УДК 81.32:811.512.154
DOI: 10.36979/1694-500X-2023-23-6-76-82

КОРПУСНАЯ ЛИНГВИСТИКА: ТЕРМИНЫ, СВЯЗАННЫЕ С МОРФОТАКТИКОЙ НОВОСОЗДАННОГО КЫРГЫЗСКОГО КОРПУСА

А.А. Касиева, Н.А. Капарова

Аннотация. Рассматриваются область корпусной лингвистики, которая сегодня делает свои первые шаги в кыргызском языкознании, а также терминология, связанная с этой областью. В работе представлен корпус кыргызского языка, который был создан в рамках совместного проекта КТУ «Манас» и Саарландского университета в Германии. На сегодняшний день он состоит из более чем двух миллионов слов. Один миллион из них составляют произведения известных авторов, занимающих важное место в кыргызской литературе, а второй миллион – тексты различных жанров, которые были опубликованы в газете «Эркин-Тоо» в 2021 году. В данном исследовании рассматриваются вопросы касательно того, чем лингвистический корпус отличается от текстов, размещённых в Интернете, и зачем создавать языковые корпуса при нынешних условиях современных высоких технологий. Основная часть статьи посвящается дискуссии о преобразовании морфологических признаков кыргызского корпуса в предлагаемые эквивалентные символы мировых лингвистических стандартов, предусмотренных для корпусов в специальных языковых порталах. В настоящее время процессу морфологического аннотирования (разметки, тегирования) подверглись более миллиона словоформ кыргызского корпуса. Поскольку процесс морфологической разметки (аннотации, тегирования) является одним из важных и первостепенных принципов языкового корпуса, его выполнение требует много труда и времени большой группы лингвистов. Также в работе подчёркиваются многие преимущества использования корпусной лингвистики в преподавании, изучении языков и её использовании для научных исследований.

Ключевые слова: кыргызский корпус; корпусная лингвистика; морфотактика; Turkic Lexicon Apertium; CQP; веб-платформа; тегирование; морфологическая аннотация; свободные и связанные морфемы.

КОРПУСТУК ЛИНГВИСТИКА: ЖАҢЫ ТҮЗҮЛГӨН КЫРГЫЗ КОРПУСУНУН МОРФОТАКТИКАСЫНА БАЙЛАНЫШТУУ ТЕРМИНДЕР

А.А. Касиева, Н.А. Капарова

Аннотация. Макалада бүгүнкү күндө кыргыз тил илиминде алгачкы кадамдарды жасап жаткан корпустук лингвистика тармагы, ошондой эле бул тармакка байланыштуу терминология каралат. Бул эмгекте Кыргыз-Түрк «Манас» университети менен Германиянын Саарланд университетинин биргелешкен долбоорунун алкагында түзүлгөн кыргыз тилинин корпусу берилген. Бүгүнкү күндө ал эки миллиондон ашык сөздөн турат. Анын бир миллиону кыргыз адабиятында маанилүү орунду ээлеген белгилүү авторлордун чыгармаларын түзсө, экинчи миллионун 2021-жылы «Эркин-Тоо» гезитине жарыяланган ар түрдүү жанрдагы тексттер түзөт. Бул изилдөөдө лингвистикалык корпус Интернетте жайгаштырылган тексттерден эмнеси менен айырмаланат жана эмне үчүн заманбап жогорку технологиялардын азыркы шартында тил корпустарын түзүү керек деген маселелер каралат. Макаланын негизги бөлүгү кыргыз корпусунун морфологиялык өзгөчөлүктөрүн атайын тил порталдарында корпустар үчүн каралган дүйнөлүк лингвистикалык стандарттардын сунушталган эквиваленттүү символдоруна айландыруу жөнүндө талкууга арналган. Азыркы учурда кыргыз корпусунун миллиондон ашык сөз формалары морфологиялык аннотациялоо (белгилөө, тегдөө) процессинен өткөн. Морфологиялык белгилөө (аннотация, белгилөө) процесси тил корпусунун маанилүү жана эң башкы принциптеринин бири болгондуктан, аны ишке ашыруу тилчилердин чоң тобунан көп эмгекти жана убакытты талап кылат. Ошондой эле макалада корпустук лингвистиканы окутууда, тилдерди үйрөнүүдө жана аны илимий изилдөө үчүн колдонууда көптөгөн артыкчылыктар баса белгиленет.

Түйүндүү сөздөр: кыргыз корпусу; корпустук лингвистика; морфотактика; Turkic Lexicon Apertium; CQP; веб-платформа; тегдөө; морфологиялык аннотациялоо; эркин жана көз каранды морфемалар.

CORPUS LINGUISTICS: TERMS RELATED TO MORPHOTACTICS OF THE NEWLY CREATED KYRGYZ CORPUS

A.A. Kasieva, N.A. Kaparova

Abstract. The purpose of this article is to introduce with corpus linguistics, which is now taking its first steps into the Kyrgyz linguistics and the terminology associated with it. Along with this, we present the Kyrgyz Corpus, which was created as a joint project between KTU «Manas» and Saarland University. Now it consists of more than two million words, which is made up of texts of various genres published in «Erkin-Too» newspaper in 2021. This study examines the difference between a linguistic corpus and texts posted on the Internet, why it is feasible to create corpora, and what the necessity is to do that in the current conditions of modern high technology. The main part of the article is devoted to the discussion of transferring the morphological features of the Kyrgyz corpus into the proposed equivalent symbols provided by special language portals for corpora according to world linguistic standards. Currently, the process of morphological annotation (mark up, tagging) has been subjected to more than one million word-forms of the Kyrgyz corpus. Since the process of morphological annotation is one of the most important principles of the linguistic corpus, its performance requires a lot of work and time of a large group of linguists. The paper also emphasizes the many benefits of using corpus linguistics in language teaching, language learning, and its use for scientific researches.

Keywords: Kyrgyz corpus; corpus linguistics; morphotactics; Turkic Lexicon Apertium; CQP; web platform; tagging; morphological annotation; free and bound morphemes.

Киришүү. “Корпустук лингвистика” жана “табигый тил иштетүү” терминдери дүйнөлүк тил илими мейкиндигинде анын заманбап парадигмалары/багыттары катары кеңири орун ээлеп, заманбап технологиялар менен тыгыз байланышта өркүндөп өсүп келе жаткан өз алдынча илим статусуна ээ болууга жетишти.

1963-жылы Х. Кучера менен В.Н. Фрэнсис америкалык англис тилинин “Браун корпусун” түзгөндөн бери [1], “тилдик корпус” тенденциясы дүйнөнүн көптөгөн тилдерин өзүнө камтып, тез жайылып бара жатканы байкалууда. Бул процесстин натыйжасында, айрыкча 1980-жылдардан баштап, дүйнөлүк тилдердин корпустары жана алардын изилдениши тууралуу көп сандагы илимий изилдөөлөр жүргүзүлүп келет. Бүгүнкү күндө биз корпустун артыкчылыктары, баалуулугу жана колдонулушу тууралуу ар кандай далилдерди келтирип, корпустук тил илимдин тарапташтарынын ар кандай пикирлерине күбө боло алабыз. Себеби кээ бир окумуштуулар корпустук лингвистиканы изилдөө куралы жана/же методология катары кабылдаса, башкалары аны дисциплина катары карашат. Мындай жагдай корпустук лингвистиканын теориясынын ар кандай чечмеленүүсүнө алып келүүдө. Ошондой эле корпустук лингвистика өзүнүн табияты боюнча мультидисциплинардык жана/же интердисциплинардык (лингвистика жана компьютердик программалоо) илим болгондугуна байланыштуу, лингвистикалык изилдөөлөргө көптөгөн илимдердин ар кыл тармактарындагы көптөгөн **жаңы түшүнүктөрдү жана терминдерди** өздөштүрүп киргизүүгө өбөлгө түздү. Орус окумуштуулары В.П. Захаров жана С.Ю. Богданованын корпустук лингвистикага берген аныктамасына таяна турган болсок: “Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий” [2], тактап айтканда, корпустук лингвистика – бул тилдик корпустарды түзүү жана аларды колдонуу боюнча жалпы принциптерди, жоболорду иштеп чыгуу менен алектенген компьютердик лингвистиканын бир тармагы катары каралат. Ал эми ушул эле окумуштуулардын пикири боюнча, тилдик же лингвистикалык корпус – бул белгилүү лингвистикалык маселелерди чечүүгө арналган, машина аркылуу окууга ыңгайлаштырылган форматтагы маалыматтардын ири, бириктирилген, структураланган, филологиялык жактан ишенимдүү массиви болуп саналат. Башкача айтканда, корпустук лингвистика багытында жүргүзүлүүчү изилдөөлөр компьютердик так методдорду колдонуу аркылуу тилдик фрагменттердин, эмпирикалык базанын негизинде ар тилдеги фонетикалык, лексикалык, грамматикалык, семантикалык, семантика-стилистикалык, прагматикалык/дискурстук моделдерди аныктап, аларды иликтөөгө кеңири мүмкүнчүлүктөрдү сунуштайт.

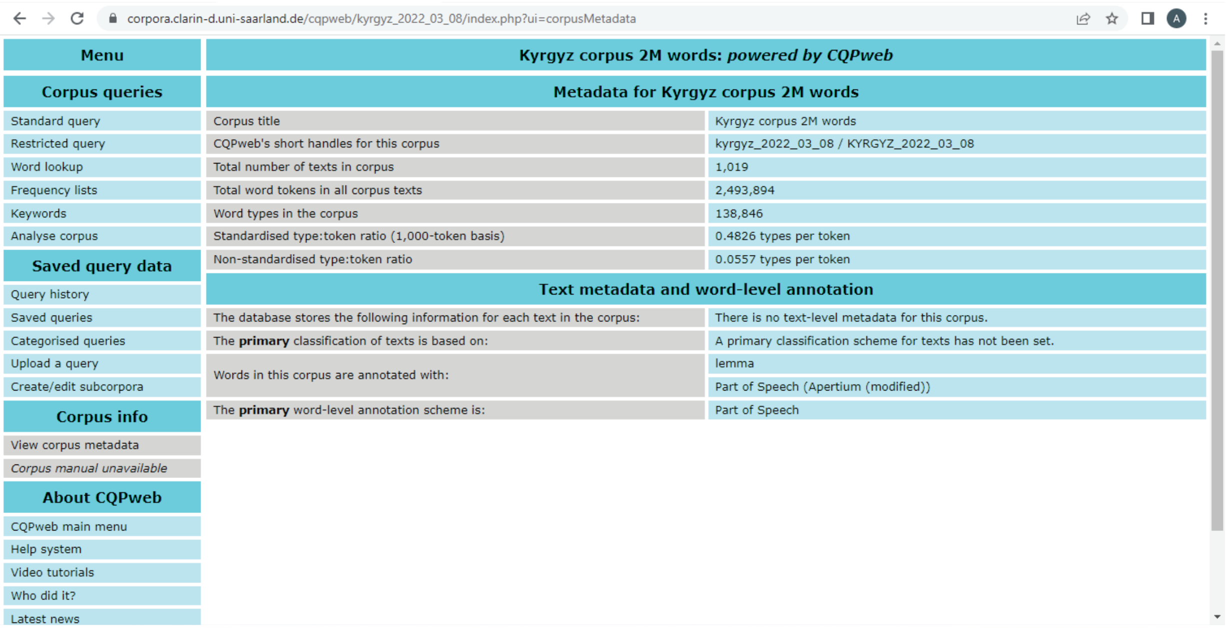
Жогоруда белгиленген маалыматка таянып, корпустук лингвистика багытына тагыраак аныктама бере турган болсок, корпустук лингвистика – бул компьютер жана электрондук тилдик корпустар аркылуу “чыныгы турмушта колдонулган тилди” **эмпирикалык** нукта изилдениши менен алектенген лингвистиканын башка тармактарына караганда салыштырмалуу жаңы багыты болуп саналат. Изилденип жаткан ар кандай тилдик маселенин табиятына жараша, аларды изилдөө үчүн тил илиминде көптөгөн изилдөө тармактары жана алардын усулдары колдонулат. Бул өңүттөн алып караганда, мындай тилдик маселелерди чечүүдө корпустук лингвистикада колдонула турган усулдар диаметралдык мүнөзгө ээ, тактап айтканда, корпустук лингвистика көп сандагы ар түрдүү изилдөө маселелерди чече ала турган усулдардын жыйындысын сунуштайт (Corpus-based Approaches), башкача айтканда, “корпуска негизделген усул”. Мындан улам, эмнеге корпустук лингвистика азыркы учурда мындай тездик менен илимпоздордун сүймөнчүлүгүнө ээ болуп, өнүгүп жаткандыгынын бирден-бир себеби экендигин аңдап түшүнсөк болот. Тил изилдөөчүлөр корпустарда жайгашкан тексттердеги маалыматтардын негизинде тилдин колдонулушунун өзгөчөлүктөрүн иликтеп, алардын кандай жолдор менен далилдене тургандыгын көрсөтүүдө алда канча алдыга карай кадам таштай алышты. Муну менен бирге, теориялык лингвистикада эмпирикалык маалыматтарга жаңыча көңүл бурулуп, корпустук лингвистика усулдарынын алкагында колдонулган техникаларга жана процедураларга кызыгуу артып келүүдө. *Корпустук лингвистика – сандык (квантитативдик) жана сапаттык (квалитативдик) анализдерди айкалыштырган жана программалык камсыздоолорго өзгөчө көңүл буруп, далилдүү лингвистикалык изилдөөлөргө өзгөчө басым жасоо деген түшүнүктү өзүнө камтыган термин катары кабыл алынат.* Анткени, жогоруда айтылгандай, “Корпустук лингвистика деген эмне?” – өз алдынча бир дисциплинабы, методологиябы, парадигмабы же булардын эч бири эмеспи – делген суроолор боюнча ар кандай жүйөлүү аныктамалар жана талкуу-тартышуулар азыркыга чейин орун алып келүүдө.

Кыргыз тилинин корпусу. Кыргыз корпусун түзүү боюнча алгачкы кадамдар Германиядагы Саарланд университети менен Кыргыз-Түрк “Манас” университетинин ортосундагы DAAD алмашуу программасынын алкагында профессор Элке Тейхтин жетекчилиги алдында башталган [3]. Жаңы түзүлгөн кыргыз тилинин корпусу жалпысынан 1019 жазуу түрүндөгү тексттерден турат жана азыркы учурда 2, 493,894 сөздү камтыйт. Ал ар кыл жанрдагы адабий чыгармалардын жана публицистикалык материалдардын санариптештирилген форматынан түзүлгөн. Кыргыз корпусунда орун алган тексттердеги сөздөрдүн кайсы сөз түркүмүнө таандык болгондугу тууралуу да маалыматтар камтылат, башкача айтканда, корпустун морфологиялык жактан белгилениши “энтектелиши” да көрсөтүлгөн. Кыргыз корпусу камтыган сөздөрдүн морфологиялык энтектелиши (эн ыйгаруу) милдети Кыргыз-Түрк «Манас» университетинин гуманитардык факультетине караштуу синхрондук котормо бөлүмүнүн жамааты, магистранттары жана студенттери аркылуу ишке ашырылууда (1–4 сүрөттөр).

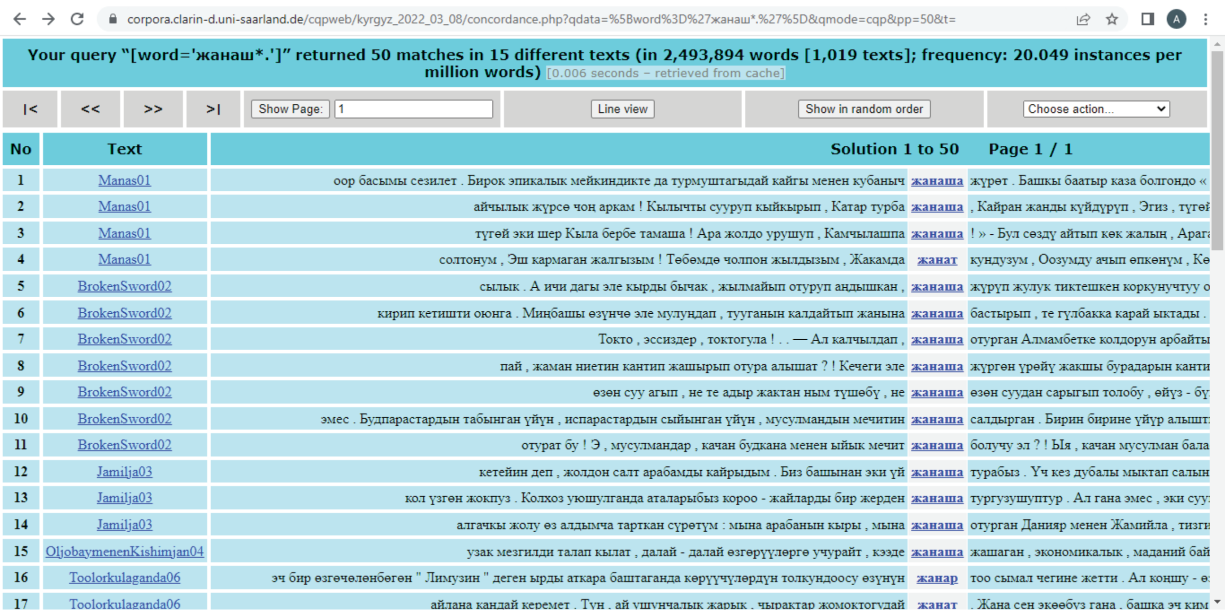
Мисалы, корпустан “жанаш” сөзүн издөөталап кылганыбызда бир-эки секундада мындай натыйжаларга ээ болобуз. Бул сөз 15 (он беш) ар кыл тексттерде 50 (элүү) жолу жолугат. Ал эми анын колдонуу жыштыгы ар бир миллион сөзгө 20.049ду түзөт:

Көрүнүп тургандай, кыргыз корпусунда маалыматтарды башкаруу жана керектүү маалыматты алууга жардам берген веб-негизделген атайын интерфейс (портал) түзүлгөн. Эн ыйгаруу процессинде корпустун бардык сүйлөмдөрү талданып, кыргыз грамматикалык эрежелери боюнча аларга морфологиялык тегдер (белгилер) коюлган. Аны аткарыш үчүн атайы дүйнөлүк тилдердин грамматикалык өзгөчөлүктөрү эске алынып тизмеленген веб-проекттер сунушталат. Кыргыз корпусундагы сөздөрдүн грамматикалык белгилерин камсыздоодо биз Turkic Lexicon Apertium веб-долбоорунун символдорун колдондук [4].

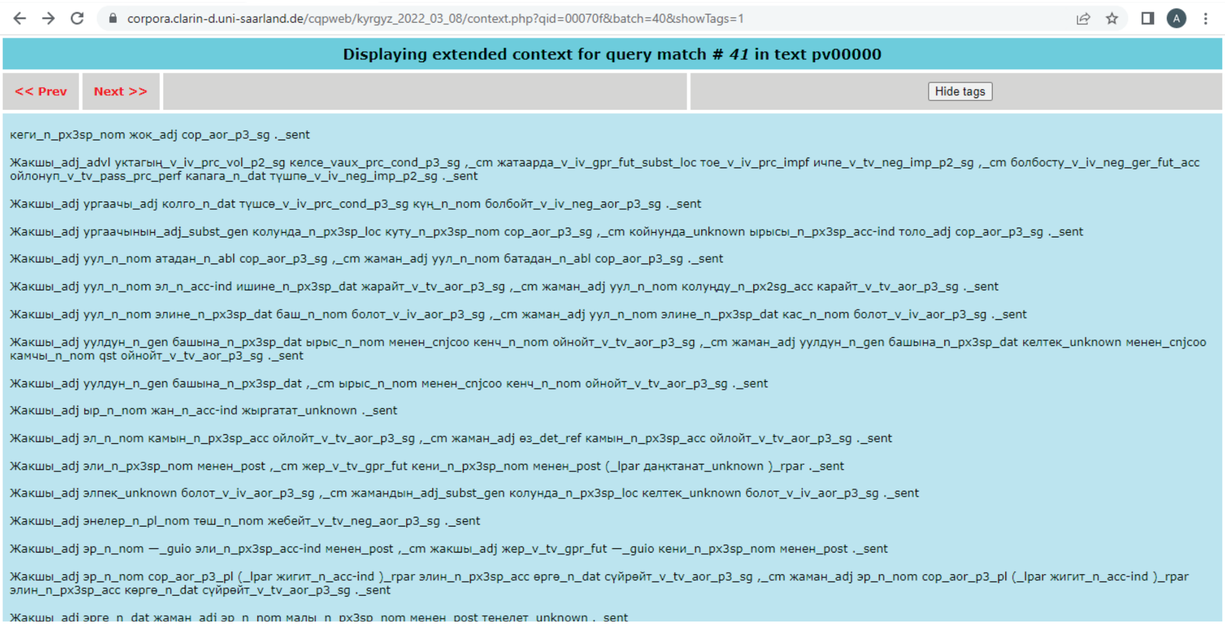
Морфотактика – сөздү түзүүдө морфемалардын бири бирине ээрчишип биригүү процессин сүрөттөө жана алардын биригүү тартибине карата киргизилген чектөөлөрдү билдирет.



Сүрөт 1 – Кыргыз корпусунун түзүлүшүнүн метадатасы



Сүрөт 2 – Кыргыз корпустан “жанаш” сөзүнүн конкордансы жана жыштыгы



Сүрөт 3 – Кыргыз корпусундагы “жанаш” сөзү кездешкен сүйлөмдөрдүн сөз түркүмүнө жараша энтектелиши

Символ	Чечмелениши	Аналог	Мисал
n	Noun	Зат атооч (Zat atooç)	алма (Alma) "apple"
v	Verb	Этиш (Etiş)	кел (Kel) "come"
adj	Adjective	Сын атооч (Sın 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"

Сүрөт 4 – Кыргыз сөз түркүмдөрү Apertium символдору боюнча

Морфотактика тил системасындагы кубулуштарды айырмалоо үчүн колдонулат.

Морфемалар мааниси жагынан бөлүнбөй турган сөздүн маани берүүчү бөлүгү жана тилдеги негизги бирдик. Этимологиялык жактан морфотактика “морфемалардын (морфтордун) бири бирине биригүүсүн (тактикасын) аныктоочу эрежелердин жыйындысы” деп которулат.

Офлазер жана Сарачлар [5] белгилеп кеткендей, морфотактика морфология менен тыгыз байланышта жана алар тилде бирге иштешет. Агглютинативдүү тилдердин көпчүлүгүндө жаңы сөздөрдү жасоо үчүн өзгөчө морфологиялык тактикалар бар.

Морфемалардын валенттүүлүгүн изилдөө үчүн тил системасына көз каранды касиеттерди аныктоо зарыл. Морфотактика бир канча процесстерге бөлүнүшү мүмкүн: флексия, деривация, клитика, бириккен сөздөрдү жана сөз айкаштарын уюштуруучу процесс.

Флексия жөндөлүш же жакталыштын натыйжасында сөздүн жеке грамматикалык маанилерин өзгөртүп, сөздүн ар түрдүү формаларын жасоо процесси аталат. Флексия процессине көптүк, таандык, жак жана чак мүчөлөрдүн, мамиле жана ыңгай мүчөлөрдүн, терс маани жана суроо маани мүчөлөрдүн сөздөргө жалганышы кирет.

Деривация латын тилинен “derivatio” кайтаруу, чектен чыгуу дегенди билдирет. Деривациялык морфемалар сөздүн баштапкы семантикалык негизин, кээде сөз түркүмүн да өзгөртөт. Ошондуктан деривация кээ бир учурларда сөз жасоо деп да айтыла берет.

Бириккен сөздөрдү уюштуруучу процесс эки же андан ашык сөздөрдү кошуу аркылуу бир бүтүн сөздү пайда кылып, ал бир маанини берип, бир басым менен айтылган татаал процесс. Бириккен сөздөрдүн тутумундагы эки сөздүн ичиндеги бир тыбыш башка бир тыбышка өтүп кетет же бир же бири нече тыбыш, кээ бир учурларда муун да түшүп калат: быйыл (бу+жыл), кайнене (кайын+эне).

Клитика – синтактикалык жактан көз карандысыз (мисалы, ат атооч же бөлүкчө), ал эми фонологиялык жактан көз каранды сөз аталат. Негизинен клитика деп муун катары саналбаган сөздөр аталат.

Машиналык котормо үчүн маанилүү дагы бир аспект – бул сөз айкаштары: Сөз айкашы белгилүү бир табигый тилдин анализинде бир сөз катары каралган сөздөрдүн топтому. Багындыруучу сөз сөз айкашынын негизги түгөйүн, багыныңкы сөз анын көз каранды түгөйүн түзөт. Негизги түгөйүнө карай сөз айкашы субстантивдик (негизги сөз – зат атооч), адективдик (негизги сөз – сын атооч), этиштик жана тактоочтук болуп бөлүнөт.

Корутунду. Макалада корпустук линвистика, корпуска аныктама берилип, корпустун Интернет талааларында жайгашкан тексттерден кандай айырмаланаарын жана ага кандай муктаждык бар деген суроолорго жооптор баяндалды. Корпуска байланыштуу жаңы түшүнүктөр жана терминдер сунушталды.

Жаңы түзүлгөн кыргыз тилинин корпусу жалпысынан бүгүнкү күндө 1019 тексттерден турат жана 2,493,894 сөздү камтыйт. Макалада кыргыз корпусунун метадатасы, белгилүү бир сөздүн конкордансы, жыштыгы, контексти жана энтектелиши кандай иштээри көрсөтүлдү. Кыргыз корпусу толугу менен Turkic Lexicon Apertium [6] долбоорунун символдору аркылуу энтектелген.

Изилдөөнүн жыйынтыгы көрсөткөндөй, кыргыз сөз түркүмдөрүн энтектөө кызыктуу натыйжаларды жана алар менен тилди моделдештирүүдө кандай кыйынчылыктар кездешээрин көрсөттүк. Жыйынтыгында кыргыз корпусунда кездешкен кыргыз тилиндеги сүйлөмдөр үчүн сүйлөө тегинин мүчөлөрүнүн кеңири модели сунушталат. Башкача айтканда, бул кадам кыргыз тилинин агглютинативдик татаалдык феноменине каршы чыгууга мүмкүн болгон корпустун сөздөрүндө учураган морфологиялык бүдөмүктүктөрдү болтурбай, так жана деталдуу энтектөө үчүн керектүү тегди колдонуу шартын мисалдардын негизинде баяндап бердик.

Алкыш. *Биз ыраазычылыгыбызды кыргыз тилинин корпусун түзүү тайпасына, өзгөчө Германиядагы Саарланд университетинин профессору Элке Тейх жана Йөрг Кнаппенге билдиребиз.*

Поступила: 28.12.22; рецензирована: 12.01.23; принята: 16.01.23.

Адабияттар

1. *Kučera H.* Computational analysis of Present-Day American English / H. Kučera & W.N. Francis // Providence, RI: Brown University Press. 1967.
2. *Захаров В.П.* Корпусная лингвистика / В.П. Захаров, С.Ю. Богданова. СПб.: СПбГУ, 2013.
3. *Kasieva A.* A New Kyrgyz Corpus / A. Kasieva, J. Knappen, S. Fischer & E. Teich. Sampling, Compilation, Annotation. Hamburg. URL: https://www.zfs.uni-hamburg.de/dgfs2020/programm/abstracts/dgfs2020-clp-kasieva.pdf (дата обращения: 12.01.2023).
4. *Касиева А.А.* Частеречные разметки для нового корпуса кыргызского языка (Инструментарий Turkic Lexicon Apertium) / А.А. Касиева, А.Т. Сатыбекова // Вестник КРСУ. 2020. Т. 20. № 6. С. 67–72. URL: http://vestnik.krsu.edu.kg/archive/154/6520 (дата обращения: 07.12.2022).

5. *Oflazer K.* Turkish Natural Language Processing / K. Oflazer & M. Saraçlar // Springer. 2018.
6. *Washington J.* A finite-state morphological transducer for Kyrgyz / J. Washington, M. Ipasov & F.M. Tyers // Proceedings of the 8th Conference on Language Resources and Evaluation, LREC'2012. С. 934–940. Istanbul. Turkey: European Language Resources Association (ELRA). URL: https://aclanthology.org/www.mt-archive.info/LREC-2012-Washington.pdf. (дата обращения: 12.01.2023).