2-Модуль. Корпустарды құру

Дәріс №7
Дәріс тақырыбы: Корпус құрудың алдын-ала жұмыстары
Сұрақтар:

1. Корпус құруды жоспарлау және технологиялық процесі
2. Дереккөздерді таңдау және таңдау критерийлері

1. Корпус құру ісінде оны жоспарлау кезеңі өте маңызды. Корпустың дұрыс құрылуы оның дұрыс жоспарлануымен байланысты. Репрезентативтік, баланс немесе корпус көлемі қандай болады деген сұрақтардың бәрі корпусты жоспарлау сатысының маңызды кезеңдері. Біз өткен дәрістен білеміз, репрезентативтік мәтіндердің жеткілікті көлемі ғана емес оның әртүрлілігі де. Әртүрлілік деген не, ол мәтіндердің жанрлық-тақырыптық алуан түрлілігі. Корпустың оңтайлы дизайны оның жұмсалу мақсатына байланысты. Корпусты құрастырушының қандай анализдер жасалатыны туралы нақты идеясы болуы керек. Егер мақсат лингвистикалық немесе әлеуметтік лингвистикалық талдаулар жасау болса, онда корпустың типі, мазмұны, көлемі және құрылымын жақсы ойластырып алу керек. Мысалы корпус құрылымын алайық, егер корпусқа ауызша және жазбаша мәтіндер де кіреді деп жоспарланса, онда олардың пропорциясын анықтап алу керек. Пропорция нәтижесі корпус мәтіндерінің балансына әкеледі. Корпустағы әртүрлі мәтін категорияларынан мәтіндер пропорцияларын құрылымдауда пилоттық сауалнама жүргізу тиімді екен. Оған мысалы Кеннеди («Корпус лингвистикасына кіріспе») Маори сөйлеу тілінің корпусын дизайндағанда радио және телевидение хабарларын құрылымдаған, сонда Бойс (1996) әр аптада телевидение мен радио жиіліктердегі әртүрлі мазмұндағы хабарлардың уақытына алдын ала анализ жасаған екен. Жаңалықтар мен сұхбаттар, спорт тақырыбындағы, музыка т.т. әрқайсысының пайызын анықтаған. Мәселен, музыка хабарларында музыка дәстүрлі болса да, олар корпусқа іріктелмеген, себебі ол қазіргі Маори сөйлеу тілін көрсете алмайтын болған.

А.Захаров пен С.Богданова корпус құруды жоспарлауда бірнеше сұрақтарды қойып, оларды жоспарлау кезеңінде жақсылап ойлап шешіп алуды ұсынады. «Жанрлық-тақырыптық құрылымнан басқа, басқа да көптеген жеке, бірақ шешуді қажет ететін маңызды мәселелер бар», мысалы: 1. Корпустағы мәтін дегеніміз не? Мысалы, газеттердегі шағын хабарландырулар – олар корпусқа жеке мәтін ретінде кіре ме, әлде біріктіруге бола ма?» деп 4 түрлі сұрақты қояды (33-б). Олар хронология да маңызды дейді. «Қазіргі орыс тілінің корпусы дегенде не ұғуға болады?» (33-б). Осы және басқа да сұрақтар жоспарлау кезеңінде қойылуы қажеттігін ескертеді.

Г.Кеннеди корпус құрастырудың (compiling a corpus) негізгі үш кезеңін атайды: корпус дизайны, мәтін жинау немесе аулау (text collection or capture) және мәтінді кодтау немесе белгілеу (text encoding or markup). (2.6 тараушасы). Сондай-ақ сақтау жүйесін жоспарлау, рұқсат алу процедураларына да арнайы тоқталады.

Randi Reppen-нің айтуында, күн сайын Интернетте әр түрлі тілдегі көбірек корпустар қолжетімді болып келеді. Дегенмен, сіз бар корпустармен толық көрсетілмеген түрлі тіл түрлерін зерттеуге қызығушылық танытуыңыз мүмкін. Бұл жағдайда сізге корпус құру қажет болады. - Each day, more and more corpora of different languages are becoming available on the web. However, you might be interested in exploring types of language that are not adequately represented by existing corpora. In this case you will need to build a corpus.

Сондай корпустарды құрастыруда Р.Реппен басты қарастырылымдарды атап, әрқайсысына жеке тоқталады.

- нақты тұжырымдалған сұрақтың болуы – корпус құрудың қажетті бірінші кезеңі, себебі бұл корпусты дизайндауға басшылық болады. корпус сол зерттелейін деп отырған тілдің көрінісі (репрезентативі) болуы шарт. Мақсатқа сәйкес материалдар жиналуы қажет.

- Қандай деректер және қанша көлемде қолдануым керек? – корпус көлемі, өткен дәрісте айтылған репрезентативтік. Корпустың көлемі зерттеушінің мақсатына және нақты тұжырымдалған сұрағына байланысты. Көлемі үлкен корпустарды, көлемі шағын корпустарды да құрастыруға болады. егер белгілі бір ақынның не жазушының шығармаларына қатысты корпус болса, ол шағын көлемді болады. дегенмен лексикалық зерттеулер, мысалы сөздік құрастыру сияқты, немесе күрделі грамматикалық қатынастарды не құбылыстарды зерттеу үшін көлемді корпустар қажет.
- Мәтіндерді қалай жинаймын? – бұл тұста рұсқат алу рәсімі маңызды. Мәтіндерді адамардан немесе қоғамдық институтардан жианғанда, ол тараптардың рұқсаттарын алу маңызды. Кейбір домендер жалпы көпшіліктің қолдануына ашық болады, яғни рұқсат алып керек емес. Ал авторлық құқықпен қорғалған мәтіндер (copyrighted texts) рұқсат алу міндетті. Кейбірінде рұқсат алу ақымен жүреді. Дәл осы сұрақта «Мәтін дегеніміз не? Файлдарды қалай атаймын? Мәтіндерді қалай сақтаймын? Сұрақтарына жауап іздейсің. Егер сыныпта жазылатын эсселерді жинасаң, мәтін белгілі бір күні сыныпта жазылған эсселердің бәрі немесе мәтін әр студенттің эссесі болуы мүмкін. Р.Реппен соңғысын дұрыс дейді. Және Р.Реппен алдымен файлдарды ең аз бірліктермен жасауды ұсынады, себебі кейін талдауда қажет болып жатса, біріктіру оңай. Үлкен бірлік сақталған файлды ашып оны қайта бөліп, әр файлды қайта атап сақтағаннан гөрі біріктіру оңай деген. Ауызша мәтіндерді жинауда мәтінді анықтау қиын деген. Қалай болса да ол зерттеушінің зерттейін деп отырған нақты сұрағына/сұрақтарына байланысты. Файлды атау да маңызды. Қарапайым сияқты оңай сияқты болғанымен, оны атау да маңызды. (өз тәжірибемізден). Мысалы, Р.Реппен: azcf108 – A letter written by a woman in a city in Arizona printed in October of 2008
- In many cases a header is included at the beginning of each corpus file. A header contains information about the file.

  Example header:

  < File name = spknnov06.mf >
  < Setting = two friends chatting at a coffee shop >
  < Speaker 1 = Male 22 years old >
  < Speaker 2 = Female 33 years old >
  < Taped = November 2006 >
  < Transcribed = Mary Jones December 2006 >
  < Notes: Occasional background traffic noise makes parts unintelligible >

- Маған қанша белгілеу керек? – белгілеу - корпус файлдарына ақпарат қосу. (Randi Reppen. 25 Mar 2010, Building and designing a corpus: What are the key considerations? The Routledge Handbook of Corpus Linguistics Routledge). POS annotation.

Корпус құрудың технологиялық процесі бар, ол белгілі бір кезеңдерден тұрады. А.Захаров пен С. Богданова айтуында мынадай кезеңдермен белгіленген:

1. Дереккөздер тізіміне сәйкес мәтіндерді алуды қамтамасыз ету.
2. Машинада оқылатын пішінге түрлендіру. Корпустарды құруға арналған электронды түрдегі мәтіндерді әртүрлі тәсілдермен алуға болады - қолмен енгізу, сканерлеу, авторлық көшірмелер, сыйға тарту және алмасу, Интернет, баспалар ұсынған түпнұсқа макеттер және т.б.
3. Мәтіндерді талдау және алдын ала өңдеу. Бұл кезеңде әртүрлі дереккөздерден алынған барлық мәтіндер филологиялық тексеруден және түзетуден өтеді. «Технологиялық» сипаттаманы дайындау мәтіннің библиографиялық және экстралингвистикалық сипаттамаларын қамтиды.
4. Түрлендіру және графематикалық талдау. Кейбір мәтіндер де машинаға дейінгі өңдеудің бір немесе бірнеше сатысынан өтеді, оның барысында қайта кодтау (қажет болған

жағдайда), сондай-ақ мәтіндік емес элементтерді (суреттер, кестелер) алып тастау немесе түрлендіру, мәтіннен екінші жолға көшірулерді (перенеосов) алып тастау, «жолдың қатты соңы» (MS -DOS мәтіндері), сызықшалардың бізді жазылуын қамтамасыз ету және т.б. Графематикалық талдау келесі операцияларды қамтиды: енгізілген мәтінді элементтерге (сөздерге, бөлгіштерге (разделители) және т.б.) бөлу, мәтіндік емес элементтерді жою, стандартты емес (лексикалық емес) элементтерді таңдау және рәсімдеу, арнайы мәтін элементтерін (аттар (аты, әкесінің аты), бас әріптермен жазылған, латын тілінде жазылған шетелдік лексемалар, суреттер атауларын, ескертулер, титул беттер, әдебиеттер тізімі және т.б.) өңдеу. Әдетте, бұл операциялар автоматты түрде орындалады. Әдетте осы кезеңде мәтінді оның құрылымдық құрамдас бөліктеріне бөлу (сегменттеу) жүзеге асырылады.

5. Мәтінді белгілеу. Мәтінді белгілеу мәтіндерге және олардың құрамдас бөліктеріне қосымша ақпаратты (метадеректер) енгізуден тұрады. Метадеректерді 3 түрге бөлуге болады: бүкіл мәтінге қатысты *экстралингвистикалық*; мәтіннің құрылымы туралы деректер; мәтіннің элементтерін сипаттайтын *лингвистикалық метадеректер*. Корпус мәтіндерінің метасипаттамасына деректердің мазмұндық элементтері де (библиографиялық деректер, мәтіннің жанрлық және стильдік ерекшеліктерін сипаттайтын белгілер, автор туралы ақпарат) және формалды (файл атауы, кодтау параметрлері, белгілеу тілінің нұсқасы, жұмыс кезеңдерінің орындаушылары) элементтері де кіреді. Бұл деректер әдетте қолмен енгізіледі. Құжаттың құрылымдық белгілеуі (абзацтарды, сөйлемдерді, сөздерді ерекшелеу) және тілдік белгілеудің өзі әдетте автоматты түрде жүзеге асырылады.

6. Автоматты белгілеудің нәтижелерін түзету: қатені түзету және екіұштылықты жою (қолмен немесе жартылай автоматты түрде).

7. Белгіленген мәтіндерді жылдам көп аспектілі іздеуді және статистикалық өңдеуді (соңғы кезең) қамтамасыз ететін мамандандырылған лингвистикалық ақпарат іздеу жүйесінің (корпус менеджері) құрылымына түрлендіру.

8. Корпусқа қолжетімділікті қамтамасыз ету. Қоршау дисплей класында қолжетімді болуы мүмкін, ықшам (компакт) дискіде таратылуы мүмкін және әлемдік желі режимінде қолжетімді болуы мүмкін. Пайдаланушылардың әртүрлі санаттарына әртүрлі құқықтар және әртүрлі мүмкіндіктер берілуі мүмкін.

9. Корпусты құру мен пайдаланудың әртүрлі аспектілерін сипаттайтын, атап айтқанда, метадеректер бойынша іздеуге мүмкіндік беретін белгілеу туралы, корпус менеджерінің сұрау тілі туралы және т.б. мәліметтер келтірілетін құжаттамалық қамтамасыз етуді жасау (34-36б.) (өздеріңіз оқисыздар)

Әрине, әрбір нақты жағдайда процедуралардың құрамы мен саны жоғарыда аталғандардан өзгеше болуы мүмкін, ал нақты технология әлдеқайда күрделі болуы мүмкін.

Корпус құруда біржақтылық **bias** мәселесін де есте ұстау керек. Біржақтылық – бұл идеяның немесе заттың пайдасына немесе оған қарсы пропорционалды емес салмақ, әдетте жабық, теріс немесе әділетсіз.

Корпусты құруда мәтіндердің таңдалуына да мұқият болу керек. Әсіресе журналистік мәтіндерде. Корпусты құрастырушы журналистік мәтіндерді жинағанда оның тақырыптарының әртүрлілігіне және нөмірлеріне назар аудару керек. Аптаның белгілі бір күндерінде немесе жылдың белгілі бір айларында белгілі бір тақырыпта көп хабарлар шығуы мүмкін. Мысалы, крисмасс туралы мамыр айына қарағанда қаңтар айында көп жазылады. Немесе қазақстандық кеңістікте мысал, наурыз жайлы. Ендеше бір тақырыптағы көп мәтіндер іріктелмес үшін корпус құрастырушы әр кезеңдегі әр тақырыптағы хабарлар мәтіндері іріктелгенін қадағалауы қажет.

Келесі кезең - *Сақтау жүйесін жоспарлау және жазбаларды сақтау* кезеңі. Мәтіндер қайда сақталады? 1 млн.нан тұратын корпус негізінен дискідегі 8-10 мегабайт орынды қажет етеді екен. Егер ол грамматикалық тұрғыдан тэгтелсе, тағы да 3-5 мегабайт қажет болады. егер парсинг жасалса, тағы 30 мегабайт керек болады.

Келесі - корпус құруда лицензияланған вебсайттардан рұқсат алу рәсімдерін ескеру керек.

Осылай корпустың дизайны, мазмұны мен құрылымы жоспарланып, шешіліп алған соң, **мәтін жинау** (text capture) кезеңі басталады. *Жазбаша мәтіндер*. Жазбаша мәтіндерді жинау қазір технологияның дамуымен мүмкіндігі арттты. Мысалы, бұрын газет журналдар баспа нұсқасында болса, қазір олар интернет арқылы элктронды нұсқада қолжетімді. Егер әрине қолмен жазылған қолжазба болса ғана, оны теріп шығуға тура келеді. Корпус құруда машинада оқылатын мәтіндер болуы шарт. Машинада оқылмайтын мәтіндер болса оларды компьютерге енгізуге көп күш жұмсалады. *Ауызша сөйлеу мәтіндерінің* басылған нұсқасы болмайтындықтан, оларды қолмен кіргізу өте қажырлы еңбек, егер қазақ тіліне қатысты айтсақ, мүмкін ағылшын тілінде автоматты оқитын бағдарламалар бар болуы мүмкін.

Systems such as Voicewalker was used for the Santa Barbara corpus and SoundScriber was used for compiling MICASE (Michigan Corpus of Academic Spoken English). Praat can be employed for phonetic analysis. Unfortunately, speech recognition software is not yet accurate enough to automatically create text from sound recordings unless they are of broadcast quality.
MICASE https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple (ашып көрсетуге болады)

Қазіргі кезде кітаптар, журналдар, газеттер компьютерленген формаларда басылып шыққан. Кейбір үлкен медиа басылымдар Британияның Гардиан немесе Индепендент газеттері материалдарын СДРомда ұсынады және оларды жартылай түрде корпусқа пайдалану керек болса ақылы түрде рұқсатын алу керек.

Жазбаша мәтіндерді жинаудың тағы бір жолы - сканерлеу. Сканерлеу 1960-80 жылдары технологияның сол кезеңінде біршама проблемалар тудырса, қазір сканерлеудің мүмкіндіктері көп. Соның өзінде әлі де қиындықтар барын зерттеушілер айтады. Мысалы газетте - дефиспен келген сөздер сканерлеп, компьютер формасына аударғанда жекелеген сөздер болып кетуі мүмкін. Сөздер I’ve, don’t тағы басқа қысқартулармен келетін сөздер сканерлеп жиналғанда кейін сөз санына кері әсерін тигізуі мүмкін екен.

Ауызша мәтіндерді жинау уақытты және қаражатты көп қажет ететін жұмыс. Себебі электроникалық құрылғылардағы жазбаларды жазбаша немесе компьютерленген формаға түсіру қосымша жұмыстарды талап етеді. Жазбашамен салыстырғанда. Ең қиыны транскрипциялау. Транскрипция жасамай ауызша мәтіндегі сөз санын анықтау мүмкін емес. Ауызша мәтін транскрипцияланып және уорд процессорда теріліп не арнайы бағдарламаларда қолмен транскрипцияланып болған соң ол электронды формаға түсті деп есептеледі. Қандай да бір линвгистикалық қателер кетпес үшін оны тексеріп те шығады, пруфридинг және өңдеу жүреді.

«Корпусты жобалау (дизайндау) мәселелері корпустың мазмұнын және деректерді ұйымдастыру үшін қолданылатын әдістерді талқылаудың алдында шешілуі керек. Көбінесе жобалау және құру принциптері жергілікті түрде анықталады (are locally determined) (Conrad 2002: 77); дегенмен, корпус жобалауына (дизайнына) қатысты айтылған Дж. Синклер принциптерін ауызша және жазбаша корпус үшін жалпы нұсқаулар ретінде қарастыруға болады» - (В книге: The Routledge Handbook of Corpus Linguistics Anne O’Keeffe, Michael McCarthy, тарау - Building a spoken corpus: What are the basics? Svenja Adolphs and Dawn Knight)

Принциптері: Джон Синклер Chapter 1: Corpus and Text — Basic Principles (John Sinclair, Tuscan Word Centre © John Sinclair 2004) -

1 The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.

2 Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.

3 Only those components of corpora which have been designed to be independently

contrastive should be contrasted.
4 Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
5 Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.
6 Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.
7 The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.
8 The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.
9 Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.
10 A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided. (Sinclair, J. (2005) 'Corpus and Text-basic Principles', in M. Wynne (ed.) Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books, pp. 1–16.)

2. Мәтіндер корпусының маңызды ерекшелігі – ол бір немесе басқа тілдегі кездейсоқ біріктірілген мәтіндердің жай ғана жиынтығы емес. Оны құруда бірқатар проблемалар бар. Олардың негізгілері мыналар (А.Захаров пен С.Богданова):
1. Мәтіндер корпусының негізгі бірлігі қандай болуы керек?
2. Мәтіндер корпусының көлемі қандай болуы керек (ол неше бірліктен тұруы керек)?
3. Мәтіндер корпусында қандай жазбаша мәтін дереккөздері және қандай көлемде көрініс табуы керек?
4. Корпусқа енгізілетін мәтіндер тілдің қай саласынан таңдалуы керек? (36-б.)

Бұл сұрақтар корпус құрудың іріктеу, іріктеу көлемі, іріктеу порциялары деген ұғымдарымен тығыз байланысты. Бұл жайында мен алдында айттым.

Мәтіндер корпусының негізгі бірлігі сөз қолданысы (әдетте оларды *сөздер* деп атайды), негіздер (түбірлер, леммалар) және сөйлемдер болуы мүмкін. аталған бірліктерде жасалатын мәтіндер корпусының көлемі құру мақсаттарына байланысты. Әріптердің, әріптер тіркестерінің, дыбыстардың, дыбыс тіркестерінің қолданылу жиілігін зерттегенде көлемі аз болуы мүмкін. Ол лексиканы, морфологиялық құбылыстарды зерттегенде және мәтіндердің синтаксистік немесе стильдік ерекшеліктерін зерттегенде әлдеқайда үлкен болуы керек.

Келесі мәселелер де проблемалық болып табылады:

1. Мәтіндер корпусына қандай функционалдық жанрлардың мәтіндері (көркем әдебиет, драма, поэзия, ғылыми мәтіндер, газеттер, журналдар, техникалық сипаттамалар және т.б.) кіруі керек?

2. Мәтіндер корпусына қандай уақыт кезеңдерінің мәтіндерін енгізу керек (қазіргі, 10 жылдық, 50 жылдық, ежелгі және т.б.)?

3. Мәтіндер тек әдеби тілдің ғана мәтіндерін қамту керек пе әлде басқа да дереккөз түрлерінде болуы керек пе? Ал әдеби тілге не жатады? (37)

Бұл сұрақтарға жауап беру кезінде мәтін корпусын әзірлеушілер әдетте лингвистика және лингвистикалық статистика мамандарының кеңестерін немесе сауалнама әдісін

пайдаланады. А.Захаров пен С.Богданова: The American Heritage Intermediate Corpus корпусын құру үшін мамандардың тәжірибесі және сауалнама әдісі қолданылған. Мамандар корпустың көлемін 5 миллион сөз (сөзқолданыс) деп анықтап, оған ағылшын тіліндегі балалар мен жасөспірімдер әдебиетінің 22 бөлімінен (жанрынан) лексика қосуды ұсынған. Соған байланысты сауалнама жүргізген. АҚШ-тың 221 мектебіне корпусқа қандай мәтіндерді қосқысы келетінін көрсетуді сұрайтын сауалнамалар жіберілді. Сауалнамаларды зерделегеннен кейін 19 мың кітап атауларының тізімі жасалды. Бұл жиынтықтан 1045 мәтін таңдап алынды. Олардың негізінде әрқайсысы 500 сөз қолданысының 10 000 элементарлы іріктемесі құрастырылды.

Жоғарыдағы Дж Синклердің принциптерінің ішінде 4-Критерий дереккөздерді таңдау критерийлері болады.

«Кез келген таңдау қандай да бір критерийлер бойынша жасалуы керек және корпусты құрудағы бірінші маңызды қадам корпусты құрайтын мәтіндер таңдалатын критерийлерді анықтау болып табылады. Жалпы критерийлерге мыналар жатады: 1. мәтін режимі; тіл сөйлеуден немесе жазудан туындады ма, әлде қазіргі уақытта электронды режимде ме; 2. мәтін түрі; мысалы, жазылған болса, кітап, журнал, хабарлама немесе хат; 3. мәтіннің домені; мысалы, академиялық немесе танымал; 4. корпустың тілі немесе тілдері немесе тілдік әртүрлілігі; 5. мәтіндердің орналасуы; мысалы (ағылшын тілі) Ұлыбритания немесе Австралия; 6. мәтіндердің күні»