Дәріс №2
2-дәріс тақырыбы: Корпус лингвистикасының ұғымдары мен бағыттары
Сұрақтар:

1. Корпус лингвистикасының негізгі ұғымдары
2. Корпус лингвистикасының міндеттері мен бағыттары

1. Корпус лингвистикасының басты ұғымы – *корпус*; *лингвистикалық корпус* деп те аталады. Ағылшынша *linguistic corpus* немесе *text corpus*, көпше *linguistic corpora* деп аталады.

Бірлескен авторлықпен шыққан «Қазақ тілінің жиілік сөздігі» еңбегінде корпусқа мынадай анықтама береді: «... корпус дегеніміз – белгілі бір ұлт тіліндегі әртүрлі жанрдағы, әртүрлі автордың, әртүрлі кезеңдегі электронды нұсқадағы белгілі бір іздеу бағдарламасы бойынша жұмыс істейтін, әртүрлі лингвистикалық анықтамалар берілген мәтіндер жинағы, мәтіндік қор» (1, 5 б. Қазақ тілінің жиілік сөздігі. А.Қ. Жұбанов, А.Ә. Жаңабекова, Б.Д. Карбозова, А.Қ. Қожахметова. – Алматы, 2016) (https://tilalemi.kz/viewer/viewer.php?file=/books/1057.pdf)

А.Жұбанов, А.Жаңабековалар (2017): «лингвистикалық немесе тілдік, мәтіндер корпусы» деп: «нақты тілдік мәселелердің шешімін табуға арналған аса үлкен көлемдегі машине (компьютер) оқи алатындай түрде көрініс табатын, бірыңғайланған, құрылымдалған, белгіленген (шартты белгілер қойылған), филологиялық тұрғыда компетентті саналатын тілдік деректер ауқымы» (12б). Бұл анықтама А.Захаров пен С.Богдановалардың анықтамасының тура аудармасы. Қараңыз: В.П. Захаров пен С.Ю. Богданова «Корпусная лингвистика: Учебник для студентов направления «Лингвистика» оқулық-еңбегінде лингвистикалық корпусты былай түсіндіреді: «Под **лингвистическим, или языковым, корпусом** текстов понимается большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [5, 7 б.]. Сөйтіп орыс лингвистерінің анықтамасын аударып пайдаланады да, 13-бетте қайтадан осы екі орыс авторының жоғарыдағы анықтамасын орыс тілінде түпнұсқада келтіреді (13б).

«...*лингвистикалық корпус* іздеу жүйесі қамтылған түрлі лингвистикалық параметрлермен белгіленген арнайы жинақталған мәтіндер жиынтығы ретінде анықталады» (Г. Б. Мәдиева, С. Б. Бектемірова, Ж.Ж. Күзембекова. Компьютерлік лингвистика. Оқу құралы 108-б).

«Корпус – белгілі бір тілдегі электронды формадағы мәтіндерді жинақтауға бағытталған ақпараттық-анықтамалық жүйе. Корпус қазіргі түсінік бойынша Корпус – ол деректердің компьютерлік базасы». (Корпустық лингвистика: негізгі терминдер мен түсініктердің оқу сөздігі / құраст.: Г.Б. Мәдиева, С.Б. Бектемірова, Н.А. Исмайлова. – Алматы: Қазақ университеті, 2018. — 40 б.)

Шетел зерттеулерінде бірқатар анықтамалар бар, соларды жинақтап айтсақ:

- корпус дегеніміз бұл тілді немесе тілдің вариациясын барынша толық бейнелеу үшін сыртқы критерийлерге сәйкес таңдалған электрондық түрдегі мәтіндер үзінділерінің жиынтығы (Джон Синклер).
- «корпус – лингвистикалық талдау мен сипаттауға негіз бола алатын жазбаша мәтін немесе транскрипцияланған сөйлеу жиынтығы» (Грейм Кеннеди An Introduction to CL. London, 1998);
- Корпус термині лингвистикада қолданғанда төрт негізгі тақырып бойынша қарастырыла алады деген: іріктеу және репрезентативтілік; соңғы өлшем; машинада оқылатын пішін; стандартты сілтеме (sampling and representativeness; finite size; machine-readible form; a standard reference) (Tony McEnery and Andrew Wilson. Corpus Linguistics: An Introduction. 2001, Оқисыздар - 29-32 б.).

Корпусты жекелеген зерттеушілер өздерінің зерттеу мақсаттарына қарай жасай алады және/немесе тілде бар дайын корпустарды зерттеу мақсатында пайдалана алады.

Корпустар ғылыми зерттеулер жүргізгеннен басқа [3, 166-167; 4, 60]:

- *лексикографияда* көпмағыналы сөздерді анықтау және сөздік жасауда;
- *грамматикада* морфемалардың жиілігін анықтау, сөз тіркестерінің және сөйлемдердің типін анықтауда;
- мәтін *лингвистикасында* мәтін типтерінің дифференциасын, азатжолдың ішкі байланыстары мен азатжол арасын анықтауда және т.б.;
- *автоматты аудармада* мәтіндердің параллель мәтіндердегі аударма баламаларын, бірнеше баламаға ие сөздерді іздеуде;
- *оқыту мақсатында* шығарма үзінділерін, оқу орындары үшін үлгілер, оқу құралдарын жасау, дәйексөз алуда және т.б.
- *тестілеуде* автоматты анализ бен сөйлеу синтезі және тағы басқа жағдайлар үшін қолданылады. (Компьютерлік лингвистика. Оқу құралы 108-б)

Басқа терминдері: аннотация, кодтау, тэг, тэгтеу, белгілеу, токен, токендеу, екіұштылық т.т. (**қараңыз:** Корпустық лингвистика: негізгі терминдер мен түсініктердің оқу сөздігі / құраст.: Г.Б. Мәдиева, С.Б. Бектемірова, Н.А. Исмайлова. – Алматы: Қазақ университеті, 2018. — 40 б.)

Аннотация – корпус дерегіне қосымша ақпаратты қолдану процесі.
encoding кодтау Әдетте кодтау корпусты құрастырудың бес кезеңінің соңғысы болып табылады, кейде оны аннотация, тэгтеу немесе белгілеу деп те атайды. **Кодтау** - мәтіндердегі элементтерді - абзац үзілістерін, айтылыс шекараларын және т.с.с. – корпуста стандартталған әдіспен көрсету тәсілі.
Тэг (ағыл. tag – затбелгі) – аталған жапсырма, дескриптор, гипермәтінді белгілеу тілінің элементі; Лингвистикалық белгілеу үдерісіндегі сөздерге арналған код. Әрбір код осы сөзді сипаттайтын грамматикалық белгілердің белгілі бір жиынына сәйкес келеді. (Корпустық лингвистика : негізгі терминдер мен түсініктердің оқу сөздігі / . — Алматы : Казахский национальный университет им. аль-Фараби, 2018. — 40 с.)
tagging тэгтеу белгіленім Корпус деректеріне аннотацияның қосымша деңгейлерін қолдану әрекеті үшін бейресми термин. Тэг әдетте кодтан тұрады, оны фонемаға, морфемаға, сөзге, сөз тіркесіне бірнеше жолдармен, мысалы, стандартты жалпылама белгілеу тілі (Standard Generalised Markup Language (SGML) элементтерін қолдану арқылы немесе сөз мен оның тэгінің арасына астыңғы сызықша таңбасын қолдану арқылы бекітуге болады (for example cat_NN1 is the word cat tagged as a singular common noun using the CLAWS C7 tagset). Тэгтеу көбінесе бағдарламалық жасақтама көмегімен автоматты түрде жүзеге асырылады. Алайда адам тарапынан редакциялау көбінесе соңғы кезеңде талап етіледі. (See also ditto tags, part-of speech tags, portmanteau tag, problem-oriented tagging, stochastic tagging, tagging errors, tagset, tag stripping.)
Токен (ағыл. tokens) – бұл сөзқолданыс, лексемаға сәйкес келетін негізгі бірлік
ambiguity екіұштылық Корпус аннотациясында мәтіннің бір жерінде екі потенциалды тэгті таңдау мүмкіндігі болған жағдайда, нақты шешім қабылдау әрдайым мүмкін бола бермейді. Мысалы, сөйлеу бөлігін тэгтеу кезінде кейбір сөздердің грамматикалық категорияларын анықтау қиынға соғады, мысалы: I put it down. (Is put the past participle or past tense form?) Bill was married. (Is married an adjective or verb?) It’s broken. (Is ’s a contraction of has or is?) There is a question on gardening. (Is gardening a noun or verb?)

Осы анықтамалардың бірінде корпус ауызша да, жазбаша да болады деп айтылады. Осыған орай бір көзқарас бар: лингвистикалық корпустар ауызша да, жазбаша да, баспа түрінде де болмайды, төртінші формасы бар, ол - машиналық ортадағы мәтіндерді -

сандық мәтінді *digital text* білдіреді. Бұл пікірмен келіспеуге де болады. Корпус мәтіндердің жиынтығы екені белгілі, онымен жұмыс жасалады.

2. Корпус тілдің көптеген салалары мен бағыттарында қолданылады. Сөздерді зерттеуде (лексикология), тіл педагогикасы мен әдістемесінде, тілді үйретуде, әдебиетте, аудармада, әлеуметтік лингвистикада, дискурсты зерттеуде т.т.

Корпус сөздерді зерттеуде таптырмас дереккөз және талдау құралы. Ағылшын тілінде қанша сөз бар деген сұрақ жауап беру мүмкін емес тіпті үлкен көлемді корпсутар бола тұрса да, Барлық корпустар сөздер қанша жиі қолданылады деген сұраққа жауап береді. Сөздер жиілігінің тізімдері (word frequency lists) жасалады. Сөздер жиілігінің тізімдері мәтіннің негізгі бөлігін салыстырмалы түрде шағын сөздер жиынтығы құрайтынын көрсететін бастапқы нүкте ретінде қызмет етеді. Өте жоғары жиілікті элементтердің көпшілігі грамматикалық сөздер болуы таңқаларлық емес: BoE-дің алғашқы он леммасы - the, be, of, және, a, in, to (infinitive particle), have, to (предлог) және ол.

Корпус арқылы сөздердің байланысы мен сәйкестігін анықтауға үлкен көмегі тиеді. Ағылшын тілінде конкорданс (Сәйкестік — контексте қолдану мысалдары). Сөздердің сөздермен сәйкестігін анықтап береді. (**көрсету**).

Лексикологияда фразалар мен идиомалар тұрақты тіркестерді, полисемияны, метафора басқа фигураларды тағы көптеген лексикалық бірліктерді зерттеуге де қолайлы. Қызықты нәтижелер алауға болады. мысалы ағылшын тілді корпус лингвистері идиомалар, мақал-мәтелдерді және басқа да осыған ұқсас бірліктерді корпустық зерттеу олардың көпшілігі сирек кездесетінін көрсетеді, негізінен журналистика мен көркем әдебиетте кездеседі.

Әлеуметтік лингвистикада корпусты пайдалану да тілдік ұтымды нәтижелер алуға мүмкіндік береді. Әлеуметтік лингвистика – тілдік вариация мен оның әлеуметтік мәнін зерттейтін тіл білімінің саласы. Социолингвистер тілді жеке қолданушылар арасындағы және тіл вариациялары арасындағы айырмашылықтарды зерттейді. Әлеуметтік лингвистика зерттейтін тақырыптар әрттүрлі. Әлеуметтік лингвистикалық факторларға жас, жыныс, әлеуметтік тап және этностың негізгі демографиялық категориялары, сондай-ақ сөйлеу жағдайының формальдылық дәрежесі, сөйлеушінің әлеуметтік желілері және т.б. жағдаяттық категориялар жатады. Әлеуметтік лингвистика сонымен қатар тіл вариацияларының жалпы сипаттамаларына, яғни аймақтық диалектілер, тілдің стандартты/стандартты емес түрлері, көптілділік, тіл саясаты, стандарттау т.б. назар аударады. Мысалы, сөйлеу тілін зерттеу (Елан таныстыру).

Біз қазіргі корпус лингвистикасының негізгі бағыттарын қысқаша сипаттайтын боламыз («Корпусная лингвистика» А.Б. Кутузов бойынша).

1. **Біріншіден**, лексикографиялық зерттеулер, сөздіктер жасау. Ағылшын тілінің барлық дерлік сөздіктері (Collins, Webster, MacMillan и т.д.) үлкен корпустар негізінде жасалған; бұл сөздікті репрезентативті етуге мүмкіндік береді.
2. **Екіншіден**, корпустарды зерттеу тілдердің лексикалық құрамы жайлы, белгілі бір сөздердің қолданылу жиілігі туралы дәлді деректер алуға мүмкіндік береді.
3. **Үшіншіден**, корпус лингвистикасы тілдердің сөздік құрамындағы өзгерістерді, оның түрлі варианттылығын (неологизмдердің пайда болуын және жоғалуын) зерттейді.
4. **Төртінші** бағыт, корпус лингвистикасы табиғи тілдердің грамматикасын зерттейді, оның ішінде қандай да бір грамматикалық құбылыстардың үйлесімділігін зерттейді.
5. **Бесіншіден**, мәтіндерді зерттейді. Мысалы корпустарды қолдана отырып мәтіндердің статистикалық сипаттамасы - сөздер мен сөйлемдердің орташа ұзындығы, сөздердің типтік тіркесі т.б. - арқылы функционалды стильді анықтай аламыз.

6. **Алтыншыдан**, КЛ лингводидактикада, яғни шет тілдерді үйренуде белсенді қолданылады. КЛ оқытылатын тіл туралы дәлді сандық мәліметтер – жиілігі жоғары лексиканың құрамы, белгілі бір грамматикалық түзілімдердің қолданылу ықтималдылығы т.т. туралы – деректер береді.
7. **Соңғысы**, аудармада көптілді корпустар аударма эквиваленттері туралы деректер береді. Корпустар автоматты машиналық аударма жүйесін жасауда маңызды.