PARALLEL KORPUS DASTURIY TA’MINOTNING TEXNIK ASOSLARI

Xolmonova Iqbola Alisher qizi,
Tayanch doktorant (PhD)
iqbolabintualisher@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada parallel korpus dasturiy ta’minotining texnik asoslari tahlil qilinadi. Tadqiqotda matnlarni raqamlashtirish, tozalash, segmentatsiya qilish, avtomatik moslashtirish, lingvistik teglash hamda ma’lumotlarni saqlash va qidiruv texnologiyalari ko‘rib chiqiladi. OCR, Unicode, indekslash va NLP vositalarining korpus yaratish jarayonidagi o‘rni ilmiy jihatdan asoslanadi. Natijalar parallel korpuslar va mashina tarjimasi tizimlari uchun sifatli lingvistik resurslar yaratishda muhim ahamiyat kasb etadi.

Kalit so‘zlar: *parallel korpus, kompyuter lingvistikasi, NLP, OCR, segmentatsiya, lingvistik annotatsiya, ma’lumotlar bazasi, qidiruv tizimlari*

Abstract. This article analyzes the technical foundations of parallel corpus software. The study examines text digitization, cleaning, segmentation, automatic alignment, linguistic annotation, as well as data storage and retrieval technologies. The roles of OCR, Unicode, indexing, and NLP tools in corpus construction are discussed from a theoretical and methodological perspective. The results highlight the importance of these technologies in building high-quality linguistic resources for parallel corpora and machine translation systems.

Keywords: *parallel corpus, computational linguistics, NLP, OCR, segmentation, linguistic annotation, database systems, information retrieval systems.*

Zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash sohalarida parallel korpuslar muhim tadqiqot obyektlaridan biri hisoblanadi. Ular turli tillardagi matnlarning moslashtirilgan (aligned) korpusi bo‘lib, mashina tarjimasi, lingvistik tahlil va ko‘p tilli axborot tizimlarini rivojlantirishda asosiy ma’lumot manbasi

sifatida xizmat qiladi. Parallel korpuslarni yaratish jarayoni matnlarni yig‘ish, raqamlashtirish, tozalash, segmentatsiya, avtomatik moslashtirish va lingvistik teglash kabi bir nechta bosqichlardan iborat. Ushbu bosqichlarning sifati korpusning umumiy aniqligi va keyingi tahlil natijalariga bevosita ta’sir ko‘rsatadi. Shu bilan birga, matnlarni raqamlashtirish va formatlash, shuningdek, ularni saqlash va tezkor qidiruv texnologiyalari korpus tizimlarining samarali ishlashini ta’minlaydi. Ayniqsa, katta hajmdagi matnlar bilan ishlashda indekslash va optimallashtirilgan ma’lumotlar bazalari muhim ahamiyat kasb etadi. Umuman olganda, parallel korpuslarni yaratish murakkab ko‘p bosqichli jarayon bo‘lib, u zamonaviy NLP va kompyuter lingvistikasi texnologiyalarining integratsiyasini talab etadi.

Matnlarni raqamlashtirish va formatlash texnologiyalari. Matnlarni raqamlashtirish va formatlash texnologiyalari zamonaviy korpus lingvistikasi hamda kompyuter lingvistikasining muhim tarkibiy qismi hisoblanadi. Ushbu jarayonlar an’anaviy shakldagi matnlarni raqamlashtirish, ularni strukturaviy jihatdan tartibga solish va lingvistik tahlil uchun tayyorlashni o‘z ichiga oladi. Raqamlashtirish bosqichi, avvalo, qog‘oz shaklidagi manbalarni skanerlash orqali ularni tasvir ko‘rinishida olishdan boshlanadi. Keyingi bosqichda Optical Character Recognition (OCR) texnologiyasi yordamida ushbu tasvirlar mashinada o‘qiladigan matn formatiga aylantiriladi. Bu jarayon kompyuter lingvistikasida muhim texnologik yutuqlardan biri bo‘lib, amaliyotda ABBYY FineReader va Tesseract kabi vositalar keng qo‘llaniladi [6:629-633].

Raqamlashtirilgan matnlar, odatda, dastlabki holatda turli xatoliklarni o‘z ichiga oladi, shu sababli ular maxsus tahrir bosqichidan o‘tkaziladi. Bu jarayonda imlo, punktuatsiya va grafik belgilar normallashtiriladi hamda matn yagona standartga keltiriladi. Shundan so‘ng formatlash bosqichi amalga oshiriladi. Formatlash matnni strukturaviy jihatdan tashkil etish, ya’ni uning ichki birliklarini (sarlavha, paragraf, gap) aniq belgilashni nazarda tutadi. Ushbu vazifani bajarishda

XML va HTML kabi belgilash tillari muhim rol o‘ynaydi. Xususan, matnni semantik va strukturaviy jihatdan belgilash tamoyillari Text Encoding Initiative tomonidan ishlab chiqilgan standartlarda o‘z ifodasini topgan.

Matnlarni raqamli muhitda to‘g‘ri aks ettirish uchun kodlash standartlariga rioya qilish zarur. Bu borada Unicode standarti alohida ahamiyat kasb etadi, chunki u turli tillarga mansub belgilarni yagona tizim asosida ifodalash imkonini beradi [7:1-20]. Ayniqsa, ko‘p tilli va parallel korpuslar yaratishda ushbu standartdan foydalanish matnlar o‘rtasidagi muvofiqlikni ta’minlaydi. Bundan tashqari, formatlash jarayonining muhim tarkibiy qismi sifatida lingvistik annotatsiya amalga oshiriladi. Bu jarayonda matnlarga morfologik, sintaktik va semantik belgilar biriktiriladi, bu esa ularni turli lingvistik tahlil qilish imkonini beradi. Korpus lingvistikasida ko‘p qatlamli annotatsiya nazariyasi Steven Bird va Mark Liberman tomonidan asoslab berilgan bo‘lib, u lingvistik ma’lumotlarni tizimli tashkil etishda muhim metodologik asos hisoblanadi [2:23-35].

1-jadval. Matnlarni raqamlashtirish va qayta ishlash bosqichlari

Bosqich	**Izoh**	**Texnologiya**	**Natija**
Skanerlash	Qog‘oz matnni raqamli shaklga o‘tkazish	Skaner qurilmalari	Raqamli tasvir
OCR	Tasvirdagi matnni tahrirlanadigan matnga aylantirish	Optical Character Recognition, ABBYY FineReader, Tesseract	Raqamli matn
Tozalash	Xatolar va ortiqcha belgilarni olib tashlash	NLP vositalari, regex	Tozalangan matn
Formatlash	Matnni strukturaviy birliklarga ajratish	XML, HTML	Tuzilgan matn
Kodlash	Belgilarni standart tizimga moslashtirish	Unicode (UTF-8)	Standart matn

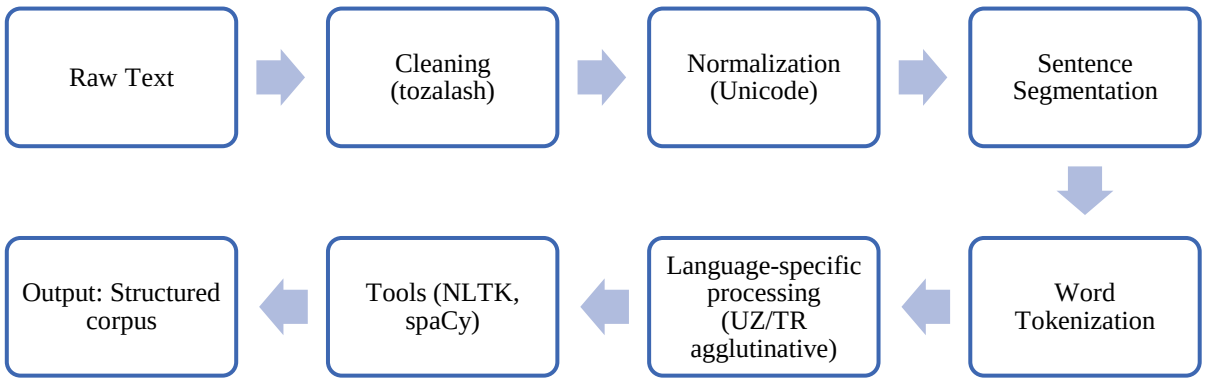
Umuman olganda, matnlarni raqamlashtirish va formatlash texnologiyalari korpus yaratish jarayonining ajralmas bosqichi bo‘lib, ayniqsa parallel korpuslar tuzishda muhim ahamiyatga ega. Ushbu texnologiyalar yordamida matnlar yagona standart asosida tartibga solinadi, ularning o‘zaro mosligi (alignment) ta’minlanadi va qidiruv hamda tahlil jarayonlarining samaradorligi oshadi. Natijada, bu jarayonlar neyron mashina tarjimasi tizimlari uchun sifatli ma’lumotlar bazasini yaratishga xizmat qiladi.

Ma’lumotlarni qayta ishlash (tozalash va segmentatsiya). Matnli ma’lumotlarni qayta ishlash jarayoni korpus lingvistikasi va tabiiy tilni qayta ishlash tizimlarida muhim bosqichlardan biri hisoblanadi. Ushbu jarayonning asosiy vazifasi xom (raw) matnlarni tahlilga tayyorlash, ya’ni ularni tozalash (cleaning) va segmentatsiya qilish orqali strukturaviy jihatdan tartibga solishdan iborat. Tozalash bosqichi matndagi ortiqcha yoki noto‘g‘ri elementlarni aniqlash va bartaraf etishga qaratilgan bo‘lib, ular orasida HTML teglarini olib tashlash, maxsus belgilarni normallashtirish, kodlash xatoliklarini tuzatish va dublikatlarni yo‘qotish kabi jarayonlar mavjud. Bu bosqichda, ayniqsa, Unicode standartiga asoslangan kodlashni bir xillikka keltirish muhim hisoblanadi, chunki turli manbalardan olingan matnlarda kodlash nomuvofiqligi uchrashi mumkin.

Tozalash jarayonidan so‘ng segmentatsiya bosqichi amalga oshiriladi. Segmentatsiya − bu matnni mantiqiy birliklarga ajratish jarayoni bo‘lib, u odatda gaplarga (sentence segmentation) va so‘zlarga (tokenization) bo‘linadi. Gap segmentatsiyasi punktuatsiya belgilariga asoslangan holda amalga oshiriladi, biroq bu jarayon har doim ham oddiy emas, chunki qisqartmalar, iqtiboslar va ko‘p nuqtali konstruksiyalar, shakl va ma’no nomutanosibligi noto‘g‘ri segmentatsiyaga olib kelishi mumkin. So‘z segmentatsiyasi matnni alohida minimal til birliklari – tokenlarga ajratish jarayonini anglatadi. Bu jarayon turli tillarning grammatik va morfologik xususiyatlariga qarab turlicha yondashuvlarni talab qiladi. Xususan,

agglutinativ tillarda, jumladan o‘zbek va turk tillarida, so‘zlar ko‘plab qo‘shimchalardan tashkil topgani uchun tokenizatsiya jarayoni murakkabroq kechadi va oddiy bo‘linish qoidalari yetarli bo‘lmaydi. Shu sababli, bunday tillarda tokenizatsiya morfologik tuzilmani hisobga olgan holda, qo‘shimchalarni asosdan ajratish yoki maxsus lingvistik qoidalarga tayangan holda amalga oshiriladi. Amaliyotda segmentatsiya va tokenizatsiya jarayonlarini avtomatlashtirish uchun turli dasturiy vositalardan foydalaniladi. Jumladan, NLTK va spaCy kutubxonalari matnlarni gap va so‘zlarga ajratish, shuningdek dastlabki lingvistik tahlilni amalga oshirish imkonini beradi [1:1-504]. Ushbu vositalar statistik va qoidaviy (rule-based) algoritmlarni birgalikda qo‘llash orqali yuqori aniqlikni ta’minlaydi.

Ma’lumotlarni tozalash va segmentatsiya qilish bosqichining samaradorligi korpus sifati va keyingi bosqichlardagi lingvistik tahlil natijalariga bevosita ta’sir ko‘rsatadi. Ayniqsa, parallel korpuslar yaratishda noto‘g‘ri segmentatsiya gaplar moslashuvi (alignment) jarayonida xatoliklarga olib kelishi mumkin. Shu sababli, ushbu bosqichda avtomatik algoritmlar bilan bir qatorda qo‘lda tekshirish (manual validation) ham muhim ahamiyat kasb etadi.



1-chizma. Korpus yaratishda matnni oldindan qayta ishlash pipeline’i

Umuman olganda, ma’lumotlarni qayta ishlashning tozalash va segmentatsiya bosqichlari korpus lingvistikasi uchun fundamental asos yaratadi. Ushbu jarayonlar matnlarni strukturaviy jihatdan tartibga solish, ularni izchil va tizimli holga keltirish hamda keyingi lingvistik annotatsiya va tahlil jarayonlari uchun tayyorlash imkonini beradi.

Avtomatik moslashtirish va lingvistik teglash mexanizmlari. Parallel korpuslar yaratishda avtomatik moslashtirish (alignment) va lingvistik teglash mexanizmlari asosiy texnologik bosqichlardan biri hisoblanadi. Ushbu jarayonlar turli tillardagi matn birliklarini o‘zaro muvofiqlashtirish hamda ularga lingvistik ma’lumotlarni biriktirish orqali korpusning funksional imkoniyatlarini kengaytiradi. Avtomatik moslashtirish algoritmlari odatda matnlarni gap yoki segment darajasida moslashtirishga qaratilgan bo‘lib, ular statistik, qoidaviy va gibrid yondashuvlarga asoslanadi.

Gap darajasidagi moslashtirishning klassik modellaridan biri William A. Gale va Kenneth W. Church tomonidan taklif etilgan uzunlikka asoslangan algoritm hisoblanadi. Ushbu model matn segmentlarining uzunligi (belgilar yoki so‘zlar soni) o‘rtasidagi korrelyatsiyaga tayanib, mos keluvchi gaplarni aniqlaydi [4:75-102]. Keyinchalik ushbu yondashuv statistik mashina tarjimasi modellarida rivojlantirildi. Xususan, Peter F. Brown va hammualliflar tomonidan ishlab chiqilgan IBM modellari so‘z darajasidagi moslashtirishni ehtimollik nazariyasi asosida amalga oshirish imkonini berdi [3:263-311]. Zamonaviy tizimlarda esa ushbu yondashuvlar neyron tarmoqlar asosidagi modellar bilan boyitilib, kontekstual moslikni aniqlash imkoniyati kengaytirildi.

Avtomatik moslashtirishdan so‘ng lingvistik teglash (annotation) jarayoni amalga oshiriladi. Lingvistik teglash matnga morfologik, sintaktik va semantik ma’lumotlarni biriktirishni nazarda tutadi. Eng keng tarqalgan teglash turlaridan biri morfologik teglash, ya’ni Part-of-Speech Tagging hisoblanadi. Ushbu jarayon har

bir so‘zga uning grammatik kategoriyasini (ot, fe’l, sifat va boshqalar) biriktirish orqali matnni chuqur tahlil qilish imkonini beradi. Zamonaviy teglash tizimlari qoidaviy va statistik yondashuvlar asosida ishlaydi hamda ko‘pincha mashinaviy o‘rganish algoritmlaridan foydalanadi.

Avtomatik moslashtirish va lingvistik teglash mexanizmlarining samaradorligi parallel korpus sifatiga bevosita ta’sir ko‘rsatadi. Noto‘g‘ri moslashtirish yoki xatoli teglash natijalarning ishonchliligini pasaytiradi va keyingi bosqichlarda, xususan mashina tarjimasi va lingvistik tahlil jarayonlarida muammolar keltirib chiqaradi. Shu sababli, zamonaviy tadqiqotlarda avtomatik usullar bilan bir qatorda yarim avtomatik va qo‘lda tekshirish mexanizmlaridan ham foydalanish tavsiya etiladi.

Ma’lumotlarni saqlash va tezkor qidiruv texnologiyalari. Korpus lingvistikasi va tabiiy tilni qayta ishlash tizimlarida ma’lumotlarni saqlash va tezkor qidiruv texnologiyalari muhim funksional komponentlardan biri hisoblanadi. Katta hajmdagi matnli ma’lumotlarni samarali boshqarish, ularni strukturaviy jihatdan to‘g‘ri tashkil etish hamda foydalanuvchining qidiruv so‘rovlariga tezkor javob qaytarish zamonaviy korpus tizimlarining asosiy talablaridan biridir. Shu sababli, ma’lumotlarni saqlashda optimallashtirilgan ma’lumotlar bazalari va qidiruv mexanizmlaridan foydalanish zarur bo‘ladi.

Matnli ma’lumotlarni saqlashda an’anaviy relyatsion ma’lumotlar bazalari bilan bir qatorda NoSQL yondashuvlari ham keng qo‘llanilmoqda. Xususan, MongoDB kabi hujjatga yo‘naltirilgan bazalar yarim strukturalangan matnlarni saqlashda qulaylik yaratadi, chunki ular moslashuvchan ma’lumotlar modelini taqdim etadi. Shu bilan birga, murakkab so‘rovlar va qat’iy strukturaga ega ma’lumotlar uchun PostgreSQL kabi relyatsion bazalar samarali hisoblanadi [1:1-504]. Korpus tizimlarida ko‘pincha ushbu ikki yondashuvning kombinatsiyasi qo‘llanilib, ma’lumotlarni saqlashning optimal modeli yaratiladi.

Tezkor qidiruvni ta’minlashda indeksatsiya (indexing) texnologiyalari muhim o‘rin tutadi. Indeksatsiya matndagi so‘z va iboralarni maxsus tuzilmalarda saqlash orqali qidiruv jarayonini sezilarli darajada tezlashtiradi. Inverted index (teskari indeks) modeli eng keng tarqalgan yondashuvlardan biri bo‘lib, unda har bir so‘zga mos ravishda uning hujjatdagi joylashuvi saqlanadi. Ushbu model asosida ishlovchi qidiruv tizimlari katta hajmdagi korpuslarda ham yuqori tezlikda natija qaytarish imkonini beradi. Amaliyotda tezkor qidiruvni tashkil etish uchun maxsus qidiruv platformalari qo‘llaniladi. Jumladan, Elasticsearch va Apache Lucene kabi texnologiyalar matnli ma’lumotlarni indekslash va qidirishda keng qo‘llaniladi [5:1-544]. Ushbu tizimlar nafaqat oddiy kalit so‘z bo‘yicha qidiruvni, balki murakkab so‘rovlar, fraza qidiruvi, lemmatizatsiya va morfologik variantlarni hisobga olgan holda qidiruvni ham amalga oshirish imkonini beradi. Bundan tashqari, zamonaviy korpus tizimlarida semantik qidiruv texnologiyalari ham rivojlanmoqda. Bu yondashuvda qidiruv faqat kalit so‘zlarga emas, balki matn mazmuniga asoslanadi. Bunda vektorli modellar va neyron tarmoqlar asosida matnlarning semantik o‘xshashligi aniqlanadi. Natijada foydalanuvchi so‘roviga yanada mos va aniq natijalar taqdim etiladi.

Ma’lumotlarni saqlash va qidiruv texnologiyalarining samaradorligi korpus tizimining umumiy ishlash tezligi va funksional imkoniyatlariga bevosita ta’sir ko‘rsatadi. Noto‘g‘ri tashkil etilgan saqlash modeli yoki yetarli darajada optimallashtirilmagan qidiruv mexanizmi tizimning sekin ishlashiga va foydalanuvchi tajribasining yomonlashishiga olib keladi. Shu sababli, zamonaviy korpuslarni yaratishda yuqori unumdorlikka ega ma’lumotlar bazalari va ilg‘or qidiruv algoritmlaridan foydalanish muhim ahamiyat kasb etadi.

Foydalanilgan adabiyotlar ro‘yxati

1. Bird S., Klein E., Loper E. Natural Language Processing with Python. O’Reilly Media, 2009, pp. 1–504.

2. Bird S., Liberman M. A Formal Framework for Linguistic Annotation. Speech Communication, 2001, pp. 23–35.

3. Brown P. F., et al. The Mathematics of Statistical Machine Translation. Computational Linguistics, 1993, pp. 263–311.

4. Gale W. A., Church K. W. A Program for Aligning Sentences. Computational Linguistics, 1993, pp. 75–102.

5. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008, pp. 1–544.

6. Smith R. An Overview of the Tesseract OCR Engine. In: Proceedings of ICDAR, 2007, pp. 629–633.

7. Stonebraker M., Hellerstein J. What Goes Around Comes Around. In: Readings in Database Systems, 2005, pp. 1–20.