КАСИЕВА А. А.
КАДЫРБЕКОВА М. К.
Бишкек, Кыргызстан
aida_kasieva@yahoo.com, medi.kadyr@gmail.com

КЫРГЫЗ ТИЛИНИН УЛУТТУК КОРПУСУН ТҮЗҮҮНҮН АЛГАЧКЫ КАДАМДАРЫ

Аннотация: Дүйнөдө акыркы 30 жылдын аралыгында корпустук лингвистика багытында илимий изилдөөлөр интенсивдүү жүргүзүлүп келүүдө. Кыргыз тил илиминде корпустук лингвистика багытында илимий изилдөө иштери жокко эсе. Ушул өңүттөн караганда, «Кыргыз тилинин улуттук корпусун түзүү" алдыңкы планга чыкты. Бул макаланын негизги максаты- 2019-жылдын 18-апрелинде Германиядагы Заарланд университети жана Кыргыз-Турк «Манас» университети тарабынан түзүлгөн Кыргыз тилинин улуттук корпусу жана аны түзүү процессинин алгачкы кадамдары менен тааныштыруу. Корпустун материалдарын морфологиялык жактан белгилөө үчүн Penn Treebank Tagset программасынын универсалдуу символдору мисал катары колдонулат. Бул макаланын жыйынтыгы кыргыз корпусун мындан ары өнүктүрүүгө мугалимдерди, изилдөөчүлөрдү, студенттерди жана кыргыз тилине кызыккандарды жана аны үйрөнүүчүлөрдү шыктандырат деген үмүттөбүз.

Ачкыч сөздөр: Кыргыз тилинин улуттук корпусу; корпустук лингвистика; сөз түркүмдөрүн белгилөө (энтектөө); корпусту түзүү; Penn Treebank Tagset программасы.

Annotation: Intensive research in the field of Corpus Linguistics has been carried out in the world over the past 30 years. There is no research has been done in the field of corpus linguistics in Kyrgyz linguistics. In this regard, the "Creation of the Kyrgyz National Corpus» came to the fore. The main purpose of this paper is to acquaint with the newly-created Kyrgyz Corpus, which was created on April 18, 2019 by the University of Zaarland in Germany and the Kyrgyz-Turkish Manas University. In this paper we make an attempt to present the initial steps in the process of its creation. Universal symbols of Penn Treebank Tagset are used as a sample for morphological markup of the corpus data. We believe that the results of this article will inspire teachers, students, researches and those, who are interested in the Kyrgyz language for further development of the Kyrgyz National Corpus.

Key words: Kyrgyz national corpus; Corpus Linguistics; part of speech tagging (POS); corpus creation; Penn Treebank Tagset.

Корпустук лингвистикага киришүү

“Корпус” термини 1960-жылдарда жашоонун ар кандай тармагына компьютердик технологиянын терең киришинен улам тил илимине киргизилген. Корпустук лингвистика-тексттерди түзүү менен бирге компьютердик технологиялардын жардамы (электрондук корпус) менен лингвистикалык изилдөөлөрдү жүргүзөт[1]. Корпустук лингвистика жана кыргыз тилинин корпусу тууралуу жалпы маалымат бере турган болсок, бул негизинен бир гана кыргыз тил илими эмес бир канча илимдердин башын камтый турган тармак деп айтсак жаңылышпайбыз. Кээ бир окумуштуулардын пикири боюнча, корпустук лингвистика – бул копьютердик лингвистиканын бир тармагы, башкалар болсо аны методология катары

көрүшсө, ал эми үчүнчүлөр болсо аны тил теориясы катары кабылдашат (лексикология, морфология, синтаксис, семантика, прагматика, дискурс, котормо) [2]. Текст корпусу – бул маалымат алып жүргүчтөргө жайгаштырылган жана автоматтык иштетүүгө арналган, тилдин баардык катмарларын камтып, аларды чагылдырган атайын жыйналган компьютердик база. Маалымат корпустарын конструкциялоодон тышкары, корпустук лингвистика текст корпусунан компьютердик аспаптар (компьютердик программалар) аркылуу ар кандай маалыматтарды алып чыгууга арналган ресурс [3].

Корпусту түзүү процессинин алгачкы баскычтары.

Захаров менен Богдановanын айтымында, корпустутүзүүнүн технологиялык процесси төмөнкү баскычтарда ишке ашуусу мүмкүн.

Тексттерди аталышы жана булактары менен топтоо;

Тексттерди компьютер окуй ала турган форматка которуу. Тексттер электрондук түрдө топтолуусу абзел. Аларды ар кандай жолдор менен алууга болот, мисалы, кол менен терип киргизүү, сканерлөө, көчүрүү ж.б.

Тексттерди алдын ала анализдөө жана алардын үстүнөн иштеп чыгуу. Бул этапта, ар кандай булактардан алынган бардык тексттер талданып, кылдаттык менен текшерилет. Ар бир тексттин аталышы жана авторлору сөзсүз көрсөтүлүүсү керек.

Тексттин конверсиясы.Айрым тексттерди алдын ала бир канча жолу карап, үстүнөн иштеп чыгуу. Конверсия процессинде тексттик эмес элементтер (таблицалар, сүрөттөр) толугу менен алынып салынышы керек.

Графематикалык анализ. Бул процессте тексттерди элементтерге бөлүү, таблица, сүрөт сыяктуу тексттик эмес элементтерди алып салуу, атайын текст элементтерин иштетүү ж.б. кирет. Бул операциялар кол менен эмес, автоматтык түрдө жүргүзүлөт.

Текстти белгилөө. Бул процессте тексттерге жана алардын компоненттерине кошумча маалымат (метадайындар) камтылат. Метадайындарды үчкө бөлүүгө болот: экстра лингвистикалык, тексттик структура боюнча маалымат, лингвистикалык метадайындар. Корпустун метадайындары библиографиялык маалыматтарды, жанрды жана стилди, автор жөнүндө маалыматты, файлдын аталышын, корпусту түзүү үчүн колдонулуучу шаймандардын аталышын жана башкаларды камтыйт. Кошумча лингвистикалык мета-маалыматтар интеллектуалдык текст иштетүүнүн натыйжасы болуп саналат жана кол менен киргизилет, бирок структуралык белгилөө абзацтагы тексттер, сөздөр, сүйлөмдөр жана лингвистиканын туура белгилери адатта автоматтык түрдө жүргүзүлөт.

Автоматтык белгилөөнүн натыйжаларын жөндөө: каталарды кол менен жана жарым автоматтык жол менен оңдоо.

Белгиленген тексттерди статистикалык натыйжаларды камсыз кылган адистештирилген лингвистикалык маалымат издөө тутумунун (корпус менеджери) структурасына айландыруу.

Корпусту колдонууга уруксат берүү. Корпус глобалдык тармакта же CDде таратылышы мүмкүн, ж.б. Ар кандай категориядагы колдонуучуларга ар кандай укуктар жана мүмкүнчүлүктөр берилиши мүмкүн.

Корпусту түзүүнүн жана пайдалануунун ар кандай аспектилерин сүрөттөгөн документти түзүү зарыл [4].

Сөз түркүмдөрүн белгилөө (энтектөө). Табигый тилди иштетүүнүн инструменттеринин бири - бул tagset PENN программасы боюнча англис тилиндеги тегдештирүү символдору төмөндөгүдөй белгиленет [5]. Ал аркылуу изилдөөчү керектүү сөздү же сүйлөмдү дароо таба алат. Эгерде студент “кат” деген зат атоочту корпустан табууну көздөсө, “кат” деген этиш чыкпайт. Компьютер ар түрдүү сөз түркүмдөрүн таап, аларды контекстке жараша зат атооч, сын атооч, этиш деп ажырата турган дараметке эгедер эмес.

Тегдештирүү үчүн универсалдуу символдор			
1	Adjective	ADJ	Сын атооч
2	Normal level	Degree=Pos	Жөнөкөй сын атооч
3	Comparative level	Degree=Comp	Салыштырма сын атооч
4	Superlative level	Degree=Sup	Күчөтмө сын атооч
5	Adverb	ADV	Тактооч
6	Conjunction	CCONJ	Байламта
7	Interjection	INTJ	Сырдык соз
8	Noun	NOUN	Зат атооч
9	Numeral	NUM	Сан атооч
10	Proper noun	PRON	Энчилүү ат
11	Verb	VERB	Этиш
12	Feminine	Fem (gender)	Эркек кишинин аты
13	Masculine	Masc (gender)	Аял кишинин аты
14	Pronoun	PRON	Ат атооч
15	Personal	Prs	Жактама ат атооч
16	Demonstrative	Dem	Шилтеме ат атооч
17	Relative	Rel	Сурама ат атооч
18	Indefinite	Ind	Белгисиз ат атооч
19	Reflexive	Reflexive=Yes	Аныктама ат атооч
20	Singular	Numb=Sing	Жекелик сан
21	Plural	Numb=Plur	Көптүк сан
22	Past simple	Tense=Past	Өткөн чак
23	Present simple	Tense=Pres	Учур чак
24	Future simple	Tense=Fut	Келер чак

Төмөндөгү мисалда “Эр Төштүк” эпосунан алынган үзүндүдө сөз түркүмдөрүн универсалдуу символдор аркылуу белгилейли:

Көп жаманды сурабай,

Бир бала сура, сен ошол.

Сөз түркүмдөрүн белгилөө үчүн атайын стандарт колдонулат

<s> <text id = l>

көп/<тактооч>

жаманды/жаман<зат атооч> + <ды><табыш жөн.>

сурабай/сура<этиш><туунду этиш> + <а><чакчыл мүчө, келер чак> + ба<таңгыч мүчө> + <й><чакчыл мүчө>

/$,/

бир/бир<сан атооч><эсептик сан атооч>

бала/бала<зат атооч><жек.сан>

сура/сура<этиш><туунду этиш><негизги мамиле><буйрук ыңгай> +<а><чакчыл мүчө, келер чак>

<$, >

сен/сен<ат атооч><жактама ат атооч>

ошол/ошол<ат атооч><шилтеме ат атооч>

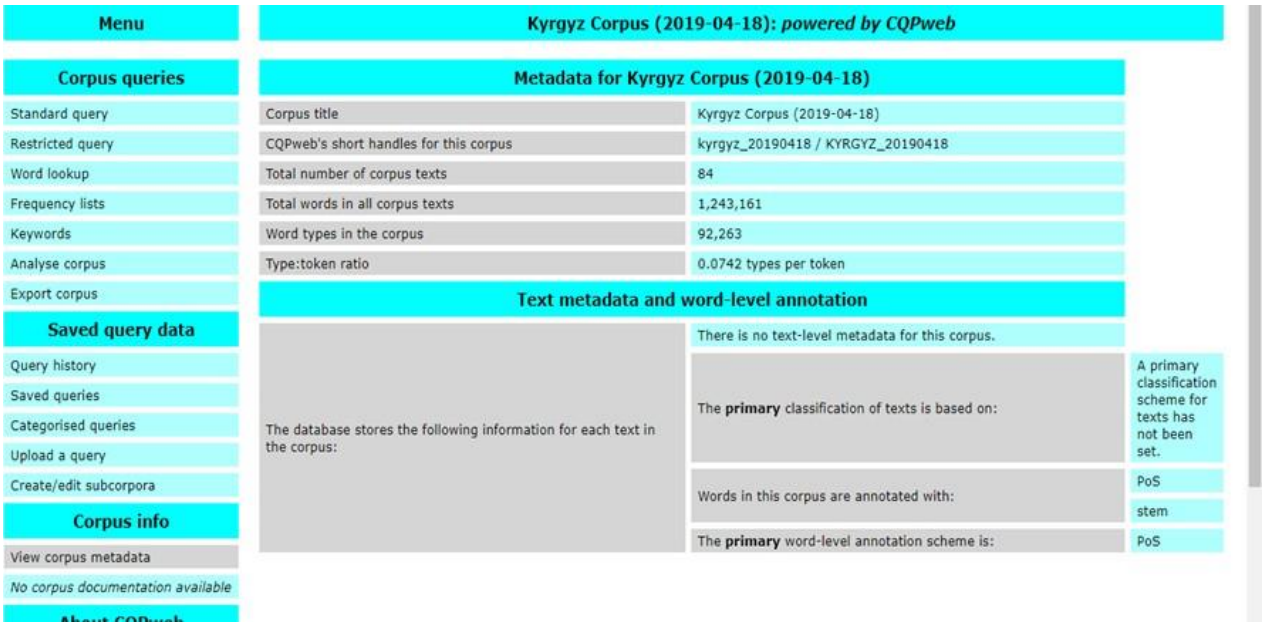
Көп жаманды сурабай, бир бала сура’ – kör (adv= many) jaman (n + dy acc= bad) sura (v +ba neg + j gerund = not ask), bir (num=one) bala (n sg nom= son) sura (v + a fv imp.= ask).

Кыргыз улуттук корпусу

Корпустун аталышы: The Kyrgyz Corpus (2019-04-18), powered by CQPweb. Проекттин алгачкы моделин түптөө боюнча иш-чаралар Германиядагы Заарланд университети жана Кыргыз-Турк «Манас» университетинин демилгелери менен башталган. Корпуста 1,205,888 сөз бар жана бул корпус Заарланд университетинин CLARIN-D лицензиясы алдында.

1-сүрөттө көрүнүп тургандай, Кыргыз корпусу 2019-жылдын 18-апрелинде түзүлгөн. Тексттердин жалпы саны 84, ал тексттерде жалпысынан бир миллиондон ашык сөз бар. Кыргыз Корпусун түзүү үчүн жыйналган, компьютер окуй ала турган тексттер www.bizdin.kg сайтынан алынган. Азыркы учурдагы кыргыз корпусу аркылуу сөздөрдүн жыштыгын, мисалдарды, белгилүү сөз түркүмдүмдөрүнө кирген сөздөрдү, сүйдөмдөрдү жана сүйлөмдөр кайсы китептен алыгандыгы тууралуу маалыматтарга ээ боло алабыз [6].

1-сүрөт. Кыргыз корпусунун үй баракчасы (негизги бети).



Корутунду

Корпустук лингвистика багытында жүргүзүлүүчү изилдөө иши компьютердик так методдорду пайдалануу менен өтө бай тилдик фрагменттердин, эмпирикалык базанын негизинде кыргыз тилинде колдонулган фонетикалык, лексикалык, грамматикалык,

семантикалык, семантика-стилистикалык моделдерди аныктоого кеңири мүмкүнчүлүктөрдү түзөт. Натыйжада, мындай өңүттөгү изилдөө иши кыргыз тилинин структуралык табиятын, анын коммуникативдик, стилистикалык, семантикалык мүмкүнчүлүктөрүн ар тараптан таанып билүүгө жана кыргыз тилин стандартташтырууга көмөкчү болот. Ошондой эле корпустук методология кыргыз тилин турмуштун ар кайсы чөйрөсүндө, ар кыл контекстте колдонууда кездешкен типтүү жана типтүү болбогон, сейрек жана жыш сөз формаларын, грамматикалык категорияларын, семантика-стилистикалык каражаттарын жана башка бүтүндөй тилдик кубулуштарын ар кандай максатта иликтөөгө мүмкүнчүлүк түзөгургандыгы анын артыкчылыгы болуп саналат.

Кыргыз тилинин улуттук корпусунун түзүлүшү тилдик кубулуштардын табиятын дагы терең иликтөөгө, кыргыз тилинин жыштык сөздүгүн түзүүгө, жалпы эле сөздүк курун аныктап тактоого, ошонун негизинде кыргыз тилин сактоого, үйрөнүүгө өбөлгө түзөт. Изилдөөнүн жыйынтыгы прагмалингвистика, корпустук жана компьютердик лингвистика, ошондой эле котормо сапатын баалоодо лингвистикалык изилдөөлөрдү жүргүзүү, тилди үйрөтүү жана үйрөнүү үчүн колдонууга мүмкүнчүлүк берет.

КОЛДОНУЛГАН АДАБИЯТТАР:

1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. –Иркутск, 2011.-5с.
2. Saldanha G. Principles of corpus linguistics and their application to translation studies research-University of Birmingham, 2009.-11 с.
3. Плунгян В.А. Корпус как инструмент и как идеология // Филологический портал. [Электронный ресурс].-2008.
4. -Режим доступа: http://www.philology.ru/linguistics2/plungyan-08.html. – Дата о
5. Захаров В.П. Введение в корпусную лингвистику. –Иркутский государственный университете, 2011.-25с.
6. Sinclair J. “Corpus and Text-Basic Principles”. [Электронный ресурс].-2008. -Режим доступа: (http://ahds.ac.uk/linguistic-corpora/). – Дата обращения: 17.05.2021.
7. Kasieva A., Knappen J., Fischer S., Teich E. A New Kyrgyz Corpus: sampling, compilation, *Annotation*. –Hamburg, 2020.-316с.