence.	89
Figure 3. 7. Example 2 of advmod in a sentence.	91
Figure 3. 8. Example of amod relation in a sentence.	93
Figure 3. 9. Example of appos relation in a sentence.	95
Figure 3. 10. Example of aux relation in a sentence.	98
Figure 3. 11. Example of case marking in a sentence.....	100
Figure 3. 12. Example of cc relation in a sentence.	102
Figure 3. 13. Example of ccomp in the sentence.	105
Figure 3. 14. Example of compound from the UD website	107
Figure 3. 15. Example of compound in the Kyrgyz sentence.	107
Figure 3. 16. Example of conj relation in a sentence.	110
Figure 3. 17. Example of cop relation in a sentence.	112
Figure 3. 18. Example 2 of cop relation in a sentence.	114
Figure 3. 19. Example of csubj in the sentence.	116
Figure 3. 20. Example of csubj:pass in a sentence.	118
Figure 3. 21. Example of det relation in a sentence.	120
Figure 3. 22. Example of discourse relation in a sentence.....	121
Figure 3. 23. Example of fixed relation in a sentence.	122
Figure 3. 24. Example of flat expression in a sentence.	125

Figure 3. 25. Example of nmod relation in a sentence.....	128
Figure 3. 26. Example of nummod relation in a sentence.	130
Figure 3. 27. Example of parataxis relation in the sentence.	133
Figure 3. 28. Example of vocative relation in a sentence.	136



ABSTRACT

DEVELOPING OF A BANK OF SYNTACTIC STRUCTURES OF THE KYRGYZ LANGUAGE WITHIN THE FRAMES OF THE “UNIVERSAL DEPENDENCIES” PROJECT

Technology has brought about a significant shift in the relationship between software and language, making them interconnected. We process natural language daily, encountering millions of digitized texts. Nowadays, understanding Natural Language Processing (NLP) is crucial for linguists worldwide. More than ever, corpus research and modern linguistics are becoming complicated yet essential fields. One or more language corpora are used to research language at core layers of it: phonetics, morphology, syntax, semantics, and pragmatics. Additionally, language also crosses borders that are only linguistic in nature when it intersects with other fields like Theoretical/Applied linguistics, Computer linguistics, Sociolinguistics, Psycholinguistics, and others. The chance to learn the language at all levels by employing a variety of online tools and linguistic resources became available.

Web projects like Universal Dependencies (UD) have emerged, creating automated syntactic parsers for dependency analysis. UD is an open community project that aims to describe grammar in different human languages consistently, organizing them into standardized categories. It is an open community project where international linguists and software engineers create treebanks for world languages.

There are very few examples of syntactic treebanks of Turkic languages on the site, there are in Turkish, Tatar, and some sentences in Kazakh. Unfortunately, there are very few syntactic structures in the Kyrgyz language on the site, which indicates the level of involvement of the Kyrgyz language with world innovations. At KTMU's Faculty of Simultaneous Translation teachers and students themselves began creating and submitting examples of UD-based syntactic markings.

Unfortunately, at the moment the Kyrgyz language has just been included in the list of selected languages of this web project, and the Kyrgyz equivalents and usage guides of terms in this new field are not well known. There are still many annotation errors along with the lack of annotation guidelines. Therefore, the aim of this work is to introduce an expanded concept of Computer Linguistics in the Kyrgyz language and to demonstrate

morphological and syntactic explanations of words based on the newly created Kyrgyz Language Corpora using the international Universal Dependencies platform. We hope that different studies in this field will help prevent many problems that arise daily in linguistics.

In our thesis, we will talk about examples of syntactic analysis models of the Kyrgyz language, which we have prepared as a proposal for the above-mentioned web project. These annotation activities will truly enhance the contribution of the Kyrgyz language by analyzing different sentences. We hope to analyze the grammar of the Kyrgyz language at a new level using new resources and platforms. This study aims to introduce an expanded concept of UD in the Kyrgyz language in order to show morphological and syntactic dependency relations.

Quantitative, qualitative, and comparative methods are currently used in our thesis on account that the specialized web platforms for language corpora and speech extraction use quantitative, statistical, and comparative methods, which will help reveal the topic of our thesis in more detail. Various articles, theses, books, and summaries were used to write the first and second parts of the thesis. Relevant literature is provided in the last section of this document. Various websites and platforms were used to write the third practical part of this study.

We hope that different studies in this area will help to avoid many problems arising in linguistics. Therefore, one of the objectives of our work is to adapt the terms of the Universal Dependencies (UD) web project to the Kyrgyz language, analyze sentences in the Kyrgyz language through this project, and upload ready-made sentences to the site. The research's findings can be applied in teaching, studying, and conducting linguistic research in Corpus Linguistics, Semantics, Syntax, Morphology, and Translation Theory, benefiting both teachers and learners.

Keywords: Corpus Linguistics, Kyrgyz Corpus, syntactic analysis, morphological tags, Natural Language Processing (NLP), Universal Dependencies, treebank

ДИССЕРТАЦИЯНЫН МАЗМУНУ

“УНИВЕРСАЛДУУ БАГЫНЫҢКЫЛЫК” ДОЛБООРУНУН АЛКАГЫНДА КЫРГЫЗ ТИЛИНИН СИНТАКСИСТИК СТРУКТУРАЛАРЫНЫН БАНКАСЫН ТҮЗҮҮ

Азыркы санариптик доор, колдонуучуга интернет мейкиндигинде жайгашкан миллиондогон жеткиликтүү форматтагы тексттерди ар кыл максатта колдонууга мүмкүнчүлүк берет. Башкача айтканда, биз колдонгон смартфондогу, планшеттеги жана компьютердеги мындай тексттер кандайдыр бир деңгээлде компьютертик иштетүүдөн өткөрүлгөн тилдик материал болуп саналат. Ушул сыяктуу эле, ар түрдүү видео материалдарды (кино, подкаст, мультфильм, ж.б.у.с.) коштогон субтитрлер да эсептөө лингвистиканын жана **табигый тилди иштетүүнүн (кыскача ТТИ)** натыйжасы катары эсептелет.

Ошентип, интернет бизге эбегейсиз көлөмдөгү даяр иштелип чыккан тексттерди жана маалыматтарды колдонууга жол ачат. А бирок, бул жөн гана анын тышкы көрүнүшү деп айтсак жаңылышпайбыз. Анткени, бул тексттер колдонуучуга сунушталганга чейин көптөгөн жана кайра-кайра кайталанган машинаны үйрөтүү жана тестирлөө процесстеринен өткөрүлөт. Өзгөчө баштапкы этапта, машина тилди таанып иштетүү жана аны тилге үйрөтүү үчүн, тил адам тарабынан кол менен иштеп чыгуусу керек. Кээде, мындай иштетүү жарым-автоматтык ыкма менен аткарылышы мүмкүн. Ал эми тил машина тарабынан жакшы деңгээлде өздөштүрүлгөн учурда (мисалы, англис, немис ж.б. тилдеринде), анда тексттер толугу менен автоматтык жол аркылуу иштетилет. Тилди иштетүүнүн мындай ыкмасы бул процесстин убактысын кыйла кыскартат жана аны жеңилдетет. Тилдин мындай автоматтык иштетүүсүн камсыздоо максатында, бул жааттар менен алектенген дүйнөлүк окумуштуулар тилдин ар бир деңгээлине жараша (мис., фонетика, морфология, синтаксис, семантика, прагматика) же болбосо анын бардык деңгээлдерин камтыган атайын веб-долбоорлорду иштеп чыгышкан. Мисалы, синтаксис багытында, сүйлөмдөрдүн автоматтык түрдө талданышын камсыздаган атайы программалар, веб-платформалар, ар кандай функцияларды аткарган колдонмолор жана инструменттер иштелип чыгууда.

Ал эми мындай тексттерди камсыздаган булактар тууралуу айта турган болсок, алар көбүнөсө атайын бул максатта колдонула турган программалар аркылуу

интернет мейкиндигинде орун алган тексттер (мис., жаңылыктар, кабарлар, репортаждар, блогдор ж.б.у.с. булактар). Булардан сырткары, мындай типтеги тексттерге илимий изилдөөлөр, ар кыл жанрдагы китептер, журналдар жана башка басылма продукциялары да карайт. Мындай тексттик маалыматтардын убакыт өткөн сайын көбөйүп жаткан көлөмү адамдарга сан жеткис мүмкүнчүлүктөрдү сунуштаганы менен, бул көрүнүш табигый тилди иштетүү (ТТИ) тармагында бир катар көйгөйлөрдү да жаратууда. Мисалы, белгилүү бир тилдин синтаксистик табиятын компьютердик иштетүү алкагында карай турган болсок, ал тилдеги өзгөчөлүктөрдү да эске алып иштете турган программалар түзүлүшү абзел.

Бул өңүттө, синтаксистик жактан иштелип чыккан тексттер, тагырак айтканда, сүйлөмдөр жана синтаксистик структуралар, атайы коддолгон энбелгилер (тегдер) менен белгиленген учурда, алардын компьютердик иштетүүсү ТТИ колдонмолору үчүн өзгөчө натыйжалуу болот. Тексттин мындай нукта талданышы негизги контекстти жана тилдик маалыматты ТТИ алгоритмдери (мисалы, морфологиялык энтектөө жана синтаксистик талдоо) аркылуу анын мындан аркы иштетүүсүнө өбөлгө түзүп, ал тексттеги керектүү жана/же маанилүү маалыматты тереңдетип талданышын, аны чечмелөөнү жана жалпы тексттен чыгарып алууну (сегментациялоону) жеңилдетет. Натыйжада, колдонуучу ар кандай тилдик операцияларды жакшы сапатта аткара ала турган түрдүү колдонмолорду пайдалана алат.

Маселен, учурда биздин түзмөктөр (гаджеттер) адамдын тилин таанып, алар менен толук баарлаша алат. Булардын катарына учурда өзгөчө популярдуу болуп келе жаткан **жасалма интеллект** (ЖИ) да кирет. Аны менен баарлашуу же башка максатта иш алып баруу үчүн бир гана зарыл нерсе - бул интернетке туташуу болуп саналат. Онлайн режимде жеткиликтүү болгон мындай курал (ЖИ), чындыгында, жогоруда баса белгиленгендей, жалаң гана аннотацияланган (тегделген же коддолгон) тексттер аркылуу мүмкүн болот. Мисалы, учурда биз тексттик маалыматтарды дүйнөнүн ар кайсы тилинде (англисче, кытайча, орусча, корейче жана башка) жакшы сапатта чагылдырган программалар менен кездеше алабыз. А бирок, тилекке каршы, кыргыз тилине келгенде биздин түзмөктөр мындай операцияларды толук кандуу же канаатандырарлык деңгээлде аткара

албай тургандыгына күбөбүз. Анткени, кыргыз тили толугу менен компьютердик иштетүүдөн өтө элек. Ошондуктан машина менен гаджеттер аны жакшы сапатта тааный албайт. Анын натыйжасында, жогорудагыдай программаларды же түзмөктөрдү кыргыз тили үчүн колдонууда бир катар одоно жана катачыл тексттер ортого чыгууда. Мындай көрүнүш, албетте, адамды өкүндүрбөй койбойт.

Бул маселени чечүү үчүн, тагырак айтканда, кыргыз тилин компьютерлештирүү жана аны иштетүү максатында 2019-жылы алгачкы 1 (бир) миллион сөздөн турган кыргыз тилинин корпусу түзүлө баштаган. Анткени тилдик корпустун материалы сөзсүз түрдө мындай иштетүүдөн өткөрүлөт жана бара-бара жасалма интеллекттин сапатын жакшыртууда пайдаланат. Ал, өз кезегинде, кыргыз тилинде сүйлөгөн “Сири”, “Алиса” ж.б.у.с. үн ассистентин түзүлүшүнө эбегейсиз чоң жардам көрсөтө алат. Ошондуктан, табигый тилди иштетүү жана анын процесстери канчалык мааниге ээ экендигине ынанып күбө болуудабыз.

Табигый тилди иштетүү чөйрөсү (ТТИ) 1900-жылдардын башында башталып, экинчи дүйнөлүк согуштан кийин олуттуу жетишкендиктерге жеткен узак жана бай тарыхка ээ. Бул тармак өзүнүн өнүгүү жолунда ар кандай олуттуу бурулуштарга жана өзгөрүүлөргө дуушар болгон.

Табигый тилди иштетүүнүн негизин түзгөн идеялар биринчи жолу 1900-жылдардын башында пайда болгон. Бул тармак швейцариялык лингвист Фердинанд де Соссюрдун эмгектеринин натыйжасында өнүккөн. Ал, тилдерди система катары мүнөздөгөн методду иштеп чыгып, биринчилерден болуп тилди илим катары кабылдап, анын үстүндө иштей баштаган.

Экинчи дүйнөлүк согуштан кийин орун алган эл аралык иш-чаралар, тилдерди которуунун баалуулугун түшүнүү концепциясына алып келген. Анын натыйжасындагы автоматтык түрдө которо турган машиналарды иштеп чыгуу аракеттери табигый тилди иштетүүнүн эң биринчи кадамдары катары таанылган. Ошентип, ал иш-аракеттер табигый тилди иштетүү чөйрөсүнө расмий түрдө жол ачкан. Экинчи дүйнөлүк согуш тилдин канчалык деңгээлде маанилүү болгондугун далилдеген көрүнүшкө айланган. Натыйжада, бул согуш табигый тилди иштетүүнүн келип чыгышына өбөлгө түзүп, ага чоң түрткү катары бааланган. Андан бери табигый тилди иштетүү тармагы тилдерди которууда, аларды

түшүнүүдө, ал тилдерде тексттерди жаратууда канчалык маанилүү экендигин практика жүзүндө тастыктап келүүдө.

Анын кээ бир маанилүү этаптарына кайрыла турган болсок, 1960-жылдары ТТИ тармагында маанилүү жетишкендиктер байкалган. Мисалы, 1964-1966-жылдары MITде (Массачусетс технологиялык институту) ELIZA аталышындагы чатбот программасы иштелип чыккан. ELIZA адам менен машинанын ортосундагы баарлашууну түшүнүүнүн алгачкы үлгүсү болуп, мындай машина жана адам ортосундагы баайланышты өнүктүргөн программа катары таанылган. Бул чатбот англис тилинде адамдын тилин түшүнүп, суроолоруна жооп бере ала турган мүмкүнчүлүккө ээ болгон. ELIZAdан кийин, 1970-2000-жылдары көптөгөн кызыктуу жана ар тараптуу иш-аракеттерди аткара алган чатботтор гана эмес, роботтор да жаратыла баштады.

1980-жылдардын аягы жана 1990-жылдардын башында статистикалык моделдер киргизилип, колго жазылган эрежелердин татаал топтомдоруна таянган табигый тилди иштетүү системаларынын көбү эскирип, бул тармакта жаңы революция башталгандыгы байкалган. Машиналык окутуу ыкмаларына өтүү жана эсептөөчү кубаттуулуктун өсүшү бул прогресстин негизги кыймылдаткычтарынын милдетин аткарган. Адам сыяктуу тексттерди жазуу, аларды түшүнүү жана алардын негизинде алгоритмдер аркылуу тилдик моделдерди түзүү, табигый тилди иштетүүнү таптакыр жаңы деңгээлге алып келгендиктен улам, жаңыча ыкмаларды колдонуу, өзгөчө LLM (Large Language Models) ‘чоң көлөмдөгү тилдик моделдер’, ТТИнү акыркы жылдарда эбегейсиз ийгиликтерге жеткирди.

Азыркы учурда табигый тилди иштетүү технологиясы адамдын жашоосунун көптөгөн аспектилерине таасирин тийгизгендиктен, ТТИ заманбап дүйнөнүн маанилүү тармагына айланды. Табигый тилди иштетүү ыкмаларынын тез өнүгүшү менен жазуу, кодго айландыруу жана бул өнүгтө ой жүгүртүү сыяктуу сапаттык иш-аракеттерде компьютер адамдар менен атаандаша ала турган абалга келди десек жаңылышпайбыз. ТТИ текстти алып же бөлүп чыгарууну, аны сегментациялоону жана автоматташтыруу жолдорун алда канча жеңилдетет. Мисалы, анын ыкмалары интернеттеги же башка онлайн платформаларга киргизилген издөө системаларын оптималдаштыруу багытында ачкыч сөздөр

аркылуу бул операцияларды аткаруу сыяктуу иш-аракеттер үчүн өтө пайдалуу жана натыйжалуу деп бааланат.

Мындан тышкары, азыркы табигый тилди иштетүү мүмкүнчүлүктөрү компьютерге текстти окуу, үндү таануу, эмоцияларды түшүнүү (сентимент анализ) жана баарлашуунун негизги компоненттерин аныктоого мүмкүндүк берет. Ошондой эле, ал компьютер менен адамдын ортосундагы баарлашууну эне тилинде жүргүзүүгө шарт түзөт. Мунун ачык далили катары, жогоруда белгиленгендей, “Сири”, “Алиса”, “Маруся” сыяктуу үн колдонмолорун атай алабыз.

Ошентип, табигый тилди иштетүү белгилүү бир тилде кездешкен түшүнүксүздүктөрдөн арылуу, аларды тактап иликтөө аркылуу маалыматтардын структурасын аныктай алат. Ошондуктан, ТТИ юридикалык, медициналык жана башка көптөгөн тармактар тарабынан көп колдонулат. Анткени ТТИнин мүмкүнчүлүктөрү көптөгөн тапшырмаларды аткарууда маалыматтарды башкаруу, аларды автоматташтыруу, байланышты жакшыртуу жана технологияны колдонуучуга жакындатууда абдан чоң роль ойнойт. Мындан улам, табигый тилди иштетүү чатботтор жана виртуалдык жардамчылар үчүн адамдын кебин түшүнүү жана тилге байланыштуу иш-аракеттерди жүргүзүү үчүн да өтө маанилүү болуп саналат. Бул алкакта, өзгөчө эсептөө лингвистика жаатындагы ТТИнин жетишкендиктери эбегейсиз кадамдарды таштап, күн сайын өркүндөп өсүп келүүдө.

Эсептөө лингвистикасы тил илими менен информатиканы айкалыштырган тармак катары белгилүү. Ал когнитивдик илим жана информатика тармагында орун алган академиялык жетишкендиктери менен баалуу болуп саналат. Анткени бул тармак табигый тилди иштетүүдө чоң жардам көрсөтөт жана ошол эле учурда, жасалма интеллекттин бир бөлүгү болуп эсептелет. Өзгөчө, ал машиналарга адамдын тилин үйрөтүүгө, аны түшүнүүгө жана коюлган сурамага же суроо-талапка жооп берүүгө мүмкүндүк берет. Ал, ошондой эле, чоң көлөмдөгү текстти иштетүү үчүн алгоритмдерди иштеп чыгууга мүмкүнчүлүк берип, аларды автоматтык түрдө талдоодо өтө маанилүү болуп саналат. Ошентип, ТТИ тилдик технологияларды өнүктүрүүгө эбегейсиз чоң салым кошот. Мунун ачык-айкын мисалы катары, тил үйрөнүү колдонмолору жана автоматташтырылган котормо

куралдары глобалдык коммуникациянын жеткиликтүүлүгүн жогорулатканын айта алабыз.

Мындай оң өнүктө орун алып жаткан көрүнүштөр тилдин структурасы жана когнитивдик процесстер алкагында байкалган жаңы ачылыштарга алып келүүдө. Анын натыйжалары искусство, саламаттыкты сактоо, билим берүү, көңүл ачуу жана көптөгөн башка тармактарда түрдүү мүмкүнчүлүктөрдү сунуштайт. Мындан улам, ар кандай деңгээлдеги уюмдарда тилди компьютердик иштетүү алкагында иш алып бара ала турган адистерге болгон талап барган сайын өсүп келүүдө.

Мисалы, ошондой эле, тилди толук аннотациялоо максатында, мындай аракеттерди так аткаруу үчүн эл аралык стандарттарга ылайык түзүлгөн онлайн форматтагы жана автоматтык түрдө сүйлөмдөрдү талдоочу атайын инструменттер жана платформалар түзүлөт. Акыркы жылдары мындай веб талдоочулар жана онлайн анализдөө куралдары тил илиминин ажырагыс бөлүгүнө айланды. Күн сайын миллиондогон “чийки” жана иштелип чыга элек метамаалыматтар тийиштүү адистер тарабынан талданып келет. Натыйжада, заманбап тилдер табигый тилди иштетүүнүн жана кепти автоматтык түрдө таануунун жаңы деңгээлине көтөрүлгөндүгү байкалууда.

Кыргыз тилинин синтаксистик жактан аннотацияланышына киришүү салыштырмалуу жакында эле (2022-ж.), дүйнөдөгү бардык тилдердин бирдиктүү стандарттуу аннотациясына багытталган **Universal Dependencies (Универсалдуу багыныңкылык)** эл аралык платформасынын жана веб-долбоорунун негизинде башталды.

Universal Dependencies (кыскача UD) долбоору дүйнөнүн бир нече тилдериндеги грамматиканы, анын түзүмүнө киргизилген сүйлөм бөлүктөрүнүн, морфологиялык мүнөздөмөлөрү жана синтаксистик багыныңкылыктарды аныктап ырастоо жана UDнин талаптарына ылайык ырааттуу нукта сүйлөмдөрдү аннотациялоо үчүн колдонулат. Бул ачык коомдук долбоорго 500дөн ашык адам салым кошуп, дүйнөнүн 100дөн ашык тилинде дарак түрүндөгү структуралар түзүлүүдө. Долбоордун негизги максаты - тилдер арасындагы аннотацияны унификациялоо жана алардын бирдей көз караштан аннотацияланышын камсыздоо болуп саналат, Бирок, ошол эле учурда, ал белгилүү бир тилге тиешелүү өзгөчөлүктөрдү сактоо сыяктуу максатын көздөйт. Udде ар кандай

тилдер арасында орун алган структуралык параллелизмди камсыз кылуу үчүн тилдер аралык ырааттуу багыныңкылык анализин түзүүгө өзгөчө көңүл бурулат.

"Универсалдуу багыныңкылык" долбоорунун алкагында кыргыз тилинин синтаксистик структураларын өнүктүрүү - кыргыз тилинин грамматикасы жөнүндөгү түшүнүгүбүздү тереңдетүүгө багытталган маанилүү жана кызыктуу иш болуп саналат. Бул долбоор лингвисттер, изилдөөчүлөр жана эсептөө лингвистика жаатындагы адистер үчүн чоң перспективаларды сунуштайт.

Кыргыз тили, Кыргызстанда гана эмес, Борбордук Азиянын ар кайсы аймактарында жана бир канча чет өлкөнүн кээ бир аймактарында да кеңири эне тил катары колдонулган түрк тили болуп саналат. Биздин тил уникалдуу агглютинативдүү грамматикага жана табияты боюнча татаал түзүлүшкө жана өзгөчөлүктөргө ээ болгон тил катары таанымал. Кыргыз тили үчүн синтаксистик деңгээлде анын структураларынын банкын иштеп чыгуу менен изилдөөчүлөр тилдин лингвистикалык касиеттерин тереңирээк түшүнүп, табигый тилди иштетүү жана эсептөө лингвистикасынын кеңири чөйрөсүнө салым кошо алышат.

Дарак түрүндөгү структура (treebank) – тийиштүү схемаларды колдонуу жана сүйлөмдөгү же сөз айкаштардагы алардын байланыштарын көрсөтүү аркылуу сүйлөмдөрдү талдоонун өзгөчө ыкмасы болуп саналат. "Treebank" термини лингвист Джеффри Линч тарабынан 1980-жылдары киргизилген. Дарак түрүндөгү структуралар көбүнчө сөз түркүмдөрүнүн тегдери менен белгиленген корпустун негизинде түзүлөт. Бул структурада сүйлөмдүн сөз түркүмдөрү (атооч, этиш, сын атооч ж.б.) жана алардын ортосундагы синтаксистик байланыштар (ээ, баяндооч, аныктооч ж.б.) атайын коддор, б.а. тегдер менен белгиленип, алардын багыныңкылык байланышы жебенин багыты менен көрсөтүлөт.

Багыныңкылыкты талдоо - бул синтаксистик анализдин жогорудан ылдыйга (top-down) карай ыкмасы болуп саналат. Талдоо алгоритми сүйлөмдөгү баш сөздү (баяндоочту) аныктап, сөздөр ортосундагы багыныңкылыкты чагылдырган дарак структурасын түзөт. Бул ыкманын негизги артыкчылыгы – өзгөчө татаал сүйлөмдүн структурасын талдоодо изилдөөчүнүн ишин кыйла жеңилдете ала тургандыгы болуп саналат.

Багыныңкылык байланышты талдоо табигый тилди иштетүү үчүн бир нече себептерден улам маанилүү. Биринчиден, бул ыкма сөз айкаштагы сөздөрдүн

ортосундагы грамматикалык байланыштарды бөлүп көрсөтүү менен синтаксистик байланыш түшүнүгүн жогорулатат жана анын татаал түзүлүшүн ачыктап көрсөтүп бере алат. Экинчиден, маалымат алуу, суроолорго жооп берүү жана аларды талдоо сыяктуу тапшырмаларды аткаруу үчүн семантикалык шилтемелерди изилдейт. Ошондой эле, ал структураланбаган (мурда иштетилбеген же белгиленбеген) тексттик материалды структураланган формага айландырууну жеңилдетип, маалыматты алуу жана талдоо процедураларын тездетет.

Багыныңкылыкты талдоо машиналык котормо системалары үчүн да өтө пайдалуу, анткени ал ар кайсы тилдердеги сөздөрдү жана сүйлөмдөрдү теңдештирип, синтаксистик жана семантикалык деталдарды сактоо менен котормонун тактыгын жогорулатууга жардам берет. Мындан тышкары, бул ыкма тексттеги элементтерди жана алардын байланыштарын салыштырып же тереңдетип изилдөөгө жардам берет.

“Универсалдуу багыныңкылык” веб-долбоорунун максаты ар кандай тилдерде синтаксистик структураларды бирдиктүү чагылдыруу, салыштыруу жана талдоону жеңилдетүү болуп саналат. Кыргыз тили үчүн синтаксистик структуралардын банкын түзүү аркылуу изилдөөчүлөр кыргыз синтаксисинин уникалдуу өзгөчөлүктөрүн жана үлгүлөрүн жакшыраак түшүнүп, аларды андан ары изилдөө жана талдоодогу маселелерди аныктай алышат. Акыры, бул изилдөө жалпы багыныңкылык долбоорунун алкагында кыргыз тили үчүн ар тараптуу жана так лингвистикалык ресурсту түзүү максатын көздөйт.

Кыргыз тилинин синтаксистик структураларын долбоор тарабынан белгиленген универсалдуу стандарттарга шайкеш келтирүү менен изилдөөчүлөр салыштырмалуу лингвистикалык изилдөөлөрдү жүргүзүп, машиналык котормо системаларын жакшыртып, тилди иштетүү алгоритмдерин иштеп чыга алышат. Диссертацияда бул аракеттердин лингвистикалык мааниси, анын кыйынчылыктары жана мүмкүнчүлүктөрү, ошондой эле кыргыз тилин изилдөө жана түшүнүү үчүн мүмкүн болгон маселелер кеңири талкууланат. Акыры, бул долбоор кыргыз тилин билүү деңгээлин жогорулатууга, лингвистикалык изилдөөлөрдүн сапатын жакшыртууга жана тилди иштетүүчү татаал

шаймандарды жана технологияларды иштеп чыгууга салым кошууга өбөлгө түзүлөт.

Бул максаттарга жетүү үчүн бир катар милдеттердин аткарылышы талап кылынат:

- Табигый тилди иштетүү жана дүйнөлүк тилчилердин багыныңкылык грамматикасы (dependency grammar) менен таанышуу;
- Кыргыз тилинин грамматикасын аннотациялоо;
- Кыргыз тилинин синтаксистик байланыштарын «Универсалдуу багыныңкылык» долбоору алкагында изилдөө;
- Кыргыз тилинин синтаксистик структураларынын банкын иштеп чыгуу (дарак түрүндөгү мисалдарды берүү);
- Кыргыз тилиндеги синтаксистик структуралардын толук колдонмосун түзүү;
- Кыргыз тилинин синтаксистик өзгөчөлүктөрүн “Универсалдуу багыныңкылык” алкагына ылайык талдоо жана категориялаштыруу;
- Тилдик талдоо үчүн кыргыз тилиндеги уникалдуу синтаксистик үлгүлөрдү жана конструкцияларды аныктоо;
- Синтаксистик касиеттерди изилдөө жана талдоо максатында кыргыз тилин изилдегендер, студенттер жана мугалимдер үчүн баалуу лингвистикалык ресурсту камсыздоо;
- Тилдердин көлөмүн кеңейтүү жана тилдер аралык салыштыруу жана талдоо үчүн билим базасын байытуу;
- Структураланган синтаксистик маалымат топтомун камсыз кылуу менен кыргыз тилинде табигый тилди иштетүү үчүн эсептөө куралдарын, терминдерин жана тиркемелерин иштеп чыгууну жеңилдетүү;
- Долбоорго даярдалган сүйлөмдөрдү расмий түрдө сунуштоо.

Булардын арасында табигый тилди иштетүү жана дүйнөлүк тилчилердин багыныңкылык грамматикасы менен таанышуу милдеттери өтө маанилүү болуп саналат. Анткени, кыргыз тилинин грамматикасын аннотациялоо, кыргыз тилинин синтаксистик байланыштарын "Универсалдуу багыныңкылык" долбоору

алкагында изилдөө, кыргыз тилинин синтаксистик структураларынын банкын иштеп чыгуу, жана башка көптөгөн милдеттер кыргыз тилинин грамматикасын жана башка аспектилерин жакшы билүү маселесин ортого коёт.

Буларга кошумча катары, табигый тилди иштетүү жаатында күн сайын жаңы терминдер жана түшүнүктөр пайда болууда. Бул терминдер жана эсептөө лингвистикасындагы жаңы түшүнүктөр тилди иштетүүдө чоң мааниге ээ. Анткени аларды билүү, которуу жана туура колдонуу, өзгөчө кыргыз тилинде, көптөгөн маселелерди, талаштарды жана талкууларды жаратууда. Бул өтүттө, кыргыз тилин аннотациялоо боюнча иш-чаралары, корпустарды түзүү жана бул долбоорго кошулуусу жаңыдан гана башталып, табигый тилди иштетүүнүн алгачкы кадамдары эми гана жасалууда. Булардан улам, дарак түрүндөгү структураларды берүү аркылуу биз кыргыз тилинин терминология жаатына да салым кошо алабыз. Бул терминдерди колдонуу аркылуу изилдөөчүлөрдүн, студенттердин жана мугалимдердин кыргыз тилинин табигый тилди иштетүү тармагына кызыгып, бул долбоорго киришүүсүнө өбөлгө түзө алат деп ишенебиз.

"Универсалдуу багыныңкылык" долбоору жана кыргыз тилинин синтаксистик түзүлүшү бул изилдөөнүн негизги предметтери болуп саналат. Изилдөө кыргыз тилинин грамматикасын иликтеп, анын синтаксистик талданышына багытталат. "Универсалдуу багыныңкылык" долбоорунун алкагында кыргыз тили үчүн синтаксистик структуралардын банкын иштеп чыгуу аркылуу, изилдөөчүлөр кыргыз тилдин синтаксисин чагылдырып, анын талданышын жакшырта алышат. Ошондой эле, табигый тилди иштетүү жана эсептөө лингвистикасы сыяктуу тармактарда жүргүзүлө турган изилдөөлөргө жол ачып, аларды өнүктүрүүгө өбөлгө түзүлөт.

Учурда кыргыз тили толугу менен иштелип чыга элек тилдердин катарында болгондуктан, бул изилдөөнүн негизги максаты - кыргыз тилинин синтаксистик структураларынын ар тараптуу жана лингвистикалык жактан так аннотациясын камсыздоо болуп саналат. Бул диссертациянын максатына ылайык, изилдөө тилдеги ар кандай синтаксистик конструкцияларды, багыныңкылыктарды жана байланыштарды аныктайт, категориялаштырат жана аларды "Универсалдуу багыныңкылык" веб-долбоорунун көрсөтмөлөрүнө ылайык стандартташтырылган форматка келтирет.

Изилдөөнүн илимий жаңылыгы катары кыргыз тилинин грамматикасын дүйнөлүк долбоордун алкагында талдоо жана ар кандай тилдердеги синтаксистик структураларды оңой жана түшүнүктүү ыкмалар аркылуу салыштырып талдоого мүмкүндүк берүүсүндө жатат. Бул көрсөтмөлөрдү кыргыз тилине колдонуу менен, изилдөө ресурстары жетишсиз тилдер үчүн синтаксистик аннотация жана багыныңкылыкты талдоо боюнча билимдин жогорулашына өбөлгө түзөт. Бул, өзгөчө, “жетишсиз ресурстуу” тил катары аныкталган кыргыз тили үчүн өтө маанилүү. Анткени, ал уникалдуу синтаксистик касиеттери бар түрк тили болсо да, учурда анын ресурстары чектелүү болгондугун моюнга алышыбыз керек.

Мындан тышкары, кыргыз тили үчүн синтаксистик структураларды иштеп чыгуу табигый тилди иштетүүдө, машиналык котормодо жана башка тилдик технология менен байланыштуу маселелерде практикалык маанилүүлүккө ээ болушу мүмкүн. Кыргыз тили үчүн жогорку сапаттагы синтаксистик жактан аннотацияланган ресурсунун болушу – тилдин өзгөчөлүктөрүнө ылайыкташтырылган жаңы тил технологиялык куралдарды иштеп чыгууга өбөлгө түзөт.

Магистрдик диссертация **үч негизги бөлүмдөн** турат. Ар бир бөлүм тил илиминдеги, табигый тилди иштетүү терминологиясындагы жана дарак түрүндөгү структураларды түзүүдөгү маанилүү аспектилерди иликтейт.

1-бөлүм изилдөөнүн өбөлгөлөрүн, маанисин, максаттарын жана негизги суроолорун чагылдырган киришүүнү камтыйт. Түрк тилдер үй-бүлөсүнө кирген кыргыз тилинде 7 (жети) миллионго жакын адам сүйлөйт. Бирок, көптөгөн адамдар үчүн баарлашуу каражаты катары маанилүү болгонуна карабастан, кыргыз тилинин синтаксистик түзүлүштөрүн изилдөө үчүн жеткиликтүү комплекстүү ресурстар жетишсиз. Бул кыргыз тилинин грамматикалык эрежелерин жана үлгүлөрүн түшүнүүгө кызыккан изилдөөчүлөр, педагогдор жана тил сүйүүчүлөр үчүн олуттуу кыйынчылыктарды жана кыргыз тилин үйрөнүүдө тоскоолдуктарды жаратат.

"Универсалдуу багыныңкылык" веб-долбоору алкагында жазылган бул эмгек, ушул кыйынчылыктар менен байланыштуу кээ бир маселелерди чечүүгө багытталат. Бул долбоордун алкагы лингвисттерди жана компьютердик илимпоздорду биргеликте иш алып баруусун талап кылат. Мында, изилдөөчүлөр, педагогдор жана мындай ресурстарды иштеп чыгуучулар тарабынан ар түрдүү

тилдердин синтаксистик структураларын изилдөө жана талдоо үчүн колдонула турган эркин жана ачык булактуу ресурсту түзүү максаты көздөлөт.

2-бөлүмдө, синтаксистик багыныңкылык жаатына тийиштүү адабияттары каралып, алардын негизги теориялары жана жыйынтыктары кеңири талкууга алынат. Багыныңкылыкты талдоо – синтаксистик анализдин бир түрү, ал сүйлөмдөгү сөздөрдүн ортосундагы мамилелерди иерархиялык байланыштар катары эмес, багыныңкылык катары моделдөөгө багытталган. Синтаксистик анализдин баштоочуларынын бири 1950-жылдары фразеологизмдердин грамматикасы деген түшүнүктү киргизген Ноам Хомский болуп саналат. Хомскийдин эмгеги негизги синтаксистик категориялардын жыйындысынан сүйлөмдөрдү түзүүнүн эрежелерин жана бул категорияларын айкалыштыруу эрежелерин аныктоо менен азыркы синтаксистик анализдин негизин түзгөн. Синтаксистик анализ тармагындагы дагы бир таасирдүү илимпоз, Мари-Кэтрин де Марнефф, багыныңкылыкты талдоо алгоритмдерин өнүктүрүүгө чоң салым кошкон. Марнеффтин эмгектери сүйлөмдөгү сөздөрдүн ортосундагы байланыштарды иерархиялык байланыштар катары эмес, багыныңкылык катары кароого багытталган. Бул, синтаксистик анализди эффективдүү жана так жүргүзүүгө мүмкүндүк берет. АКШда бир нече изилдөөчүлөр синтаксистик жана багыныңкылыкты талдоо тармагына салым кошкон. Алардын арасына табигый тилдеги текстти талдоо үчүн татаал алгоритмдерди иштеп чыккан жана талдоо моделдерин окутуу жана баалоо үчүн дарак структураларын өнүктүргөн Дэниел Журафский, Дэн Клейн жана Кристофер Мэннинг сыяктуу көрүнүктүү окумуштуулар кирет. Алар грамматикада сүйлөмдөрдүн иерархиялык түзүлүшүн сөз айкаштары аркылуу сүрөттөгөн. Фразеологиялык грамматикада сүйлөмдөр иерархиялык түзүлүшкө уюшулган составдык бөлүктөрү боюнча талданат. Сөз айкаштардын түзүлүшүнүн грамматикасынын эрежелери сөз айкаштарды жана алардын кандай жол менен байланышып, алардын сүйлөмдөрдү түзүү жолун аныктайт. Багыныңкылыктын структурасы - багыныңкылык катары сүйлөмдөгү сөздөрдүн ортосундагы мамилелерди моделдөөгө багытталган синтаксистик анализдин дагы бир ыкмасы болуп саналат. Багыныңкылыктын түзүлүшүндө сүйлөмдөгү ар бир сөз анын баш сөзүнө (мис., баяндоочко) байланышып, анын синтаксистик ролун башкарат. Багыныңкылыкты талдоо – сүйлөмдөгү сөздөрдүн ортосундагы багыныңкылыктарды автоматтык түрдө анализдөө жана аларды

багыныңкылыктын структурасында көрсөтүү процесси болуп саналат. Синтаксистик талдоодо колдонулган дарак структуралардын эки негизги түрү бар: ыкташууга негизделген жана багыныңкылыкка негизделген. Ыкташууга негизделген дарактар сөз айкаштары боюнча сүйлөмдөрдүн иерархиялык түзүлүшүн билдирет, ал эми багыныңкылыкка негизделген дарактар болсо, багыныңкылык катары сөздөрдүн ортосундагы байланышты аныктайт. Дарак структуралары - сүйлөмдөрдүн аннотацияланган корпусу, алар кол менен аннотацияланат, мисалы, сөз түркүмдөрдүн тегдери, фразанын түзүлүшү же багыныңкылык маалыматы. Дарак структуралары синтаксистик талдоочуларды окутуу жана баалоо үчүн, ошондой эле тилдик кубулуштарды изилдөө жана лингвистикалык теорияларды иштеп чыгуу үчүн колдонулат.

Диссертациянын 3-бөлүмү иштин практикалык өзгөчөлүктөрүн жана ыкмаларын карайт. Бул бөлүмдө биз маалыматтарды чогултуу жана кыргызча сүйлөмдөрдү талдоо процесстери жөнүндө баяндайбыз. Эсептөө лингвистикасы жана табигый тилди иштетүү үчүн маалыматтарды чогултуу процесси, ошондой эле, дарак түрүндөгү структураларды колдонуу жөнүндө кеңири маалымат берилет. Иштин практикалык бөлүгүндө колдонгон маалыматтар 2019-жылы түзүлгөн биринчи кыргыз тилинин корпусунан алынгандыгы баса белгиленет. Бул сүйлөмдөр “Универсалдуу багыныңкылык” долбоорунда болгон морфологиялык тегдерге жана синтаксистик байланыштар аркылуу аныкталып, дарак түрүндөгү структураларга келтирилген. Алдыдагы мисалдарда долбоордогу кээ бир синтаксистик структуралар жана алардын түшүндүрмөлөрү берилген.

Синтаксистик байланыштар	Түшүндүрмө	Кыргыз тилиндеги мисалдар
acl	сүйлөмдүк аныктооч	“ Прокурор берген көрсөтмөлөр милдеттүү болуп саналат.”
acl:relcl	сүйлөмдүк аныктооч (сүйлөмдүн	“ Биз апам алып келгендерди ашканага койдук.”

	ээсине тиешеси жок)	
advcl	бышыктоочтук багыныңкы сүйлөм	“ Кимдир бирөө чуркап бир кесе суу апкеле калды.”
advmod	бышыктоочтук аныктооч	“Албетте, тоолуу өлкөлөрдүн айта берсе көйгөйлөрү абдан көп.”
amod	сын атоочтук аныктооч	“ Тескей жээктеги Тоң, Жети-Өгүз райондорунун табигый кооз жерлери абдан көп.”
appos	кошумча аныктооч	“Кезинде бу саптарды акын Жолон Мамытов жазып калтырган эле.”
aux	жардамчы этиш	“Тегерете койлору жуушап жатат. ”
case	жөндөмө көрсөткүчү	“Дал ушундай кадамга бир кылымдан кийин Сталин барган экен да.”
cc	байламта	“Коргошун, цинк жана башка темир кеңдерин ачканга чоң өбөлгө жаратканы кашкайган чындык!”
ccomp	сүйлөмдүк толуктооч	“ Жамгыр жаап жатат деп ойлойм.”
compound	татаал сөздөр	“Туркиянын айыл чарба кредиттик кооперативдер бирлигинин 600 миңден ашуун мүчөсү бар.”
conj	бир өңчөй сөздөр	“Февралдан тарта Кыргызстан менен Россиянын ортосундагы авиарейстердин

		саны өсөт.”
cop	байламта этиш	“Алар Арстанбапка жетиш үчүн дагы 5 саат жол жүрүш керек эле.”
csubj	сүйлөмдүк ээ	“ Чоң Аярдын айтканы төгүн эмес, чын экен.”
ccomp	сүйлөмдүк толуктооч (негизги сүйлөмдү толуктап турган багыныңкы сүйлөм)	“Жамгыр жаап жатат деп ойлойм. ”
discourse	дискурс байланыштыргыч	“ Ии, кокуй, ошону табышмактатып жаттың беле?”
fixed	туруктуу сөз айкашы (өзгөрбөс)	“анча-мынча, ошондой эле, тилекке каршы”
flat	жалпак түзүлүш бир деңгээлде каралган сөз айкаштары	“Садыр Жапаров, Алпаслан Жейлан, Леонардо ди Каприо”
parataxis	татаал сүйлөмдүн өзгөчө структурасы	“Кыш кырга конбо , жай сайга конбо. ”
vocative	каратма сөздөр	“Жүр, балам , жумуш жасайлы.”

Чогултулган тексттик маалымат морфологиялык жана синтаксистик талдоо үчүн колдонулат. Мындай талдоо өзүнө төмөнкү процесстерди камтыйт:

1. Кептин бөлүктөрүнүн тегдери: Ар бир сөзгө анын грамматикалык категориясын көрсөткөн тег ыйгарылат (мисалы, зат атооч, этиш, сын атооч ж.б.);
2. Синтаксистик структуралар: Сүйлөмдөгү сөздөрдүн ортосундагы грамматикалык байланыштар аныкталат;
3. Семантикалык ролдор: Сүйлөмдөгү ар бир сөздүн же сөз айкашынын маанилик функциясы белгиленет (мисалы, аткаруучу, объект, ж.б.).

Эсептөө лингвистикасында жана табигый тилди иштетүүдө (ТТИ) маалыматтарды чогултуунун эң маанилүү аспектилеринин бири болуп, дарак түрүндөгү структуралар (treebanks) колдонулат. Дарак түрүндөгү структуралар - бул тилдин синтаксистик түзүлүшүн көрсөткөн, башкача айтканда, аннотацияланган сүйлөмдөрдүн топтому. Алар төмөнкү өзгөчөлүктөргө ээ:

1. Иерархиялык түзүлүш: Сүйлөмдөр дарак түрүндөгү иерархиялык структура катары көрсөтүлөт;
2. Синтаксистик байланыштар: Сөздөрдүн ортосундагы грамматикалык мамилелер так көрсөтүлөт;
3. Лингвистикалык аннотациялар: Ар бир сөзгө жана түйүнгө тиешелүү лингвистикалык маалымат кошо берилет.

Мунун негизинде, дарак түрүндөгү структуралар синтаксистик анализ үчүн өтө баалуу ресурска айланат жана алар:

- Сүйлөмдөгү сөздөрдүн ортосундагы татаал мамилелерди визуалдаштырууга мүмкүндүк берет;
- Тилдин грамматикалык структурасын терең түшүнүүгө жардам берет;
- Машиналык окутуу алгоритмдерин синтаксистик талдоо үчүн үйрөтүүдө колдонулат;
- Ар кандай тилдердин синтаксистик өзгөчөлүктөрүн салыштырууга мүмкүндүк түзөт.

Булардын натыйжасында, дарак түрүндөгү структураларды колдонуу аркылуу, изилдөөчүлөр тилдин грамматикалык түзүлүшүн так жана системалуу түрдө

изилдей алышат. Кыргыз тилинин синтаксисин мындай нукта изилдөө кыргыз тилинин синтаксистик структураларын "Универсалдуу багыныңкылык" веб-долбоорунун алкагында иликтөө аркылуу, кыргыз тилин табигый тилди иштетүү жана эсептөө лингвистикасы тармактарында өнүктүрүүгө салым кошуу максатын көздөйт. Бул изилдөө кыргыз тилинин синтаксисин терең түшүнүүгө, аны башка тилдер менен салыштырып иликтөөгө жана келечекте кыргыз тил технологияларын өнүктүрүүгө негиз болуп бере алат.

Изилдөөнүн негизги жыйынтыктары:

Бул диссертациялык иште бир катар теориялык жана практикалык маселелер кеңири талкууга алынды. Табигый тилди иштетүүнүн тарыхы жана маанилүүлүгү изилденип, көрүнүктүү лингвисттердин жана окумуштуулардын синтаксиске байланыштуу теориялары каралды. Синтаксистин жана багыныңкылыкты талдоонун грамматикалык өзгөчөлүктөрүн изилдөө аркылуу "Универсалдуу багыныңкылык" долбоорунун алкагында синтаксистик мамилелерди талдап, кыргыз тилиндеги мисалдары иштелип чыкты. Мындан тышкары, стандартташтырылган эрежелерди колдонуу менен кыргыз тили үчүн дарак түрүндөгү структураларын түзүүнүн маанилүүлүгү жана зарылчылыгы баяндалды.

Изилдөөнүн маанилүүлүгү:

"Универсалдуу багыныңкылык" долбоорунун алкагында кыргыз тилинин синтаксистик структураларынын банкын изилдөө жана иштеп чыгуу бир нече себептерден улам маанилүү экендиги далилденди:

- Табигый тилди иштетүү тапшырмаларын (маалымат алуу, машиналык котормо, ж.б.) жакшыртуу.
- Ар кандай тилдердин синтаксистик структураларын салыштыруу мүмкүнчүлүгүн түзүү.
- Синтаксистин универсалдуу принциптерин терең түшүнүүгө салым кошуу.
- Тилди окутуу жана үйрөнүү процесстерин жакшыртуу.

Теориялык жана практикалык салым:

Теориялык жактан, бул иш тилдин түзүлүшү жана функциясы жөнүндөгү түшүнүгүбүздү тереңдетет. Кыргыз синтаксисинин уникалдуу өзгөчөлүктөрүн аныктоо менен, лингвистикалык теорияга салым кошот.

Практикалык жактан, бул изилдөө:

- Табигый тилди иштетүү системаларын жана машиналык окутуу алгоритмдерин жакшыртууга негиз түзөт.
- Тил үйрөнүүчүлөр, мугалимдер жана изилдөөчүлөр үчүн баалуу ресурс болот.
- Кыргыз тилине мүнөздүү грамматикалык үлгүлөрдү жана конструкцияларды түшүнүүгө жардам берет.

Келечектеги изилдөөлөр үчүн сунуштар:

- Кыргыз тилинин синтаксистик структураларынын банкын кеңейтүү жана аны тереңдетүү.
- Башка түрк тилдери менен салыштырма изилдөөлөрдү жүргүзүү.
- Кыргыз тили үчүн өзгөчө табигый тилди иштетүү куралдарын иштеп чыгуу.

Ошондой эле, кыргыз тилиндеги синтаксистик структуралар тил үйрөнүүчүлөр, мугалимдер жана тилге кызыккан изилдөөчүлөр үчүн баалуу ресурс боло алат. Ал кыргыз тилине мүнөздүү грамматикалык калыптарга жана конструкцияларга түшүндүрмө берип, тилди окутууга жана үйрөнүүгө жардам берет. Мындан тышкары, ал кыргыз тилинин синтаксисин изилдеген изилдөөчүлөр үчүн маалымдама материал катары кызмат кылып, тилдик структураларды так жана деталдуу талдоого мүмкүндүк берет.

Жыйынтыктап айтканда, "Универсалдуу багыныңкылык" долбоорунун алкагында кыргыз тилинин синтаксистик структураларын изилдөө теориялык жана практикалык мааниге ээ. Бул изилдөө тилдин түзүлүшүн жана функцияларын түшүнүүгө, ошондой эле табигый тилди иштетүүдө, тилди үйрөнүүдө жана лингвистикалык изилдөөлөрдү натыйжалуу нукта жүргүзүүдө олуттуу салым кошот.

Ачкыч сөздөр: Корпус лингвистикасы, кыргыз корпусу, синтаксистик талдоо, морфологиялык тегдер, табигый тилди иштетүү (ТТИ), Универсалдуу багыныңкылык, дарак түрүндөгү структура.



GENİŞ ÖZET

“EVRENSEL BAĞIMLILIKLAR” PROJESİ ÇERÇEVESİNDE KIRGIZ DİLİNİN SÖZDİZİMSEL YAPILARININ VERİTABANININ OLUŞTURULMASI

Günümüzün dijital çağı, kullanıcılara internet ortamında bulunan milyonlarca erişilebilir formattaki metni çeşitli amaçlarla kullanma imkanı sunmaktadır. Başka bir deyişle, akıllı telefonlarımızda, tabletlerimizde ve bilgisayarlarımızda kullandığımız bu metinler, belirli bir düzeyde bilgisayar işlemlerinden geçmiş dilsel materyallerdir. Benzer şekilde, çeşitli video materyallerini (film, podcast, animasyon vb.) destekleyen altyazılar da aslında hesaplamalı dilbilim ve **doğal dil işleme (kısaca DDI)** ürünü olarak kabul edilir.

Böylece internet bize muazzam miktarda hazır işlenmiş metin ve veri kullanma imkanı sunmaktadır. Ancak bunun sadece yüzeysel görünümü olduğunu söylemek yanlış olmaz. Çünkü bu metinler kullanıcıya sunulmadan önce birçok ve tekrarlanan makine öğrenimi ve test süreçlerinden geçirilir. Bazen, özellikle başlangıç aşamasında, makinenin dili tanıyıp işlemesi ve onu dile öğretmesi için, dilin insan tarafından manuel olarak işlenmesi gerekebilir. Bazen, bu işleme yarı otomatik yöntemle gerçekleştirilebilir. Dil makine tarafından iyi düzeyde öğrenildiğinde ise (örneğin İngilizce, Almanca vb.), metinler tamamen otomatik yolla işlenir. Dili işlemenin bu yöntemi, bu sürecin süresini önemli ölçüde kısaltır ve kolaylaştırır. Dilin bu tür otomatik veya bilgisayarlı işlemlerini sağlamak amacıyla, bu alanlarla ilgilenen dünya çapındaki bilim insanları dilin her düzeyine göre (örn. fonetik, morfoloji, sözdizimi, anlambilim, edimbilim) veya tüm düzeylerini kapsayan özel web projeleri geliştirmişlerdir. Örneğin, sözdizimi alanında, cümlelerin otomatik olarak analizini sağlayan özel programlar, web platformları çeşitli işlevleri yerine getiren uygulamalar ve araçlar geliştirilmektedir.

Bu tür metinleri sağlayan kaynaklara gelince, bunlar çoğunlukla bu amaçla kullanılmak üzere tasarlanmış programlar aracılığıyla internet ortamında yer alan metinlerdir (örn. haberler, bildirimler, röportajlar, bloglar vb. kaynaklar). Bunların dışında, bu tür metinlere bilimsel araştırmalar, çeşitli türlerde kitaplar, dergiler ve diğer basılı ürünler de dahildir. Zamanla artan bu metinsel verilerin hacmi insanlara sayısız fırsatlar sunsa da, bu olgu doğal dil işleme (DDİ) alanında bir dizi sorunu da beraberinde getirmektedir. Örneğin, belirli bir dilin sözdizimsel doğasını bilgisayarlı işleme

çerçevesinde ele alacak olursak, o dildeki özellikleri de dikkate alarak işleyebilecek programlar geliştirilmelidir.

Bu bağlamda, sözdizimsel olarak işlenmiş metinler, daha spesifik olarak cümleler ve sözdizimsel yapılar, özel kodlanmış etiketlerle (tag'lerle) işaretlendiğinde, bunların bilgisayarlı işlemesi DDİ uygulamaları için özellikle etkili olur. Metnin bu şekilde analizi, ana bağlamı ve dilsel bilgiyi DDİ algoritmaları (örneğin, morfolojik etiketleme ve sözdizimsel analiz) aracılığıyla daha ileri işlemeye zemin hazırlayarak, metindeki gerekli ve/veya önemli bilginin derinlemesine analizini, yorumlanmasını ve genel metinden çıkarılmasını (segmentasyonunu) kolaylaştırır. Sonuçta kullanıcı çeşitli dilsel işlemleri iyi kalitede gerçekleştirebilen çeşitli uygulamaları kullanabilir.

Örneğin, günümüzde cihazlarımız (gadget'larımız) insan dilini tanıyabilir ve onlarla tam olarak iletişim kurabilir. Bunlar arasında günümüzde özellikle popüler hale gelen **yapay zeka** (YZ) da yer almaktadır. Onunla iletişim kurmak veya başka amaçlarla çalışmak için tek gerekli şey internete bağlanmaktır. Çevrimiçi modda erişilebilir olan bu araç (YZ), aslında, yukarıda vurgulandığı gibi, sadece açıklamalı (etiketlenmiş veya kodlanmış) metinler aracılığıyla mümkün olmaktadır. Örneğin, günümüzde dünyanın çeşitli dillerinde (İngilizce, Çince, Rusça, Korece ve diğer) metinsel verileri iyi kalitede yansıtan programlarla karşılaşabiliyoruz. Ancak ne yazık ki, Kırgızca söz konusu olduğunda cihazlarımızın bu tür işlemleri tam olarak veya tatmin edici düzeyde gerçekleştiremediğine tanık oluyoruz. Çünkü Kırgızca henüz tamamen bilgisayarlı işlemekten geçmemiştir. Bu nedenle makine ve gadget'lar onu iyi kalitede tanıyamamaktadır. Sonuçta yukarıda bahsedilen programları veya cihazları Kırgızca için kullanırken bir dizi kaba ve hatalı metin ortaya çıkmaktadır. Bu durum elbette insanı üzmeden edemez.

Bu sorunu çözmek için, daha spesifik olarak Kırgızca'yı bilgisayarlaştırmak ve işlemek amacıyla 2019 yılında ilk 1 (bir) milyon kelimelik Kırgızca derlemi oluşturulmaya başlandı. Çünkü dilsel derlemin materyali kaçınılmaz olarak bu tür bir işlemekten geçirilir ve zamanla yapay zekanın kalitesini iyileştirmede kullanılır. Bu, sırasıyla, Kırgızca konuşan "Siri", "Alice" vb. ses asistanının oluşturulmasına muazzam bir yardım sağlayabilir. Bu nedenle, doğal dil işlemenin ve süreçlerinin ne kadar önemli olduğuna tanık oluyoruz.

Doğal dil işleme (DDİ) alanı, 1900'lerin başında başlayan ve İkinci Dünya Savaşı'ndan sonra önemli başarılarla ulaşan uzun ve zengin bir tarihe sahiptir. Bu alan gelişim yolunda çeşitli önemli dönüm noktaları ve değişikliklerle karşı karşıya kalmıştır.

Doğal dil işlemenin temelini oluşturan fikirler ilk olarak 1900'lerin başında ortaya çıkmıştır. Bu alan, İsviçreli dilbilimci Ferdinand de Saussure'ün çalışmaları sonucunda gelişmiştir. O, dilleri bir sistem olarak karakterize eden bir yöntem geliştirerek, ilk kez dili bir bilim olarak ele alıp üzerinde çalışmaya başlamıştır.

İkinci Dünya Savaşı'ndan sonra gerçekleşen uluslararası etkinlikler, dilleri çevirmenin değerini anlama kavramına yol açmıştır. Bunun sonucunda otomatik çeviri makineleri geliştirme çabaları, doğal dil işlemenin ilk adımları olarak kabul edilmiştir. Böylece, bu çabalar doğal dil işleme alanına resmi olarak yol açmıştır. İkinci Dünya Savaşı, dilin ne kadar önemli olduğunu kanıtlayan bir fenomen haline gelmiştir. Sonuçta bu savaş doğal dil işlemenin ortaya çıkmasına zemin hazırlamış ve ona büyük bir itici güç olarak değerlendirilmiştir. O zamandan beri doğal dil işleme alanı, dilleri çevirmede, onları anlamada, o dillerde metinler üretmede ne kadar önemli olduğunu pratikte doğrulamaktadır.

Bazı önemli aşamalarına değinecek olursak, 1960'larda DDİ alanında önemli ilerlemeler gözlenmiştir. Örneğin, 1964-1966 yıllarında MIT'de (Massachusetts Teknoloji Enstitüsü) ELIZA adlı bir sohbet botu programı geliştirilmiştir. ELIZA, insan-makine arasındaki iletişimi anlamının ilk örneği olarak, bu tür makine ve insan arasındaki etkileşimi geliştiren bir program olarak tanınmıştır. Bu sohbet botu İngilizce olarak insan dilini anlama ve sorularına cevap verme yeteneğine sahipti. ELIZA'dan sonra, 1970-2000 yılları arasında sadece ilginç ve çok yönlü görevleri yerine getirebilen sohbet botları değil, aynı zamanda robotlar da oluşturulmaya ve geliştirilmeye başlanmıştır.

1980'lerin sonu ve 1990'ların başında istatistiksel modeller tanıtılmış, elle yazılmış karmaşık kural setlerine dayanan doğal dil işleme sistemlerinin çoğu eskimiş ve bu alanda yeni bir devrim başladığı gözlemlenmiştir. Makine öğrenimi yöntemlerine geçiş ve hesaplama gücünün artması bu ilerlemenin ana itici güçleri olmuştur. İnsan gibi metinler yazma, onları anlama ve bunlara dayanarak algoritmalar aracılığıyla dilsel modeller oluşturma, doğal dil işlemeyi tamamen yeni bir düzeye getirdiğinden, yeni yöntemlerin kullanımı, özellikle LLM (Large Language Models) 'büyük ölçekli dil modelleri', DDİ'yi son yıllarda muazzam başarılarla ulaştırmıştır.

Günümüzde doğal dil işleme teknolojisi insan yaşamının birçok yönünü etkilediğinden, DDİ modern dünyanın önemli bir alanı haline gelmiştir. Doğal dil işleme yöntemlerinin hızlı gelişimi ile yazma, kodlama ve bu bağlamda düşünme gibi niteliksel faaliyetlerde bilgisayarın insanlarla rekabet edebilecek duruma geldiğini söylemek yanlış olmaz. DDİ metin çıkarma, segmentasyon ve otomasyon yollarını önemli ölçüde kolaylaştırır. Örneğin, yöntemleri internetteki veya diğer çevrimiçi platformlardaki arama sistemlerini optimize etme alanında anahtar kelimeler aracılığıyla bu işlemleri gerçekleştirmek gibi faaliyetler için çok yararlı ve etkili olarak değerlendirilir.

Ayrıca, günümüzün doğal dil işleme yetenekleri bilgisayarın metni okumasına, sesi tanımasına, duyguları anlamasına (duygu analizi) ve iletişimin temel bileşenlerini belirlemesine olanak tanır. Aynı zamanda, bilgisayar ile insan arasındaki iletişimin anadilde gerçekleştirilmesine zemin hazırlar. Bunun açık kanıtı olarak, yukarıda belirtildiği gibi, "Siri", "Alice", "Marusya" gibi ses uygulamalarını sayabiliriz.

Böylece, doğal dil işleme belirli bir dilde karşılaşılan belirsizlikleri gidererek, bunları netleştirerek inceleyerek bilgilerin yapısını belirleyebilir. Bu nedenle, DDİ hukuk, tıp ve diğer birçok alan tarafından yaygın olarak kullanılmaktadır. Çünkü DDİ'nin yetenekleri birçok görevi yerine getirirken veri yönetimi, otomasyon, iletişimi iyileştirme ve teknolojiyi kullanıcıya yakınlaştırmada çok önemli bir rol oynamaktadır. Bu nedenle, doğal dil işleme sohbet botları ve sanal asistanlar için insan konuşmasını anlama ve dille ilgili faaliyetleri yürütme açısından da çok önemlidir. Bu bağlamda, özellikle hesaplamalı dilbilim alanındaki DDİ'nin başarıları muazzam adımlar atmakta ve her geçen gün gelişmektedir.

Hesaplamalı dilbilim, dilbilim ve bilgisayar bilimini birleştiren bir alan olarak bilinir. Bu alan, bilişsel bilim ve bilişim alanındaki akademik başarıları ile değerlidir. Çünkü doğal dil işlemede büyük yardım sağlar ve aynı zamanda yapay zekanın bir parçası olarak kabul edilir. Özellikle, makinelerin insan dilini öğrenmesine, anlamasına ve verilen sorgu veya talebe cevap vermesine olanak tanır. Ayrıca, büyük miktarda metni işlemek için algoritmaların geliştirilmesine imkan vererek otomatik analizde çok önemli bir rol oynar. Böylece, DDİ dil teknolojilerinin gelişimine muazzam katkı sağlar. Bunun açık bir örneği olarak, dil öğrenme uygulamaları ve otomatik çeviri araçlarının küresel iletişimin erişilebilirliğini artırdığını söyleyebiliriz.

Bu olumlu gelişmeler, dilin yapısı ve bilişsel süreçler çerçevesinde yeni keşiflere yol açmaktadır. Sonuçları sanat, sağlık, eğitim, eğlence ve diğer birçok alanda çeşitli fırsatlar sunmaktadır. Bunun sonucunda, çeşitli seviyelerdeki kuruluşlarda bilgisayarlı dil işleme alanında çalışabilecek uzmanlara olan talep giderek artmaktadır.

Örneğin, dili tam olarak açıklamak amacıyla, bu tür faaliyetleri doğru bir şekilde gerçekleştirmek için uluslararası standartlara uygun olarak oluşturulmuş çevrimiçi formatta ve cümleleri otomatik olarak analiz eden özel araçlar ve platformlar geliştirilmektedir. Son yıllarda, bu tür web analizcileri ve çevrimiçi analiz araçları dilbiliminin ayrılmaz bir parçası haline gelmiştir. Her gün milyonlarca "ham" ve işlenmemiş meta veri ilgili uzmanlar tarafından analiz edilmektedir. Sonuçta modern dillerin doğal dil işleme ve otomatik konuşma tanıma konusunda yeni bir seviyeye ulaştığı gözlemlenmektedir.

Kırgız dilinin sözdizimsel açıdan açıklanmasına yönelik çalışmalar nispeten yakın zamanda (2022'de), dünyadaki tüm dillerin birleşik standart açıklamasını hedefleyen **Universal Dependencies (Evrensel Bağımlılıklar)** uluslararası platformu ve web projesi temelinde başlatıldı.

Universal Dependencies (kısaca UD) projesi, dünyanın birçok dilindeki grameri, yapısına dahil edilen cümle öğelerinin (treebank), morfolojik özelliklerini ve sözdizimsel bağımlılıkları tanımlayıp doğrulamak ve UD'nin gerekliliklerine uygun olarak tutarlı bir şekilde cümleleri açıklamak için kullanılır. Bu açık topluluk projesine 500'den fazla kişi katkıda bulunmuş ve dünyada 100'den fazla dilde ağaç yapıları oluşturulmaktadır. Projenin temel amacı, diller arasındaki açıklamaları birleştirmek ve onların aynı bakış açısından açıklanmasını sağlamaktır. Ancak, aynı zamanda belirli bir dile özgü özellikleri korumayı da amaçlamaktadır. UD'de, özellikle farklı diller arasındaki yapısal paralelliği sağlamak için diller arası tutarlı bağımlılık analizi oluşturmaya odaklanılmaktadır.

"Evrensel Bağımlılıklar" projesi kapsamında Kırgızcanın sözdizimsel yapılarını geliştirmek, Kırgızca grameri hakkındaki anlayışımızı derinleştirmeye yönelik önemli ve ilginç bir çalışmadır. Bu proje, dilbilimciler, araştırmacılar ve hesaplamalı dilbilim alanındaki uzmanlar için büyük perspektifler sunmaktadır.

Kırgız dili, sadece Kırgızistan'da değil, Orta Asya'nın çeşitli bölgelerinde ve birkaç yabancı ülkenin bazı bölgelerinde de yaygın olarak ana dil olarak kullanılan bir Türk

dilidir. Dilimiz, benzersiz bir eklemeli gramere ve doğası gereği morfolojik açıdan karmaşık bir yapıya ve özelliklere sahip bir dil olarak tanınmaktadır. Kırgızca için sözdizimsel düzeyde yapı bankasını geliştirerek, araştırmacılar dilin dilbilimsel özelliklerini daha derinden anlayabilir ve doğal dil işleme ve hesaplamalı dilbilimin geniş alanına katkıda bulunabilirler.

Ağaç yapısı (treebank), uygun şemaları kullanarak ve cümledeki veya kelime öbeklerindeki ilişkilerini göstererek cümleleri analiz etmenin özel bir yöntemidir. "Treebank" terimi dilbilimci Geoffrey Leech tarafından 1980'lerde tanıtılmıştır. Ağaç yapıları genellikle kelime türlerinin etiketleriyle işaretlenmiş bir derlem temelinde oluşturulur. Bu yapıda, cümlenin kelime türleri (isim, fiil, sıfat vb.) ve aralarındaki sözdizimsel ilişkiler (özne, yüklem, tamlayan vb.) özel kodlar, yani etiketlerle işaretlenir ve bağımlılık ilişkileri ok yönüyle gösterilir.

Bağımlılık analizi, sözdizimsel analizin yukarıdan aşağıya (top-down) bir yöntemidir. Analiz algoritması, cümledeki baş kelimeyi (yüklemi) belirler ve kelimeler arasındaki bağımlılığı yansıtan bir ağaç yapısı oluşturur. Bu yöntemin temel avantajı, özellikle karmaşık cümle yapısını analiz ederken araştırmacının işini önemli ölçüde kolaylaştırabilmesidir.

Bağımlılık ilişkisi analizi, doğal dil işleme için birkaç nedenden dolayı önemlidir. İlk olarak, bu yöntem kelime öbeklerindeki kelimeler arasındaki dilbilgisel ilişkileri ayırt ederek sözdizimsel bağlantı anlayışını artırır ve karmaşık yapısını açıklayabilir. İkincisi, bilgi çıkarma, soru cevaplama ve analiz gibi görevleri gerçekleştirmek için anlamsal bağlantıları inceler. Ayrıca, yapılandırılmamış (önceden işlenmemiş veya etiketlenmemiş) metin materyalini yapılandırılmış forma dönüştürmeyi kolaylaştırarak bilgi çıkarma ve analiz prosedürlerini hızlandırır.

Bağımlılık analizi, makine çeviri sistemleri için de çok faydalıdır, çünkü farklı dillerdeki kelimeleri ve cümleleri eşleştirerek, sözdizimsel ve anlamsal detayları koruyarak çeviri doğruluğunu artırmaya yardımcı olur. Ayrıca, bu yöntem metindeki öğeleri ve ilişkilerini karşılaştırmaya veya derinlemesine incelemeye yardımcı olur.

Araştırmanın Amacı

“Evrensel Bağımlılıklar” web projesinin amacı, çeşitli dillerdeki sözdizimsel yapıları tek tip bir şekilde yansıtmak, karşılaştırmak ve analiz etmeyi kolaylaştırmaktır. Kırgızca

için sözdizimsel yapılar bankası oluşturarak araştırmacılar, Kırgızca sözdiziminin benzersiz özelliklerini ve kalıplarını daha iyi anlayabilir ve bunları daha ileri düzeyde inceleme ve analizdeki zorlukları belirleyebilirler. Nihai olarak, bu araştırma genel bağımlılık projesi çerçevesinde Kırgızca için kapsamlı ve doğru bir dilbilimsel kaynak oluşturmayı amaçlamaktadır.

Kırgızca sözdizimsel yapılarını proje tarafından belirlenen evrensel standartlara uygun hale getirerek, araştırmacılar karşılaştırmalı dilbilim çalışmaları yürütebilir, makine çeviri sistemlerini iyileştirebilir ve dil işleme algoritmaları geliştirebilirler. Tezde, bu çabaların dilbilimsel önemi, zorlukları ve fırsatları, ayrıca Kırgızca'yı inceleme ve anlama için olası sorunlar geniş çapta tartışılmaktadır. Sonuçta bu proje Kırgızca'yı bilme düzeyini artırmaya, dilbilim araştırmalarının kalitesini iyileştirmeye ve karmaşık dil işleme araçları ve teknolojileri geliştirmeye katkıda bulunmaya zemin hazırlamaktadır.

Bu amaçlara ulaşmak için bir takım görevin yerine getirilmesi gerekmektedir:

- Doğal dil işleme ve dünya dilbilimcilerinin bağımlılık grameri ile tanışma;
- Kırgızca gramerini açıklama;
- Kırgızca sözdizimsel ilişkilerini "Evrensel Bağımlılık" projesi çerçevesinde inceleme;
- Kırgızca sözdizimsel yapılar bankasını geliştirme (ağaç yapısı örnekleri verme);
- Kırgızcadaki sözdizimsel yapıların kapsamlı bir kılavuzunu oluşturma;
- Kırgızca sözdizimsel özelliklerini "Evrensel Bağımlılık" çerçevesine göre analiz etme ve kategorize etme;
- Dilsel analiz için Kırgızcadaki benzersiz sözdizimsel kalıpları ve yapıları belirleme;
- Sözdizimsel özellikleri inceleme ve analiz etme amacıyla Kırgızca'yı araştıranlar, öğrenciler ve öğretmenler için değerli bir dilbilimsel kaynak sağlama;
- Dillerin kapsamını genişletme ve diller arası karşılaştırma ve analiz için bilgi tabanını zenginleştirme;
- Yapılandırılmış sözdizimsel veri seti sağlayarak Kırgızcadaki doğal dil işleme için hesaplama araçları, terimleri ve uygulamaları geliştirmeyi kolaylaştırma;
- Proje için hazırlanan cümleleri resmi olarak sunma.

Bunlar arasında doğal dil işleme ve dünya dilbilimcilerinin bağımlılık grameri ile tanışma görevleri çok önemlidir. Çünkü Kırgızca gramerini açıklama, onun sözdizimsel ilişkilerini "Evrensel Bağımlılık" projesi çerçevesinde inceleme, Kırgızca sözdizimsel yapılar bankasını geliştirme ve diğer birçok görev, Kırgızca gramerinin ve diğer yönlerinin iyi bilinmesi sorununu ortaya koymaktadır.

Bunlara ek olarak, doğal dil işleme alanında her gün yeni terimler ve kavramlar ortaya çıkmaktadır. Bu terimler ve hesaplamalı dilbilimdeki yeni kavramlar dil işlemede büyük öneme sahiptir. Çünkü bunları bilmek, çevirmek ve doğru kullanmak, özellikle Kırgızcada, birçok sorunu, tartışmayı ve müzakereyi beraberinde getirmektedir. Bu bağlamda, Kırgızca'yı açıklama çalışmaları, derlemler oluşturma ve bu projeye katılım yeni başlamış olup, doğal dil işleminin ilk adımları henüz atılmaktadır. Bundan dolayı, ağaç yapıları sunarak Kırgızca terminoloji alanına da katkıda bulunabiliriz. Bu terimleri kullanarak araştırmacıların, öğrencilerin ve öğretmenlerin Kırgızca doğal dil işleme alanına ilgi duymasına ve bu projeye katılmasına zemin hazırlayabileceğimize inanıyoruz.

Araştırmanın Konusu

"Evrensel Bağımlılık" projesi ve Kırgızca sözdizimsel yapısı bu araştırmanın ana konularıdır. Araştırma, Kırgızca gramerini inceleyerek sözdizimsel analizine odaklanmaktadır. "Evrensel Bağımlılık" projesi çerçevesinde Kırgızca için sözdizimsel yapılar bankası geliştirerek, araştırmacılar Kırgızca sözdizimini yansıtabilir ve analizini iyileştirebilirler. Ayrıca, doğal dil işleme ve hesaplamalı dilbilim gibi alanlarda yürütülecek araştırmalara yol açarak bunların geliştirilmesine katkıda bulunacaktır.

Şu anda Kırgızca tam olarak işlenmemiş diller arasında olduğundan, bu araştırmanın ana amacı Kırgızca sözdizimsel yapılarının kapsamlı ve dilbilimsel açıdan doğru bir açıklamasını sağlamaktır. Bu tezin amacına uygun olarak, araştırma dildeki çeşitli sözdizimsel yapıları, bağımlılıkları ve ilişkileri belirler, kategorize eder ve bunları "Evrensel Bağımlılık" web projesinin yönergelerine uygun standartlaştırılmış bir formata getirir.

Araştırmanın bilimsel yeniliği, Kırgızca gramerini küresel proje çerçevesinde analiz etmesi ve farklı dillerdeki sözdizimsel yapıları kolay ve anlaşılır yöntemlerle karşılaştırmalı olarak inceleme imkanı sunmasında yatmaktadır. Bu yönergeleri Kırgız diline uygulayarak, araştırma kaynakları yetersiz diller için sözdizimsel açıklama ve

bağımlılık analizi konusundaki bilginin artmasına katkıda bulunur. Bu, özellikle "yetersiz kaynaklı" dil olarak tanımlanan Kırgız dili için çok önemlidir. Çünkü benzersiz sözdizimsel özelliklere sahip bir Türk dili olmasına rağmen, şu anda kaynakları sınırlı olduğunu kabul etmeliyiz.

Ayrıca, Kırgızca sözdizimsel yapıların geliştirilmesi doğal dil işleme, makine çevirisi ve diğer dil teknolojisiyle ilgili konularda pratik öneme sahip olabilir. Kırgız dili için yüksek kaliteli sözdizimsel olarak açıklanmış bir kaynağın varlığı, dilin özelliklerine uyarlanmış yeni dil teknolojisi araçlarının geliştirilmesine katkıda bulunur.

Araştırmanın Yapısı

Yüksek lisans tezi üç ana bölümden oluşmaktadır. Her bölüm dilbilimi, doğal dil işleme terminolojisi ve ağaç yapılarının oluşturulmasındaki önemli yönleri inceler.

1. bölüm, araştırmanın öncüllerini, önemini, amaçlarını ve temel sorularını yansıtan bir giriş içerir. Türk dilleri ailesine ait olan Kırgızca yaklaşık 7 (yedi) milyon kişi konuşmaktadır. Ancak, birçok insan için iletişim aracı olarak önemli olmasına rağmen, Kırgızca sözdizimsel yapılarını incelemek için mevcut kapsamlı kaynaklar yetersizdir. Bu, Kırgızcanın dilbilgisel kurallarını ve kalıplarını anlamak isteyen araştırmacılar, eğitimciler ve dil severler için ciddi zorluklar ve Kırgız dilini öğrenmede engeller yaratmaktadır.

"Evrensel Bağımlılık" web projesi çerçevesinde yazılan bu çalışma, bu zorluklarla ilgili bazı sorunları çözmeyi amaçlamaktadır. Bu projenin çerçevesi, dilbilimcilerin ve bilgisayar bilimcilerinin birlikte çalışmasını gerektirir. Burada, araştırmacılar, eğitimciler ve bu tür kaynakları geliştirenler tarafından çeşitli dillerin sözdizimsel yapılarını incelemek ve analiz etmek için kullanılabilecek ücretsiz ve açık kaynaklı bir kaynak oluşturma amacı güdülmektedir.

2. bölümde, sözdizimsel bağımlılık alanıyla ilgili literatür gözden geçirilir ve temel teorileri ve sonuçları geniş çapta tartışılır. Bağımlılık analizi, sözdizimsel analizin bir türüdür ve cümledeki kelimeler arasındaki ilişkileri hiyerarşik bağlantılar olarak değil, bağımlılıklar olarak modellemeyi amaçlar. Sözdizimsel analizin öncülerinden biri, 1950'lerde öbek yapı grameri kavramını ortaya atan Noam Chomsky'dir. Chomsky'nin çalışması, temel sözdizimsel kategorilerin bir setinden cümleleri oluşturma kurallarını ve bu kategorileri birleştirme kurallarını tanımlayarak modern sözdizimsel analizin

temelini oluřturmuřtur. Sözdizimsel analiz alanındaki bir diğerk etkili bilim insanı, bağımlılık analizi algoritmalarının geliştirilmesine büyük katkı sağılayan Marie-Catherine de Marneffe'dir. Marneffe'nin çalışmaları, cümledeki kelimeler arasındaki ilişkileri hiyerarşik bağlantılar olarak değil, bağımlılıklar olarak ele almaya odaklanmıştır. Bu, sözdizimsel analizin etkili ve doğru bir şekilde yapılmasını sağılar. ABD'de birkaç arařtırmacı sözdizimsel ve bağımlılık analizi alanına katkıda bulunmuřtur. Bunlar arasında doğıal dildeki metni analiz etmek için karmařık algoritmalar geliřtiren ve analiz modellerini eğıtmek ve değıerlendirmek için ağa yapılarını geliřtiren Daniel Jurafsky, Dan Klein ve Christopher Manning gibi önde gelen bilim insanları yer almaktadır. Onlar, gramerde cümlelerin hiyerarşik yapısını öbek yapılar aracılığıyla tasvir etmişlerdir. Öbek yapı gramerinde cümleler, hiyerarşik bir yapıda düzenlenmiş bileřenleri açısından analiz edilir. Öbek yapı gramerinin kuralları, öbekleri ve bunların nasıl birleşerek cümleleri oluřturduğunu tanımlar. Bağımlılık yapısı, sözdizimsel analizin cümledeki kelimeler arasındaki ilişkileri bağımlılıklar olarak modellemeyi amaçlayan bir başka yaklaşımdır. Bağımlılık yapısında, cümledeki her kelime baş sözcüğüne (örn. yükleme) bağlanır ve onun sözdizimsel rolünü yönetir. Bağımlılık analizi, cümledeki kelimeler arasındaki bağımlılıkları otomatik olarak analiz etme ve bunları bir bağımlılık yapısında gösterme sürecidir. Sözdizimsel analizde kullanılan iki ana ağa yapısı türü vardır: öbek tabanlı ve bağımlılık tabanlı. Öbek tabanlı ağalar, cümlelerin öbek yapıları açısından hiyerarşik yapısını temsil ederken, bağımlılık tabanlı ağalar kelimeler arasındaki ilişkileri bağımlılıklar olarak tanımlar. Ağa yapıları, cümlelerin elle açıklanmış bir derlemesidir, örneğın kelime türü etiketleri, öbek yapısı veya bağımlılık bilgisi ile. Ağa yapıları, sözdizimsel analizcileri eğıtmek ve değıerlendirmek için, ayrıca dilsel olguları incelemek ve dilbilimsel teoriler geliřtirmek için kullanılır.

Tezin 3. bölümü, çalışmanın pratik özelliklerini ve yöntemlerini ele alır. Bu bölümde veri toplama ve Kırgızca cümleleri analiz etme süreçleri hakkında bilgi vereceğiz. Hesaplamalı dilbilim ve doğıal dil işleme için veri toplama süreci ve ayrıca ağa yapılarının kullanımı hakkında kapsamlı bilgi verilecektir. Çalışmanın pratik kısmında kullanılan verilerin 2019 yılında oluřturulan ilk Kırgız dili derleminden alındığı vurgulanır. Bu cümleler, "Evrensel Bağımlılıklar" projesindeki morfolojik etiketler ve sözdizimsel bağlantılarla tanımlanıyor ve ağa benzeri yapılarda sunuluyor. Ařağıdaki örnekler bazı sözdizimsel yapıları ve bunların açıklamalarını göstermektedir.

Sözdizimsel bağlantılar	Açıklama	Kırgızca örnekler
acl	cümle belirleyici	“ Прокурор берген көрсөтмөлөр милдеттүү болуп саналат.”
acl:relcl	cümle belirleyici	“ Биз апам алып келгендерди ашканага койдук.”
advcl	zarf cümlesi	“ Кимдир бирөө чуркап бир кесе суу апкеле калды.”
advmod	zarf değiştirici	“Албетте, тоолуу өлкөлөрдүн айта берсе көйгөйлөрү абдан көп.”
amod	sıfat değiştirici	“ Тескей жээктеги Тоң, Жети-Өгүз райондорунун табигый кооз жерлери абдан көп.”
appos	ek tanımlayıcı	“Кезинде бу саптарды акын Жолон Мамытов жазып калтырган эле.”
aux	yardımcı fiil	“Тегерете койлору жуушап жатат .”
case	ismin halleri	“Дал ушундай кадамга бир кылымдан кийин Сталин барган экен да.”
cc	eşgüdüm konjonktörü	“Коргошун, цинк жана башка темир кеңдерин ачканга чоң өбөлгө жаратканы кашкайган чындык!”
ccomp	cümlecik tümleci	“ Жамгыр жаап жатат деп ойлойм.”
compound	bileşik kelime	“Туркиянын айыл чарба кредиттик кооперативдер бирлигинин 600 миңден

		ашуун мүчөсү бар.”
conj	eşlenik	“Февралдан тарта Кыргызстан менен Россиянын ортосундагы авиарейстердин саны өсөт.”
cop	kopula	“Алар Арстанбапка жетиш үчүн дагы 5 саат жол жүрүш керек эле.”
csubj	cümlecik öznesi	“ Чоң Аярдын айтканы төгүн эмес, чын экен.”
ccomp	cümle tamamlayıcısı (ana cümleyi tamamlayan alt cümle)	“Жамгыр жаап жатат деп ойлойм. ”
discourse	söylem bağlayıcı	“ Ии, кокуй, ошону табышмактатып жаттың беле?”
fixed	sabit ifade (değişmez)	“анча-мынча, ошондой эле, тилекке каршы”
flat	Düz bir yapı, bir düzeyde ele alınan bir dizi ifadedir	“Садыр Жапаров, Алпаслан Жейлан, Леонардо ди Каприо”
parataxis	karmaşık bir cümlelerin özel yapısı	“Кыш кырга конбо , жай сайга конбо. ”
vocative	vokatif, bir metinde hitap edilen diyalog katılımcısını işaretlemeye	“Жүр, балам , жумуш жасайлы.”

Toplanan metin verisi morfolojik ve sözdizimsel analiz için kullanılır. Bu tür bir analiz aşağıdaki süreçleri içerir:

1. Konuşma bölümlerinin etiketlenmesi: Her kelimeye dilbilgisel kategorisini gösteren bir etiket atanır (örneğin, isim, fiil, sıfat vb.);
2. Sözdizimsel yapılar: Cümledeki kelimeler arasındaki dilbilgisel ilişkiler belirlenir;
3. Anlamsal roller: Cümledeki her kelimenin veya kelime öbeğinin anlamsal işlevi işaretlenir (örneğin, yapan, nesne, yer vb.).

Hesaplamalı dilbilim ve doğal dil işlemede (DDİ) veri toplamanın en önemli yönlerinden biri ağaç yapılarının (treebanks) kullanılmasıdır. Ağaç yapıları, dilin sözdizimsel yapısını gösteren, başka bir deyişle açıklanmış cümlelerin bir koleksiyonudur. Bunlar aşağıdaki özelliklere sahiptir:

1. Hiyerarşik yapı: Cümleler ağaç benzeri hiyerarşik bir yapı olarak gösterilir;
2. Sözdizimsel ilişkiler: Kelimeler arasındaki dilbilgisel ilişkiler açıkça gösterilir;
3. Dilbilimsel açıklamalar: Her kelimeye ve düğüme ilgili dilbilimsel bilgi eklenir.

Buna dayanarak, ağaç yapıları sözdizimsel analiz için çok değerli bir kaynak haline gelir ve şunları sağlar:

- Cümledeki kelimeler arasındaki karmaşık ilişkileri görselleştirme imkanı sunar;
- Dilin dilbilgisel yapısını derinlemesine anlamaya yardımcı olur;
- Makine öğrenimi algoritmalarını sözdizimsel analiz için eğitmede kullanılır;
- Farklı dillerin sözdizimsel özelliklerini karşılaştırma imkanı sağlar.

Bunların sonucunda, ağaç yapılarını kullanarak araştırmacılar dilin dilbilgisel yapısını kesin ve sistematik bir şekilde inceleyebilirler. Kırgızca sözdizimini bu şekilde incelemek, Kırgız dilinin sözdizimsel yapılarını "Evrensel Bağımlılık" web projesi çerçevesinde araştırarak, Kırgız dilini doğal dil işleme ve hesaplamalı dilbilim alanlarında geliştirmeye katkıda bulunmayı amaçlamaktadır. Bu araştırma, Kırgız dilinin sözdizimini derinlemesine anlamaya, onu diğer dillerle karşılaştırmalı olarak incelemeye ve gelecekte Kırgız dil teknolojilerini geliştirmeye temel oluşturabilir.

Sonuç

1. Araştırmanın Temel Bulguları:

Bu tez çalışmasında, bir dizi teorik ve pratik mesele kapsamlı bir şekilde ele alınmıştır. Doğal dil işlemenin tarihi ve önemi incelenmiş, önde gelen dilbilimcilerin ve bilim

insanlarının sözdizimi ile ilgili teorileri gözden geçirilmiştir. Sözdiziminin ve bağımlılık analizinin gramatikal özelliklerini inceleyerek, "Evrensel Bağımlılıklar" projesi çerçevesinde sözdizimsel ilişkiler analiz edilmiş ve Kırgızca örnekler geliştirilmiştir. Ayrıca, standartlaştırılmış kurallar kullanarak Kırgız dili için ağaç yapılarının oluşturulmasının önemi ve gerekliliği açıklanmıştır.

2. Araştırmanın Önemi:

"Evrensel Bağımlılıklar" projesi çerçevesinde Kırgızca sözdizimsel yapılarının bankasını incelemek ve geliştirmek, aşağıdaki nedenlerden dolayı önemli olduğu kanıtlanmıştır:

- Doğal dil işleme görevlerini (bilgi çıkarımı, makine çevirisi vb.) iyileştirmek.
- Farklı dillerin sözdizimsel yapılarını karşılaştırma imkanı sağlamak.
- Sözdiziminin evrensel ilkelerinin derin anlaşılmasına katkıda bulunmak.
- Dil öğretimi ve öğrenme süreçlerini geliştirmek.

3. Teorik ve Pratik Katkı:

Teorik açıdan, bu çalışma dilin yapısı ve işlevi hakkındaki anlayışımızı derinleştirmektedir. Kırgız sözdiziminin benzersiz özelliklerini tanımlayarak dilbilim teorisine katkıda bulunmaktadır.

4. Pratik açıdan, bu araştırma:

- Doğal dil işleme sistemlerini ve makine öğrenimi algoritmalarını geliştirmek için bir temel oluşturmaktadır.
- Dil öğrenenler, öğretmenler ve araştırmacılar için değerli bir kaynak olmaktadır.
- Kırgız diline özgü gramatik kalıpları ve yapıları anlamaya yardımcı olmaktadır.

5. Gelecekteki Araştırmalar için Öneriler:

- Kırgızca sözdizimsel yapılar bankasını genişletmek ve derinleştirmek.
- Diğer Türk dilleriyle karşılaştırmalı çalışmalar yürütmek.
- Kırgız dili için özel doğal dil işleme araçları geliştirmek.

Ayrıca, Kırgızca sözdizimsel yapılar, dil öğrenenler, öğretmenler ve dile ilgi duyan araştırmacılar için değerli bir kaynak olabilir. Kırgız diline özgü gramatik kalıpları ve yapıları açıklayarak dil öğretimi ve öğrenimine yardımcı olur. Bunun yanı sıra, Kırgızca sözdizimini inceleyen araştırmacılar için referans materyal olarak hizmet ederek, dilsel yapıların kesin ve ayrıntılı analizine olanak tanır.

Sonuç olarak, "Evrensel Bağımlılıklar" projesi çerçevesinde Kırgızca sözdizimsel yapılarını incelemek hem teorik hem de pratik öneme sahiptir. Bu araştırma, dilin yapısını ve işlevlerini anlamaya, ayrıca doğal dil işlemede, dil öğreniminde ve dilbilimsel araştırmaların etkili bir şekilde yürütülmesinde önemli katkılar sağlamaktadır.

Anahtar Kelimeler: Derlem dilbilimi, Kırgızca derlem, sözdizimsel analiz, morfolojik etiketler, doğal dil işleme (DDİ), Evrensel Bağımlılıklar, ağaç yapısı.



АБСТРАКТ

РАЗРАБОТКА БАНКА СИНТАКСИЧЕСКИХ СТРУКТУР КЫРГЫЗСКОГО ЯЗЫКА В РАМКАХ ПРОЕКТА "УНИВЕРСАЛЬНЫЕ ЗАВИСИМОСТИ"

Технологии привели к значительному изменению отношений между программным обеспечением и языком, сделав их взаимосвязанными. Мы ежедневно обрабатываем естественный язык, сталкиваясь с миллионами оцифрованных текстов. В настоящее время понимание технологии обработки естественного языка (ОЕЯ) имеет решающее значение для лингвистов всего мира. Как никогда ранее, корпусные исследования и современная лингвистика становятся сложными, но необходимыми областями. Языковые корпуса применяются для исследования языка на основных его уровнях: фонетике, морфологии, синтаксисе, семантике и прагматике. Кроме того, язык пересекает границы исключительно лингвистического характера, когда он сталкивается с другими областями, такими как теоретическая/прикладная лингвистика, компьютерная лингвистика, социолингвистика, психолингвистика и другие. Появилась возможность изучать язык на всех уровнях, используя различные онлайн-инструменты и лингвистические платформы.

Появились такие веб-проекты, как Universal Dependencies (UD), создающие автоматические парсеры для анализа зависимостей. UD - это проект открытого сообщества, целью которого является последовательное описание грамматики языков мира и их объединение в стандартизированные категории. Это открытый проект сообщества, в рамках которого международные лингвисты и инженеры-программисты создают банки деревьев для мировых языков.

На сайте представлено очень мало примеров аннотаций на тюркских языках, есть на турецком, татарском и ряд предложений на казахском. К сожалению, на сайте очень мало синтаксических структур на кыргызском языке, что говорит о недостаточном уровне вовлеченности кыргызского языка в мировые инновации.

На факультете синхронного перевода КТМУ преподаватели и студенты сами начали создавать и присылать примеры синтаксической разметки на основе UD.

К сожалению, на данный момент кыргызский язык только вошел в список отобранных языков этого веб-проекта, и кыргызские эквиваленты и руководства по использованию терминов в этой новой области не очень известны. До сих пор

существует множество ошибок в аннотациях, а также отсутствуют справочники. В связи с этим целью данной работы является введение расширенного понятия «Компьютерная лингвистика» для кыргызского языка и демонстрация синтаксических описаний на основе недавно созданных корпусов кыргызского языка с использованием международной платформы UD.

В нашей диссертации мы обсудим примеры моделей синтаксического анализа кыргызского языка, которые мы подготовили в качестве предложения для вышеупомянутого веб-проекта. Эта деятельность по аннотированию позволит значительно расширить потенциал кыргызского языка за счет анализа различных предложений. Мы надеемся проанализировать грамматику на новом уровне, используя новые ресурсы и платформы. Цель данного исследования - ввести расширенное понятие универсальных зависимостей в кыргызском языке, чтобы показать морфологические и синтаксические связи.

В диссертации используются различные исследовательские методы. При написании первой и второй частей диссертации были использованы различные статьи, диссертации, книги и краткие обзоры. Соответствующая литература приведена в последнем разделе данного документа. Для написания третьей практической части исследования были использованы различные веб-сайты и платформы.

Мы надеемся, что различные исследования в этой области помогут избежать многих проблем, возникающих в лингвистике. Поэтому одной из задач нашей работы является адаптация терминов веб-проекта Universal Dependencies (UD) к кыргызскому языку, анализ предложений на кыргызском языке с помощью этого проекта, а также загрузка готовых предложений на сайт. Результаты исследования могут быть использованы в преподавании, изучении и проведении лингвистических исследований в области корпусной лингвистики, семантики, синтаксиса, морфологии и теории перевода с пользой как для преподавателей, так и для учащихся.

Ключевые слова: Корпусная лингвистика, корпус кыргызского языка, синтаксический анализ, морфологические теги, обработка естественного языка (NLP), Универсальные Зависимости, банк деревьев.

INTRODUCTION

DEVELOPING OF A BANK OF SYNTACTIC STRUCTURES OF THE KYRGYZ LANGUAGE WITHIN THE FRAMES OF THE “UNIVERSAL DEPENDENCIES” PROJECT

There are millions of raw and annotated texts available online for us to explore every day. These writings cover a wide range of subjects, languages, and genres of media, from news stories and blog posts to research papers and product evaluations. The increasing amount of textual data offers a wealth of ideas, expressions, and information about people, but it also poses a number of possibilities and problems for the field of natural language processing (NLP).

Since annotated texts have been fully tagged with linguistic annotations such as part-of-speech tags, named entities, syntactic structures, and semantic labels, they are especially useful for natural language processing (NLP) applications. The annotations supply fundamental context and metadata, which facilitate the analysis, interpretation, and extraction of significant information from the text by NLP algorithms.

However, the majority of textual material accessible on the internet consists of raw texts, or texts that have not been manually annotated. Even without the organized annotations included in annotated texts, raw texts offer a variety of language features, trends, and patterns. NLP has particular difficulties when mining knowledge from raw texts, such as resolving obstacles, ambiguity, and linguistic variance across different genres, languages, and registers.

Precisely to perform the actions above for full language annotation, various online parsers and platforms have been created to annotate sentences within international standards. Over the years, these web parsers and online parsing tools have emerged in the life of linguistics. Millions and millions of raw metadata are being analyzed every day by language specialists. Modern languages have stepped into a new level of Natural Language Processing and automatic speech recognition. Even our gadgets are able to recognize human language and have a full conversation with them. The only necessary thing to communicate with Artificial Intelligence is to have access to internet. We have been able to use these annotated data as much as we want in advanced languages as English, Chinese, Russian, Korean, and others. Unfortunately, things are very different

with the Kyrgyz language. New all kinds of recent steps towards annotation of the Kyrgyz language started only from 2019 with the creation of the very first corpus. But these processes covered only the morphological level of the language at that time. The transition to syntactic annotation of the Kyrgyz language started relatively recently with the introduction of the international platform and web project Universal Dependencies, aimed at one-standard annotation of all languages of the world.

The uniqueness of this project is that you can put any currently existing language of the world under its standards and freely analyze the sentences. These standardized features enable linguists from all countries to unite and try to put the grammar of any language under UD standards. Special tags unique to UD allow a thorough analysis of language and grammatical dependencies. A bank of tree structures of all the world's languages also allows cross-linguistic analysis of a particular language.

Enhancing our comprehension of the grammar and syntax of the Kyrgyz language is the goal of the significant and captivating work being done on the development of the language's syntactic structures within the "Universal Dependencies" project. Under this project, we would be able to create a comprehensive and linguistically accurate annotation of the syntactic structures of the Kyrgyz language. This involves identifying and categorizing the various syntactic constructs, dependencies, and relationships within a language and expressing them in a standardized format that aligns with the UD guidelines.

We truly believe that we could elaborate the grammar and the usage of our mother tongue to another international level. Many significant works in this field have already been done due to the diligence of initiatives and professors of our Simultaneous Translation and Computer Linguistics departments of Kyrgyz-Turkish Manas University.

CHAPTER 1. THE GENERAL OUTLOOK OF THE THESIS PAPER AND INTRODUCTION TO UNIVERSAL DEPENDENCIES WEB PROJECT

1.1 Background and significance of the study

The development of a bank of syntactic structures of the Kyrgyz language within the frames of the "Universal Dependencies" project is a significant and exciting endeavor that aims to enhance our understanding of the grammar and syntax of the Kyrgyz language. This project, which focuses on annotating and parsing natural language data to create a unified framework of syntactic structures, holds great promise for linguists, researchers, and computational linguists alike. The Kyrgyz language, a Turkic language spoken primarily in Kyrgyzstan, presents a unique set of challenges and nuances in its syntax and grammar. By developing a bank of syntactic structures for the Kyrgyz language, researchers can not only gain a deeper insight into the language's linguistic properties but also contribute to the broader field of natural language processing and computational linguistics.

The "Universal Dependencies" project, which aims to create cross-linguistically consistent treebanks of syntactic dependencies for a wide range of languages, provides a valuable framework for annotating and analyzing the syntactic structures of the Kyrgyz language. By aligning the syntactic structures of the Kyrgyz language with the universal standards established by the project, researchers can facilitate comparative linguistic studies, improve machine translation systems, and enhance language processing algorithms. In this introduction, we will explore the importance of developing a bank of syntactic structures for the Kyrgyz language within the frames of the "Universal Dependencies" project. We will discuss the linguistic significance of this endeavor, the challenges and opportunities it presents, and the potential implications for the study and understanding of the Kyrgyz language. Ultimately, this project has the potential to advance our knowledge of the Kyrgyz language, promote cross-linguistic research, and contribute to the development of sophisticated language processing tools and technologies.

Having a standardized framework for analyzing Kyrgyz language data would have several positive impacts on research and data analysis in the field. Firstly, it would ensure consistency and comparability across different studies and data sets. Researchers would be able to use the same standards and methodologies for analyzing language data, making it easier to compare results and draw meaningful conclusions. Secondly, a

standardized framework would help to improve the quality of research and data analysis in the field. Researchers would be required to follow established guidelines and best practices, reducing the risk of errors and inconsistencies in their analysis. Additionally, a standardized framework would provide a common language and set of tools for researchers working in the field of Kyrgyz language analysis. This would facilitate collaboration and knowledge-sharing, ultimately advancing the overall understanding of the Kyrgyz language and its different linguistic features. So, having a standardized framework for analyzing Kyrgyz language data would help to promote rigor and consistency in research, improve the quality of data analysis, and facilitate collaboration among researchers in the field.

Motivation for research

There are several reasons why researching and developing a bank of syntactic structures of the Kyrgyz language within the frames of the "Universal Dependencies" project would be valuable and important. First and foremost, having a comprehensive bank of syntactic structures for the Kyrgyz language would greatly benefit natural language processing tasks such as text analysis, information retrieval, machine translation, and sentiment analysis. This is especially important in the context of the increasing digitalization of information and the growing importance of computational linguistics in various fields. Furthermore, contributing to the "Universal Dependencies" project by developing a bank of syntactic structures for the Kyrgyz language would help facilitate cross-linguistic research and comparison of syntactic structures across different languages. This can lead to a better understanding of universal principles of syntax and contribute to the advancement of linguistic theory. Additionally, developing a bank of syntactic structures for the Kyrgyz language can also have practical applications in language teaching and learning, as well as in the development of language processing tools for speakers of Kyrgyz. The motivation for researching and developing a bank of syntactic structures for the Kyrgyz language within the framework of the "Universal Dependencies" project is multifaceted and significant.

First and foremost, despite its cultural and historical significance, the resources and tools available for the computational analysis of the Kyrgyz language are limited. Therefore, the development of a bank of syntactic structures will provide a valuable resource for linguistic research, natural language processing, and the development of

language technology for Kyrgyz. Secondly, the "Universal Dependencies" project aims to create cross-linguistically consistent treebanks of syntactic structures for a wide range of languages. By participating in this project, and specifically focusing on the Kyrgyz language, researchers can contribute to the broader goal of creating a comprehensive and diverse dataset that facilitates the comparison of syntactic structures across languages. This comparative approach not only enriches our understanding of language universals and typological diversity but also provides valuable insights for language-independent natural language processing applications. Furthermore, the development of a bank of syntactic structures for the Kyrgyz language holds practical implications for various fields, including machine translation, information retrieval, and language learning applications. Access to consistent and well-annotated syntactic structures can significantly enhance the accuracy and performance of computational systems that rely on linguistic analysis, ultimately benefiting both the academic community and the general public. Lastly, this research project offers an opportunity to collaborate with linguists, computational linguists, and language technologists, both within and outside the Kyrgyzstan region. By fostering collaborations and knowledge exchange, the project can contribute to the advancement of linguistic research and technology development in the Kyrgyz language, ultimately strengthening linguistic diversity and inclusivity within the broader field of natural language processing. In conclusion, the motivation for research on developing a bank of syntactic structures for the Kyrgyz language within the "Universal Dependencies" project is driven by the need to address the scarcity of linguistic resources and tools for Kyrgyz, contribute to cross-linguistic research efforts, enhance language technology applications, and foster collaboration and knowledge exchange in the field of computational linguistics.

Topicality of the research

The research on developing a Bank of Syntactic Structures of the Kyrgyz Language within the frames of the "Universal Dependencies" project is highly topical and significant. This is because such a project can greatly contribute to the field of natural language processing (NLP) and computational linguistics by providing a comprehensive and standardized dataset of syntactic structures in the Kyrgyz language. With the increasing focus on building language resources for under-resourced languages, the development of a Bank of Syntactic Structures for the Kyrgyz language can help

improve the accuracy and performance of NLP tools and applications for this language. This research can also facilitate cross-linguistic research and comparisons, as well as support the development of machine translation and other language technology applications for Kyrgyz. Furthermore, participating in the "Universal Dependencies" project can help standardize the annotation and representation of syntactic structures in the Kyrgyz language, making the data more interoperable and accessible for researchers and developers working on multilingual NLP projects. This research is highly topical and has the potential to make a significant contribution to the field of linguistics and NLP.

1.2 Research objectives and research questions

The objective of this research is to develop a bank of syntactic structures of the Kyrgyz language using the Universal Dependencies framework. This framework aims to provide a unified representation of syntactic structures across different languages, allowing for easier comparison and analysis. By compiling a bank of syntactic structures for the Kyrgyz language within this framework, researchers can better understand the unique features and patterns of Kyrgyz syntax, as well as identify areas for further study and analysis. Ultimately, this research aims to contribute to the broader goal of building a comprehensive and accurate linguistic resource for the Kyrgyz language within the Universal Dependencies project.

The following tasks had to be accomplished in order to achieve these set goals:

- Compile a comprehensive inventory of syntactic structures present in the Kyrgyz language.
- Analyze and categorize syntactic features of the Kyrgyz language according to the Universal Dependencies framework.
- Identify unique syntactic patterns and constructions in the Kyrgyz language for linguistic research and analysis.
- Establish a standardized representation of Kyrgyz syntactic structures within the Universal Dependencies framework for consistency and comparability.
- Provide a valuable linguistic resource for researchers, students, and speakers of the Kyrgyz language to study and analyze syntactic properties.

- Contribute to the broader Universal Dependencies project by expanding the scope of languages represented and enriching the knowledge base for cross-linguistic comparison and analysis.
- Facilitate the development of computational tools and applications for natural language processing in Kyrgyz by providing a structured syntactic dataset.
- Enhance understanding of the syntactic properties of the Kyrgyz language, contributing to linguistic typology and theoretical research in syntax.

The subject of the research is the development of a bank of syntactic structures for the Kyrgyz language as part of the "Universal Dependencies" project. This project aims to create a universal set of syntactic annotations for a wide range of languages, allowing for cross-linguistic comparison and analysis. The focus of this research is specifically on the Kyrgyz language, a Turkic language spoken primarily in Kyrgyzstan. By developing a bank of syntactic structures for Kyrgyz within the framework of the "Universal Dependencies" project, researchers aim to improve the representation and analysis of the language's syntax, facilitating future research in areas such as natural language processing and computational linguistics.

The object of the research on the topic of developing a bank of syntactic structures for the Kyrgyz language within the framework of the "Universal Dependencies" project is to create a comprehensive and linguistically accurate annotation of the syntactic structures of the Kyrgyz language. This involves identifying and categorizing the various syntactic constructions, dependencies, and relations within the language, and representing them in a standardized format that is compatible with the "Universal Dependencies" guidelines.

The scientific novelty of the research: The scientific novelty of the research lies in its focus on creating a comprehensive and linguistically accurate annotation of the syntactic structures of the Kyrgyz language within the framework of the "Universal Dependencies" project. This project aims to develop cross-linguistically consistent syntactic annotation guidelines for a wide range of languages, allowing for easier comparison and analysis of syntactic structures across different languages. By applying these guidelines to the Kyrgyz language, the research contributes to the growing body of knowledge on syntactic annotation and dependency parsing for under-resourced languages. This is particularly significant for the Kyrgyz language, as it is a Turkic

language with unique syntactic properties that may not be fully represented in existing resources. Furthermore, the development of a bank of syntactic structures for the Kyrgyz language can have practical applications in natural language processing, machine translation, and other language technology tasks. The availability of a high-quality syntactic annotation resource for Kyrgyz can facilitate the development of language technology tools that are tailored to the specific characteristics of this language.

The theoretical importance of developing a bank of syntactic structures of the Kyrgyz language within the frames of the "Universal Dependencies" project lies in advancing our understanding of the structure and function of the language. By systematically analyzing and categorizing the syntactic structures of Kyrgyz, researchers can gain insights into the underlying principles that govern sentence construction and meaning in this particular language. This can contribute to linguistic theory and help to elucidate the unique features of Kyrgyz syntax. From a practical perspective, the development of a syntactic structure bank for Kyrgyz can have a range of applications. First and foremost, it can improve natural language processing systems and machine learning algorithms that rely on syntactic analysis to understand and generate human language. By annotating and categorizing the syntactic structures of Kyrgyz, researchers can build more accurate and efficient language processing tools that can benefit a wide range of applications, from machine translation to information retrieval.

Also, the bank of Kyrgyz syntactic structures can be a valuable resource for language learners, teachers, and researchers interested in the language. It can provide insight into the grammatical patterns and constructions characteristic of the Kyrgyz language, helping in teaching and learning the language. In addition, it can serve as a reference material for researchers studying the syntax of the Kyrgyz language, allowing for a more accurate and detailed analysis of language structures. So, the development of a bank of syntactic structures of the Kyrgyz language within the framework of the Universal Dependencies project has both theoretical and practical significance, contributing to our understanding of the structure and function of language, as well as contributing to progress in natural language processing, Language Learning and Linguistic Research.

1.3 Overview of the structure of the thesis

Chapter 1 provides an introduction to the topic, outlining the background and significance of the study, as well as the research objectives and questions. The Kyrgyz language, a member of the Turkic language family, is spoken by approximately 5 million people primarily in Kyrgyzstan but also in neighboring countries such as Kazakhstan, Uzbekistan, and China. Despite its importance as a means of communication for many people, there is a lack of comprehensive resources available for studying the syntactic structures of the Kyrgyz language. This presents a significant challenge for researchers, educators, and language enthusiasts interested in understanding the grammatical rules and patterns of Kyrgyz. The "Universal Dependencies" project aims to overcome this challenge by creating a bank of syntactic structures for a wide range of languages, including Kyrgyz. This project is a collaborative effort that brings together linguists, computer scientists, and language enthusiasts to create a free, open-source resource that can be used by researchers, educators, and developers to study and analyze the syntactic structures of different languages. The development of a bank of syntactic structures for the Kyrgyz language within the frames of the "Universal Dependencies" project is essential for several reasons. Firstly, it will provide researchers with a valuable resource for studying the grammatical rules and patterns of Kyrgyz. This will enable researchers to analyze the syntax of Kyrgyz more effectively, leading to a deeper understanding of the language's structure. Furthermore, a bank of syntactic structures for Kyrgyz will be a valuable resource for educators who teach the language. By providing teachers with a comprehensive resource for studying the syntactic structures of Kyrgyz, they will be better equipped to design effective language learning programs and materials.

In Chapter 2, we review the existing literature, exploring key theories and findings in this area. Dependency parsing is a type of syntactic analysis that focuses on modeling the relationships between words in a sentence as dependencies, rather than as hierarchical relationships. One of the pioneers of syntactic analysis is Noam Chomsky, who introduced the concept of phrase structure grammar in the 1950s. Chomsky's work laid the foundation for modern syntactic analysis by defining rules for generating sentences from a set of basic syntactic categories and rules for combining these categories. Another influential figure in the field of syntactic analysis is Marie-Catherine de Marneffe, who has made significant contributions to the development of

dependency parsing algorithms. Marneffe's work focuses on modeling the relationships between words in a sentence as dependencies, rather than as hierarchical relationships, which allows for more efficient and accurate syntactic analysis. In the USA, several researchers have contributed to the field of syntactic analysis and dependency parsing. Some notable contributors include *Daniel Jurafsky, Dan Klein, and Christopher Manning*, who have developed sophisticated algorithms for parsing natural language text and have contributed to the development of treebanks for training and evaluating parsing models. Phrase structure grammar is a formalism for describing the hierarchical structure of sentences in terms of phrases. In phrase structure grammar, sentences are analyzed in terms of their constituent parts, which are organized into a hierarchical structure. The rules of phrase structure grammar define how words can be combined to form phrases and how phrases can be combined to form sentences. Dependency structure is a different approach to syntactic analysis that focuses on modeling the relationships between words in a sentence as dependencies. In dependency structure, each word in a sentence is connected to its head word, which governs its syntactic role. Dependency parsing is the process of automatically analyzing the dependencies between words in a sentence and representing them in a dependency structure. There are two main types of treebanks used in syntactic analysis: constituency-based and dependency-based. Constituency-based treebanks represent the hierarchical structure of sentences in terms of phrases, while dependency-based treebanks represent the relationships between words as dependencies. Treebanks are annotated corpora of sentences that have been manually annotated with syntactic information, such as part-of-speech tags, phrase structure, or dependency information. Treebanks are used for training and evaluating syntactic parsers, as well as for studying linguistic phenomena and developing linguistic theories.

Chapter 3 deals with the practical part of the work. In this chapter, we tried to collect certain data and start analyzing the Kyrgyz sentences. The data collection process for computational linguistics and natural language processing (NLP) involves gathering large quantities of text data from various sources such as books, articles, websites, social media, and other text-based sources. The data we used in the practical part comes from the first Kyrgyz Language Corpora created in 2019 (Corpora, 2019). This text data is then annotated with linguistic information, such as part-of-speech tags, syntactic structures, semantic roles, and more. One key aspect of data collection in computational

linguistics and NLP is the use of Treebanks. Treebanks are collections of annotated sentences that capture the syntactic structure of a language. These tree structures represent the relationships between words in a sentence, providing valuable information for syntactic analysis.

The Role of Treebanks in Syntactic Analysis in NLP. Treebanks play a crucial role in syntactic analysis in NLP. They serve as training data for various NLP tasks such as parsing, part-of-speech tagging, named entity recognition, and more. By using treebanks, NLP researchers can build and evaluate syntactic parsers that automatically analyze the syntactic structure of sentences.

The Importance of Treebanks in the Creation of Language Corpora. Treebanks are essential for the creation of language corpora, which are large collections of text data used for linguistic research. Treebanks provide a structured representation of language data, making it easier to analyze and extract linguistic information. They serve as a valuable resource for building and evaluating NLP models, improving the accuracy and performance of various NLP tasks.

Annotation Guidelines and Criteria for Syntactic Structures. When creating treebanks, researchers follow annotation guidelines and criteria for syntactic structures. These guidelines specify how to annotate different syntactic categories, such as nouns, verbs, adjectives, prepositions, and more. They also define the criteria for identifying syntactic relationships between words in a sentence, ensuring consistency and accuracy in the annotation process.

Explanation of Dependency Parsing Techniques and Tools Used. Dependency parsing is a common technique in NLP that analyzes the relationships between words in a sentence by representing them as directed links (dependencies) between words. These tools automatically analyze the syntactic structure of sentences, creating dependency trees that capture the grammatical relationships between words.

In summary, the data collection process and annotation in computational linguistics and NLP involves gathering text data and annotating it with linguistic information. Treebanks play a crucial role in syntactic analysis, providing structured representations of language data for building and evaluating NLP models. Annotation guidelines and criteria ensure consistency and accuracy in syntactic annotation, while dependency parsing techniques and tools analyze the relationships between words in a sentence.

Deduction to Chapter 1

The chapter discusses the creation of a syntactic structure bank for the Kyrgyz language, a project aimed at improving language processing algorithms and facilitating cross-linguistic research. It highlights the need for linguistic resources in Kyrgyz, practical applications, and the potential for collaboration. The research's relevance lies in natural language processing, computational linguistics, and under-resourced language technology development. It aims to ensure consistency, comparability, and quality of research, ultimately enhancing the Kyrgyz language's understanding.

The paper aims to develop a bank of syntactic structures for the Kyrgyz language based on the Universal Dependencies web project, focusing on linguistic resources, practical applications, cross-linguistic research, and knowledge exchange in computational linguistics. This research depicts the significance of natural language processing and computational linguistics. This involves compiling, categorizing, and representing syntactic features to enhance linguistic resources and facilitate cross-linguistic comparison. The project's scientific novelty lies in its focus on creating a comprehensive and linguistically accurate annotation for the Kyrgyz language, contributing to theoretical and practical advancements in linguistic research and natural language processing. The theoretical importance of this research lies in advancing our understanding of Kyrgyz syntax and contributing to linguistic theory, while the practical significance extends to language learning, teaching, and the development of language technology tools. Overall, the project seeks to contribute to the broader goal of building a comprehensive linguistic resource for the Kyrgyz language within the Universal Dependencies project.

CHAPTER 2. SYNTACTIC ANALYSIS AND DEPENDENCY PARSING

OVERVIEW

2.1 Phrase Structure, Constituency, and Its Rules

Phrase Structure and Constituency are fundamental concepts in syntactic analysis within the field of linguistics. Phrase Structure refers to the hierarchical arrangement of words and phrases within a sentence. This hierarchical structure helps to determine the relationships between different elements of a sentence, such as subjects, verbs, objects, and modifiers. The most common way to visually represent phrase structure is through tree diagrams, which illustrate the hierarchical relationships between the different elements of a sentence. Constituency refers to the grouping of words and phrases together to form larger units known as constituents (Chomsky, 1957). These constituents can be hierarchical in nature, with smaller constituents combining to form larger ones. For example, in the sentence "The cat chased the mouse," the phrase "the cat" is a constituent that forms part of the larger constituent "chased the mouse."

There are several rules that govern constituency and phrase structure. Some of the most common rules include:

- *The Phrase Structure Rule.* This rule states that each phrase in a sentence has a specific structure that can be represented as a tree diagram. For example, a simple sentence may be structured as $S \rightarrow NP VP$, where S is the sentence, NP is the noun phrase, and VP is the verb phrase.
- *Head-Phrase Rule.* This rule states that each phrase has a head word that determines the category of the phrase. For example, in a noun phrase like "the big red apple," the head word is "apple," which determines that it is a noun phrase.
- *Complement Rule.* This rule states that certain phrases may require complements to complete their meaning. For example, a verb phrase like "saw the movie" requires a direct object complement ("the movie") to make the sentence grammatically complete.
- *Adjunction Rule.* This rule states that additional modifiers can be added to a phrase through adjunction. For example, in the sentence "She ate the delicious cake quickly," the adverbs "delicious" and "quickly" are adjuncts that modify the verb phrase "ate the cake."

By understanding the concepts of Phrase Structure and Constituency, linguists are able to analyze and describe the structure of sentences in a systematic and structured way. These concepts are essential in syntactic analysis and help to uncover the underlying rules and patterns that govern the ways in which language is structured.

Phrase Structure is a key concept in generative grammar, specifically in the context of syntax. It refers to the hierarchical structure of a sentence or phrase, breaking it down into smaller constituents or elements. Constituency is the idea that certain words or groups of words form a unit or constituent within a sentence. These units can be further divided into smaller units, creating a hierarchy of structure within a sentence. The rules of Phrase Structure dictate how constituents are combined to form a grammatically correct sentence (Chomsky, 1957). These rules determine the order and organization of elements within a sentence, ensuring that it adheres to the syntax of a particular language. Some common rules include the Subject-Verb-Object (SVO) order, which is typical in English sentences, as well as rules for creating subordinate clauses and coordinating multiple phrases within a sentence.

Syntactic analysis involves the study of the structure of sentences, focusing on how words and phrases are combined to convey meaning. This analysis is key to understanding the rules and patterns of a language's grammar. Dependency parsing, on the other hand, is a method of syntactic analysis that focuses on the relationships between words in a sentence. Instead of using a hierarchical structure like Phrase Structure, dependency parsing represents the connections between words as a network of dependencies. *Noam Chomsky*, a prominent linguist, is credited with developing the theory of Generative Grammar, which includes Phrase Structure rules as a central concept (Kübler, McDonald & Nivre, 2009). Chomsky's work has had a significant influence on the field of syntax and has shaped our understanding of how language is structured and produced. Dependency parsing has also gained attention in recent years, with researchers like *Marie-Catherine de Marneffe* and others contributing to its development and application in natural language processing tasks.

So, both Phrase Structure and dependency parsing are important tools in syntactic analysis, offering different perspectives on how sentences are structured and how meaning is conveyed through language. By studying these concepts and applying them to real-world language data, linguists and researchers can gain valuable insights into the underlying mechanisms of syntax and language processing.

Syntactic analysis, the study of the structure and arrangement of words in sentences, has a long and complex history that stretches back to the ancient grammarians of Greece and Rome. However, it wasn't until the 20th century that significant advancements were made in this field, particularly with the contributions of linguists such as Noam Chomsky and Emily M. Bender. *Noam Chomsky* is perhaps the most influential figure in the history of syntactic analysis. In the 1950s and 1960s, Chomsky revolutionized the field with his theories of transformational generative grammar, which posited that all human languages share a common underlying structure and that these structures can be described using a formal system of rules. Chomsky's work laid the foundation for modern syntactic analysis, and his ideas continue to shape the field to this day (Chomsky, 1957).

One of Chomsky's most important contributions to syntactic analysis was the concept of universal grammar, which suggests that all human languages are governed by a set of innate rules and principles that are hardwired into the human brain. According to Chomsky, these rules are responsible for the generation of sentences in a given language and determine the limits of what is grammatically possible. This idea has had a profound impact on our understanding of language and has inspired decades of research into the nature of language and human cognition. Another important figure in the history of syntactic analysis is *Emily M. Bender*, a contemporary linguist known for her work on computational linguistics and natural language processing. Bender has made significant contributions to the field by developing new computational tools and algorithms for analyzing the syntactic structure of languages. Her work has helped to advance our understanding of how syntax operates in different languages and has opened up new avenues for research in the field.

Other pioneering linguists who have made important contributions to syntactic analysis include *Roman Jakobson*, who developed the theory of markedness and played a key role in the development of structuralist linguistics, and *Charles Fillmore*, who introduced the concept of frame semantics and revolutionized the study of lexical semantics. Also, the history of syntactic analysis is a rich and diverse one, filled with contributions from many different thinkers and researchers. From the groundbreaking work of Noam Chomsky to the innovative approaches of Emily M. Bender and others, the field has been shaped by a wide range of ideas and perspectives that continue to

inform our understanding of language and cognition (Marcus, Santorini & Marcinkiewicz, 1993).

Noam Chomsky's theory of phrase structure grammar is a foundational framework in the field of linguistics that seeks to describe the structure and organization of natural language. This theory is based on several key concepts and principles that have significantly influenced the study of language. Phrase structure rules: In Chomsky's theory, sentences are analyzed in terms of structured phrases (Chomsky, 1957). These phrases are constructed based on a set of phrase structure rules that specify how words can be combined to form larger units. The rules define the hierarchical relationships between different elements of a sentence and help generate grammatically correct sentences.

- *Recursive structures.* Chomsky's theory of phrase structure grammar allows for the generation of an infinite number of sentences through recursive structures. Recursive rules enable the repeated embedding of phrases within other phrases, leading to the creation of complex sentences with multiple levels of nesting.
- *Syntactic categories.* Chomsky's theory categorizes words into different syntactic categories based on their grammatical functions. These categories include nouns, verbs, adjectives, prepositions, conjunctions, and determiners. Each category plays a specific role in the structure of a sentence and has associated rules governing its use.
- *Transformational rules.* Chomsky introduced transformational rules as a way to account for the movement and transformation of linguistic elements within a sentence. These rules allow for the transformation of a sentence from one form to another while preserving its underlying meaning. For example, passive voice constructions and question formation can be described using transformational rules.
- *Deep structure and surface structure.* Chomsky's theory distinguishes between deep structure and surface structure in language. Deep structure represents the underlying meaning or semantics of a sentence, while surface structure refers to the actual arrangement of words and phrases in a sentence. Transformational rules are used to derive surface structure from deep structure.
- *Generative capacity.* One of the central principles of Chomsky's theory is the idea of generative capacity, which refers to the ability of a grammar to produce a

potentially infinite number of well-formed sentences. Phrase structure grammar is designed to be generative, allowing speakers to create new sentences that adhere to the rules of the language.

It may be noted, that Chomsky's theory of phrase structure grammar provides a formal framework for understanding the structure of sentences in natural language. By identifying key concepts such as phrase structure rules, syntactic categories, and transformational rules, Chomsky's theory has significantly contributed to the study of language and has shaped the way linguists analyze and describe linguistic structures.

Marie-Catherine de Marneffe is a prominent researcher in the field of natural language processing who has made significant contributions to both constituency and dependency parsing theories. Her work has advanced our understanding of how language is structured and how parsers can effectively analyze and model it. Marneffe's work on constituency parsing has focused on developing more accurate and efficient models for representing the syntactic structure of sentences. One of her key contributions in this area is the development of probabilistic models that incorporate lexical and syntactic information to predict the structure of a sentence. These models have been shown to outperform previous approaches by providing more accurate and robust analyses of the syntax of sentences (Green, 2011).

In addition to her work on constituency parsing, Marneffe has also made important contributions to dependency parsing. Dependency parsing is a type of syntactic analysis that represents the relationships between words in a sentence as directed links, with each link indicating a syntactic relationship between two words. Marneffe has developed novel algorithms and models for dependency parsing that improve the accuracy and efficiency of parsing systems. Her work has also explored the use of semantic information in dependency parsing, which has led to more comprehensive and accurate analyses of sentence structure.

We'd like to point out that, Marneffe's contributions to constituency and dependency parsing have greatly advanced the field of natural language processing. Her research has provided new insights into how parsers can effectively analyze and represent the syntax of sentences, leading to more accurate and robust parsing systems (Green, 2011). Her work continues to influence the development of parsing algorithms and models, shaping the way we understand and process language.

Phrase structure, constituents, and their rules are key aspects in the study of syntax, which is the branch of linguistics that deals with the structure of sentences. Over the years, various theories have been proposed by renowned linguists like Noam Chomsky, Daniel Jurafsky, Christopher Manning, and other pioneering authors from the USA. These theories have greatly influenced our understanding of how language is structured and processed by the human mind (Perlmutter, 1983).

Noam Chomsky's theory of Universal Grammar posits that all languages share a common underlying structure, and that this structure is innate to humans. According to Chomsky, there are certain rules and principles that govern the way in which words are combined to form meaningful sentences. These rules involve the arrangement of words into phrases and clauses, and the relationships between these elements. Chomsky also introduced the concept of phrase structure, which refers to the hierarchical organization of words within a sentence. Phrases are made up of one or more words, and can be further divided into smaller constituents (Zhang & Nivre, 2011). These constituents are the building blocks of sentences, and can be categorized as either heads or modifiers. *Daniel Jurafsky and Christopher Manning's* theory of Constituency Parsing builds upon Chomsky's ideas and aims to automatically identify the constituents of a sentence using computational methods. In this theory, constituents are defined as groups of words that function as a single unit within a sentence. These constituents can be identified based on their syntactic properties, such as their position in the sentence or their grammatical category.

So, the theories of Chomsky, Jurafsky, Manning, and other pioneering authors from the USA have greatly contributed to our understanding of phrase structure, constituent, and their rules. By studying the way in which words are combined to form meaningful sentences, linguists can gain insights into the cognitive processes involved in language production and comprehension. These theories continue to be a central focus of research in the field of syntax, and will likely shape our understanding of language for years to come.

2.2 Dependency Structure and Dependency Parsing Algorithm

Syntactic analysis is the process of analyzing the structure of sentences in natural language to determine the relationships between words and phrases. Two main approaches to syntactic analysis are phrase structure parsing and dependency parsing. Phrase structure parsing, also known as constituency parsing, is a bottom-up approach

to analyzing the syntactic structure of a sentence (Weiss, Stern & Goldberg, 2015). In phrase structure parsing, a sentence is parsed into a hierarchical structure of phrases and constituents, with each phrase representing a group of words that function together as a unit in the sentence. The main advantage of phrase structure parsing is that it provides a clear and intuitive representation of the syntactic structure of a sentence, making it easier to analyze and understand.

Dependency parsing, on the other hand, is a top-down approach to syntactic analysis that focuses on identifying the relationships between words in a sentence. In dependency parsing, the parsing algorithm identifies the head word in a sentence and creates a tree structure that represents the dependencies between the head word and its dependents. The main advantage of dependency parsing is that it provides a more direct and efficient representation of the syntactic relationships between words in a sentence, making it easier to analyze complex sentence structures (Zhang & Nivre, 2011). In terms of complexity, phrase structure parsing tends to be more complex than dependency parsing because it involves parsing a sentence into a hierarchical structure of phrases and constituents. This can be more computationally expensive and may require more sophisticated parsing algorithms to accurately analyze the syntactic structure of a sentence. Dependency parsing, on the other hand, is generally simpler and more efficient because it focuses on identifying the dependencies between words without the need to parse the sentence into a hierarchical structure.

Chomsky's transformational grammar is a prominent linguistic theory that revolutionized the field of linguistics in the mid-20th century. At the core of Chomsky's theory is the idea that language is governed by a universal grammar that underlies all human languages. This universal grammar consists of a set of rules and principles that guide how language is structured and generated in the human mind. One key aspect of Chomsky's transformational grammar is its approach to sentence structure. Chomsky proposed that sentences are not only made up of words strung together randomly but that they have a hierarchical structure with distinct levels of organization. This hierarchical structure is represented by phrase structure rules, which specify how words are combined to form phrases, and how phrases are combined to form sentences (Weiss, Stern & Goldberg, 2015).

In terms of dependency structure, Chomsky's transformational grammar focuses more on constituent structure – the relationships between words and phrases within a sentence

– rather than on explicit dependencies between individual words. According to Chomsky, sentences are built up from smaller units, or constituents, which are formed through the application of phrase structure rules. These rules specify the syntactic categories of words and their hierarchical relationships within a sentence. Chomsky's transformational grammar also introduced the concept of transformational rules, which allow for the transformation of one type of sentence structure into another. These rules operate on underlying structures generated by phrase structure rules and can account for variations in word order, passive voice constructions, and other syntactic transformations in language.

Marneffe's theory of dependency grammar is a structural approach to syntax that focuses on the relationships between words in a sentence. Unlike traditional phrase structure grammars, which analyze sentences in terms of phrases and hierarchical structures, dependency grammars represent sentences as directed graphs, where each word in the sentence is a node and the relationships between words are represented as directed edges. In Marneffe's theory, the main focus is on the syntactic relationships between words, which are called dependencies (Charniak, 2001). Dependencies are asymmetrical relationships between words that represent syntactic links between them. There are different types of dependencies in Marneffe's theory, including subject-verb, object-verb, modifier-head, and head-complement dependencies. Each dependency represents a specific syntactic relationship between the two words involved.

One key aspect of Marneffe's theory is the concept of head-first parsing, which means that the head of a phrase is processed before its dependents. This approach ensures that dependencies are correctly identified and processed in a way that reflects the syntactic structure of the sentence. Another important aspect of Marneffe's theory is the notion of valency, which refers to the number of arguments that a verb can take. Verbs with different valencies will have different dependency structures in a sentence (Green, 2011).

Chomsky and Marneffe both focus on dependency structures as a way to understand and analyze language, but they approach the concept in different ways. Noam Chomsky, a renowned linguist, is known for his theory of generative grammar, which proposes that all human languages share a universal grammar that is hardwired in our brains. Chomsky's theory of dependency structures is rooted in his idea of a hierarchical structure in language, where words are arranged in a specific order that determines their

relationships to one another. According to Chomsky, these relationships can be represented as directed arrows between words, with one word being dependent on another. This creates a tree-like structure where words are connected by branches that show their syntactic relationships.

On the other hand, Marie-Catherine de Marneffe, an influential linguist, focuses on the syntactic structure of language as a way to understand the relationships between words. Marneffe's theory of dependency structures, known as Universal Dependency (UD), is based on the idea that all languages can be analyzed using a consistent set of syntactic dependencies. Unlike Chomsky, Marneffe's approach to dependency structures is more data-driven, using large corpora of language data to identify universal patterns in language. Chomsky's theory emphasizes the innate nature of language structure, suggesting that all languages share the same underlying principles (Green, 2011). In contrast, Marneffe's theory is more empirically driven, looking at the patterns that emerge from natural language use across different languages.

Another difference between Chomsky and Marneffe's theories is their approach to analyzing dependency structures. Chomsky's theory is more theoretical, focusing on abstract syntactic relationships between words in a hierarchical structure. Marneffe's theory, on the other hand, is more practical, aiming to develop a universal set of syntactic dependencies that can be applied to analyze any language. In summary, Chomsky and Marneffe both offer valuable insights into the structure of language through their theories on dependency structures. While Chomsky's theory is more focused on the innate principles of language and abstract syntactic relationships, Marneffe's theory is more data-driven and aims to identify universal patterns in language using a universal set of syntactic dependencies. Both approaches contribute to our understanding of language and how it is structured.

Dependency analysis in literature is a critical lens through which scholars and critics examine the power dynamics, interconnections, and influences among various elements in a text. The United States has been home to many pioneering authors who have contributed significantly to the development of dependency analysis in literature. One of the most prominent figures in dependency analysis from the USA is Fredric Jameson. Jameson is a renowned literary critic and cultural theorist whose work has had a profound impact on the field of literary criticism (McDonald, Lerman & Pereira, 2006). His seminal book, "The Political Unconscious: Narrative as a Socially Symbolic Act,"

revolutionized the way scholars think about literature and its relationship to politics, culture, and society. Jameson's work is characterized by its rigorous analysis of power structures, ideology, and symbolism in literature, and his insights have influenced generations of scholars working in the field of literary theory.

Another pioneering author in dependency analysis from the USA is Sandra Gilbert. Gilbert is a leading feminist literary critic who has written extensively on the intersection of gender, power, and language in literature. Her groundbreaking book, "The Madwoman in the Attic: The Woman Writer and the Nineteenth-Century Literary Imagination," co-authored with Susan Gubar, explores the ways in which women writers have been marginalized and silenced in literary history. Gilbert's work challenges traditional literary canons and offers new perspectives on the relationships between authors, texts, and readers (McDonald, Lerman & Pereira, 2006).

Additionally, Edward Said is another influential figure in dependency analysis from the USA. Said was a Palestinian-American literary critic and political activist whose work on postcolonialism, Orientalism, and the representation of the "other" in literature has had a lasting impact on the field of literary studies. His seminal book, "Orientalism," critiques Western representations of the Middle East and challenges scholars to rethink their assumptions about power, identity, and difference in literature. Said's work continues to shape the way scholars think about colonialism, imperialism, and globalization in literature and culture.

Dependency analysis is a crucial aspect of linguistic research that involves the identification and analysis of grammatical relationships between words in a sentence. There are several algorithms used in linguistic research to perform dependency analysis, some of which are detailed below (Nivre, 2003):

- *Transition-based parsing algorithms.* Transition-based parsing algorithms are a class of dependency parsing algorithms that construct parse trees through a series of transitions. These algorithms move from an initial configuration of the input sentence to a final configuration representing the complete parse tree. The most common transition-based parsing algorithm is the arc-eager algorithm, which performs a left-to-right scan of the input sentence and builds the dependency tree incrementally by predicting the next transition based on the current state of the parse.

- *Graph-based parsing algorithms.* Graph-based parsing algorithms represent parse trees as graphs, where nodes represent words in the sentence and edges represent dependency relationships. These algorithms typically involve scoring each possible dependency arc in the tree and selecting the highest-scoring arcs to construct the final parse tree. One popular graph-based parsing algorithm is the MaltParser, which uses a second-order learning algorithm to construct dependency trees.
- *Recursive neural network models.* Recursive neural network models are deep learning models that can capture syntactic dependencies between words in a sentence. These models typically use recursive neural networks to encode the syntax of a sentence and predict the dependency relationships between words. Recursive neural network models have been shown to achieve state-of-the-art performance in dependency parsing tasks by capturing long-distance dependencies and complex syntactic structures.
- *Dependency pattern matching algorithms.* Dependency pattern matching algorithms identify syntactic patterns in a sentence by matching predefined dependency structures. These algorithms typically use hand-crafted rules or machine learning techniques to identify common patterns in dependency trees. For example, the Stanford Typed Dependencies system uses a set of predefined dependency types and rules to parse and annotate sentences.

So, dependency analysis algorithms play a crucial role in linguistic research by helping researchers understand the syntactic structure of sentences and extract useful information from text data. These algorithms vary in complexity and performance, making them suitable for different types of linguistic research tasks. Dependency analysis is a fundamental aspect of NLP that involves identifying the relationships between words in a sentence. This analysis plays a crucial role in various NLP tasks, such as parsing, machine translation, information extraction, and sentiment analysis. Some of the key applications of dependency analysis in NLP include:

- Dependency analysis is essential for syntactic parsing, which involves analyzing the structure of a sentence and determining the relationships between words. By identifying the dependencies between words, parsers can accurately analyze the grammatical structure of a sentence and generate a parse tree that represents the syntactic relationships between words.

- Dependency analysis is also important for machine translation systems, which translate text from one language to another. By understanding the dependencies between words in a sentence, machine translation systems can accurately capture the meaning of the source text and generate a more accurate translation. Dependency analysis helps ensure that the translated text is syntactically and semantically correct.
- Dependency analysis is used in information extraction tasks, where the goal is to automatically extract relevant information from text. By identifying the dependencies between words, information extraction systems can accurately extract entities, relationships, and events mentioned in the text. This information can then be used for various applications, such as data mining, knowledge graph construction, and question-answering.
- Dependency analysis is also applied in sentiment analysis, which involves analyzing the sentiments expressed in text. By identifying the dependencies between words in a sentence, sentiment analysis systems can accurately determine the sentiment polarity of the text, such as positive, negative, or neutral. Dependency analysis helps capture the relationships between sentiment-bearing words and their modifiers, resulting in a more accurate sentiment analysis.
- Dependency analysis is used in named entity recognition, which involves identifying and classifying named entities such as persons, organizations, locations, and dates in text. By understanding the dependencies between words, named entity recognition systems can accurately identify named entities and their relationships with other words in the text. This information is important for various NLP applications, such as information retrieval, knowledge graph construction, and entity linking.

Constituency and dependency parsing are two crucial techniques in linguistic research and natural language processing (NLP) applications. They play a significant role in understanding the structure and relationships within sentences and texts, which are essential for extracting meaning and information from language data. Constituency parsing involves breaking down a sentence into its grammatical components or phrases, such as noun phrases, verb phrases, and prepositional phrases (Nivre, 2003). This method helps identify the syntactic structure of a sentence and how its parts relate to

each other. Constituency parsing is important in linguistic research as it provides insights into the underlying grammar of a language and the rules that govern the formation of sentences. In NLP applications, constituency parsing is essential for tasks such as information extraction, sentiment analysis, and machine translation, as it helps in identifying and understanding the syntactic structure of texts (Toutanova & Manning, 2000).

Dependency parsing, on the other hand, focuses on the relationships between words in a sentence by identifying the dependencies or links between them. This involves determining which words are dependent on or governed by other words in the sentence. Dependency parsing is crucial for understanding the semantic relationships within sentences and how words interact with each other. In linguistic research, dependency parsing is used to study sentence structure and analyze language patterns. In NLP applications, dependency parsing is essential for tasks like named entity recognition, semantic role labeling, and question answering, as it helps in determining the relationships between words and extracting relevant information from texts (Toutanova & Manning, 2000). Both constituency and dependency parsing are important tools in linguistic research and NLP applications because they help in extracting meaning and information from language data. By analyzing the structure and relationships within sentences, these parsing techniques enable researchers and developers to better understand and process natural language text. Ultimately, constituency and dependency parsing contribute to improving the accuracy and effectiveness of NLP systems and facilitate various language-related tasks such as information retrieval, sentiment analysis, and machine translation (Handler & Zeman, 2018).

Syntactic analysis, also known as syntax, is the branch of linguistics that studies the rules for combining words into meaningful sentences. The development of syntactic analysis in the United States has been influenced by a number of major figures who have made significant contributions to the field. One of the most influential figures in the development of syntactic analysis in the United States is Noam Chomsky. Chomsky's groundbreaking work in the 1950s and 1960s laid the foundation for modern syntactic theory. In his seminal book "Syntactic Structures" (1957), Chomsky introduced the concept of transformational grammar, which posited that there are deep structures underlying surface structures in language. He also proposed the idea of a

universal grammar, a set of innate rules that govern the structure of all human languages.

Another major figure in the development of syntactic analysis in the United States is *Zellig Harris*. Harris was a leading proponent of linguistics, which focused on the analysis of language structure without reference to underlying mental processes. He developed the method of distributional analysis, which involves studying the distribution of linguistic elements in texts to uncover underlying syntactic patterns. Other important contributors to the development of syntactic analysis in the United States include Gerald Gazdar, who developed the theory of Generalized Phrase Structure Grammar (GPSG), and William Labov, who pioneered the study of sociolinguistics and the role of social factors in language structure.

Chomsky's generative grammar, first proposed in the 1950s, revolutionized the field of linguistics by introducing the concept of a generative grammar, which aims to explain the implicit knowledge speakers have about their language and how they are able to produce and understand an infinite number of sentences. Chomsky's theory was based on the idea that language is innate and that humans are born with a universal grammar that guides language acquisition (Chomsky,1959). Within Chomsky's generative grammar framework, the focus was mainly on phrase structure rules, which described the hierarchical structure of sentences and the relationships between words within a sentence. This led to the development of formal languages such as Chomsky's own transformational-generative grammar, which introduced the concept of deep structure and surface structure in language.

As research in linguistics progressed, alternative theories began to emerge that challenged the focus on phrase structure rules and hierarchical relationships in Chomsky's generative grammar. One such theory was dependency grammar, which posits that the relationship between words in a sentence can be described in terms of directed dependencies rather than hierarchical relationships. This shift in perspective led to the development of modern dependency parsing models, which aim to analyze the syntactic structure of sentences by representing the relationships between words as dependencies (Chomsky,1959).

Dependency parsing models have gained popularity in recent years due to their ability to capture syntactic phenomena such as long-distance dependencies and non-projective structures more effectively than traditional phrase structure-based models. These

models have been applied in various natural language processing tasks, such as part-of-speech tagging, parsing, and machine translation. One of the key advantages of dependency parsing models is their simplicity and transparency, as they directly capture the relationships between words in a sentence without the need for complex hierarchical structures. This makes them easier to interpret and analyze compared to traditional generative grammar models. Dependency structure and dependency analysis are crucial aspects in the field of linguistics, particularly in the study of syntactic structures and language processing. The dependency structure refers to the relationships between words in a sentence, where one word depends on another for its meaning or function. Dependency analysis, on the other hand, involves the identification and representation of these dependencies in a systematic manner (Bohnet, 2010).

Several prominent linguists have contributed to the development of theories and algorithms related to dependency structure and analysis. Noam Chomsky, a pioneering figure in the field of linguistics, introduced the notion of transformational grammar and generative grammar, which revolutionized the study of syntax. His work laid the foundation for modern theories of syntax and provided valuable insights into the ways in which words are combined to form coherent sentences. Similarly, Marie-Catherine de Marneffe, a linguist known for her research on dependency parsing, has made significant contributions to the field by developing algorithms for automatically analyzing the syntactic structures of sentences. Her work has helped researchers in achieving more accurate and efficient analyses of linguistic data, leading to advancements in natural language processing and machine translation.

In the United States, several other authors have also played a key role in shaping the field of dependency structure and analysis. Their research has focused on topics such as the representation of dependencies in parsing algorithms, the incorporation of semantic information into syntactic analyses, and the development of robust linguistic resources for various languages (Bohnet, 2010). So, the theories and algorithms proposed by Chomsky, Marneffe, and other pioneering authors from the USA have significantly advanced our understanding of how words are organized in sentences and how they depend on each other for meaning. Their work has paved the way for the development of innovative tools and methods for studying and analyzing language, and their contributions continue to shape the field of linguistics to this day.

2.3 Constituency-based and Dependency-based Treebanks

Treebanks are linguistic resources that annotate natural language texts with syntactic structures, allowing for detailed analysis and comparison of sentence structure. Two common approaches to representing syntax in treebanks are constituency-based and dependency-based annotations. Constituency-based treebanks, also known as phrase structure treebanks, are based on the principles of constituency grammar. In this approach, sentences are analyzed as hierarchically structured phrases, where larger constituents are made up of smaller constituents. The treebank consists of trees that represent the hierarchical structure of a sentence, with nodes representing constituents such as phrases and words. Each node in the tree represents a constituent, with the root node representing the entire sentence (Collins, 2003).

Dependency-based treebanks, on the other hand, are based on dependency grammar, which focuses on the relationships between words in a sentence rather than hierarchical phrase structures. In this approach, each word is linked to a head word that governs its syntactic role in the sentence. Dependency trees in these treebanks represent these relationships, with arrows indicating the direction of the relationship and the nature of the dependency (Bohnet, 2010). Dependency-based treebanks aim to capture the syntactic relations between words more directly and are often considered more efficient for certain linguistic analyses. Both constituency-based and dependency-based treebanks have their strengths and weaknesses. Constituency-based treebanks are well-suited for capturing hierarchical structures and are often easier to interpret for linguists trained in phrase structure rules. Dependency-based treebanks, on the other hand, are more focused on the relationships between words and can be more efficient for certain natural language processing tasks such as parsing and machine translation.

In practice, both constituency-based and dependency-based treebanks are widely used in linguistic research and natural language processing applications. They provide valuable resources for analyzing sentence structure, developing computational models of language, and training machine learning algorithms for various language tasks. The choice between constituency-based and dependency-based annotations depends on the specific research questions and application requirements.

Noam Chomsky's Generative Grammar is a theory of language structure that seeks to explain how native speakers are able to produce and understand an infinite number of sentences. The theory posits that language is governed by a universal grammar, which is

a set of principles and rules that are inherent to all human languages. According to Chomsky, this innate language faculty allows individuals to generate and comprehend complex sentences, even ones they have never heard before (Chomsky,1965). One of the key concepts in Generative Grammar is the idea of constituency, which refers to the hierarchical structure of sentences. In this framework, sentences are broken down into smaller units, or constituents, that combine to form larger phrases and sentences. For example, in the sentence "*The cat chased the mouse*," the constituents would be "*the cat*," "*chased*," and "*the mouse*." These constituents can be further broken down into smaller units, such as "the" and "cat."

Constituency-based treebanks are a tool used in computational linguistics to analyze and annotate sentences according to their constituency structure. In a constituency-based treebank, each word in a sentence is assigned a syntactic label, such as noun, verb, or adjective, and these labels are used to build a tree-like structure that represents the hierarchical relationships between the words. The treebank can then be used to analyze the syntax and grammar of a language, as well as to develop computational models for natural language processing tasks. By applying Generative Grammar principles to constituency-based treebanks, researchers can gain insights into the underlying structures of language and the rules that govern sentence formation. This approach allows linguists to study how different languages organize their syntax, as well as how speakers parse and interpret sentences. This can have important implications for language learning, artificial intelligence, and machine translation, as it enables researchers to develop more accurate and efficient language processing algorithms (Collins, 2003).

Marneffe's Dependency Grammar is a formalism for representing the grammatical structure of sentences in a language. It is based on the concept of dependency, which is a syntactic relationship between words in a sentence where one word (the head) governs or determines the grammatical properties of another word (the dependent). In a dependency-based approach, the structure of a sentence is represented as a set of binary relationships between words, where each word is connected to its head in a directed acyclic graph. In Marneffe's Dependency Grammar, the relationships between words are categorized into different types, such as subject, object, modifier, and complement, among others. These relationships capture the syntactic dependencies between words in

the sentence, providing a more fine-grained analysis of the grammatical structure compared to traditional constituency-based models.

One of the key advantages of dependency grammar is its ability to capture non-projectivity, or crossing dependencies, in a sentence. This means that words in a sentence can be related to each other in a non-linear fashion, which is not always easily captured in phrase structure-based grammars. Dependency-based treebanks are annotated corpora of sentences where each word is assigned a syntactic label indicating its grammatical role and its relationship to other words in the sentence. These treebanks serve as training data for computational systems that use dependency grammar for tasks such as parsing, machine translation, and information extraction.

Marneffe's Dependency Grammar and dependency-based treebanks have been widely used in natural language processing research, particularly in recent years with the development of more sophisticated parsing algorithms and machine learning techniques. Dependency grammar offers a simple and elegant representation of syntactic structure that is well-suited for computational models, making it a valuable tool for a wide range of natural language processing tasks.

Treebanks are annotated collections of syntactic structures derived from text corpora, which are used for natural language processing tasks such as parsing, machine translation, and information extraction. In the United States, the development of constituency-based and dependency-based treebanks has evolved over the years, each with its own set of challenges and advantages. Constituency-based treebanks, also known as phrase structure treebanks, represent the syntactic structure of a sentence using labeled brackets that indicate the hierarchical relationships between words. These treebanks have been in use since the 1980s, with some of the earliest examples being the Penn Treebank and the Wall Street Journal Treebank. The Penn Treebank, developed at the University of Pennsylvania, consists of parsed sentences from the Wall Street Journal and has been widely used as a benchmark for evaluating parsing algorithms.

Dependency-based treebanks, on the other hand, represent the syntactic structure of a sentence using directed graphs where each word is a node and the relationships between words are represented as edges. Dependency-based treebanks have gained popularity in recent years due to their simplicity and efficiency in capturing syntactic dependencies. One of the most well-known dependency treebanks in the United States is the Universal

Dependencies project, which provides parsed data in multiple languages for research and development purposes (Chomsky,1965). The evolution of constituency-based and dependency-based treebanks in the United States has been driven by advances in natural language processing research and the growing demand for high-quality annotated data. Convergence towards universal standards, such as the Universal Dependencies framework, has also played a role in facilitating interoperability between different treebanks and promoting collaboration among researchers and developers.

So, constituency-based and dependency-based treebanks have both made significant contributions to the field of natural language processing in the United States and beyond. While constituency-based treebanks have a long history and have been widely used in research, dependency-based treebanks are gaining popularity for their simplicity and efficiency. Both types of treebanks continue to evolve and improve, providing valuable resources for researchers, developers, and language technology applications (Collins & Duffy, 2002).

Constituency-based treebanks and dependency-based treebanks are two different approaches to annotating syntactic structures in a corpus. Each approach has its own strengths and weaknesses, and the choice of which type of treebank to use depends on the specific research goals and linguistic theories being employed. Constituency-based treebanks, also known as phrase structure treebanks, represent sentences as hierarchical structures consisting of phrases that are nested within each other. The tree structures are typically rooted at the sentence level and branch out into smaller phrases, such as noun phrases and verb phrases. Constituency-based treebanks are often used in formal linguistic theories, such as Generative Grammar, which analyze language as a set of rules governing the hierarchical structure of sentences.

On the other hand, dependency-based treebanks represent sentences as directed graphs where each word is a node and the relationships between words are labeled as dependencies. In dependency-based treebanks, the syntax of a sentence is represented by how words are connected to each other in a tree structure, with one word typically being the head or governor of another word. Dependency-based treebanks are often used in Dependency Grammar, which focuses on the relationships between words and the head-dependent structures that govern sentence syntax. In terms of accuracy and efficiency, dependency-based treebanks are often considered to be more precise and easier to annotate than constituency-based treebanks. Dependency structures are more

straightforward to parse and align with the underlying syntactic relationships in a sentence. However, constituency-based treebanks can provide more detailed information about the hierarchical structure of a sentence, which can be useful for certain types of linguistic analyses.

Another key difference between constituency-based and dependency-based treebanks is in how they handle coordination and subordination structures. Constituency-based treebanks typically represent coordination and subordination using complex tree structures, while dependency-based treebanks capture these relationships through the dependencies between words (Collins & Duffy, 2002). This difference in representation can impact the complexity and accuracy of parsing and annotation in each type of treebank. In conclusion, constituency-based and dependency-based treebanks offer different perspectives on the syntactic structures of language. The choice between the two types of treebanks depends on the specific research goals and theoretical frameworks being employed, as well as considerations of accuracy, efficiency, and the level of detail required in the annotation. Both types of treebanks have their own strengths and weaknesses, and researchers should carefully consider these factors when selecting a treebank for their linguistic analyses. Constituency-based and dependency-based treebanks are two common methods for syntactic parsing in natural language processing. These treebanks provide annotated representations of sentences' syntactic structures, which can be used for a variety of NLP tasks. Here are some of the key applications of constituency-based and dependency-based treebanks in NLP (Collins & Duffy, 2002):

- Constituency-based and dependency-based treebanks are used for parsing sentences to determine their syntactic structure. Parsing involves assigning a tree structure to a sentence that represents the relationships between words in the sentence. This can be used for tasks such as sentence understanding, machine translation, and information extraction.
- Treebanks can be used to extract useful information from text by identifying the relationships between words in a sentence. For example, treebanks can be used to extract named entities, relationships between entities, and other structured information from unstructured text.
- Treebanks are used in machine translation systems to help generate more accurate translations. By understanding the syntactic structure of the source

language and the target language, machine translation systems can produce more accurate translations.

- Treebanks can also be used in sentiment analysis to understand the syntactic structure of text and determine the sentiment expressed in a sentence. By analyzing the relationships between words in a sentence, sentiment analysis systems can determine whether a sentence is positive, negative, or neutral.
- Treebanks can be used in question answering systems to parse and analyze the structure of questions and find relevant information in a large corpus of text. By understanding the syntactic structure of questions, question answering systems can determine the most relevant answers.
- Treebanks can be used in automatic summarization systems to identify the most important information in a text and generate concise summaries. By understanding the syntactic structure of the text, automatic summarization systems can identify key concepts and relationships between words to produce accurate summaries.

Constituency-based and dependency-based treebanks are valuable resources in natural language processing for a wide range of tasks, including parsing, information extraction, machine translation, sentiment analysis, question answering, and automatic summarization. By providing annotated representations of syntactic structures, treebanks enable more accurate and efficient processing of natural language text. Constituency-based treebanks and dependency-based treebanks are tools used in natural language processing to annotate linguistic data. Annotation guidelines establish a standard for annotators to follow when creating these treebanks, ensuring consistency and accuracy in the annotation process. Here are some detailed annotation guidelines for both constituency-based and dependency-based treebanks:

Constituency-based Treebanks:

1. Phrase Structure:
 - Identify and label phrases in the sentence according to their syntactic categories (e.g., noun phrase, verb phrase).
 - Ensure that each phrase is properly nested within the larger syntactic structure of the sentence.
2. Tree Structure:

- Create a tree structure representing the syntactic relationships between words in the sentence.
 - Use brackets to enclose each phrase and indicate its syntactic category.
3. Head-directionality:
- Identify the head of each phrase and mark it with a specialized annotation symbol.
 - Pay attention to the directionality of modifiers (e.g., left-branching vs. right-branching).
4. Coordination:
- Properly annotate coordination structures, ensuring that all elements are grouped together and connected with appropriate conjunctions.

Dependency-based Treebanks:

1. Dependency Relations:
- Identify the relationships between words in the sentence, where one word serves as the head of another.
 - Use labeled arcs to represent these dependencies, indicating the type of grammatical relationship (e.g., subject, object).
2. Root Word:
- Designate a root word for each sentence, representing the main verb or the main noun phrase.
 - Ensure that all other words in the sentence are connected to the root word through a series of dependencies.
3. Directionality:
- Pay attention to the direction of dependencies, indicating whether they are incoming (modifiers) or outgoing (heads).
 - Follow a consistent directionality pattern throughout the treebank.
4. Multi-word Expressions:
- Properly annotate multi-word expressions (e.g., idiomatic phrases) as a single unit, maintaining their internal syntactic structure.
 - Use special annotations to indicate that certain words form a cohesive, non-constituent group.

Also, annotation guidelines for constituency-based and dependency-based treebanks should be detailed, precise, and consistent to ensure the creation of high-quality annotated data for natural language processing tasks. annotators should be well-trained and familiar with the guidelines to produce accurate and reliable annotated treebanks. Constituency-based and dependency-based treebanking are two popular methods used in natural language processing for annotating linguistic data. While both approaches have their own strengths and weaknesses, they also present several challenges and future directions for improvement.

Challenges in constituency-based treebanking:

- *Lack of consistency in annotation.* Constituency-based treebanking relies on annotating hierarchical structures in sentences, which can sometimes lead to discrepancies in how different annotators interpret the same sentence. This lack of consistency can affect the overall quality of the treebank.
- *Complexity of annotation.* Constituency-based treebanking involves annotating multiple layers of information in sentences, such as phrases, clauses, and sub-clauses. This complexity can make the annotation process time-consuming and labor-intensive.
- *Scale limitations.* Creating a constituency-based treebank for a large corpus of text can be resource-intensive, requiring a significant amount of manual annotation. This can limit the scale of the treebank and the variety of linguistic phenomena that can be captured.

Challenges in dependency-based treebanking:

- *Ambiguity in dependency relations.* Dependency-based treebanking relies on annotating the syntactic relationships between words in a sentence, which can sometimes be ambiguous. For example, a word may have multiple dependents or could be assigned different roles in the dependency structure.
- *Lack of standardization.* There is currently no standardized framework for dependency annotation, leading to variations in how different treebanks are annotated. This lack of standardization can make it difficult to compare and combine treebanks from different sources.
- *Maintenance and scalability.* Dependency-based treebanks require regular updates and revisions to keep up with changes in the language or the annotation

guidelines. This can pose challenges in terms of maintaining the treebank and scaling it up to include new linguistic phenomena.

Future directions in constituency-based and dependency-based treebanking:

- *Annotation guidelines.* Developing more detailed and standardized annotation guidelines for constituency-based and dependency-based treebanking can improve consistency and interoperability among different treebanks. This can help researchers compare and combine treebanks more effectively.
- *Automation and semi-automation:* Developing automated or semi-automated tools for annotation can help speed up the treebanking process and reduce the manual labor involved. This can enable the creation of larger and more comprehensive treebanks.
- *Cross-linguistic treebanking.* Expanding constituency-based and dependency-based treebanking to cover a wider range of languages can help improve the representation and analysis of linguistic diversity. This can involve developing language-specific annotation guidelines and tools for cross-linguistic comparison.

So, constituency-based and dependency-based treebanking play a crucial role in advancing natural language processing research. Overcoming the challenges and exploring new directions in treebanking can help enhance the quality and scope of linguistic data available for various computational tasks.

Constituency-based and dependency-based treebanks are two prominent approaches in computational linguistics for representing the syntactic structure of sentences. These two approaches have their roots in the theories of *Noam Chomsky*, *Ivan A. Sag*, and *Tom Wasow*, among others. Chomsky's transformational-generative grammar introduced the concept of constituent structure, where sentences are analyzed in terms of their hierarchical structure of phrases. This approach gave rise to constituency-based treebanks, where sentences are parsed into a tree structure with nodes representing constituents such as noun phrases, verb phrases, and clauses (Chomsky, 1965).

On the other hand, dependency-based treebanks stem from the theories of *Lucien Tesnière* and others, which focus on the relationships between words in a sentence rather than their hierarchical structure. In dependency-based treebanks, sentences are represented as a set of head-dependent relationships, where each word is connected to its governing word in the sentence.

Marneffe's theory of universal dependencies builds on the work of both Chomsky and Tesnière by seeking to create a universal set of dependency relations that can be applied across languages. This approach has been instrumental in the development of multilingual dependency-based treebanks, which allow for cross-linguistic comparison and analysis. The pioneering work of researchers from the USA, such as Chomsky, Marneffe, and their colleagues, has been fundamental in shaping the field of computational linguistics and the development of treebanks. Their theories have provided the theoretical underpinnings for the construction of syntactic annotation schemes and the development of parsing algorithms for natural language processing tasks (Chomsky, 1965).

In conclusion, both constituency-based and dependency-based treebanks have their own strengths and weaknesses, and researchers continue to explore the potential benefits of integrating these two approaches. The work of Chomsky and other pioneering authors from the USA has laid the foundation for these advancements in the field of computational linguistics, driving progress in syntactic analysis and natural language processing.

2.4 Applications of Treebanks in Computational Linguistics

Treebanks are collections of parsed sentences that represent the structure of a language. In computational linguistics, treebanks serve as valuable resources for various applications, such as natural language processing (NLP) and machine learning. One of the main applications of treebanks is in the development and evaluation of natural language processing tools, such as parsers and part-of-speech taggers. By providing annotated data, treebanks enable researchers to train and test the accuracy of these tools on real-world language data. This helps improve the performance of NLP systems for tasks such as text classification, information extraction, and machine translation.

Treebanks also play a crucial role in linguistic research, enabling scientists to study the syntax and semantics of a language. By analyzing the tree structures generated from treebanks, researchers can uncover patterns and rules that govern the way language is structured. This can lead to new insights in areas such as syntax, semantics, and discourse analysis. Furthermore, treebanks are used in machine learning algorithms, such as supervised and unsupervised learning techniques. By leveraging the annotated data in treebanks, machine learning models can be trained to perform a variety of

language processing tasks, such as sentiment analysis, named entity recognition, and syntactic parsing.

In addition, treebanks are essential for the development of language resources, such as lexicons, dictionaries, and ontologies. These resources rely on accurate linguistic annotations provided in treebanks to enhance the quality and coverage of language data. Also, treebanks are indispensable tools in computational linguistics, providing a rich source of linguistic data that can be utilized for a wide range of applications. Their influence extends beyond academia, with industries such as information retrieval, text mining, and chatbot development benefiting from the insights gained from annotated language data (Zhang & Zhang, 2015).

Treebanks serve as a valuable resource for linguistic analysis within the generative grammar framework established by Noam Chomsky. Generative grammar is a theoretical framework that posits that language is a rule-governed system, and that the structure of sentences can be represented as hierarchical structures. Chomsky's generative grammar approach emphasizes the study of syntax, or the rules that govern the structure of sentences, as a key component of understanding language. In the context of generative grammar, Treebanks are corpora of parsed sentences that represent the syntactic structure of sentences in a hierarchical manner, typically using tree structures. Each node in the tree represents a word in the sentence, with branches indicating the relationships between words and their syntactic categories. By analyzing and annotating text with these tree structures, linguists and computational linguists can gain insights into the underlying syntactic rules that govern the formation of sentences (Chomsky, 1957).

Treebanks provide a systematic and structured way of capturing the syntactic properties of language, allowing researchers to analyze and compare patterns across different languages and text types. In the generative grammar framework, Treebanks can be used to test and refine linguistic theories, as well as to build computational models that can automatically generate or analyze sentences based on syntactic rules. Chomsky's work has been influential in shaping the field of computational linguistics, and Treebanks play a crucial role in applying his theoretical insights to practical tasks such as parsing, language generation, and machine translation (Zhang & Zhang, 2015). By utilizing Treebanks within the generative grammar framework, researchers can gain a deeper

understanding of the structure of language and develop more sophisticated models for natural language processing tasks.

Treebanks are a fundamental resource in natural language processing and machine learning, providing annotated linguistic data that can be used for a variety of tasks, such as parsing, language modeling, and sentiment analysis. Essentially, a treebank is a collection of sentences that have been manually annotated with syntactic tree structures, representing the grammatical structure of the sentences. These annotated tree structures can be used to train and evaluate parsing algorithms, which automatically analyze the grammatical structure of a sentence and assign a tree structure to it. By comparing the output of parsing algorithms with the annotated tree structures in a treebank, researchers can evaluate the accuracy of their algorithms and make improvements to their models.

Treebanks are typically created by linguists or trained annotators who manually annotate each word in a sentence with its syntactic role, such as subject, verb, object, etc. The annotated words are then connected by labeled arrows to represent the relationships between them. This process results in a tree structure that captures the hierarchical relationships between words in a sentence. One of the most well-known treebanks is the *Penn Treebank*, which contains over 4.5 million words of English text annotated with syntactic tree structures. The Penn Treebank has been widely used in research projects in computational linguistics and natural language processing, and has been instrumental in the development of parsing algorithms and other language processing tools. In addition to the *Penn Treebank*, there are treebanks available for many other languages, including *Chinese*, *Arabic*, and *German*. These treebanks play a crucial role in the development of natural language processing tools for a diverse range of languages and are an essential resource for researchers working in the field.

Treebanks are collections of parsed sentences in which each word is assigned a syntactic label based on its grammatical role in the sentence. These treebanks are used for syntactic and semantic analysis to help linguists and computational scientists better understand the structure and meaning of language. One approach to syntactic analysis using treebanks is Dependency Grammar, a theory that posits that the structure of a sentence can be represented as a set of relationships between individual words. The theory of *Joakim Nivre and Joel Tetreault* takes this approach a step further by incorporating ideas from Head-Driven Phrase Structure Grammar (HPSG) and Lexical

Functional Grammar (LFG) to create a unified theory of dependency that captures both syntactic and semantic information (Green, 2011).

In this theory, each word is assigned a head, which is typically the main verb in the sentence. The head governs other words in the sentence, forming a dependency structure that reflects the relationships between words. For example, the word "dog" in the sentence "The dog chased the cat" would be dependent on the word "chased," as it is the subject of the verb. By using treebanks annotated with these dependency structures, linguists and computational scientists can analyze the relationships between words in a sentence and uncover patterns that reveal underlying syntactic and semantic principles. This analysis can help improve natural language processing tasks such as machine translation, information extraction, and sentiment analysis.

In addition to syntactic analysis, treebanks can also be used for semantic analysis, which seeks to understand the meaning of words and how they combine to create meaningful expressions. The theory of *Eliyahu Kiperwasser and Yoav Goldberg* extends the dependency-based approach to incorporate semantic information, allowing for a deeper analysis of the relationships between words in a sentence. By annotating treebanks with both syntactic and semantic information, researchers can explore the interactions between form and meaning in language, shedding light on how words are structured and interpreted in context. This integrated approach to syntactic and semantic analysis using treebanks offers a rich resource for linguists and computational scientists to deepen their understanding of language and improve the performance of natural language processing systems.

HPSG was developed by Carl Pollard and Ivan Sag in the 1980s as a theory of grammar that describes the structure of sentences in terms of hierarchical phrase structures and the relations between those phrases. In HPSG, each word in a sentence is associated with a set of features that determine its syntactic and semantic properties. The theory is "head-driven" in that it posits that the head of a phrase determines the structure and behavior of the whole phrase. HPSG also incorporates a feature-based formalism, which allows for a more fine-grained analysis of linguistic phenomena.

LFG, on the other hand, was developed by Joan Bresnan and Ronald Kaplan in the 1980s as a theory of grammar that emphasizes the role of lexical entries in determining syntactic structure. In LFG, each word in a sentence is associated with a set of lexical and functional information that determines its grammatical properties. The theory is

"lexicalist" in that it posits that the lexicon plays a crucial role in determining syntactic structure, and that syntactic phenomena can be accounted for by relating the structures of lexical entries. LFG also incorporates a constraint-based formalism, which allows for a more flexible and expressive analysis of linguistic phenomena (Green, 2011). Both HPSG and LFG have been influential in the field of syntax and have been applied to a wide range of languages. Despite their differences in emphasis and formalisms, both theories share a commitment to a lexicalist approach to grammar and have contributed to our understanding of the complex relationships between words in a sentence.

In conclusion, the applications of treebanks in computational linguistics have been instrumental in advancing the understanding of language structure and processing. Through the creation and analysis of annotated tree structures, researchers have been able to develop and refine linguistic theories proposed by influential figures such as *Noam Chomsky*, *Christopher D. Manning*, and other pioneering authors from the USA. Chomsky's transformational-generative grammar theory laid the foundation for modern computational linguistics by emphasizing the importance of formal grammatical structures in understanding language. Treebanks have provided empirical evidence for Chomsky's theories by allowing researchers to analyze the syntactic and semantic patterns in large corpora of text.

Manning and his colleagues have further expanded on Chomsky's work by developing probabilistic models and machine learning algorithms that can automatically annotate and parse text, leading to more accurate and efficient natural language processing systems. Treebanks have played a crucial role in training and evaluating these systems, enabling researchers to test the validity of linguistic theories and improve the performance of computational models (Nivre, Hall & Nilsson, 2006). Overall, the integration of treebanks in computational linguistics has not only advanced our understanding of language processing but has also facilitated the development of practical applications such as machine translation, information retrieval, and sentiment analysis. By building on the theories of Chomsky, Marneffe, and other influential authors, researchers continue to push the boundaries of linguistic research and technological innovation in the field of computational linguistics.

Deduction on Chapter 2

The chapter on syntactic analysis and dependency parsing provides a comprehensive understanding of the foundational theories and methodologies in linguistics, focusing on

the contributions of prominent figures like Chomsky and Marneffe. Chomsky's transformational-generative grammar and Marneffe's dependency-based approach emphasize the importance of understanding the relationships between words in a sentence and how these relationships can be represented through syntactic structures.

The chapter also discusses the challenges and debates within the field, such as the tension between formalism and empiricism, and the ongoing efforts to develop more accurate and efficient parsing algorithms. Syntax is a crucial aspect of linguistics that studies the structure and rules governing sentence composition in natural language. In the United States, several pioneering authors have made significant contributions to the field of syntax, influencing the development of linguistic theories and methodologies.

Noam Chomsky, Ray Jackendoff, Barbara Partee, Joan Bresnan, and Laura Michaelis have all made significant contributions to the field of syntax. Their work has significantly influenced the Universal Dependencies project, which aims to create a unified framework for annotating and analyzing the syntactic structure of languages. By drawing on the insights and theories developed by these authors, the project seeks to establish a common framework for syntactic analysis that can be applied to a wide range of languages.

In conclusion, the contributions of pioneering authors from the USA in the field of syntax have had a profound impact on the development of linguistic theories and methodologies, shaping our understanding of syntax and language structure.

CHAPTER 3. DEPENDENCY RELATIONS IN THE KYRGYZ LANGUAGE BASED ON UNIVERSAL DEPENDENCIES TAGSETS

3.1 Parts-of-Speech-tags and Tools Used in Universal Dependencies

The global collaboration known as **Universal Dependencies**, sometimes shortened to UD, aims to build treebanks of all languages spoken on Earth. These treebanks are freely available and accessible. Fundamental uses include automated text processing in the domain of natural language processing (NLP) and natural language syntax and grammar research, particularly in the context of linguistic typology. While allowing for language-specific additions as needed, the project's main goal is to establish cross-linguistic standardization in annotation. The three linked initiatives that gave rise to the annotation method are the Stanford Dependencies, Google Universal Part-of-Speech Tags, and the Intersect Interlingua for Morphosyntactic Tagsets. Unlike phrase structure trees, the UD annotation technique employs a representation in the form of dependency trees. Unfortunately, at present, the Kyrgyz language has just entered the list of accepted languages of the Universal Dependencies web project, and the first steps are being taken in this area. Therefore, in this work, we will discuss the examples we have prepared as proposals for syntactic analysis models of the Kyrgyz language for the above-mentioned web project, and we hope to carry out the actions of introducing those sentences to the site later.

Sentences are syntactically analyzed using the UD annotation scheme according to the dependencies of dependency grammar. A label on the dependency edge indicates the syntactic function that each dependence is described as a result of. The syntactic annotation scheme of UD uses numbered nodes to represent words and punctuation marks, with each node dependent on a governor and one dependent on a fake ROOT node. Let us see one example of annotation from UD in Kyrgyz language:

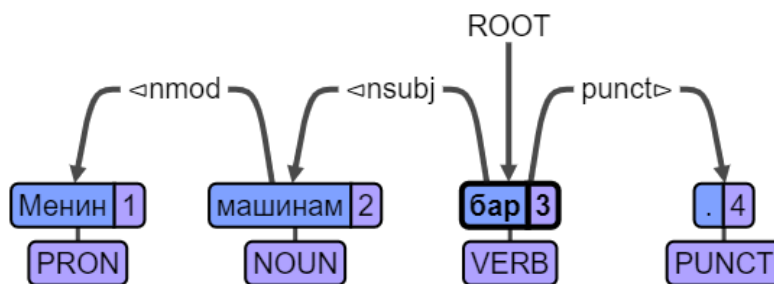


Figure 3. 1. A picture of a dependency structure in a simple sentence.

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Менин	Мен	PRON			2	nmod	2:nmod	
2	машинам	машина	NOUN			3	nsbj	3:nsbj	
3	бар	бар	ADJ			0	root	0:root	
4	.	.	PUNCT			3	punct	3:punct	

Table 3.1. A table representation of a simple sentence.

sent_id = 001

text = Менин машинам бар.

1	Менин	Мен	PRON	PRP	Case=Gen Number=Sing Person=1 PronType=Prs				
	2	nmod	_	_					
2	машинам	машина	NOUN	NNN	Case=Gen Number=Sing Number[psor]=Sing Person[psor]=1	3	nsbj		
	_	_							
3	бар	бар	VERB	VB	Aspect=Perf Mood=Ind Polarity=Pos Tense=Pres				
	0	root	_		SpaceAfter=No				
4	.	.	PUNCT	PCT	_	3	punct	_	_

This example shows that there are 4 components in this sentence which are called **nodes** (nodes are usually numbered) and each of them is connected in a dependency relation. *Менин* is the first node and is **dependent** on node #2 *машинам* (NOUN) which is a **governor** of the dependent node. Node #2 is dependent on node #3. The third node is *бар* (VERB) and it is the **root** of the sentence. The root is considered to be the head of the whole sentence. Additionally, there is a fake node ROOT which is numbered 0, the

unseen intentional node. The numbers in the sentence are arranged in order as 1, 2, 3, 4, and so on. Those relations can be described as mother-daughter/father-son relations.

There is a huge importance in the annotation techniques and tools in UD to apply one dependency structure. There are special and unified UD POS tags used for all the languages in the world. Let us first see tags that mark the core part-of-speech categories in Table:

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

Table 3.2. Core POS categories in UD

Besides the core POS tags, there are relation tags that are marked to show the nodes' dependency. These tags are frequently updated according to specific language features.

Below in Figure 2 we can observe those tags:

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	Headless	Loose	Special	Other
conj cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

Figure 3. 2. 37 universal syntactic relations used in UD

The table lists core, non-core and nominal dependents, which contain nominal, clauses, modifier words and function words. Additionally, this figure shows other parts of the dependency parsing, such as coordination, loose and special relations.

Languages define specific relations as subtypes of universal types, starting with the basic type and a colon. Subtypes are language-specific and optional, but some are semi-mandatory if the language focuses on a particular phenomenon. If language requires, those additional tags should be used, but they are not mandatory.

3.2 Dependency parsing techniques and dependency relations in the Kyrgyz language

Generating meaning and information from textual content is crucial for dependency parsing, and this is valuable for many language processing applications. It makes it easier to standardize a language to a worldwide framework and conduct scientific research on that language. As computerized sentence parsing and annotation become more common, linguists need to adapt and seize these new technical opportunities. These days, automated sentence parsing plays a major role in Natural Language Processing (NLP) and Natural Language Understanding (NLU).

These procedures help to digitalize the language so that computers can read and comprehend it easily. For the Kyrgyz language to continue developing and being preserved, academics' efforts in this area are vital. Regretfully, the lack, absence, or variation in the syntax of the Kyrgyz language made it impossible to identify any of the 37 dependent ties from UD that were indicated. Nonetheless, we will attempt to demonstrate in this work how syntactic treebanks are created and how they may be enhanced further. We attempt to analyze through treebanks these syntactic relations and depict dependency treebanks as a guideline for scholars who are interested in Kyrgyz language research.

acl: clausal modifier

acl is a clausal modifier of the noun (adnominal clause). The acl relation is used for finite and non-finite clauses that modify nominal or nominal phrases, unlike the **advcl** relation for adverbial clauses that modify predicates. See an example of acl in the Kyrgyz sentence:

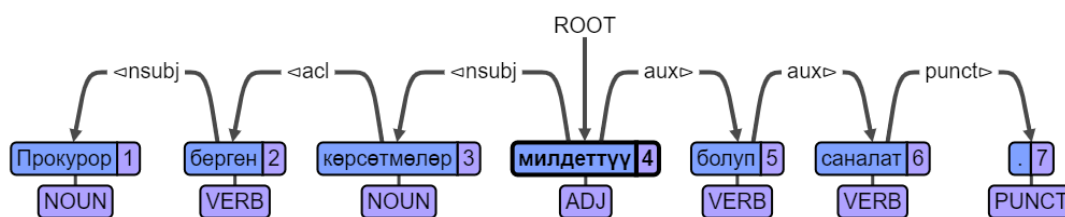


Figure 3. 3. Example of acl relation in the sentence

“*Прокурор берген көрсөтмөлөр милдеттүү болуп саналат*”. Its translation - “The testimony given by the prosecutor is considered to be mandatory.”

"<Прокурор>"

"прокурор" NOUN @nsbj #1->2

"<берген>"

"бер" VERB @acl #2->3

"<көрсөтмөлөр>"

"көрсөтмө" NOUN @nsbj #3->4

"<милдеттүү>"

"милдеттүү" ADJ @root #4->0

"<болуп>"

"бол" VERB @aux #5->4

"<саналат>"

"сана" VERB @aux #6->5

"<.>"

"." PUNCT @punct #7->6

In this sentence, “Прокурор берген” is considered to be the clausal modifier of the noun of the sentence “көрсөтмөлөр”. The testimony is given by the prosecutor, i.e. what kind of testimony? It modifies the noun and plays the role of an adjective (*testimony given by the prosecutor*).

acl:relcl is an extra feature of acl in a number of languages. The acl:relcl connection links the changed nominal to the relative clause head. It is a relative clause modifier that changes a noun that is coreferential with a component inside the relative clause. Multiple languages have several types of relative clauses: finite and non-finite. For instance, non-finite sentences in English are not regarded as relative, although this difference might not be made in other languages.

advcl: adverbial clause

Adverbial clause modifiers (**advcl**) alter verbs or predicates with a clausal dependent as the primary predicate, such as temporal, consequence, conditional, or purpose clauses. A nominal is never modified by Advcl; on the other hand, a clausal predicate that is realized as a nominal may occasionally be modified, but only in part. It needs to distinguish between changing the nominal entity and the whole matrix clause.

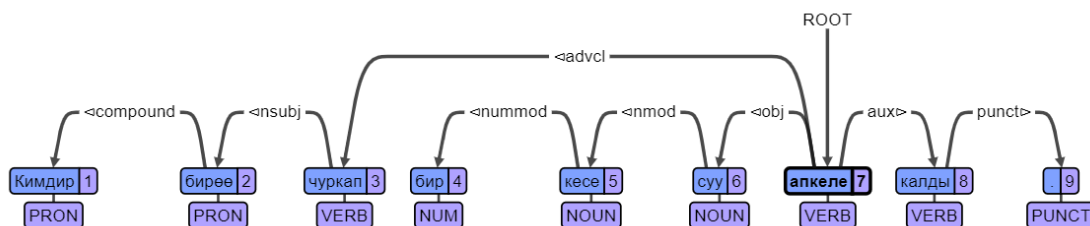


Figure 3. 4. Example of advcl relation in the sentence.

“Кимдир бирөө чуркап бир кесе суу апкеле калды.” Its translations – “Someone brought a cup of water, running.” Here, in this sentence, we observe the **advcl** relation, where the “Кимдир бирөө чуркап” is an adverbial modifier of the predicate. *How* was the action completed? The answer is by running someone brought a cup of water. Usually, advcl can be easily identified by asking the simple question: *How was the given action completed?* Now, let us analyze the sentence by its CG-3 format showing its dependency (Toktonaliev & Tokoev, 2015).

"<Кимдир>": PRON (pronoun) @compound (compound word) #1->2

"<бирөө>": PRON @nsubj (nominal subject) #2->3

"<чуркап>": VERB @advcl (adverbial clause modifier) #3->7

"<бир>": NUM (numeral) @nummod (numeric modifier) #4->5

"<кесе>": NOUN @nmod (nominal modifier) #5->6

"<суу>": NOUN @obj (object) #6->7

"<апкеле>": VERB @root (root) #7->0

"<калды>": VERB @aux (auxiliary) #8->7

"<.>": PUNCT (punctuation) @punct (punctuation) #9->8

"<Кимдир>": PRON (pronoun) @nmod (nominal modifier) #1->2

This annotation indicates that "<Кимдир>" is a pronoun and serves as a nominal modifier related to the word at position 2 in the sentence. "<бирөө>" is a pronoun and functions as the nominal subject related to the word at position 3 in the sentence.

"<чуркан>" is a verb and acts as an adverbial clause modifier related to the word at position 7 in the sentence. "<бир>" is a numeral and serves as a numeric modifier related to the word at position 5 in the sentence. "<кече>" is a noun and functions as a nominal modifier related to the word at position 6 in the sentence. "<сүү>" is a noun and acts as the object related to the word at position 7 in the sentence. "<апкеле>" is a verb and serves as the root of the sentence. "<калды>" is a verb and functions as an auxiliary related to the word at position 7 in the sentence. "<.>" is a punctuation mark and acts as punctuation related to the word at position 8 in the sentence. These annotations provide a structured representation of the **advcl** (adverbial clause modifier) in the given sentence. Additionally, there is a special tag **advcl:relel** for a relative clause that modifies a clause. These types of sentences exist in German, English, Japanese, Italian, and some others, examples of treebanks are given in the UD page.

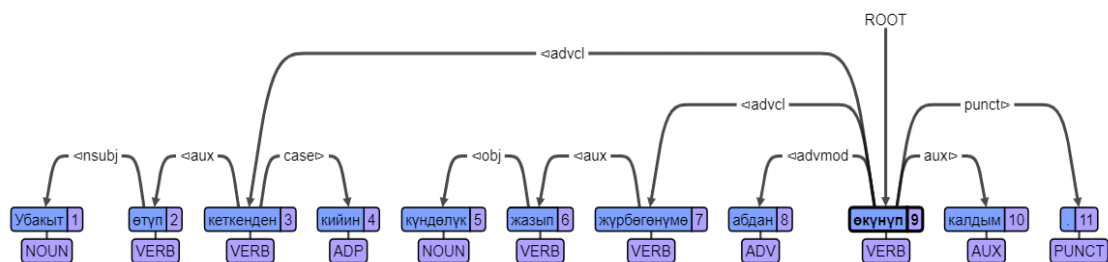


Figure 3. 5. Example 2 of advcl relation in the sentence.

“Убакыт өтүп кеткенден кийин күндөлүк жазып жүрбөгөнүмө абдан өкүнүп калдым.” Its translation: “I really regret not having a diary *after some time passed*.” Here, we are able to observe another **advcl** relation in the Kyrgyz sentence. When do I regret or how do I regret not having a diary? “Убакыт өтүп кеткенден кийин” is advcl of the sentence, meaning (when?) *after some time passed*. It is a whole adverbial clause within another clause with its own subject and predicate relation, forming advcl dependency.

"<УБАКЫТ>"

"убакыт" NOUN @nsbj #1->2

"<ӨТҮП>"

"өт" VERB @aux #2->3

"<КЕТКЕНДЕН>"

"кеткен" VERB @advcl #3->9

"<КИЙИН>"

"кийин" ADP @case #4->3
 "<күндөлүк>"
 "күндөлүк" NOUN @obj #5->6
 "<жазып>"
 "жаз" VERB @aux #6->7
 "<жүрбөгөнүмө>"
 "жүр" VERB @advcl #7->9
 "<абдан>"
 "абдан" ADV @advmod #8->9
 "<өкүнүп>"
 "өкүн" VERB @root #9->0
 "<калдым>"
 "кал" AUX @aux #10->9
 "<.>"
 " ." PUNCT @punct #11->9

In the sentence “Убакыт өтүп кеткенден кийин күндөлүк жазып жүрбөгөнүмө абдан өкүнүп калдым.” the dependency relations can be explained as follows:

"убакыт" NOUN @nsubj #1->2: The word "убакыт" (time) is the subject of the adverbial clause.

"өтүп" VERB @aux #2->3: The word "өт" (pass) is an auxiliary verb linked to the main verb at position 3.

"кеткенден" VERB @advcl #3->9: The word "кеткен" (leave) is a verb that introduces an adverbial clause linked to the main verb at position 9.

"кийин" ADP @case #4->3: The word "кийин" (after) is a postposition and is linked to the word at position 3.

"күндөлүк" NOUN @obj #5->6: The word "күндөлүк" (diary) is the object of the verb at position 6.

"жазып" VERB @aux #6->7: The word "жаз" (write) is a main verb linked to the auxiliary verb at position 7.

"жүрбөгөнүмө" VERB @advcl #7->9: The word "жүр" (go) is an auxiliary verb that introduces an adverbial clause linked to the main verb at position 9.

"абдан" ADV @advmod #8->9: The word "абдан" (very) is an adverb that modifies the verb at position 9.

"өкүнүп" VERB @root #9->0: The word "өкүн" (regret) is the main verb of the sentence and serves as the root.

"калдым" VERB @aux #10->9: The word "кал" (stay) is an auxiliary verb that is linked to the main verb "өкүнүп", which is a root of the sentence.

"." PUNCT @punct #11->9: The period serves as punctuation and is linked to the main verb at position 9, indicating the end of the sentence.

advmod: adverbial modifier

A (non-clausal) adverb or adverbial phrase that modifies a predicate or a modifier word is called an adverbial modifier (**advmod**). In many linguistic practices, adverbial modifiers serve the same purpose as adverbs, whether they are realized by nouns, adverbs, or adpositional phrases. There is no distinction made between modifiers of predicates and modifier terms like adverbs or adjectives. Adverbial modifiers are essential in Universal Dependencies (UD) because they change predicates or other modifier terms such as adjectives or adverbs. They clarify issues like "How? When? Where? " and offer further details on the action, description, or modifier in a phrase.

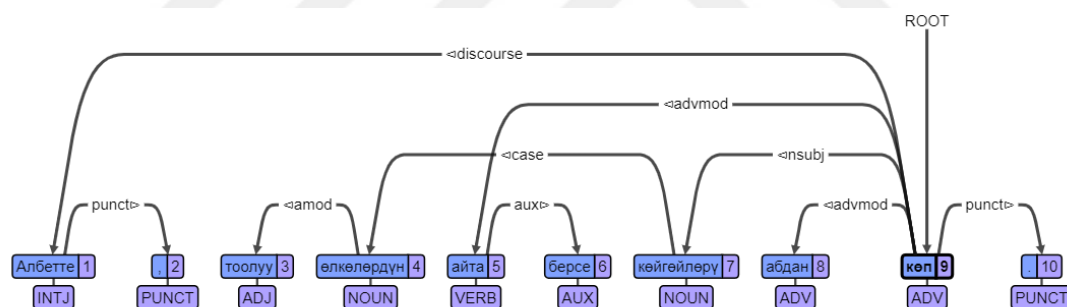


Figure 3. 6. Example of advmod in a sentence.

“Албетте, тоолуу өлкөлөрдүн айта берсе көйгөйлөрү абдан көп”. Its translation: “Of course, there are still (if say) a lot of problems in mountainous countries.” In this sentence “көп” is the root and predicate, and it is modified by the word “абдан” and the phrase “айта берсе”, meaning there are *still a lot of* problems. “Problems” (көп) is a predicate of the sentence and these two advmod relations emphasize its context (Nuruev, 2005).

"<Албетте>"

"Албетте" INTJ @discourse #1->9

"<, >"

"," PUNCT @punct #2->1
 "<тоолуу>"
 "тоолуу" ADJ @amod #3->4
 "<өлкөлөрдүн>"
 "өлкөлөрдүн" NOUN @case #4->7
 "<айта>"
 "айта" VERB @advmod #5->9
 "<берсе>"
 "берсе" AUX @aux #6->5
 "<көйгөйлөрү>"
 "көйгөйлөрү" NOUN @nsubj #7->9
 "<абдан>"
 "абдан" ADV @advmod #8->9
 "<көп>"
 "көп" ADV @root #9->0
 "<.>"
 "." PUNCT @punct #10->9

The Kyrgyz sentence “Албетте, тоолуу өлкөлөрдүн айта берсе көйгөйлөрү абдан көп.” has following relations:

"Албетте" INTJ @advmod #1->9: The word "Албетте" (Of course) is an interjection that modifies the entire sentence.

"," PUNCT @punct #2->1: The comma serves as punctuation and is linked to the interjection at position 1.

"тоолуу" ADJ @amod #3->4: The word "тоолуу" (mountainous) is an adjective that modifies the noun at position 4.

"өлкөлөрдүн" NOUN @case #4->7: The word "өлкөлөрдүн" (countries) is a noun in the genitive case, linked to the noun at position 7.

"айта" VERB @advmod #5->9: The word "айта" (say) is a verb that modifies the entire sentence.

"берсе" AUX @aux #6->5: The word "берсе" (if) is an auxiliary verb linked to the verb at position 5.

"көйгөйлөрү" NOUN @nsubj #7->9: The word "көйгөйлөрү" (problems) is the subject of the sentence.

"абдан" ADV @advmod #8->9: The word "абдан" (a lot of) is an adverb that modifies the verb at position 9.

"көп" ADV @root #9->0: The word "көп" (many) is the main verb of the sentence and serves as the root.

"." PUNCT @punct #10->9: The period serves as punctuation and is linked to the main verb at position 9, indicating the end of the sentence.

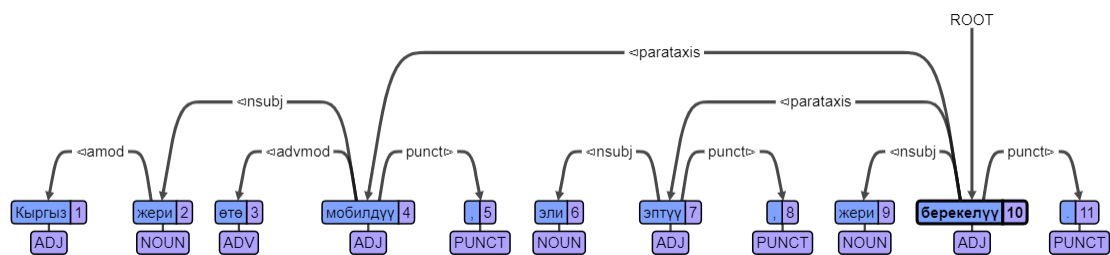


Figure 3. 7. Example 2 of advmod in a sentence.

“Кыргыз жери өтө мобилдүү, эли энтүү, жери берекелүү.” Its translation: “The Kyrgyz land is very mobile, the people are active, the land is abundant.”

"<Кыргыз>"

"кыргыз" ADJ @amod #1->2

"<жери>"

"жер" NOUN @nsubj #2->4

"<өтө>"

"өтө" ADV @advmod #3->4

"<мобилдүү>"

"мобилдүү" ADJ @conj #4->10

"<,>"

"," PUNCT @punct #5->4

"<эли>"

"эл" NOUN @nsubj #6->7

"<энтүү>"

"энтүү" ADJ @conj #7->10

"<,>"

"," PUNCT @punct #8->7
 "<жери>"
 "жер" NOUN @nsubj #9->10
 "<берекелүү>"
 "берекелүү" ADJ @root #10->0
 "<.>"
 "." PUNCT @punct #11->10

In the provided dependency parsing for a sentence in the Kyrgyz language, the words are annotated with their respective parts of speech and dependency relations. Here's a description of the parsing in a sentence:

The sentence consists of several words, including adjectives, nouns, adverbs, and punctuation marks, each with specific dependency relations. The word "кыргыз" (Kyrgyz) is an adjective modifying the noun "жери" (place). The word "эптүү" (active) is an adjective modifying the noun "эли" (people). Additionally, the adverb "өтө" (very) modifies the adjective "мобилдүү" (mobile), indicating the manner of an adjective – very mobile. The punctuation marks "," and "." serve to separate the different elements within the sentence. The adverb "өтө" in the Kyrgyz language is used to modify adjectives or predicates by indicating the manner in which the action is performed. It serves to intensify the meaning of the adjective or the predicate, conveying the idea of "mobile" in a strong or emphatic manner. This **adverb** plays a crucial role in providing additional detail and emphasis to the action described by the verb, enhancing this expression (Toktonaliev, 2015).

Adverbial modifiers can also have additional relationships, such locative adverbial modifier (**advmod:lmod**) and emphasizing word (**advmod:emph**). In the analytical layer of PDT, **advmod:emph** is a special class of words that often modify noun phrases, including prepositional phrases. For instance, the word "specifically" in the English phrase "specifically large" below has the function of **advmod:emph**, which modifies the adjective "large".

A subclass of the **advmod** relation is the locative adverbial modifier (**advmod:lmod**); if the modifier specifies a place, it is designated as **lmod**. For example, the place "жерге" in the sentence "Ал жерге отурду." (He sat down) is labeled as **advmod:lmod**.

amod: adjectival modifier

Any adjectival phrase that functions to modify a noun (or pronoun) is known as an adjectival modifier (**amod**). The relationship holds true regardless of how the noun's meaning is altered—for example, through composition (big house) or idiom (hot dogs). Although amod dependent can have its own modifiers (such as "very large house"), it shouldn't function as a clause. If it is a clause, then the correct tag is **acl**.

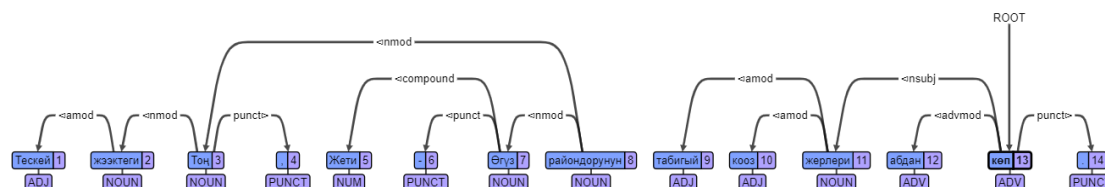


Figure 3. 8. Example of amod relation in a sentence.

“Тескей жээктеги Тоң, Жети-Өгүз райондорунун табигый кооз жерлери абдан көп.” Its translation: “Ton, Jeti-Oguz districts on the Northern coast have a lot of natural beautiful spots”. In the provided dependency relation sentence, 3 adjectives are used to modify nouns and convey specific attributes. “Тескей” (Northern) is an adjective that modifies the noun “жээктеги” (coast), “табигый” and “кооз” are adjectives that modify the noun “жерлери” (spots). Furthermore, there is a CG-3 format of this sentence, containing its detailed dependency parsing.

"<Тескей>"

"Тескей" ADJ @amod #1->2

"<жээктеги>"

"жээк" NOUN @nmod #2->3

"<Тоң>"

"Тоң" NOUN @nmod #3->8

"<,>"

"," PUNCT @punct #4->3

"<Жети>"

"Жети" NUM @compound #5->7

"<->"

"-" PUNCT @punct #6->7

"<Өгүз>"

"Өгүз" NOUN @nmod #7->8

"<райондорунун>"

"район" NOUN #8->
 "<табигый>"
 "табигый" ADJ @amod #9->11
 "<кооз>"
 "кооз" ADJ @amod #10->11
 "<жерлери>"
 "жер" NOUN @nsubj #11->13
 "<абдан>"
 "абдан" ADV @advmod #12->13
 "<көп>"
 "көп" ADV @root #13->0
 "<.>"
 "." PUNCT @punct #14->13

The adjectives in this context play a crucial role in providing detailed descriptions and characteristics of the nouns they modify, enriching the whole expression and conveying specific attributes of the nominal mentioned in the sentence. Adjectives in dependency syntax are represented by the "**amod**" relation, which indicates the dependency between the adjective and the noun or pronoun it modifies. This relation helps in defining the role of adjectives in modifying nouns or pronouns within the syntactic structure. Adjectives can have their own modifiers, providing additional detail.

Adjective modifiers, like adjectives and adjectival phrases, enhance descriptive information by providing detailed descriptions of nouns, improving language processing and comprehension. They have 4 major functions in a sentence: enhancing descriptive information, clarifying relationships, improving precision, and avoiding ambiguity. They clarify relationships between nouns and associated attributes, establishing specific characteristics and distinctions. Adjective modifiers improve language precision by adding depth and clarity to elements they modify, enabling accurate communication and conveying specific details. They also help avoid ambiguity by providing clear and unambiguous descriptions of nouns, ensuring accurate meaning is conveyed without confusion or misinterpretation (Imanov, 2009).

appos: appositional modifier

A nominal that comes just after the initial noun and functions to *define, alter, name, or characterize* that noun is called an appositional modifier. In one of these structures, it defines abbreviations and provides examples in parenthesis. Appos is a grammatical construction used between two nominals, which can be switched in modulo punctuation. Similar constructions exist for titles with less than a full nominal, like state senator Paul Mnuchin. Some grammatical traditions call these "fixed (or close) appositions" and use the name as the head. If two nominals go together and can be swapped, and the meaning isn't changed, then these nominals are appositional modifiers (Universal Dependencies).

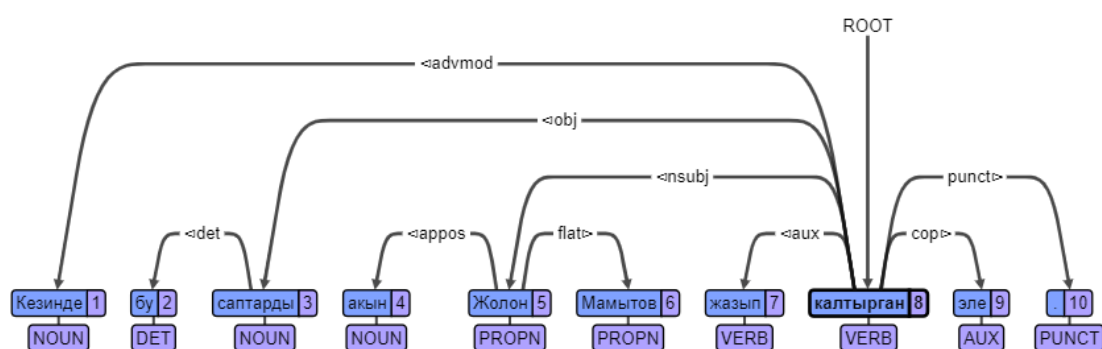


Figure 3.9. Example of appos relation in a sentence.

“Кезинде бу саптарды акын Жолон Мамытов жазып калтырган эле.”. The translation of the following Kyrgyz sentence into English: “Once these lines (poems) were left written down by the poet Jolon Mamytov.” Here, “акын” is an appositional modifier that comes just before the initial proper name “Жолон Мамытов”, clarifying it, *who is Jolon Mamytov in real life*. He is a poet, i.e. “акын” plays a role of **appos** relation, altering the main nominal (subject) of the sentence (Akunova et al., 2010).

"<Кезинде>"

"кез" NOUN @advmod #1->8

"<бу>"

"бу" DET @det #2->3

"<саптарды>"

"сап" NOUN @obj #3->8

"<акын>"

"акын" NOUN @appos #4->5

"<Жолон>"

"Жолон" PROPN @nsubj #5->8
 "<Мамытов>"
 "Мамытов" PROPN @flat #6->5
 "<жазып>"
 "жаз" VERB @aux #7->8
 "<калтырган>"
 "калтыр" VERB @root #8->0
 "<эле>"
 "эле" AUX @cop #9->8
 "<.>"
 " ." PUNCT @punct #10->8

In the Kyrgyz sentence “Кезинде бу саптарды акын Жолон Мамытов жазып калтырган эле.”, the dependency relations can be described as follows:

"кез" NOUN @advmod #1->8: The word "кезинде" (once/time) is a noun and functions as an adverbial modifier in the sentence, relating to the word at position 8.

"бу" DET @det #2->3: The word "бу" (these) is a determiner and serves as the determiner of the word at position 3.

"сап" NOUN @obj #3->8: The word "сап" (lines) is a noun and functions as the object of the verb at position 8.

"акын" NOUN @appos #4->5: The word "акын" (poet) is a noun and serves as an apposition to the word at position 5.

"Жолон" PROPN @nsubj #5->8: The word "Жолон" (Jolon) is a proper noun and functions as the subject of the verb at position 8.

"Мамытов" PROPN @flat #6->5: The word "Мамытов" is a proper noun and is linked as a flat dependency to the word at position 5.

"жаз" VERB @aux #7->8: The word "жаз" (write) is a verb and functions as an auxiliary to the verb at position 8.

"калтыр" VERB @root #8->0: The word "калтыр" (leave) is a verb and serves as the root of the sentence.

"эле" AUX @cop #9->8: The word "эле" (cop) is an auxiliary and functions as the copula in relation to the word at position 8.

" ." PUNCT @punct #10->8: The period serves as punctuation and is linked to the word at position 8.

aux: auxiliary

An **aux** (auxiliary) of a phrase is a function word that is connected to a verbal predicate and denotes features like as tense, mood, aspect, voice, or evidentiality. Nonverbal TAME indicators are found in many languages and are also regarded as instances of aux, even though they are often verbs (and may also have non-auxiliary purposes). Although those auxiliaries that are needed to construct the passive voice are now also classed as aux, UD guidelines strongly advise using the subtype aux:pass in languages with a grammaticalized (periphrastic) passive. Sometimes called auxiliary verbs, helper verbs are crucial for conveying grammatical distinctions that the main verb is unable to. They create questions, negatives, and passive structures. In a verb phrase, auxiliary verbs follow the root verb in a dependent relationship and offer distinct grammatical variations. Copulas can also be included in this list as they are pure linking words for nonverbal predication. Auxiliary verbs, or function words, are designated as AUX in dependency relations and are terms that support the lexical verb in a verb phrase. They indicate grammatical differences like person, number, tense, mood, aspect, voice, or evidentiality that are not conveyed by the lexical verb. It is worth noting that many languages do not include grammaticalized auxiliaries, and in those that do, there may be differences in the distinction between auxiliaries and complete verbs (Lyukina, 2018). All things considered, auxiliary verbs serve an important function for conveying complicated grammatical notions, producing a variety of grammatical tenses, and greatly influencing the structure and meaning of sentences. The future tense in English is called the analytical tense because it is constructed with the use of an auxiliary verb. In this case, the auxiliary word "will" is used to show that the action—writing—is done in the future rather than being translated. Auxiliary verbs in English can be classified as either primary or modal.

Main auxiliary verbs in English are:

- *be*
- *do*
- *have*

Modal auxiliary verbs in English are:

- *can*
- *could*
- *dare*

- *may*
- *might*
- *must*
- *should*
- *will*
- *would*

However, in the Kyrgyz language **aux** are mostly varied and play a huge role in sentence structure. The tenses in Kyrgyz express an action that is being performed at the moment with another action. For instance, in the Present tenses, auxiliary verbs are common:

- Мен окуп жатам. I'm reading (now).
- Алар карап турат. They're looking (now).

The main auxiliary verbs in the Kyrgyz are:

жат lie

жүр go

тур stand

отур sit

The auxiliary verbs *тур* and *отур* are archaic; they are used in this role much less frequently than *жат* and *жүр*. The main (not auxiliary, semantic) verb when forming the present tense is in the form of a gerund with *-(ы)п* ending.

There is a specific tag **aux:pass** for passive auxiliaries in English, such as in the sentence “*This book will be read*”, it is more sufficient to tag *be* as *aux:pass*.

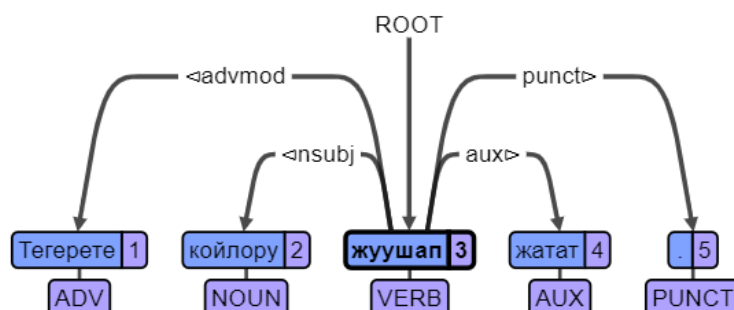


Figure 3. 10. Example of aux relation in a sentence.

“Тегерете койлору жуушан жатат.” Its translation: “The sheep are being herded together.”

"<Тегерете>"

"тегерете" ADV @advmod #1->3

"<койлору>"

"кой" NOUN @nsubj #2->3

"<жуушан>"

"жууша" VERB @root #3->0

"<жатат>"

"жат" AUX @aux #4->3

"<.>"

"." PUNCT @punct #5->3

In the sentence “Тегерете койлору жуушан жатат.”, the syntactic relations can be explained as follows:

"тегерете" ADV @advmod #1->3: The word "тегерете" (round) is an adverb that modifies the verb at position 3, indicating the manner in which the action is performed.

"кой" NOUN @nsubj #2->3: The word "кой" (sheep) is a noun and functions as the subject of the verb at position 3, indicating the doer of the action.

"жууша" VERB @root #3->0: The word "жууша" (to herd) is the main verb of the sentence and serves as the root of the sentence, expressing the action being performed.

"жат" AUX @aux #4->3: The word "жат" (lie) is an auxiliary verb and is linked to the main verb at position 3, indicating the tense or aspect of the action.

"." PUNCT @punct #5->3: The period serves as punctuation and is linked to the word at position 3, indicating the end of the sentence.

This breakdown provides a clearer understanding of how auxiliary verbs function in sentences providing a detailed description of the action’s tense, mode, and aspect.

case: case marking

A case is applied to case-marking components that are regarded to be separate grammatical words, such as prepositions, postpositions, and clitic case markers. These elements are considered dependent on the nouns to which they are linked or introduced. The function of the case relation is to provide a consistent analysis of nominal components, prepositions, and cases in morphologically rich languages such that

nominals in oblique situations acquire the same dependency structure as nominals introduced by adpositions. Case-marking components are regarded as dependents of the noun or sentence they attach to or introduce in universal dependencies. This method seeks to provide a more consistent understanding of nominal elements, prepositions, and case in morphologically rich languages by doing away with the idea that a preposition acts as a link between a changed word and its object (Djumaliev et al., 2023).

Case marking in the Kyrgyz language is usually accomplished by agglutination; in this scenario, the case-marked word will have the relevant **case** suffix. Postpositions and some case-like suffixes that are tokenized as distinct syntactic tokens (inflectional groups) are marked by the case relation. For postpositions derived from nouns with possessive suffixes (*кументердин, кыздардын*), **nmod:poss** relation is used. In this case, the head is the derived postposition.

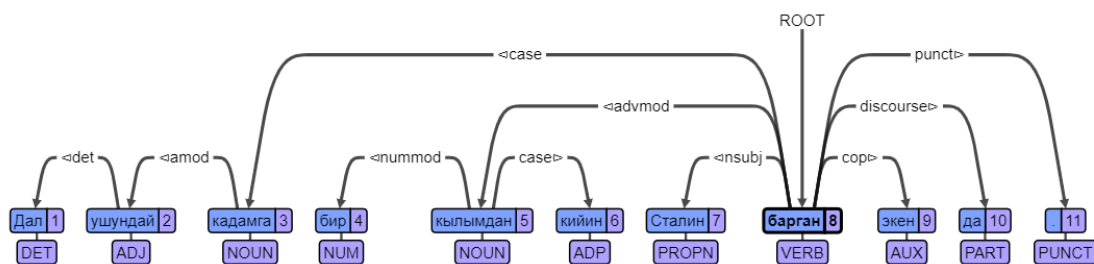


Figure 3. 11. Example of case marking in a sentence.

“Дал ушундай кадамга бир кылымдан кийин Сталин барган экен да.” The translation of this sentence into English: “Stalin took exactly the same step a century later.”

"<Дал>"

"Дал" DET @det #1->2

"<ушундай>"

"ушундай" ADJ @amod #2->3

"<кадамга>"

"кадамга" NOUN @case #3->8

"<бир>"

"бир" NUM @nummod #4->5

"<кылымдан>"

"кылымдан" NOUN @advmod #5->8
 "<кийин>"
 "кийин" ADP @case #6->5
 "<Сталин>"
 "Сталин" PROPN @nsubj #7->8
 "<барган>"
 "барган" VERB @root #8->0
 "<экен>"
 "экен" AUX @cop #9->8
 "<да>"
 "да" PART @discourse #10->8
 "<.>"
 "." PUNCT @punct #11->8

The sentence *"Дал ушундай кадамга бир кылымдан кийин Сталин барган экен да."* can be analyzed as follows:

"Дал" DET @det #1->2: The word "Дал" (the) functions as a determiner modifying the adjective at position 2.

"ушундай" ADJ @amod #2->3: The word "ушундай" (similar/same) is an adjective that modifies the noun at position 3.

"кадамга" NOUN @case #3->8: The word "кадамга" (step) is a noun in the **dative case**, linked to the verb at position 8.

"бир" NUM @nummod #4->5: The word "бир" (one) is a numeral modifier linked to the noun at position 5.

"кылымдан" NOUN @advmod #5->8: The word "кылымдан" (century) is a noun serving as an adverbial modifier linked to the verb at position 8. It is in the **ablative case**.

"кийин" ADP @case #6->5: The word "кийин" (after) is a postposition that indicates the relationship between the noun at position 5 and the following words.

"Сталин" PROPN @nsubj #7->8: The word "Сталин" (Stalin) is a proper noun serving as the subject of the sentence.

"барган" VERB @root #8->0: The word "барган" (go) is the main verb of the sentence and serves as the root.

"экен" AUX @cop #9->8: The word "экен" is a copula verb linked to the main verb at position 8.

"да" PART @discourse #10->8: The word "да" is a particle indicating the completion of the sentence.

"." PUNCT @punct #11->8: The period serves as punctuation and is linked to the main verb at position 8, indicating the end of the sentence.

cc: coordinating conjunction

A **cc** is the relation between a conjunct and an associated coordinating conjunction. A sentence's opening may also contain a coordinating conjunction. In linguistic dependency structure, coordinating conjunctions play a crucial role in connecting two elements without subordinating one to the other. Despite the fact that the phrase does not have numerous conjuncts connected with a conj connection, this is equally tagged as **cc**. If “менен, жана, да, бирок, деген, үчүн, анда, же” and other conjunction words are connected with nominal, they are tagged as **cc**. Let us see examples of coordinating conjunctions in the Kyrgyz language. Coordinating conjunctions are joined to the conjunct that comes just after in a dependency structure. This indicates that the coordinating conjunction is associated with the element that comes after it, as opposed to the element that comes before. With this method, sentences headed by conjunctions can be treated in parallel while maintaining the conjunction's dependence on the sentence's root and attachment to the conjunct that comes just after it. An example of coordinating conjunctions in dependency structure is: *"The book is delightful and intriguing,"* where "delightful" and "intriguing" are connected by the coordinating conjunction "and," and the conjunction is attached to the immediately succeeding conjunct. In some languages there is cc:preconj: preconjunct that shows the relationship between the head of coordination and the word at the beginning of a multi-word coordinating conjunction, such as *either ... or, neither ... nor, both ... and*.

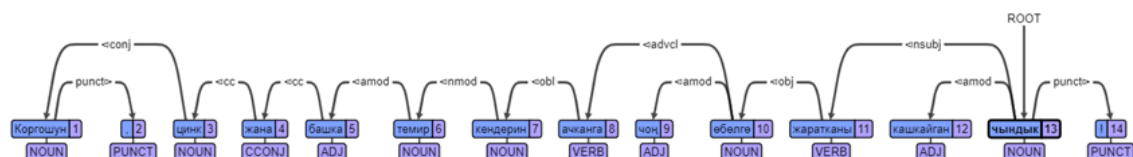


Figure 3. 12. Example of cc relation in a sentence.

“Коргошун, цинк жана башка темир кеңдерин ачканга чоң өбөлгө жаратканы кашкайган чындык!” and its translation into English: “It is a sad fact that lead, zinc, and other irons greatly contributed to the discovery!”

"<Коргошун>"

"Коргошун" NOUN @conj #1->3

"<,>"

"," PUNCT @punct #2->1

"<цинк>"

"цинк" NOUN @cc #3->4

"<жана>"

"жана" CCONJ @cc #4->5

"<башка>"

"башка" ADJ @amod #5->6

"<темир>"

"темир" NOUN @nmod #6->7

"<кендерин>"

"кендерин" NOUN @obl #7->8

"<ачканга>"

"ачканга" VERB @advcl #8->10

"<чоң>"

"чоң" ADJ @amod #9->10

"<өбөлгө>"

"өбөлгө" NOUN @obj #10->11

"<жаратканы>"

"жаратканы" VERB @nsubj #11->13

"<кашкайган>"

"кашкайган" ADJ @amod #12->13

"<чындык>"

"чындык" NOUN @root #13->0

"<!>"

!" PUNCT @punct #14->13

The Kyrgyz sentence "*Коргошун, цинк жана башка темир кендерин ачканга чоң өбөлгө жаратканы кашкайган чындык!*" can be analyzed as follows:

"Коргошун" NOUN @conj #1->3: The word "Коргошун" (a lead) functions as a noun and is linked to the following word as a conjunction.

"," PUNCT @punct #2->1: The comma serves as punctuation and is linked to the noun "Коргошун" at position 1.

"цинк" NOUN @cc #3->4: The word "цинк" (zinc) is a noun and is coordinated with the following word.

"жана" CCONJ @cc #4->5: The word "жана" (and) is a coordinating conjunction linking the noun "цинк" and the word "башка".

"башка" ADJ @amod #5->6: The word "башка" (other) is an adjective modifying the noun at position 6.

"темир" NOUN @nmod #6->7: The word "темир" (metal) is a noun functioning as a nominal modifier linked to the noun at position 7.

"кендерин" NOUN @obl #7->8: The word "кендерин" (mines) is a noun serving as an oblique nominal dependent.

"ачканга" VERB @advcl #8->10: The word "ачканга" (open) is a verb serving as an adverbial clause modifier linked to the root of the sentence.

"чоң" ADJ @amod #9->10: The word "чоң" (big) is an adjective modifying the noun at position 10.

"өбөлгө" NOUN @obj #10->11: The word "өбөлгө" (pillar, foundation) is a noun serving as the direct object of the verb at position 11.

"жаратканы" VERB @nsubj #11->13: The word "жаратканы" (creat) is a verb serving as the subject of the sentence.

"кашкайган" ADJ @amod #12->13: The word "кашкайган" (clear) is an adjective modifying the noun at position 13.

"чындык" NOUN @root #13->0: The word "чындык" (truth) is a noun serving as the root of the sentence.

!" PUNCT @punct #14->13: The exclamation mark serves as punctuation and is linked to the noun "чындык" at position 13, indicating the end of the sentence.

ccomp: clausal complement

A dependent phrase that serves as the central argument of a verb or adjective is called a clausal complement (**ccomp**). In other words, it acts as an adjective or verb's object. When the subject is in control and the higher subject or object has no additional interpretation that would be classified as xcomp, clausal complements—whether finite or nonfinite—are suitable. Only sentences headed by (non-finite) verbal nouns in accusative or nominative cases are right for the usage of ccomp. Additional clausal arguments should be annotated as advcl connection. In most languages, reported speech phrases often apply the ccomp connection.

However, sentences in UD site are not fully annotated yet. Therefore, we would try to annotate them ourselves and see how clausal complements function in the sentences.

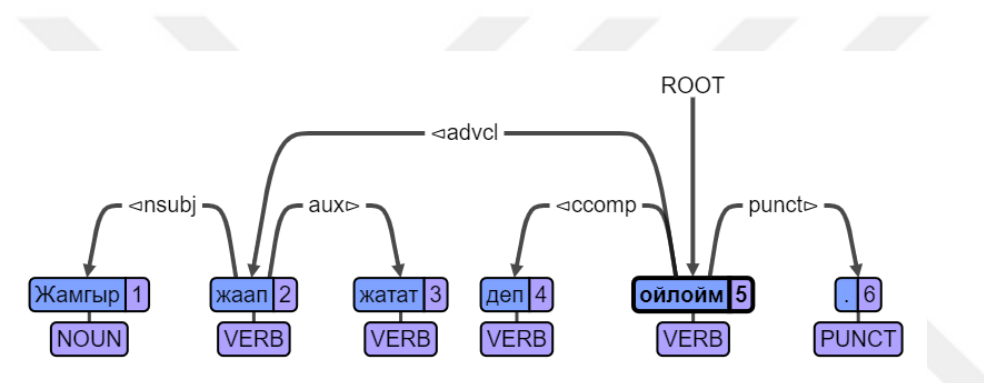


Figure 3. 13. Example of ccomp in the sentence.

“Жамгыр жаап жатат деп ойлойм.” And here is its translation into English: “I think it's raining.”

"<Жамгыр>"

"Жамгыр" NOUN @nsbj #1->2

"<жаап>"

"жаа" VERB @advcl #2->5

"<жатат>"

"жат" VERB @aux #3->2

"<деп>"

"де" VERB @ccomp #4->5

"<ойлойм>"

"ойло" VERB @root #5->0

"<.>"

"." PUNCT @punct #6->5

The Kyrgyz sentence "*Жамгыр жаап жатат деп ойлойм*" can be analyzed as follows:

"Жамгыр" NOUN @nsubj #1->2: The word "Жамгыр" (rain) functions as a noun and is linked to the following word as the subject of the adverbial clause.

"жаап" VERB @advcl #2->5: The word "жаап" (to rain) is a verb serving as an adverbial clause modifier linked to the following words.

"жатат" VERB @aux #3->2: The word "жатат" (to lie) is a verb serving as an auxiliary to the verb at position 2.

"деп" VERB @ccomp #4->5: The word "деп" (say that) is a verb serving as a clausal complement linked to the verb at position 5.

"ойлойм" VERB @root #5->0: The word "ойлойм" (think) is a verb serving as the root of the sentence.

"." PUNCT @punct #6->5: The period serves as punctuation and is linked to the verb "ойлойм" at position 5, indicating the end of the sentence.

compound

Along with fixed and flat relations, there are three relations for multiword expressions (MWEs): the **compound** relation. It may be applied to many other kinds of compounding, such as particle verbs, serial verbs, verb and adjective compounds, and noun compounds (like "real estate"). The syntactic relationship between the elements of these multiword sentences is depicted by the compound relation. Dependency structures are used to annotate compounds to illustrate the relationship between the compound's component words. The compound connection for "video games developer", for example, captures the link between "video" and "games" as well as the relationship between "games" and "developer". The compound relation has been integrated into the universal dependency framework to take into consideration for syntactic dependencies seen in multiword statements. This method guarantees that the relationships between compounds' constituent parts, including their modifiers and heads, are precisely reflected in the structure.

In Kyrgyz compound relation is used mainly noun-verb compounds (*жер жүзү, көз айнек*), and numbers, such as *он миң беш жүз алтымыш жети* (ten thousand five hundred and sixty-seven).

Compounds have various distinctions and subtypes as **compound:lvc**, **compound:prt**, **compound:redup** and **compound:svc**.

Compound: light verb construction (**compound:lvc**) category includes light verbs. The verb in a light-verb composition has very little semantic meaning. The semantics of the construction are determined by the non-head word, often a noun or adjective. Although the semantically loaded component of a light-verb construction is generally an adjective or a noun, it is common to observe verbs in this position particularly in code-switching settings.

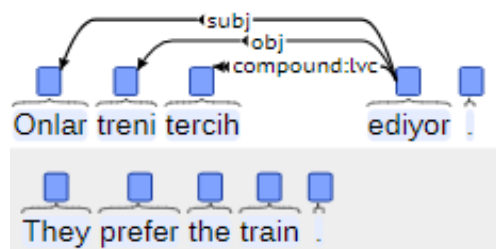


Figure 3. 14. Example of compound from the UD website

compound:prt (phrasal verb particle) is the phrasal verb particle relation that identifies an idiomatic phrasal verb, and holds between the verb and its particle (tagged as ADP). It is a subtype of the compound relation. In English **compound:prt** are mostly phrasal verbs such as *shut down*, *get along*, *run out*, *get behind*, *take off* and etc. **compound:redup** are reduplicated compounds. This subtype of compound covers a range of reduplicated forms in Kyrgyz. Reduplication is a common process especially for adverbs and adjectives, e.g. *көн-көн*, *чоң-чоң*, *аста-аста*, *чык-чык*, *секин-секин*, *кыбыр-кыбыр*, *кош-кош*, *мый-мый* and etc. **compound:svc** are serial verb compounds that are used for serial verb constructions. In this type of construction, several verbs are combined to describe the same action. In a serial verb formation, the verbs have the same subject, but they might not have the same object.

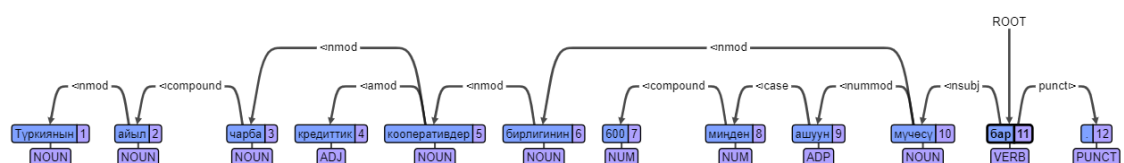


Figure 3. 15. Example of compound in the Kyrgyz sentence.

“Туркиянын айыл чарба кредиттик кооперативдер бирлигинин 600 миңден ашуун мүчөсү бар.” The translation of the following sentence into English: “The Union of Agricultural Credit Cooperatives of Turkiye has more than 600,000 members.”

"<Түркиянын>"

"Түркия" NOUN @nmod #1->2

"<айыл>"

"айыл" NOUN @compound #2->3

"<чарба>"

"чарба" NOUN @nmod #3->5

"<кредиттик>"

"кредит" ADJ @amod #4->5

"<кооперативдер>"

"кооператив" NOUN @nmod #5->6

"<бирлигинин>"

"бирлик" NOUN @nmod #6->10

"<600>"

"600" NUM @compound #7->8

"<миңден>"

"миң" NUM @case #8->9

"<ашуун>"

"ашуун" ADP @nummod #9->10

"<мүчөсү>"

"мүчө" NOUN @nsubj #10->11

"<бар>"

"бар" VERB @root #11->0

"<.>"

"." PUNCT @punct #12->11

This sentence consists of a series of words with their respective parts of speech and dependency relations. Here is an explanation of two compound words in the sentence:

"<Түркиянын>" (NOUN) is a noun that serves as a modifier (nmod) for the compound word "<айыл чарба>" (NOUN). It indicates a possessive relationship in genitive case, meaning "of Turkiye".

"<айыл>" (NOUN) is a noun that is compounded (**compound**) with the word "<чарба>" (NOUN), forming the phrase "айыл чарба" which means "agriculture".

"<чарба>" (NOUN) is a noun along with its compound "<айыл>" serves as a modifier (nmod) for the word "<кооперативдер>" (NOUN). It indicates the type or purpose of the credit, meaning "agriculture credit".

"<кредиттик>" (ADJ) is an adjective that modifies (amod) the word "<кооперативдер>" (NOUN). It describes the type of cooperatives, meaning "credit cooperatives".

"<кооперативдер>" (NOUN) is a noun that serves as a modifier (nmod) for the word "<бирлигинин>" (NOUN). It indicates a relationship, meaning "cooperative union".

"<бирлигинин>" (NOUN) is a noun that serves as a modifier (nmod) for the compound word "<600 миңден>" (NUM). It indicates a quantity, meaning "of 600 thousand".

"<600>" (NUM) is a number that is compounded (compound) with the word "<миңден>" (NUM), forming the phrase "600 миңден" which means "above 600 thousand".

"<миңден>" (NUM) is a case marker (case) that indicates the ablative case of the number, meaning "from 600 thousand".

"<ашуун>" (ADP) is a preposition that modifies (nummod) the word "<мүчөсү>" (NOUN). It indicates a measurement unit, meaning "per".

"<мүчөсү>" (NOUN) is a noun that serves as the subject (nsubj) of the verb "<бар>" (VERB). It represents the entity that is "going" or "existing". "<бар>" (VERB) is a verb that serves as the root of the sentence. It indicates the action of "going" or "existing".

"<.>" (PUNCT) is a punctuation mark (punct) that marks the end of the sentence.

conj: conjunct

A conjunct is the relationship between two parts in linguistic dependence relations that are joined by a coordinating conjunction, such "and" or "or." Coordinate structures are symmetrical, yet through the conj connection, the initial conjunction is often regarded as the parent of every subsequent coordinated phrase. In the sentence "Alice is kind and persistent," for instance, the relationship between the words "kind" and "persistent" is captured by the conjunct relation. There are two kinds of coordination: conjunctions between every pair of conjuncts and asyndetic coordination, which does not employ the coordinating conjunction (Djumalieva et al., 2023). Conjuncts are often separated by

commas or other punctuation when there are no coordinating conjunctions. Thus, conjuncts do not need to have the same type as the coordination structure; instead, the dependent structure represents a coordination structure with a conjunction like "and" or "so on". A conjunct in dependency relations provides a structure for expressing the syntactic links inside coordinated structures by symbolizing the relationship between items connected by coordinating conjunctions. Furthermore, the enhanced representation might include the connection between every conjunct and the parent of the coordination. The effective parents may be found algorithmically, and the information regarding shared dependents is unique; their explicit display is only for convenience. The enhanced representation displays all dependents, whereas the basic representation always ties a shared dependence among conjuncts to the initial conjunct (coordination head).

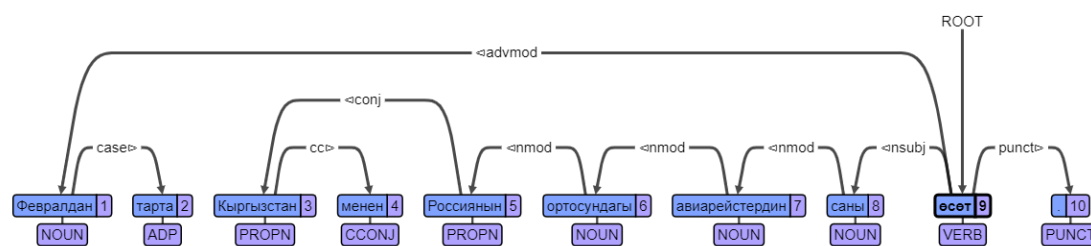


Figure 3. 16. Example of conj relation in a sentence.

“Февралдан тарта Кыргызстан менен Россиянын ортосундагы авиарейстердин саны өсөт.” In English: “From February, the number of flights between Kyrgyzstan and Russia will increase.”

"<Февралдан>"

"Февраль" NOUN @advmod #1->9

"<тарта>"

"тарта" ADP @case #2->1

"<Кыргызстан>"

"Кыргызстан" PROPN @conj #3->5

"<менен>"

"менен" CCONJ @cc #4->3

"<Россиянын>"

"Россия" PROPN @nmod #5->6

"<ортосундагы>"

"орто" NOUN @nmod #6->7
 "<авиарейстердин>"
 "авиарейс" NOUN @nmod #7->8
 "<саны>"
 "сан" NOUN @nsubj #8->9
 "<өсөт>"
 "өс" VERB @root #9->0
 "<.>"
 " ." PUNCT @punct #10->9

Here is the breakdown of the dependency relations and the explanation of conjuncts in the sentence:

"<Февралдан>" (From February) is a noun modifying the verb "<өсөт>" (grew) and is connected to it with the dependency relation "@advmod" (#1->9).

"<тарта>" (from) is a case marker indicating the source and is connected to "<Февралдан>" with the dependency relation "@case" (#2->1).

"<Кыргызстан>" (Kyrgyzstan) is a proper noun and is connected to "<Россиянын>" (Russia's) with the dependency relation "@conj" (#3->5).

"<менен>" (and) is a coordinating conjunction connecting "<Кыргызстан>" (Kyrgyzstan) and "<Россиянын>" (Russia's) with the dependency relation "@cc" (#4->3).

"<ортосундагы>" (middle) is a noun modifying "<авиарейстердин>" (airline's) and is connected to it with the dependency relation "@nmod" (#6->7).

"<авиарейстердин>" (airline's) is a noun modifying "<саны>" (number) and is connected to it with the dependency relation "@nmod" (#7->8).

"<саны>" (number) is the subject of the verb "<өсөт>" (grow) and is connected to it with the dependency relation "@nsubj" (#8->9).

"<өсөт>" (grew) is the main verb of the sentence and is connected to the root of the sentence with the dependency relation "@root" (#9->0).

"<.>" (.) is a punctuation mark indicating the end of the sentence and is connected to "<өсөт>" (grow) with the dependency relation "@punct" (#10->9).

cop: copula

"**Cop**" refers to the link between a function word and a nonverbal predicate, as in the context of the predication phrase "Kylie *is* an influencer" (thus for example, "*is*" here is a copula). In addition to conventional copulas, nonverbal (pronominal) copulas are widely used throughout the world's languages. Verbal copulas are tagged as AUX rather than VERB. Nominal copulas are tagged with PRON or DET. The cop relation should only be used in situations when the nonverbal predicate is considered to be the head of the phrase and in pure copulas, which most languages have at most TAME categories to the meaning of the predicate (changing its tense, for instance). A copula, also known as *copulas* or *copulae* in the plural, is a term or phrase used in linguistics to connect the subject and subject complement of a sentence. The words "is" in the sentence "The sky is blue" and "was not being" in the sentence "He was not being helpful" are two instances of this. Copula is a Latin word for a "link" or "tie" that joins two distinct objects. A copula is typically a verb or a word that appears like a verb, however, this isn't always the case. A copula is commonly called a *linking verb* in grammar courses for beginner and elementary students learning English. There is debate concerning the origins of the copula and how copulas are used in various languages. In the Kyrgyz, copulas are *эле, эмес, экен, бар, жок*, and others, just like cop relations in other Turkic languages. Let us observe one example of the copula word "эле" in a Kyrgyz sentence.

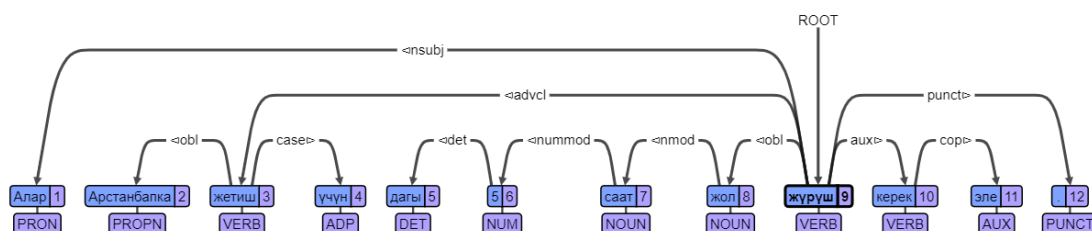


Figure 3. 17. Example of cop relation in a sentence.

"Алар Арстанбапка жетиш үчүн дагы 5 саат жол жүрүш керек эле." is the given sentence in Kyrgyz. Here is its translation: "They had to travel another 5 hours to reach Arslanbob."

"<Алар>"

"Алар" PRON @nsbj #1->9

"<Арстанбапка>"

"Арстанбапка" PROPN @obl #2->3

"<жетиш>"

"жетиш" VERB @advcl #3->9
 "<үчүн>"
 "үчүн" ADP @case #4->3
 "<дагы>"
 "дагы" DET @det #5->6
 "<5>"
 "5" NUM @nummod #6->7
 "<саат>"
 "саат" NOUN @nmod #7->8
 "<жол>"
 "жол" NOUN @obl #8->9
 "<жүрүш>"
 "жүрүш" VERB @root #9->0
 "<керек>"
 "керек" VERB @aux #10->9
 "<эле>"
 "эле" AUX @cop #11->10
 "<.>"
 "." PUNCT @punct #12->9

Here is the explanation of the dependency relations in the sentence:

"<Алар>" (They) is a pronoun and is the subject of the verb "<жүрүш>" (walk) and is connected to it with the dependency relation "@nsubj" (#1->9).

"<Арстанбапка>" (to Arslanbob) is a proper noun and is connected to "<жетиш>" (arrive) with the dependency relation "@obl" (#2->3).

"<жетиш>" (arrive) is the main verb of the sentence and is connected to the root of the sentence with the relation "@advcl" (#3->9).

"<үчүн>" (for) is a case marker indicating the purpose and is connected to "<жетиш>" (arrive) with the relation "@case" (#4->3).

"<дагы>" (other) is a determiner modifying "<5>" (5) and is connected to it with the dependency relation "@det" (#5->6).

"<5>" (5) is a numeral modifying "<саат>" (hours) and is connected to it with the dependency relation "@nummod" (#6->7).

"<саат>" (hours) is a noun and is connected to "<жол>" (road) with the dependency relation "@nmod" (#7->8).

"<жол>" (road) is an oblique noun and is connected to "<жүрүш>" (walk) with the dependency relation "@obl" (#8->9).

"<жүрүш>" (walk) is the main verb and the root of the sentence with the dependency relation "@root" (#9->0).

"<керек>" (must) is an auxiliary verb indicating necessity and is connected to "<жүрүш>" (walk) with the dependency relation "@aux" (#10->9).

"<эле>" is a copula verb indicating emphasis and is connected to "<керек>" (must) with the dependency relation "@cop" (#11->10).

"<.>" (.) is a punctuation mark indicating the end of the sentence and is connected to "<жүрүш>" (walk) with the dependency relation "@punct" (#12->9).

According to Kelepir (2003), in Turkish language sentences with non-verbal predicates have copulas. These predicates can be participles, adjectives, or nouns. In English, for instance, the verb *to be* has the inflected verb role in these kinds of phrases, but it adds no meaning to the clause. In Kyrgyz, the copula is a syntactic requirement of the functional heads when the role of a predicate is usually played by nominal; it has no semantic role. They have a distinct position inside the syntactic framework. In Kyrgyz, they often appear toward the conclusion of the phrase and add an emphasis to the actual predicate.

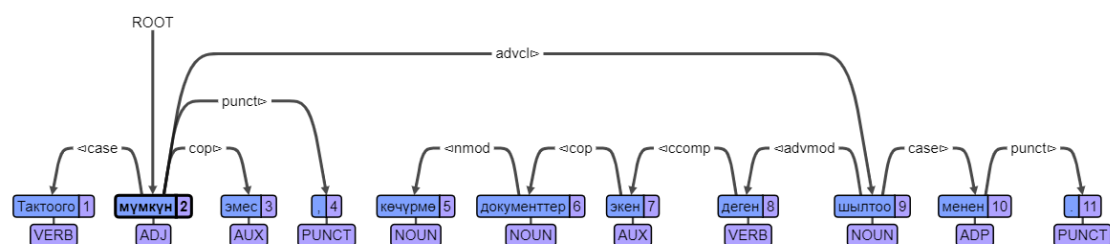


Figure 3. 18. Example 2 of cop relation in a sentence.

"Тактоого мүмкүн эмес, көчүрмө документтер экен деген шылтоо менен." is the sentence in Kyrgyz, containing 2 copula relations. Its translation into English: "It is impossible to verify, under the pretext that they are duplicate documents."

"<Тактоого>"

"Тактоого" VERB @case #1->2

"<мүмкүн>"

"мүмкүн" ADJ @root #2->0
 "<эмес>"
 "эмес" AUX @cop #3->2
 "<,>"
 ", " PUNCT @punct #4->2
 "<көчүрмө>"
 "көчүрмө" NOUN @nmod #5->6
 "<документтер>"
 "документтер" NOUN @cop #6->7
 "<экен>"
 "экен" AUX @ccomp #7->8
 "<деген>"
 "деген" VERB @advmod #8->9
 "<шылтоо>"
 "шылтоо" NOUN @advcl #9->2
 "<менен>"
 "менен" ADP @case #10->9
 "<.>"
 ". " PUNCT @punct #11->10

Here is the explanation of copulas that come after adjectives and nouns:

"<Тактоого>" (To clarify) is a verb and is connected to "<мүмкүн>" (possible) with the dependency relation "@case" (#1->2).

"<мүмкүн>" (possible) is the main adjective of the sentence and is the root with the dependency relation "@root" (#2->0).

"<эмес>" (not) is an auxiliary verb indicating negation and is connected to "<мүмкүн>" (possible) with the dependency relation "@cop" (#3->2).

"<,>" (,) is a punctuation mark and is connected to "<мүмкүн>" (possible) with the dependency relation "@punct" (#4->2).

"<көчүрмө>" (copy) is a noun and is connected to "<документтер>" (documents) with the dependency relation "@nmod" (#5->6).

"<документтер>" (documents) is a noun and is connected to "<экен>" (are) with the dependency relation "@cop" (#6->7).

"<экен>" (are) is an auxiliary verb indicating existence and is connected to "<деген>" (say) with the dependency relation "@ccomp" (#7->8).

"<деген>" (saying) is a verb and is connected to "<шылтоо>" (excuse) with the dependency relation "@advmod" (#8->9).

"<шылтоо>" (inspection) is a noun and is connected to "<менен>" (with) with the dependency relation "@advcl" (#9->2).

"<менен>" (with) is a case marker indicating association and is connected to "<шылтоо>" (excuse) with the dependency relation "@case" (#10->9).

"<.>" (.) is a punctuation mark indicating the end of the sentence and is connected to "<менен>" (with) with the dependency relation "@punct" (#11->10).

csubj: clausal subject

Syntactically, clausal subjects are those in which the subject is a phrase in and of itself. The root of the sentence is the copular verb's complement when the verb is a copular verb, meaning that the governor of this relation does not have to be a verb. The primary lexical verb or another predicate in the subject phrase is the dependent. The clausal subject of a passive verb or verb group can also be expressed using the csubj connection. In languages where there is a grammaticalized passive transformation, using the subtype csubj:pass is highly advised. csubj:outer may be used if the subject of a copular clause whose predicate is itself a clause.

csubj:outer is outer clause clausal subject. In order to indicate that it is not the subject of the contained clause, this relation indicates the clausal subject of a copular clause whose predicate is also a clause. The majority of clausal subjects of copular clauses are not meant for the:outer subtype; only those where the predicate is a sentence in and of itself. In the case where the copular clause predicate is a nominal, adjective, etc., plain **csubj** (or another subtype) will be suitable.

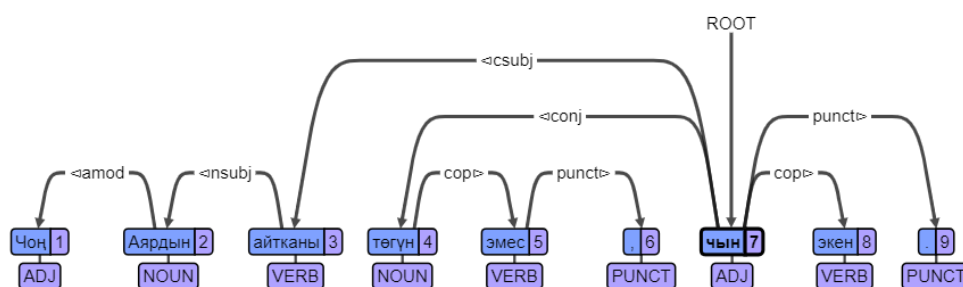


Figure 3. 19. Example of csubj in the sentence.

“*Чоң Аярдын айтканы төгүн эмес, чын экен.*” – is a sentence in the Kyrgyz language that contains clausal subject in it. “What Ayar said was not a lie, but truth.” is its translation into English.

"<Чоң>"

"Чоң" ADJ @amod #1->2

"<Аярдын>"

"Аяр" NOUN @nsubj #2->3

"<айтканы>"

"айт" VERB @csubj #3->7

"<төгүн>"

"төгүн" NOUN @conj #4->7

"<эмес>"

"эмес" VERB @cop #5->4

"<,>"

"," PUNCT @punct #6->5

"<ЧЫН>"

"ЧЫН" ADJ @root #7->0

"<экен>"

"экен" VERB @cop #8->7

"<.>"

"," PUNCT @punct #9->7

In this sentence, the token "<Чоң>" is an adjective modifying the noun "Аярдын". "<Аярдын>" is a noun and serves as the subject of the verb "айтканы". "<айтканы>" is a verb and has "Аярдын" as its subject. These tokens together make a clausal subject for the root of the sentence. “*Чоң Аярдын айтканы*” is a clausal subject that has a nsubj relation and can be a separate sentence itself. "<төгүн>" is a noun and is connected to "эмес" through the copula "экен". "<эмес>" is a copula and is connected to "төгүн". "<ЧЫН>" is an adjective and is the root of the sentence. "<экен>" is a verb and serves as the copula for "ЧЫН". "<.>" is a punctuation mark, indicating the end of the sentence.

csubj:pass is a clausal passive subject. A clausal passive subject is a clausal syntactic subject of a passive clause. This subtype is very common for the Indo-European language family, for instance, in English. See the example below.

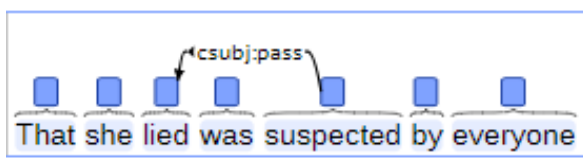


Figure 3. 20. Example of csubj:pass in a sentence.

det: determiner

In dependency parsing, a determiner is the relation between the head of a noun phrase and its determiner. In linguistics a word that comes before a noun and indicates something about its ownership, definiteness, or quantity is known as a determiner. It gives context that makes an explanation regarding the noun's quantity, definiteness, or ownership. Determiners add distinctiveness to nouns through a subtractive process that limits their identity. They serve to both denote quantity and number and to make clear what a word refers to. There are various categories of determiners, such as possessives, demonstratives, articles, and quantifiers.

Types of Determiners in English:

- Articles: These include "a," "an," and "the." They identify a noun as specific or nonspecific.
- Demonstratives: Words like "this," "that," "these," and "those" communicate the placement of a noun in space or time.
- Possessives: These include "my," "your," "his," "her," "its," "our," and "their," and are used to describe ownership or possession.
- Quantifiers: Examples include "some," "a few," "many," "a little," "little," "few," and "none," and they indicate the quantity or amount of the noun.

Determiners are essential for defining and changing nouns because they give context, which is necessary to comprehend the significance and traits of the nouns they come before. They add to the overall meaning and distinction of the language and are a crucial part of the structure of noun phrases. They limit or "determine" a noun phrase in some way and are mutually-exclusive, meaning only one determiner can be used in the same noun phrase. They come before adjectives and play a significant role in making sentences sound and look sensible and meaningful.

Now in an updated version of Universal Dependencies tagsets, there are 3 subtypes of a determiner (det): det:numgov, det:nummod, and det:poss. **det:numgov** is a pronominal

quantifier governing the case of the noun. For instance, since pronominal quantifiers in Slavic languages often do not agree with the quantified noun in case (unlike non-quantifying determiners), they are called **det:numgov** instead of **det**. The counted noun must be in its genitive form in order for the quantifier to work. The entire sentence—quantifier with noun—is regarded as a single, neuter noun phrase that can be used in place of expected nominative, accusative, or vocative noun phrases. The quantifier is not labeled as the head of the nominal in order to promote parallelism between languages (and even between morphological situations within a single language). Nonetheless, the information on the case conditions is preserved using the **det:numgov** label.

det:nummod is a pronominal quantifier agreeing in case with the noun. Since pronominal quantifiers in Slavic languages typically disagree with the quantified noun in question (unlike non-quantifying determiners), in Universal Dependencies they are designated *det:nummod* or *det:numgov* rather than **det**. They tend to agree when the entire phrase—quantifier with a noun—plays a part that would normally be filled by genitive, dative, locative, or instrumental noun phrases. These are marked with the term **det:nummod**.

det:poss is a possessive determiner. **Det:poss** should be used in place of **det** if a possessive determiner is present. Possessive is an aspect shared by all possessive determiners. The kind of pronoun that is used to convey possession or belongingness is the possessive determiner. It assists in identifying the owner of the noun that it determines. In English, possessive determiners like "my," "your," "his," "her," "its," "our," and "their" are frequently used. Possessive determiners modify nouns and precede them in sentences, while possessive pronouns replace nouns and stand alone. For instance, in the sentence “Бул менин китебим.” “менин” is a possessive determiner, modifying the noun “китебим”, however in the following sentence “Бул меники.”, the word “меники” considered to be a possessive pronoun in Universal Dependencies syntactic relations tagsets.

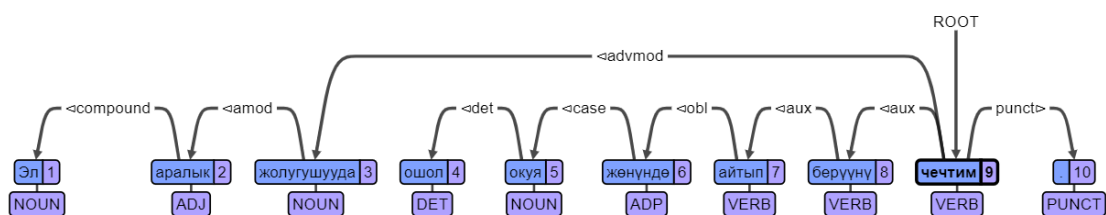


Figure 3. 21. Example of det relation in a sentence.

“Эл аралык жолугушууда ошол окуя жөнүндө айтып берүүнү чечтим.” is a sentence in Kyrgyz that has a det relation. Here is the translation of it: “I decided to tell about that event at the international meeting.”

"<Эл>"

"Эл" NOUN @compound #1->2

"<аралык>"

"аралык" ADJ @amod #2->3

"<жолугушууда>"

"жолугушууда" NOUN @advmod #3->9

"<ошол>"

"ошол" DET @det #4->5

"<окуя>"

"окуя" NOUN @case #5->6

"<жөнүндө>"

"жөнүндө" ADP @obl #6->7

"<айтып>"

"айтып" VERB @aux #7->8

"<берүүнү>"

"берүүнү" VERB @aux #8->9

"<чечтим>"

"чечтим" VERB @root #9->0

"<.>"

"." PUNCT @punct #10->9

In this sentence, the tokens <ошол> and <окуяны> are connected with the **det** relation and the word “ошол” is a determiner for the word “окуя”. The token <окуя> is a noun

that means “story” in the Kyrgyz language. The token <ошол> indicates the determiner “this” and it becomes “this story”.

discourse

discourse is a discourse element. Interjections and other discourse particles and elements—which, aside from being expressive—do not directly relate to the sentence's structure—are included in this way. According to Penn’s Treebanks definition, this includes non-adverbial discourse markers (*well, like, of course, but you know or actually*), fillers (*um, ah*), and interjections (*oh, uh-huh, Welcome*).

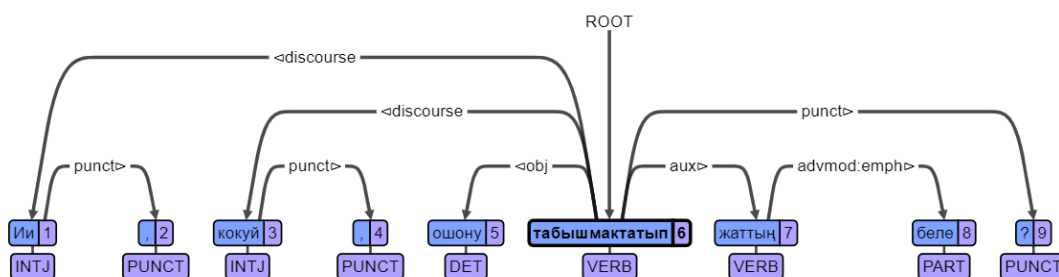


Figure 3. 22. Example of discourse relation in a sentence.

“Ии, кокуй, ошону табышмактатып жаттың беле?” is a sentence that has two discourse elements that are “Ии” and “кокуй”. The sentence’s translation into English: “Oy, well, you were hiding it as a secret?”

"<Ии>"

"Ии" INTJ @discourse #1->6

"<,>"

"," PUNCT @punct #2->1

"<кокуй>"

"кокуй" INTJ @discourse #3->6

"<,>"

"," PUNCT @punct #4->3

"<ошону>"

"ошо" DET @obj #5->6

"<табышмактатып>"

"табышмакта" VERB @root #6->0

"<жаттың>"

"жат" VERB @aux #7->6

"<беле>"

"беле" PART @advmod:emph #8->7

"<?>"

"?" PUNCT @punct #9->6

fixed: multiword expression

fixed is a fixed multiword expression. A particular kind of multiword expression known as a fixed multiword expression has a fixed or rigid structure and functions similarly to a single unit. These phrases act similarly to function words and are frequently grammaticized. With few exceptions, they usually do not permit intervening words and are not morphosyntactically flexible. Some examples of fixed multiword formulations include the preposition "in spite of" or "as well as" in English. Because these phrases have a definite shape and are difficult to change or rearrange, they are regarded as fixed expressions. It is significant that the notion of fixed multiword phrases might show slight variations depending upon the particular linguistic framework or research perspective. Overall in, though, the main idea is that multiword expressions that are fixed are different from other kinds of multiword expressions that could have greater structural flexibility. The annotation of fixed MWEs is done in a flat structure, where the fixed label is used to tie all following words in the expression to the initial word. There shouldn't be any internal modifications made to fixed MWEs. Since punct, goeswith, and reparandum dependents are not actual syntactic connections, a word that attaches as fixed should not have any dependents. The fixed relation replaces the old **mwe** relation to prevent misunderstanding regarding its scope.

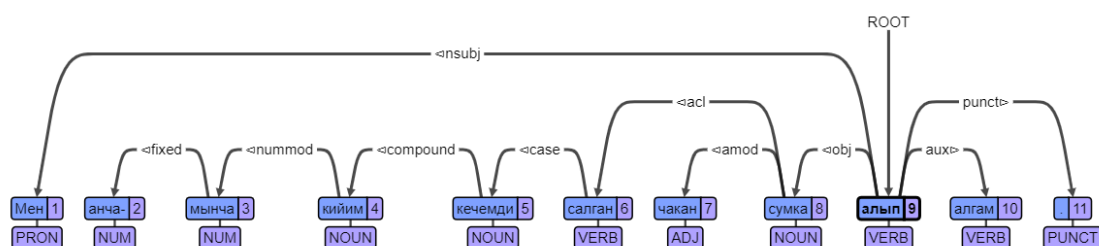


Figure 3. 23. Example of fixed relation in a sentence.

"Мен анча-мынча кийим кечемди салган чакан сумка алып алгам." is a sentence that contains a multiword expression "анча-мынча". The sentence in English as follows: "I bought a small bag in which I put my few clothes."

"<Мен>"

"Мен" PRON @nsubj #1->9
 "<анча->"
 "анча-" NUM @fixed #2->3
 "<мынча>"
 "мынча" NUM @nummod #3->4
 "<кийим>"
 "кийим" NOUN @compound #4->5
 "<кечемди>"
 "кечемди" NOUN @case #5->6
 "<салган>"
 "салган" VERB @acl #6->8
 "<чакан>"
 "чакан" ADJ @amod #7->8
 "<сумка>"
 "сумка" NOUN @obj #8->9
 "<алып>"
 "алып" VERB @root #9->0
 "<алгам>"
 "алгам" VERB @aux #10->9
 "<.>"
 " ." PUNCT @punct #11->9

Here's a breakdown of the relationships:

1. "<Мен>" ("I") is the subject of the sentence, indicated by the @nsubj (nominal subject) relation to the verb "алып" ("took") at position #9.
2. "<анча->" ("a few") is fixed to the following word "<мынча>" ("a few"), forming a numerical phrase "a few/few," indicated by the @fixed relation between #2 and #3.
3. "<мынча>" ("few") modifies "<кийим>" ("clothes"), indicating that the clothes were a few, shown by the @nummod (numerical modifier) relation between #3 and #4.
4. "<кийим>" ("clothes") is compounded with "<кечемди>" ("item") to form "clothes" shown by the @compound relation between #4 and #5.
5. "<кечемди>" ("items") is in the genitive case, indicating possession or relationship, related to the action of taking, shown by the @case relation between #5 and #6.

6. "<салиган>" ("put") is the verb of the sentence, related to the noun "<сумка>" ("bag") through the action of putting some clothes, shown by the @acl (adjectival clause modifier) relation between #6 and #8.
7. "<чакан>" ("big") modifies "<сумка>" ("bag"), indicating the size of the bag taken, shown by the @amod (adjectival modifier) relation between #7 and #8.
8. "<сумка>" ("bag") is the direct object of the verb "алып" ("took"), shown by the @obj (object) relation between #8 and #9.
9. "<алып>" ("took") is the main verb of the sentence, with "<Мен>" ("I") as the subject and "<сумка>" ("bag") as the direct object, shown by the @root relation to #0 (root of the sentence).
10. "<алгам>" ("did") is an auxiliary verb indicating past tense, related to the main verb "алып" ("took"), shown by the @aux (auxiliary) relation between #10 and #9.
11. "<.>" (".") is a punctuation mark indicating the end of the sentence, related to the root of the sentence, shown by the @punct (punctuation) relation between #11 and #9.

flat: flat expression

flat is a flat expression. When normal substitution tests fail to identify any of the immediate components as the single head, the flat relation is employed to combine the constituents of the phrase. This covers instances where many elements satisfy the head test, such in the case of *Sadyr Japarov*, where either Sadyr or Japarov can function as the whole in most situations, as well as instances where none do, like in *Чоң Алай* (in Kyrgyz). Keep in mind that in these situations, the flat connection is only applicable if no further particular relation exists. For instance, any of the conjuncts can typically take the place of the total in coordination structures annotated with the conj connection.

A flat structure is used to annotate flat expressions, where the **flat** tag is used to connect each existing component in the expression to the initial one. It is assumed that the structural annotation in these expressions is, in principle, independent and that the flat relations are not syntactic head-modifier connections. A flat expression's constituent parts might each have nested flat structures as dependents. For instance, the first name *Батма Зуура* and the last names *Исабекова* and *Камчыбековна* are both flat expressions that are merged to form a bigger flat name in the name *Исабекова Батма Зуура Камчыбековна*.

Personal names, foreign phrases, iconic sequences, and things divided for readability are the prototypes for flat. Beyond these prototypes, flat may be used to other forms of name and number expressions, for example. But instead of using flat, every attempt should be made to discover a head, even if an expression is peculiar or follows a certain pattern. **Dep** can be used if a head can be located but no substantive dependent relation is suitable. The scope of **flat** may extend beyond names of persons to names of other kinds of entities that depart from general headed structure. The expressions under this category must be established by language-specific criteria.

Some examples of sequences for which neither head-dependent nor coordination connections remain accurate include "filler" words (do re mi), onomatopoeia (boom boom), and gibberish (blah blah blah). These sequences possess a **flat** annotation as well.

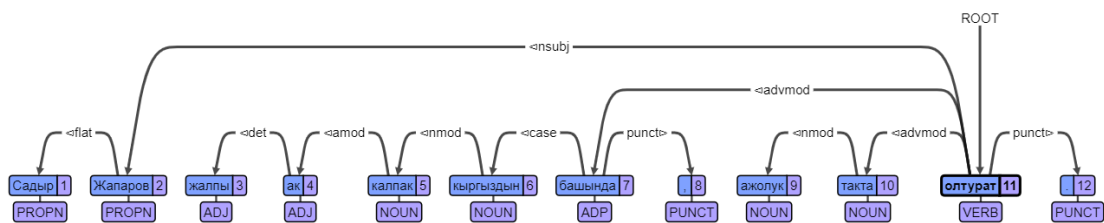


Figure 3. 24. Example of flat expression in a sentence.

"Садыр Жапаров жалпы ак калпак кыргыздын башында, ажолук такта олтурат."- is a sentence that has a **flat** expression structure "". Its translation into English: "Sadyr Japarov sits at the head of the general white kalpak Kyrgyz nation, at the azho throne."

"<Садыр>"

"Садыр" PROPN @flat #1->2

"<Жапаров>"

"Жапаров" PROPN @nsubj #2->11

"<жалпы>"

"жалпы" ADJ @det #3->4

"<ак>"

"ак" ADJ @amod #4->5

"<калпак>"

"калпак" NOUN @nmod #5->6

"<кыргыздын>"

"кыргыздын" NOUN @case #6->7
 "<башында>"
 "башында" ADP @advmod #7->11
 "<,>"
 ", " PUNCT @punct #8->7
 "<ажолук>"
 "ажолук" NOUN @nmod #9->10
 "<такта>"
 "такта" NOUN @advmod #10->11
 "<олтурат>"
 "олтурат" VERB @root #11->0
 "<.>"
 ". " PUNCT @punct #12->11

In the provided Kyrgyz sentence:

"<олтурат>" ("leads") is the main verb of the sentence, indicated by the relation "@root" to the word at position #0.

"<Жапаров>" ("Zhaparov") is the subject of the sentence, linked to the main verb "<олтурат>" ("sits") through the relation "@nsubj" (nominal subject) with the word at position #11.

"<Садыр Жапаров>" ("Sadyr Zhaparov") is a compound proper noun where "<Садыр>" ("Sadyr") is linked to "<Жапаров>" ("Zhaparov") via the relation "@flat" (flat multiword expression).

"<жалпы>" ("general") is an adjective modifying the noun "<калпак>" ("hat"), forming the phrase "general hat", shown by the relation "@det" (determiner) between #3 and #4.

"<ак>" ("white") is an adjective modifying the noun "<калпак>" ("hat"), forming the phrase "white hat", indicated by the relation "@amod" (adjectival modifier) between #4 and #5.

"<калпак>" ("hat") is modified by the adjectives "<жалпы>" ("general") and "<ак>" ("white"), and is also related to "<кыргыздын>" ("of Kyrgyz") through the relation "@nmod" (nominal modifier) between #5 and #6.

"<кыргыздын>" ("Kyrgyz") is in the genitive case, indicating possession or relationship, related to "<калпак>" ("hat"), shown by the relation "@case" between #6 and #7.

"<башында>" ("on the head") is an adverbial phrase modifying the verb "<олтурат>" ("leads"), indicating the location where Zhaparov wears the hat, shown by the relation "@advmod" (adverbial modifier) between #7 and #11.

"<,>" (" , ") is a punctuation mark separating parts of the sentence. It is related to the adverbial phrase "<башында>" (" , ") through the relation "@punct" (punctuation) between #8 and #7.

"<ажолук>" ("public") is a noun modifying the noun "<такта>" ("platform"), forming the phrase "public platform", indicated by the relation "@nmod" (nominal modifier) between #9 and #10.

"<такта>" ("platform") is modified by the noun "<ажолук>" ("public"), and it is also an adverbial modifier of the verb "<олтурат>" ("leads"), indicating the function or location of Zhaparov's leadership, shown by the relation "@advmod" between #10 and #11.

"<.>" (" . ") is a punctuation mark indicating the end of the sentence, related to the main verb "<олтурат>" ("leads"), shown by the relation "@punct" (punctuation) between #12 and #11.

list

For chains of related objects, the **list** relation is employed. All items in lists containing more than two should alter the first item. A list that resembles a list of paragraphs, such as a catalog of things, will include one or more sentences for each item, and as there are no relationships between the sentences, there won't be any list relations. On the other hand, casual and online writing frequently includes sections that are processed as single sentences but are intended to be read as lists. These structures, for instance, are frequently seen in email signatures where the various contact information pieces are categorized as lists.

Another place where list has been used is for a sequence of attributes or descriptive terms used as the title line of a review (such as product or restaurant reviews, etc.): Long Lines, Silly Rules, Rude Staff, Ok Food.

The use of **list** shouldn't be excessive, though. Even if a composition is typographically organized as a list, it should be evaluated using the more common grammatical connections of standard sentences if this can be done with ease, usually as a coordinated structure. Specifically, the list should be examined as a coordinated structure when it is expressed as a single phrase with commas and overt coordination.

mark

The word "marker" indicates a clause as coming under the authority of another clause. This includes terms like [en] *that* or *whether* in a complement clause. The marker for an adverbial phrase is usually a subordinating conjunction, such as [en] *although* or *while*. The subordinate clause head is the dependant of the marker. It is important to differentiate relative pronouns like [en] *who* or *there*, which complete a regular verbal argument or modify grammatical connection, from relative clause markers, which are **mark**. In the Kyrgyz language markers are эгерде, эгер, да, бы particle.

nmod: nominal modifier

A noun (or noun phrase) that serves as an adjunct or non-core (oblique) argument is referred to as a **nominal modifier** in dependency relations. This indicates that when it joins a verb, adjective, or other adverb, it functions similarly to an adverbial. On the other hand, it relates to a quality or genitive complement when appended to a noun. In morphologically rich languages, the case can be used to further specify the nmod relation, giving nominal and case-marking elements a uniform analysis. Nmod has two additional subtypes: **nmod:poss** and **nmod:tmod**. Treebanks may optionally use nmod:poss to distinguish non-adpositional possessives such as Айзаттын китеби (Aizat's book), Айбектин тобу (Aibek's ball).

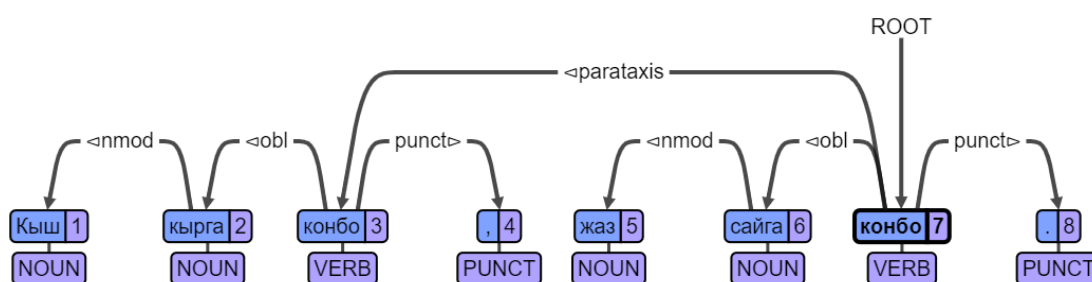


Figure 3. 25. Example of nmod relation in a sentence.

"Кыш кырга конбо, жаз сайга конбо." is a proverb in the Kyrgyz Language that contains 2 nominal modifiers "кыш" and "жаз", its translation into English: "Don't land on the winter field, don't land on the spring valley."

"<Кыш>"

"Кыш" NOUN @nmod #1->2

"<кырга>"

"кырга" NOUN @obl #2->3

"<конбо>"
 "конбо" VERB @parataxis #3->7
 "<,>"
 "," PUNCT @punct #4->3
 "<жаз>"
 "жаз" NOUN @nmod #5->6
 "<сайга>"
 "сайга" NOUN @obl #6->7
 "<конбо>"
 "конбо" VERB @root #7->0
 "<.>"
 "." PUNCT @punct #8->7

Here is the breakdown of the Kyrgyz sentence with two nominal modifiers:

"<конбо>" ("do not land") is the main verb of the sentence, indicated by the relation "@root" to the word at position #0.

"<Кыш>" ("winter") is a noun that modifies the noun "< кырга >" ("edge/hill"), indicating the direction or target of the action, forming the phrase "winter edge/hill", shown by the relation "@nmod" (nominal modifier) between #1 and #2.

"<кырга>" ("to the edge") is a noun indicating direction or destination, related to the verb "<конбо>" ("do not land") through the relation "@obl" (oblique nominal) between #2 and #3.

"<,>" (" , ") is a punctuation mark separating parts of the sentence. It is related to the verb "<конбо>" ("do not land") through the relation "@punct" (punctuation) between #4 and #3.

"<жаз>" ("spring") is a noun that modifies the noun "<сайга>" ("to the valley"), indicating the direction or target of the action, forming the phrase "spring valley", shown by the relation "@nmod" between #5 and #6.

"<сайга>" ("to the valley") is a noun indicating direction or destination, related to the verb "<конбо>" ("do not land") through the relation "@obl" (oblique nominal) between #6 and #7.

"<.>" (" . ") is a punctuation mark indicating the end of the sentence, related to the main verb "<конбо>" ("do not land"), shown by the relation "@punct" between #8 and #7.

nsubj: nominal subject

A nominal subject is represented by the symbol **nsubj**. This argument is the more agentive, the do-er, or the proto-agent of the clause; that is, it is in the position that passes the standard grammatical test for subjecthood. This nominal can have a noun at its head, a pronoun or relative pronoun, or in situations with an ellipsis, other items like an adjective (Musazhanova et al., 2023). Even though the subject is usually not the proto-agent argument because of valency-changing operations, the nsubj relation is still employed for the nominal subject of a passive verb or verb group. In languages where there is a grammaticalized passive transformation, using the subtype **nsubj:pass** is highly advised. **nsubj:outer** can be used when the subject of a copular clause is itself a clause, i.e., when the predicate is a clause.

See examples of nsubj in the above annotated sentences, each of them contains various nsubj relations.

nummod: numeric modifier

Any number or number phrase that modifies the noun's meaning by adding a quantity is called a numeric modifier (nummod).

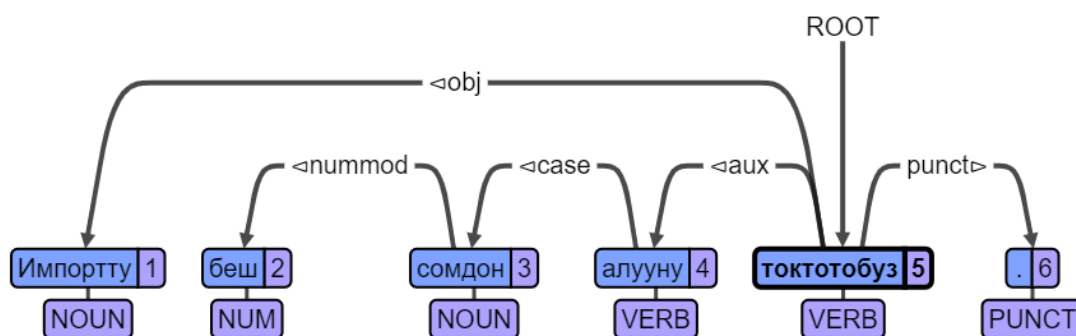


Figure 3. 26. Example of nummod relation in a sentence.

“Импортту беш сомдон алууну токтотобуз.” is a Kyrgyz sentence that has a numeric modifier “беш” that modifies the noun “сомдон”. Its translation into English: “We will stop taking imports from five soms.”

"<Импортту>"

"Импортту" NOUN @obj #1->5

"<беш>"

"беш" NUM @nummod #2->3

"<сoмдoн>"
 "сoмдoн" NOUN @case #3->4
 "<aлyунy>"
 "aлyунy" VERB @case #4->5
 "<тoктoтoбyз>"
 "тoктoтoбyз" VERB @root #5->0
 "<.>"
 " ." PUNCT @punct #6->5

In the provided Kyrgyz sentence there is a **nummod** relation, which means that it contains a numerical modifying a noun:

"<тoктoтoбyз>" ("we will stop") is the main verb of the sentence, indicated by the relation "@root" to the word at position #0.

"<Импорттy>" ("import") is the direct object of the verb "<тoктoтoбyз>" ("we will stop"), shown by the relation "@obj" (object) between #1 and #5.

"<беш>" ("five") is a numeral modifier of the noun "<сoмдoн>" ("som", a Kyrgyz currency), forming the phrase "five som", shown by the relation "@nummod" (numerical modifier) between #2 and #3.

"<сoмдoн>" ("from som") is in the ablative case, indicating the unit of currency, related to the verb "<тoктoтoбyз>" ("we will stop"), shown by the relation "@case" between #3 and #4.

"<aлyунy>" ("taking") is a verb indicating the case, related to the main verb "<тoктoтoбyз>" ("we will stop"), shown by the relation "@case" (case) between #4 and #5.

"<.>" (".") is a punctuation mark indicating the end of the sentence, related to the main verb "<тoктoтoбyз>" ("we will stop"), shown by the relation "@punct" (punctuation) between #6 and #5.

obj: object

After the subject, the object is a verb's most important argument. Usually, the noun phrase refers to the object being acted upon or changing in state or motion (the proto-patient). The accusative case is frequently used to indicate the object in languages that distinguish between morphological cases. The main question is whether the cases that a verb dictates (dative, genitive, etc.) are core in the language. Whether an adposition is

present or not, these cases are frequently indirect. As a result, even though the conventional grammar refers to these dependents as "objects," they are forced to use the **obl** relation instead of the **obj** relation. In the scenario that there are multiple objects, one should be designated as **iobj** (indirect object) and the others as **obj**. In these situations, determining which object (the patient) is most directly impacted is necessary. In most cases, if there is only one object, it should be **obj** unless it is morphosyntactically closer to standard patient arguments than it is to instances of **iobj** in the language (Djumaliev et al., 2023).

Direct **objects** usually take either the accusative or nominative (unmarked) cases in Kyrgyz, such as “Алина китеп сатып алды.” (Alina bought a book.). Here in this sentence, the word “китеп” is unmarked, i.e. it is an object in a nominative case.

If the predicate just affects part of an item, it can also be in the ablative case. For all "non-core" arguments and adjuncts, we employ the **obl** relation; we do not label arguments of verbs in other circumstances with **obj**. See the examples of object relation in the above given sentences.

obl: oblique nominal

A noun phrase that functions as a complement or modifier in a sentence is known as a "oblique nominal" in linguistics. It usually denotes a connection, such as one of possession, direction, location, or the recipient. It is referred to as "oblique" because it performs a different grammatical function instead of acting as the main verb's subject or direct object. For instance, "to her" is an oblique nominal in the sentence "*He gave the book to her*," designating the person to whom the action is being given. Similar to this, in "*She walked through the park*," the phrase "through the park" specifies the action's location in an indirect manner.

In languages that have them, oblique nominals are frequently indicated by prepositions or case endings (dative, ablative, or genitive). The noun must be indicated as an object (**obj**) whether it is in the accusative or nominative case. They convey further details about the relationship or action that the statement describes. When a nominal (noun, pronoun, or noun phrase) serves as an adjunct or non-core (oblique) argument, the **obl** relation is employed. This indicates that it is equivalent to an adverbial attached to a verb, adjective, or other adverb in terms of functionality. The annotated sentences above may provide instances of oblique sentences.

parataxis

Parataxis is a relation between a word and other elements, such as a parenthetical or clause after a ":" or ";", placed side by side without explicit coordination. It is a discourse-like equivalent of coordination and usually obeys an iconic ordering. The first part of a sentence is usually the head, while the second part is parataxis dependent, regardless of the language's headedness properties.

Syntactic parallelism is frequently found in the sentence structure of compound sentences, which are composed of two simple sentences joined by the use of conjunctions and intonation. The proverb's meaning—the comparison of two unlike things—is revealed by contrasting the two simple sentences to create a compound sentence. "The repetition of individual words in components, the drop-out of individual members of a sentence, the commonality of the individual members of a sentence for both components, the coincidence of the form of time and person of verbal predicates" are examples of syntactic parallelism in compound sentences (Thompson, 2021).

The parataxis relation refers to treating two separate sentences as a single sentence, often due to sentence segmentation following sentence-final punctuation or utterance turns in spoken corpus. This can occur when there is a clear discourse relation linking two clauses, and sometimes more than two sentences are joined in this way. To maximize similarity to conjunction analysis, all later sentences are made dependent on the first sentence.

Parataxis contrasts with hypotaxis, which subordinates ideas and exhibits juxtaposition, transition, and connection, by presenting ideas without explicit subordination or coordination. Examples of parataxis always juxtapose brief, straightforward, independent phrases without the use of subordinating conjunctions. As parataxis is a parataxic device, coordinated conjunctions establish equal links between the various elements of a phrase (Thompson, 2021).

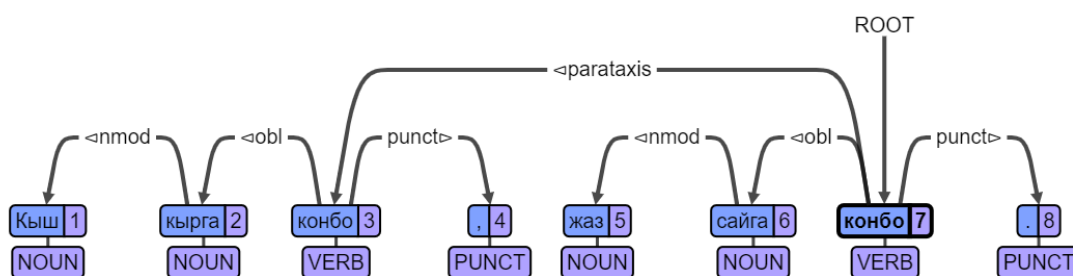


Figure 3. 27. Example of parataxis relation in the sentence.

“Кыш кырга конбо, жаз сайга конбо.” – “Don't land on the winter field, don't land on the spring valley.”

In the provided Kyrgyz proverb, the relation labeled as "@parataxis" is associated with the word "<конбо>" ("do not land") at positions #3 and #7.

Parataxis refers to a syntactic relationship where two or more clauses are juxtaposed or coordinated without the use of a conjunction or subordinating element. In this context, "<конбо>" ("do not land") is related to the preceding words "<кырга>" and "<Кыш>" as well as the following words "<сайга>" and "<жаз>" through parataxis.

This implies that the action described by "<конбо>" is directly linked or coordinated with the nominal phrases "<Кыш кырга>" and "<жаз сайга>", indicating the connection of these two clauses without any explicit conjunction or subordinating element.

punct: punctuation

This is applied to any punctuation in a sentence, provided that it is kept in the typed dependencies. Keep in mind that symbols with the symbol SYM cannot be associated using the punct relation since they are not punctuation. With the exception of ellipses, tokens with the relation punct always attach to content words and are never reliant on anything else. Because punct is not a regular dependence connection, it does not meet the standard criterion for becoming the head word. Rather, we apply the subsequent guidelines:

- The following conjunct has a punctuation mark appended to it that separates coordinated components.
- A dependent unit has a punctuation mark added to it before or after it.
- A punctuation mark is affixed at the topmost node that maintains projectivity inside the pertinent unit.
- If not doing so will result in non-projectivity, paired punctuation marks (such as quotes and brackets, occasionally dashes, commas, and other symbols) should be affixed to the same word. Typically, this word is the first in the sentence contained in the paired punctuation.

Punctuations in the Kyrgyz language are as follows: *quotation marks* (“ ”), *apostrophe* (’), *hyphen* (-), *brackets* [], *parentheses* (), *en dash* (/), *em dash* (\), *ellipsis* (...), *comma* (,), *colon* (:), *semicolon* (;), *exclamation mark/exclamation point* (!), *question mark* (?),

full stop/period (.) and others. You may observe the punctuation relation in each annotated sentence above.

root

The **root** grammatical relationship indicates the basis of the phrase. The governor is a ROOT fake node. The ROOT node is indexed with 0, as the indexing of the sentence's tokens starts at 1. The root dependency relation in every tree should have just one node. If there are several orphaned dependents and the main predicate is absent (due to ellipsis), then one of the orphans is elevated to the head (root) position, and the remaining dependents are attached to it (Djumaliev et al., 2023). The root examples are provided in each sentence above.

vocative

The vocative case, also known as the vocative construction, is a syntactic case that is used to explicitly address or call upon someone. It's a type of direct communication. For instance, "Kalys" appears in the vocative case in the statement "Kalys, turn on the light," suggesting that the speaker is speaking to Kalys directly.

In a number of languages, including Kyrgyz, the vocative case is often conveyed by distinctive forms or intonation rather than being separated by separate inflectional endings. The vocative case is typically used in speech or writing to call attention to oneself or to address a specific person directly. It indicates to the intended recipient that the person being addressed is the one who should get it.

The vocative relation is used in texts when addressing a conversation participant (common in talks, dialogue, emails, newsgroup notifications, etc.). The addressee's name and the host phrase are connected. A vocative and a null subject commonly occur together. If the nominal is meant to be vocative, then implement the vocative connection.

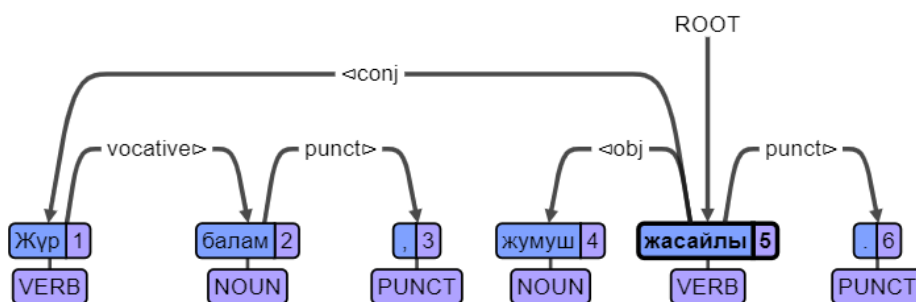


Figure 3. 28. Example of vocative relation in a sentence.

“Жүр балам, жумуш жасайлы.” is an example of a sentence in the Kyrgyz language, containing a vocative relation. Its translation into English: “Come on, my son, let's do some work.”

"<Жүр>"

"Жүр" VERB @conj #1->5

"<балам>"

"балам" NOUN @vocative #2->1

"<,>"

"," PUNCT @punct #3->2

"<жумуш>"

"жумуш" NOUN @obj #4->5

"<жасайлы>"

"жасайлы" VERB @root #5->0

"<.>"

"." PUNCT @punct #6->5

Here is the explanation of a sentence that contains a vocative in it based on CG-3 format:

#1->5: The verb "<жүр>" (to walk) is connected to the verb "<жасайлы>" (to do) in a conjunctive relation.

#2->1: The noun "<балам>" (my son) is vocatively addressed in connection with the verb "<жүр>" (to walk).

#3->2: The punctuation mark "<,>" follows the noun "<балам>" (my son).

#4->5: The noun "<жумуш>" (work) appears to be the object of the verb "<жасайлы>" (to do).

#5->0: The verb "<жасайлы>" (to do) is the root of the sentence.

#6->5: The punctuation mark "<.>" follows the verb "<жасайлы>" (to do), marking the end of the sentence.

Deduction to Chapter 3

In this chapter, we first examined that the Universal Dependencies project uses a special list of symbols characteristic of annotating sentences in any world language. We analyzed that in UD, each constituent of a sentence, either a word, a numeral, or a punctuation, is a token of that sentence. The tokens are numbered in order, e.g. as 1, 2, 3, etc. Each token is connected to another token to form a syntactic dependency. The sentence necessarily has a root, which is marked by the numeral 0. Often the root is the predicate itself. After we broke down the symbols for major and minor parts of speech, we listed all 37 existing syntactic dependencies in UD.

Next, we tried to give an alphabetical definition of all these dependencies with examples in Kyrgyz. All of the above examples of sentences were taken from the existing corpus of the Kyrgyz language. Several of these dependencies do not exist in the Kyrgyz language due to the lack of some grammatical features. With illustrative examples of a bank of tree structures, we tried to give a thorough analysis of special syntactic dependencies. Each tree structure was made on the UD Annotatrix open-access website, which was created by Associate Professor of Linguistics at Swarthmore College Jonathan North Washington. Each dependency was given a treebank structure and its CG-3 format. We attempted to provide a meticulous analysis of each treebank, breaking it down into components and their relations.

CONCLUSION

It is essential to develop syntactic treebank structures in Kyrgyz inside the Universal Dependencies (UD) for a number of reasons. It is essential to create syntactic treebank structures in the Kyrgyz language for cross-linguistic research, linguistic documentation, language comprehension, and Natural Language Processing applications. It advances and preserves linguistic variety, which is a larger purpose, in addition to improving our knowledge of Kyrgyz syntax. Establishing a Kyrgyz language treebank helps with the language's linguistic historical documentation and preservation. The nuances of Kyrgyz syntax can be preserved for future generations by linguists and language lovers through methodical analysis and annotation of Kyrgyz written works. Syntactic treebanks offer an organized depiction of the grammatical relationships and attributes found in a language. By creating a treebank for Kyrgyz, we may better comprehend the language's syntactic structures and do more in-depth linguistic research, including computer modeling, dependency analysis, and syntactic parsing.

Creating and enhancing NLP tools and algorithms specifically for the Kyrgyz language can be greatly aided by making use of syntactic treebanks. Accurate syntactic annotations supplied by the treebank are useful for applications like information retrieval, text summarization, sentiment analysis, and machine translation.

The inclusion of Kyrgyz in the Universal Dependencies project facilitates research and cross-linguistic comparison. Researchers can examine linguistic universals and variations, compare Kyrgyz with other languages, and find typological trends by aligning Kyrgyz syntactic structures with the universal criteria set by UD.

The first chapter is devoted to the analysis of the significance and history of this research work. The importance and need to create a bank of tree structures for the Kyrgyz language using a standardized framework were described. The first chapter describes the main reasons for improving the Kyrgyz language through this project, and it also reveals the relevance, goals, and novelty of this work.

The second chapter is devoted to a review of the existing literature. Dependency syntactic analysis is fundamental in modeling the structure of a language. This chapter describes the grammatical features of syntactic and dependency analysis. The rules of phrasal structures and compounds are considered based on the research of N. Chomsky, M. C. de Marneffe, and other outstanding linguists. The works on syntax by Roman

Jacobson, Charles Fillmore, and Daniel Jurafsky were also reviewed. The first steps and algorithms for automated sentence parsing were studied. We studied the important algorithms for detailed sentence parsing based on linguist Nivre and step-by-step rules created by him were discussed. Additionally, we disassembled the Constituency-based treebanks and studied tree structures based on dependency relations. In this chapter, the topic of the importance of using tree structures in computational linguistics was revealed.

The third chapter explores the Universal Dependencies project, which applies universal symbols to annotate sentences in any language. It identifies 37 existing syntactic dependencies in UD, which are connected to form a sentence. This chapter provides an alphabetical definition of these dependencies, with examples from the Kyrgyz language corpus. We also tried to provide a thorough analysis of tree structures, created on the UD Annotatrix open-access website, and provide a detailed breakdown of each dependency's components and relations. This paper developed a treebank of syntactic structures for the Kyrgyz language, aligning its structures with universal standards. This research facilitated an understanding of Kyrgyz syntax and NLP tools. The treebank ensured interoperability and comparability with other languages, fostering collaboration. Challenges included complexities of Kyrgyz syntax and scarcity of resources. Future efforts should expand and refine the treebank to enhance its utility and accuracy. This dissertation contributes to linguistic typology, cross-linguistic research, and multilingual natural language processing.

Overall, we truly believe that the outcomes of the paper may contribute to the future language research. The syntactic treebank enhances linguistic resources for the Kyrgyz language by providing a structured dataset of annotated texts. It improves language understanding by allowing detailed analysis of grammatical patterns and syntactic constructions. It also empowers Natural Language Processing (NLP) tools for Kyrgyz.

The Kyrgyz treebank aids in computational research, enhancing machine learning, linguistics, and artificial intelligence. It integrates Kyrgyz into global standards, promoting its visibility. The treebank also empowers language development initiatives, aiding teachers, curriculum developers, and activists in preserving and revitalizing the Kyrgyz language. We hope that the treebanks provided in this paper aid in linguistic research by enabling comparative studies, typological analysis, and theoretical investigations into the structure and properties of the Kyrgyz language.

BIBLIOGRAPHY

“Kyrgyz Language Corpora”, Saarland University, the Department of Language and Science (2019), <https://corpora.clarin-d.uni-saarland.de/cqpweb/>.

“Universal Dependencies”, 2004. Available: <https://universaldependencies.org/>

Aizhan Satybekova, and Aida Kasieva, “Parts-of-Speech Annotation of the Newly Created Kyrgyz Corpus (Turkic Lexicon Apertium Tools)”, (Vestnik KRSU, 2020), <http://vestnik.krsu.edu.kg/archive/154/6520>

Akesh Imanov, “Syntax of the Kyrgyz language”, Bishkek, 2009

Akunova A., Raimbekova M., Karamendeeva Ch., “Contemporary Kyrgyz Language. Syntax”, 2010

Amantur Japarov, “Kyrgyz language syntax”; Frunze, 1979

Anna Thompson, “Syntactic Parallelism and Structure in Kyrgyz Proverbs”, (Masters dissertation paper, Bryn Mawr College) Pennsylvania, 2021

Attokurov A., Osh, 2006, “Syntax of the Kyrgyz language”

Ben Lutkevich, “Natural Language Processing (NLP)”, 2021, DOI:10.1109/ICPCSN58827.2023.00238

Bernd Bohnet, (Beijing, 2010). “Very High Accuracy and Fast Dependency Parsing is not a Contradiction”, file:///C:/Users/musaz/Downloads/Very_high_accuracy_and_fast_dependency_parsing_is_.pdf

Dan Jurafsky and James H. Martin, “Speech and Language Processing”, 2021.

Daniel Zeman et al., “CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies”, 2018, DOI 10.18653/v1/k17-3021

David M. Perlmutter, “Disjoint Reference and Backward Anaphora”, In Language and Cognitive Style: Papers in Honor of David McNeill, 1983

Elena Grudeva, “Corpus Linguistics”, 2012, https://www.academia.edu/24711344/Grudeva_e_v_korpusnaya_lingvistika

Elena Lyukina, “*Using Universal Dependencies in a grammatical analysis of a multilingual text (using the example of an impersonal predicate)*”, 2018, DOI 10.25205/1818-7935-2018-16-2-19-33

Eugene Charniak, “*Immediate-head parsing for language models*”, 2001, DOI 10.3115/1073012.1073029

Gulnura Jumalieva, Aida Kasieva, and Seide Musazhanova; “*Adaptation of the terms of the web project Universal Dependencies into the Kyrgyz language*”; Vestnik KRSU, 2023), DOI 10.36979/1694-500X-2023-23-6-71-75

Joakim Nivre et al., “*Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection*”, 2004. <https://arxiv.org/pdf/2004.10643>

Joakim Nivre, 2003, “*An Efficient Algorithm for Projective Dependency Parsing*”. In Proceedings of the Eighth International Conference on Parsing Technologies, pages 149–160, Nancy, France. <https://aclanthology.org/W03-3017>

Joakim Nivre, Johan Hall, and Jens Nilsson, “*MaltParser: A Data-Driven Parser-Generator for Dependency Parsing*”, 2006, http://www.lrec-conf.org/proceedings/lrec2006/pdf/162_pdf.pdf

Jonathan N. Washington, “*UD Annotatrix*”, 2021, <https://jonorthwash.github.io/ud-annotatrix/src/server/public/html/annotatrix.html#1>

Julia Hockenmaier, and Mark Steedman, “*CCG bank: A corpus of CCG derivations and dependency structures extracted from the Penn treebank*”, 2007, DOI 10.1162/coli.2007.33.3.355

Kristina Toutanova, and Christopher D. Manning, “*Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger*”, 2000, DOI 10.3115/1117794.1117802

Kurmanbek Toktonaliev and Taalaibek Tokoev, “*Contemporary Kyrgyz language (Syntax)*”, Bishkek, 2015

Masato Hagiwara, “*Real-World Natural Language Processing*”, 2021.

McEnery et al., “*Universal Dependencies*”, Columbus: Creative Commons Attribution-NonCommercial-NoDerivatives, 2021.

- Meishan Zhang and Yue Zhang, “Combining discrete and continuous features for deterministic transition-based dependency parsing”, 2015, DOI 10.18653/v1/d15-1153
- Michael Collins and Nigel Duffy, USA, 2002. “New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron”, DOI 10.3115/1073083.1073128
- Michael Collins, “Head-driven Statistical Models for Natural Language Parsing”, 2003, DOI 10.1162/089120103322753356
- Michael Weiss, Alexander Stern, and Yoav Goldberg, “Structured Training for Large-Scale Syntactic Distillation”, 2015, DOI 10.4324/9781315664163-23
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. “Building a Large Annotated Corpus of English: The Penn Treebank”. Computational Linguistics, 19(2):313–330. <https://aclanthology.org/J93-2004>
- Nathan David Green, “Dependency Parsing”, 2011, DOI 978-80-7378-184-2
- Niladri Sekhar Dash, “Language Corpora Annotation and Processing”, 2021, DOI 10.1007/978-981-16-2960-0
- Noam Chomsky, “Aspects of the Theory of Syntax”, 1965
- Noam Chomsky, “Review of B. F. Skinner's Verbal Behavior”, 1959
- Noam Chomsky, “Syntactic Structures”, 1957
- Ryan McDonald and Fernando Pereira, “Online Learning of Approximate Dependency Parsing Algorithms”, 2006, <https://api.semanticscholar.org/CorpusID:802998>
- Ryan McDonald, “Discriminative dependency parsing with support vector”, 2006
- Ryan McDonald, Kevin Lerman, and Fernando Pereira, “Multilingual Dependency Analysis with a Two-Stage Discriminative Parser”, 2006, DOI 10.3115/1596276.1596317
- Sandra Kübler, Ryan McDonald, and Joakim Nivre, (2009). “Dependency Parsing”. DOI 10.1007/978-3-031-02131-2_2
- Seide Musazhanova, Aida Kasieva, and Gulnura Jumalieva; “Syntactic Annotation of the Newly-created Kyrgyz Corpus”, Vestnik of IKSU, 2023, DOI 811.512.154

Simon Šuster, Madhumita Sushil, and Walter Daelemanson, “*Revisiting neural relation classification in clinical notes with external information*”, 2018, DOI 10.18653/v1/w18-5603

Takuya Matsuzaki, Yusuke Miyao, and Jun’ichi Tsujii, “*Probabilistic CFG with Latent Annotations: More Right-branching Parsing with Less Training Data*”, 2005, DOI 10.3115/1219840.1219850

Tony McEnery and Andrew Wilson, “*Corpus Linguistics*”, 2001.

Tynybek Nuruev, “*Kyrgyz language syntax*”; Bishkek, 2005

Veronika Garshina and Julia Bogoyavlenskaya, “*Development of a linguistic parser for the Russian language*”, 2012, DOI 004.912

Walter de Gruyter, “*Corpus Linguistics and Linguistic Theory*”, 2020, DOI 10.1515/cllt

Yue Zhang and Joakim Nivre, “*Transition-Based Dependency Parsing with Rich Non-Local Feature*”, 2011, <https://aclanthology.org/P11-2033>

