Дәріс №3-4
Дәріс тақырыбы: Корпус лингвистикасының қалыптасуы
Сұрақтар:

1. Корпус лингвистикасының пайда болуы: картотекадан корпусқа
2. Лингвистикалық корпустарды құрудың тарихы
3. Қазақ тілі корпустарын құру тәжірибесі

1. КЛ зерттеулерге деген қызығушылық әрине компьютерлердің дамуымен байланысты. Компьютерлер өте үлкен көлемдегі материалдарды өңдеуге мүмкіндік берді. Дегенмен айта кету керек КЛ компьютерлер пайда болғанға дейін қалыптасты. Шын мәнінде, адамдар корпус лингвистикасы пайда болмай тұрып-ақ, 18 ғасырдан бастап оны құрды және зерттеді. Мысалдар: Круден/ Cruden және басқалардың Библияны зерттеулері, сөздіктерді құрастыру (Johnson, Oxford English Dictionary, Webster Dictionary /Джонсон, Оксфорд ағылшын сөздігі, Вебстер сөздігі), тілдерді оқыту (1921 жылғы Торндайктың жиілік корпусы/Thorndike'a), дескриптивті грамматика (Фрис, 1940, Квирк, 1968/// Fries, 1940, Quirk, 1968) осыған мысал.

Корпусқа негізделген зерттеулер әдетте 1960 жылдардың басынан электронды, машина оқитын корпустардан басталды деп айтылады. Дегенмен одан бұрын корпусқа негізделген лингвистикалық талдау болды, ол негізінен бес түрлі салада іске асты: 1) Библияны және әдебиетті зерттеу; 2) лексикография, 3) диалект зерттеулері 4) тілді оқыту зерттеулері 5) грамматикалық зерттеулер (Г.Кеннеди, 13-б).

1) Корпусқа негізделген зерттеулердің алғашқы маңызды бөліктерінің бірі Библияны қолданып жасалған лингв.лық зерттеулер. Александр Круден, Лондондық кітап сатушы, пруфридер, түрме реформаторы, 1701 жылы Абердинде дүниге келген, Библияның алғаш әйгілі нұсқасын *Authorised (King James) Version of Bible* жасаған. Алғаш 1736 жылы жарық көрген Круденнің Concordance еңбегі монументалды (маңызы зор) болды, 1879 жылға дейін 42 басылымы шыққан. Бұл Библиядағы конкорданстарды талдаған. (әрі қарай осы еңбектен G.Kennedy An Introduction to Corpus Linguistics)

Торндайк-Лорге тізімі

1944 жылы құрастырылған ағылшын тіліндегі сөз жиіліктерінің алғашқы және ықпалды тізімі. Кейінгі көптеген жаңартылған тізімдер болды, бірақ бұл термин әлі де кейде тілдегі сөз жиіліктерінің эмпирикалық тізімін білдіру үшін жалпы түрде пайдаланылады. [Эдвард Л. Торндайк және Ирвинг Д. Лорге (1905–1961), АҚШ психологтары]

Квирк корпусы (Survey of English Usage **сайтын көрсету** **https://www.ucl.ac.uk/english-usage/**) бір миллион сөзқолданысын қамтыды. Британдық ағылшын тілінің ауызша және жазбаша үлгілерін 1955 және 1985 жылдар аралығын қамтыған. Корпус әрқайсысы 5000 сөзден тұратын 200 мәтіннен тұрған. Ауызша мәтіндер диалогты да, монологты да қамтиды, ал жазбаша мәтіндер тек баспа және қолжазба материалдарын ғана емес, сонымен қатар эфирлік жаңалықтардағы және сөйлеу сценарийлеріндегі дауыстап оқылатын ағылшын тілінің мысалдарын қамтиды. Survey Corpus бастапқыда кәдімгі қағазда егжей-тегжейлі грамматикалық аннотациялары бар мыңдаған бланктар түрінде құрастырылған. Бұл енді компьютерлендірілді және әрбір лексикалық элемент сөз класы үшін автоматты түрде белгіленді.

«Бұл корпус электронды емес соңғы сөздік болды. Оны құру 25 жылды алған, ал 1989 жылы, ол аяқталған кезде, технологиялар әлдеқайда озып кетті» (А.Б.Кутузов). Корпусты жедел цифрландыру керек болды. Бұл корпус енді UCL Лондондағы Университет колледжінде қол жетімді.

КЛ тілдегі эмпирикалық деректермен байланысты қарастырылады. Кейбір орыс тілді дереккөздер лингвистика тарихының эмпирикалық ғылым ретінде басталуы КЛ қалыптасу тарихына да ықпал етті деп көрсетеді. Бұл тұрғыдан А.Захаров пен С.Богдановалар КЛ әдістану ретіндегі (ғалымдардың ойынша әдіснама) пайда болуын

лингвистиканың осы эмпирикалық ғылым ретіндегі тарихымен тығыз байланысты деп таниды.

Корпус лингвистикасында қазірде қолданылатын технологиялар компьютерлер пайда болғаннан бұрын да болды деген пікірді орыс зерттеушілері А.Захаров пен С.Богдановалар да құптайды. «КЛ қолданылатын технологиялар электронды компьютерлерден әлдеқайда бұрын пайда болған: олардың көпшілігінің түп тамыры XYIII соңы және XIX ғ. дәстүрлерінде жатыр, осы кезеңдерде лингвистика бірінші рет «реалды/шынайы» немесе эмпирикалық ғылым деп танылған болатын» дейді. Осылай дей келе олар КЛ негізі болып табылатын лингвистикалық зерттеулердің үшеуіне тоқталады. Олар:

1. Историческая лингвистика: изменения в языке и реконструкция (сравнительно-исторический метод).
2. Написание грамматик, лексикография и обучение языку.
3. Социолингвистика: языковое многообразие.

Осы үш салада қолданылған технологиялар қазіргі КЛ дамуына және керісінше де әсер етті деген ой білдіреді (қараңыздар: 11-14 бб.).

Компьютерлердің пайда болуы мәтінге тілдік сипаттамалар жасауды жеңілдету арқылы және талдауға қажетті деректердің көлемін көбейту арқылы КЛ қарқындап дамуына ерекше серпін берді.

2. Квирк корпусы (Survey of English Usage) бір миллион сөзқолданысын қамтыды. Бұл корпус электронды емес соңғы сөздік болды. Оны құру 25 жылды алған, ал 1989 жылы, ол аяқталған кезде, технологиялар әлдеқайда озып кетті. Корпусты жедел цифрландыру керек болды. Бұл корпус енді Лондондағы Университет колледжінде қол жетімді.

Электронды корпустардың бірінші буыны:

1. 1960 жылдар: Браун корпусы - The Brown Corpus (АҚШ) 1 млн сөз. Нақты атауы: Brown University Standard Corpus of Present-Day American English. 1961 жылдан 1964 жылға дейін құрастырылған. Корпустың тілі: американдық ағылшын тілі, жазбаша мәтіндер, 1 млн сөзқолданыстар (бұл сан корпустардың бірінші буыны үшін стандарт болған). Корпустың жасаушылары - Nelson Francis және Henry Kucera//Генри Кучера. Корпус әрқайсысы 2000 сөзден тұратын 500 мәтінен тұрады. Осы корпустың негізінде 1969 жылы Американдық мұра сөздігі/American Heritage Dictionary жасалды. https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
2. 1970 жылдар The Lancaster-Oslo/Bergen Corpus (LOB) - Lancaster, Oslo and Bergen зерттеушілері жасаған корпус. 1 млн сөзден тұратын Британдық ағылшын мәтіндерінен тұрған. Бұл 1970-78 жылдар аралығында жасалған корпус, Ланкастер, Осло университеттері мен Бергендегі ғылыми орталықтың жобасы. Корпустың тілі Британдық ағылшын тілі, 1 млн сөзқолданыс, құрылымы Браун корпусына ұқсас. Ғалымдар 1 млн сөзқолданыстың аз екеніне көз жеткізе бастаған. Сайт - http://khnt.hit.uib.no/icame/manuals/lobman/
3. Лондон-Лунд Корпус-London-Lund Corpus (LLC). 1975 жылы аяқталған ауызша ағылшын сөйлеу тілінің корпусы. Ол шамамен 500 мың сөз қолданыстарын қамтыған, орфографиялық транскрипциямен, фонетикалық және просодикалық белгіленімдермен қамтылған. Бұл жұмыс алдымен University College London қызметкерлерімен қағаз вариантында орындалып, кейіннен швед қаласы Лундтың лингвистерімен компьютерлік формаға түсірілген. Сайт проекта - http://korpus.uib.no/icame/manuals/LONDLUND/INDEX.HTM
   Жалғасы: London-Lund Corpus 2. **A corpus of spoken British English** https://projekt.ht.lu.se/llc2
4. 1980 жылдар Орыс тілінің машиналық қоры (Машинный Фонд русского языка). Бұл

корпустың жасалуы 1985 жылы СССР Ғылым академиясының Орыс тілі институтында басталған. 1991 жылы қаржыландыру тоқтап, жұмыс та тоқтаған. http://cfrl.ruslang.ru/ (сайт түсініксіз)

5. Орыс тілінің Уппсала корпусы (Швеция). 1980 жылдары Швецияның Уппсала унверситетінде славистика институтында қазіргі орыс тілі мәтіндерінің Уппсала корпусы жасалған. 1 млн сөзқолданыс, шамамен 600 мәтін. Сайт - http://www.slaviska.uu.se/korpus.htm (сайт ашылмады)

1990 жылдардан бастап корпустардың *екінші буыны* деп аталатын кезеңі басталады. Технология әлдеқайда дамыған, 100 млн. және одан асатын корпустарды жасауға мүмкіндік туды.

1. 1990 жылдар Британдық ұлттық корпус - British National Corpus, 100 млн. сөз. 90 пайызы жазбаша мәтіндер, 10 пайызы ауызша. Жасауға Британ үкіметін қоса алғанда көптеген ұйымдар қатысқан. 1995 жылы корпусты жасау аяқталған. Корпус 4124 мәтіннен тұрады, оның ішінде 863-і ауызша әңгімелесу немесе монологтардан транскрипцияланған. Әрбір мәтін орфографиялық сөйлемдерге сегменттелген, әрбір сөзге автоматты түрде сөздің класс коды тағайындалған (сөз табы). Тұтас корпуста 6,4 млн. орфографиялық сөйлем бар. Сайт - http://www.natcorp.ox.ac.uk
2. The Bank of English, Birmingham (Collins Cobuild), 600 млн. сөз **COBUILD, an acronym for Collins Birmingham University International Language Database**. Бұл 1980 жылы басталған. Collins баспасы жаңа сөздік жасап шығару үшін корпусты жасауды қолға алған. 1990 жылы Collins баспасы мен Бирмингем университеті бірігіп жұмыс жасауды бастап, The Bank of English бастамасын бастайды. Жоба жетекшісі - Джон Синклер. 1997 жылы шамамен 300 млн сөзқолданыс болса, 2005 жылы 525 млн.-ға жеткен. Әр ай сайын корпусқа 2,5 млн жаңа сөзқолданыс қосып отырады. Корпустың 25 пайызы ауызша сөйлеу тілі, 75 пайызы жазбаша. http://www.collins.co.uk/Corpus/CorpusSearch.aspx
3. 2000 жылдар American National Corpus, 100 млн. сөз деп жоспарланған, қазір 11 млн.. 2003 жылы бірінші басылымы шыққан. Ақылы негізде пайдалануға ғана болады. http://www.americannationalcorpus.org/
4. Corpus of Contemporary American English, (Корпус современного американского английского) 400 млн. сөз. https://www.english-corpora.org/coca/
5. Национальный корпус русского языка, 1,5 млрд. сөз. https://ruscorpora.ru/new/
6. Gigaword corpora: ағылшын, араб, қытай 50 тіл шамамен, 2 млрд. сөз. Еуропалық Одақ қаржыландырады. Linguistic Data Consortium компаниясы басқарады. Негізінен мәтіндер публистикадан, жаңалықтардан. http://www.ldc.upenn.edu

Корпус лингвистикасы әлем бойынша жеке бағыт ретінде 1990 жылдары қалыптасты. Корпус жасау ісін орыс тіл білімі кешіректеу бастады дегенмен жақсы қарқынмен дамытуда. 2004 жылы жасаған Орыс тілінің ұлттық корпусы http://ruscorpora.ru маңызды орын алады.

Английская лингвистическая служба Lexical Computing Ltd. (A. Kilgarriff) предоставляет на коммерческой основе доступ более чем к 40 корпусам различных языков.

3. Қазақ тілі корпустарын құру тәжірибесі. «Ұлттық корпус – қандай да бір елдің тілінде бар барлық жазбаша және ауызша дискурстарды (жарнамадан бастап көркем әдебиет мәтініне дейін) толық, теңбе-тең және шамалас көрсететін, тілдің нақты қолданылуы мен өзгеруі жайлы мәліметтердің (оның ішінде статистикалық) түбегейлі жаңа дереккөзі болып қызмет ететін көлемі жағынан ең үлкен корпус» (Сулейменова Э.Д. Қазақ тілі үшін Ұлттық корпус керек пе? // ҚазҰУ хабаршысы. Филология сериясы, No 1(131). 2011. – Б. 77https://philart.kaznu.kz/index.php/1-FIL/article/view/643/618)

Тілдердің ұлттық корпусы не үшін жасалады деген сұраққа да жауапты осы мақаладан табасыздар.

Алматы қазақ тілі корпусы. http://web-corpora.net/KazakhCorpus/search/?interface_language=kz Корпусты жасау әл-Фараби атындағы Қазақ ұлттық университетінің жалпы тіл білімі және шетел филологиясы кафедрасына тиесілі. Осы кафедраның күшімен 2012 жылдың мамыр айында басталған. Қазіргі таңда корпустың көлемі 40 миллионнан астам сөзқолданыстарынан тұрады. Корпус мәтіндері автоматты морфологиялық талдағыш көмегімен белгіленген, корпустағы 86% сөзформаларына грамматикалық талдау жасалынған. Корпуста омонимия алынған жоқ, яғни әрбір сөзформа талдауының барлық мүмкін деген нұсқалары беріліп, контексті ескермей тіркелген. Аталмыш корпусты жасауда Шығыс армян ұлттық корпусының (EANC) іздеу жүйесі бейімделген болатын.

Қазақ тілінің ұлттық корпусы. https://qazcorpus.kz/ 30 млн сөзқолданысты қамтиды, оның ішінде 14 млн сөзқолданыстан тұратын мәтінге метабелгіленім, яғни мәтіннің авторы, автордың жасы, мәтін тақырыбы, стилі, жанры т.б., енгізілген. Мәтіндер көркем әдебиет, ғылыми стиль, публицистикалық стиль, ресми және сөйлеу стилінен алынған. Әрине көлемі жағынан көркем әдебиет пен публицистикалық стиль мәтіндері көп (https://qazcorpus.kz/about/1/). Сөйлеу стилі газет журналдардағы сайттардағы сұхбаттар алынған, мұны айта кету керек нақты табиғи сөйлеудің көрінісі емес.

Kazakh language Text-to-Speech – 2 https://issai.nu.edu.kz/tts2-eng/

Сайттарында қазақша **Қазақ мәтінді сөзге түрлендіру – 2 (Kazakh TTS2)** деп аударған.

Зерттеулерді ынталандыру және цифрлық технологияларды қазақшалау мақсатында 2021 жылы Ақылды жүйелер мен жасанды интеллект институты “KazakhTTS” атты деректер жиынтығын әзірледі.

KazakhTTS - жалпы ұзақтығы 90 сағаттан астам қазақ тіліндегі аудиожазбалардан тұратын жоғары сапалы деректер жиынтығы. Бұл деректер жиынтығы кәсіби спикерлердің көмегімен жазылған ер мен әйел дауыстарынан тұрады. Деректер жиынтығы ғылым және өнеркәсіп өкілдерінің тарапынан үлкен сұраныс тудыра отырып, бір жылдың ішінде 500 астам рет жүктелген болатын.

Жұмысымызды жалғастыру үшін біз KazakhTTS2 деп аталатын жаңа деректер жиынтығын ұсынамыз. KazakhTTS2 жиынтығы көбірек деректер мен кәсіби спикерлер дауыстарымен қатар бірнеше жаңа тақырыптарды қамтиды. Атап айтқанда, бұл жиынтықта біз деректер көлемін 271 сағатқа дейін арттырдық. Үш жаңа спикер – екі әйел мен бір ер адамды қостық. Әр спикердің оқыған деректер үлесі 25 сағаттан асады. Тақырыптардың қамтылу аясын кітап пен Уикипедия мақалаларымен әртараптандырдық.

Kazakh Speech Corpus https://issai.nu.edu.kz/kz-speech-corpus/

Қазақ тілінің сөйлеу корпусы (KSC) - шамамен 335 сағаттық транскрипцияланған аудионы қамтиды, ол әртүрлі аймақтардан, жас топтарынан және жыныстағы қатысушылар айтқан 154 000-нан астам айтылыстарды қамтиды. Сапасы жоғары болуы үшін оны қазақ тілінде сөйлейтіндер мұқият тексерді. KSC — сөйлеуді тану, сөйлеу синтезі және сөйлеушіні тану сияқты қазақ сөйлеуін және тілін өңдеуге арналған әртүрлі қосымшаларды жетілдіру үшін әзірленген ең үлкен көпшілікке қолжетімді деректер қоры. KSC дерекқоры Creative Commons Attribution 4.0 халықаралық лицензиясы бойынша сұрау бойынша жалпыға ортақ және коммерциялық пайдалану үшін қол жетімді (яғни лицензия керек!) (A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline *In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 697–706. Association for Computational Linguistics, 2021.* Yerbolat Khassanov, Saida Mussakhojayeva, Almas Mirzakhmetov, Alen Adiyev, Mukhamet Nurpeiissov and Huseyin Atakan Varol, Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, Nur-Sultan, Kazakhstan мақала). Бұл аудио жазбаларды

авторлар интернет арқылы краудсорсинг болды, онда волонтерлерден веб-браузер арқылы ұсынылған сөйлемдерді оқуды сұрады.

Осы корпус әрі қаарй жетілдірілген де 2022 жылы екінші нұсқасы жасалған. Қараңыз төменде. «Although several Kazakh speech corpora have been presented (Makhambetov et al., 2013; Shiet al., 2017; Mamyrbayev et al., 2019), there is no generally accepted common corpus. Most of them are either publicly unavailable or contain an insufficient amount of data to train reliable models».

Kazakh Speech Corpus 2 https://issai.nu.edu.kz/kz-speech-corpus/

Kazakh Speech Corpus 2 (KSC 2) – өнеркәсіптік ауқымдағы алғашқы ашық бастапқы қазақ тіліндегі сөйлеу корпусы. KSC2 корпусы бұрын ұсынылған екі корпусты қамтиды: **Қазақша сөйлеу корпусы (Kazakh speech corpus)** және **Қазақша мәтіннен сөзге 2 (Kazakh Text-To-Speech 2)** және телебағдарламалар, радио, журналдар ? (орысшасында сайтта) (**senate)** және подкасттар сияқты басқа көздерден алынған қосымша деректерді толықтырады. Барлығы KSC2 құрамында 600 000-нан астам айтылымдарды қамтитын 1,2 тыс. часов 1200? сағатқа жуық жоғары сапалы транскрипцияланған деректер бар.