Дәріс №6
Дәріс тақырыбы: Корпустардың типологиясы
Сұрақтар:

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

1. Алдыңғы сабақтан еске түсірейік. Корпустардың жіктелісінде корпустар паралельдік критерийі бойынша жіктеледі. Қандай?

«**Параллелдік**» критерийі бойынша корпустар біртілді, екітілді және көптілді корпустар болып жіктеледі.(?)

Бүгінгі сабақтағы паралель корпус деп нені танимыз дегенді анықтап алайық. Көптілді корпустар кемінде үш тілді қамтуы керек. Екі тілді қамтыса, екітілді корпустар болуы қажет. Дегенмен зерттеулерде мысалы, Тони МакЭнери, Ричард Сяо, Юкио Тоно «Corpus-based Language Studies: An Advanced Resource» (Tony McEnery, Richard Xiao, Yukio Tono) кітабында, екі не одан да көп тілді қамтыған корпустарды көптілді корпустар деп алған. паралелль корпус деп түпнұсқа мәтіндер мен олардың бір не бірнеше тілдерге аудармаларынан тұратын корпусты, *салыстырылатын* (comparable/сопоставимый) корпусқа бірдей іріктеу техникасын пайдалана отырып әр түрі тілдерден жиналған L1 деректерінен тұратын корпусты атаған. (L1 дегенді авторлар аударылмаған L2 болса аударылған деген). Бұл зерттеушілердің айтуынша корпустардың әр түрлі типтерін анықтау түрлі критерий қолдануымызбен байланысты. Параллель корпус деп егер корпус түпнұсқа мәтіндер мен олардың аудармаларынан паралель тұрса (бұл МакЕнеру мен Уэлсон кітабында: «бірдей мәтіннің түпнұсқасы (L1) мен оның аудармасы (L2)), атаймыз, немесе салыстырылатын корпус деп егер корпустың субкорпусы бірдей іріктеу техникасын қолдана отырып салыстырылса, атаймыз деген (). (семинарға дайындалғанда зерттеу авторларының осы мәселеге көзқарасын тануларың керек).

Әрі қарай паралель корпустар екі тілді және көп тілді бола алады дейді. Салыстырылатын корпустар деп МакЕнери (2003) корпус компоненттері бірдей іріктеу техникасын пайдалана отырып жиналатын және ұқсас баланс пен репрезентативтік корпустар анықталады. Яғни *әр түрлі тілдерден бірдей іріктеу уақыт кезеңіндегі бірдей домендегі бірдей жанрдағы* мәтіндердің *бірдей пропорциясы*. (McEnery, A. 2003. 'Corpus linguistics' in R. Mitkov (ed.) The Oxford Handbook of Computational Linguistics, pp. 448-463. Oxford: Oxford University Press.)

Параллель корпус әр түрлі тілдерден бірдей үлгіде іріктелген екі немесе одан да көп корпустарын тұрады. Параллель корпустың прототипі бірнеше тілдердегі бірдей құжаттардан, яғни мәтіндер мен олардың аудармаларынан тұрады. Ресми құжаттар (техникалық нұсқаулықтар, үкіметтік ақпараттық парақшалар, парламенттік істер т.б.) жиі аударылатындықтан, мәтіннің бұл түрлері параллель корпустарда жиі кездеседі. Мысалы, *Корпус ресурстары мен терминологиясын шығару The Corpus Resources and Terminology Extraction* (CRATER) https://catalogue.elra.info/en-us/repository/browse/ELRA-W0003/ https://catalogue.elra.info/en-us/repository/browse/ELRA-W0033/ корпусы корпустың осы түрінің мысалы болып табылады. Бұл корпус үш тілдегі – ағылшын, француз және испан – жоба. Корпус таза техникалық мәтіндерден тұрады, мәтіндер Халықарлық Телекоммуникациялар Одағынан алынған. Жоба Ланкастер университеті, Англия және Мадрид Автономия университетінің (Universidad Autónoma de Madrid) біріккен жұмысы. Корпус 5,5 млн сөзден тұрады. Корпуста сөйлемнен сөзге дейінгі теңестіру алгоритмдерді қолданған. Мәтіндер сөз таптары және морфологиялық аннотациямен тэгтелген.

A parallel corpus consists of two or more corpora that have been sampled in the same way from different languages. The prototypical parallel corpus consists of the same documents in a number of languages, that is a set of texts and their translations. Since official documents (technical manuals, government information leaflets, parliamentary proceedings etc.) are frequently translated, these types of text are often found in parallel corpora. The Corpus Resources and Terminology Extraction (CRATER) corpus is an example of this type of corpus.

Дегенмен, параллель корпустың басқа түрі (кейде «салыстырмалы корпус» деп аталады) әр түрлі тілдегі әр түрлі мәтіндерден тұрады: тек іріктеу әдісі бірдей. Мысалы, корпус әр тілде белгілі бір уақыт кезеңінде жарық көрген көркем әдебиеттің 100 000 сөзінен құрылуы мүмкін.

However, another type of parallel corpus (sometimes called a 'comparable corpus') consists of different texts in each language: it is merely the sampling method that is the same. For instance, the corpus might contain 100,000 words of fiction published in a given timeframe for each language.

Параллельді корпустардың қосымшалары әр түрлі тілдердің лексикасын немесе грамматикасын салыстыруды, аударылған мәтіндердің тілдік ерекшеліктерін қарауды және машиналық аудармамен жұмысты қамтиды. Осы мақсаттардың көпшілігі үшін параллель корпусты өңдеудің маңызды бірінші қадамы *теңестіру* **(alignment)** болып табылады. Теңестіру дегеніміз не? Параллельді корпуспен жұмыс істегенде, А тіліндегі мәтіннің қандай бөліктері В тіліндегі баламалы/эквивалентті мәтінге сәйкес келетінін дәл білу пайдалы. Мұндай ақпаратты параллель мәтіндерге қосу процесі *теңестіру/туралау* деп аталады. Теңестіру сөйлем деңгейінде жүзеге асырылуы мүмкін, бұл жағдайда әрбір сөйлем басқа тіл(лер)дегі сәйкес келетін сөйлеммен байланысады. Бұл қиын, өйткені сөйлем үзілістері бастапқы мәтіндегідей аудармада міндетті түрде бірдей жерде болмайды.. Немесе, теңестіру сөз деңгейінде орындалуы мүмкін, бұл жағдайда әрбір сөз параллель мәтіндегі сөзбен немесе сөздермен байланыстырылуы керек. Бұл әлдеқайда күрделі, өйткені берілген сөз басқа тілдегі бір сөзге, бір сөзден көп сөздерге сәйкес келуі мүмкін немесе басқа тілде сөзге мүлде сәйкес келмеуі мүмкін және сөздің орын тәртібі де әртүрлі болуы мүмкін. For example, English *I saw it* would correspond to French *je l'ai vu*, where *I* = *je*, *saw* = *ai vu*, and *it* = *l'*.

The applications of parallel corpora include comparing the lexis or grammar of different languages (see comparative linguistics), looking at the linguistic features of translated texts, and work on machine translation. For many of these purposes, an important first step in processing the parallel corpus is alignment.

alignment When working on a **parallel corpus**, it is useful to know exactly which parts of a text in language A correspond to the equivalent corresponding text in language B. The process of adding such information to parallel texts is called alignment. Alignment can be carried out at the *sentence level*, in which case each sentence is linked to the sentence it corresponds to in the other language(s). This is not straightforward, as the sentence breaks are not necessarily in the same place in a translation as they are in the original text.

Alternatively, alignment can be done at the *word level*, in which case each word must be linked to a word or words in the parallel text. This is much more complex, as a given word may correspond to one word, more than one word, or no word at all in the other language, and the word order may be different as well. For example, English *I saw it* would correspond to French *je l'ai vu*, where *I* = *je*, *saw* = *ai vu*, and *it* = *l'*. However, word alignment is also much more useful than sentence alignment, for example, for finding translation equivalents and compiling bilingual **lexicons**.

A Glossary of Corpus Linguistics by Paul Baker, Andrew Hardie and Tony McEnery, 2006 Edinburgh University Press Ltd

Бірінші түрдегі мәтіндердің паралельді корпустарын дайындауда және оларды өңдеуге арналған бағдарламалық пакеттерді әзірлеуде проблема туындайды, ол түпнұсқа мәтін мен оның аударма мәтінінің сәйкестігін орнату. Бұл мәселені шешу үшін мәтіндерді

автоматты **мәтінді автоматты түрде теңестіру** деп аталатын әдіс қолданылады (метод автоматического выравнивания (alignment) текстов.). Бұл әдістің мәнісі түпнұсқа мәтінді параллель сегментациялауда және оны сөйлемдерге, клауздарға (грамматикалық конструкцияларға), сөз тіркестеріне және сөздерге аударуда жатыр. Сөйлем деңгейінде теңестіру кезінде олар А.В. Зубов пен И.И. Зубованың оқулығында сипатталғандай екі мәтіннің сөйлемдері арасындағы алты мүмкін теңестіру қолданылуы мүмкін:.

1) одно исходное предложение переводится одним предложением;
2) два исходных предложения переводятся одним предложением;
3) одно исходное предложение переводится двумя предложениями;
4) два исходных предложения переводятся двумя предложениями, но внутренние границы этих предложений в тексте оригинала и тексте перевода не совпадают;
5) предложение исходного текста не переводится;
6) предложение в тексте перевода не имеет эквивалента в тексте оригинала.

«Параллель корпус» термині әдетте әртүрлі тілдердегі корпустарға қатысты болса да, бір тілдің әртүрлі аймақтық диалектілеріндегі корпустар (мысалы, Браун және Ланкастер–Осло/Берген (LOB) корпусы) немесе әртүрлі уақытта бір тілдегі әртүрлілікті (мысалы, LOB және Freiburg–LOB Corpus of British English (FLOB) корпусы) де «параллель» деп санауға болады дейді зерттеушілер.

> Although the term ‘parallel corpus’ usually refers to corpora in different languages, corpora in different regional dialects of the same language (for example, the Brown and Lancaster–Oslo/Bergen (LOB) corpora) or in the same language variety at different times (for example, the LOB and Freiburg–LOB Corpus of British English (FLOB) corpora) can also be considered to be ‘parallel’ in a similar sense.

Паралельді корпустар көптілді қоғамдарда, айталық БҰҰ, ЕО елдерінде және ресми екі тілді, мысалы Канада сияқты, елдерде жиі жасалады.

Параллель корпустарды екі негізгі түрге бөлуге болады: 1) кез келген бастапқы тілде жазылған түпнұсқа мәтіндер, және осы бастапқы мәтіндердің бір немесе бірнеше тілдердегі мәтін-аудармаларының жиынтығын білдіретін корпус; 2) екі немесе одан да көп тілде жазылғанына қарамастан бір тақырыптағы мәтіндерді біріктіретін корпустар. (Захаров, Богданова). Бұл екі корпустың екеуі де тілдерді салыстырмалы зерттеуде сонымен қатар аударманың оның ішінде машиналық аударманың тиімді әдістерін дамыту мақсаттарында жасалады және қолданылады.

Параллель корпустар тізім ашып көру:

https://www.clarin.eu/resource-families/parallel-corpora

Europarl (Koehn, 2002) - **Philipp** Koehn Филипп Коэн құрастырған. https://opus.nlpl.eu/Europarl-v3.php

United Nations Parallel Corpus https://conferences.unite.un.org/uncorpus

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B., (2016), The United Nations Parallel Corpus, Language Resources and Evaluation (LREC’16), Portorož, Slovenia, May 2016. (пдф бар)

2. Ауызша сөйлеу корпусын жасау өте күрделі жұмыс. Сондықтан да алдымен жазбаша мәтіндердің корпусы жасалды сосын барып ауызша сөйлеу тілінің корпустары жасала бастады. Соның бірі өзіміз білетін - The London-Lund Corpus of Spoken English (LLC) Ауызша

ағылшын тілінің Лондон-Лунд корпусы. Жобаның мақсаты оқыған ересек сөйлеушілердің сөйлеуіндегі ағылшын тілінің грамматикалық жүйесінің ерекшеліктерін барынша толық қамту болды. Ағылшын тілінің Лондон-Лунд корпусы екі жобадан тұрады. Біріншісі 1959 жылы Рэндольф Квирк бастаған Лондон университеті колледжіндегі (UCL) ағылшын тілін қолдану сауалнамасы the Survey of English Usage (SE). Екінші жоба – 1975 жылы Лунд университетінде Ян Свартвик Лондон сауалнамасының бауырлас жобасы ретінде бастаған The Survey of Spoken English «Ауызша ағылшын тілін зерттеу» (SSE).

> The London-Lund Corpus of Spoken English derives from two projects. The first is the Survey of English Usage (SE) at University College London, launched in 1959 by Randolph Quirk. The second project is the Survey of Spoken English (SSE), which was started by Jan Svartvik at Lund University in 1975 as a sister project of the London Survey.

Корпустың көлемі 1 миллион сөзді қолданысты құрайды. Ауызша сөйлеу мәтіндері радиохабарлардың, ресми құрылымдардың жиналыстарының, сондай-ақ бейресми әңгімелердің жазбалары болды. Корпустың машиналық нұсқасы Лунд университетінде (Швеция) жасалды және 1979 жылы пайдалануға дайын болды. Дәл осы Лондон-Лунд ауызша сөйлеудің корпусы машинада оқылатын корпустардың бірі болды. Ол жасырын жазылған сөйлесулерден құралған 34 мәтіннен тұрды, кейіннен олар Дж. Свартвик пен Р. Квирктің «Ағылшын сөйлеуінің корпусы» «Корпус английского разговора» (1980) кітабында жарияланды.

Осы корпустың жалғасы *Лондон-Лунд Корпус 2 Британдық ағылшын сөйлеу тілінің корпусы* (London-Lund Corpus 2 A CORPUS OF SPOKEN BRITISH ENGLISH). Корпус 2014-2019 жылдар аралығында британдық ағылшын ересек оқыған сөйлеушілерінен жазылып алынған 500 000 сөзден тұратын ауызша тілдің корпусы. Корпус, бір жағынан, қазіргі сөйлеуді синхрондық тұрғыдан және әртүрлі регистрде және сөйлеушілер топтары тұрғысынан зерттеудің ресурсы, екінші жағынан, бұл корпус 1950-1980 жылдары жасалған Лондон Ланд 1 корпусын жасау принциптерін қолданып жасалған. Сондықтан бұл шамамен арасы 50 жыл айырмамен ағылшын тілінің әртүрлі уақыт кезеңін салыстыра зерттеуге мүмкіндік береді. Корпус дизайны мынадан тұрады: бетпе бет әңгімелесу, телефон/смс әңгімелесу, БАҚ хабарлары, парламенттік жинақтар, комментарий, заң жинақтары және даяр сөйлеулер.

Егер Лондон Лунд корпусы лексикалық, грамматикалық және дискурс талдауымен қатар просодикалық зерттеу болса, *Ланкастер ағылшын сөйлеу корпусы (Lancaster/IBM Spoken English Corpus (SEC) т*ек просодикалық зерттеуге арналған корпус болды. Корпус 52 600 сөзден тұратын ересектердің британдық стандартты ағылшын сөйлеу тілін құрады. 1984 мен 1987 жылдар аралығында 11 категорияда - радио хабарлар, университет дәрістері, діни хабарлар, поэзия, диалог және пропаганда - іріктелді. Бұл корпустың бір артқышылығы бірнеше нұсқалар бар соның ішінде тыныс белгілермен немесе т белгілерсіз орфографиялық транксрипция, грамматикалық тегтелген, парсингтелген, екпін, интонация және паузалары көрсетіліп просодикалық транксрипцияланған. Фонетикалық транскрипциясы және СД ромда жазылған сандық жазбасы бар бұл корпус көптеген фонологиялық зерттеулер жасауға пайдасы тиді, дегенмен мұнда сөйлеушілерінің әлеуметтік немесе білімі туралы ақпарат жоқ болғандықтан әлеуметтік тілдік талдаулар жасауда шектеулер бар екенін айтады зерттеушілер.

Осыдан кейін көптеген ауызша сөйлеудің корпустары жасалды және жасалып жатыр. Жазбаша корпустарды жасаумен салыстырғанда ауызша сөйлеу корпустары баяу жүреді. Себебі оған бірнеше себептер бар. сөзді таспаға жазып алу, сосын оны транскрипциялау, ол өте қиын жұмыс. Егер жазба жақсы жазылған болса, алайда сапасы нашар болса, шу, сыртқы ортаның әртүрлі дыбыстары оның транксрипциясын жазуды қиындата түседі. (Захаров пен Богданова). Зерттеушілердің айтуынша, таспаға жазылған 1 сағаттық жазба 7000-9000 сөзден тұрады екен. 1 сағаттық жазбаны минималды

просодикалық ақпаратпен орфографиялық транскрипциялау шамамен 10 сағатты алады екен.