• ФИЗИКО-МАТЕМАТИЧЕСКИЕ НАУКИ

ӘОЖ 378:625.162.22 https://doi.org/10.51301/vest.su.2021.v143.i1.06

З.А. Маханова*, П.А. Қожабекова, М.А. Сейтжаппар, Н.Е. Сабит
М.О.Әуезов атындағы ОҚМУ, Шымкент, Қазақстан
*e-mail: zlikha70@bk.ru

ҚАЗАҚ ТІЛІНІҢ АВТОМАТТАНДЫРЫЛҒАН МАРКЕРЛІК КОРПУСЫН ӘЗІРЛЕУ

Андатпа. Мақала қазақ тілін технологиялармен жақындастыру туралы. Себебі болашақта бізді қоршаған барлық дүниелер технологиялармен тығыз байланыста болмақ. Күнделікті қолданыстағы жаңа сөздер, қалыптасып жатқан жаңа қызметтік лауазымдар айтылып отырған трансформацияның жаршысы секілді. Ақпараттық технологиялар және интернеттің үдере дамуы қоғам мүшелері арасындағы коммуникациялық байланысты арттыруда. Бұл өз кезегінде жоғары көлемді сандық ақпараттың топтастырылуы мен жинақталуына себеп болды. Іс жүзінде, ақпарат алмасу тек қана технологиялық байланыс қана емес, сонымен бірге күрделі лингвистикалық құбылыс. Адамдардың лингвалды құралдарды, тілді пайдалануы, сөз тіркестері мен сөзді қолдануы, мәліметтердің құрылымдық ортасын түсінуі сияқты мәселелер лингвистика білімінің мәнді саласына айналып, лингвистика мен компьютер ғылымының тоғысқан жерінде компьютерлік лингвистика пәндік аймағы пайда болды.

Негізгі сөздер: корпус, маркерленген корпус, лингвистика, корпустық лингвистика, корпустық технология, токенизация, лемматизация.

Корпус — белгілі бір ережелер бойынша таңдалған және өңделген, тілді зерттеу үшін база ретінде қолданылатын мәтіндер жиынтығы. Олар статистикалық талдау және статистикалық гипотезаларды тексеру, осы тілдегі лингвистикалық ережелерді растау үшін пайдаланылады.

Корпус мәтіндердің шектеулі санынан тұрады, бірақ ол тиісті тілдегі (немесе тіл астындағы) мәтіндердің барлық көлемі үшін типтік лексикограммалық феномендерді барабар көрсетуге арналған. Функционалдылық үшін корпустың көлемі мен құрылымы маңызды. Ұсынылған өлшем міндетке байланысты, өйткені ол зерттелетін феномендер үшін көптеген мысалдар табылуы мүмкін. Статистикалық тұрғыдан қарағанда тіл салыстырмалы түрде сирек сөздердің көп санын (Ципф Заңы) қамтитындығына байланысты, алғашқы бес мың ең жиіліктік сөздерді (мысалы, шығын, кешірім) зерттеу үшін 10-20 миллионға жуық сөз қолданылатын корпус талап етіледі, ал алғашқы жиырма мың сөзді (атаусыз, жүрек қағу, қобалжу) сипаттау үшін жүз миллионнан астам сөз қолданылатын корпус талап етіледі.

Мәтіндердің бастапқы маркерлеуде әрбір корпус үшін міндетті кезеңдер жатады:

- токенизация (орфографиялық сөздерді бөлу)
- лемматизация (сөз қорын сөздік формасына келтіру)
- морфологиялық талдау

Нәтижелерді ұсыну мәселесі.

Үлкен корпустарда бұрын өзекті емес мәселе пайда болады: сұрау бойынша іздеу шектелген уақытта физикалық тұрғыдан қарауға мүмкін емес жүздеген және тіпті мыңдаған нәтижелер (қолдану контекстері) бере алады. Бұл проблеманы шешу үшін іздеу нәтижелерін топтастыруға және оларды ішкі жинақта автоматты түрде бөлуге мүмкіндік беретін (іздеу

нәтижелерін кластерлеу) не олардың маңыздылығын статистикалық бағалаумен неғұрлым тұрақты сөз тіркестерін (коллокация) беретін жүйелер әзірленеді [1].

Кейбір корпустарда талдаудың одан әрі құрылымдық деңгейлері қолданылады. Атап айтқанда, кейбір шағын корпустар толығымен синтаксистік түрде таңбалануы мүмкін. Мұндай корпустар әдетте терең аннотацияланған немесе синтаксистік деп аталады. Морфологияның, семантиканың және прагматиканың аннотациясын қоса алғанда, лингвистикалық құрылымдық талдаудың басқа да деңгейлері болуы мүмкін.

Корпус – корпустық лингвистиканың негізгі түсінігі мен деректер базасы. Корпустардың әртүрлі типтерін талдау және өңдеу компьютерлік лингвистика, сөйлеу және машиналық аударма саласындағы жұмыстардың көпшілігінің мәні болып табылады, онда корпустар сөйлеудің бөліктерін және басқа да міндеттерді таңбалау үшін жасырын маркалық модельдерді жасау кезінде жиі қолданылады. Корпустар мен жиілік сөздіктер шет тілдерін оқытуда пайдалы болуы мүмкін.

Мәтіндік корпустарды құрудың мақсаттылығы:

- нақты контексте лингвистикалық деректерді ұсыну;

- деректердің үлкен өкілеттілігі (корпус үлкен көлемде);

-мысалы, мәтінді графикалық және лексикалық-грамматикалық талдауды жүзеге асыру және т. б. сияқты әртүрлі лингвистикалық міндеттерді шешу үшін бір рет құрылған корпусты бірнеше рет қолдану мүмкіндігі.

Корпус құрудың технологиясы келесі қадамдардан тұрады:

1. Мәтіндердің дереккөз талаптарына сай жинақталуы.

2. Машинаға ыңғайлы форматқа келтіру-құрылымдау. Корпусқа қажетті мәтіндер әртүрлі тәсілмен алынуы мүмкін: сканерлеу, қолмен теру, авторлық көшірме, интернет, түпнұсқалық макет т.с.с.

3. Талдау және мәтінді алдын ала өңдеу. Бұл қадамда әртүрлі дереккөздерден алынған мәтіндер филологиялық түзетулерден өтеді. Корпустың техникалық сипаттамасы мәтіннің библиографиялық және экстралингвистикалық сипаттамасын қамтиды.

4. Конверттеу және графематикалық талдау. Кейбір мәтіндер алдын ала машиналық өңдеудің бірнеше сатысынан өтеді. Бұл орайда (қажет болса) қайта кодтау, мәтіндік емес элементтерді (кесте, сурет) жою не өзгерту, жаңа азат жолдарды өшіру, сызықшаның бірегей қойылуын қадағалау (мысалы: <<–>>, <<–>>) т.б. әрекеттер орындалады. Графикалық талдауда болса: кіріс мәтінін элементтерге (сөйлем, сөз т.б.) бөлу, лексикалық емес элементтерді табу және рәсімдеу, арнайы символдарды (атаулар (аты-жөн инициалдары)) өңдеу, өзге тіл лексемдерін бақылау, сурет атауларын өңдеу жүзеге асырылады.

5. Мәтінді таңбалау (разметка). Таңбалау кезінде мәтін сөздеріне қосымша мәліметтер (метмәлімет) тағайындалады. Метамәліметтерді 3 типке бөлуге болады: экстралингвистикалық-барлық мәтінге қатысты, мәтін құрылымы бойынша мәлімет, лингвистикалық метамәлімет мәтін элементтерін сипаттайды. Метамәліметтер элементтің библиографиялық, жанрлық сипаттамасына қатысты, стиліне және авторына қатысты мәліметтермен қоса файл аты, код тәсілі, таңбалау нұсқасы сынды формалды ақпараттарды да қамтиды. Әдетте айтылған мәліметтерді қолмен теріп енгізеді. Құжатты құрылымдық талдау (абзацтарға бөлу, сөйлемдер ажырату т.б.) автоматты түрде жүзеге асырылады.

6. Автоматты таңбалау нәтижелерін түзету: қателерді жөндеу, бірегей емес бөліктерді теңгеру.

7. Көп аспектілі іздеу мен статистикалық талдауға жағдай жасайтын арнайы лингвистикалық ақпараттық іздеу жүйесіне конвертациялау (қорытқы этап).

8. Корпусқа ену мүмкіндігін жүзеге асыру. Корпус жеке есептеуіш құрылғы жадынан бөлек басқа да тасымалдағыштар арқылы тіпті, глобальды желі арқылы таралуы мүмкін.

9. Сұраныстар тіліне, таңбалау тәсіліне және корпусты қолдану туралы нұсқалық бойынша құжаттамалар даярлау [2].

Іс жүзінде кейбір қадамдарды жүзеге асыру үшін күрделірек технологиялар қажет болуы мүмкін.

Қазіргі уақытта корпустық лингвистика проблематикасындағы қолданбалы шешімдер саласы құралдардың кең спектрімен өңделуде. Алайда, ТТӨ(Техникалық тілді өңдеу) құралдарының кез – келген заманауи іске асырылуы ішінара шешім және толық, әмбебап шешім бірнеше бағыттарда дамуы шығармашылық іздестіруді ынталандыратын осындай құралдардың болашақ әзірлемелерінің қайнар көзі болып табылатынын ескеру қажет. Көптеген практикалық шешімдердің негізі статистикалық тәсілдер болып табылады. Сондықтан қандай да бір математикалық модельдер мен әдістерді қолдана отырып, табиғи тіл жүйесін теориялық ұғыну талпыныстарына негізделген тілдік қызметті зерттеуде аналитикалық бағытты дамыту өзекті болып табылады.

Қазіргі таңда компьютерлік лингвистика саласының жетістіктері екі басыңқы бағытта қолданылуда:

– тілдер семантикасын ұғыну мен зерттеу әдістерін жетілдіру. Бұрын белгісіз болып келген құбылыстар мен заңдылықтарды ашу.

– машина мен адам арасындағы қатынасты жеңілдету. Осы бағытта технологиялық шешімдердің өнімдеріне назар аударылады. Айтып кеткеніміздей, компьютерлік лингвистиканың өнімдері жасанды интеллектіні оқытуда негізгі құрамдас бөлік.

Мәтіндік корпустарды құрудың маңыздылығы:

– нақты контексте лингвистикалықдеректердіұсыну;

– деректердің үлкен өкілеттілігі (корпус үлкен көлемде);

– мысалы, мәтінді графикалық және лексикалық-грамматикалық талдауды жүзеге асыру және т. б. сияқты әртүрлі лингвистикалық міндеттерді шешу үшін бір рет құрылған корпусты бірнеше рет қолдану мүмкіндігі.

Қорыта айтқанда, корпустарды қолдану арқылы зерттеудің статистикалық әдістерін қолдана отырып, тілдік құбылыстар туралы болжамдарды растауға немесе жоққа шығаруға болады. Алдымызға қойған міндеттерді шешу үшін корпустың болуы жеткіліксіз. Мәтінде лингвистикалық ақпараттың болуы қажет. Осылайша белгіленген корпус идеясы пайда болды. Таңбалар сөздердің жиілігін және әр түрлі тіл бөліктері өкілдерінің жиілігін есептеуге көмектеседі. Лингвистикалық таңба сөзге кодты (тэг) беру үшін пайдаланылады, ол сөзді сипаттайтын грамматикалық белгілер жиынтығын білдіреді.

Корпустар пайдаланушылардың бірнеше қайталап пайдалануына арналған, тиісінше, олардың таңбалануы және олардың бағдарламалық жабдықтамасы белгілі бір түрде біріздендірілуі тиіс. Таңбаларға қатысты айтар болсақ, лингвистикалық да, экстралингвистикалық да таңба мәтіндер мен тілдік бірліктерді сипаттаудың кейбір кең таралған және қабылданған қағидаттарына негізделуі тиіс. Таңбалардың параметрлері мен олардың мәндері "табиғи" жеткілікті болуы тиіс, яғни жалпы қабылданған ғылыми сипаттамаға сәйкес болуы тиіс. Бағдарламалық қамтамасыз етуге келетін болсақ, ол типтік сұраныстарды өңдеуді және типтік тапсырмаларды шешуді қолдайды. Форматтарды толтыру мен құрылымдарды біріздендіру үлкен маңызға ие. Деректерді ұсынудың бірыңғай пішімдері көптеген жағдайларда бірыңғай бағдарламалық жасақтаманы пайдалануға және корпустық деректермен алмасуға мүмкіндік береді. Корпустарға қатысты стандарттау, деректер типтерінің үйлесімділігі әр түрлі корпустарды салыстыру тұрғысынан да маңызды.

ӘДЕБИЕТТЕР

[1] Архипов А.В. Разметка лингвистическая [Электронный ресурс]. – Режим доступа:http: //www.lomonosov-fund.ru/enc/ru/encyclopedia:0127221#r1_1.Дата доступа: 15.03.2014.

[2] Азарова И.В., Алексеева К.Л., Захарова Л.А. Разметка текстовых фрагментов в корпусе агиографических текстов СКАТ//Труды международной конференции «Корпусная лингвистика – 2006». – СПб: изд-во С.-Петерб. ун-та, Изд-во РХГА, 2006. – С. 16-24.

REFERENCES

[1] Arhipov A.V. Razmetka lingvisticheskaja [Jelektronnyj resurs]. – Rezhim dostupa: http: //www.lomonosov-fund.ru/enc/ru/encyclopedia:0127221#r1_1.Data dostupa: 15.03.2014.

[2] Azarova I.V., Alekseeva K.L., Zaharova L.A. Razmetka tekstovyh fragmentov v korpuse agiograficheskih tekstov SKAT//Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika – 2006». – SPb: izd-vo S.-Peterb. un-ta, Izd-vo RHGA, 2006. – S. 16-24.

З.А. Маханова*, П.А. Кожабекова, М.А. Сейтжаппар, Н.Е. Сабит
ЮКГУ имени М.Ауэзова, Шымкент, Казахстан
*e-mail: zlikha70@bk.ru

РАЗРАБОТКА АВТОМАТИЗИРОВАННОГО МАРКЕРНОГО КОРПУСА КАЗАХСКОГО ЯЗЫКА

Аннотация. Статья о сближении казахского языка с технологиями. Потому что в будущем все окружающие нас мир будут тесно связаны с технологиями. Как будто новые слова в повседневной жизни, новые формируемые служебные должности-это вестник трансформации. Информационные технологии и процессное развитие интернета увеличивают коммуникационные связи между членами общества. Это, в свою очередь, послужило поводом для консолидации и накопления высокоразвитой цифровой информации. На самом деле, обмен информацией не только технологическая связь, но и сложное лингвистическое явление. Такие проблемы, как использование людьми лингвальных средств, языка, употребление словосочетаний, понимание структурной среды данных, стали существенной сферой лингвистических знаний, в сочетании с лингвистикой и компьютерной наукой возникла предметная зона компьютерной лингвистики.

Ключевые слова: корпус, маркированный корпус, лингвистика, корпусная лингвистика, корпусная технология, токенизация, лемматизация.

Z.A. Makhanova*, P.A. Kozhabekova, M.A. Seitzhappar, N.E. Sabit
SKSU named after M.Auezov, Shymkent, Kazakhstan
*e-mail: zlikha70@bk.ru

DEVELOPMENT OF AN AUTOMATED MARKER CORPUS OF THE KAZAKH LANGUAGE

Abstract. Article about the convergence of the Kazakh language with technologies. Because in the future, all the world around us will be closely connected to technology. It is as if new words in everyday life, new positions being formed, are the messenger of transformation.Information technologies and the development of the Internet strengthen communication links between members of society. This, in turn, led to the consolidation and accumulation of highly developed digital information. In fact, information exchange is not only a technological connection, but also a complex linguistic phenomenon.Problems such as people use of lingual means tongue, the use of phrases, understanding the structural data environment, have become a significant field of linguistic knowledge, combined with linguistics and computer science arose the subject area of computational linguistics.

Keywords: corpus, labeled corpus, Linguistics, corpus linguistics, corpus technology, tokenization, lemmatization.