Дәріс №15
Дәріс тақырыбы: Әр типті корпустарға шолу
1. Шетел ұлттық корпустары
2. Орыс тілінің ұлттық корпустары
3. Қазақ тілінің ұлттық корпустары

1. **Британдық ұлттық корпус** http://www.natcorp.ox.ac.uk/, https://www.english-corpora.org/bnc/

Британдық ұлттық корпус (BNC) алғашында Оксфорд университетінің баспасымен 1980-1990 жылдардың басында құрылған және ол жанрлардың кең ауқымындағы (мысалы, ауызша, көркем әдебиет, журналдар, газеттер және академиялық) мәтіннің 100 миллион сөзін қамтиды. Британдық ұлттық корпуста іздеу мен зерттеу мүмкіндіктерін түрлі интерфейстер береді. Солардың бірімен BNCWeb at Lancaster University біз жұмыс жасаған болатынбыз.

Американдық ұлттық корпус (American National Corpus) https://anc.org/

Ашық Американдық ұлттық корпусы (OANC) — 1990 жылдан бастап жасалған барлық жанрлар мәтіндері мен ауызша деректердің транскрипттерін қамтитын американдық ағылшын тілінің ауқымды электронды жинағы. Барлық деректер мен аннотациялар толығымен ашық және пайдаланудың қай түріне де шектеусіз.

Қолжетімді деректер және аннотациялар мыналар:

OANC: түрлі тілдік құбылыстар (сайттан қараңыз!) үшін автоматты түрде жасалған аннотациялары бар қазіргі американдық ағылшын тілінің 15 миллион сөзі.

MASC: Ашық Американдық ұлттық корпустың (OANC) тіл құбылыстарының бірнеше қабаттары үшін қолмен жасалған немесе расталған аннотациялары бар 500 000 сөз деректері американдық ағылшын тілінің 19 жанрына (сайттан қараңыз!) теңдей таратылған.

Венгер тілінің ұлттық корпусы (Hungarian National Corpus)
http://corpus.nytud.hu/mnsz/index_eng.html

Венгер Ұлттық корпусы (HNC) бойынша жұмыс 1998 жылы Венгрия ғылым академиясының Тіл білімі ғылыми-зерттеу институтының Корпус лингвистикасы кафедрасында Тамаш Варадидің жетекшілігімен басталған. Мақсаты қазіргі венгр тілінің 100 миллион сөзден тұратын теңдестірілген анықтамалық корпусын құру болған. 2002 жылдан бастап Карпат бассейнінің венгр тіліндегі корпусы жобасының бүкіл Карпат бассейнінің венгр тілін қолдануына деректерді жинау аймағын кеңейту бойынша жаңа жұмыстар басталған. Мақсаты Венгрияның шекарасынан тыс жердегі венгр тілінің 15 миллион сөзден тұратын корпусын құру болған. 2005 жылы қарашада Словакия, Субкарпатия, Трансильвания және Воеводинағы тіл варианттылықтарын қамтитын нағыз Венгр тілінің ұлттық корпусы жасалған.

Корпусқа тіркелу қажет сонан кейін пайдалануға болады.

Өзбек тілінің Ұлттық корпусы https://uzbekcorpus.uz/enVer

Татар тілінің ұлттық корпусы Tatar National Corpus "Tugan tel" http://web-corpora.net/TatarCorpus/search/index.php?interface_language=en

Объем корпуса на сентябрь 2013 года составляет более 26 миллионов словоупотреблений. Корпус содержит тексты различных жанров (художественная литература, тексты СМИ, тексты официальных документов, учебная литература, научные публикации и др.). Каждый документ имеет метаописание (авторы, их пол, выходные данные, даты создания, жанры, части, главы и др.). Тексты, включенные в корпус, снабжены морфологической разметкой (информация о части речи и грамматических характеристиках словоформы). Морфологическая разметка текстов корпуса выполняется автоматически с использованием модуля двухуровневого морфологического анализа татарского языка, реализованного в программном инструментарии PC-KIMMO.

Түрік тілінің ұлттық корпусы Turkish National Corpus (TNC) – Türkçe Ulusal Derlemi (TUD) https://www.tnc.org.tr/

Түрік тілінің ұлттық корпусы - қазіргі түрік тілінің теңдестірілген (балансированный), ауқымды (50 миллион сөз) және жалпы мақсаттағы корпусы. ТҰК негізінен Британдық ұлттық корпустың фреймуокін ұстанады, бірақ қажет болған жағдайда ТҰК корпусының дизайнында қажетті түзетулер енгізіледі. корпус коммерциялық емес пайдалануға арналған тегін ресурс болып табылады.

50 миллион сөзді құрайтын Түрік тілінің ұлттық корпусы теңдестірілген және қазіргі түрік тілін танытатын корпус болып табылады. Ол 24 жылдық кезеңді (1990-2013) қамтитын алуан түрлі жанрлар бойынша мәтіндік деректер үлгілерінен тұрады. Жазбаша құрамы (98%) әртүрлі тақырыптардағы әртүрлі салаларда жазылған мәтіндерден тұрады. Ауызша деректерден алынған транскрипциялар ТҰК деректер қорының 2%-ын құрайды, ол спонтанды, күнделікті сөйлесулерді және белгілі бір коммуникативті жағдайларда жиналған сөйлеулерді қамтиды.

Корпус мәтіндері және метадеректер туралы мынадай оқисыздар https://www.tnc.org.tr/about-the-corpus/object/

Өте ұтымды тұсы осы корпус жайлы жазылған Жарияланымдар бөлімді ашсаңыз, Google Scholar-ға апарады.

2. **Национальный корпус русского языка**, 2 млр. сөз https://ruscorpora.ru/new/

Национальный корпус русского языка — представительная коллекция текстов на русском языке общим объемом более 2 млрд слов, оснащенная лингвистической разметкой и инструментами поиска.

Основной (374 млн)	*Устный* (13 млн)	*Параллельные* (173 млн) • Английский(44 млн) • Армянский(1,6 млн) • Башкирский(550 тыс) • Белорусский(10 млн) • Болгарский(5,2 млн) • Бурятский(401 тыс) • Испанский(5,4 млн) • Итальянский(4,9 млн) • Китайский(4,4 млн) • Корейский(73 тыс) • Латышский(4,4 млн) • Литовский(702 тыс) • Немецкий(30 млн) • Польский(6,4 млн)	*Поэтический* (13 млн)

		• Португальский(566 тыс) • Румынский(903 тыс) • Сербский(1,9 млн) • Словенский(2 млн) • Украинский(9,4 млн) • Финский(3,7 млн) • Французский(7,1 млн) • Хинди(123 тыс) • Чешский(3,9 млн) • Шведский(16 млн) • Эстонский(2,2 млн) • Многоязычный(5 млн)	
*Газетные*2 (790 млн) • Центральные СМИ(765 млн) • Региональные СМИ(24 млн)	*Акцентологический* (133 млн)	*Диалектный* (599 тыс)	*Русская классика β* (17 млн)
	Мультимедийный (5,7 млн)	*Обучающий* (664 тыс)	*Исторические*4 (14 млн) • Древнерусский(781 тыс) • Берестяные грамоты(23 тыс) • Старорусский(8,8 млн) • Церковнославянский(5,3 млн)
Синтаксический (1,5 млн)	*МультиПАРКи* (458 тыс) • Русский(229 тыс) • Англо-русский(229 тыс)	*От 2 до 15* (4,4 млн)	• *Панхронический* (383 млн)
Социальные сети (160 млн)			

3. Қазақ тілінің ұлттық корпусы. «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.

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.