Дәріс №3-4
Корпус лингвистикасының қалыптасуы

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

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
7. Sketch Engine 90+ тілдегі 600 пайдалануға дайын корпусты қамтиды, олардың әрқайсысында тілдің шынайы өкілін көрсету үшін өлшемі 60 миллиард сөзге дейін бар. https://www.sketchengine.eu

Корпус лингвистикасы әлем бойынша жеке бағыт ретінде 1990 жылдары қалыптасты. Корпус жасау ісін орыс тіл білімі кеширектеу 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.