Дәріс №5
Дәріс тақырыбы: Корпустардың негізгі сипаттамалары
Сұрақтар:

1. Корпустардың репрезентативтілігі
2. Корпустардың классификациясы

1. Репрезентативтілік ұғымы корпус жасауда дизайн ұғымымен байланысты болады. дизайн ауқымды Р соның біреуі. Дизайн – корпусқа қандай мәтіндер кіреді, мәтіндердің саны, белгілі бір мәтіндерді іріктеу, мәтіндердің өз ішінен мәтін үлгілерін таңдау және мәтін үлгілерінің көлемін таңдау. Осылардың әрқайсысы Д.Бибердің айтуынша, саналы немесе санасыз түрде іріктеу шешімін қамтиды.

Д.Бибер 1993 жылғы мақаласында корпус репрезентативті болуы керек деген. «жалпы тілге қатысты жалпылаулардың негізі ретінде дұрыс пайдаланылуы үшін» корпус репрезентативті болуы керек («для того, чтобы надлежащим образом использоваться в качестве основы для обобщений, касающихся языка в целом» // `in order to be appropriately used as the basis for generalizations concerning a language as a whole); мысалы, корпусқа негізделген сөздіктер, грамматикалар және сөз таптары тэгтері репрезентативтік негізді талап ететін қосымшалар (аппликейшны) (Бибер, 1993в)

Негізінен зерттеушілер репрезентативлікке қол жеткізудің ең маңызды қарастырылымы деп іріктеу көлеміне баса көңіл бөледі: корпусқа қанша мәтіндер қамтылуы және әр іріктеу мәтінінде қанша сөз қамтылуы керек. Дегенмен репрезентативлікке қол жеткізудің ең маңызды қарастырылымы іріктеу көлемі емес дейтіндер де бар. іріктеуді зерттеушілер мақсатты поплуцияны нақты анықтау және іріктеудің әдісіне қатысты шешімдер корпусты құруда маңызды қарастырылымдар болатындығын айтады. Бұған Д.Бибер де келіседі. Д.Бибердің түсіндіруінде, «репрезентативтік іріктеудің популяциядағы әртүрліліктің толық ауқымын қаншалықты қамтитынын білдіреді. - `Representativeness refers to the extent to which a sample includes the full range of variability in a population` (
Literary and Linguistic Computing, Vol. 8, No. 4, 1993

Әртүрлілік деген не? Оны Бибер жағдаяттық және лингвистикалық перспективалардан қарастырыла алатынын, және екеуі де репрезентативтікті анықтауда маңызды екенін айтқан. Мәтіндердің кез-келген таңдауы - іріктеу. Іріктеу репрезентативті ме, бұл ең алдымен ол мақсатты популяциядағы түрлі мәтін түрлерінен қаншалықты таңдалып алғандығына байланысты; сондықтан оның репрезентативтігін бағалау мақсатты популяцияның (себебі іріктеу популяцияның репрезентативті/көрінісі) толық анықтамасына және сол іріктеуді популяциядан таңдап алудың техникаларына/әдістеріне тәуелді. (мәтін-іріктеу-популяция). Мақсатты популяцияның анықтамасының екі аспектсі бар: 1) популяцияның шектері - популяцияға қандай мәтіндер кіреді жәнне қандай мәтіндер алынып тасаталады, 2) популяцияның өз ішінде иерархиялық ұйымдастырылуы – популяцияға қандай мәтін категориялар ы кіреді және олардың анықтамалары қандай. Д.Бибер корпустарды құруда осы айтылғандарға жете мән берілмейтінін және іріктеу мақсатты популяция алдын-ала анықталып алмай жиналатынын айтады. Сондықтан ондай корпустарға репрезентативтігі немесе толықтығы туралы баға беру мүмкін емес дейді.

Популяциясы анықталған корпустарға екі корпусты көрсеткен: Броун (Фрэнсис пен Кучера 1964/79) және ЛОБ (Джоносон т.б. 1978) корпустары. Бұл корпустардың мақсатты популяциялары екеуіне де қатысты анықталған (1) шекараларына – барлығы 1961 жылы жарық көрген Ағылшын мәтіндері АҚШ-та (Браун) және Ұлыбританияда (ЛОБ) және (2) иерархиялық ұйымдастырылуы – басты 15 мәтін категориясы және осы категриялардың өз ішінде көптеген субжанрлар. Бұл корпустарды құруда корпус құрушылары сонымен қатар жақсы іріктеу фреймдерін қолданған. Іріктеу фреймдері дегенміз популяцияның операционалды анықтамасы, популяция мүшелерінің тізімі (itemised list), содан

репрезенативтік іріктеу таңдап алынады. ЛОБ корпусының іріктеу фреймі: кітаптар үшін 1960-1964 жылдардағы «The British National Bibliography Cumulated Subject Index, 1960-1964» 1961 жылы жарық көрген барлық жарияланымдар және гаезт-журналдар үшін мақсатты популяция - Willing's Press Guide-тағы (1961) 1961 жылғы жарияланымдар. Браун корпусында іріктеу фреймі Браун Университетінің кітапханасы мен Providence Athenaeum-дегі кітаптар мен газет журналдар.

Іріктек фреймі жақсы анықталса, ықтимал іріктеуді (probabilistic sampling/ вероятностная выборка) sampling таңдауға болады. ықтимал іріктеудің бірнеше түрлері бар. Бірақ олардың бәрі кездейсоқ таңдауға негізделген. Қарапайым кездейсоқ іріктеу. Қарапайым кездейсоқ іріктеуде популяциядағы барлық мәтіндердің бірдей таңдалуына мүмкіндік бар. мысалы, Британдық Ұлттық Библиографиядағы барлық аталымдар тізбеленіп нөмірленсе, кездейсоқ нөмірлер кестесі кітаптардың кездейсоқ іріктеуін таңдауда қолданыла алады. Стратификацияланған іріктеу. Бұл ЛОБ пен Браун корпустарын құруда пайдаланылған екен. Бұл әдісте популяцияның өз ішінен ішкі топтар (бұл жағдайда жанралар) анықталады (субгруппы) сонан кейін бұл страталардың әрқайсысы кездейсоқ техникаларын пайдалана отырып іріктеледі. Бұл әдістің артықшылығы барлық страталар толық қамтылады, бір стратаға көп қамтылып кетпейтіндей. Браун мен ЛОБ-та жанр категориялары деңгейінде 100 пайыз және әр жанр өз ішінде мәтіндер дәл таңдалған).

Репрезентативтік деп әртүрлі кезеңдегі, жанрдағы, стильдегі, авторлардың және т.б. мәтіндердің корпуста қажетті-жеткілікті және пропорционалды берілуі, яғни проблемалық аймақтың барлық қасиеттерін көрсете алуы түсініледі (*Рыков В.В.* Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Международного семинара Диалог-2002. – М.: Наука, 2002). Репрезентативтілікті анықтаудың әртүрлі тәсілдемелері бар.

Literary and Linguistic Computing, Vol. 8, No. 4, 1993

Репрезентативтілік - тіл әртүрлілігін қамтуы не көрсетуі керек (вариациясы). Мұны Лич 1991 жылы еңбегінде айтып өткен. Бұл жайлы Доуглас Бибер де 1993 жылғы арнайы репрезентативтілік жайлы жазған еңбегінде де атап өтеді. «Representativeness refers to the extent to which a sample includes the full range of variability in a population.» Biber (1993: 243). Tony McEnery, Richard Xiao, Yukio Tono (2006): «A corpus is essentially a *sample* of a language or language variety (i.e. *population*).».

(the extent to which a sample includes the full range of variability in a population) дегенді білдіреді.

Biber, D. (1993) 'Representativeness in corpus design', *Literary and linguistic computing* 8: 4, 243–57. (пдф бар!)

> One of the key claims it should be possible to make of a corpus is that it is a representative sample of a particular language variety. There are many safeguards that may be applied in sampling to ensure maximum representativeness in corpus design. Biber (1993) emphasises that the limits of the population that is being studied must be defined as clearly as possible before sampling procedures are designed. One way to do this is to use a comprehensive bibliographical index – this was the approach taken by the Lancaster–Oslo/Bergen (LOB) Corpus builders who used the British National Bibliography and Willing's Press Guide as their indices.

Тони Мак Енери мен Эндри Уэлсон бұл аталған тәсілдемелерді өте жақсы және ол жарық көрген кітаптармен және газеттермен жақсы дейді. Алайда формалды емес тілде, айталық күнделікті әңгімелесулер мен жеке хат алмасуларда, мүмкін еместігін айтқан. Себебі тілдің мұндай түрлері әдетте индекстелмейді не кітапханада сақталмайды. Мұндай жағдайларда индекс іріктеу шеңберін қолданғанның орнына демографиялық іріктеу типін қолдану дұрыс. *Демографиялық іріктеу* дегеніміз информанттарды жасы, жынысы, әлеуметтік табы, аймағы т.т. бойынша іріктеу. Бұл Британдық ұлттық корпусты (БҰК) жасауда ауызша сөйлеу бөлігін жинауда пайд.н тәсілдеме: информанттар демографиялық іріктеу бойынша таңдалған, оларға кассета типті жазу құралдары беріліп, күнделікті сөйлеулерін екі күннен жеті күнге дейін жазған. Дегенмен Крауди (Crowdy 1993)

демографиялық іріктеудің өзін ғана алу көптеген маңызды лингвистикалық типтерді қамтымауы мүмкін, сондықтан демографиялық іріктеуді басқа да тәсілдемелермен толықтыру қажет деген. Бұл тағы да БҰК жасауда пайдаланылған, маңызды ауызша әрекеттер, мысалы медиадағы сұхбаттар және заң жинақтары, информанттар жазып алатын күнделікті сөйлеулерге енбеуі мүмкін, сондықтан сөйлеудің мұндай типтері анықталып алынып, олар да іріктеуге қосылған.

Популяцияны анықтауда Бибер сонымен қатар популяцияның иерархиялық құрылымын - стратаны – анықтауға көңіл бөледі. Страта дегеніміз қандай әртүрлі жанрлардан, қандай каналдардан т.б. алынады. мысалы, газет мақалалары ма, романтикалық фильмдер ме, ғылыми мәтіндер ме т.т. солар анықталады.

«Корпус» термині әдетте шектелген белгіленген өлшемдегі мәтіндер жинағын білдіреді. Уақыт өте келе корпустың көлемі мен құрамы өзгеруі мүмкін, бірақ бұл өзгерістер оның құрылымын өзгертпеуі керек, немесе оны негізді түрде өзгертуі керек. Корпустың берілуі, әртүрлі белгілері бойынша оның жеке бөліктерінің арақатынасы *репрезентативтілік* немесе *тепе-теңдік* деп аталады. Бірінші корпустың көлемі, алдыңғы дәрісте айтылғандай, 1 миллион сөз қолданысын құрады (Браун корпус, Ланкастер-Осло-Берген корпусы, орыс тілінің Упсала корпусы). Мұндай көлем тілді барлық алуан түрлілігімен көрсетуге мүмкіндік бермеді. Қазіргі уақытта жалпы тілдік (ұлттық) корпусқа кемінде 100 миллион сөз қолданысы кіруі керек деп есептеледі. Ұлттық корпус белгілі бір тілді өзінің өмір сүруінің белгілі бір кезеңінде немесе кезеңдерінде барлық жанрларда, стильдерде, аумақтық және әлеуметтік варианттар және т.б. жағынан көрсетеді. (мысалы, NCRP, http://ruscorpora.ru сайтында қолжетімді, BNC шектеулі түрде http://www.natcorp.ox.ac.uk/ сайтында немесе http://sara.natcorp .ox. ac.uk). Қазіргі барлық лингвистикалық зерттеулер мен сөздіктер мен грамматикаларды құрастыру жұмыстары қандай да бір түрде көрнекті (репрезентативті) мәтіндік корпусты пайдалануға бағытталған деп айтуға болады.

Корпус лингвистикасы нысанның кем дегенде екі типіне (мәтіндердің корпустарына) сүйенетінін тәжірибе көрсетіп отыр. Олар:

1. Бірінші типтегі корпустар әмбебап, олар сөйлеу әрекетінің барлық көптүрлілігін көрсетеді.
2. Екінші типтегі корпустар сөйлеу тәжірибесінде кейбір лингвистикалық немесе мәдени құбылыстардың болуын көрсетеді, олар арнайы мақсатта ad hoc құрылады (они построены ad hoc (для специальной цели), мысалы, мақал-мәтелдер корпусы немесе газет сөзіндегі саяси метафоралар корпусы [31]. (***Ad hoc*** is a Latin phrase meaning literally 'to this'.)

Екі жағдайда да репрезентативтілік проблемалық аймақтың барлық қасиеттерінің мәтіндер корпусында көрініс табуын статистикалық бағалау ретінде ғана қарастырылады.

2. Корпустардың әртүрлілігіне қарамастан, оларды кластарға бөлудің екі негізгі жолы бар:
1) тұтас тілге (көбінесе белгілі бір кезеңнің тіліне) қатысты корпустарды қандай да бір қосалқы тілге (подъязыку) (жанр, стиль, белгілі бір жастағы немесе әлеуметтік топ тілі, жазушы немесе ғалым тілі және т.б.) қатысты корпустарға қарсы қою;
2) корпустарды лингвистикалық белгілеу типіне қарай бөлу. Белгілеудің көптеген типтерінің болуына қарамастан, іс жүзінде бар корпустардың көпшілігі морфологиялық немесе синтаксистік типтегі корпустарға жатады (соңғылары ағылшын әдебиетінде treebanks деп аталады, оларды «синтаксистік құрылымдардың банктері» деп аударуға болады).

> 1) противопоставление корпусов, относящихся ко всему языку (часто к языку определенного периода), корпусам, относящимся к какому- либо подъязыку (жанр, стиль, язык определенной возрастной или социальной группы, язык писателя или ученого и т.д.);
> 2) разделение корпусов по типу лингвистической разметки. Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа (последние в англоязычной литературе называют treebanks, что можно перевести как «банки синтаксических структур»).

А.Захаров, Богданова топтастыруы (кестеде). Сонымен, **тілдік деректер түріне қарай** корпус *жазбаша, ауызша* және *аралас* болып бөлінеді. Жазбаша корпустарда ауызша сөйлеу берілмейді (Браун корпусы, LOB), ауызша корпустарда ауызша сөйлеу ғана беріледі, аралас корпустар әдетте ұлттық корпустар болады, олар белгілі бір уақыт кезеңінде тілдің көрінісін береді (НКРЯ, BNC және т.б.).

«**Параллелдік**» критерийі бойынша корпустар *біртілді, екітілді* және *көптілді* болып бөлінеді. *Біртілдік корпуста* тілдің диалектілері, вариванттары қарсы тұрады. *Екітілді* және *көптілді* корпустар екі немесе одан да көп тілде жазылғанына қарамастан бір тақырыптық саладағы мәтіндерді біріктіреді (мысалы, әртүрлі елдерде және әртүрлі тілдерде өткізілген белгілі бір ғылыми мәселе бойынша конференциялар материалдарының корпусы). Мұндай корпус терминологиямен жұмыста көмектеседі және аудармашылар жиі пайдаланады. Екітілді немесе көптілді корпустың тағы бір нұсқасы – бір бастапқы тілде жазылған мәтіндер-түпнұсқаларының жиынтығы және осы бастапқы мәтіндердің бір немесе бірнеше басқа тілдерге мәтіндер-аудармалары. Мұндай корпус салыстырмалы-салғастырмалы зерттеулер жүргізуге, аударма теориясы бойынша зерттеулерге және адамды және компьютерді аударманы үйрету үшін баға жетпес материал береді.

«**Әдебиеттілік**» критерийі бойынша *әдеби, диалектілік, ауызекі тіл, терминологиялық* және *аралас* корпустар ажыратылады. Ауызекі тіл корпусының мысалы ретінде Санкт-Петербургте әзірленген Один Речевой День (ОРД) [38], терминологиялық корпустың үлгісі - тікелей «тірі» мәтіндік материал бойынша терминологиялық сөздікті жасауға мүмкіндік беретін корпус лингвистикасы бойынша мәтіндер корпусы [54].

Құрылу мақсаты бойынша корпустар *көпмақсатты* және *арнайы* болып бөлінеді. *Көпмақсатты* корпустар әдетте әртүрлі жанрдағы мәтіндерді қамтиды (бұған ұлттық корпустар жатады), ал *арнайы* корпустар бір жанр немесе жанрлар тобымен шектелуі мүмкін.

Мәтіндер корпусын **жанр бойынша** *әдеби, фольклорлық, драмалық, публицистикалық* және т.б. бөлуге болады. *Публицистикалық* корпустың мысалдары *20 ғасырдың соңындағы орыс газеттерінің мәтіндерінің компьютерлік корпусы* мысал бола алады (http://www.philol. msu.ru/~lex /corpus/) және саяси метафоралар корпусы [2].

Қолжетімділік критерийі бойынша корпустар *еркін қолжетімді, коммерциялық, жабық* корпустар болып бөлінеді. *Еркін қолжетімді* корпус кез келген уақытта корпустың барлық мәтіндеріне толық онлайн режимінде қол жеткізуге мүмкіндік береді. Кейбір жағдайларда еркін қол жеткізу корпус деректерінің бір бөлігіне қамтамасыз етілуі мүмкін. *Коммерциялық* корпустармен жұмыс істегенде оны on-line пайдалану немесе компакт-дискідегі көшірме құқығын сатып алу қажет. Алдымен корпустың аннотациясымен танысуға болады немесе, мүмкін, тіпті сынақ режимінде (в пробном режиме), бірақ, әдетте, барлық мәтіндермен емес, тек шағын субкорпуспен корпуспен жұмыс істей аласыз. *Жабық* корпустар тар белгілі бір мақсаттарға арналған, жалпы көпшілікке пайдалану үшін арналмаған.

Тағайындау бойынша *зерттеу* және *иллюстрациялық* корпустар ажыратылады. *Зерттеу корпустары* тілдің қолданысының әртүрлі аспектілерін зерттеу үшін құрылады. корпустардың бұл типі лингвистикалық міндеттердің кең класына бағытталған. *Иллюстрациялық корпустар* ғылыми зерттеулерден кейін жасалады: олардың мақсаты жаңа фактілерді ашу емес, бұрыннан алынған нәтижелерді растау және негіздеу. Олар бұрын басқа лингвистикалық әдістермен ашылған белгілі бір тілдік (сөйлеу, мәтіндік) фактілерді растайтын лингвистикалық мысалдарды бөліп көрсетуге қызмет етеді. Иллюстративті корпустың типтік мысалы - «Путеводитель по дискурсивным словам русского языка» [3], мұнда бөлшектердің (частиц) семантикалық талдауы және ерекшеленген мағыналар маңызды мәтіндік материалмен қамтаамсыз етіледі, ол оқырманға семантикалық интерпретацияларды тексеруге мүмкіндік береді [17; 2].

«**Динамикалық**» критерийі корпустарды *динамикалық* және *статикалық* деп бөледі. Алғашқыда корпустар тіл жүйесінің белгілі бір уақыттық жағдайын көрсететін статикалық құрылымдар ретінде жасалды. *Статикалық* корпустар белгілі бір шағын уақыт кезеңіндегі мәтіндерді қамтиды [17]. Бұл корпус түрінің типтік түрлері авторлық корпустар – жазушы мәтіндерінің жиынтығы болып табылады. (әрі қарай өздеріңіз қаарйсыздар)

Признак	Типы корпусов
Тип языковых данных	Письменные Устные Смешанные
«Параллельность»	Одноязычные Двуязычные Многоязычные
«Литературность»	Литературные Диалектные Разговорные Терминологические Смешанные
Цель	Многоцелевые Специализированные
Жанр	Литературные Фольклорные Драматургические Публицистические
Доступность	Свободно доступные Коммерческие Закрытые
Назначение	Исследовательские Иллюстративные
Динамичность	Динамические (мониторные) Статические
Разметка	Размеченные Неразмеченные
Характер разметки	Морфологические Синтаксические Семантические Просодические и т.д.
Объем текстов	Полнотекстовые «Фрагментнотекстовые»

Corpus-based Language Studies: An Advanced Resource Book
Tony McEnery, Richard Xiao, Yukio Tono
Published Routledge 2006
https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm