3 Оқу пәнінің тақырыптары бойынша дәріс тезистері

1-бөлім. Корпус лингвистикасына кіріспе

Дәріс №1

Корпус лингвистикасының зерттеу нысаны мен негізгі ұғымдары

1. Корпус лингвистикасы туралы
2. Корпус лингвистикасының зерттеу нысаны мен негізгі ұғымдары

Корпустық лингвистика-компьютерлік технологияларды қолдана отырып, лингвистикалық корпустарды (мәтін корпустарын) құру мен пайдаланудың жалпы принциптерін әзірлеумен айналысатын компьютерлік лингвистиканың бөлімі. Мәтіндердің лингвистикалық немесе тілдік корпусы деп нақты лингвистикалық мәселелерді шешуге арналған, бірыңғай, құрылымдалған, белгіленген, филологиялық құзыретті тілдік мәліметтердің үлкен жиынтығы.

Корпустарды құрудың орындылығы мен пайдаланудың мәні мынадай алғышарттармен айқындалады: 1) корпустың жеткілікті үлкен (репрезентативті) көлемі деректердің типтілігіне кепілдік береді және тілдік құбылыстардың барлық спектрін толық көрсетуді қамтамасыз етеді;

2) әртүрлі типтегі деректер корпуста өзінің табиғи контекстік нысанында болады, бұл оларды жан-жақты және объективті зерделеуге мүмкіндік береді; 3) бір рет жасалған және дайындалған деректер массивін бірнеше рет, әртүрлі зерттеушілер және әртүрлі мақсаттарда пайдалана алады. Корпусты тілдік және сөйлеу бірліктері туралы әртүрлі анықтамалар мен статистикалық мәліметтер алу үшін пайдалануға болады. Атап айтқанда, корпустар негізінде сөз формаларының, лексемалардың, грамматикалық категориялардың жиілігі туралы мәліметтер алу мақсатында қолданылуы мүмкін.

Дәріс №2

Корпус лингвистикасының қалыптасуы

1. Корпус лингвистикасының пайда болуы: картотекадан корпусқа (Жұбанов жаңабековалардың кітабы Захаров пен Богдановалардан алынған)
2. Корпус лингвистикасының басқа салалармен байланысы

Корпустық лингвистика лингвистикадағы эмпирикалық деректермен айналысатын әдістер, процедуралар мен ресурстар жиынтығы түрінде ұсынылуы мүмкін. Қазіргі корпустық лингвистиканың методология ретінде көтерілуі лингвистиканың эмпирикалық ғылым ретіндегі тарихымен тығыз байланыстылығын көрсетеді. Тарихи лингвистика: тілдегі өзгерістер және қайта құру (Салыстырмалы тарихи әдіс). Қазіргі корпустық лингвистикаға әсер еткен басты бағыттардың бірі *салыстырмалы-тарихи тіл білімінен* келді. Бұл таңқаларлық емес, өйткені тарихи зерттеулермен айналысатын лингвистер әрдайым мәтіндерді немесе мәтіндер жинақтарын негізгі дәлелдер ретінде қолданған. XIX ғасырда дамыған көптеген технологиялар ежелгі тілдерді (праязыков) қайта құру немесе тілдер арасында байланыс орнату үшін қазіргі уақытта қолданылады. Үнді-еуропалық дәстүрде тілдік өзгерістерді зерттеу және қайта құру әрекеттері мәтіндерге немесе корпустарға (тарихи ескерткіштерге) байланысты болды. Я. Гримм және одан кейінгі кіші грамматиктер мәтіндерден алынған цитаталармен тілдердің тарихы мен грамматикасы туралы пікірлерін қолдады. Жас графиктер өздерінің Манифестінде диалектілерде жазылған қазіргі тілді зерттеу жүргізгендерін жариялады (тек ежелгі мәтіндерді зерттеу ғана емес) және бұл да үлкен маңызға ие болды. XIX ғасырдан бері дамып келе жатқан көптеген идеялар мен технологиялар корпус лингвистикасында қолданылып, содан кейін дамыды. Тарихи корпустарды құрастыру әлі де үлкен қызығушылық тудырады. Шынында да, электронды түрде қол жетімді алғашқы корпустардың ішінде тарихи корпустар да

болды. Электрондық форматта қол жетімді көптеген мәтіндердің пайда болуы көптеген деректерді тез жинауға мүмкіндік берді. Бұл лингвистерге лингвистикалық талдаудағы статистикалық әдістер арқылы жеңіске жетуге, сонымен қатар зерттеудің жаңа әдістері мен модельдерін жасауға және дамытуға мүмкіндік берді. Бүгінгі таңда тілдік өзгерістердің математикалық күрделі модельдерін электронды денелердегі мәліметтер арқылы есептеуге болады.

Дәріс №3
Корпус лингвистикасының тарихы

1. Линвгистикалық корпустарды құрудың тарихы
2. Қазақ тілінің ұлттық корпустарын құрастыру мәселесі

Лингвистер 1960 жылдары компьютерленген мәтіндердің алғашқы корпустарын жинады. Алғашқы компьютерленген корпус- Брауның корпусы (The Brown Corpus1) – 1961 жылы АҚШ-та алғаш рет жарияланған американдық кітаптардан, газеттерден, журналдардан 500 мәтінді қамтыған. Браун корпусындағы әр мәтіннің ұзындығы 2000 сөзден тұрады (қолданылатын сөздер – tokens) және бүкіл жинаққа 1 миллион сөз кіреді (әрқайсысында 2000 сөзден тұратын 500 мәтін). Корпус авторлары В. Френсис (W. Francis) және Г. Кучера (H. Kucera) бастапқы статистикалық өңдеудің көптеген материалдарымен бірге жүрді: жиілік және алфавиттік-жиілік сөздігі, әртүрлі статистикалық үлестірімдер. Браун корпусын құрудың *мақсаты*-жазбаша ағылшын тілінің жеке жанрларын жүйелі түрде зерттеу және жанрларды салыстыру. Оның пайда болуы жалпы қызығушылық пен қызу пікірталастарды тудырды. Біріншіден, олар мәтіндерді таңдау принциптері мен осындай корпуста шешілуі мүмкін міндеттердің құрамына тоқталды. Бір жағынан, ол статистикалық процедуралар негізінде құрылды; екінші жағынан, статистика корпус авторларының кәсіби түйсікке негізделген ерікті шешімдерімен бірге қолданылды. Бұл күрделі процестің максималды объективтілігіне қол жеткізу үшін тексеру және бақылау үшін барынша формализацияланған процедураларды құру қажет болды.

Дәріс №4
Корпустардың негізгі сипаттамалары

1. Корпустардың репрезентативтілігі
2. Корпустардың классификациясы

"Корпус" термині, әдетте, түпкілікті белгіленген мөлшердегі мәтіндер жиынтығын білдіреді. Уақыт өте келе корпустың көлемі мен құрамы өзгеруі мүмкін, бірақ бұл өзгерістер оның құрылымын өзгертпеуі керек. Ұсынылған корпустар әр түрлі яғни репрезентативті сипаттамаларға ие. Алғашқы корпустардың көлемі, жоғарыда айтылғандай, 1 миллион қолдануды құрады (Браунов корпусы, Ланкастер-Осло-Берген корпусы, Уппсала орыс тілі корпусы). Мұндай көлем тілді оның барлық алуан түрлілігінде көрсетуге мүмкіндік бермеді. Қазіргі уақытта жалпы тілдік (ұлттық) корпуста кемінде 100 миллион сөз қолданылуы керек деп саналады. Ұлттық корпус бұл тілді оның өмір сүруінің белгілі бір кезеңінде (немесе кезеңдерінде) жанрлардың, стильдердің, аумақтық және әлеуметтік нұсқалардың және т. б. барлық түрлерінде ұсынады. Барлық заманауи лингвистикалық зерттеулер мен сөздіктер мен грамматиканы құрастыру жұмыстары мәтіндердің өкілді (өкілдік) корпусын қолдануға бағытталған деп айта аламыз. Корпус авторларының міндеті- мүмкіндігінше көп мәтіндерді жинау. Корпус-бұл тілдің немесе тілдің қысқартылған моделі деп айта аламыз. Тәжірибе көрсеткендей, корпус лингвистикасы кем дегенде екі түрлі объект (мәтін корпусы) жұмыс істейді: 1.Бірінші типтегі корпустар әмбебап, олар сөйлеу әрекетінің барлық түрлерін көрсетеді. 2.Екінші типтегі корпустар қоғамдық сөйлеу практикасында кейбір лингвистикалық немесе мәдени құбылыстың болуын көрсетеді.

Дәріс №5
Корпустардың типологиясы

1. Корпустардың ерекше типтері
   1.1 Паралельді корпустар
   1.2 Сөйлеу тілінің корпустары

Параллелизм критерийі бойынша корпустар бір тілді, екі тілді және көп тілді болып бөлінеді. Көптілді корпустарда диалектілер мен тіл нұсқалары қарама-қайшы келеді. Мысалы, ағылшын тілінің ана тілі және ағылшын тілі шет тілі сияқты түрлері жаңа технологиялар пайда болғанға дейін ғылыми қызығушылықтан тыс қалды, бұл салыстырмалы сөйлеу жұмыстарының едәуір көп санын контрастты талдауға тартуға мүмкіндік берді. Екі тілді және көп тілді корпустар екі немесе бірнеше тілде тәуелсіз жазылған бір тақырыптық аймақтағы мәтіндерді біріктіреді(мысалы, әртүрлі елдерде және әртүрлі тілдерде өткен белгілі бір ғылыми мәселе бойынша конференция материалдарының корпусы). Мұндай корпустар терминологиямен жұмыс істеуге көмектеседі және оларды аудармашылар жиі қолданады. Екі тілді немесе көп тілді корпустың тағы бір нұсқасы – кез-келген бастапқы тілде жазылған түпнұсқа мәтіндер мен осы бастапқы мәтіндердің бір немесе бірнеше басқа тілдерге аударылған мәтіндер. Мұндай корпус салыстырмалы салыстырмалы зерттеулер жүргізу, аударма теориясын зерттеу және адам мен компьютерді аударуды үйрету үшін баға жетпес материал ұсынады.

Параллель корпустарды екі негізгі түрге бөлуге болады: 1) қандай да бір бастапқы тілде жазылған түпнұсқа мәтіндердің жиынтығын және осы бастапқы мәтіндердің бір немесе бірнеше басқа тілдерге аударылған мәтіндерін білдіретін корпустар; 2) екі немесе бірнеше тілде тәуелсіз жазылған, бір тақырыптық саладағы мәтіндерді біріктіретін корпустар. Осы және басқа да корпустар тілдерді салыстырмалы зерттеу үшін (лексикология, грамматика, стилистика, аударматану саласында және т.б.), сондай-ақ аударманың тиімді әдістерін, оның ішінде машиналық әдістерді әзірлеу мақсатында құрылады және пайдаланылады.

Прагматика компьютерлік лингвистика мен корпустық зерттеулерде лингвистиканың басқа салалары сияқты мұқият зерттелмеген, өйткені сөйлеу тілінің корпусын құру қиын болды. Корпусты құрастырушылар әрдайым оның көмегімен шешуге болатын лингвистикалық мәселелердің барлық түрлерін елестете алмайды. Олардың ішінде тілді жалпы түсіну үшін ерекше маңызды сала-ауызша мәтіндерді зерттеу. Лондон-Лунд корпусы (the London-Lund Corpus) "ағылшын тілін қолдануға шолу" (The survey of English Usage) жобасының аясында әзірленген. Жобаның мақсаты ана тілінде ағылшын тілінің грамматикалық жүйесінің ерекшеліктерін мүмкіндігінше толық бекіту болды. Жоба 1960 жылдан бастап Лондон университетінің колледжінде Р.Квирктің жетекшілігімен жасалды. Корпустың көлемі - 1 миллион сөз. Ауызша сөйлеу мәтіндері радиохабарлар, ресми құрылымдардың отырыстары, сондай-ақ бейресми әңгімелер болды. Корпустың машиналық нұсқасы 30 Лунд университетінде (Швеция) жасалды және 1979 жылы пайдалануға дайын болды. Дәл осы Лондон-Лунд ауызша сөйлеу корпусы машина оқитын алғашқы корпустардың бірі болды. Ол жасырын жазылған әңгімелерді білдіретін 34 мәтіннен тұрды, олар Дж. Свартвик және Р. Квирка" Ағылшын әңгіме корпусы " (1980) . Бұл кітап компьютерлік корпустар кең таралмаған кезде өте пайдалы болды және күрделі ауызша транскрипцияны қолдану қиын болды. Ауызша сөйлеу корпусын құрастырудағы қиындықтарға байланысты бұл корпус ұзақ уақыт бойы ауызша ағылшын тілін компьютерлік зерттеудің маңызды көзі болып қала берді.

2-бөлім. Корпустарды құру
Дәріс №6

Корпус құрудың алдын-ала жұмыстары

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

Кез келген корпустың жобасы оны құру кезеңдерін және оны одан әрі дамыту жолдарын көздеуі тиіс. Корпус ұғымы-лингвистер әрдайым жұмыс істейтін дәстүрлі картотекалардың жалғасы. XX ғасырда бұл картотекалар компьютерлік және көпшілікке қол жетімді болды. Корпустық тәсілдің қалыптасуында интернет маңызды рөл атқарды, оның даму барысында әртүрлі лингвистикалық зерттеулер жүргізуге жарамды мәтіндік материалдардың үлкен көлемі қол жетімді болды. Бұл жағдайда тілдік материалдың тепе-теңдігі туралы сұрақ туындайды. Бұл мәселе әсіресе ұлттық корпусты қалыптастыру кезінде басты кезекте тұр. Корпустың өкілдігі мәтіндік материалдың жеткілікті көлемімен де, оның әртүрлілігімен де қамтамасыз етілуі керек. Жанрлық-тақырыптық құрылымнан басқа, көптеген басқа, жеке, бірақ маңызды мәселелерді шешуге тура келеді, мысалы: 1. Корпустағы мәтін дегеніміз не? Мысалы, газеттердегі кішігірім жарнамалар-олар корпусқа жеке мәтіндер ретінде кіреді ме немесе оларды біріктіруге бола ма?
2. Газеттегі мақала мәтіні ме? Немесе газеттің бір шығарылымын бір мәтін ретінде қарастыру керек пе?
3. Жеке мәтін дегеніміз не?
4. Жарияланған хат-хабарлардағы әр хат жеке мәтін бола ма, хаттар бірыңғай дискурсты немесе осы хаттардың жиынтығын құрайды ма?
Мәтіндер корпусының маңызды ерекшелігі-бұл бір немесе басқа тілдің кездейсоқ біріктірілген мәтіндерінің жиынтығы ғана емес. Оны құру кезінде бірқатар проблемалар туындайды. Олардың негізгілері мыналар болып табылады:
1. Мәтіндер корпусының негізгі бірлігі не болуы керек?
2. Мәтін корпусының көлемі қандай болуы керек (оның құрамында қанша бірлік болуы керек)?
3. Мәтін корпусында қандай жазбаша мәтін көздері ұсынылуы керек және қандай мөлшерде?
4. Корпус құрамына кіретін мәтіндер қандай бастапқы тіл саласынан таңдалуы керек? Бұл сұрақтарға алғашқы жауаптар профессор Р.г. Пиотровский мен оның студенттерінің 1965-1980 жылдардағы көптеген зерттеулерінде берілді, олар жиілік сөздіктерін құрастыру және лингвостатистикалық зерттеулер жүргізу үшін мәтіндерді таңдауға қатысты болды. Сол проблемалар жиілік сөздігінің алғы сөзінде талқыланды, Л. Н. Засорина (1977). Дәл сол кезде 37 іріктеменің жалпы жиынтығын, іріктеменің көлемін, іріктеменің бөлігін (Элементарлық іріктеме) және т.б. бағалау үшін әртүрлі статистикалық әдістер қолданылды. Мәтіндер корпусының негізгі бірлігі пайдалану (әдетте сөздер деп аталады), негіздер (тамырлар, леммалар) және сөйлемдер болуы мүмкін. Қабылданған бірліктерде жасалған мәтіндер корпусының көлемі құру мақсаттарына байланысты. Әріптерді, әріптерді, дыбыстарды, дыбыстық тіркестерді қолдану жиілігін зерттеуде ол аз болуы мүмкін. Лексиканы, морфологиялық құбылыстарды зерттеуде және мәтіндердің синтаксистік немесе стилистикалық ерекшеліктерін зерттеуде ол әлдеқайда үлкен болуы керек . *Мынадай мәселелер де проблемалық болып табылады:*
1. Мәтіндер корпусына қандай функционалды жанрлар кіреді (көркем проза, драма, өлеңдер, ғылыми мәтіндер, газеттер, журналдар, техникалық сипаттамалар және т. б.)?
2. Мәтіндер корпусына қандай уақыт аралықтарын енгізу керек (қазіргі, 10 жыл, 50 жыл, Ежелгі және т. б.)?
3. Мәтіндер тек әдеби тілде немесе басқа дереккөздерде бола ма?
Бұл сұрақтарға жауап беру кезінде мәтін корпусын жасаушылар әдетте тіл білімі және лингвостатистика саласындағы мамандардың кеңестерін немесе сауалнама әдісін қолданады. Зерттеу тәжірибесіне сүйене отырып, мамандар мәтіндер корпусының жалпы көлемін, мәтіндерді шығару уақытын, мәтіндер саны мен қарапайым таңдау мөлшерін,

таңдалған мәтіндердің жанрларын және олардың санын, әр жанрдан алынған қарапайым үлгілер санын анықтайды.

Дәріс №7
Деректер жинау режимдері

1. Монитор корпусы тәсілдемесі (кітапта корпус линвгистикасы)
2. Веб корпус концептісі
3. Корпус үлгісі тәсілдемесі

Адам мен компьютердің өзара әрекеттесуі тіл білімінде қолданбалы принципті күшейтеді. Бодуэн де Куртенэ 20 ғасырда шешілуі қажет Тіл білімінің мәселелері туралы былай деп атап өтті: "Тіл білімінде сандық, математикалық ойлауды жиі қолдану керек және осылайша оны ғылымдарға дәл жақындату керек" . Қолданбалы лингвистикалық міндеттер олардың реттелген сипатымен ерекшеленеді. Көбінесе олар белгілі бір әлеуметтік тапсырысты білдіреді. Оларды іске асыру "Тапсырыс беруші-әзірлеуші" диалогында жүріп жатыр. Қолданбалы міндеттердің екінші ерекшелігі-олардың тексерілуі, ал тексерілуі қайталанатын, қайталанатын және әр уақытта жаңа материалда болады. Тіл-иерархиялық құрылымы бар семиотикалық жүйе. Бұл жүйе зерттеуді және ресімдеуді қажет етеді, бұл оны танудың тәсілдері мен әдістерін әзірлеуді, іске асыруды және біріктіруді өте өзекті етеді. Біріктіру дегеніміз-әр деңгейді зерттеу үшін бірдей әдістер жиынтығын қолдану. Табиғи тілді формализациялау тривиалды емес міндет болып табылады және нашар құрылымдалған проблемалардың барлық ерекшеліктеріне ие. Тілдің жабық ішкі жүйелердің ашық жүйесі екенін қолдану орынды болып көрінеді. Әрбір Ішкі жүйе шексіз, сондықтан оны модельдеуге болады, содан кейін ішкі жүйелер арасында белгілі бір қатынастар орнатылады.

Белгілеу мәтіндерге және олардың компоненттеріне арнайы тегтерді белгілеуден тұрады: лингвистикалық және сыртқы (экстралингвистикалық). Таңбалаудың келесі лингвистикалық түрлері бөлінеді: морфологиялық, семантикалық, синтаксистік, анафориялық, прозодикалық, дискурстық және т.б. талдаудың одан әрі құрылымдық деңгейлері кейбір денелерге қолданылады. Атап айтқанда, кейбір кішкентай корпустар толығымен синтаксистік түрде белгіленуі мүмкін. Мұндай корпустар әдетте терең аннотацияланған немесе синтаксистік деп аталады, ал синтаксистік құрылымның өзі тәуелділік ағашы болып табылады.

Мәтіндерді қолмен белгілеу (Аннотация) қымбат және уақытты қажет ететін міндет. Қазіргі уақытта корпустарды белгілеу үшін әртүрлі бағдарламалық құралдар ашық қол жетімді. Шартты түрде оларды бөлек (stand-alone) және веб-бағдарланған (web-based) деп бөлуге болады. Сонымен қатар, соңғы жылдары әзірлеушілердің назарын веб-қосымшаларға аударды. Бұл жүйелер бірқатар артықшылықтарға ие:

Көптеген интернет-сөздіктер пайда болды, қазақ тілін меңгеру және қазақ тілінде әртүрлі ақпараттар берілген әртүрлі веб-сайттар жасалды. Қазіргі уақытта, келесідей сайттар, платформалар, порталдар, блогтар, білім беретін сайттар жасалды және іске асырылды: Уикипедия (Википедия), Викибілім, Сөздік.кз, Тұған тіл және т.б.; көңіл көтеретін сайттар: Мәссаған,Қазақша КВН, ІргеТас және т.б.; танымдық сайттар: www.tanym.kz және т.б.; діни бағыттағы сайттар: Дін ислам, діни сайты және т.б. Біз қазақ тілін тасымалдаушыларға арналған және қаза тілі мен басқа халықтар тілі мен мәдениетіне қызығатын және меңгеретіндерге арналған интернет-ресурстардың пайда болуы мен дамуына куәгер бола аламыз, себебі осындай сайттар қазақ тілінде әртүрлі ақпарат береді. Осындай сайттардың, веб-парақшалардың, платформалардың, блогтардың толығуы мен өзіндік сипаты: қарапайым мазмұннан маңызды, өзіндік дизайн түрімен ерекшеленген ақпараттық толықтыруға дейінгі сипатымен ерекшеленеді. Қазақ тілінің әрқилы корпустары да жасалынды, ол өз кезегінде олардың жақсы динамикасы мен дамуын көрсетеді. Осыған байланысты корпустық лингвистиканың терминологиясын

меңгеру маңызды. Жобалық режимдегі оқыту мақсатында жасалған осы сөздік игеру үшін қажетті терминологиялық аппараттың ядролық қорын құрайтын, корпустық лингвистиканың тиісті терминдерін жүйелі беру тапсырмасын көздейді

Дәріс №8
Табиғи тілді өңдеудің негізгі рәсімдері

1. Токен және токендеу
2. Лемма және леммалау
3. Стемминг
4. Парсинг

Шын мәнінде, қазіргі кездегі корпус әрқашан компьютерлік мәліметтер базасы болып табылады және оны құру барысында арнайы процедуралар мен бағдарламаларды қолдану табиғи болып табылады. Мысалы, ***токендеу***, яғни табиғи тілдегі таңбалар ағынын жеке маңызды бірліктерге (токендер, сөз формалары) бөлу табиғи тілді одан әрі өңдеудің қажетті шарты болып табылады. Егер тілдерде керемет тыныс белгілері болса, токендеу қиын болмас еді – тіпті қарапайым бағдарлама. Бос орындар мен тыныс белгілерін басшылыққа ала отырып, мәтінді сөздерге бөле алады. Бірақ іс жүзінде тілдерде мұндай тыныс белгілері жоқ, бұл токендеу міндетін қиындатады. Мысалы, ағылшын тілінде бірегей таңбалануы мүмкін емес жағдайлар бар: *chap* бұл *chapter* сөзінің қысқартылған түрі немесе сөйлемнің соңында орналасқан *chap* сөзі болуы мүмкін. *Jan* сөзі *January* сөзінің қысқартылған түрі немесе сөйлемнің соңында орналасқан тиісті атау ретінде қарастыруға болады. Бірінші жағдайда нүкте сөзбен бірдей таңбалауышқа жатқызылуы керек, ал екінші жағдайда оны бөлек тегке бөлу керек. Алайда, мұндай қиындықтар өте шектеулі екенін және мәтінді өңдейтін көптеген қосымшалар көбінесе оларды елемейтінін (мысалы, қысқартулар мен күрделі сөздерді ескермейді) немесе оларды жеке алгоритммен өңдейтінін байқауға болмайды. Морфологиялық талдаудың тағы бір нақты міндеті – ***леммалау***, яғни басқа сөз формаларына негізделген 39 сөздің бастапқы формасын қалыптастыру процесі. Көптеген тілдерде бұл сөз әртүрлі ауытқулармен бірнеше формада болуы мүмкін. Мысалы, ағылшынша *'walk'* етістігі келесі формалармен ұсынылуы мүмкін:' *walk',' walked',' walks','walking'*. Сөздікте жазылған "*walk"* негізгі формасы сөз леммасы деп аталады. ***Леммалау дегеніміз****-талдау кезінде бір сөз ретінде өңделетін етіп бір сөздің әртүрлі флективтік формаларын топтастыру процесі*. Леммалаудан біршама өзгеше процесс **стемминг** деп аталады, ол сөздің стемін (негізін) табудан тұрады. Айырмашылық мынада, стеммер контекстті білмей жеке сөзді өңдейді, сондықтан сөйлеудің әртүрлі бөліктеріне байланысты әртүрлі мағынаға ие сөздерді ажырата алмайды. Алайда, стеммерлерді орындау оңайырақ және деректерді тезірек өңдейді, ал олардың төменгі дәлдігі көптеген қосымшалар үшін шешуші болмауы мүмкін. Мысалы, "жақсы" таңбалауышы "жақсы" леммасына сәйкес келеді, бірақ бұл стемминг кезінде төмендейді. "Walk" леммасы - "walking" таңбалауышының негізгі формасы және бұл сәйкестік стемминг кезінде де, лемматизация кезінде де анықталады. Төменде стемминг және лемматизация мысалдары келтірілген. Келесі сөйлем берілген: [the] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs]. Ең танымал стеммерлердің бірі SnowballAnalyzer келесі стемаларды шығарады: [quick] [brown] [fox] [jump] [over] [lazy] [dog].
Бұл сөйлемнің сөздерінің леммалары келесідей болады: [the] [quick] [brown] [fox] [jump] [over] [lazy] [dog]. Леммалау сөйлеу бөліктерін сәйкестендірумен байланысты және корпустан сөздерді тиісті таңбалауыштарға дейін қысқартуды қамтиды. Бұл леммалау, зерттеушіге барлық мүмкін нұсқаларды енгізуді қажет етпестен, жеке таңбалауыштың барлық нұсқаларын бөліп алуға және зерттеуге мүмкіндік береді.
Компьютерлік бағдарлама бойынша, қандай да бір сөзді іздегенде, экранға ең алдымен сол сөз кездесетін (метабелгіленімі берілген) мәтіндер, яғни мысалдар тізімі шығады.

Сонымен қатар экранның екінші бетіне әртүрлі ұяшықта сол сөз туралы лингвистикалық ақпараттар беріледі. Мысалы, «бала» сөзін алайық:
Мәтінде «баласына» формасында кездескенде, ең алдымен программа оның түбірін табады. Мұны *лемматизация* деп атайды.
Баласына: лемма: «бала» (түбірі)
лексикалық мағынасы: 1. Ата-ананың перзенті, ұрпағы, тұқым.
2. Нәресте, сәби, бөбек. 3 Жан-жануардың күшігі, құстардың балапаны.
семантикалық белгіленім: зат есім, дара, деректі, жалпы
морфологиялық белгіленім: бала/зт: сы/ТЖ-3+на/БС
Мұндағы ТЖ-3 – тәуелдік жалғау 3 жағы, БС – барыс септігі.
Егер іздеп отырған сөз мәдени тілдік бірлік болса, онда лексикалық мағынадан кейін, мәдени семантикасы да көрсетіледі, сонымен қатар нақтырақ түсіндіру үшін суреттері қоса шығады. Мәселен:
ЗҰЛПЫҚАРДЫҢ : *лемма:* «Зұлпықар» (түбірі)
Лексикалық мағынасы: Көктен түскен төрт қылыштың бірі.
Мәдени семантикасы: Көктен түскен төрт қылышты *хәмкам, сәмсам, зұлқажа, зұлпықар* деп атаған.
Семантикалық белгіленім: зат есім, күрделі, деректі, жалқы
Морфологиялық белгіленім: зұлпықар/зт+дың/ІС
Мұндағы ІС-ілік септігі.
Міне, корпус мәтіндерінен ізделген сөзге осылайша лингвистикалық сипаттамалар беріледі.

Парсинг (Талдау)-бұл тілдің лексемаларының (сөздердің, таңбалауыштардың) сызықтық тізбегін оның ресми грамматикасымен салыстыру үдерісі. Нәтиже әдетте тәуелділік ағашы (синтаксистік ағаш) болып табылады. Үлкен корпустар үшін автоматты талдағыштарды (талдаушыларды) құру компьютерлік лингвистиканың маңызды бағыттарының бірі болып табылады. Көптеген тәсілдер сапалық және сандық өлшемдерді біріктіреді.

Дәріс №9
Корпустарды белгілеу құралдары

1. Белгіленім ұғымы туралы
2. Лингвистикалық белгіленім және оның түрлері
   2.1 Морфологиялық белгіленім
   2.2 Синтаксистік белгіленім

Табиғи тілдерді өңдеуге арналған арнайы бағдарламалардың ішінде автоматты белгілеу бағдарламалары ерекше орын алады. Корпусты белгілеу (tagging, annotation), әсіресе қазіргі заманғы корпустың өлшемдерін ескере отырып, көп уақытты қажет ететін операция болып табылады. Егер таңбалаудың кейбір түрлері үшін, атап айтқанда анафориялық, просодикалық, автоматты жүйелерді құру әлі де күрделі болып көрінсе және жұмыстың негізгі бөлігі қолмен жасалса, онда морфологиялық және синтаксистік талдау үшін әр түрлі бағдарламалық құралдар бар, оларды сәйкесінше тегтер (Тегтер) және парсерлер (парсерлер) деп атайды. Автоматты морфологиялық талдау бағдарламаларының (теггерлердің) жұмысы нәтижесінде әрбір лексикалық бірлікке грамматикалық сипаттамалар, оның ішінде сөйлеу бөлігі, лемма және граммдар жиынтығы (мысалы, жыныс, Сан, жағдай, анимация/жансыздық, ауысу және т.б.) жатады. Автоматты синтаксистік талдау бағдарламаларының нәтижесінде сөздер мен сөз

тіркестерінің арасындағы синтаксистік байланыстар жазылады, ал тиісті сипаттамалар (сөйлем түрі, фразаның синтаксистік функциясы және т.б.) синтаксистік бірліктерге жатады. Алайда, табиғи тілді автоматты түрде талдау ақылға қонымды және мағыналы емес – ол, әдетте, бір лексикалық бірлікке (сөздер, сөз тіркестері, сөйлемдер) бірнеше талдау нұсқаларын береді. Бұл жағдайда олар грамматикалық омонимия туралы айтады. Жалпы түсініксіздікті (морфологиялық, синтаксистік) жою компьютерлік лингвистиканың маңызды және күрделі міндеттерінің бірі болып табылады. Корпусты құру кезінде түсініксіздікті жою үшін автоматты және қолмен әдістер қолданылады. Жаңа буын корпустарында жүздеген миллион сөздер бар, сондықтан адамның араласуын барынша азайтатын жүйелерді дамытудың принциптері ұсынылған. Морфологиялық немесе синтаксистік түсініксіздікті автоматты түрде шешу, әдетте, статистикалық әдістерді қолдана отырып, жоғары деңгейдегі ақпаратты (синтаксистік, семантикалық) қолдануға негізделген. Әр түрлі лингвистикалық мәселелерді шешу үшін мәтіндер жиынтығы жеткіліксіз. Сондай-ақ, мәтіндерде әр түрлі қосымша лингвистикалық және экстралингвистикалық ақпаратты нақты көрсету қажет. Сонымен, браунов сияқты корпус материалында сөздердің жиілігін оңай анықтауға болады-оларды белгілі бір контексте үнемі қолдану. Алайда, бұл токендердің жиілігі болады (сөз формасы). Таңбалауыштардың жиілігін анықтау үшін әр сөзге оның леммасы берілуі керек.

Сонымен, *белгілеу мәтіндерге* және оларға қатысты арнайы тегтер компоненттері: лингвистикалық, лексикалық, грамматикалық және басқа да сипаттамаларды сипаттайтын және сыртқы, экстралингвистикалық автор және мәтін туралы ақпарат: автор, атауы, жарияланған жылы және орны,жанр, тақырып).

Лингвистикалық таңбалау- таңбалаудың лингвистикалық түрлерінің ішінде: морфологиялық, синтаксистік, семантикалық, анафориялық, прозодикалық, дискурстық және т. б. ерекшеленеді.Олардың барлығы мынадай қағидаттарға сәйкес жүзеге асырылады:

1) таңбалау схемасын сипаттау (негіздеу);
2) лингвистикалық ұғымдардың жалпы қабылданған жүйесі;
3) пайдаланушыға белгілі талдау схемасы;
4) параметрлерді енгізудің уәжділігі;
5) теориялық бейтарап (дәстүрлі) белгілеу схемасы;
6) халықаралық стандарттарды ұстану.

Морфологиялық белгілеу -шетел терминологиясында part-ofspeech tagging (POS-tagging), сөзбе – сөз-ішінара белгілеу термині қолданылады. Шындығында, морфологиялық белгілерге сөйлеу бөлігінің белгісі ғана емес, сонымен қатар сөйлеудің осы бөлігіне тән грамматикалық категориялардың белгілері де кіреді. Бұл таңбалаудың негізгі түрі: біріншіден, үлкен денелердің көпшілігі морфологиялық белгіленген корпустар, екіншіден, 46 морфологиялық талдау талдаудың одан әрі формаларының негізі ретінде қарастырылады-синтаксистік және семантикалық, үшіншіден, компьютерлік морфологиядағы жетістіктер үлкен денелерді автоматты түрде дұрыс белгілеуге мүмкіндік береді.

Дәріс №10

Лингвистикалық белгіленім және оның түрлері

1. Семантикалық белгіленім
2. Анафоралық белгіленім және просодикалық белгіленім
3. Экстралингвистикалық белгіленім

Семантикалық белгілеу - семантикалық тегтер көбінесе осы сөзді немесе сөз тіркесін қамтитын семантикалық категорияларды және оның мағынасын сипаттайтын тар ішкі категорияларды білдіреді. Корпустардың семантикалық таңбалануы сөздердің мағынасын нақтылауды, омонимия мен синонимияны шешуді, сөздерді санаттауды (категорияларды),

тақырыптық сыныптарды, каузативтілік белгілерін, бағалау және туынды сипаттамаларды және семантикалық белгілеудің өзіндік нұсқасын ұсынады. Лексика-семантикалық тегтер келесі өрістер бойынша топтастырылған: * таксономия (лексеманың тақырыптық класы) – зат есімдер, сын есімдер, етістіктер мен үстеулер үшін; * мереология ("бөлік – бүтін", "элемент – жиын" қатынастарын көрсету) – пәндік және объективті емес атаулар үшін; * топология (таңбаланатын объектінің топологиялық мәртебесі) – пәндік атаулар үшін; • каузация – етістіктер үшін; • қызметтік мәртебе – етістіктер үшін; • бағалау-пәндік және бейметалдық атаулар, Сын есімдер мен үстеулер үшін. Сөзжасамдық сипаттамалар бірнеше түрді қамтиды: • морфо-семантикалық сөзжасамдық белгілер (мысалы, "каритив", "семельфактив") • * сөзжасамдық категория (мысалы, ауызша зат есім); * сөзжасамдық лексика-семантикалық (таксономиялық)тип (мысалы, сын есімнің өлшемінен құралған жарнама); * сөзжасамның морфологиялық түрі (субстантивация, күрделі сөз).

Белгілеудің басқа түрлері бар, атап айтқанда: анафориялық белгілеу. Ол анықтамалық байланыстарды, мысалы, есімдік байланыстарды бекітеді. просодикалық белгілеу- екпін мен интонацияны білдіретін тегтер қолданылады. Ауызша сөйлеу корпусында просодикалық таңбалау көбінесе үзілістерді, қайталануларды, ескертпелерді және т.б. білдіретін дискурстық белгілеумен бірге жүреді.

Экстралингвистикалық таңбалар немесе метадеректер "сыртқы", "интеллектуалдық" белгілеулерді (библиографиялық сипаттамалар, типологиялық сипаттамалар, тақырыптық сипаттамалар, әлеуметтанушылық сипаттамалар), "формальды" құрылымдық белгілеулерді (мәтін, бөлім, тарау, бөлім, абзац, сөйлем), сондай-ақ техникалық-технологиялық белгілеулерді (кодтау, өңдеу күндері, орындаушылар, электрондық нұсқаның көзі) қамтиды. Метадеректер жиынтығы көбінесе корпустардың зерттеушілерге беретін мүмкіндіктерін анықтайды. Бұл деректерді таңдағанда, зерттеу мақсаттары мен лингвистердің қажеттіліктерін, сондай-ақ мәтінге белгілі бір қосымша белгілерді енгізу мүмкіндіктерін басшылыққа алу керек.

Дәріс №11
Корпус лингвистикасындағы стандарттау

1. Белгіленген мәтіндерді таныстырудың тілдік құралдары
2. Халықаралық стандарттар және жобалар

Корпустар, әдетте, көптеген пайдаланушылар бірнеше рет қолдануға арналған, сондықтан оларды белгілеу және лингвистикалық қолдау белгілі бір түрде біріктірілуі керек. *Корпусқа қатысты стандарттар* әдетте белгілеу түрлерінің үйлесімділігіне әсер етеді. Оларды кейде *"кодтау стандарттары"* деп атайды. Сондай-ақ, әртүрлі корпустардың салыстырылуына, соның ішінде олардың әртүрлі тапсырмаларға жарамдылығы туралы бағалауларға байланысты мәселе маңызды. Олар*" бағалау стандарттары "* деп аталады. Таңбалауға келетін болсақ, лингвистикалық және экстралингвистикалық таңбалар мәтіндер мен тілдік бірліктерді сипаттаудың кең таралған және жалпы қабылданған принциптеріне негізделуі керек. Белгілеу параметрлері және олардың мәні жеткілікті "табиғи" болуы керек, яғни.жалпы қабылданған Ғылыми жіктеулерге сәйкес келуі керек. Корпус-менеджерлерді лингвистикалық және бағдарламалық қамтамасыз ету үлгілік сұрау салуларды өңдеуді және үлгілік міндеттерді шешуді қолдауы тиіс. Деректерді ұсынудың бірыңғай форматтары көптеген жағдайларда бірыңғай бағдарламалық жасақтаманы пайдалануға және Корпус деректерімен алмасуға мүмкіндік береді. Бір жағынан, деректерді ұсыну форматтарын оларды толтыру тұрғысынан, екінші жағынан олардың құрылымы тұрғысынан Стандарттау туралы айтуға болады. Ең үлкен қиындық-ауызша сөйлеуді транскрипциялауды стандарттау. Ауызша сөйлеуді графикалық бекіту саласында, тіпті бірыңғай және міндетті стандарт болмаса да, белгілі бір прогреске қол жеткізілді (ең алдымен прецеденттердің болуымен байланысты), онда табиғи тілдік қарым-қатынастың вербальды емес компонентін сипаттауда стандарттар әзірленбеген, бұл

осы салада одан әрі ілгерілеуді қиындатады. Корпустарға қатысты стандарттау, мәліметтер типтерінің үйлесімділігі әр түрлі корпустардың салыстырмалылығы тұрғысынан да маңызды. Сонымен қатар, корпустар сандық және сапалық бағалауға ұшырауы мүмкін. Корпустар туралы сандық деректер олардың көлемін, корпустың әртүрлі критерийлер бойынша толтырылуын, корпустың немесе қосалқы корпустың лингвистикалық параметрлерін бағалауға мүмкіндік береді. Сапалы бағалау дегеніміз-нәтижелерді талдау негізінде корпустарды бағалау және салыстыру.

3-бөлім. Корпустарды қолдану

Дәріс №12

Корпус менеджерлері

1. Корпус іздеу жүйесі ретінде
2. Сұрау салу (сауал) тілдері

Еркін қол жетімді пайдалану жеткілікті көптеген мәтінді өңдеу құралдары электрондық өнімдерге мәтіндер жиынтығы лингвистикалық ақпаратты жинақтау және өңдеу зерттеушінің міндеттері.

"Мәтіндер корпусы" ұғымының ажырамас бөлігі мәтіндік және лингвистикалық деректерді басқару жүйесі,жақында ол көбінесе корпус деп аталады.Менеджер (немесе корпус менеджері). Корпус менеджері-бұл бағдарламалық жасақтаманы қамтитын мамандандырылған іздеу жүйесі
корпуста деректерді іздеуге, статистикалық деректерді алуға арналған құралдарақпаратты ұсыну және нәтижелерін пайдаланушыға ыңғайлы нысан.

Корпус менеджері:

-KWIC (мәтінмәнде кілт сөз) және толық құрыңыз конкорданттық тізімдер;

-жеке сөздерді ғана емес, сөз тіркестерін де іздеңіз;

-шаблондар бойынша іздеуді жүзеге асыру (күрделі сұраулар);

-таңдалған бірнеше критерийлер бойынша тізімдерді сұрыптаңыз пайдаланушы;

- табылған сөз формаларын көрсетуге мүмкіндік береді шексіз контексте;

-жеке элементтер бойынша статистикалық ақпарат беруге;

корпус;лемманы, сөз формаларының морфологиялық сипаттамаларын және байланысты метадеректер (библиографиялық, типологиялық)корпустың бөліну дәрежесі;

-нәтижелерді сақтау және басып шығару;

- жеке файлдармен де, корпустармен де жұмыс істеуге,
неограниченными мөлшері бойынша;

-сұрауларды жылдам өңдеу және нәтиже беру;

- әр түрлі мәтіндік деректер форматтарын (txt, doc, rtf) қолданыңыз,html, xml және т. б.);

пайдалану оңай (интуитивті) болуы керек тәжірибелі және жаңадан келген пайдаланушы үшін

Ең танымал әмбебап корпус менеджерлері SARA, SARA (BNC), Manatee/Bonito, CQP, DC сияқты. Өңдеу үшін корпус деректерін менеджерлер келесілер негізінде әзірлей алады. деректер базасын басқару жүйелері (ДҚБЖ) немесе іздеу жүйелері.Мысалы, орыс тілінің ұлттық корпусы бойынша іздеуindex іздеу жүйесі арқылы жүзеге асырылады.Server Professional. Грамматикалық және метатекстік ақпаратты іздеу үшін index қабілеттері қатысады.Жасырын қасиеттерді іздеу сервері(атрибуттар) және мәтін фрагменттері. Іздеу жүйесін беру index құралдарының көмегімен қалыптасады.Server, ол ескере отырып, толық мәтінді ақпаратты іздеуді қамтамасыз етеді веб-серверде немесе корпоративтік желіде орыс тілінің морфологиясы. Іздеу орыс, ағылшын және орыс тілдерінің морфологиясын ескере отырып жұмыс істейді.

Дәріс №13

Корпус менеджерлері

1. Шығарылым интерфейстері (выходные интерфейсы)
2. Лингвистикалық емес корпустардың корпус менеджерлері

Интернет желісін ақпараттық толықтыру (веб-кеңістік). Оны үлкен көптілді корпус ретінде қарастыруға болады.Лингвистикалық талдаудың негізгі материалы-тіл,сөйлеу туындылары түрінде жазылған интернетте үлкен көлемде және әртүрлілікте және тікелей ұсынылған машинамен өңдеуге болады. Бұл факт лингвистерге маңызы зор, өйткені мәтіндерді машиналық аудармаға аудару ғимараттардың пішіні мен құрылысы уақытша және материалдық талдауды талап етеді.

Мәтіндік массивтері интернетте корпусты қалыптастыру үшін деректер көзі ретінде кеңінен қолданыладыСондай-ақ интернетте кең ұсынылған мәтіндер тест ретінде қолданылады.Мәтінді талдау мен өңдеудің әртүрлі бағдарламаларына арналған материал ақпарат (әсіресе статистикалық және ықтималды әдістері).

Бұл ретте веб-кеңістікті тікелей дене ретінде қарауға болады.Бұл мәселе әсіресе 2001 жылы А.Килгарифтің баяндамасынан кейін белсенді бола бастады және талқыланды [45]. Оған ешбір корпус тең келе алмайтыны анық.Қамтитын тілдік материалдың вебпен репрезентативтілігі материалдар және басқа интернет қызметтері (мысалы, электронды пошта). Сонымен бірге, веб-корпустың тепе-теңдігі туралы сұрақ туындайды. Әлбетте, интернетте әлі де тіл сөйлеудің белгілі бір түрлері шығармалар бұрынғыға қарағанда жиірек ұсынылған.

Веб-кеңістікті корпус ретінде пайдаланған кезде, рөлкорпус менеджерлері іздеу жүйелері арқылы орындалады. Ғаламторда 72 еске түсіретін классификациялық типті жүйелер бар кітапхана каталогтары (анықтамалар, орысша ортақ атауы «анықтамалар-анықтамалар»).

Бұл жүйелердің деректер базасы белгілі бір мағынада корпустар деп санауға болады. Семантикалық түрі, бірақ негізгі іздеу құралы желідегі ақпарат жаһандық ақпарат болып табылады.Ауызша түрдегі іздеу жүйелері (іздеу жүйелері – іздеу қозғалтқыштар) бүкіл интернет кеңістігін индекстеу. Сонымен бірге бұл сөздік жүйе көрсеткіштерінің қалай құрастырылғанын түсіну пайдалы және сәйкесінше, негіздерді пайдалану кезінде осы мүмкіндіктерді ескеріңіз лингвистикалық материал ретінде іздеу жүйесінің деректері зерттеуге болады.

Мұндай жүйелердің көптеген түрлері бар, олар әртүрлі және бір-бірінен ерекшеліктері сұрау тілі, дизайн, сервис және т.б.Вербалды түрдегі негізгі іздеу жүйелерінің арасында (ең алдымен деректер қорының көлемі бойынша) мыналарды жатқызу керек: Google, Fast Search (AllTheWeb), AltaVista, WiseNut,HotBot, MSN іздеу, Teoma. Орыс жүйелерінің ішінде негізгі үшеуі бар: Яндекс (Яндекс, Яндекс), Рамблер (Rambler), Апорт!

Кез келген іздеу жүйесінің негізгі бөліктері ретінде үшеуі бар:

1. Робот – қарауды (сканерлеуді) қамтамасыз ететін ішкі жүйе. Интернет және инверттелген файлды сақтау (индекс базасы деректер) жаңартылған. Бұл бағдарламалық пакет болуы және желінің ақпараттық ресурстарының жағдайы туралы ақпарат жинаудың негізгі құралы болып табылады.
2. Іздеу деректер қоры – индекс деп аталатын – арнайы ұйымдастырылған деректер құрылымы (индекстік деректер базасы),оның ішінде негізінен тұратын инверттелген файл индекстелген веб-құжаттардан алынған лексикалық бірліктер,және осы бірліктер туралы әртүрлі ақпаратты қамтитын (73 атап айтқанда, олардың құжаттардағы ұстанымдары), сондай-ақ құжаттардың өздері туралы және жалпы сайттар.
3. Іздеу жүйесі – қамтамасыз ететін іздеу ішкі жүйесі пайдаланушының сұранысын өңдеу (рецепт іздеу), іздеу дерекқор және іздеу нәтижелерін пайдаланушыға жеткізу. Іздеу жүйесі жүйе пайдаланушымен пайдаланушы арқылы байланысады.Интерфейстер – браузер бағдарламаларының экрандық формалары: интерфейс сұрауларды және іздеу нәтижелерін қарауға арналған интерфейсті қалыптастыру.

Индекс файлы (немесе жай ғана индекс) жылдам іздеуге бағытталған байланысты файлдар сұрау бойынша деректер жиыны болып табылады.Индекс әрқашан инверттелгенге негізделген файл. Төңкерілген іздеу массивін ұйымдастыру схемасы арқылы құжаттарға қолжетімділікті қамтамасыз ету принципіне негізделген мазмұн идентификаторлары. Мұндай схема құжаттардың дәйекті массивін өңдеу үшін арнайы көмекші инверттелген файлдар – нүктелер қол жеткізу арқылы алынған

Дәріс №14
Корпус зерттеулері

1. Корпустарды қолданушылар және корпустарды қолдану жолдары
2. Корпусқа негізделген лексикографиялық зерттеулер
3. Корпусқа негізделген грамматикалық зерттеулер
4. Корпусқа негізделген дискурс зерттеулері

Лексикология үшін электронды корпустардың көмегімен зерттеу жүргізу жалпы қабылданған стандартқа айналды. Осыған байланысты корпустық талдау деректеріне негізделген және соңғысының сөздік мақалалардың құрылымына және иллюстрациялық мысалдарды таңдауға әсерін анықтауға мүмкіндік беретін лексикографиялық зерттеулер перспективалы болып табылады. Корпустық лингвистика әдістері Электронды сөздіктердегі мағынаның анықтамаларын оңтайландыру үшін сөздік бірліктердің мағыналық парафраздарын талдау үшін қолданылады. Корпустағы опциялар тақырыбы пікірталас тудырады. Ғалымдарды көп функциялы стандартталған көлемді корпуста зерттелетін құбылыстың нұсқаларын қаншалықты ескеру керек және сөздіктің сәулеті мен дизайны опция факторларымен қаншалықты анықталуы керек деген сұрақ қызықтырады. Қолданыстағы корпус белгілі бір лингвистикалық қауымдастықтың тілдік қолданысы үшін жеткіліксіз болуы мүмкін. Алайда, жеке пропорционалды түрде ұйымдастырылған пропорционалды корпус ретінде ол коммуникативті жанрлардың алуан түрлілігінің бөлімі болып табылады. Мұндай жағдайда опция кез-келген жағдайда өңделмеген тілдік фактілер деңгейінде болады. Басқа мәдени аймақтың тілдік материалымен жұмыс істеу кезінде мәдени фондық білімді талдау үшін қажет, әсіресе тілдердің аймақтық түрлерін зерттеу үшін маңызды тиісті құжаттама мәселесі туындайды. Көптеген лингвистердің пікірінше, мәтіндік корпус-бұл ғылыми зерттеулерде қолданылатын көптеген мәтіндер.

- лексикалық дағдылар өзіне әр түрлі деңгейдегі сипаттамаларды қосады;
- белсенді және белсенді емес дағдыларға әр түрлі түсіндірме беріледі;
- сөз мағынасы, қалыптастыру, сәйкестік деңгейі, мысалдарда келтіру лексикалық ережелердің міндетті сипаттамасы болып табылады;
- лексикалық ережелер салыстыруда тіларалық және ішкі тілдік деңгейде құралады;
- лексикалық ережелер және оны құрайтын сипаттамалар оқушылармен бірге эвристикалық жолмен шығарылады;
- лексикалық ережелер ішінде бірліктік оқытулық лексемді семантикалық бірліктерді анықтап білу мен қолдануға нұсқаулық-ережеге және белгілі лексикалық бірліктер тобының жүйеліленген ерекшеліктері жинақталған ереже бойынша анықталады;
- лексикалық ережелер әр түрлі түрде қарастырылады: модельдік фразада, контекстік жағдайда немесе иллюстрацияларда, әрекетке нұсқаулық ескертпе, алгоритмдерде.

Лексикалық дағдыларды дамытуға арналған жаттығулар құрастыру кезінде лексикалық дағдылардың барлық бөліктерін ескеру қажет.
Лексикалық тапсырмалардың көптеген көзқарастарға байланысты классификациялауға болады.

1. Сөйлеу әрекетінің түрлеріне байланысты рецептивті мен репродуктивті жаттығулар, олар тағы репродуктивті және продуктивті болып бөлінеді.

2. Лексикалық дағдылар даму сатыларына, яғни сөзді меңгеру сатысы.
3. Лексикалық дағдылардың бағыттылығына байланысты бөлек аспектілерге фонетикалық, сызбалық және т.б. бөлінеді.
4. Белгілі сөздердің байланысын құруға бағытталған жаттығулар.
5. Бөлектенген сөзбен немесе сөз тіркестегі, сөйлемдегі сөзбен жатығулармен жұмыс жасау болжалады (И.В Рахманова контексті және контексті емес жаттығулар).
6. Оқытылуға тиіс сөзбен жасалған жұмыс түрлері (ауыстыру, кірістіру және т.б.).
7. Лексикалық жаттығулардың меңгеруге бағыттылығы.

Тілдің жиілік грамматикасы. Тілдің толық жиіліктік грамматикасын құру соңғы кезге дейін қиын міндет болып көрінді, дегенмен оны шектеулі материалда шешетін көптеген зерттеулер бар

Дәріс №15
Әр типті корпустарға шолу

1. Шетел корпустары
2. Орыс тілінің корпустары
3. Қазақ тілінің ұлттық корпустары

Шетел корпустары. Қазіргі американдық ағылшын корпусы (Қазіргі Американдық Ағылшын – COCA корпорациясы) - бұл ағылшын тіліндегі ең үлкен корпус http://corpus.byu.edu/coca/, және ағылшын тілінің американдық нұсқасының жалғыз үлкен және теңдестірілген корпусы. Оны 2008 жылы М.Дэвис (Brigham Young University, АҚШ) жасаған. COCA құрамында 410 миллион сөз бар және 1990 жылдан бастап қазіргі уақытқа дейін ауызша сөйлеуді, көркем прозаны, танымал журналдарды, газеттерді және 76 ғылыми әдебиетті білдіретін мәтіндерді қамтиды. Ол жылына екі рет жаңартылады және тілде болып жатқан өзгерістерді бақылау үшін ыңғайлы. Неміс мәтіндік корпустарынан dereko корпусын (das Deutsche Referenz Korpus) мекен-жайы бойынша атап өту керек http://www.ids-mannheim.de/kl/projekte/korpora/. Мангеймдегі (Германия) неміс тілі институтының жобасы аясында құрылған электрондық жиналыс көркем әдебиеттен, ғылыми және публицистикалық мәтіндерден тұрады және 4 миллиардтан астам сөздерді (16-дан астам) қамтиды. 08. 2010). Бұл, мүмкін, әлемдегі ең үлкен корпус, бірақ ол жеке неміс тілді қосалқы корпорациялардың жиынтығы ретінде безендірілген. Корпуста TEI ұсыныстарына сәйкес жасалған SGML негізіндегі морфосинтактикалық түзету бар. Неміс корпусымен жабдықталған Cosmas II корпус менеджері лексикалық бірліктер мен сөз формаларының морфологиялық белгілері бойынша іздеуге мүмкіндік береді. *Британдық ұлттық корпус* (BNC) - бұл 100 миллионнан астам ауызша және жазбаша ағылшын сөздері бар үлкен анықтамалық ғимараттардың бірі. Ол Оксфорд университетінде Ланкастер университеті мен Британ кітапханасының қатысуымен әзірленді. Корпусты құру бойынша жұмыс 1991 жылдан 1994 жылға дейін жалғасты. Жазбаша ағылшын тілін ұсынатын субкорпус бүкіл корпустың 90% құрайды және әртүрлі жастағы газеттер, мерзімді ғылыми басылымдар мен журналдарды, танымал ғылыми фантастиканы, жарияланған және жарияланбаған хаттарды, мектеп пен университет жазбаларын және т.б. қамтиды. т. Ауызша сөйлеудің қосалқы корпусы жобаға өз еркімен қатысқан, Ұлыбританияның әртүрлі бөліктерінде тұратын және әртүрлі әлеуметтік таптарға жататын әртүрлі жастағы адамдардың сөйлеуін қамтиды. Ауызекі сөйлеу көптеген контекстермен қоршалған: Ресми іскери немесе үкіметтік кездесулерден бастап радио шоулары мен телефон сөйлесулеріне дейін

Орыс тілінің корпустары. Ұзақ уақыт бойы 87 лингвист жұмыс істей алатын орыс тілінің қоғамдық, өкілді және белгіленген корпусы болған жоқ. Мұндай корпусты құру бойынша тікелей жұмыс тек 2000 жылы басталды, дегенмен 1980-ші жылдардан бастап белгілі бір жетістіктер болған . Орыс тілінің ұлттық корпусы (ОТҰК) – бұл электронды түрдегі орыс мәтіндерінің жиналысына негізделген ақпараттық-анықтамалық жүйе. Ол

алғаш рет сайтқа орналастырылды http://ruscorpora.ru / 2004 жылдың сәуір айында. Корпус орыс тіліне қатысты әртүрлі мәселелерге қызығушылық танытқандарға арналған: кәсіби лингвистер, Тіл оқытушылары, мектеп оқушылары мен студенттер, орыс тілін үйренетін шетелдіктер. Орыс тілінің ұлттық корпусы өкілдік критерийіне және қазіргі заманғы ғимараттарға қойылатын басқа да талаптарға жауап береді, бұған оның келесі сипаттамалары дәлел бола алады: 1) 176 миллионға жуық сөздерді қолданатын ОТҰК көлемі (сайт деректері бойынша http://ruscorpora.ru / 2011 жылдың ақпан айына); 2) орыс тілін қолданудың барлық негізгі салаларына жататын оны құрайтын мәтіндердің жанрлық әртүрлілігі (ғылыми, ресми-іскерлік, публицистикалық, шіркеулік-буындық, көркемдік, ауызекі және электрондық коммуникацияны қоса алғанда, ауызекі тіл-тұрмыстық); 3) негізгі әлеуметтік параметрлері бойынша (жасы, білім деңгейі және тілді меңгеру, кәсіби тиесілігі, сөйлеу мәдениетінің түрлері)шығармалары корпусқа енген авторлардың құрамы (кемінде 20 мың) өте алуан түрлі.; 4) ОТҰК -та тілдік құбылыстарды қолданудағы өзгерістерді қадағалауға және осы өзгерістердің динамикасын белгілеуге мүмкіндік беретін, құрудың әртүрлі кезеңдеріне жататын мәтіндердің болуы

Қазақ тілінің ұлттық корпустары. Ұлттық корпус ғылыми зерттеулердің түр-түрін жүргізуді қамтамасыз етеді: лексикографияға, жасанды интеллектіге, әдебиеттануға, сөйлеу тілін талдау мен жинақтауға және лингвистиканың барлық салаларына қатысты зерттеу түрлері. Сонымен бірге беделді академиялық сөздіктер мен ғылыми грамматикаларды құрастыру да корпустар негізінде жүзеге асады. Ұлттық корпусты пайдаланушылар — әр түрлі саладағы тілшілер, әдебиеттанушылар, тарихшылар және гуманитарлық білім салаларының өкілдері. Ұлттық корпустың ана тілі мен шетел тілін оқытуда, оқулықтар мен бағдарламалар құрастыруда маңыздылығы аса зор деуге болады.
Бұл жоба жасалуының соңғы кезеңінде (9 жылда) терең аннотацияланған (әр сөзіне белгіленім қойылып, тілтанымдық және энциклопедиялық ақпараттар берілген) қазақ тілінің электрондық мәтіндерінің 300 миллион сөзқолданыстағы көлемін қамтитын Қазақ тілінің ұлттық корпусы деп аталатын кеңауқымды инновациялық-ақапарттық ашық жүйе ретіндегі мегажобаның қарапайым бастапқы нұсқасы болып табылады. Жалпы, идеалды түрдегі Қазақ тілінің ұлттық корпусында ұсынылып отырған қарапайым бастапқы нұсқаға қарағанда бірнеше жүздеген есе кең ауқымды және әлдеқайда ұзын тереңдіктегі (ортағасырлардан бастап бүгінге дейінгі) мәтіндер корпусы тілдегі пропорциясына қарай жанр-жанрмен, стиль-стильмен қамтылады, сондай-ақ ондағы белгіленім (разметка) түрлері де алуан-алуан болмақ. Ал мына ұсынылып отырған қарапайым бастапқы нұсқада біз 10 миллиондай сөзқолданыстағы мәтін көлемін қамтып отырмыз, оның тереңдік деңгейі де үлкен емес, негізінен осы заманғы қазақ мәтіндері қамтылды. Бұл жобаның мәтіндік базасына 15 томдық «Қазақ әдеби тілі сөздігінен» иллюстрациясынан мысалдар қосылды. 15 томдық сөздік мысалдары әртүрлі дереккөздерден алынғандықтан, қазақ әдеби тілінің барлық жанрларын қамтиды деуге болады. Бұған қоса 5 млн. сөзқолданыс көркем проза, поэзия, драматургия, ғылыми-гуманитарлық, публицистикалық стильдер бойынша алынып отыр.

Лингвистикалық және экстралингвистикалық белгіленімдердің бастапқы әзірлемесі жасалды. Атап айтқанда, метабелгіленімдер корпус жадына салынған мәтіндердің дереккөздері туралы мәліметтермен жабдықталған. Сонымен бірге бұл жобада осы мәтіндерге қатысты қолданылған қарапайым формадағы белгіленім түрлері де шектеулі, атап айтқанда: морфологиялық белгіленім, морфо-семантикалық белгіленім берілді, сонымен бірге лексикалық семантика мен мәдени семантика көрсетілді. Соңғысының қойылуы қолмен жүзеге асырылғандықтан (болашақта жартылай автоматты формасына қатысты бағдарлама әзірленеді), мұндай семантика тек белгілі бір тізімдегі сөздер қатарымен шектелді. Жобада олардың тізімі көрсетілді. Демек бастапқы қарайым нұсқадағы мәдени семантика тек осы тізімдегі сөздерде берілгендіктен, іздеуші оларды тізім бойынша теру арқылы ғана көре алады. Ал корпустың базасындағы мәдени семантикасы бар тізімге енбей қалған басқа сөздерге алдағы уақытта осы ақпарат

бойынша репрезенттелу жолдары қарастырылып, енгізіледі. Сонымен, ұсынылып отырған Қазақ тілінің ұлттық клорпусы атты мегажобаның қарапайым бастапқы нұсқасына неғұрлым жеңіл әрі қарапайым белгіленім салынып, мейлінше аз (10 млн.) мәтін көлемі қамтылды. Алдағы уақытта бұл ақпараттар кеңдігі, тереңдігі жағынан да толықтырылып, өңделіп, белгіленім сапасы да артады, соған қарай Қазақ тілінің ұлттық корпусының барлық параметрлері мен корпусшалары (подкорпус) ресурстары түгенделіп, жасақталады.