№ 14 дәріс. Цифрлық түркітану және корпустық лингвистика (Digital Turkology)

Дәрістің мақсаты: Түркітанудағы заманауи компьютерлік-технологиялық парадигманың мәнін ашу; түркі тілдерінің электронды корпустарын құру принциптерін, көне жазбаларды цифрландыру, OCR технологиялары мен табиғи тілді өңдеу (NLP) модельдерін қолдану әдіснамасын саралау.

1. Цифрлық гуманитаристика (Digital Humanities) аясындағы цифрлық түркітанудың қалыптасуы

- **Цифрлық түркітану (Digital Turkology) ұғымы:** Дәстүрлі түркологиялық зерттеу әдістерін (филологиялық, текстологиялық, салыстырмалы-тарихи) компьютерлік лингвистика, информатика және жасанды интеллект алгоритмдерімен интеграциялайтын пәнаралық бағыт.
- **Негізгі міндеттері:**
  - Түркі жазба жәдігерлерін сандық форматқа түсіру және сақтау;
  - Түркі тілдерінің электронды корпустарын (жазбаша және ауызша) әзірлеу;
  - Табиғи тілді өңдеу (NLP) жүйелерінде түркі тілдерінің (агглютинативті құрылымын) толыққанды жұмыс істеуін қамтамасыз ету.

2. Түркі тілдерінің ұлттық корпустарын құру принциптері

- **Корпустық лингвистика негіздері:** Корпус — белгілі бір ғылыми-лингвистикалық критерийлерге сай іріктелген, компьютерлік өңдеуден өткен мәтіндердің ірі электронды базасы.
- **Корпустарды белгілеу (Разметка / Annotation) түрлері:**

1. **Морфологиялық аннотация (POS-tagging):** Түркі тілдерінің агглютинативті табиғатына байланысты сөз түбірі мен қосымшалар тізбегін (аффикстерді) грамматикалық тұрғыдан автоматты түрде ажырату.
2. **Семантикалық аннотация:** Сөздердің контекстік мағынасын, концептуалды өрісін және сөз тіркесімділігін (коллокациясын) белгілеу.
3. **Синтаксистік аннотация (Treebanks):** Сөйлемдегі сөздердің бағыныңқылық пен сыңарылық байланысын иерархиялық ағаш сияқты құрастыру.

- **Түркі тілдік корпустардың тәжірибесі:** *Қазақ тілінің ұлттық корпусы*, *Türkçe Ulusal Derlemi (TUD)*, сондай-ақ халықаралық ғылыми бастамалар.

3. Түркі жазбаларын цифрландыру, OCR және NLP модельдері

- **Көне жазбалар мен қолжазбаларды цифрландыру:**
  - Руникалық (Орхон-Енисей, Талас), Ескі ұйғыр, Араб графикалы (Шағатай, Ортағасырлық қыпшақ) ескерткіштерін 3D-сканерлеу, спектрлік фотоға түсіру және Юникод (Unicode) стандарттарына енгізу.
  - Халықаралық цифрлық жобалар тәжірибесі (мысалы, Германиядағы Берлин-Бранденбург академиясының *Turfanforschung* жобасы).
- **Оптикалық таңбаларды тану (OCR — Optical Character Recognition) технологиялары:**
  - Көне қолжазбалар мен сирек кездесетін кітаптардың графемаларын, шрифтері мен пиксельдік кескіндерін машиналық оқыту (Machine Learning) арқылы автоматты түрде мәтіндік форматқа айналдыру.
- **Табиғи тілді өңдеу (NLP) және Жасанды интеллект (AI):**
  - **Нейрондық желілер мен Үлкен тілдік модельдер (LLM):** Түркі тілдерінің морфологиялық бағанының күрделілігін (агглютинациясын) ескеретін арнайы токенезация (subword tokenization) әдістерін қолдану.
  - Автоматты аударма, семантикалық іздеу, мәтіндердің авто-аннотациясы мен транслитерациясы (мысалы, араб/кирилл/латын графикалары арасындағы кедергісіз автоматты ауысу).

Семинар тапсырмалары мен ДОӨЖ (СӨЖ) бағыттары:

1. Заманауи түркі тілдерінің ұлттық корпустарына салыстырмалы шолу жасау (көлемі, разметка деңгейі, функционалы).
2. Көне руникалық жазбаларды цифрлық Юникод жүйесінде таңбалаудың графикалық-техникалық мәселелерін талқылау.
3. Түркі тілдеріндегі морфологиялық анализаторлардың (Morphological Analyzers) жұмыс істеу алгоритмдерін саралау.

Әдебиеттер тізімі:

1. Захаров В. П., Богданова С. Ю. *Корпусная лингвистика.* — СПб., 2020.
2. Aksan Y., et al. *Construction of the Turkish National Corpus (TNC).* — 2012.
3. Дерлемдер мен корпустар бойынша халықаралық түркологиялық цифрлық жобалар материалдары.