



**КЫРГЫЗ - ТҮРК “МАНАС” УНИВЕРСИТЕТИ ТАБИГЫЙ
ИЛИМДЕР ИНСТИТУТУ КОМПЬЮТЕРДИК ИНЖЕНЕРИЯ
БӨЛҮМҮ**

**ТЕКСТТИ АВТОМАТТЫК ТҮРДӨ ИШТЕТҮҮДӨ
ЛЕММАТИЗАЦИЯ МАСЕЛЕЛЕРИ**

**Даярдаган
Муканбетова Айкерим Муканбетовна**

**Илимий жетекчи
доц.др. Бакыт Шаршембаев**

Магистрдик диссертация

**Май 2022
БИШКЕК, КЫРГЫЗСТАН**

ИЛИМИЙ ЭТИКАГА ШАЙКЕШТИГИ

Бул диссертацияда колдонулган бардык маалыматтар академиялык жана этикалык эрежелерге жооп берет. Ошондой эле, башка илимий булактардан алынган маалыматтарга “АДАБИЯТ” бөлүмүндө шилтемелер көрсөтүлдү.

Айкерим Муканбетова

Колу:

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.

Aykerim Mukanbetova

İmza:

ЖОЛ-ЖОБОГО ШАЙКЕШТИГИ

“Текстти автоматтык түрдө иштетүүдө лемматизация маселелери“ аталышындагы магистрдик диссертация, Кыргыз-Түрк “Манас” университетинин магистрдик диссертация жазуу жол-жобосуна ылайык даярдалды.

Муканбетова Айкерим

доц.др. Бакыт Шаршембаев

Колу

Колу

Компьютердик инженерия бөлүм башчысы

Доц. др. Райымбек Султанов

Колу

YÖNERGEYE UYGUNLUK

“Otomatik metin işlemede lemmatizasyon problemleri“ adlı Yüksek Lisans Tezi, Kırgızistan-Türkiye Manas Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazım Yönergesi’ne uygun olarak hazırlanmıştır.

Aykerim Mukanbetova

Doç.Dr.Bakıt Şarşembaev

İmza

İmza

Bilgisayar Mühendisliği ABD Başkanı

Doç. Dr. Rayimbek SULTANOV

İmza

КАБЫЛ АЛУУ ЖАНА ЧЕЧИМ

Доц.др. Бакыт Шаршембаевдин жетекчилиги астында Муқанбетова Айкерим тарабынан даярдалган “ Текстти автоматтык түрдө иштетүүдө лемматизация маселелери ” темасындагы диссертация, калыстар тарабынан Кыргыз-Түрк “Манас” университети, Табигый илимдер институту Компьютердик инженерия бөлүмүндө Магистрдик диссертация катары кабыл алынды.

..... – – 2022

КАЛЫСТАР:

Төрагасы	ф.и.к., доц. Рысбек Исмаков	
Илимий жетекчи	ф.и.к., доц. Шаршембаев Б. Д.	
Мүчө	т.и.д., проф. Бримкулов У.Н.	
Мүчө	ф.и.к., доц. Замиргүл Казакбаева	
Мүчө	Док. Жумабаева Чынара Н.	

..... – – 2022

Доц. Док. Исмет Алтынташ
Институт мүдүрүнүн м.а.

KABUL VE ONAY

Doç. Dr. Bakıt Şarşembaev danışmanlığında Aykerim Mukanbetova tarafından hazırlanan “ Otomatik metin işlemede lemmatizasyon problemleri” adlı bu çalışma, jürimiz tarafından Kırgızistan-Türkiye Manas Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Ana Bilim Dalında Yüksek Lisans Tezi olarak kabul edilmiştir.

... 2022

JÜRİ:

Jüri başkanı	Doç.Dr. Rısbek İSKAKOV	
Danışman	Doç.Dr. Bakıt ŞARŞEMBAYEV	
Üye	Prof.Dr. Ulan BRİMKULOV	
Üye	Doç.Dr. Zamirgül KAZAKBAEVA	
Üye	Öğr.Gör.Dr. Çınara CUMABAYEVA	

..... – – 2022

Doç. Dr. İsmet ALTINTAŞ
Enstitü Müdür V.

ЫРААЗЫЧЫЛЫК

Диссертацияны жазууда ар кандай кыйынчылыктар кезиккен учурларда маселелерди чечүүдө көмөк көрсөткөн, практикалык жана теориялык бөлүмүн жазууда чоң салым кошкон, ар дайым колдоп келген доц.др. Бакыт Шаршембаев агайыма жана ошондой эле ар тараптуу илимий кеңештерин аябаган доц.др Райымбек СУЛТАНОВ агайыма, терең ыраазычылык билдиремин.

Айкерим Муканбетова

Бишкек , Май 2022

**ТЕКСТТИ АВТОМАТТЫК ТҮРДӨ ИШТЕТҮҮДӨ ЛЕММАТИЗАЦИЯ
МАСЕЛЕЛЕРИ**

Муканбетова Айкерим

**КЫРГЫЗ-ТҮРК «МАНАС» УНИВЕРСИТЕТИ, ТАБИГИЙ ИЛИМДЕР
ИНСТИТУТУ**

МАГИСТРДИК ИШ, 2022

Илимий жетекчи: доц. др. Бакыт Шаршембаев

АННОТАЦИЯ

Илимдин, техниканын жана маданияттын өнүгүшүнүн азыркы этабында коомдук өндүрүштүн процессинде алдыңкы орунду ээлеген жана адамдын ишинин бардык чөйрөлөрүнө кирип жаткан информацияны иштеп чыгуу процесстерине артыкчылык берилип жатат. Адамзатка ыңгайлуулукту түзүү үчүн маалыматты табигый тилде иштетүү каражаттарына көбүрөөк маани берилүүдө. Бүгүнкү күндө табигый тилди иштетүү (Natural Language Processing NLP) көптөгөн тармактарда колдонулат, анын ичинде үн жардамчылары, автоматтык текст котормолору жана текстти филтрлөө. Негизги үч багыт болуп төмөнкүлөр саналат: кеп таануу (Speech Recognition), табигый тилди түшүнүү (Natural Language Understanding) жана табигый тилдин өндүрүлүшү (Natural Language Generation). [1]

Көптөгөн ушул сыяктуу системаларда табигый тилди иштетүүнүн негизин морфологиялык анализ түзөт. Ал эми илимпоздор буга чейин кеңири таралган тилдер үчүн табигый тилди иштетүүчү көптөгөн системаларды иштеп чыгышса да, кыргыз тили үчүн көлөмдүү базасы бар системалар жок жана бул иштин актуалдуулугун көрсөтүп турат.

Бул изилдөөдө кыргыз тилинин жетиштүү базасы бар жаңы анализатору сунушталган. Колдонуучулук жана жеткиликтүүлүк жагынан окшош системаларда бир катар анализдер жүргүзүлдү жана алардын натыйжалары бул анализаторду жакшыртуу үчүн колдонулган.

Изилдөөнүн объектиси болуп С.Каралаев, С.Орозбаков, Ж.Мамай Манас, Семетей, Сейтек эпосторундагы тексттеринин корпусу алынган.

Изилдөөнүн предмети - текстти автоматтык лемматизациялоо жана аны

автоматтык түрдө иштетүү, сөз формаларын морфологиялык анализдөө ыкмасы болуп саналат.

Максатка жетүү үчүн төмөнкү маселерлерди чечүү зарылчылыгы жаралган:

1. Текстти автоматтык иштетүүнүн (АОТ) [2] заманбап чет элдик ыкмаларын талдоо.
2. «Электрондук текст» жана «тексттердин корпусу» түшүнүктөрүн карап көрүү.
3. С.Каралаев, С.Орозбаков, Ж.Мамай Манас, Семетей, Сейтек эпосторундагы тексттеринин корпусун түзүү.
4. Сөз формаларын баштапкы сөз формасына келтирүү эрежелерин талдоо.
5. С.Каралаев, С.Орозбаков, Ж.Мамай Манас, Семетей, Сейтек эпосторундагы тексттердин корпусун автоматтык түрдө иштеп чыгуу.

Ачкыч сөздөр: Лемматизация, сөздүн баштапкы формасы, компьютердик лингвистика, автоматтык түрдө текстти иштеп чыгуу, корпус, морфологиялык талдоо, токенизация.

OTOMATİK METİN İŞLEMEDE KÖK ÇÖZÜMLEME PROBLEMLERİ

Aykerim MUKANBETOVA

**KIRGIZİSTAN-TÜRKİYE MANAS ÜNİVERSİTESİ, FEN BİLİMLERİ
ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ, 2022

Danışman: Doç. Dr. Bakıt ŞARŞEMBAEV

GENİŞ ÖZET

Dil, insanlar arasında anlaşmayı sağlayan bir iletişim aracıdır. Sosyal yaşamın en büyük gereksinimlerinden bir tanesi dildir. Dünya’da halihazırda konuşulmakta olan 7000’in üzerinde dil olması bu kavramın ne kadar önemli olduğu gösterilmiştir.

Son zamanlarda hayatımızın vazgeçilmez bir parçası olan bilgisayar dünyasında ise iki farklı dil vardır. Bunlar makine dilleri olan programlama dilleri ve doğal dillerdir. İnsanların kullandığı diller doğal dil olarak adlandırılır. Makine ile insan arasındaki iletişimin bağının arttığı bu günlerde, blog yazıları, sosyal medya mesajları ve içerikleri ile metin verileri artarak büyümeye devam etmekte, bu verilerden anlamlı ve istenilen bilgilerin bilgisayarın anlayabileceği biçimde çıkarılması için Doğal Dil İşleme tekniği kullanılmaktadır. Bu teknik, insan dilini bilgisayarın anlayabileceği bir formata dönüştürme işlemini yapmaktadır. Dilbilim ve yapay zekanın bir alt dalı olan Doğal Dil İşleme, genellikle Natural Language Processing yani NLP olarak bilinmektedir.

Ana dilde konuşmak her ne kadar kolay gibi görünse de, dil edinimi gerçekten de zorlayıcı ve zaman alan bir süreçtir. Farklı bir yabancı dili ne gibi süreçler atlatarak öğrendiğinizi ya da öğrenemediğinizi düşünebilirsiniz. Ayrıca dilin canlı bir organizma gibi düşünülmesi gerekir. Gündelik dilde bile değişen zamanla birlikte değişen birçok kelime ve durum ifadesi vardır. Doğal dili anlama ve doğal dilde düşünebilme süreci beyin içinde karmaşık bir yapıya sahipken, bunu hemen anlayarak bilgisayar ortamında yazılıma dökkebilmek için de çok fazla mesai harcanmıştır. 50 yılı geçen bir süre sonucunda yapay zeka modüllerinin geliştirilmesi ile doğal dil işleme ortaya çıkmıştır.

Doğal dil işleme, insan dillerinin otomatik hesaplama olduğundan, sürece insan tarafından üretilen metni girdi olarak alan algoritmalar ve çıktılar olarak doğal görünümlü metinler üreten algoritmalar dahildir.

Diyalog temelli bir yapay zeka uzman sisteminden karar almak istediğinizde,

robotun sizin talimatlarınız doğrultusunda işlemleri gerçekleştirmesi gibi akıllı bir sistemin çalışması için sizin konuştuğunuz doğal dilin işlenmesi gerekir.

NLP alanı, bilgisayarları insanların kullandığı doğal dillerle yararlı görevleri yerine getirmeyi gerektirir. NLP'nin girdisini doğal konuşma, çıktısını da yazılı metin oluşturur. NLP'nin temel iki bileşeni vardır. Bunlardan ilki NLU olarak kısaltılan doğal dil anlayışıdır. Anlama şu görevleri içerir; verilen girdinin doğal dilde yararlı temsillerle eşlenmesi ve dilin farklı yönlerini analiz etmek. Diğer bileşen ise NLG kısaltması ile anılan doğal dil üretimidir.

NLG, bazı iç temsillerden doğal dil biçiminde anlamlı söz öbeklerini ve cümleleri üretme sürecidir. Bu süreçte, yapay zeka tarafından içerik algılanır, ilgili içeriği bilgi tabanından alınması sağlanarak metin planlaması gerçekleşir. Planlamaya göre istenilen kelimeler seçilir ve cümlelerin oluşturulmasıyla cümlenin tonu belirlenmeye başlanır. Ardından cümle planının cümle yapısıyla eşleştirilmesi gerçekleşir.

Yapay zekanın iki bileşeninden NLU [2], NLG'den daha zordur. Doğal dil anlayışındaki zorluklar, doğal dilin son derece zengin bir yapıya sahip olmasından ileri gelmektedir. Daha önce doğal dil bölümünde detaylıca açıkladığımız nedenlerden dolayı dilin farklı seviyelerinde belirsizlik olabilir. Sözcüksel belirsizlik, kelime seviyesi gibi çok ilkel düzeyde olabilir.

Örneğin “yaz” kelimesinin isim olarak mı yoksa fiil olarak mı kullandığının algılanması zor olabilir. Söz dizimi seviyesinde “Doktor ona baktı.” gibi bir cümlede doğru tonlama yapılmadıysa ses algısında ayrıştırma zor olabilir. Özellikle Türkçe ve Kırgızca gibi eril/dişil ayrımı olmayan dillerde zamir belirsizliği yaşanabilir. Bunların önüne geçebilmek için işlenen dilin kodları maksimum derecede detaylandırılarak işlenmelidir.

Cümlelerin ve söz öbeklerinin kodlandığı dil bilimsel ağaçlar ne kadar doğru çizilirse sonuçlar da o kadar güvenilir olur. Ayrıştırma ağacı, cümleyi yapılandırılmış parçalara ayırır, böylece bilgisayar dili kolayca anlayabilir ve işleyebilir. Ayrıştırma algoritmasının bu ayrıştırma ağacını oluşturması için yapay zeka kodlamasında hangi ağaç yapılarının doğru olduğunu açıklayan bir dizi yeniden yazma kuralları oluşturulmalıdır.

Doğal Dil İşlemede Sosyal Faydalar çoktur. Şirketlerin verileri işlemesine yardımcı olmanın yanı sıra, duygu analizi de toplumu anlamamıza yardımcı olmaktadır, bu bir nevi genel müşteri profilinin tanınması için şirketlere de dolaylı fayda sağlar

diyebiliriz. Örneğin, Periscope, ABD Başkanı Donald Trump'ın duygusal durumunu izlemek için dil ve yüz ifadelerini işleyen bir veri motoru olan Trump-Emoticoaster'ı oluşturmak için görsel tanıma ile eşleştirilmiş NLP'ye sahiptir. Benzer teknoloji okullardaki şiddet olaylarını da engelleyebilir: Columbia Üniversitesi'nde, araştırmacılar risk altındaki 9.000 genç üzerinden kaydedilen 2 milyon tweet'i işleyerek "Gençlerdeki dil değişimi nasıl onları şiddet olaylarına yaklaştırıyor?" gibi bir soruyu cevaplandırdılar. Böylece sosyal hizmet uzmanları, risk altındaki gençler için tedbirler alabiliyor.

Doğal dil işleme, kendi duygusal durumunuzu izlemenize de yardımcı olabilir. Woebot, bir Facebook Messenger, chatbot veya tek başına bir uygulama ile kullanıcılara bağlanan bir elektronik terapisttir. Henüz çok yüksek hassasiyetlere sahip olarak çalışmamaktadır. Ancak Woebot, kullanıcılarının endişe ve depresyon düzeylerini takip edebilmektedir.

Doğal dil işleme, aynı zamanda iş arayanların öz geçmişleri arasında çeşitlendirmelerine, farklı adayları çekmelerine ve daha nitelikli çalışanlar işe almasına yardımcı olur. Bu, metin analizi gerektiren türden bir dil çözümlemesidir ve bu sayede örneğin X üniversitesinden mezun olan adaylar ya da Y dilini bilen kişiler öncelik sırasındaysa, gelen CV'lerde yer alan metinleri tarayarak onları sizin için düzenleyebilir.

Spam algılama, istenmeyen e-postayı gelen kutunuzdan uzak tutmak için de aynı mantıkla NLP kullanılabilir. Outlook ve Gmail gibi programlar, metinleri algılayarak belirli kişilerden gelen mesajları oluşturduğunuz klasörlere ayırır. Örneğin tanıtımlar, sosyal medya ya da birincil klasörleri oluşturur ve sizin için hemen her şeyi derleyip toplayarak işlerinizi daha hızlı halletmenize yardım eder.

İş hayatında duyarlılık analizi gibi araçlar, şirketlerle ilgili tweet'lerin iyi ya da kötü olup olmadıklarını hızlı bir şekilde fark etmelerine yardımcı olur, böylece müşteri memnuniyeti araştırması ve marka itibarına dair analizler yapılabilir. Analiz firması Periscope'ten veri görselleştiricisi Skye Morét, İngilizce kelimelerin %30'unun pozitif olduğunu söylüyor. Geri kalanlarsa nötr ya da negatif diyerek ekliyor ve bu noktada NLP dönüştürücülerinin ifadelerle, giflerle durumun daha iyi analiz edildiğini ekliyor. Evet, bugüne kadar geleneksel olarak, şirketler geri bildirim olumlu veya olumsuz olarak sınıflandırmak için doğal dil işlemeyi kullandılar. Ancak, FleishmanHillard'daki sosyal ve inovasyondan sorumlu kıdemli başkan yardımcısı Ryan Smith, bugünün araçlarının yapay zeka sayesinde en fazla kullanılan üzüntü, öfke ve korku gibi daha kesin duyguları

tanımladığını ve bunların marka için büyük değer taşıdığını söylüyor.

Doğal dil işleme, şirketler tarafından dokümantasyon süreçlerinin verimliliğini artırmak, belgelerin doğruluğunu geliştirmek ve büyük veri tabanlarından en uygun bilgileri belirlemek için kullanılabilir. Örneğin, bir hastane hekimin yapılandırılmamış notlarından belirli bir teşhis almak ve bir fatura kodu atamak için bile doğal dil işlemeyi kullanabilir. Özetle doğal dil işlemenin yararları sayısızdır ve hayatın hemen her anındadır.

Bir çok bunun gibi sistemlerde doğal dili işlemenin temelini morfolojik analiz oluşturur. Ancak, bilim adamları bugüne kadar ortak diller için Doğal Dil İşlemesini yapabilen bir çok sistemleri geliştirmiş olsa bile, Kırgız dili için hacimsel bir tabana sahip sistemler yok ve bu işin gerekliliğini göstermekte.

Bu araştırmada kırgız dilinin yeterli tabana sahip yeni analizör sunulmakta. Kullanıcı dostu ve erişebilirlik açısından benzer sistemler incelendi ve bu incelemelerin sonucu bu analizörü iyileştirmede kullanıldı.

Araştırmanın nesnesi olarak Sayakbay Karalaryev, Sagınbay Orozbekov, Cusup Mamay'ın söylemiş olduğu MANAS, SEMETAY, SEYTEK destanında bulunan metinleri alındı.

Araştırmanın konusu, metnin otomatik lemmatizasyonu ve otomatik işlenmesi, kelime biçimlerinin morfolojik analiz yöntemidir.

Bu amaca ulaşmak için aşağıdaki konuların ele alınması gerekir:

1. Diğer diller için kullanılan modern Kelime İşlem Yöntemlerinin analizi.
2. "Elektronik metin" ve "metin gövdesi" kavramlarının incelenmesi.
3. S. Karalaev, S. Orozbekov, J. Mamai Manas, Semetey, Seitek destanlarındaki metinlerin gövdesini oluşturmak.
4. Kelime formlarını orijinal kelime formlarına indirgeme kurallarının analizi.
5. S. Karalaev, S. Orozbekov, J. Mamai Manas, Semetey, Seitek destanlarındaki metinlerin gövdesinin otomatik gelişimi.

Anahtar kelimeler: Lemmatizasyon, kelimenin temel biçimi, bilgisayar dilbilimi, otomatik kelime işleme, derlem, morfolojik analiz, simgeleştirme.

PROBLEMS OF LEMMATIZATION IN AUTOMATIC TEXT PROCESSING

Aikerim MUKANBETOVA

**KYRGYZ-TURKISH MANAS UNIVERSITY, GRADUATE SCHOOL OF
NATURAL AND APPLIED SCIENCES**

M.S.Thesis, 2022

Supervisor: Assoc. Prof. Bakyt SHARSHEMBAEV

ABSTRACT

Emphasis is placed on information processing processes, which stand out at the current level of the development of science, technique and culture, and which have entered every business area of humanity. In order to provide convenience to humanity, more importance is given to the means of processing information in natural language. Today, natural language processing (NLP) is used in many fields, such as voice control, voice assistants, automatic text translators, text filtering systems. This scientific research area consists of three main aspects: Speech Recognition, Natural Language Understanding and Natural Language Generation.

In many such systems, morphological analysis is the basis for natural language processing. However, although scientists have developed so far many systems capable of Natural Language Processing for common languages, systems with a volumetric base for the Kyrgyz language do not exist and this shows the necessity of the work.

In this research, a new analyzer of Kyrgyz language with sufficient base is presented. Similar systems in terms of user friendliness and accessibility were reviewed and the results of these reviews were used to improve this analyzer.

As the object of the research, the body of the texts in the MANAS, SEMETEV, SEYTEK epic, which were said by Sayakbay Karalaryev, Saginbay Orozbekov, and Cusup Mamay, was taken.

The subject of the research is the automatic lemmatization and automatic processing of the text, the method of morphological analysis of word forms.

To achieve this goal, the following issues need to be addressed:

1. Analysis of modern Word Processing Methods used for other languages.
2. Examining the concepts of "electronic text" and "text body".

3. To form the body of texts in the epics of S. Karalaev, S. Orozbekov, J. Mamai Manas, Semetei, Seitek.
4. Analysis of the rules for reducing word forms to original word forms.
5. Automatic development of the body of texts in the epics of S. Karalaev, S. Orozbekov, J. Mamai Manas, Semetei, Seitek.

Keywords: Lemmatization, primary form of word, computer linguistics, automatic word processing, corpus, morphological analysis, symbolization.



ПРОБЛЕМЫ ЛЕММАТИЗАЦИИ ПРИ АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТА

Муқанбетова Айкерим

**КЫРГЫЗКО-ТУРЕЦКИЙ УНИВЕРСИТЕТ «МАНАС», ИНСТИТУТ
ЕСТЕСТВЕННЫХ НАУК**

МАГИСТРСКАЯ ДИССЕРТАЦИЯ, 2022

Научный руководитель: Доц. к.т.н. Бакыт Шаршембаев

АННОТАЦИЯ

На современном этапе развития науки, техники и культуры приоритет отдается процессам обработки информации, которые занимают ведущее положение в процессе общественного производства и проникают во все сферы деятельности человека. Больше внимания уделяется средствам обработки информации на естественном языке для удобства человечества. Сегодня обработка естественного языка (NLP) используется во многих отраслях, включая голосовые помощники, автоматический перевод текста и фильтрацию текста. Три основные области: распознавание речи, понимание естественного языка и генерация естественного языка.

Во многих таких системах морфологический анализ является основой обработки естественного языка. Хотя учеными уже разработано множество систем обработки естественного языка для наиболее распространенных языков, систем с большой базой для кыргызского языка нет, и это говорит об актуальности работы.

В данном исследовании был предложен новый анализатор с достаточной базой кыргызского языка. Был проведен ряд анализов аналогичных систем с точки зрения удобства использования и доступности, и их результаты были использованы для улучшения этого анализатора.

Объектом исследования является корпус текстов в эпосах Манас, Семетей, Сейтек С. Каралаева, С. Орозбакова, Дж. Мамай.

Предметом исследования является метод автоматической лемматизации текста и его автоматической обработки, морфологический анализ словоформ.

Для достижения поставленной цели необходимо решить следующие вопросы:

1. Анализ современных зарубежных методов обработки текстов (АОТ) [3].
2. Рассмотреть понятия «электронный текст» и «корпус текстов».
3. Создание корпуса текстов в эпосах Манас, Семетей, Сейтек С. Каралаева, С. Орозбакова, Дж. Мамай.
4. Анализ правил приведения словоформ к исходной словоформе.
5. Автоматическое развитие корпуса текстов в эпосах Манас, Семетей, Сейтек С. Каралаева, С. Орозбакова, Дж. Мамай.

Ключевые слова: лемматизация, начальная форма слова, компьютерная лингвистика, автоматическая обработка текстов, корпус, морфологический анализ, токенизация.

МАЗМУНУ

ТЕКСТТИ АВТОМАТТЫК ТҮРДӨ ИШТЕТҮҮДӨ ЛЕММАТИЗАЦИЯ МАСЕЛЕЛЕРИ

ИЛИМИЙ ЭТИКАГА ШАЙКЕШТИГИ.....	II
BİLİMSEL ETİĞE UYGUNLUK	II
ЖОЛ-ЖОБОГО ШАЙКЕШТИГИ.....	III
YÖNERGEYE UYGUNLUK.....	III
КАБЫЛ АЛУУ ЖАНА ЧЕЧИМ	IV
КАЛЫСТАР:.....	IV
KABUL VE ONAY	V
JÜRİ:	V
ЫПРАЗЫЧЫЛЫК.....	VI
ТЕКСТТИ АВТОМАТТЫК ТҮРДӨ ИШТЕТҮҮДӨ ЛЕММАТИЗАЦИЯ МАСЕЛЕЛЕРИ.....	VII
АННОТАЦИЯ	VII
GENİŞ ÖZET	IX
ABSTRACT	XIII
АННОТАЦИЯ	XV
МАЗМУНУ	XVII
ЖАДЫБАЛДАРДЫН ТИЗМЕСИ	XIX
КЫСКАРТМАЛАР ТИЗМЕСИ.....	XX

I БӨЛҮМ

КИРИШҮҮ

1.1. Текстти автоматтык түрдө иштетүү системалары тууралуу жалпы маалымат	1
1.2. Текстти автоматтык түрдө иштетүү системаларынын түрлөрү.....	1
1.3. Текстти автоматтык түрдө иштетүүдөгү текстти талдоо методдору	2
1.4. Текстти автоматтык түрдө иштетүү системаларынын колдонулушу ...	3
1.5. Электрондук текст	3

II БӨЛҮМ

КОРПУСТУК ЛИНГВИСТИКА . СӨЗДҮ БАШТАПКЫ ФОРМАГА КЕЛТИРҮҮ ПРОБЛЕМАСЫ.....	5
2.1. Корпус тууралуу жалпы маалымат	5

2.2. Тексттердин корпусу. Түзүү принциптери	5
2.3. Электрондук текст корпустун негизи.....	9
2.4. Лемматизация жана нормалдаштыруу. Сөздү баштапкы формасынына келтирүү проблемасы.....	11
2.5. Сөз формаларын сөздүн түпнуска (сөздүк) формасына келтирүү эрежелери	12

III БӨЛҮМ

КЫРГЫЗ ТИЛИ ЖАНА АНЫН ЖАЛПЫ МҮНӨЗДӨМӨСҮ. КЫРГЫЗ ТИЛИНИН МОРФОЛОГИЯЛЫК МОДЕЛИ	14
3.1. Кыргыз тилинин жалпы мүнөздөмөсү	14
3.2. Кыргыз тилинин морфологиялык модели.....	14
3.3. Кыргыз тилиндеги сөздөрдүн морфологиялык структурасы.....	15
3.4. Сөздүн морфемалык курамы	15

IV БӨЛҮМ

МАТЕРИАЛДАР ЖАНА МЕТОДДОР	17
4.1. Материалдар.....	17
4.2. Методдор	19

V БӨЛҮМ

ЖЫЙЫНТЫКТАР	41
КОЛДОНУЛГАН БУЛАКТАРДЫН ТИЗМЕСИ:.....	42
ӨМҮР БАЯН.....	Hata! Yer işareti tanımlanmamış.
БИЛИМ ДЕҢГЭЭЛИ.....	Hata! Yer işareti tanımlanmamış.
МАКАЛАЛАРЫ	Hata! Yer işareti tanımlanmamış.

ЖАДЫБАЛДАРДЫН ТИЗМЕСИ

Жадыбал 1. Кептин негизги бөлүктөрү	21
Жадыбал 2. Сүйлөөнүн өзгөчө бөлүктөрү	21
Жадыбал 3. Кызматчы сөз түркүмдөрү.....	22
Жадыбал 4. Зат атоочтун түрлөрү	22
Жадыбал 5. Сан атооч.....	22
Жадыбал 6. Учур чак, өткөн чак келээр чак.....	22
Жадыбал 7. Жактар	23
Жадыбал 8. Таандык мүчөлөр	23
Жадыбал 9. Жөндөмөлөр.....	24
Жадыбал 10. Жандуу жансыз заттар.	24
Жадыбал 11. Мамилелер.	24
Жадыбал 12. Ыңгайлар.....	24
Жадыбал 13. Чектүү этиш формалары.....	25
Жадыбал 14. Чектүү эмес этиш формалары.	25
Жадыбал 15. Сын атоочтун даражалары.	26
Жадыбал 16. Терс маани мүчөсү.	26
Жадыбал 17. Менен.	26
Жадыбал 18. Сөз түркүмдөрүнүн жана омонимдердин индекси.	27

КЫСКАРТМАЛАР ТИЗМЕСИ

Кыскартылышы

NLP

NLU

NLG

AOT

IDE

NLTK

Чечмелениши

Natural Language Processing

Natural Language Understanding

Natural Language Generation

Автоматическая обработка текста

Integrated development environment

Natural Language Toolkit



I БӨЛҮМ

КИРИШҮҮ

1.1. Текстти автоматтык түрдө иштетүү системалары тууралуу жалпы маалымат

Азыркы учурда тил илиминин негизги проблемалары лексиканы жана семантиканы изилдөө, ошондой эле тез автоматташтырылган которуу болуп саналат. 20-кылымдын орто ченинде, өндүрүлгөн маалыматтын көлөмүнүн тынымсыз өсүшү эбегейсиз чоң көлөмдүү берилиштерден маалымат издөө, аны тандоо жана белгилүү критерийлер боюнча иреттөө актуалдуу маселе катары каралган.

Бул изилдөөлөрдө сөздүктөр, энциклопедиялар, архивдерсиз иштөө мүмкүн эмес. Бирок, тилекке каршы, окумуштуулардын керектүү маалыматтык ресурстарга жетүү мүмкүнчүлүгү дайыма эле боло бербейт. Азыркы лингвисттерге бул жагынан табигый тилди иштетүү үчүн ар кандай системаларды түзүү менен алектенген компьютердик, прикладдык лингвистика сыяктуу илимдин тармагы жардам бере алат. Бирок лингвистикалык маалыматтык ресурстарсыз бул иштетүү мүмкүн эмес.

Акыркы ондогон жылдар ичинде өзүнчө чакан маселелерди чечүүгө арналган көптөгөн автоматтык текстти иштетүү системалары пайда болду. Жогоруда айтылгандарга байланыштуу азыркы прикладдык лингвистикадагы эң актуалдуу маселе текстти автоматтык түрдө иштетүүнүн ыкмалары жана анализи болуп саналат.

1.2. Текстти автоматтык түрдө иштетүү системаларынын түрлөрү

Текстти автоматтык түрдө иштетүү системаларын төмөндөгүдөй классификациялоого болот:

- лицензиялоонун түрүнө жараша (менчик – көбүнчө белгилүү өндүрүүчүлөргө таандык, ал эми академиялык иштеп чыгуулар – бекер таратылат);
- жеткилүүлүк боюнча (системалар адамдардын аз өлчөмдүү чөйрөсүнө гана жеткиликтүү болушу мүмкүн, же коомдук доменде болушу мүмкүн);
- көп тилдүүлүк боюнча (колдоого алынган тилдердин санындагы айырмачылык);

- универсализм боюнча (конкреттүү маселелерди чечүү, жалпысынан тексттерди иштетүү);
- колдонулган маалыматтарга жараша (иштелген маалыматтардын түрү жана көлөмү).

Көп тилдүү системалар көбүнчө коммерциялык жактан жагымдуураак жана колдонууга оңой. Өз кезегинде, белгилүү бир тилге же тилдердин бир бөлүгүнө багытталган системалар тилдин өзгөчөлүктөрүн эске алуу менен көптөгөн милдеттерди аткарууда сапаты кичине болсо да, бирок абдан маанилүү өсүшүн камсыз кылат. Көп тилдүү системанын классикалык мисалы Google Translator болуп саналат.

Кеңири колдонууга багытталган системалар (коммерциялык), акыркы колдонуучуларга ылайык адатта ыңгайлуу жакшы интерфейске ээ (мисалы, Microsoft Bing Translator жана Google translator, ОРФО, “Яндекс.Текст” ар түрдүү утилиталары жана башкалар.). Бирок бул учурда, бул системаларды колдонуунун негизги жолуна караганда жагымдуу кошумча болуп саналат. Тескерисинче, изилдөөчүлөр үчүн гана иштелип чыккан системалар үчүн же ири долбоорлордун бир бөлүгү катары программалык камсыздоо интерфейси өз ара аракеттенүүнүн негизги (жана көбүнчө жалгыз) жолу болуп калат. Бул системалардагы акыркы колдонуучу интерфейстери көбүрөөк тестирлөөгө негизделген жана көбүнчө консолго негизделген. Мындай системалардын мисалдарына mystem, АОТ, ruMorpho 1 жана 2, Tomita талдоочу, OpenXerox, Snowball кирет. Алардын дээрлик бардыгы текстти талдоонун ар кандай этаптарында пайда болгон конкреттүү маселелерди чечүүгө арналган: тексттен сөздөрдү алуу (токенизация), морфологиялык талдоо (кептин бөлүктөрүн жана башка грамматикалык мүнөздөмөлөрдү аныктоо), сүйлөмдөрдүн синтаксистик түзүлүшүн куруу ж.б.

1.3. Текстти автоматтык түрдө иштетүүдөгү текстти талдоо методдору

Автоматтык түрдө текстти иштетүүдө көбүнчө учурда негизги чечим катары текстти талдоо кабыл алынат. Ал төмөндөгүдөй деңгээлдерди камтыйт:

Графематикалык анализ. Берилиштер массивинен сүйлөмдөрдү жана сөздөрдү (токендерди) тандоо.

Морфологиялык анализ. Сөздүн грамматикалык негизин бөлүп алуу, сөз

мүчөлөрүн аныктоо, сөздү сөздүк формасына жеткирүү.

Синтаксистик анализ. Сүйлөмдөгү сөздөрдүн ортосундагы синтаксистик байланыштарды аныктоо, сүйлөмдүн синтаксистик түзүлүшүн куруу.

Семантикалык анализ. Сөздөрдүн жана синтаксистик топтордун ортосундагы семантикалык байланыштарды аныктоо, семантикалык байланыштарды чыгаруу.

Ар бир мындай талдоо өз алдынча практикалык колдонууга ээ эмес, бирок көбүрөөк жалпы маселелерди чечүү үчүн активдүү колдонулат. Көптөгөн изилдөө системалары так жардамчы милдеттерди чечүү үчүн иштелип чыккан. Мындай системалар же методдорду сыноо жана эсептөө эксперименттерин жүргүзүү үчүн, же болбосо белгилүү бир прикладдык маселени чечүүчү системалар үчүн компоненттер (же китепканалар) катары колдонулат. Мындай системалардын мисалы болуп графематикалык анализ жана токенизация үчүн NLTK куралы, морфологиялык анализатор MyStem саналат.

1.4. Текстти автоматтык түрдө иштетүү системаларынын колдонулушу

Азыркы тапта коммерциялык кызыкчылык туудурган текстти автоматтык түрдө иштетүү системалары төмөнкү маселелерди чечүүдө колдонулат:

1. Machine Translation and Translation Aids [4]– машиналык котормо жана котормо куралдары;
2. Text Generation - текстти түзүү;
3. Localization and Internationalization - локализация жана интернационалдаштыруу;
4. Controlled Language – чектелген тилде иштөө;
5. Word Processing and Spelling Correction– текстти түзүү документтер (киргизүү, оңдоо, каталарды оңдоо);
6. Information Retrieval – маалыматты издөө жана ага байланыштуу милдеттер.

Белгилей кетсек, бул бөлүнүү кыйла шарттуу жана реалдуу системаларда функциялардын айкалышы көп кездешет [5].

1.5. Электрондук текст

Заманбап коммуникация каражаттарынын өнүгүшүнө жана жаңы типтеги –

электрондук тексттин барган сайын кеңейишине байланыштуу тил илиминде тексттин белги бирдиктеринин бирдиктүү семантикалык байланыш ырааттуулугу катары салттуу түшүнүгүнөн чегинүү байкалууда, негизги касиеттери ырааттуулугу жана бүтүндүгү болуп саналат. Электрондук тексттердин эволюциясы текст лингвистикасында алардын өзгөчөлүктөрүн так изилдей турган өзүнчө бир тармакты бөлүп алуу зарылчылыгы келип чыккан. Интернетте, дүйнөлүк маалыматтык социалдык-маданий мейкиндикте маалыматтар жайгаштырылат, анын ичинде электрондук тексттер бар. Электрондук текст заманбап маданияттын жаңы феномени катары электрондук китепканаларды активдүү түзүүнүн аркасында пайда болгон жана изилдене баштаган. Мазмундук жана маалыматтык мазмуну боюнча салттуу тексттин сапаттары менен жабдылган электрондук текст, тармактык чөйрөнүн өзгөчөлүгү менен байланышкан бир катар өзгөчөлүктөр менен, ошондой эле маалыматты берүүнүн айырмаланган формасы менен мүнөздөлөт. «Электрондук текст» түшүнүгү электрондук чөйрөдө, тармакта болгон текст катары чечмеленет. Салттуу текст көбүнчө белгилүү бир басылма же китеп менен байланыштырылат, ал эми электрондук текст мүмкүн болгон жанрдын ар кандай басылмаларын чагылдырышы мүмкүн. Салттуу формага гезит, журнал, энциклопедия, альбом ж.б. сыяктуу типологиялар жана классификациялар кирет. Бул басылмалардын атын атаганда китеп түрүндөгү текстти түшүнөбүз. Электрондук форма үчүн жанр экинчи даражадагы мааниге ээ. Электрондук текст - бул фактыларды жана цитаталарды табууга жана бөлүп көрсөтүүгө мүмкүндүк берген ар кандай маалымат, ал эми шилтемелер жана гипершилтемелер системасы электрондук тексттерди активдүү окууга мүмкүндүк берет. «Электрондук документ» деген термин да колдонулат. Электрондук документ - бул символдордун, сүрөттөрдүн жыйындысы түрүндө жазылган маалымат же материалдык алып жүрүүчүгө үн жазуу жана сактоо жана жалпы пайдалануу үчүн телекоммуникация жана компьютердик технологиялардын жардамы менен мейкиндикте жана убакытта жеткирүүгө арналган. Электрондук документтеги маалымат электрондук цифралык формада берилет.

II БӨЛҮМ

КОРПУСТУК ЛИНГВИСТИКА . СӨЗДҮ БАШТАПКЫ ФОРМАГА КЕЛТИРҮҮ ПРОБЛЕМАСЫ

2.1. Корпус тууралуу жалпы маалымат

Көптөгөн текст иштетүү системаларынын ажырагыс бөлүгү болуп - корпустар саналат. Корпустагы ар бир сөз толук грамматикалык мүнөздөмөлөр менен камсыз кылынат: ал сөздүн кайсы бөлүгүнө кирет, кандай формада турат, анын синтаксистик ролу кандай. Корпустар [5] тексттерди темалар жана жанрлар боюнча классификациялоо тапшырмаларын үйрөтүү үчүн, омонимияны жок кылуу жана анафораны чечүү үчүн колдонулган синтаксистик талдоочуларды жана программаларды үйрөтүү үчүн салым катары кызмат кылат. Машина котормочуларын даярдоо үчүн түрдүү тилдердеги окшош тексттерден турган параллелдүү корпустар колдонулат. Эреже катары, корпустар ондогон жылдар бою чогулуп, аларды түзүүгө чоң изилдөө топтору тартылат – мисалы, “Орус тилинин улуттук корпусу” долбоору 13 жылдан бери иштеп келе жатат жана аны Яндекс колдойт.

2.2. Тексттердин корпусу. Түзүү принциптери

«Маалыматтар корпусу» түшүнүгүн түшүндүрбөй туруп «тексттердин корпусу» түшүнүгүн ачуу мүмкүн эмес. Маалымат корпусу – бул тил системасынын аймагынан берилген эрежелерге ылайык түзүлгөн маалыматтардын үлгүсү. Бул үлгү лингвистикалык сүрөттөлүүгө тийиш болгон кубулуштарды камтыйт. Маалымат корпусу бир өлчөм менен мүнөздөлөт. Аны сөз менен гана өлчөөгө болот, анткени анын компоненттерин чыгарууга жөндөмү жок. Бирок, бул маалымат корпусу тилди система катары кайра куруу үчүн колдонулушу мүмкүн эмес дегенди билдирбейт. Тескерисинче, бул корпусту лингвистикалык изилдөөнүн негизги милдеттеринин бири. Ошентип, биз тил системасынын продуктуларына мүнөздүү болгон глобалдык карама-каршылыктардын бирин көрөбүз. Тилчи өз ишмердүүлүгүнүн жеке натыйжаларына таянып, тилдин система катары функционалдуулугу жөнүндө тыянак чыгарышы керек. Тексттик корпус маалымат корпусунун бир түрү. Анын бирдиктери тексттер же жетиштүү

узундуктагы текст фрагменттери болуп саналат, алар көйгөйлүү аймактагы тексттердин каалаган фрагменттерин камтыйт. Жалпысынан алганда, корпус (латын тилинен *corpus* - "дене") [6] бирден ашык текстти камтыган ар кандай жыйын деп атоого болот. Көбүнчө айрым тексттер ар кандай типтеги лингвистикалык жана адабий талдоо үчүн пайдаланылган фактыга туш болушу мүмкүн. Айтылгандарга карабастан, «электрондук лингвистиканын негизи катары корпус» түшүнүгү жалгыз текстти текшерүүдөн айырмаланат. В.В. Рыковдун айтымында, тексттик корпус бул тексттердин айрым жыйнактары. Бул жыйнактардын негизин логикалык идея, бул тексттерди бириктирген идея түзөт. Логикалык идея тексттер айкалыштырылган эрежелерде, ошондой эле корпустарды талдоо программасында жана алгоритминде, ага байланыштуу методологияда жана идеологияда камтылган.

Максатына жараша текст корпусунун бир нече түрү бар:

1. Сактоо формасы боюнча:

- үн түрүндө;
- жазылган;
- аралаш;

2. Тексттердин берилүүчү тилине ылайык:

- бир тилдүү;
- көп тилдүү;

3. Жанр боюнча:

- адабий;
- диалект;
- оозеки;
- журналисттик;
- аралаш;

4. Кирүү ыкмалары боюнча:

- эркин жеткиликтүү;
- коммерциялык;
- жабык;

5. Дайындоо боюнча:

- изилдөө;

- иллюстративдик;
- 6. Динамизм боюнча:
 - динамикалык (монитор);
 - статикалык;
- 7. Кошумча маалыматтардын болушуна жараша:
 - аннотацияланган (белгиленген);
 - бөлүштүрүлбөгөн [5].

В.В. Рыков текст корпусунун бир аз башкача классификациясын берет:

1. Уюштуруу даражасы жана түзүмү боюнча:
 - электрондук архив - бул электрондук алып жүрүүчүдөгү тексттер, бирок алардын машиналык алып жүрүүчүдө берилген формасы стандартташтырылбаган жана бирдиктүү эмес;
 - электрондук китепкана - бул жерде тексттер бир тектүү жана стандартташтырылган түрдө берилет;
 - тексттердин корпусу - форма стандартташтырылган жана унификацияланган, тексттер тилдик чындыктын бир бөлүгүн чагылдырууга арналган;
 - субкорпус – бул корпусун кандайдыр бир автономдуу бөлүгү.
2. Хронологиялык:
 - синхрондуу;
 - мониторинг (тилдин учурдагы абалын көзөмөлдөйт);
 - диахрондук.
3. Индексөө аркылуу:
 - жөнөкөй;
 - аннотацияланган.
4. Тил боюнча:
 - бир тилдүү;
 - эки тилдүү;
 - көп тилдүү.
5. Ишти колдонуу жана колдонуу ыкмасы боюнча:
 - изилдөө;

- иллюстративдик;
- параллелдүү.

6. Корпустун абалына жараша:

- динамикалык;
- статикалык

Корпустарды долбоорлоо жана пайдалануу бирдиктуу метод боюнча жүргүзүлбөйт. Бул тилдердин, салттардын, технологиялык процесстердин айырмачылыгына байланыштуу. Бирок, негизги талаптарды бөлүп көрсөтүүгө болот. В.В. Рыковдун тизмесине ылайык, негизги талаптар катары төмөнкү факторлор кирет:

1. Колдонуучунун түрү – топ, лингвистикалык коом, индивид.
2. Негизги логикалык ой.
3. Корпусту түзүүдө колдонула турган маалыматтардын көлөмү.
4. Корпусту түзүү канчалык реалдуу жана зарыл.
5. Кандай компоненттер колдонулат - тексттик үзүндүлөр, толук тексттер, экөө тең бир убакта.
6. Текстти тандоо процедурасы кандайча ишке ашат, анткени ал ар кандай максаттар үчүн ар кандай жолдор менен жүзөгө ашырылат:
 - сүйлөө материалы каралат;
 - тексттер сканерден өткөрүлөт;
 - корпустун калыптанышын аяктоо.
7. Продукт катары корпусту төмөндөгүдөй көрсөтсө болот:
 - бүтүндөй тексттин аннотациясын;
 - тексттин сөздүк материалынын бирдиктүү чагылдырылышы.
8. Аннотация, вербалдык маалыматты индекстөө

Корпус үчүн тандалган тексттердин көлөмү жөнүндөгү маселе эң негизги маселелердин бири болуп саналат. Тил үчүн сейрек болсо да изилденип жаткан кубулуш корпуста чагылдырылса абдан жакшы болмок. Комплекстүүлүк корпустун курамына жана структурасына коюлган маанилүү талаптардын бири болуп саналат.

Тандоосуна жараша корпустар тең салмактуу жана монитордук болуп бөлүнөт. Биринчиси оозеки жана жазуу жүзүндөгү тексттерди камтыган

дискурстун ар кандай модулдарын билдирген тексттерден турат. Алар стилдери, жанрлары жана темалары боюнча такыр башкача болушу мүмкүн. Корпусту түзүү процессинде изилдөөчүлөр ар кандай типтеги тексттер бериле турган пропорцияларды белгилешет. Бул корпустар белгиленген көлөмү менен мүнөздөлөт. Аларды толуктоо жаңы тексттерди кылдат тандап алгандан кийин гана ишке ашырылат. Монитордук корпустар дайыма толуктоо менен мүнөздөлөт. Берилген тилдеги жаңы тексттер аларга стилдердин, режимдердин жана жанрлардын тең салмактуулугун сактабастан дайыма кошулуп турат.

Лингвистикалык маселелерди чечүү тексттерде жана тексттердин ичиндеги айрым тилдик бирдиктерде камтыла турган кошумча лингвистикалык жана металлингвистикалык маалыматты (белгилөө/аннотация) талап кылат. Текст жөнүндө ар кандай экстралингвистикалык маалыматтарды, анын ичинде автордун аты-жөнүн, жанрын, аталышын чагылдырган мета-белгиден тышкары, заманбап корпус лингвистикалык мүнөздөмөнүн ар кандай деңгээлдерине туура келген белгилерди камтыйт. Бул морфологиялык, фонетикалык, синтаксистик же башка белгилөө болушу мүмкүн.

Ар кандай лингвистикалык маселелерди чечүү үчүн тексттер жана тексттин ичиндеги айрым тилдик бирдиктер кошумча лингвистикалык жана металлингвистикалык маалыматты - белгини (аннотацияны) камтышы зарыл. Заманбап корпуста мета-белгилөөдөн тышкары, тилдин ар кандай деңгээлдерине туура келген белгилөөлөр бар - морфологиялык, синтаксистик, фонетикалык жана башкалар.

2.3. Электрондук текст корпустун негизи

Электрондук тексттер басылмаларга салыштырмалуу принципиалдуу жаңы касиеттери менен мүнөздөлөт. Бул аларды сактоого жана таратууга заманбап мамиле менен шартталган. Электрондук тексттер тилчилер үчүн кеңири перспективаларды ачат. Бул кошумча классификацияны жана салттуу маселелерди чечүүгө жаңы ыкмаларды эске алуу менен чоң көлөмдөгү маалыматты иштетүүгө тиешелүү.

Компьютердик технологиялар тил илиминде көптөн бери колдонулуп келет. Р.Буса өткөн кылымдын 40-жылдарынын аягында Фома Аквинскийдин

чыгармаларынын конкордансын түзүүнүн үстүндө иштей баштаган. Бул иш бир нече ондогон жылдардан бери уланып келе жатат. Азыркы этапта ар кандай тилдик тармактарда колдонулган көптөгөн долбоорлор бар. Бул долбоорлор компьютердин жардамы менен иштейт. Мисал катары басып чыгарууга автоматтык түрдө даярдоо жана индекстерди түзүү, компьютердин жардамы менен тексттердин мазмуну жана стилистикалык анализи, электрондук басылмалар, кепти таануу жана генерация кирет. Бул изилдөө аймагы жалпысынан “Гуманитардык информатика” деп аталат.

Электрондук тексттер компьютердик лингвистика жана филология менен байланышкан ар кандай ишмердүүлүктүн негизин түзөт. Алгач электрондук текст басма текстти түзүүнүн этаптарынын бири болгон. Бара-бара электрондук тексттер акыркы көз карандысыз продуктка, мисалы, лексикалык маалымат базаларына, электрондук басылмаларга айланган. Текстти электрондук түргө өткөрүү (санариптештирүү) үчүн ар бир символ тууралуу маалыматты, ошондой эле мета-маалыматтарды электрондук алып жүрүүчүгө өткөрүү зарыл. Электрондук тексттердеги мета-маалымат бул текст камтыган кошумча маалыматтар. Буга астын сызуу, кара тамга, бөлүмгө бөлүү, строфа, ыр саптары жана башка маалыматтар, чыгарманын аталышына чейин, автордун аты-жөнү, кошумча лингвистикалык маалыматтар кирет.

Басылып чыккан текстти (кайра) санариптештирүү процесси басып чыгарылган оригиналды санариптештирүүнү жана аны кошумча мета-маалымат менен камсыздоону камтыйт.

Текстти киргизүү процессинде белгилөө жарым-жартылай аткарылышы мүмкүн - белгилөө, метаберилиштерди киргизүү. Бул аракеттер тексттик нюанстардын негизинде түз же кыйыр түрдө автоматтык режимде аткарылат. Татаал түзүлүштөгү тексттердин көбү кесипкөй филолог тарабынан кошумча иштетүүнү талап кылат. Адис комментарийлерди жана толуктоолорду киргизет, критикалык аппаратты түзөт же жеке фрагменттерди гипершилтемелер менен байланыштырат. Электрондук текстти жарыялоо үчүн, ал көрсөтүлүүчү, навигацияланган жана изделген программалык камсыздоого ылайыкташтырылышы керек.

Электрондук текст менен иштеген лингвисттер үчүн анын узак мөөнөткө

сакталышы олуттуу көйгөй. Көпчүлүк заманбап электрондук басылмалар аларда көрсөтүү жана издөө үчүн иштелип чыккан программалар менен тыгыз байланышта жана бул программалардын иштөө мөөнөтүнөн түздөн-түз көз каранды. Text Encoding Initiative (TEI, www.tei-c.org) — бул программалык камсыздоо ОСден көз карандысыз коддоого мүмкүндүк берген, тилчилер тобу тарабынан иштелип чыккан филологиялык текстти белгилөө системасы. Бул коддоо узактыгы боюнча басылган текст менен салыштырууга болот. Текстти коддоо демилгеси эл аралык стандарт болгон универсалдуу белгилөө тилине (XML) негизделген. Ал тексттеги лирика, проза, драматургияга мүнөздүү өзгөчөлүктөрдү бөлүп көрсөтүү жөндөмү менен жабдылган, сөздүктөрдү, сөздүн транскрипцияларын, терминологиялык базаларды белгилей алат. Белгилөө ар кандай символдорду коддоочу татаал гипертексттик шилтемелерди түзүү үчүн зарыл болгон механизмдерди ишке ашырат. Text Encoding Initiative жана XML анын презентациясын сүрөттөгөн типографиялык маалыматтан бөлүнгөн семантикалык белгилөө концепциясына негизделген.[6]

Текстти көрсөтүү өз тилинде берилген маалыматтарды колдонуу менен жүзөгө ашырылат. Бул маалыматтар форматтоону сүрөттөө жана маалымат чагылдырылган медианын өзгөчөлүктөрүн эске алуу үчүн зарыл. Text Encoding Initiative артыкчылыгы - маалыматты узак мөөнөткө сактоо үчүн бирдиктүү стандартты колдонуу мүмкүнчүлүгү. Филологдор тарабынан электрондук тексттер менен иштөөдө колдонулган бул стандарт 10 жылдан ашык убакыттан бери дүйнөнүн булуң-бурчунан келген көз карандысыз лингвисттердин тобунун өкүлдөрү тарабынан иштелип чыккан жана сыналып келген.

2.4. Лемматизация жана нормалдаштыруу. Сөздү баштапкы формасына келтирүү проблемасы

Грамматикалык сөздүктө ишке ашырылган нормалдаштыруу процесси баштапкы тексттен грамматикалык маалыматты алып салууга мүмкүндүк берет (жөндөмөлөр, сандар, этиштин түрлөрү жана чактары, бөлүкчөлөрдүн үндөрү ж.б.у.с.).

Текстти нормалдаштыруу стемпингди колдонбойт. Лемматизатор өз натыйжалары боюнча нормализаторго алда канча жакын. Бирок, контекстти эске албастан, сөздөрдүн жөнөкөйлөштүрүлгөн анализин колдонот. Бул сөз мүчөсүн

аныктоодо түшүнүксүздүктөргө алып келет. Мисалы, биз чуңкур казабыз деген сөз айкашындагы сөздөрдү лемматташтыруу экинчи сөзгө эки лемматташтыруу вариантын берет: каз зат атооч жана казуу этиши. Бул түшүнүксүздүктү морфологиялык анализатордун катышуусусуз чечүү мүмкүн эмес.[7]

Лемматизация – морфологиялык анализдин бир ыкмасы, ал сөз формасын баштапкы сөздүк формасына (лемма) келтирүү методу.

Ар кандай табигый тилде лемматташтыруу процессинде эки ача натыйжаларды бере турган сөздөрдүн белгилүү бир пайызы бар, мисалы, “каз” сөз формасы акырында эки леммага – казуу (этиш) жана каз (зат атооч) кыскарышы мүмкүн.

Сөздүн түпкү формасы — лингвистикалык статистикада кээде «сөздүн негизги формасы» деген мааниде колдонулган термин. Бул таптакыр туура эмес, анткени тил илиминде сөздүн баштапкы формасы жөнөкөй (же туунду) негиз катары түшүнүлөт, ал сөздүн андан аркы жасалышынын жана калыптанышынын булагы болуп саналат.

Лемматизацияда биз сөздү алып, ал үчүн лемманы алабыз - нормалдуу (баштапкы, сөздүк) форма: китепке - китеп. Кыргыз тилинин лексикасы сөздүктөгү жазуулардын аталыштары зат атоочтун баштапкы формасына ар дайым туура келгидей кылып түзүлгөн. Демек, лемматизация морфологиялык анализдин кошумча продуктусу болушу мүмкүн. Бирок, морфологиялык анализдин өзү бир топ татаал жана өтө чоң сөздүк базаны талап кылат.

2.5. Сөз формаларын сөздүн түпнуска (сөздүк) формасына келтирүү эрежелери

Морфологиялык анализдин милдети – сөздүн нормалдуу формасы кандай сөз формасынан түзүлгөнүн аныктоо. Бул талдоо бардык мүмкүн болгон сөз формаларына эмес, нормалдуу формалардын негизинде гана ориентацияга өбөлгө түзөт.

Нормалдуу форма сөздүн лингвистикалык салтына жараша айырмаланган формаларынын бири катары каралышы керек. Форма сөзүнүн астында сөздүн (саптын) нормалдуу формасы менен байланышып, сөздүн колдонулуш өзгөчөлүктөрүн көрсөткөн формасы түшүнүлөт. Изилдөө чөйрөсүндө сөздүн расмий мүчөлөрүнүн курамы боюнча талаш-тартыштар дагы эле уланып келе жатканына карабастан, жалпысынан негизги сөз мүчөлөрүнүн тизмеси орун алган.

Белгилүү бир морфологиялык сөздүктү колдонуу процессинде эң негизгиси тизмени тандоо болуп саналат, анткени кийинчерээк анын өзгөртүү операциялары көбөйөт.



III БӨЛҮМ

КЫРГЫЗ ТИЛИ ЖАНА АНЫН ЖАЛПЫ МҮНӨЗДӨМӨСҮ. КЫРГЫЗ ТИЛИНИН МОРФОЛОГИЯЛЫК МОДЕЛИ

3.1. Кыргыз тилинин жалпы мүнөздөмөсү

Түрк тилдеринин бири болгон кыргыз тили Өзбекстандын Наманган, Анжиян, Фергана облустарында, Тажикстандын кээ бир тоолуу аймактарында, Казакстандын Кыргыз Республикасы менен чектеш аймактарында, КНРдин Синьцзян-Уйгур автономиялуу району, Монгол Эл Республикасынын батыш райондорунда, Афганистандын жана Пакистандын тундук-чыгышында, Турцияда жарым-жартылай сүйлөгөн Кыргыз Республикасынын түпкү калкынын улуттук тили. Кыргыз тилинде сүйлөгөндөрдүн жалпы саны 4,8 миллиондон ашык адамды түзөт. Кыргыз Республикасында эки диалектилик топ бар – түндүк жана түштүк, башка классификация боюнча – түндүк, түштүк-батыш жана түштүк-чыгыш, алар негизинен фонетикасы жана лексикасы боюнча айырмаланат.[8]

Типологиялык же морфологиялык белгилери боюнча тилдер төмөнкү топторго бөлүнөт: аморфтук, агглютинациялоочу (агглютинативдик), флекциялык жана полисинтетикалык.

Кыргыз тили типологиялык классификациясы боюнча агглютинативдик же агглютинациялык тилдерге кирет. Агглютинативдик тилдер (латын тилинен *agglutination* - жабыштыруу) - флексиянын басымдуу түрү агглютинация болгон структурага ээ тилдер - ар бири бирден гана маани алып жүрүүчү ар кандай форманттарды (суффикс же префикс) жабыштыруу менен жасалат.

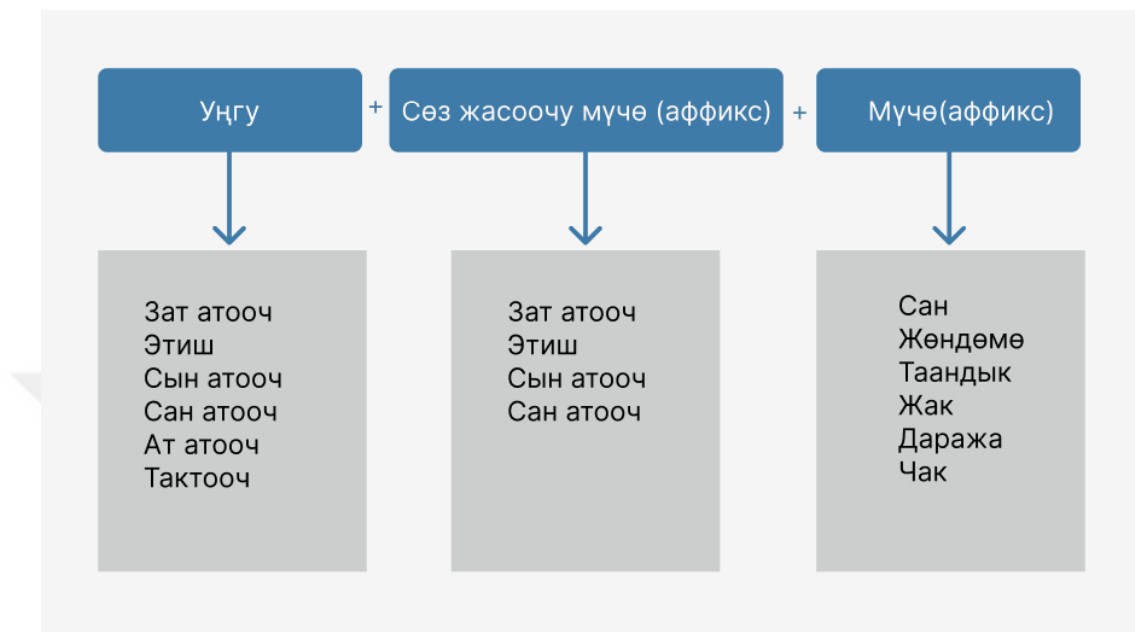
Агглютинативдик структура флекциялык структурага карама-каршы келет, мында ар бир формант дароо бир нече ажырагыс маанилерди алып жүрөт (мисалы, учур, жыныс, сан ж. б.). Андан да маанилүүсү, агглютинативдик тилдерде форманттар бөлүнгүс структураларды түзбөйт жана башка форманттардын таасиринен өзгөрбөйт[9].

3.2. Кыргыз тилинин морфологиялык модели

Кыргыз тили аффикстерди уңгуга (же негизге) кошуунун катаал ырааттуулугу менен мүнөздөлөт, мисалы, зат атооч жасоо үчүн адегенде уңгуга сөз

жасоочу суффикс, андан соң форма түзүүчү аффикстер уланат. Бул аффикстердин бирин калтыруу алардын кошулуу тартибин өзгөртпөйт.[10]

3.3. Кыргыз тилиндеги сөздөрдүн морфологиялык структурасы.



Сүрөт 1 . Кыргыз тилиндеги сөздөрдүн морфологиялык структурасы

3.4. Сөздүн морфемалык курамы

Морфема (грекче morphе - форма) - сөз жана сөз формаларынын эң кичине маанилүү бөлүгү, б.а. бул берилген сөз формасынын курамындагы уңгулардын, сөз жасоочу жана форма түзүүчү аффикстердин жалпыланган аталышы. [11][12][13][14]

Морфемалардын ичинен уңгулар менен аффикстер функционалдык жактан карама-каршы келет. Тамыр (унгу) - борбордук жана милдеттүү элемент, сөздүн лексикалык маанисин алдын ала аныктоочу морфемикалык түзүлүш; тектеш сөздөрдүн жалпы мүчөсү: ыр (песня), ыр + ым (моя песня), ыр + да (петь).

Уңгулар эркин (эркин унгу) жана туташкан унгу болуп экиге бөлүнөт. Эркин уңгулар сөз жасоочу аффикстерге менен же аларсыз өз алдынча сөздүн негизин катары колдонулат: Жол (дорого), жыл (год), бала (ребенок), ыр + да (петь).

Туташкан унгулар туунду аффикстер менен айкалышып гана колдонулуп, сөздүн уңгусуна эч качан тең келбейт: ом+ур (разворотить), ом+кор (выворачивать, переворачивать), үйрө+н (научиться), үйрө+т (научить), суй+даң (редкий, жидкий), суй+ул (делать редким), суй+ук (редкий, жидкий).

Аффикс (лат. affixis - тиркелген) (Мүчө) - өз алдынча уңгусуз колдонулбаган, бир эле сөздүн түрдүү формаларын жасоого же башка бир уңгулуу сөздү жасоого кызмат кылган морфема. Ошентип, аффикстер форма жасоочу же сөз өзгөртүүчү мүчөлөр жана туунду болуп экиге бөлүнөт. Форма жасоочу аффикстер бир эле сөздүн формаларын түзгөн аффикстер. Форма жасоочу аффикстерге зат атоочтун жөндөмө аффикстери (жөндөмө мүчөлөрү), зат атоочтун көптүк аффикстери (көптүк сан мүчөлөрү), этиштин өздүк аффикстери (этиштин жак мүчөлөрү) ж.б. кирет: окуучу (ученик), окуучу+нун (ученика), окуучу+га (ученику), окуучу+лар (ученики); жаза+мын (я напишу), жаза+сың (ты напишешь), жаза+сыз (напишете), жаза+т (напишет).

Сөз жасоочу аффикстердин жардамы менен берилген уңгудан жаңы сөздөр түзүлөт. Мисалы: сура (спрашивать), сура+к (допрос), сура+к+кана (судилище), сура+к+чы (следователь); тиз (выстраивать в ряд), тиз+ме (ряд, список), тиз+ме+чи (составитель списков).

Сөздүн негизи (Негиз) – форма жасоочу аффикстери болбогон бүтүндөй сөз. Негиз уңгудан гана турушу мүмкүн: кол (рука), баш (голова), ыр (песня), ак (белый); уңгу жана сөз жасоочу аффикстерден: колдо-ду (кол – рука, кол+до – поддерживать, колдо+ду – поддержал, негиз – колдо), ырчы-га (ыр – песня, ыр+чы – певец, ырчы-га – певцу, негиз – ырчы), акта-ган (ак – белый, ак+та – белить, акта+ган – беливший), негиз – акта).[15]

Ошентип, уңгулар туунду эмес (тубаса негиз) жана уңгу менен сөз жасоочу аффикстерден турган туундулар (тууунду негиз) болуп бөлүнөт. [16] [17] [18] [19]

IV БӨЛҮМ

МАТЕРИАЛДАР ЖАНА МЕТОДДОР

4.1. Материалдар

Бул изилдөөдө колдонулган компьютерлер:

Ноутбук

Процессор: Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz

Оперативдик эс (RAM): 16.0 GB

Операциондук система: 64-bit Operating System

Python тили - жогорку деңгээлдеги жалпы максаттуу программалоо тили, жана бул веб-тиркемелерди иштеп чыгуу үчүн дагы кеңири колдонулат. Бул тил программа иштеп чыгуучунун өндүрүмдүүлүгүн жогорулатууга жана коддун окулушун жакшыртууга багытталган. Программалоо тилинин аталышынын туура айтылышы – Пайтон, бирок көбүрөөк – Питон деп бурмаланганы колдонулат. Python: структуралык, объектиге багытталган, функционалдык, императивдик жана аспектке багытталган программалоолорду колдойт. Бул тилде динамикалык типизация, эстутумду автоматтык түрдө башкаруу, толук интроспекция, өзгөчөлүктү иштетүү механизми, көп агымдуу эсептөөлөрдү колдоо жана ыңгайлуу жогорку деңгээлдеги маалымат структуралары бар. Python коду функцияларга жана класстарга уюшулган, алар модулдарга бириктирилет, алар өз кезегинде пакеттерге бириктирилет. Python адатта чечмеленүүчү катары колдонулат, бирок аны Java байт кодуна жана MSILге (.NET платформасынын ичинде) түзсө болот.

Python тилинде китепканалар

Программалоо тилинин китепканасы – бул программалоо тилин колдонуу менен кандайдыр бир конкреттүү операцияларды жеңилдеткен модулдардын жана функциялардын жыйындысы. Биз өзүбүздүн ишибизде текст менен иштөөдө NLTK китепканасын колдондук.[20]

NLTK

NLTK - бул Pythonдо классификациялоо, стемминг, маркерлөө, талдоо жана семантикалык ой жүгүртүү сыяктуу иштерди колдогон маанилүү китепкана. Бул

табигый тилди иштетүү жана машина үйрөнүү үчүн негизги курал болуп саналат. Бүгүнкү күндө ал NLP жана машина үйрөнүү менен жаңыдан баштап жаткан Python иштеп чыгуучулары үчүн билим берүү негизи катары кызмат кылат.

Китепкана Пенсильвания университетинде Стивен Бирд жана Эдвард Лаупер тарабынан иштелип чыккан. Ал NLP тин ачылышында негизги ролду ойногон. NLTK, башка Python китепканалары жана куралдары менен бирге, азыр дүйнө жүзүндөгү университеттер тарабынан окуу пландарында колдонулат.

Китепкана абдан ар тараптуу, бирок аны табигый тилди иштетүү үчүн колдонуу кыйын. NLTK абдан жай болушу мүмкүн жана тез өнүгүп жаткан өнөр жай колдонуу талаптарына жооп бербейт. Бирок иштеп чыгуучулар бул пайдалуу китеп (<https://www.nltk.org/book/ch01.html>) сыяктуу ресурстарды пайдалана алышат.

IDE - Integrated development environment - интеграцияланган иштеп чыгуу чөйрөсү, тиешелүү программалоо тилинде ыңгайлуураак иштеп чыгууга мүмкүндүк берүүчү программалык каражаттардын жыйындысы. Адатта IDE-де тексттик редактор, компилятор же интерпретатор, мүчүлүштүктөрдү оңдоочу жана башка программалык камсыздоо бар. IDE программ иштеп чыгуу ылдамдыгын жогорулатууга мүмкүндүк берет (албетте, IDE менен иштөө боюнча алдын ала окутуу шартында)

PyCharm - бул Python программалоо тили үчүн иштеп чыгарылган интегралдык өнүктүрүү чөйрөсү жана эффективдүү программалоо үчүн зарыл болгон куралдардын толук топтомун өзүнө камтыйт. Бул IDE-нин биринчи версиясы 2010-жылы чыккан.

Азыр PyCharm эки версияда таратылат:

- акы төлөнүүчү (PyCharm Professional Edition) жана
- акысыз (PyCharm Community Edition).

Акысыз версиясы ачык булак жана Apache 2 лицензиясы боюнча лицензияланган. Бул Python иштеп чыгуу үчүн ылайыктуу жеңилдетилген чөйрө.

Акы төлөнүүчү версия - көп тилдүү веб-тиркемелерди камтыган иштеп чыгуу мүмкүнчүлүгү бар кыйла өркүндөтүлгөн жана функционалдык версия жана төмөнкү фреймворкторду колдойт:

- Django,
- Flask,
- Google App Engine,
- Pyramid,
- web2py

Ал алыстан иштеп чыгууга, ошондой эле маалымат базалары менен иштөөгө мүмкүндүк берет.

PyCharm артыкчылыктары

PyCharm бардык пайдалуу функциялары бар ыңгайлуу код редакторуна ээ: синтаксисти бөлүп көрсөтүү, авто-форматтоо, толтуруу жана тиешелүү боштуктарды калтыруу. PyCharm тилдин интерпретаторунун версияларынын шайкештигин текшерүүгө, ошондой эле код калыптарын колдонууга мүмкүндүк берет.

Документацияны көп колдонгондор үчүн, аны редактордун терезесинде эле (элементтер үчүн) же браузерде (тышкы документациялар үчүн) көрүү ыңгайлуу болот.

PyCharm кодду тез рефакторлоого, ошондой эле ыңгайлуу мүчүлүштүктөрдү оңдоонун графикалык түрүн колдонууга мүмкүндүк берет.

Бул утилитта Django'нун бардык акыркы версияларын, ошондой эле IronPython, Jython, Cython, PyPy wxPython, PyQt, PyGTK жана башка көптөгөн куралдарды колдойт.

PyCharmда сиз камтылган Unit тестин өткөрө аласыз, Python, Django, SSH, мүчүлүштүктөрдү оңдоочу жана маалымат базалары үчүн интерактивдүү консолдорду колдоно аласыз.

PyCharm плагиндердин чоң коллекциясына ээ жана аны JIRA, Youtrack, Lighthouse, Redmine, Trac жана башкалар сыяктуу ар кандай трекерлер менен бирге колдонсо болот.

PyCharm кайчылаш платформа иштеп чыгуу чөйрөсү болуп саналат: Linux, Windows жана Mac OS колдонсо болот.

4.2. Методдор

Токенизация

Токенизация чоң көлөмдөгү текстти токендер деп аталган кичинекей бөлүктөргө бөлөт. Бул үзүндүлөр же токендер үлгүлөрдү табуу үчүн абдан пайдалуу жана стемминг менен лемматизация үчүн негизги кадам катары каралат. Токенизация ошондой эле сезимтал маалымат элементтерин сезгич эместерге алмаштыра алат.

Табигый тилди иштетүү (NLP) текстти классификациялоо, сезимди талдоо, акылдуу чатбот, тилди которуу жана башка көптөгөн колдонмолорду куруу үчүн колдонулат. Ошентип, жогорудагы максатка жетүү үчүн текст үлгүсүн түшүнүү маанилүү болуп калат.

Текстти классификациялоо же спам чыпкалоо сыяктуу тапшырмаларды Keras жана Tensorflow сыяктуу терең үйрөнүү китепканалары менен бирге NLP иштетет.

NLPдеги Токенизациянын мааниси

Табигый тилди иштетүүдөн мурун, символ сапты түзгөн сөздөрдү аныкташыбыз керек. Ошентип, токенизация табигый тилди иштетүүнү (NLP) улантуу үчүн эң маанилүү кадам болуп көрүнөт.

Бул кадам зарыл, анткени тексттин чыныгы маанисин тексттеги ар бир сөздү талдоо аркылуу чечмелесе болот.

Эми мисал катары төмөнкү сапты алалы:

Менин атым Айжамал Ишенова

Жогорудагы сапта токенизацияны аткаргандан кийин, биз төмөндө көрсөтүлгөндөй натыйжаны алабыз:

['Менин', 'аты-жөнүм', 'бул', ' Айжамал ', ' Ишенова ']

Операцияны колдонуунун ар кандай варианттары бар. Биз токенизацияланган форманы төмөнкү үчүн колдоно алабыз:

- тексттеги сөздөрдүн жалпы санын эсептөө;
- бир сөздүн жыштыгын, башкача айтканда, белгилүү бир сөздүн жалпы санын аныктоо ж.б.

Кыргыз тилинин эрежелеринин тег менен белгилениши

Эркин жана байламталуу морфемалардын түрлөрү англис жана кыргыз тилдеринде бирдей. Эркин морфемалардын лексикалык жана функционалдык түрлөрү кыргызча маани берүү (эркин морфемалар) жана жардам (функционалдык морфемалар) деп аталат. Байланыштуу морфемалардын кыргызча уланды мүчө (флексиялык аффикс) жана куранды мүчө (тууунду аффикс) деп аталган.

Уңгулоо жана лемматташтыруу - морфологиялык талдоо процесстери. Уңгулоо (стемминг) каражаттары сөз формаларын (псевдо) уңгуга чейин кыскартат, ал эми лемматизация сөз формаларын лингвистикалык жактан жарактуу леммаларга чейин кыскартат.

Мисалы: (сиздер) камсыздандырылбагандардансыздарбы

Камсыздандырылбагандардансыздарбы/кам <n> <nom> +сыз <adj> <neg> +да <v> <imp> +н <v> <reflex> <tv> +дыр <v> <caus><iv> + ыл <v> <pass>+ба<neg> +ган <adj> <pastpart><pst> +дар <pl> + дан <n> <abl> +сыздар <a2pl>+бы<ques>

Камсыздандырылбагандардансыздарбы/кам <з.ат> <ат.жөн><абстр.> +сыз <сын ат.> + да <эт.> <буйр.эт.> +н <эт.> <өзд.м.> <өтп.эт> +дыр <эт.> <арк.м> <өтм.эт> + ыл <эт.> <туюк м.>+ба<терс м> +ган <сын ат.> <атоочт.> <жай д.> <сап.с.ат> + дар <көп.с> +дан <чыг.жөн> + сыздар <2ж.к.с>+бы <сур.м>

Төмөндө кыргыз тилинин грамматикалык өзгөчөлүктөрүн жана алардын мисалдарын түрк лексикасынын апертиумунун символдору менен таблицаларда көрсөтүлгөн.

Жадыбал 1. Кептин негизги бөлүктөрү

Символ	Мааниси	Мааниси	Мисал
n	Noun	Зат атооч (Zat atooç)	алма (Alma) "apple"
v	Verb	Этиш (Etiş)	кел (Kel) "come"
adj	Adjective	Сын атооч (Sin atooç)	жакшы (Jakşy) "good"
adv	Adverb	Тактооч (Taktooç)	бүгүн (Bügün) "today"
prn	Pronoun	Ат атооч (At atooç)	мен (men) "I"
num	Numeral	Сан атооч (San atooç)	үч (üç) "three"

Жадыбал 2. Сүйлөөнүн өзгөчө бөлүктөрү

Символ	Мааниси	Мааниси	Мисал
ij	Interjection	Сырдык сөздөр (Sırdık sözdör)	катыгүн (katıgün) "alas"

ideo	Ideophone	Тууранды сөздөр (Ideophone)	тарс (tars) "bang"
------	-----------	-----------------------------	--------------------

Жадыбал 3. Кызматчы сөз түркүмдөрү

Символ	Мааниси	Мааниси	Мисал
conj	Conjunction	Байламта (Baylamta)	менен (menen) "with"
post	Postposition	Жандооч (Jandooç)	кийин (kiyin) "after"
mod	Modal words	Модалдык сөздөр (Modal words)	албетте (albette) "of course"
-----	Particle	Бөлүкчөлөр (Bölükçölör)	го (go) "it seems"

Жадыбал 4. Зат атоочтун түрлөрү

Символ	Мааниси	Мааниси	Мисал
pers	Personal pronoun	Жактама (Jaktama)	биз (biz) "we"
dem	Demonstrative pronoun	Шилтеме (Şilteme)	тигил (tigil) "that"
itg	Interrogative pronoun	Сурама (Surama)	ким (kim) "who"
neg	Negative pronoun	Тангыч (Tangiç)	эч бир (eç bir) "none"
qnt	Quantifier pronoun	Аныктама (Anıktama)	бүт (büt) "whole"
ind	Indefinite pronoun	Белгисиз (Belgisiz)	кээ бир (kee bir) "some"

Жадыбал 5. Сан атооч

Символ	Мааниси	Мааниси	Мисал
sg	Singular	Жекелик сан (Jekelik san)	китеп (kitep) "book"
pl	Plural	Көптүк сан (Köptük san)	китеп+лар=китептер (kitepter) "books"

Жадыбал 6. Учур чак, өткөн чак келээр чак

Символ	Мааниси	Мааниси	Мисал
pres	Present tense	Учур чак (Uçur çak)	Жер айдал+ууда (Jer aydaluuda) "The land is being plowed"
past	Past tense	Өткөн чак (Ötkön çak)	Генерал кат жаз+ды (General kat jazdı) "The general wrote a letter"
fut	Future tense	Келер чак (Keler çak)	Эми өз журтуңа бар+а+сың (Kel-kel) (Emi öz jurtunga barasing) "Now you will go to your homeland"

Кыргызча грамматикада жыныс(род) жок. Демек, үчүнчү жактын жекелик мүчөсүн кабыл алган этиштердин ар башка жыныстагы аяктоолору бирдей. Үчүнчү

жак жекелик жана көптүк сандагы зат атоочтор эч кандай аффиксти албайт. Экинчи жак жекелик же көптүк санда алардын сылык формалары да бар.(-сыз,-сыздар)

Жадыбал 7. Жактар

Символ	Мааниси	Мааниси	Мисал
p1sg	1st person singular	1-жак жекелик сан (1. jak jekelik san)	сурады+м (suradım) "I have asked" окуучу+мун(okuuçumun) "I am a pupil"
p2sg	2nd person singular	2-жак жекелик сан (2.jak jekelik san)	алса+ң (alsang)"If you take" шер+сиң (şersing) "You are a hero"
p3sg	3rd person singular	3-жак жекелик сан (3.jak jekelik san)	келе+т (kelet) "he/she will come) ал бала+zero affix (al bala) "he is a boy"
p1pl	1st person plural	1-жак көптүк сан (1. jak köptük san)	келди+к (keldik) "we came " бардыгы+быз (bardıgıbiz) "all of us"
p2pl	2nd person plural	2-жак көптүк сан (2. jak köptük san)	окуучу+суңар (okuuçu+sungar) "You are pupils" отурасы+ңар (Oturasingar) "You are sitting"
p3pl	3rd person plural	3-жак көптүк сан (3. jak köptük san)	Ал мугалим+zero affix (mugalim) " she/he is a teacher"

Жадыбал 8. Таандык мүчөлөр

Символ	Мааниси	Мааниси	Мисал
px1sg	1st person singular possessive	1-жактын жекелик сандагы таандык мүчөсү (1. jaktın jekelik sandağı taandık müçösü) (-ым)(ım)	кеп+ым=кебим (kep+ım=kebim) "my speech"
px2sg	2nd person singular possessive	2-жактын жекелик сандагы таандык мүчөсү (2. jaktın jekelik sandağı taandık müçösü) (-ың) (-ing)(-ыңыз) (-i ngiz)	жер+ың=жерин (jer+ıng=jering) "your land"
px3sg	3rd person singular possessive	3-жактын жекелик сандагы таандык мүчөсү (3.jaktın jekelik sandağı taandık müçösü) (-ы(-сы)) (-ı(-sı))	кебете+сы = кебетеси (kebete+sı = kebetesi) " his/her appearance"
px1pl	1st person plural possessive	1-жактын көптүк сандагы таандык мүчөсү (1. jaktın köptük sandağı taandık müçösü) (ыбыз)(-ıbiz)	жер+ыбыз=жерибиз (jer+ıbiz=jeribiz) "our land"
px2pl	2nd person plural possessive	2-жактын көптүк сандагы таандык мүчөсү (2. jaktın köptük sandağı taandık müçösü)	марттык+ыңар = марттыгыңар (martık+ıngar =

		(-ыңар) (-ingar) (ыңыздар)(-ingizdar)	marttıgingar) "your generosity"
px3pl	3rd person plural possessive	3-жактын көптүк сандагы таандык мүчөсү (3.jaktın köptük sandağı taandık müçösü) (-ы(-сы)) (-1(-sı))	олжо+ сы = олжосу (oljo+sı=oljosu) "their mining"

Жадыбал 9. Жөндөмөлөр

Символ	Мааниси	Мааниси	Мисал
nom	Nominative	Атооч (Atooç)	алма+zero affix (alma) "an apple"
gen	Genitive	Илик (Ilik)	алма+ нын (alma+nın) "of the apple"
dat	Dative	Барыш (Barış)	алма+ га (alma+ga) "to the apple"
acc	Accusative	Табыш (Tabış)	алма+ ны (alma+nı) "apple as definite object"
loc	Locative	Жатыш (Jatış)	алма+ да (alma+da) "on the apple"
abl	Ablative	Чыгыш (Çığış)	алма+ дан (alma+dan) "from the apple"

Жадыбал 10. Жандуу жансыз заттар.

Символ	Мааниси	Мааниси	Мисал
aa	Animate	Жандуу (Janduu)	адам (adam) "person" ат (at) "horse"
nn	Inanimate	Жансыз (Jansız)	дептер (depter) "notebook"

Жадыбал 11. Мамилелер.

Символ	Мааниси	Мааниси	Мисал
act	Active	Негизги мамиле (Negizgi mamile)	оку (oku) "read"
coop	Cooperative	Кош мамиле (Koş mamile)	таб+ ыш (tabış) "help to find"
ref	Reflexive	Өздүк мамиле (Özdük mamile)	бат+ ын +ды (batındı) "he/she dared"
pass	Passive	Туюк мамиле (Tuuyuk mamile)	оку+ ыл +ган=окулган (oku+ıl +gan=okulgan) "was read"
caus	Causative	Аркылуу мамиле (Arkıluu mamile)	ойно+т+уу (oinotuu) "making play"

Жадыбал 12. Ыңгайлар.

Символ	Мааниси	Мааниси	Мисал
pri	Present indicative	Баяндагыч ыңгай учур чакта (Bayandagıç ingai uçur çakta) (a+жак мүчө) (a+jak müçö)	Такта Шерали отур+а+т (Takta Şerali oturat) "Sheraly is sitting on the throne " (Broken sword)
fti	Future indicative	Баяндагыч ыңгай келер чакта (Bayandagıç ingai	Таз мен дагы бар+а+мын деди .(Taz men dagı

		keler çakta) (-a+жак мүчө) (-a+cak müçö)	baramın dedi) "Taz said that he also will go "
ifi	Past indicative	Баяндагыч ыңгай өткөн чакта (Bayandagıç ingai ötkön çakta) (-ды,-ган,- ыптыр,-чу)	Башын ийкеп отур+ду Акунжан .(Kel -kel)(Başın iykep oturdu Akunjan) "Akunjan sat nodding his head"
imp	Imperative	Буйрук ыңгай (Buyruk ingaai) (-сын,-гын,- гыла,-ыңыз) (-sın,-gın,- gıla,-ingız)	Тилмеч орусча оку+сун (Tilmeç orusça okusun) "Let the translator read it in Russian"
cnd	Conditional	Шарттуу ыңгай (Şarttuu ingai) (-ca)(-sa)	Арсен Саманчин бил+се гана ! (Arsen Samançin bilse gana! (When mountains fall) "If Arsen Samanchin would know !"
prc_vol		Тилек ыңгай (Tilek ingai) (-са экен,-са,-гай элек,- айык,-алы,-гым келет) (- sa eken,-gai elek-ayık,- alı,-gım kelet)	Мен оку+гум келет. (Men okugum kelet) (Kel-kel) "I want to read"
prc_fplan?	Imperfect subjunctive	Ният ыңгай (Niyet ingai) (-мак,-макчы) (-mak,- makçı)	Эмне айт+макчы? (Emne aitmakçı?) (Kel-kel) "What he would like to say ?"

Жадыбал 13. Чектүү этиш формалары.

Символ	Мааниси	Мисал
<cop> <ifi>	Copula past indicative	эле(+agreement) (ele)
<cop> <neg>	Copula negative	эмес(+agreement) emes
<cop> <ger>	Copula gerund	экен(+noun stuff) eken

Жадыбал 14. Чектүү эмес этиш формалары.

	Символ	Мааниси	Мисал
Verbal nouns	1)<ger> gerund	Кыймыл атооч (Kıymıl atooç)	ойлон+уу =ойлонуу (oilon+uu=oilonuu) "thinking"
	2)<ger_perf> perfect gerund		кура+ым =курам (kura+ım=kuram) "composition"
	3)<ger_impf> imperfect gerund		айт+ал+бай =айталбай (ait+al+bai=aytalbay) "being not able to say"
	5)<ger_pres> present gerund		жаз+ыш =жазыш (jaz+ış=jazış) "writing"
	6)<ger_fut>		көр+ар =көрөр (kör+ar=körör)
Verbal adjectives	1)<gpr_pres>	Атооч тук (Atooç tuk)	
	2)<gpr_past>		"???"
	3)<gpr_impf>		бил+мак+сан =билмексен

Verbal adverbs			(bil+mak+san=bilmeksen) "like she/he doesn't know"
	4)<gpr_perf>		шыбыра+ган+дай = шыбырагандай (şıbıra+gan+day) "like whispering"
	1)<gna_perf>	Чакчыл (Çakçıl)	шыбыра+ган = шыбыраган (şıbıra+gan=şıbıragan) "person who whispers"???
	2)<gna_impf>		Алар сүйлөш+а = сүйлөшө кетишти (Alar süylôş+a=süylôşö ketiştı) "They started talking"

Жадыбал 15. Сын атоочтун даражалары.

Символ	Мааниси	Мааниси	Мисал
pst	Positive	Жай даража (Jay daraja)	кооз (kooz) "beautiful"
comp	Comparative	Салыштырма даража (Salıstırma daraja)	коозураак (koozuraak) "more beautiful"
sup	Superlative	Күчөтмө даража (Küçötmö daraja)	эң кооз (eng kooz) "the most beautiful"

Терс маани мүчөсү

Терс маани мүчөсү аффикс (-ба) кыймыл-аракеттин терс белгисин билдирген этиштер менен колдонулат. Дагы бир аффикс (-сыз) зат атоочтордун терс жактуулугун билдирет.

Жадыбал 16. Терс маани мүчөсү.

Символ	Мааниси	Терс маани мүчөсү	Мисал
<neg>	Negative	(-ба) (-ba)	бар+ ба (barba) "do not go"
		(-сыз) (-sız)	суу+ сыз =суусуз (suu +sız=suusuz) "without water"

Жадыбал 17. Менен.

Символ	Мааниси	Мисал
ins	Instrumental	Ал балта менен отун жарып жатат. (Al balta menen otun jarıp jatāt) – "He chops wood with an ax"
soc	Socialative	Айжан классташы менен окуп жатат. (Aijan klasstaşı menen okup jatāt.) "Aijan is studying with her classmate."
con	Conjunctive	Алина менен Актан эшикте ойноп жатышат. (Alina menen Aktan eşikte oynop jatıřat) "Alina and Aktan are playing outside"

Дипломдук ишибиздин алкагында биз “морфологиялык анализатор” программасын иштеп чыктык, аны кыргыз тилиндеги чыгармалардын тексттеринин корпусунун негизинде сынап көрдүк. Бул программа кирүүчү маалымат катары сөз, сүйлөм же кандайдыр бир файлды кабыл алат. Эгер сүйлөм болсо, ал биринчи токенделип, андан соң, ар бир сөз базадан изделет. Базада болсо, анын кайсы сөз түркүмү болгондугу анык болот, эгер жок болсо программанын жардамы менен морфологиялык анализ жүргүзүлөт.

База жогоруда айтылган чыгармалардагы сөздөрдөн турат жана сөздөр төмөнкү таблицада берилгендей сөз түркүмдөрүнө жараша индекстелген.

Жадыбал 18. Сөз түркүмдөрүнүн жана омонимдердин индекси.

Омонимсиз		1 омоним		2 омоним	
Зат атооч	01	Зат атооч	11	Зат атооч	21
Сын атооч	02	Сын атооч	12	Сын атооч	22
Сан атооч	03	Сан атооч	13	Сан атооч	23
Ат атооч	04	Ат атооч	14	Ат атооч	24
Этиш	05	Этиш	15	Этиш	25
Кыймыл атооч	06	Кыймыл атооч	16	Кыймыл атооч	26
Атоочтук	07	Атоочтук	17	Атоочтук	27
Тактооч	08	Тактооч	18	Тактооч	28
Сырдык сөз	09	Сырдык сөз	19	Сырдык сөз	29
Тууранды сөз	0a	Тууранды сөз	1a	Тууранды сөз	2a
Жандооч	0б	Жандооч	1б	Жандооч	2б
Бөлүкчө	0в	Бөлүкчө	1в	Бөлүкчө	2в
Байламта	0г	Байламта	1г	Байламта	2г
Модалдык сөз	0д	Модалдык сөз	1д	Модалдык сөз	2д
3 омоним		4 омоним		5 омоним	
Зат атооч	31	Зат атооч	41	Зат атооч	51
Сын атооч	32	Сын атооч	42	Сын атооч	52
Сан атооч	33	Сан атооч	43	Сан атооч	53
Ат атооч	34	Ат атооч	44	Ат атооч	54
Этиш	35	Этиш	45	Этиш	55
Кыймыл атооч	36	Кыймыл атооч	46	Кыймыл атооч	56
Атоочтук	37	Атоочтук	47	Атоочтук	57
Тактооч	38	Тактооч	48	Тактооч	58
Сырдык сөз	39	Сырдык сөз	49	Сырдык сөз	59
Тууранды сөз	3a	Тууранды сөз	4a	Тууранды сөз	5a
Жандооч	3б	Жандооч	4б	Жандооч	5б
Бөлүкчө	3в	Бөлүкчө	4в	Бөлүкчө	5в
Байламта	3г	Байламта	4г	Байламта	5г
Модалдык сөз	3д	Модалдык сөз	4д	Модалдык сөз	5д
6 омоним		7 омоним		8 омоним	

Зат атооч	61	Зат атооч	71	Зат атооч	81
Сын атооч	62	Сын атооч	72	Сын атооч	82
Сан атооч	63	Сан атооч	73	Сан атооч	83
Ат атооч	64	Ат атооч	74	Ат атооч	84
Этиш	65	Этиш	75	Этиш	85
Кыймыл атооч	66	Кыймыл атооч	76	Кыймыл атооч	86
Атоочтук	67	Атоочтук	77	Атоочтук	87
Тактооч	68	Тактооч	78	Тактооч	88
Сырдык сөз	69	Сырдык сөз	79	Сырдык сөз	89
Тууранды сөз	6a	Тууранды сөз	7a	Тууранды сөз	8a
Жандооч	6б	Жандооч	7б	Жандооч	8б
Бөлүкчө	6в	Бөлүкчө	7в	Бөлүкчө	8в
Байламта	6г	Байламта	7г	Байламта	8г
Модалдык сөз	0д	Модалдык сөз	1д	Модалдык сөз	2д

Эсеп боюнча 30 миндей сөзгө индекс коюлуп чыккан. Бул санды дагы жогорулатып ишти улантууга болот. Индексстер айрым учурлар үчүн кол менен бирден коюлуп чыкса, кээ бир учурлар үчүн Excel де, кандайдыр бир мүнөзгө ээ болгон сөз түркүмдөрүнүн тобуна атайын формуланын жардамы менен автоматтык түрдө коюлуп чыккан. Мисалы, айрым сын атоочтор, аягында төмөндөгүдөй аффикстердин кошулушу менен жасалат, алар: -дай, -тай, -дой, -той, -дөй, -төй, -дей, -тей, -дуу, -дүү, -түү, -луу, -лүү сыяктуу. Эрежеге ылайык, эгерде сөздүн аягында ушундай аффикс бар болсо, анда ал сын атооч 02 деп белгилейбиз. Албетте кээ бир учурлар дал келбей калат бул эрежеге, мисалы Ботой деген сөз сын атооч эмес зат атооч, мындай учурлар өзүнчө каралган.

1	b3	
2	ЁЛКАДАЙ	/02
8	ААРДУУ	/02
12	ААРЧЫЛУУ	/02
43	АБАЖУРДУУ	/02
44	АБАЖУРЛУУ	/02
117	АБЖАЛУУ	/02
153	АБЖОЙЛУУ	/02

Программа, анализ жасоо үчүн кыргыз тилиндеги сөз формаларын түзүү эрежелери бар питон файлдарын жана маалымат базаларын колдонот. Тилекке каршы, кемчиликсиз иштеген анализаторду түзүү абдан татаал иш. Көптөгөн сөз жасоо эрежелерин жазып алгандан кийин да, бүт текст катасыз анализденет деген

кепилдик жок. Көптөгөн өзгөчөлүктөр бар жана жазылган эрежелер сөздү туура эмес окуп, туура эмес сөз формасына алып келиши мүмкүн.

Төмөндө питондо жазылган коддун айрым бөлүктөрү көрсөтүлгөн. Adjectives.py файлында сын атоочту аныктоо үчүн, сын атооч жасалуучу мүчөлөрдүн топтому белгиленген, def get_info_adj_ending(); функциясы, параметрге берилген аффиксти аныктоо менен, берилген сөз сын атооч экени жана анын түрү белгилөө болоёт, болбосо False кайтарат.

```
if root in adj_sup_word:
    return "adj_sup"
elif root in adj_pst_kosh_words or adj_pst_koshmok_words:
    return "adj"
def get_info_adj_ending(ending):
    if ending in adj_noun_to_adj or ending in adj_verb_to_adj or
ending in adj_adj_to_adj:
        return "adj"
    elif ending in adj_comp_ending:
        return "adj_comp"
    elif ending in adj_sup_ending:
        return "adj_sup"
    else:
        return False

adj_noun_to_adj = {'луу', 'туу', 'лүү', 'түү', 'дуу', 'дүү',
                  'сыз', 'суз', 'сиз', 'сүз',
                  'дай', 'тай', 'дой', 'дөй', 'дей', 'той', 'төй', 'тей',
                  'чыл', 'чул', 'чил', 'чүл',
                  'гы', 'кы', 'ки', 'ги',

                  'лык', 'дык', 'дүк', 'лүк', 'лик', 'лук', 'дук', 'дик', 'тык',

                  'лаш', 'леш', 'лош', 'лөш', 'даш', 'деш', 'дош', 'дөш', 'төш', 'таш',
                  'чан', 'чен', 'чон', 'чөн',
                  'дак', 'так',
                  'мер', 'мөр',
                  'кор', 'көр',
                  'кер', 'көр',
                  'көй',
                  'поз', 'боз',
                  'дар',
                  'кел', 'саак',
                  'ий', 'ый' #?
                  }
adj_verb_to_adj = {'гыч', 'гич', 'гүч', 'гуч', 'кыч', 'кич', 'күч', 'куч',
                  'ыңкы', 'өңкү', 'үңкү', 'иңки',
                  'ык', 'ук', 'үк', 'ик', 'өк', 'к', #?
                  'гыр', 'күр',
                  'чаак', 'чээк',
                  'көлөк', 'калак', 'калаң', #уганаак
                  'гыс', 'кыс', 'гүс', 'күс', #гы + сыз
                  'гын', 'гин', 'күн', 'кын',
                  'арман', 'ерман', 'өрмөн', 'ормон', #ар + ман
                  'анаак', 'өнөөк', 'енээк', #аган + ак
```

```

        'ма', 'мо', 'мө', 'ме',
        'ынды', 'анды', 'унду', 'инди',
        'ыш', 'ош',
        'аак', 'ээк',
        'гак', 'гок'}
adj_adj_to_adj = {'гыл', 'гил',
                  'гылт', 'гулт', 'гүлт',
                  'гылтым', 'гүлтүм', 'гултум',
                  'ылжым', 'үлжүм', #каралжым
                  'ылтыр', 'үлтүр',
                  'ыш', 'үш',
                  'гыч',
                  'омук', 'омтук',
                  'чыл', 'чүл',
                  'ча'
                  }

adj_comp_ending =
{'ыраак', 'ураак', 'раак', 'рээк', 'үрөөк', 'ирээк', 'араак', 'рөөк'}
adj_sup_word =
{'эң', 'өтө', 'абдан', 'эбегейсиз', 'ашкан', 'мүлдө', 'чымкый', 'чылк'}
adj_sup_ending = {'кап', 'кып'}
adj_pst_kosh_words = {'улуу-кичүү', 'жаш-кары', 'узун-кыска', 'ачуу-
таттуу', 'жоон-ичке', 'арык-семиз', 'кең-тар',
                     'эстүү-баштуу', 'жылуу-жумшак', 'илимдүү-
билимдүү', 'ак-кара', 'кызыл-тазыл', 'эски-уску', 'майда-чүйдө',
                     'жаман-жуман', 'оңой-олтоң', 'жоон-жолпу', 'одур-
бодур', 'оргу-баргы', 'жакын-жукун',
                     'тааныш-туунуш', 'ачык-айрым', 'жөө-жалаң'}
adj_pst_koshmok_words= {'кара күрөң', 'кара тору', 'кара көк', 'ак
куба', 'жашыл ала', 'сары ала', 'көк ала',
                       'кызыл ала', 'кара кашка', 'жээрде кашка', 'ак
пейил', 'ак көңүл', 'кара көз', 'кырдач мурун',
                       'теке сакал', 'кыргый кабак', 'ак жуумал'}

```

Noun.py файлында `def get_info_noun_ending_from_noun()` ;
 функциясы камтылган жана зат атооч жасоочу аффикстер белгиленген:

```

def get_info_noun_ending_from_noun(ending):
    if ending in noun_noun_to_noun_more_used:
        return "n"
    elif ending in agent_noun:
        return "agnt"
    elif ending in noun_noun_to_noun_less_used:
        return "n"
    elif ending in noun_noun_to_noun_useless:
        return "n"
    else:
        return 'none'
def get_info_noun_ending_from_verb(ending):
    if ending in noun_verb_to_noun_more_used:
        return "n"
    elif ending in noun_verb_to_noun_less_used:
        return "n"
    elif ending in noun_verb_to_noun_useless:
        return "n"

```

```

    else:
        return 'none'
agent_noun = {
    'чы', 'чи', 'чу', 'чү'
}

noun_noun_to_noun_more_used = {
    'кана',
    'лык', 'лик', 'лук', 'лүк',
    'тык', 'тик', 'тук', 'түк',
    'дык', 'дик', 'дук', 'дүк',
    'чылык', 'чилик', 'чулук', 'чүлүк',
    'лаш', 'лөш', 'лош', 'лөш',
    'даш', 'дөш', 'дош', 'дөш',
    'таш', 'төш', 'тош', 'төш',
}
noun_noun_to_noun_less_used = {
    'кер', 'гер', 'көр', #соодагер
    'кор', #ашкор
    'поз', 'боз',
    'стан',
    'ча', 'че', 'чо', 'чө', #бакча
    'чык', 'чик', 'чук', 'чүк', #көлчүк
    'ак', 'ек', 'ок', 'өк', #кезек
    'дырык', 'дурук', 'дүрүк', 'турук', 'түрүк', #ооздурук
    'кеч', #арабакеч
    'тай', #агатай
}
noun_noun_to_noun_useless = {
    'а', 'е', #карта
    'ай', 'өй', #көкөй
    'аке', #тайаке
    'алак', 'олок', 'елек', #кызалак
    'ан', 'ен', 'он', #эрен
    'аң', 'оң', 'ең', 'өң', #жашаң
    'ас', 'ес', 'ос', 'өс', #белес
    'ат', 'от', 'ет', #канат
    'гай', 'кей', 'көй', #карагай
    'гак', 'как', 'кок', #капкак
    'гый', #кудагый
    #'дан', #чыгдан
    'ек', #коёнек
    'жын', 'жин', 'жүн', #кунажын
    'күр', #түпкүр
    'күч', #түпкүч
    'лак', #тайлак
    'мар', 'мер', 'мөр', #башмак
    'мач', 'меч', #жармач
    'са', 'сө', 'се', 'со', #дөңсө
    'той', #жолтой
    'так', 'тек', #жеңилтек
    'чак', 'чек', 'чөк', #келинчек
    'чар', 'чер', 'чор', #букачар
    'чын', #кулакчын
    'ык', 'ик', 'ук', 'үк', #көчүк
    'үр', #дүмүр
}

```

```

noun_verb_to_noun_more_used = {
    "ак", "ек", "ук", "үк", "ык", "ок", "ик", "өк", #калак
    'арман', 'ерман', 'өрман', 'ирмен', 'урман', #аларман
    'гы', 'кы', 'ки', 'ги', 'ку', 'кү', 'гу', 'гү', #ачыткы
    'гын', 'кын', 'гин', 'кин', 'гун', 'гүн', 'кун', 'күн', #качкын
    'гыч', 'гуч', 'гүч', 'гич', 'кыч', 'куч', 'күч', 'кич', #аткыч
    "ым", "им", "ум", "үм", 'ам', 'ом', 'өм', #агым
    "ма", "ме", 'мо', 'мө', #басма
    "мак", "мек", 'мок', 'мөк', #иймек
    'ын', 'ин', 'ун', 'үн', 'н', #агын
    "ынды", "инди", "унду", "үндү", 'енди', #кесинди
    'ооч', 'өөч', #баяндооч
    "ыш", "иш", "уш", "үш", #альш
    'ыч', 'ич', 'уч', 'үч', 'ч', #багыныч, урунч
    'ылдак', 'улдак', 'үлдөк', #арылдак
}

noun_verb_to_noun_less_used = {
    "аан", "оон", "өөн", #жыгаан
    "гак", "гек", "гок", "гөк", "как", "кек", "кок", "көк", #баткак
    'ган', 'кан', 'гөн', 'көн', #жараткан
    'ды', 'ты', 'ту', 'ду', 'тү', 'дү', #бышты
    'ал', 'ыл', 'өл', 'ул', 'ол', #курал
    "ылга", "илге", "алга", "улга", 'өлгө', 'олго', #демилге
    "макей", "мекей", 'мокей', 'мөкей', #алмакей
    "мал", 'мол', 'мөл', #ташымал
    'мыш', 'муш', 'мүш', 'миш', #кылмыш
    "ныч", "нүч", "нч", #таяныч, суранч
    'оо', 'өө', #жайлоо, суроо
    'ыт', 'ит', 'ут', 'үт', 'т', #жайыт
    "уу", "үү", #сүйүү
    'чак', 'чек', 'чок', 'чөк', #бөлчөк
}

noun_verb_to_noun_useless = {
    "а", "ө", "өөн", #жара
    "ага", "ого", "еге", "өгө", #босого
    'ака', #жарака
    'аке', 'еке', #жалжаке
    'алак', 'елек', 'өлөк', #жапалак
    "ан", #жылан
    "анак", "өнөк", #көзөнөк
    "аса", 'есе', 'өсө', #бересе
    'ат', 'өт', #өлөт
    'ач', 'өч', #калач, кемөч
    'аша', 'еше', 'ошо', #киреше
    'га', 'ка', 'ке', 'ге', #тутка
    'галак', 'калак', 'голок', #кайкалак
    'гылык', 'килик', 'гүлүк', #ичкилик
    'дак', 'так', 'дөк', #быштак
    'ды', 'ди', #каралды
    'дык', 'дик', 'тик', #калдык
    'й', #куркулдай
    'лжы', #аралжы
    'ман', 'мөн', #көчмөн
    'мач', #жаңылмач
    'мжы', 'мжу', #сурамжы
    'мта', #байламта
    'мча', #кошумча
    'нак', 'нөк', #жарнак
    'оок', #тырмоок
}

```

```

'оол', 'еел', #кароол
'оон', 'еен', #сүрөөн
'пооч', #айланпооч
'ты', #жарты
'уур', 'үүр', #чукуур
'ы', 'у', #соку
'ымта', #барымта
'ымча', 'умча', #алымча
'ың', 'ең', 'оң', #козголоң
}

```

Pronoun.py файлында ат атоочтор белгиленген:

```

def get_info_pronoun_root(root):
    if root in personal_pronoun:
        result = is_sg_or_pl(root)
        return "prn_pers" + result
    elif root in Demonstrative_pronoun:
        return "prn_dem"
    elif root in Interrogative_pronoun:
        return "prn_itg"
    elif root in Negative_pronoun:
        return "prn_neg"
    elif root in Quantifier_pronoun:
        return "prn_qnt"
    elif root in Indefinite_pronoun:
        return "prn_ind"
def is_sg_or_pl(root):
    if root in pronoun_singular:
        if root in Personal_pronoun_first_person:
            result = "_p1_sg"
        elif root in Personal_pronoun_second_person:
            result = "_p2_sg"
        else:
            result = "_p3_sg"
    else:
        if root in Personal_pronoun_first_person:
            result = "_p1_pl"
        elif root in Personal_pronoun_second_person:
            result = "_p2_pl"
        else:
            result = "_p3_pl"
    return result
personal_pronoun =
{'мен', 'сен', 'сиз', 'ал', 'биз', 'силер', 'сиздер', 'алар'
}
pronoun_singular = {'мен', 'сен', 'сиз', 'ал'
}
pronoun_plural = {'биз', 'силер', 'сиздер', 'алар'
}
Personal_pronoun_first_person = {'мен', 'биз'
}
Personal_pronoun_second_person = {'сен', 'сиз', 'силер', 'сиздер'
}
Personal_pronoun_third_person = {'ал', 'алар'
}

```

```

}

Demonstrative_pronoun =
{'бу', 'бул', 'ушу', 'ушул', 'ошо', 'ошол', 'тиги', 'тигил', 'тетиги', 'тетигил',
',
',
'tээтетиги', 'тээтетигил'
}

Interrogative_pronoun =
{'ким', 'эмне', 'не', 'кандай', 'кайсы', 'кай', 'канча', 'нече', 'кайдан', 'кач
ан',
',
',
'кайда', 'кана', 'кайсыл'
}

Negative_pronoun = {'эч ким', 'эч нерсе', 'эчтеме', 'эч кандай', 'эч
качан', 'эч кайда', 'эч кайдан', 'эч бир'
}

Quantifier_pronoun = {'бүт', 'бүтүн', 'бүткүл', 'баары', 'баардык', 'ар
ким', 'ар нерсе', 'ар бир',
',
',
'ар кайсы', 'ар кандай', 'өз'
}

Indefinite_pronoun = {'алда ким', 'алда эмне', 'алда не', 'алда
кандай', 'алда качан', 'алда канча',
',
',
'алда нече', 'алда кайда', 'алда кайдан', 'кимдир
бирөө', 'кандайдыр бир', 'качандыр бир',
',
',
'эмнегедир', 'негедир', 'бир нерсе', 'бир
неме', 'бирөө', 'бир деме', 'бир демке',
',
',
'бирдеме', 'неме', 'кайсы бир', 'кай бир', 'кээ
бир', 'бири', 'кайсы бирөө'
}

```

Verb.py файлында этиш жасоочу аффикстердин топтому бар жана аны аныктоочу функция камтылган:

```

def get_info_verb_ending(ending):
    result = get_voice(ending)
    result = result + get_mood(ending)
    if ending in verb_other_to_verb_more_used or ending in
verb_verb_to_verb_more_used:
        return "v"

    elif ending in verb_other_to_verb_less_used or ending in
verb_verb_to_verb_less_used:
        return "v"

    elif ending in verb_other_to_verb_useless or ending in
verb_verb_to_verb_useless:
        return "v"

    elif ending in verb_ideophone_to_verb:
        return 'v'

    elif

```

```

else:
    return 'none'

def get_voice(ending):
    if ending in v_voice_reflexive:
        return 'ref'
    elif ending in v_voice_cooperative:
        return 'coop'
    elif ending in v_voice_passive:
        return 'pass'
    elif ending in v_voice_causative:
        return 'caus'
    else:
        return 'act'

def get_mood(ending):
    if ending in v_tense_present:
        return 'ind_pres'
    elif ending in v_tense_past:
        return 'ind_past'
    elif ending in v_tense_future:
        return 'ind_fut'
    elif ending in v_mood_imperative:
        return 'imp'
    elif ending in v_mood_conditional:
        return 'cnd'
    elif ending in v_mood_subjunctive:
        return 'niet'
    elif ending in v_mood_imperfect:
        return 'tilek'
    else:
        return 'imp'

#этиштин мамилелери
v_voice_active = {      # негизги мамиле
}
v_voice_reflexive = {    #оздук мамиле
    'ын', 'ан', 'ун', 'үн', 'ың',   #эгин жүгүнөт
    'ыл', 'ол', 'өл', 'ул', 'үл',   #күн бүркөлдү
    'ылда', 'улда',               #жыркылда
}

v_voice_cooperative = {  #кош мамиле
    'ыш', 'иш',           #кылышты
    'аш', 'еш', 'уш', 'үш' #иштештик
}
v_voice_passive = {      #туюк мамиле
    'ыл', 'ул', 'үл', 'ил',   #жабыл
}
v_voice_passive_exception = {  #туюк мамиле исключение (л тыбыш менен
    буткон этиштер)
    'ын', 'ин', 'ун', 'үн', 'ың'   #тилин
}
v_voice_causative = {     #аркылуу мамиле
    'дыр', 'тыр', 'дур', 'дүр', 'тур', 'түр', 'дир', 'тир', #бастыр
    'кар', 'гар', 'көр', 'тар', #башкар
    'каз', 'коз', 'көз', 'гөз', #көргөз
    'кыз', 'киз', 'гуз', 'гиз', 'гыз', 'из', 'ыз', #тургуз
    'кыр', 'кир', #жаткыр

```

```

    'ыр', 'ир', 'ур', 'үр', #учур
    'сет', #көрсөт
    'т' #байкат
}

#этиштин ыңгайлары
v_mood_indicative_present = { #Баяндагыч ыңгай учур чакта
}
v_mood_indicative_past = { #Баяндагыч ыңгай өткөн чакта
}
v_mood_indicative_future = { #Баяндагыч ыңгай келер чакта
}

v_mood_imperative = { #буйрук ыңгай

    'гын', 'гының', 'гун', 'гүн', 'гунуң', 'гүнүң', 'гин', 'гиниң',
#жүргүн - p2sg
    'кын', 'кының', 'кун', 'күн', 'кунуң', 'күнүң', 'кин', 'киниң',
    'ыңыз', 'иңиз', 'үңүз', 'уңуз', #келиңиз - p2sg politely
    'гыла', 'гиле', 'гула', 'гүлө', #тургула - p2pl
    'кыла', 'киле', 'кула', 'күлө',
    'сын', 'син', 'сун', 'сүн', #уксун - p3
    'чы', 'чи', 'чу', 'чү' #кетчи
}
v_mood_conditional = { #шарттуу ыңгай
    'са', 'се', 'со', 'сө' #келсе
}
v_mood_imperfect = { #каалоо-тилек ыңгай - v_fut
    'гай', 'гей' #баргаймын
    'алы', 'ели', #баралы - p1pl
    'айын', 'ейин', #телейин
}
v_mood_subjunctive = { #максат-ниет ыңгай - v_fut
    'мак', 'мек', 'макчы', 'мекчи' #бармакчы
}

#этиштин чактары
v_tense_present = { #учур чак
    'ууда', 'үүдө', #айдалууда
    'дыр' #арбындыр
}
v_tense_past = { #өткөн чак
    'ыптыр', 'иптир', 'уптур', 'үптүр', 'аптыр', #жамандаптыр
    'ып', 'ип', 'уп', 'үп', 'ап' #келипмин
    'уучу', 'үүчү', 'оочу',
    'чу', 'чү', 'ышчу', #барышчу
    'ды', 'ди', 'ду', 'дү', #келди
    'ган', 'ген', 'гон', 'гөн', #калган
    'кан', 'кен', 'кон', 'кен'
}
v_tense_future = { #келер чак
    'а', 'е', 'й', #окуймун, барат
    'ар', #айтар
    'бас', 'бес', 'бос', 'бөс', #өлбөс
    'пас', 'пес', 'пос', 'пөс'
}

```

```

#этиштин өзгөчө формалары
v_chakchyl_gna_perf = {      #чакчыл
    'ып',

    ''      #
}
v_chakchyl_gna_intent = {
    #ганы
}
v_chakchyl_gna_negative = {
    #майынча
}
v_chakchyl_gna_suc = {
    #гыча
    #ганча
}
v_chakchyl_gna_impf = {      #чакчыл
    'а', 'е', 'ы', 'о', 'ө',
    'й'
}

#атоочтук
v_atoochtuk_gpr_perf = {
    'ган', 'ген', 'гон', 'гөн',      #эгилген
}
v_atoochtuk_gpr_fut_neg = {
    'бас', 'бес', 'бос', 'бөс',      #өлбөс-
    'пас', 'пес', 'пос', 'пөс',
}
v_atoochtuk_gpr_impf = {
    'ар', 'өр', 'ор', 'ер',      #агар-

    'гыдай', 'гидей', 'гудай', 'гүдөй',      #жаагыдай
    'чудай', 'чидей', 'чыдай', 'чүдөй',      #айтчудай
    'гандай', 'гендей', 'гондой', 'гөндөй',      #шыбырагандай
}
v_atoochtuk_gpr_pres = {
    'гыс', 'гус', 'гис', 'гүс'      #түгөнгүс
    'мыш',      #мыш
    #аган
    'максан', 'мексен', 'мөксөн', 'моксон'      #билмексен
}
v_atoochtuk_gpr_past = {
    'уучу', 'үүчү', 'оочу'      #куйкалоочу
}

#кыймыл атооч
v_gerund_ger_perf = {
    '', '', '', ''      #
}
v_gerund_ger_impf = {
    'мак', 'мек', 'мок', 'мөк'      #отурмак
    'май', 'мей', 'мөй', 'мой'      #бермей
}
v_gerund_ger_pres = {
    'аш', 'ыш', 'иш', 'уш', 'үш'      #караш-
}
v_gerund_ger_fut = {

```

```

    '','','','' #
}
v_gerund= {
    'оо','уу','өө','үү' #уруу- жана толтура
}

#how verb is made
verb_other_to_verb_more_used = {
    'а','о','ө','е', #сана-
    'ай','ой','өй','ей', #азай-
    'ар','өр','ир','ор','ер', #агар-

    'ла','да','де','до','дө','та','тө','то','тө','лө','ло','лө', #дарыла-
    'лан','лон','лөн','дан','ден','тан','тен', #тынчтан-
    'лаш','леш','лош','лөш', #акылдаш-
    'даш','деш','дош','дөш',
    'таш','теш','тош','төш',
    'сыра','сире','сура','сүрө', #алсыра-
    'ык','ик','ук','үк', #кечик-
    'ырка','урка','үркө','арка','ирке' #жабырка-
}

verb_other_to_verb_less_used = {
    'ал','ыл','ол','ул','ил', #тирил
    'ке', #теске-
    'сы','си','су','сү', #кыйынсы-
    'сын','син','сун','сүн', #тыңсын-
    'шы','шу','шү','ши', #быкшы-
    'ы','у','и' #бийы-
}

verb_other_to_verb_useless = {
    'ан', #оозан-
    'ыт','от','өт', #жогот-
    'бар','быр', #албыр-
    'лук','дык', #ардык- ,соолук-
    'ра','ро', #куура-
    'са','сө','се', #сууса-
    'сун', #бойсун-
    'ча', #кырча-
    'чий','чий','чуй','чүй', #тикчий-
    'чыла','чиле','чүлө', #кедейчиле
    'ша', #какша-
    'шырай','ширей' #акшырай-
}

verb_verb_to_verb_more_used = {
    'гансы','генси','гөнсү','гонсу', #ачкансы-
    'кансы','кенси','көнсү','консу',
    'гыла','гула','гүлө','гиле', #кескиле-
    'кыла','кула','күлө','киле',
    'ла','ло','лө', #сыйпала-
    'лык','лик','лук','лүк', #буулук-
    'тык','тик','тук','түк',
    'дык','дик','дук','дүк',
    'мала','меле','моло','мөлө', #бурмала-
    'ын','ин','ун','үн', #сүйүн-
    'алан','елен','олон','өлөн' #созолон-
}

verb_verb_to_verb_less_used = {
    'а','ө', #кура-

```

```

'ала', 'еле',      #тебеле-
'ар', 'ер', 'ор', 'өр',      #чыгар-
'тар', 'тор', 'төр',      #кыстар-
'жы', 'жу',      #чоюлжу-
'ка', 'ке', 'кө', #ийке-
'ра',      #кыйра-
'шта',      #ыргышта-
'ык', 'үк', 'ук' #канык-
}
verb_verb_to_verb_useless = {
'ай', 'ей',      #кечирей-
'гала', 'голо',  #тайгала-
'к',      #тынык-
'ман', 'мен',    #иймен-
'маш', 'мыш',    #таймаш-
'па',      #жайпа-
'са', 'се', 'сө', #өксө-
'чы',      #чапчы-
'чыла', 'чула', 'чоло', #барчыла-
'шыр'      #жапшыр-
}

verb_ideophone_to_verb = {
'ай', 'ой', 'ей', 'ый', 'өй',      #торсой-
'ый', 'ий', 'уй', 'үй',      #балкый-
'ылда', 'улда', 'үлдө',      #арсылда-
'ыра', 'ире',      #жаркыра-
'а', 'е',      #чыркыра-
'шы', 'шый'      #быкшы- , быкшый-
}

```

Программанын иштеши

Программа веб тиркеме катары даярдалган. Ал Башкы бет, Морфология, Кыргыз тили, Биз жөнүндө аттуу баракчалардан турат. Негизги бөлүгү катары анализ жүргүзүүчү Морфология бөлүгү эсептелет. Төмөндө көрсөтүлгөндөй эле, анализ жүргүзүү үчүн формага сөз жасабыз жана Анализ кнопкасын басып текшеребиз:

Морфологиялык анализ жасоо үчүн сөз жазыңыз жана анализ кнопокасын басыңыз

Сөздү териңиз

Анализ

Мисалы, *китептер* деген сөздү терели. Анда жыйынтык төмөндөгүдөй болот:

Китептер деген сөздүн анализи

Китептер

Анализ

Берилген сөз	китептер
Сөздүн уңгусу	китеп
Сөз түркүмү	зат атооч
Сөздүн мүчөлөрү	тер. көптүк түр

V БӨЛҮМ

ЖЫЙЫНТЫКТАР

Эсептөө лингвистикасынын узак убакытка созулганына карабастан, анын өнүгүшүнүн ушул этабында көптөгөн идеялар программалык продуктыларда колдонулушун таба элек. Бул илимдин куралдарынын өзү программаларды түзүүдө көптөгөн маселелерди чечүүдө олуттуу жардам бере алат. Бул диссертацияны жазуу процессинде тексттик корпус, текст, компьютердик лингвистика, лемматизатор, лемматизация жана башка түшүнүктөр теориялык жактан түшүнүлгөн. Бул диссертацияда биз кыргыз тилиндеги тексттердин корпусун колдонуп анын баштапкы формасын аныктоочу автоматташтырылган анализаторду түздүк.

Лемматизация издөө алгоритмдеринде кеңири колдонулат. Демек, бул сөздүн суроосу боюнча жыйынтыктарды гана эмес, ал киргизилген формада гана эмес, көбүрөөк натыйжаларды табууга мүмкүндүк берет. Лемматизация тексттин уникалдуулугун текшерүүдө, веб-иштеп чыгууда, программалоодо жана семантикалык өзөктү түзүүдө да колдонулат. Бул программа кыргыз тилиндеги бардык тексттер үчүн колдонулушу мүмкүн. Корпусту анализдөө үчүн, мисалы, кыргыз тилинде сөз түзүү эрежелери менен маалымат базаларын түзүү керек. Лемматизатор кемчиликсиз иштебей тургандыгын белгилей кетүү маанилүү. Бул сөз жасоо эрежелери бар маалымат базаларынын толук эместигинен. Туура натыйжаларды бериши үчүн, жылдарды камтыган убакыт талап кылынышы мүмкүн болгон абдан оор жумушту аткаруу керек. Бул үчүн, сөз түзүүнүн бардык эрежелерин жана бардык өзгөчөлүктөрдү эске алуу керек. Мисалы, англис тилинде лемматизациялоодо, туура эмес этиштер жөнүндө бардык маалыматты маалымат базасына киргизүү керек, ал үчүн зат атоочтор, сын атоочтор жана этиштер ар кандай маалымат базаларын түзүшөт, анда сөздүн ушул конкреттүү бөлүктөрүнүн сөз түзүү эрежелери жазылат. Сөздү леммага алып барууда, кээде ал сөздүн кайсы бөлүгүнө, зат атоочко же этишке тиешелүү экендигин аныктоо маселеси келип чыгат. Бул маселенин абсолюттук чечими азырынча жок, анткени ал үчүн сөз жасоонун бардык өзгөчөлүктөрүн эске алуу зарыл.

КОЛДОНУЛГАН БУЛАКТАРДЫН ТИЗМЕСИ:

- [1] Natural Language Processing in Action Understanding, analyzing, and generating text with Python Hobson Lane, Cole Howard, Hannes Max Hapke, 2019
- [2] Natural Language Processing Explore NLP with machine learning and deep learning techniques, Jalaj Thanaki, 2017, ISBN 978-1-78712-142
- [3] Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
- [4] MachineTranslation and Translation Theory Christa Hauenschild , Susanne Heizmann, 1997
- [5] Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. — СПб.: Питер, 2019. — 368 с.: ил. — (Серия «Бестселлеры O'Reilly»). ISBN 978-5-4461-1153-4
- [6] Грант С. Ингерсолл, Томас С. Мортон, Эндрю Л. Фэррис Обработка неструктурированных текстов. Поиск, организация и манипулирование. / Пер. с англ. Слинкин А. А. — М.: ДМК Пресс, 2015. — 414 с.: ил. ISBN 978-5-97060-144-0
- [7] Astika, Gusti. (2018). Lemmatizing textbook corpus for learner dictionary of basic vocabulary. Indonesian Journal of Applied Linguistics. 7. 630. 10.17509/ijal.v7i3.9813.
- [8] Труды конференции. В 2-х томах. Т 2. — Казань: Издательство Академии наук Республики Татарстан, 2017. — 327 с.
- [9] Биялиев К.А.Справочник по грамматике кыргызского языка. /КыргызскоРоссийский Славянский университет. — Бишкек, 2013г. — 128 с.
- [10] Алгоритм синтаксического анализатора для машинного перевода текстов Б. О. Кочконбаева // Труды VМеждунар. науч.-практ. Конф. Информатизация общества. — Астана, 2016. — С.92-95.
- [11] О морфологическом анализе в приложениях автоматической обработки текста (АОТ) // Б. О. Кочконбаева//Бюллетень науки и практики. — 2018. — Т. 1, № 12.— С.608-612.
- [12] Об оптимизации алгоритма морфологического анализа // Б. О. Кочконбаева, Т. Садыков. — Ташкент, 2018. — С. 293-299 [13] Модель морфологического анализа кыргызского язык// Б. О. Кочконбаева, Т. С. Садыков // Издательство Академии наук Республики Татарстан — Казань, 2017. — С.135-155.
- [14] Математическое моделирование и алгоритм морфологического анализа кыргызского языка [Текст] / Б. О. Кочконбаева, А. Дж. Сатыбаев // Бюллетень науки и практики. — 2019. — Т. 5, № 3. — С. 220-224.
- [15] . Кыргыз тили үчүн сөздүн негизин аныктоо модели [Текст]/ Б. О.

Кочконбаева // Изв. ОшТУ. – 2018. – № 1. – С.24-30.

[16] Программа для ЭВМ «Natural Language Processing. Морфологический анализатор кыргызского языка» [Текст] / Свидетельство КР, №537-Кыргызпатент, 19.12.2018.

[17] Лексический анализатор естественного текста //Б. О. Кочконбаева, Н. Р. Абдыраева // Изв. ОшТУ. – 2014. – С.207-209.

[18] Automatic processing of text in natural language [Текст] /Б. О. Кочконбаева, А. Алдосова // Бюллетень науки и практики. – 2018. – Т.4, № 7. – С. 216-221.

[19] Тестирование программы морфологического анализатора естественного языка [Текст] / Б. О. Кочконбаева, А. Дж. Сатыбаев //Бюллетень науки и практики. – 2019. – Т. 5, № 3. – С. 215-219.

[20] Text Analytics With Python, A Practitioner's Guide to Natural Language Processing, Second Edition. ISBN-13 (electronic): 978-1-4842-4354-1, Dipanjan Sarkar, Bangalore, Karnataka, India.

