9-апта
Дәріс №17-18
Корпусқа негізделген зерттеу әдісі

1. Корпусқа негізделген талдау әдісі туралы
2. Корпус құру және аннотациялау

1. Корпус лингвистикасы лингвистиканың танымал саласы, ол компьютерлік бағдарламалық құралдың көмегімен электронды түрде сақталған мәтіндердің өте үлкен жиынтығын талдауды қамтиды. «Корпус» сөзі латын тілінен аударғанда «дене» *body* дегенді білдіреді, сондықтан корпус мәтіндердің «денесі» болып табылады (a 'body' of texts.). Ғалым А.Қ. Жұбанов «Қазақ тілінің жиілік сөздігі» еңбегінде корпус терминіне 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.].

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» [7 б.]. – қазақ тіл біліміндегі еңбек тура осы кітаптың ізімен шыққан.

Жалпы бұл пәннің аталуына қатысты пікірлер басқа тілдің салалары сияқты бірізді емес. Сол себептен «Корпус лингвистикасы дегеніміз не немесе не болуы керек?» деген сұрақ лингвистер арасында көп талқыланған. Бұл сұраққа жауаптардың, корпус линвгистикасына берілген анықтамалар мен сипаттамалардың әртүрлілігі таңқаларлық дейді Шарлотта Тейлор. Корпус лингвистикасы - бұл құрал, әдіс, әдістеме, әдістанымдық тәсіл, пән, теория, теориялық көзқарас, парадигма (теориялық немесе әдіснамалық) немесе осылардың жиынтығы теория ма, модель ме, немесе әдіс пе, не ол? Осы сұрақ әлі күнге дейін зерттеушілер пікірлерінің қайшылығын тудырып отыр.

«Корпустық лингвистика дегеніміз не?» деген мақаласында Шарлотта Тейлор «бұл пән бе, әдістану ма/методология ма, парадигма ма, осылардың ешқайсысы емес пе немесе осылардың барлығы ма? деген сұраққа жауап іздейді. Бірақ анық жауаптар жоқ дейді. (Charlotte Taylor «What is corpus linguistics?" – is it a discipline, a methodology, a paradigm or none or all of these? – but does not attempt to offer any definitive answers). (2008, 179 What is corpus linguistics? What the data says ICAME Journal No. 32 2008 179-200)

Корпустық лингвистиканың негізін салушылардың бірі Аартс өзінің Мэйдспен бірігіп жазған 1984 жылғы кітабында (Aarts and Meijs (1984), оның алдында 1982 жылы ван ден Хеувельмен (Aarts and van den Heuvel (1982) бірігіп жазған зерттеулерінде *корпустық лингвистика* терминін қолданған. Аартс бұл терминнің біраз дүдәмалдықпен туындағанын айтады, өйткені «бұл өте жақсы атау емес деп ойладық (және мен әлі де ойлаймын): бұл оның басты зерттеу құралымен және деректер көзімен аталатын таң қаларлық пән (странная дисциплина). Мүмкін, бұл термин өзінің пайдалылығынан әлдеқашан өтіп кеткен шығар/Возможно, этот термин к настоящему времени изжил себя». Corpora Jul 1998 to -: Corpora: First use of the term 'corpus linguistics' (uib.no). Сонда «корпус лингвистикасы» пәнінің атауы өзінің басты зерттеу құралының және деректер көзінің атауымен аталатын бірден-бір пән. Міне бұл сондықтан корпус лингвистикасы туралы жиі айтылатын мәселелердің бірін тудырып отыр, және баламаларға басымдық беруге мүмкіндік беріп отыр.

(On the Corpora List, Aarts is reported as commenting that the term was coined with some hesitation "because we thought (and I still think) that it was not a very good name: it is an odd discipline that is called by the name of its major research tool and data source. Perhaps the term

has outlived its usefulness by now". This raises one of the recurrent concerns over talking about corpus linguistics, and may account for the preference for alternatives.).

Aarts, Jan and Willem Meijs (eds.). 1984. Corpus linguistics: Recent developments in the use of computer corpora in English language research. Amsterdam: Rodopi.

Aarts, Jan and Theo van den Heuvel. 1982. Grammars and intuitions in corpus linguistics. In S. Johansson (ed.). Computer corpora in English language research, 66–84. Bergen: Norwegian Computing Centre for Humanities.

Aarts, Jan. 2002. Does corpus linguistics exist? Some old and new issues. In L. E. Breivik and A. Hasselgren (eds.). From the COLT's mouth… and others': Language corpora studies in honour of Anna-Brita Stenström, 1–19. Amsterdam: Rodopi.

Лийч 1992 жылы бұл саланы «компьютерлік корпус лингвистикасы» деп атап «компьютерлік корпус лингвистикасы тілді үйренудің жаңадан пайда болған әдістануын ғана емес, сонымен қатар іс жүзінде бұл пәнге деген жаңа философиялық көзқарасты анықтайды» деп тұжырымдаған. Компьютерлік корпус лингвистикасының сипаттамаларын жаңа парадигма ретінде сипаттаған (Leech 1992: 106). Яғни Лийч компьютерлік корпус лингвистикасын жаңа философиялық көзқарас деп анықтаған. (Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In J. Svartvik (ed.). 105–122)

McEnery және Wilson корпус лингвистикасын тіл білімінің семантика, грамматика, фонетика немесе әлеуметтік лингвистика сияқты дәстүрлі саласы ретінде емес, «әдіснама» ретінде сипаттаған (1996: 1, McEnery, T. and Wilson, A. (1996), *Corpus Linguistics*. Edinburgh: Edinburgh University Press. McEnery, Tony, and Andrew Wilson. 2001. Corpus linguistics, 2nd edn. Edinburgh: Edinburgh University Press;).

Стефан Грис (Stefan Th. Gries) корпус лингвисі, бірнеше кітаптардың авторы «Корпус лингвистикасы дегеніміз не?» (2009) мақаласында сұрақтарға жауап беру арқылы корпус лингвистикасының мәнісін жақсы түсіндірген. Зерттеуші корпус линвгисі ретінде бұл сала не туралы деген сұраққа «бұл шын мәнісінде тәсілдеме» деген. «Тәсілдеме дедіңіз дегенмен мен оны ешқашан түсінген емеспін. Корпус лингвистикасы теория ма немесе модель ме немесе әдіс пе, не ол?» деген сұраққа Грис адамдардың пікірі әртүрлі, бірі деп мысалы Лийчтің пікірін айтқан: *компьютер корпус лингвистикасы* деп атап оны жаңа философиялық тәсілдеме» (Leech (1992:106 Leech, Geoffrey N. 1992. Corpora and theories of linguistic performance. Directions in corpus linguistics. Proceedings of Nobel Symposium 82, ed. by Jan Svartvik, 105–22. Berlin, New York: Mouton de Gruyter). Көптеген басқа зерттеушілер, Гристің өзі де, оны әдіс(нама) деп қарастырады (cf. McEnery et al. 2006:7f McEnery, Anthony, Richard Xiao, and Yukio Tono. 2006. Corpus-based language studies: an advanced resource book. London, New York: Routledge.).

Егер де КЛ әдіс(нама) десек, онда **қандай теория** негізінде сіз зерттеу нәтижелерін түсіндіре аласыз? деген сұраққа зерттеуші бұның жауабы қарапайым емес, бірі - көптеген корпус лингвистер лингвистикалық теория басты маңыздылыққа ие емес деуі мүмкін. Лийчтің ойынша, корпус-лингвистикалық жұмыстардың басым бөлігі дискриптивті немесе қолданбалы сипатқа ие және іс жүзінде көп лингвистикалық теорияны қамтымайды. Екіншісі – кез келген теориялық көзқарасты қолдаушы/ұстанушы лингвистер корпустық деректерді қолдана алады. Егер лингвист неміс екінші тілді ағылшын тілін үйренушілер күрделі сөйлемдердің қалыптасуын қалай меңгеретінін зерттесе, онда ол оны екінші тілді меңгеру теориясының шеңберінде түсіндіреді және т.б. Лийчтің ойынша, лингвистикалық теорияның белгілі бір түрі корпуслингвистикалық әдістерімен үйлесімді.

Стаббс 1993, корпус лингвистикасының әдістану ретіндегі шектеулі анықтамасын жоққа шығарады және Синклердің 1991 ж. зерттеуіне пікір білдіре отырып, «пәннің осы көзқарасы бойынша корпус лингвистикалық талдау құралы ғана емес, лингвистикалық теориядағы маңызды ұғым» деп атап өткен (1993: 23–24). (Stubbs, Michael. 1993. British traditions in text analysis: From Firth to Sinclair. In M. Baker, F. Francis and E. Tognini-Bonelli (eds.). Text and technology: In honour of John Sinclair, 1–36. Amsterdam: John Benjamins.)

Вольфганг Теуберт/Teubert (2005) те теориялық концептуализацияға баса назар аударады және корпус лингвистикасын «тілді зерттеудің теориялық тәсілдемесі» деп сипаттайды (2005: 2). (Teubert, Wolfgang. 2005. My version of corpus linguistics. International Journal of Corpus Linguistics 10(1): 1–13.)

2001 жылы Тогнини-Бонелли корпус лингвистикасын «теориялық мәртебеге» ие «қолдану алдындағы әдістану» ретінде сипаттады (2001: 1). (Tognini-Bonelli, Elena. 2001. Corpus linguistics at work. Amsterdam: John Benjamins). Дәл сол сияқты, Микаэла Мальберг корпус лингвистикасын «ағылшын тілін өзіндік теориялық фреймокпен сипаттайтын тәсілдеме» деп сипаттайды (2005: 2) (Mahlberg, Michaela. 2005. English general nouns: A corpus theoretical approach. Amsterdam: John Benjamins) және мұны баса атап көрсету үшін «корпус теориялық тәсілдемесі» "corpus theoretical approach" (2005, 2006) терминін қолданады (Mahlberg, Michaela. 2006. Lexical cohesion: Corpus linguistic theory and its application in English language teaching. International Journal of Corpus Linguistics 11(3): 363–383.).

Британдық ғалымдар Тони МакЭнери мен Эндрю Харди (Tony McEnery мен Andrew Hardie What is corpus linguistics?) «Корпус лингвистикасы дегеніміз не?» мақаласында «бұл, әрине, сіз тіл білімінде зерттеуге болатын көптеген тақырыптардан айтарлықтай ерекшеленеді, өйткені бұл тілдің қандай да бір аспектісін зерттеуге тікелей байланысты емес, керісінше, бұл тілді зерттеуге арналған процедуралардың немесе әдістердің жиынтығына бағытталған сала (бірақ, біз көретініміздей, корпус лингвистерінің ең аз дегенде бір негізгі мектебі корпус лингвистикасын әдістану деген сипаттаумен келіспейді)» деген. Осы процедураларды ескере отырып, біз лингвистиканың көптеген салаларына корпусқа негізделген тәсілдемені (a corpus-based approach) қолдана аламыз. Дәл осы себептен, осы кітапта көрсетілгендей, корпус лингвистикасы біздің тілді зерттеуге деген тәсілдемеміздің бағытын өзгерте алады. Ол тілдің бірқатар теорияларын нақтылап, қайта анықтауы мүмкін. Бұл сондай-ақ бізге қолайлы өлшемдегі корпустар мен оларды пайдалану үшін жеткілікті қуатты машиналар дамығанға дейін зерттеу қиын болған тіл теорияларын қолдануға мүмкіндік береді. (2011 https://www.cambridge.org/core/books/corpus-linguistics/what-is-corpus-linguistics/C16393FB65F2BA9D7C7EFF9284F99EAE)

Енді корпус лингвистикасы дегеніміз не дегенді аздап түсініп алсақ, көзбен көрейік корпустарды. Және корпустарға негізделген талдау зерттеушілерге қандай мүмкіндіктер береді танысытырп көрейік. Тілдік элементтердің қолданылу жиілігі, яғни морфемалардың, сөздердің, грамматикалық үлгілердің т.б., корпустың (бөліктерінде) қаншалықты жиі кездесетінін және т.б. анықтай аласың; бұл ақпарат әдетте жиілік тізімдерінде көрсетіледі; осы элементтердің қатар кездесу жиілігі, яғни морфемалардың белгілі бір сөздермен қаншалықты жиі кездесетіні, белгілі бір грамматикада нақты сөздердің қаншалықты жиі кездесетіні зерттеледі; бұл ақпарат негізінен, мысалы, ізделетін сөздің барлық кездесулері болатын конкорданс деп аталатындарда көрсетіледі. (frequency list and concordance)

ҚА Тіл білімі институтында 1969 жылы «Түркі тілдерін статистикалық және ақпараттық зерттеу» атты Бүкілодақтық ғылыми жиыны өткен. 1973 жылы «Қазақ тексінің статистикасы» атты ғылыми жинақ жарық көрген [2]. Сонымен қатар 1973-1974 жж. Одақтық «Тіл статистикасы» тобының ғылыми жетекшісі Раймонд Генрихович Пиотровский мен Қазақстандық профессор ғалым Қалдыбай Бектайұлы Бектаевтың «Математические методы в языкознании» атты екі бөлімнен тұратын, жоғары оқу орындарына арналған оқу құралы қазақ елінде жарық көрген [3]. Міне, сол кездердегі (1978 жылы) «Тіл статистикасы және автоматтандыру» ғылыми-зерттеу тобына жетекшілік еткен Қ.Б. Бектаевтың «Статистико-информационная типология тюркского текста» атты докторлық монографиясы қазақ тіл білімінің статистикалық зерттеу саласына қосқан сүбелі еңбегі болды [4]. XX ғасырдың 90 жылдары 20 томнан тұратын М.Әуезовтің шығармалар жинағы ЭЕМ жадына енгізіліп, соның нәтижесінде 1995 жылы «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» атты жиілік сөздіктер жүйесі жарық көрген [5].

2 Қазақ тексінің статистикасы. Статистика казахского текста. – Алматы: Қаз ССР-ның «Ғылым» баспасы, 1973. – 731 б.

3 Бектаев К.Б., Пиотровский Р.Г. Математические методы в языкознании. Ч.1. Теория вероятностей и моделирование нормы языка. – Алматы, 1973. – 281 с.; Ч.2. Математическая статистика и моделирование текста. – Алматы, 1974. – 334 с.
4 Бектаев К.Б. Статистико-информационная типология тюркского текста. – Алма-Ата, «Наука» КазССР, 1978. – 167 с.
5 Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С., Белботаев А.Б. М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. – Алматы-Түркістан, 1995. – 346 б.

2. Корпус мөлшері қандай болу керек деген сұрақ. Корпустың қаншалықты үлкен болуы керек екеніне қатысты қатаң ережелер жоқ, оның орнына өлшем бірқатар критерийлер арқылы белгіленеді. Бұл критерийлердің бірі корпус зерттеу үшін қолданылатын тіл аспектілеріне қатысты. Кеннеди (1998: 68) просодияны яғни сөйлеу ырғағын, екпін және интонацияны зерттеу үшін» «100 000 сөзден тұратын корпус әдетте көптеген сипаттау мақсаттары үшін жалпылау жасау үшін жеткілікті болады» деп болжайды. Дегенмен, Кеннеди етістік-форма морфологиясын талдау, яғни, етістіктің шақтарын білдіру үшін -ed, -ing және -s сияқты жалғауларды пайдалануды талдау жарты миллион сөзді қажет ететінін айтады. Лексикография үшін, яғни сөздерді талдау және олардың қолданылуы, көбінесе сөздік құру үшін миллион сөздің жеткілікті болуы екіталай, өйткені сөздердің жартысына жуығы бір рет кездеседі және олардың көпшілігі көп мағыналы болуы мүмкін, яғни бірнеше түрлі мағынаға ие. Сонымен, жазбаша және ауызша тіл жанрларының өте кең ауқымын қамтитын және британдық ағылшын тілі үшін стандартты анықтама ретінде әрекет етуге арналған Британдық Ұлттық корпус көлемі 100 миллион сөзді құрайды (сайтын ашу).

(Paul Baker Corpus Methods in Lnguistics тарауы бойынша) Корпус лингвистикасында негізгі теориялық концептілер іріктеу (Sampling), тепе-теңдік және репрезентативтілік болып табылады. Корпус белгілі бір тілдің, тілдің әртүрлілігінің немесе белгілі бір тақырыптың көрінісі болуы керек болғандықтан, кейбір мәтіндер корпусты тұтастай бұрмаламау үшін оның ішіндегі мәтіндер мұқият таңдалып, теңестірілуі керек. Сондықтан корпуста толық мәтіндер болмай, мәтіндердің бөліктері болуы мүмкін. Мысалы, егер біз Виктория көркем әдебиетінің корпусын құрғымыз келсе, біз сол кезеңнің 30 авторын таңдап, корпусқа қосу үшін олардың әрқайсысының 3 романынан аламыз. Дегенмен, кейбір авторлар басқаларға қарағанда ұзағырақ роман жазуы мүмкін, нәтижесінде олардың жазу стилі корпуста шамадан тыс көрсетіледі. Нәтижесінде біз әр романнан бірдей өлшемді үлгілерді ғана алуды шеше аламыз (айталық, 30 000 сөз). Дегенмен, біз бұл үлгілерді романдардың әртүрлі орындарынан алу арқылы теңестіруіміз керек - егер біз әрбір романнан тек алғашқы 30 000 сөзді алсақ, бізде романдардың басталуының корпусы болар еді. Сондықтан біз мәтіннің әртүрлі романдардың басынан, ортасынан және аяғынан бірдей таңдалғанын қамтамасыз етуіміз керек. Міне бұл іріктеу. Басқа жағдайларда, іріктеуді соншалықты мұқият қарастырудың қажеті жоқ - егер біз тек бір автордан мәтін жинайтын болсақ немесе тұтас мәтіндерді қарастырғымыз келсе немесе мәтіндер өте қысқа болса, онда бұл мәтін бөліктерін емес, тұтас мәтіндерді қамтиды.

Аннотация.

The annotation part of a file refers to elements added to provide specifically linguistic information (e.g., part of speech, semantic information, and pragmatic information). Most commonly, annotation takes the form of part-of-speech tagging of words. The first sentence of the Brown Corpus is shown in a parts-ofspeech annotated form in (1a). The tags used in this sentence are explained in (1b) – full details can be found in the Brown Corpus Manual (khnt.aksis.uib.no/ icame/manuals/brown/INDEX.HTM). Other

Корпустар көбінесе қосымша ақпаратпен түсіндіріледі (немесе тегтеледі), олар бойынша күрделі есептеулерді орындауға мүмкіндік береді. Мұндай ақпарат бірнеше пішінде болуы мүмкін, мысалы, корпус ішіндегі жеке мәтіндер көбінесе жеке файлдар ретінде сақталады және олардың әрқайсысында мәтін туралы оның авторы, жарияланған күні, жанры және т.б. сияқты ақпаратты беретін «тақырып» болуы мүмкін. Бұл ақпарат

зерттеушілерге мәтіндердің белгілі бір түрлеріне (мысалы, газет мақалалары) назар аударуға немесе әртүрлі типтерді салыстыруға (мысалы, ер және әйел авторлар) мүмкіндік беру үшін пайдалы болуы мүмкін. Мұндай аннотацияда кейде стандартты жалпыланған белгілеу тілі (SGML) қолданылады, соның арқасында тегтер < > сәйкес бұрыштық жақшалардың ішіндегі *кодтар* (*элементтер* ретінде белгілі) пішінін алады. Сондай-ақ белгілі бір таңбалар, екпіні бар әріптер сияқты, нысандар ретінде белгілі кодтармен (known as entities) ұсынылған. Бұлар әрқашан амперсанд & таңбасынан басталып, нүктелі үтірмен аяқталады. Мысалы, екпінді é әрпі SGML нысаны é ретінде ұсынылуы мүмкін. Тегтеу әртүрлі. Мысалы Қазақ тілінің ұлттық корпусындағы тегтеуді қараңыз.