PARALLEL KORPUSNING REPREZENTATIVLIGI VA BALANSLASH MUAMMOLARI

Hamroyeva Shahlo Mirdjanovna,
Filologiya fanlari doktori., prof.v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Xolmonova Iqbola Alisher qizi,
tayanch doktorant
iqbolabintualisher@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada parallel korpuslarning reprezentativligi va balanslash muammolari lingvistik hamda korpus lingvistikasi nuqtai nazaridan tahlil qilinadi. Parallel korpusning til materialini real kommunikativ holatda qanchalik to‘liq aks ettirishi, uning ilmiy va amaliy qiymatini belgilovchi asosiy omillardan biri sifatida qaraladi. Tadqiqotda reprezentativlik tushunchasining nazariy asoslari, korpusni balanslash metodlari hamda amaliyotda uchraydigan asosiy muammolar yoritiladi. Shuningdek, ushbu muammolarning neyron mashina tarjimasi tizimlariga ta’siri tahlil qilinadi. Maqolada korpus sifatini oshirishga qaratilgan stratifikatsiya, metadata asosida boshqarish va preprocessing yondashuvlari ham ko‘rib chiqiladi.

Kalit so‘zlar: *Parallel korpus, reprezentativlik, balanslash, korpus lingvistikasi, alignment, janr xilma-xilligi, metadata, Neyron mashina tarjimasi, data sparsity, subword modeling.*

Abstract: This article analyzes the issues of representativeness and balancing in parallel corpora from both linguistic and corpus linguistics perspectives. The extent to which a parallel corpus reflects language material in real communicative contexts is considered one of the key factors determining its scientific and practical value. The study addresses the theoretical foundations of representativeness, methods of corpus balancing, and major practical challenges encountered in practice. Furthermore, the impact of these issues on Neural Machine Translation systems is analyzed. The article also examines approaches aimed at improving

corpus quality, including stratification, metadata-based management, and preprocessing techniques.

Keywords: *Parallel corpus, representativeness, balancing, corpus linguistics, alignment, genre diversity, metadata, Neural Machine Translation, data sparsity, subword modeling.*

Parallel korpus bu ikki yoki undan ortiq tillardagi matnlar va ularning o‘zaro tarjimalaridan tashkil topgan, segment darajasida moslashtirilgan lingvistik resursdir. Korpus lingvistikasida bunday resurslar tarjima jarayonini empirik o‘rganish, tillararo kontrastiv tahlil va statistik asoslangan mashina tarjima tizimlarini ishlab chiqishda asosiy manbalardan biri hisoblanadi [3:1-294].

Reprezentativlik esa korpusning muayyan til yoki til juftligining real lingvistik qo‘llanishini qay darajada to‘liq va ishonchli aks ettirishini bildiradi. Sinclair ta’kidlaganidek, korpusning ilmiy qiymati uning hajmiga emas, balki uning tilning tabiiy va xilma-xil qo‘llanishini qamrab olish darajasiga bog‘liqdir. Reprezentativ korpus tilning leksik, grammatik va pragmatik qatlamlarini muvozanatli tarzda o‘z ichiga olishi lozim.

Parallel korpuslarda reprezentativlik masalasi yanada murakkablashadi, chunki bu yerda nafaqat bitta tilning ichki xilma-xilligi, balki ikki til o‘rtasidagi semantik va strukturaviy muvofiqlik ham hisobga olinadi. Brown va boshqalar statistik mashina tarjima modellarida korpus sifati va uning reprezentativligi natijaviy tarjima aniqligiga bevosita ta’sir qilishini ko‘rsatib bergan. Shu sababli, parallel korpusda janrlar, domenlar va uslubiy qatlamlar o‘rtasidagi muvozanat ilmiy jihatdan muhim hisoblanadi.

Zamonaviy tadqiqotlarda reprezentativlik va balanslash masalalari neyron mashina tarjimasi tizimlarining sifatini belgilovchi asosiy omillardan biri sifatida qaraladi. Yetarli darajada reprezentativ bo‘lmagan korpuslar modelda “data sparsity” muammosini kuchaytirib, tarjima sifatini pasaytiradi.

Reprezentativlikni ta’minlovchi asosiy omillar. Korpus lingvistikasida reprezentativlik – korpusning real til ishlatilishini ishonchli va muvozanatli aks ettira olish darajasi sifatida talqin qilinadi. Sinclair ta’kidlaganidek, korpusning ilmiy qiymati uning hajmidan ko‘ra, tilning tabiiy va xilma-xil qo‘llanishini qanchalik to‘liq qamrab olishiga bog‘liq. Shu nuqtai nazardan, korpus reprezentativligini ta’minlash bir nechta asosiy omillarga tayanadi.

Birinchi muhim omil – janr va uslub xilma-xilligidir. Korpus tarkibida badiiy, ilmiy, rasmiy va publitsistik matnlar muvozanatli tarzda kiritilishi kerak. Agar korpus faqat bitta janrga (masalan, faqat yangiliklar yoki badiiy matnlarga) asoslangan bo‘lsa, u tilning umumiy funksional ko‘lamini to‘liq aks ettira olmaydi [3:1-294].

Ikkinchi omil – tematik (domen) qamrov kengligi. Til turli kommunikativ sohalarda turlicha namoyon bo‘ladi. Shuning uchun korpus siyosat, iqtisod, ta’lim, texnologiya va kundalik hayotga oid matnlarni o‘z ichiga olishi zarur. Bu til birliklarining turli kontekstlarda ishlatilishini modellashtirish imkonini beradi.

Uchinchi omil – leksik va grammatik xilma-xillikning ta’minlanishi. Reprezentativ korpusda so‘z boyligi, grammatik shakllar va sintaktik konstruksiyalar keng qamrovda aks etishi lozim [2:223-243].

To‘rtinchi omil – autentiklik va zamonaviylik. Korpus real kommunikativ materiallarga asoslangan bo‘lishi va zamonaviy til holatini aks ettirishi kerak. Sun’iy yoki eskirgan matnlar tilning hozirgi ishlatilish tendensiyalarini buzishi mumkin.

Beshinchi omil – balanslash. Korpusga kiritiladigan matnlar tasodifiy emas, balki ma’lum mezonlar asosida tanlanishi lozim. Bu jarayonda stratifikatsiya usuli keng qo‘llaniladi, ya’ni matnlar janr, mavzu va uslub bo‘yicha oldindan belgilangan nisbatlarda taqsimlanadi [4:1-386].

1-jadval. Balanslash strategiyasi

Janr	Ulushi
Badiiy matnlar	30%

Ilmiy matnlar	25%
Publitsistik matnlar	25%
Rasmiy matnlar	20%

Umuman olganda, reprezentativlikni ta’minlash bu faqat katta hajmli korpus yaratish emas, balki tilning real ishlatilish holatini muvozanatli, tizimli va ilmiy asoslangan tarzda modellashtirish jarayonidir.

2-jadval. Reprezentativlikni ta’minlovchi asosiy omillar

№	Omil	Qisqa tavsif
1.	Janr xilma-xilligi	Turli uslubdagi matnlar (badiiy, ilmiy, rasmiy)
2.	Tematik qamrov	Turli sohalar (siyosat, iqtisod, ta’lim va b.)
3.	Lingvistik xilma-xillik	So‘z va grammatik shakllarning keng qamrovi
4.	Autentiklik	Real va tabiiy til materiallari
5.	Balanslash	Matnlarning muvozanatli tanlanishi

Korpusni balanslash: mohiyati va metodlari. Korpusni balanslash – bu korpus tarkibidagi matnlarni turli lingvistik, tematik va funksional kategoriyalar bo‘yicha muvozanatli taqsimlash jarayonidir. Korpus lingvistikasida balanslashning asosiy maqsadi til materialining bir tomonlama (bias) bo‘lib qolishining oldini olish va tilning real ishlatilish holatini ishonchli aks ettirishdir . Balanslangan korpus tilning turli janr, uslub va domenlarini nisbatan teng qamrab oladi, bu esa lingvistik tahlil va avtomatik modellar uchun ishonchli asos yaratadi. Nazariy jihatdan balanslash reprezentativlik tushunchasi bilan bevosita bog‘liq bo‘lib, korpusning ilmiy qiymatini oshiruvchi asosiy omillardan biridir. Agar korpusda ma’lum bir janr yoki mavzu haddan tashqari ustun bo‘lsa, bu tilning umumiy strukturasi haqida noto‘g‘ri xulosalarga olib kelishi mumkin [6:1-224].

Korpusni balanslashda bir nechta metodlar qo‘llaniladi. Stratifikatsiya usuli eng keng tarqalgan yondashuv bo‘lib, unda matnlar oldindan belgilangan kategoriyalar (janr, mavzu, uslub) bo‘yicha qatlamlarga ajratiladi va har bir qatlamdan ma’lum miqdorda namuna olinadi. Bu usul korpus ichida ichki muvozanatni ta’minlashga yordam beradi. Bundan tashqari, tasodifiy tanlash

(random sampling) usuli ham qo‘llaniladi, bunda matnlar katta to‘plamdan statistik tasodif asosida tanlanadi. Bu yondashuv subyektiv tanlash xatolarini kamaytiradi, biroq har doim ham to‘liq balansni kafolatlamaydi. Yana bir muhim metod – metadata asosida balanslash bo‘lib, bunda har bir matn haqida janr, muallif, vaqt, domen kabi ma’lumotlar (metadata) saqlanadi va shu asosda korpus tarkibi boshqariladi. Bu usul ayniqsa yirik va ko‘p qatlamli korpuslar uchun samarali hisoblanadi.

Parallel korpuslarda reprezentativlik va balanslash muammolari. Parallel korpuslar ikki yoki undan ortiq tillardagi matnlar va ularning o‘zaro tarjimalaridan tashkil topgan lingvistik resurs bo‘lib, ular tarjima tadqiqotlari va avtomatik tarjima tizimlari uchun asosiy ma’lumot manbai hisoblanadi. Bunday korpuslarning ilmiy qiymati ularning reprezentativligi va balanslanganligiga bevosita bog‘liq.

Reprezentativlik muammosi shundan iboratki, parallel korpus ko‘pincha tilning real qo‘llanish holatini to‘liq aks ettira olmaydi. Buning asosiy sabablari – matnlarning cheklangan manbalardan olinishi, ayrim janr yoki domenlarning ustunligi hamda tarjima qilingan matnlarning tabiiy emasligi (translationese) hisoblanadi. Natijada korpus tilning barcha funksional qatlamlarini bir xil darajada qamrab olmaydi, bu esa lingvistik tahlil va statistik modellar uchun xatoliklarni keltirib chiqaradi [5:1-179].

Balanslash muammosi esa korpus tarkibidagi turli kategoriyalar o‘rtasidagi nomutanosiblik bilan bog‘liq. Amaliyotda ko‘pincha badiiy yoki yangilik matnlari ko‘p bo‘lib, ilmiy, texnik yoki huquqiy matnlar yetarli darajada kiritilmaydi. Shuningdek, ayrim tillar juftligida parallel materiallar soni teng emasligi ham muhim muammo hisoblanadi. Bu holat modelning ayrim domenlarda yaxshi ishlashi, boshqalarida esa past natija berishiga olib keladi. Bundan tashqari, parallel korpuslarda alignment (moslashtirish) muammolari ham reprezentativlik va balanslashga ta’sir qiladi. Gaplar darajasida birga bir moslik har doim ham mavjud

bo‘lmaydi; birga ko‘p yoki ko‘pga ko‘p holatlar esa strukturaviy nomutanosiblikni kuchaytiradi. Bu esa korpusni avtomatik qayta ishlash jarayonini murakkablashtiradi.

Ushbu muammolar ayniqsa neyron mashina tarjimasi tizimlarida sezilarli ta’sir ko‘rsatadi. Noto‘g‘ri balanslangan yoki reprezentativligi past korpuslar modelning umumlashtirish qobiliyatini kamaytiradi va tarjima sifatini pasaytiradi [2:223-243]. Shu sababli zamonaviy tadqiqotlarda korpusni yaratish jarayonida janrlar bo‘yicha stratifikatsiya, metadata asosida boshqarish va sifat nazorati kabi yondashuvlar keng qo‘llanilmoqda. Bu usullar korpusning ilmiy ishonchliligini oshirish va avtomatik tarjima tizimlarida barqaror natijalarga erishish imkonini beradi.

3-Jadval. Parallel korpuslarda reprezentativlik va balanslash: muammo va yechimlar

№	Muammo	Yechim
1	Janr va domen nomutanosibligi	Stratifikatsiya asosida janrlarni muvozanatli tanlash
2	Cheklangan parallel matnlar	Korpusni kengaytirish va yangi manbalarni qo‘shish
3	Translationese (sun’iy tarjima tili)	Original va tabiiy matnlarga ustuvorlik berish
4	Alignment muammolari (1:1, 1:N, N:M)	Yarim avtomatik + qo‘lda tekshirish usullarini qo‘llash
5	Leksik va morfologik disbalans	Subword (BPE) va normalizatsiya texnikalaridan foydalanish

Reprezentativlik va balanslash NMT tizimlarining sifatini belgilovchi asosiy omillardan biri hisoblanadi. Neyron mashina tarjimasi katta hajmdagi parallel ma’lumotlarga tayanadi, shuning uchun korpusning sifati modelning o‘rganish jarayoniga bevosita ta’sir qiladi [1:1-15].

Birinchi navbatda, reprezentativ bo‘lmagan korpus NMT modelida noto‘g‘ri umumlashtirish (poor generalization) muammosini keltirib chiqaradi. Agar korpus faqat ayrim janr yoki domenlarni qamrab olsa, model boshqa sohalardagi matnlarni

tarjima qilishda xatolarga yo‘l qo‘yadi. Bu holat “domain bias” deb ataladi. Ikkinchidan, balanslanmagan korpus “data skewness” muammosini yuzaga keltiradi. Ya’ni, model ko‘p uchraydigan strukturani yaxshi o‘rganadi, kam uchraydigan leksik yoki grammatik shakllarni esa yomon o‘zlashtiradi. Natijada tarjimada bir xillik va takroriy konstruktsiyalar paydo bo‘ladi. Uchinchidan, parallel korpusdagi alignment xatolari NMT modelining encoder-decoder mexanizmida noto‘g‘ri mapping hosil bo‘lishiga olib keladi. Bu esa semantik yo‘qotish va tarjima aniqligining pasayishiga sabab bo‘ladi. Shuningdek, morfologik jihatdan boy tillarda, masalan o‘zbek va turk tillarida, balanslanmagan korpus data sparsity muammosini kuchaytiradi. Kam uchraydigan so‘z shakllari model tomonidan yetarlicha o‘rganilmaydi, bu esa tarjima sifatini pasaytiradi. Shu sababli zamonaviy NMT tizimlarida korpusni oldindan qayta ishlash (preprocessing), stratifikatsiya asosida balanslash va subword texnologiyalari (BPE, SentencePiece) keng qo‘llaniladi. Bu yondashuvlar modelning umumlashtirish qobiliyatini oshirib, tarjima aniqligini sezilarli darajada yaxshilaydi.

Foydalanilgan adabiyotlar ro‘yxati

1. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. International Conference on Learning Representations (ICLR), 2015, pp. 1–15.

2. Baker M. Corpora in Translation Studies. Target, 1995, Vol. 7, No. 2, pp. 223–243.

3. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2012, pp. 1–294.

4. McEnery T., Xiao R., Tono Y. Corpus-Based Language Studies: An Advanced Resource Book. Routledge, 2006, pp. 1–386.

5. Sinclair J. Corpus, Concordance, Collocation. Oxford University Press, 1991, pp. 1–179.

6. Sinclair J. Trust the Text: Language, Corpus and Discourse. Routledge, 2004, pp. 1–224.