Глоссарий

Ағылшын тілінің халықаралық корпусы (International Corpus of English – ICE, http://www.ucl.ac.uk/englishusage/projects/ice.htm) (20 мың сөз) – қолданысын көрсететін ұлттық субкорпус жиынтығы ағылшынның түрлі нұсқаларын (Австралия, Ұлыбритания, Гонконг, Үндістан, Ирландия, Канада, Кения, Малайзия, Жаңа Зеландия, Сингапур, АҚШ, Танзания, Филиппин, Шри-Ланка, Оңтүстік Африка, Ямайка) қаласында. Әрбір субкорпус 1 миллион сөзден тұратын жазбаша және ауызша мәтіндерден тұрады. Корпустың британдық компоненті (ICE-GB) толығымен дайындалған, оның мәтіндері морфологиялық және синтаксистік белгілермен жабдықталған. Http://www.ucl.ac.uk/english-usage/projects/ice.htm веб-сайты іс бойынша 20 000-нан астам сөздікке қол жеткізуге мүмкіндік береді.

Ақпараттық сұраныс – белгіленген ақпараттық қажеттілігінің сөздік көрінісі. Сұраныстар формалды мен пәндік мазмұнымен талданады және корпуспен жұмыс істейтін қолданбалы бағдарламалардың сұраныстарымен сөздіктегі терминдермен сипатталады.

Аралас корпус – белгілі бір уақыт кезеңі ішінде тілдің болуын білдіретін ұлттық корпустар (NCRC, BNC және т.б.)

Ауызекі корпус – корпус үлгілерін қамтитын қолдану сөйлеу тілі, практикалық транскрипция ұсынылған және жазылуы көмегімен ауызекі корпусының үлгісі Санкт-Петербургте корпусында көрсетілген .

Анафорикалық белгі – референттік, яғни есімдіктік байланыстарды жазып алатын белгі

Браун корпусы (The Brown Corpus, толық атауы – The Brown Standard Corpus of American English, корпус автор- 9 лары: W. Francis және H. Kucera, сайт адресі: http://icame.uib. no/brown/bcm.html) (1 млн сөз) – АҚШ-та 1963 жылы Браун университетінде (Brown University) құрастырылған алғашқы компьютерленген корпус. Б.к. құрастырудың мақсаты – жазба ағылшын тілінің әртүрлі жанрын және жанрларды салыстыруда жүйелі зерттеуді қамтамасыз ету. Б.к. статистикалық процедуралар негізінде құрастырылды; статистика корпус авторларының кәсіби интуициясымен сәйкестікте қолданылды. Б.к. 60 жылдардағы АҚШ баспа сөйленісін бейнелеу мақсатында жасалды, ол 1961 жылы АҚШ-та алғаш жарияланған американдық кітаптар, газет, журналдардан алынған 500 мәтінді қамтиды. Б.к. әрбір мәтін 2000 сөзді құрайды (сөзқолданыстар – tokens). Б.корпусындағы көптеген деректер саны алғашқы статистикалық өңдеуден өткен: жиілік және алфавиттік-жиілік сөздік, әртүрлі статистикалық үлестірулер.

Британ Ұлттық корпусы (British National Corpus – BNC, сайт адресі: http://corpus.byu.edu/bnc; http://www.natcorp.ox. ac.uk/) (100 млн сөз) – XX ғасыр аяғындағы британ ағылшын тілінің ауызекі және жазба үлгілерін қамтитын біртілді, синхронды мәтін корпусы. Бұл корпус әртүрлі жанрларда ұсынылған сол уақыттағы ауызекі және жазба британ ағылшын тілінің жалпы үлгісі болып табылады. BNC сөз таптарының белгіленімдерін, толық лингвистикалық аннотация мен контекстік ақпаратты қамтиды.

Жабық корпус – тар спецификалық мақсатқа арналып жасалған корпус, сонымен қатар ол жалпы қауымға, кең қолданысқа арналмайды.

Жазбаша корпус – ауызша сөйлеу ұсынылмаған корпус (Браундық корпус, LOB). З Зерттеу корпусы – бұл корпус тілдің қызмет етуінің әртүрлі аспектілерін зерттеу мақсатында жасалады. З.к., ереже бойынша, лингвистикалық міндеттердің кең ауқымға бағытталған ондаған миллионнан жүздеген миллионға дейінгі сөзқолданысты қамтиды.

Иллюстративті корпус – бұл корпус ғылыми зерттеуді жүргізгеннен соң құрастырады. И.к. мақсаты – алынған нәтижелерді дәлелдеу және дәйектеу, яғни И.к. басқа да лингвис- 11 тикалық тәсілдер арқылы анықталған белгілі бір тілдік (сөйленістік, мәтіндік) фактілерді лингвистикалық мысалдарды белгілеп көрсету үшін қызмет етеді. И.к. әдеттегі мұндай мысалдары «Орыс тілінің дискурсивті сөздерінің жолкөрсеткішінде» ұсынылған. Мұнда оқырманға ұсынылған семантикалық интерпретацияны тексеруге мүмкіндік беретін маңызды мәтіндік материалдардың белгіленген мағыналарымен және бөлшектердің семантикалық анализі беріледі.

Коммерциялық корпус (on-line немесе компакт-дискідегі көшірме) – сатып алу арқылы пайдалануға болатын корпус. Корпустың алдын ала аннотациясымен танысуға болады немесе бір реттік режимде корпуспен жұмыс істеуге болады, бірақ, ереже бойынша, барлық мәтіндерге толыққанды пайдаланбай, тек қана кішкене көлемдегі бөлігімен қолдануға болады.

Корпус – белгілі бір тілдегі электронды формадағы мәтіндерді жинақтауға бағытталған ақпараттық-анықтамалық жүйе. Корпус қазіргі түсінік бойынша Корпус – ол деректердің компьютерлік базасы. Корпустың негізгі бірлігі болып сөзқолданыстар (сөздер), негіздер (түбір, лемма) және сөйлемдер танылады. Корпусты жасау үшін компьютерлік бағдарламалар мен арнайы процедуралар қолданылады. Корпус тілдік бірліктер мен тілдік деректердің статистикасы мен әртүрлі анықтамаларын алу үшін маңызды болып табылады. Корпустың әртүрлі анықтамалары бар: 1) мәтін болып қалыптасқан жағдаят туралы ақпаратты көрсететін мәтіндерді репрезентативті жинақтау (жеткізуші, автор, адресат немесе аудитория туралы ақпарат) [Э. Финеган]; 2) Статистикалық талдау мен болжамды дәйектеуге, не бір саладағы тілдік ере- 12 желерді негіздеуге немесе бір жағдайдың кездесушілігін тексеру үшін қолданылатын көлемді және құрылымдалған мәтіндер жиынтығы (көп жағдайда элетронды түрде) [Википедия 62]; 3) тіл моделі сипатын қолдану үшін нақты тілдік белгілердің сәйкестілігімен таңдалған тілдің бір фрагментін жинақтау [Т. МакЭнери и Э. Вилсон]; негізінде логикалық ой, логикалық идеясы бар, сонымен қатар бұл мәтіндерді біріктіретін және корпустағы мәтіндерді ұйымдастыру ережесін жүзеге асыратын, корпус мәтіндерін талдау бағдарламалары мен алгоритмдерінің осы идеология мен әдіснамамен байланыстыратын кейбір мәтіндер жинағы [В.В. Рыков]. Корпус жасаудың негізіне зерттеуші филологтардың қағаз күйіндегі дәстүрлі карточкалардың құрастыру тәжірибесі жатыр. XX ғасырда компьютерлік технологиялардың кең қолданылуы картотеканың компьютерлік және жалпыға қолжетімді сипатқа енді. Корпустық амал-тәсілдің дамуына Интернеттің берері мол, себебі ғаламтор арқылы кең ауқымды мәтіндер деректеріне қолжеткізуге мүмкіндік берді

Корпустар классификациясы – әртүрлі негіздеріне байланысты корпустарды кластарға (топтарға) жүйелеу. Корпустарды кластарға бөлудің екі негізгі тәсілі анықталған: 1) барлық тілдерге қатысты корпустарды қарама-қарсы қою (немесе тілдің бір кезеңіне қою), сондай-ақ, тілдің ішкі құрылымына қатысты корпустарға қатысты (жанр, стиль, белгілі бір жас ерекшелік не әлеуметтік тортар тілі, жазушы не ғалымның тілі және т.б.); 2) Лингвистикалық белгіленім типіне қарай корпустарды бөлу: морфологиялық немесе синтаксистік тип корпустары (treebanks, яғни «синтаксистік құрылым банкі»). Синтаксистік белгіленімді қамтитын корпус лексикалық бірліктің морфологиялық сипатын да қарастырады. Корпустар өз белгілері бойынша әртүрлі критерийлермен ерекшеленеді: а) тілдік деректердің типі бойынша жазба, ауызша және аралас; «параллельдік» қасиеті бойынша біртілдік, екітілдік және көптілділік; «әдебилік» бойынша әдебиеттік, диалекті- 13 лік және ауызекі сөйлеу; мақсаты бойынша көпмақсаттылық және арнайы мамандырылған; жанры бойынша әдеби, фольклорлық, драматургиялық, публицистикалық; қолжетімділік бойынша еркін қолжеткізушілік, коммерциялық, жабық; белгіленуі бойынша зерттеуге, иллюстративтік; динамикалық қасиеті бойынша динамикалық (мониторлық) және статистикалық; белгіленімнің бар болуы бойынша белгіленген және белгіленбеген; белгіленім сипаты бойынша морфологиялық, синтаксистік және семантикалық, просодикалық және т.б.; мәтін көлемі бойынша толықмәтінді, «мәтін фрагменті» және т.

Корпустың қолжетімділігі – корпус тұтынушылары үшін ең маңызды өлшемі: еркін қолданыстағы корпустар кез келген уақытта on-line режимінде корпустың барлық мәтіндеріне кең ауқымда қолдануға мүмкіндік береді. Еркін қолданыс корпус мәліметтерінің біршама бөлігіне де қолдануға мүмкіндік туғызады. Корпустардың Таңбасы (tagging, annotation) – ол өте қиын операция, себебі ол тілдік бірліктер әр түрлі сипаттамаларымен атайды: морфологиялық, синтаксистік, лексикалық, просодикалық, анафорикалық. Егер анафорикалық, просодикалық белгілеулер автоматты жүйелерді жасау өте қиын және негізгі бөлігі жұмыс қолмен жүргізілсе, онда морфологиялық және синтаксистік талдау әртүрлі бағдарламалық құралдары деп атауға болатын парсеры болып табылады (parsers). Корпус мәтіндерінің мета сипаттамасы – корпус мәтіндерінің мета сипаттамасына мыналар кіреді: 1) мағыналы деректер элементтері – мәтіннің жанры мен стилінің ерекшеліктері, автор туралы мәліметтер, 2) ресми – файл атауы, кодтау параметрлері, тілдік нұсқасы, жұмыс кезеңдерінің орындаушылары. Бұл деректер әдетте қолмен енгізіледі.

Корпус лингвистикасындағы стандарттау – корпустарды белгілеу мен лингвистикалық қолдауды біріктіру; Деректер ұсыну форматтарын біріктіру және олардың құрылымы. Мақсаты – таңбалау түрлерінің үйлесімділігіне қол жеткізу (кодтау стандарты); түрлі органдардың салыстырмалылығы (стандарттау). Корпустардың мазмұны – 1) деректердің әдеттегі болуын қамтамасыз ететін және лингвистикалық құбылыстардың бүкіл спектрін ұсынудың толықтығын қамтамасыз ететін жеткілікті үлкен (өкілді) орган көлемі; 2) әр түрлі мәліметтер органда олардың жан-жақты контекстік нысандағы болып табылады, бұл оларды жан-жақты және объективті зерттеу мүмкіндігін тудырады; 3) дайындалған және дайындаған деректер массивін әртүрлі зерттеушілер бірнеше мақсатта қолдануға болады. Корпустық лингвистика – компьютерлік технологияларды қолданудағы лингвистикалық

корпустарды пайдалану және құрастырудың жалпы қағидаларын өңдеумен айналысатын компьютерлік лингвистиканың бағыты (XX ғ.).

Корпустық лингвистика электронды корпустар көмегімен «шынайы өмірдегі» тілді қолдануды меңгеруге бағытталған. К.л. пәні – лингвистикалық зерттеулер үшін және жалпы қауымдағы тұтынушылар мүддесіндегі оқу әдістеріне арналған теориялық және практикалық механизмдерінің негізін жасау және тілдік мәліметтерінің толыққанды ауқымын қолдану.

Корпустық менеджер (ағыл. corpus manager) – деректерді іздестіруге арналған бағдарламалық қамтамасыз етуді, статистикалық ақпаратты алуды және пайдаланушыны ыңғайлы пішінмен қамтамасыз етуді қамтитын мамандандыр- 15 рылған іздеу жүйесі. Корпустық менеджер – мәтіндік және лингвистикалық деректерді басқару жүйесі. Корпустық менеджердің міндеттері әртүрлі: келісім тізімдерін құру; жеке сөздерді, сөз тіркестерін іздеу; үлгілер арқылы іздеу (күрделі сұраулар); тізімдерді бірнеше критерий бойынша сұрыптау; шексіз контекстте сөз формаларын анықтау; статистикалық ақпаратты ұсыну; леммаларды бейнелеу, сөз формалары мен метадеректердің морфологиялық сипаттамалары (библиографиялық, типологиялық); нәтижелерді сақтау және басып шығару; бөлек файлдар ретінде жұмыс істеу және т.б. Ең әйгілі әмбебап менеджерлер SARA, XAIRA (BNC), Manatee / Bonito, CQP, DDC болып табылады

Корпустардың түрлері – корпустар әр түрлі негіздер бойынша жіктеледі , түрлі түрлері бар: сөйлеу толықтығы туралы әмбебап және т.б. 1) әмбебап корпус сөйлеу әрекетінің алуан түрлілігін көрсетеді; 2) жеке тұрғын үй қоғамның сөйлеу практикасында лингвистикалық немесе мәдени құбылыстың болуын көрсетеді; Корпусты құру мақсатында көп мақсатты және мамандандырылған деп бөледі; Қолжетімділік критерийіне сәйкес қолда бар корпустар жабық, коммерциялық корпустар деп бөлінеді; зерттеу және иллюстрациялық корпустарды тағайындау; динамизм критерийі бойынша корпус динамикалық және статикалық болып бөлінеді; корпус белгіленген және белгіленбеген деп бөлінеді.

Көпфункционалды корпус – әртүрлі жанрлардың (ұлттық корпустардың) мәтіндерін қамтитын корпус.

Қазіргі заманғы американдық ағылшын корпусы (Corpus of Contemporary American English – COCA, http://corpus.byu.edu/coca/) (520 млн сөз) – корпустық лингвистика профессоры Марк Дэвистің көмегімен 2008 ж. Brigham Young University (АҚШ) құрылған американдық ағылшын корпусы – ірі, еркін қолжетімді болып табылады. Корпус 16 1990 жылдан бастап бүгінгі күнге дейін жылына 20 миллион сөзбен толтырылды. COCA – аралас типтегі корпус, онда 5 жанрдың жазбаша мәтіндері бар: көркем әдебиет, танымал журналдар мен газеттерден алынған медиа-мәтіндер, ғылыми әдебиет, ауызша сөйленіс. Іздеу интерфейсі көптеген мүмкіндіктерді ұсынады: сөздерді, сөз тіркестерін, леммаларды, грамматикалық пішіндерді, синонимдік сериялады және т.б. іздеу. Корпус жылына екі рет жаңарып отырады және тілдік өзгерістер динамикасын бақылауға ыңғайлы.

Қазақ тілінің Алматы корпусы (ҚТАК, аталмыш корпус әл-Фараби атындағы Қазақ ұлттық университетінің филология және әлем тілдері факультетінің жалпы тіл білімі және

еуропа тілдері кафедрасы оқытушыларымен, Мәдиева Г.Б. басшылығымен және ЭЖМ НИУ (Мәскеу) филология қызметкерлерінің қатысуымен құрастырылуда), сайт адресі: http://web-corpora.net/KazakhCorpus/s) (40 млн сөзқолданысы бар) – Қазақстан Республикасының мемлекеттік тілі қазақ әдеби тіліндегі кең ауқымда белгіленген мәтіндер қоры негізінде анықтағыштық-ақпараттық жүйе ретіндегі қазақ тіліндегі Ұлттық корпустың бір нұсқасы болып табылуы мүмкін. ҚТАК үшін Шығыс армян ұлттық корпусының (EANC) іздеу жүйесіне бейімделген. Корпус үздіксіз сандық, сапалық жағынан жаңартылып, корпустың іздеу қызметі едәуір жоғарылауда. Корпус мәтіндері автоматтандырылған морфологиялық анализатор көмегімен белгіленген, корпустың 86% сөзформалары грамматикалық талдауға түскен. Корпуста омонимия алынбаған, яғни әрбір сөзформаның контекстке байланыссыз барлық мүмкін болатын талдаулары берілген. ҚТАК көркем әдебиет, БАҚ, ғылыми әдебиеттер мәтіндерін құрайды.

Қалыпты корпус – кішігірім уақыт интервалдарының мәтіндері бар орган, мысалы, авторлық корпустар – жазушылардың мәтіндерін жинақтау болып табылады. Қалыпты корпус мысалы, белгілі бір хронологиялық кезеңде лингвистикалық құбылыстардың жұмыс істеуін анықтау: сөздердің 17 мағынасын өзгерту, белгілі синтаксистік конструкцияларды пайдалану жиілігі және т.б. сияқты тілдік жүйенің белгілі бір уақыттық күйін көрсетеді. Құрылымдық таңбалау – мәтіннің құрылымдық белгілеуі болып табылады, оның барысында параграфтар, сөйлемдер, сөздер таңдалады және әдетте автоматты түрде орыналады.

Лексиканың автоматты классификациясы – мәтінді түсінудің автоматты процедураларының негізгі кілттерінің бірі. Лексиканың автоматты классификациясы мәтін құрылымының формалануы және мәтін элементтерінің арасындағы семантикалық байланыстың сандық бағалануы аясында жүзеге асады (леммалар мен сөзформалармен ұсынылған сөздер).

Лемма – бастапқы, сөздік форма. Сөздікке негізделген сөздердің негізгі түрі. Мысалы: орыс тілінде: форма именительного падежа, ал қазақ тілінде – атауыш септік. 18 Лемматизатор – леммалдандыруды жүзеге асыратын минималды ресурстар мен сыртқы тәуелділіктерді талап ететін ықшам әрі қарапайым модуль.

Леммалдандыру – бұл басқа сөздік формаларынан сөздің өзіндік формасын қалыптастыру үдерісі. Леммалдандыру сөйлеу бөліктерінің сәйкестендірілуімен байланысты және корпустағы тиісті сөздердің қысқартылуын қамтиды. Леммалдандыру барлық ықтимал нұсқаларын енгізудің қажеті жоқ жеке белгісін оқшаулауға және оқуға мүмкіндік береді. Мысалы, ағылшын тіліндегі «Серуен» етістігі келесі формалармен ұсынылады: «серуендеу», «серуендеген», «серуендейді». Леммалдандыру – бұл талдау кезінде бір сөз ретінде қарастырылатын, сөздің әртүрлі формаларын топтастыру үдерісі. Мысалы, сөйлемде: [The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs] келесі леммалар көрсетіледі: [the] [quick] [brown] [fox] [jump] [over] [the] [lazy] [dog].

Лингвистикалық белгіленім (лингвистическая разметка) – лингвистикалық бірлік процесі негізіндегі лингвистикалық сипаттамалар атрибуциясы. Лингвистикалық белгілердің келесі түрлері бөлінеді: морфологиялық, синтаксистік, семантикалық, анафориялық, просодикалық, дискурстық және т.б. Олардың барлығы келесі қағидаттарға

сәйкес жүзеге асырылады: 1) орналасу схемасының сипаттамасы (негіздемесі); 2) лингвистикалық тұжырымдамалардың жалпы қабылданған жүйесі; 3) пайдаланушының таныс талдау схемасы; 4) параметрлерді енгізуге уәждеме; 5) теориялық бейтарап (дәстүрлі) орналасу схемасы; 6) халықаралық стандарттарға сәйкестік.

Мамандандырылған корпус – бір жанр немесе жанрлар тобының мәтіндерін қамтитын корпус. Жанр көрнекті әдеби, фольклор, драмалық, журналистік, және т.б. байланысты. Мысалы, журналистік корпусы үшін саяси метафораларды, XX ғасырдың аяғында орыс газеттері ішінде компьютерлік корпусын жатқызуға болады. (Http://www.philol.msu.ru/~lex/ corpus/). М.К. кез келген кәсіби секторының терминологиялық аппараты болуы мүмкін. Контекстте мерзімге пайдалану болып табылады, өйткені ол ұлттық корпустан кем емес болып табылады. **Мәтіндердің арнайы корпусы** – нақты ғылыми-зерттеу тапсырмалары үшін мақсаттар мөлшерін, әдетте шағын теңестірілген корпусы болып табылады. Ресей оқушыларының ағылшын мәтіндерінің ерекшеліктерін зерттеу - басты мақсаты. Түпнұсқа мәтін материалы Санкт-Петербургтегі мектептерде (9-11 сынып 78 оқушы) 2007 жылы қараша – жел- 20 тоқсан айларында жиналды. Ағылшын деңгейі: аралық / орта (26%) орта есеппен / Upper-Intermediate (74%) жоғары. Нәтижесінде, сөйлеу тілдері жұмыстарының «құрылымдық кедейлік» қалыптасқаны анықталды, студенттер стандартты ағылшын тілінің дамыған және табиғи үлгілері емес, ең қарапайым құрылымдарын жүйелі түрде көреді. «Санкт-Петербург мектебінің ағылшын мәтіндерін», «өтпелі грамматика» жағдайлары (екінші тіл дамыту барысында аралық тілі, ана тілі мен екінші арасындағы өтпелі, аралық жүйесі, оларға әлі жеткілікті иелігін жеткен жоқ) қалыптастыру басталды. Осылайша, стандартты ағылшын негізгі грамматикалық құрылымдармен сәйкес келмейді шет тілі (EFL). Ол «жаһандық ағылшын» дамушы ережелері негізінен «тасқа айналған» interyazyka үлгісіне негізделген деп күтілуде.

Метадеректер – мәтіндер мен олардың құрамдас бөліктеріне жататын қосымша ақпарат болып табылады. Метадеректердің үш түрі бар: жалпы мәтінге қатысты экстравингвистикалық; мәтін құрылымы туралы мәліметтер; мәтіннің элементтерін сипаттайтын лингвистикалық метадеректер.

Морфологиялық белгіленім (part-of-speech tagging (POS-tagging) – бұл сөйлеу бөлігінің белгісі ғана емес, сонымен қатар сөйлеу бөлігіне тән грамматикалық санаттардың белгілерін қамтитын морфологиялық белгі, лингвистикалық таңбаның негізгі түрі. Көптеген ірі корпустар морфологиялық жағынан белгіленеді.

Морфологиялық талдау – синтаксистік және семантикалық талдаудың негізі.

Мультимедиялық корпус – сөйлеу жағдайында «дыбысталған» сөздерді зерттеуге арналған электронды ресурс. Жазу мәтіндерінен басқа осы түрдегі корпусқа мәтінге сілтеме жасай отырып, байланыс процесінің бейне және аудио жазбаларын кіргізуге болады. Мәтіндер тек қана лингвистикалық бірліктерді ғана емес, сонымен қатар әр түрлі коммуникациялық жағдайларда сөйлеушінің сөйлеу әрекеттерін, оның мінез-құлқын зерттеуге мүмкіндік беретін транскрипт- 21 термен теңестіріледі. Мультимедиялық корпус табиғи диалогтың ауызша және ауызша емес компоненттерінің өзара әрекеттесуін зерттеу тұрғысынан перспективалы болып табылады.

Орыс тілінің ұлттық корпусы (ОТҰК, сайттың мекенжайы: http://ruscorpora.ru/) (сөз қолданысы 176 млн) – элек- 22 трондық нысанда жиналған орыс мәтіндеріне негізделген, репрезентативтік ақпараттық-анықтама жүйесі. ОТҰК алғаш рет 2004ж сәуір айында орналастырылған. Корпус орыс тіліне байланысты әр түрлі қызықтыратын мәселелермен барлық жандарға арналған. ОТҰК жанрлық алуан мәтіндер (ғылыми, ресми-іскерлік, публицистикалық, шіркеу-құдайға сенетін, көркем, ауызша-тұрмыстық саланы қоса алғанда, ауызша және электронды коммуникацияны) ерекшеленеді; өте алуан түрлі негізгі социологиялық өлшемдер (жасы, білім деңгейі және меңгеру, кәсіби керек-жарақтары, түрлері сөйлеу дақылдар) корпусқа шығармалары кірген авторлардың саны (кемінде 20 мың). Корпус қазіргі заманғы мәтіндерді құрайды , құру кезеңінде олардың 1951-2007 жылдардағы мәтіндерді қамтиды (сөз қолданысы 97,5 млн). Диахрондық бөлігі XVIII ғасырдың мәтіндерін біріктіреді (сөз қолданысы 1,1 млн) XIX ғасырдың (прозалық мәтіннің сөз қолданысы 23,3 млн Және поэтикалық мәтіндердің сөзқолданысы 2,5 млн), XX ғасырдың 1-ші жартысынан (сөз қолданысы 25,4 млн).

Параллельді корпус – параллельді белгісі бойынша сипатталатын корпус: біртілді, қостілді немесе көптілді. Біртілді корпус қарама h-қайшылығы тараған нұсқалардың тілі бойынша ажыратылады. Мысалы, ағылшын тілі туған тіл ретінде, ағылшын тілі шет тілі ретінде. Екі тілді және көптілді корпустары: а) екі немесе бірнеше тілдерде жазылғанына қарамастан бір тақырыптық облыстағы мәтіндерді (мысалы, әр түрлі елдерде және әр түрлі тілдерде өткен корпус материалдары бойынша ғылыми конференцияларды) біріктіреді. Мұндай корпустар терминологиямен жұмыс істеуге көмектеседі және аудармашылар жиі пайдаланады; б) көптеген мәтіндер-түпнұсқаларын жазылған қандай да бір бастапқы тілінде, және мәтіндер-аударымдарын бастапқы мәтіндер бір 23 немесе бірнеше басқа тілдерде болуы. Баға жетпес нақты материал жүргізу үшін салыстырмалы-салғастырмалы зерттеулер үшін зерттеу аударма теориясы және оқыту үшін аудару, адам және компьютерлер үшін корпус ұсынылады.

Парсер (parser) – бағдарламалық құралы автоматты синтаксистік анализатор. Оның құрылуы компьютерлік лингвистика облысы үшін маңызды үлкен корпустарының бірі болып саналады. **Парсинг** – 1) терминге лексем (сөз, токен) тілінің салыстыру процесі, оның формальды тілінің грамматикасы, әдетте, ағаш тәуелділіктері нәтижесі болып табылады (синтактикалық ағаш); 2) автоматтандырылған үдеріс көшіру материалдарды бір сайт (немесе бірнеше), басқа сайтқа (немесе деректер базасына, кейіннен құю жеке сайттар немесе сату).

Просодикалық таңба – тэгтер, түбір білдіруші және екпінді пайдаланатын таңбалау түрі. Корпустарда, П.т-лық ауызша сөйлеуде қайталаңба белгілеу үшін арналған дискурстық таңба, үзіліс, ескертпелер және т.б. сөйлеулер жиі сүйемелденеді.

Робот – кіші жүйешені (бағдарламалық кешен) қамтамасыз ететін қарап шығу, Интернет және инвертировандық файлдарды қолдау (индекстік деректер базасын) актуалды жай-күйде сақтау. Робот – негізгі ақпаратты құралдарды жинау мен ақпараттық ресурстар желісінің жай күйін анықтайды.

Семантикалық таңбалау – семантикалық тегтерді қолданумен берілген сөз немесе сөз тіркесіне жататын семантикалық санаттар. С.т. Корпус сөздердің мәнін, үнділік пен

синонимияның шешілуін, сөздерді (категорияларды) санаттауды, тақырыптық кластарды таңдауды, пайда болу белгілерін, бағалау мен деривациялық сипаттамаларын және т.б. сипатта.

Синтаксистік түзету – морфологиялық талдау деректері негізінде жасалған талдаудың нәтижесі. Бұл таңбаның түрі лексикалық бірліктер мен әр түрлі синтаксистік құрылымдар арасындағы синтаксистік байланыстарды сипаттайды (мысалы, бағынатын тармақ, етістік, сөз тіркесі және т.б.) Стем Сөз – сөздің негізі, сөздің өзгермейтін бөлігі.

Стеммер – мөртабанның нақты сатысы. **Стеммер** аффикс арқылы сөз түрлендіруін жүзеге асыратын тілдермен жұмыс істей алады. Мұндай тілдердің мысалдары – орыс және ағылшын тілдері болып табылады. Классикалық Стеммер (мысалы, Портер stemmer) негізгі кемшіліктерді бірі, олар жиі ұқсас синтаксис сөздер, бірақ мүлдем басқа құндылықтармен ажыратылады. **Стеммер** контексті білместен бір сөзді қабылдайды, сөйлеудің әртүрлі бөліктеріне сілтеме жасайтын және әртүрлі мағыналарға ие сөздерді ажыратпайды. **Стеммер** деректерді өңдеу үшін әдетте іске асыруы жылдамырақ болады. Көптеген қосымшалар үшін олардың жұмысының төменгі дәлдігі шешуші мәнге ие болмауы мүмкін. Мысалы, «жақсы» белгісі леммаға «жақсы» сәйкес келеді, бірақ бұл мөртабандарда жоқ. Лемма «серуен» белгісі «жаяу» базалық нысаны болып табылады, және осы морфологиялық және лемматизациясы болып табылды.

Стемминг – сөздің негізін табу процесін, қалған бөлігі сөздің барлық грамматикалық формалары бірдей болады. Мысалы, «әдемі» сөзі үшін негіз «әдемі» болады, бірақ «сұлулық» сөзінің тамыры болады. Стемминг – лемматизациядан ерекшеленеді.

Тарихи корпус – репрезентациялау мен стандарттау тұрғысынан жасауда ерекше қиындық тудыратын белгілі бір тарихи кезеңдегі мәтіндер негізінде жасалады. Тарихи корпусқа Санкт-Петербургтің XVI-XVII ғ. Мәтіндердің агиографиялық корпусы (СКАТ, СПбМУ филология факультетінің 25 мастематикалық лингвистика кафедрасында құрастырылды, корпус көлемі 2006 ж. 500 мың сөзқолданысын қамтыды. Сайт адресі: http://project.phil.pu.ru/skat). СКАТ – көне орыс агиографиялық әдебиеттің еркерткіштері бойынша мәтіндердің электронды корпусы. Тілдің ұлттық корпусы (ТҰК) – бұл белгілі бір тілдің белгілі бір кезеңінде (кезеңдерінде), оның өмір сүруінің барлық сан алуан жанрлар, стильдер, аумақтық және әлеуметтік нұсқаларының корпусы. Мысалы, Орыс тілі ұлттық корпусы, Британ ұлттық корпусы, Алматы қазақ тілі корпусы. ТҰК – Корпустық лингвистиканың мамандандырылған лингвистерімен құрылады. Токен (ағыл. tokens) – бұл сөзқолданыс, лексемаға сәйкес келетін негізгі бірлік.

Токенизация – табиғи тілдің бөлек маңызды бірлікке бөлу (белгіше, сөздік формалар). Токенизация – табиғи тілді әрі қарай өңдеудің қажетті шарты. Егер тілдер мінсіз тыныс белгілеріне ие болса, токенизация қиын болмас еді – тіпті қарапайым бағдарлама мәтінді сөздерге, кеңістіктерге және тыныс белгілеріне қарай бөлуі мүмкін. Шындығында, тілдерде Токенизацияның тапсырмасын күрделендіретін пунктуация жоқ, сондықтан ағылшын тілінде бірден-бір таңбаланбайтын жағдайлар бар. Мысалы, ол сөйлемнің соңында орналасқан сөздің қысқартылған формасы немесе сол сөз деген болуы мүмкін; Мұндай қиындықтар шектеулі, мәтінді өңдейтін көптеген қосымшалар оларды жиі

елемейді (мысалы, қысқартулар мен күрделі сөздерді есепке алмайды) немесе оларды бөлек алгоритм арқылы өңдеген жөн.

Тэг (ағыл. tag – затбелгі) – аталған жапсырма, дескриптор, гипермәтінді белгілеу тілінің элементі; Лингвистикалық белгілеу үдерісіндегі сөздерге арналған код. Әрбір код осы сөзді сипаттайтын грамматикалық белгілердің белгілі бір жиынына сәйкес келеді. 26 Тэггер (tagger) – мәтіннің автоматты морфологиялық талдауын жасайтын және оны белгілейтін бағдарламалық құрал, әр сөзге грамматикалық тегті немесе грамматикалық тегтер жиынтығын тағайындайды. POS-тегтеуді қолдану ең көп таралған. Бейтаныс мәтіндерді белгілеу үшін, басым бөлігі алдымен қолмен белгіленген корпустарда үйрену керек. Автоматты түрде белгілеу оқу үрдісінде алынған модель негізінде жүргізіледі. Егер бір пішінге бірнеше талдау жүргізілсе, олардың біреуі солға немесе дұрыс контекстке қарай таңдалады. Теггерлердің кейбірі барлық ықтимал талдауды көруге мүмкіндік береді.

Файл индексі (индекс) – сұраныс бойынша мәліметтерді жылдам іздеуге негізделген өзара байланыстағы файлдар. Индекс негізін инвертирленген файлдар қамтиды. Іздеу ауқымының ұйымдастыруының инвертирленген сызбасы мазмұн идентификаторы арқылы құжаттарға қолжетімділікті қамтамасыз ету принципіне негізделеді. Мұндай сызбаны арнайы қосымша инвертирленген файлдарды, қолжетімділік нүктені, құрастыру мақсатында ауқымды құжаттарды дәйектелген түрде өңдеу жолымен алады

Іздеу деректер базасы (ағыл. index database) – проиндексталған веб-құжаттарды қамтитын алуан түрлі туралы ақпаратты осы бірліктерде арнайы түрде ұйымдастырылған мәліметтер құрылымынан алынған, инвертировандық файлдан тұратын лексикалық бірліктерді қамтиді, алынған проиндекстік веб-құжаттарды қамтитын алуан түрлі туралы ақпаратты осы бірліктерде (мысалы, олардың ұстанымының құжаттары) туралы өздерінің құжаттарында және сайттарында жарияланады.

Іздеу жүйесі – өңдеу сұрау, пайдаланушының деректер базасында іздеу және беруді , іздеу нәтижелерін пайдаланушыға қамтамасыз ететін іздеу жүйесі. І.Ж. пайдаланушылық интерфейстер арқылы қарым-қатынас жасайды – экрандық нысандары бағдарламалар-браузерлер: интерфейс қалыптастыру сұрау салулары және интерфейс көру іздеу нәтижелері. І.Ж көптеген түрлері бар, олар тілін сұраныстары, дизайн, сервис және т.б түрлерімен ерекшеленеді. Негізгі іздеу жүйелеріне вербалды түрі (бірінші кезекте көлемі бойынша деректер базасын) Google, Fast Search (alltheweb іздеу жүйелерін иеленеді), AltaVista, WiseNut, HotBot, MSN Search, Teoma. Ресейлік жүйелердің басты болып: Яндекс (Яndex,Yandex), Рамблер (Rambler), Апорт (Aport) кіреді

Экстралингвистикалық таңба (метадеректер) – "сыртқы", "зияткерлік" белгілеулер (библиографиялық сипаттамалары, типологиялық сипаттамалары, тақырыптық сипаттамасы, социологиялық сипаттамалары), "формальды" құрылымдық белгілері (мәтін, бөлім, тарау, бөлім, абзац, сөйлем), техникалық-технологиялық таңбалау (кодтау, қайта өңдеу күнін, орындаушылар көзі, электрондық нұсқалары) тұрады. Э.т. өзара байланысын, тілі мен шарттарын және оны жұмыс істеуін; оқып-үйрену үшін жекелеген көпшілік тілін анықтау үшін қажет.

Эмпирикалық қолдау – пайдалану корпусының сапалық әдісі. Корпустар ақпаратты жиілік сөздер, фразалар мен конструкциялар өндірісімен, онымен пайдаланылуы мүмкін сандық зерттеулерді қамтамасыз етеді. Көптеген салаларда теориялық және компьютерлік лингвистика сандық зерттеулерді пайдаланады. Ұқсастықтары мен айырмашылықтары әр түрлі сөйлейтін топтар арасында немесе әр түрлі типті мәтіндерді қамтамасыз етуі туралы деректер жиілік үшін коррекциялаудың зерттеулері және т. б. көрсетеді.