IQRO JURNALI , IQRO JOURNAL, ИКРО ЖУРНАЛИ
ISSN:2181-4341| wordlyknowledge.uz/index.php/iqro
Volume 22 Issue 01 IYUN 2026

DIPLOMATIK MATNLAR KORPUSINI SHAKLLANTIRISHNING LINGVISTIK VA TEXNIK MEZONLARI

Usmonova Mexrikamol Odiljonovna
Andijon davlat chet tillari instituti
1 kurs 10.00.11 – “Amaliy va kompyuter lingvistikasi”
sohasi tayanch doktaranti
Email: usmonovamehrikamol@gmail.com

Annotatsiya

Ushbu maqolada zamonaviy kompyuter tilshunosligining dolzarb yo‘nalishlaridan biri bo‘lgan diplomatik matnlar korpusini loyihalash, uning lingvistik va texnik mezonlarini ishlab chiqish masalalari tadqiq etilgan. Tadqiqot doirasida korpusga matnlarni saralash prinsiplari, muvozanatlanganlik va reprezentativlik mezonlari, shuningdek, diplomatik terminlarning o‘ziga xos xususiyatlarini saqlagan holda matnlarga meta va lingvistik belgi qo‘yish (annotation) tizimi IMRAD formati talablari asosida tahlil qilingan. Maqola yakunida o‘zbek-ingliz parallel diplomatik korpusini yaratishning konseptual modeli taklif etilgan.

Kalit so‘zlar: korpus lingvistikasi, diplomatik matnlar korpusi, lingvistik mezonlar, texnik mezonlar, reprezentativlik, tokenizatsiya, lemmatizatsiya, XML format.

Abstract

This article investigates the issues of designing a diplomatic text corpus and developing its linguistic and technical criteria, which is a topical direction in modern computational linguistics. Within the research, the principles of text selection, balance, and representativeness criteria, as well as the system of meta and linguistic annotation while preserving the specific features of diplomatic terms, are analyzed based on the IMRAD format requirements. At the end of the article, a conceptual model for creating an Uzbek-English parallel diplomatic corpus is proposed.

Keywords: corpus linguistics, diplomatic corpus, linguistic criteria, technical criteria, representativeness, tokenization, lemmatization, XML format.

Introduction (Kirish)

Bugungi kunda jahon tilshunosligida raqamli texnologiyalar va sun’iy intellekt tizimlarining rivojlanishi matnlarni avtomatik tahlil qilish, mashina tarjimasi va raqamli leksikografiya sohalari oldiga yangi vazifalarni qo‘ymoqda. Xususan, muayyan soha diskursiga oid matnlar korpusini yaratish ixtisoslashgan terminologiyani tizimlashtirishning eng samarali usuli hisoblanadi. Diplomatiya sohasi o‘zining qat’iy qoliplari, xalqaro huquqiy normalari va lisoniy etiketi bilan ajralib turuvchi murakkab tizimdir.

O‘zbek tili jahon sahnasida munosib o‘rin egallayotgan va xalqaro aloqalar faollashayotgan hozirgi davrda o‘zbek va ingliz tillaridagi diplomatik hujjatlarning parallel hamda qiyosiy korpuslarini yaratish ehtiyoji yuqori. Biroq, diplomatik matnlar korpusini shakllantirishda qaysi xalqaro standartlarga tayanish, matnlarni qanday lingvistik filtrlardan o‘tkazish va ularga raqamli ishlov berish (markup) masalalari tizimli tadqiq etilmagan. Ushbu maqolaning maqsadi diplomatik matnlar korpusining lingvistik va texnik mezonlarini ilmiy asoslash hamda ularning takomillashtirilgan modelini ishlab chiqishdan iborat.

Methods (Metodlar)

Tadqiqot jarayonida korpus lingvistikasining tizimli modellashtirish, distributiv va statistik tahlil metodlaridan foydalanildi. Diplomatik matnlarning o‘ziga xos xususiyatlarini aniqlash uchun quyidagi metodologik bosqichlar amalga oshirildi:

1. **Matnlarni tanlash va filtrlash metodi:** O‘zbekiston Respublikasi Tashqi ishlar vazirligi rasmiy bayonotlari, xalqaro shartnomalar, konvensiyalar va memorandumlar matnlari tahlil obyekti sifatida olindi.

2. **Lingvistik taksonomiya:** Matnlardagi diplomatik terminlar, klišelar va barqaror birikmalarni ajratish uchun struktur-funksional tahlil qo‘llandi.

3. **Texnik standartlashtirish:** Korpus arxitekturasini loyihalashda xalqaro **TEI (Text Encoding Initiative)** va **LMF (Lexical Markup Framework)** standartlari metodologik asos qilib olindi.

Results (Natijalar)

Olib borilgan tadqiqot natijasida diplomatik matnlar korpusini shakllantirishning ikki asosiy ustuni – lingvistik va texnik mezonlar tasnifi ishlab chiqildi.

1. Lingvistik mezonlar

Diplomatik korpusning lingvistik asosini matnlarning **reprezentativligi** (til tizimini to‘liq aks ettira olishi) va **muvozanatlashganligi** tashkil etadi. Diplomatik matnlar quyidagi janriy mezonlar asosida tasniflandi:

- **Rasmiy-shartnomaviy matnlar:** Konvensiyalar, xalqaro bitimlar, paktlar. Ushbu matnlarda terminologik qat’iyat va bir ma’nolilik (monosemiya) yuqori.
- **Hujjatli-diplomatik matnlar:** Notalar, memorandumlar, rasmiy maktublar. Bu yerda diplomatik etiket va klišelar ustunlik qiladi.
- **Deklarativ-axborot matnlari:** Sammit deklaratsiyalari, qo‘shma bayonotlar, TIV press-relizlari. Ushbu matnlarda neologizmlar va ijtimoiy-siyosiy leksika ko‘p uchraydi.

2. Texnik mezonlar

Texnik mezonlar matnlarni kompyuter o‘qiydigan formatga keltirish va ularni metadata bilan ta’minlash jarayonini qamrab oladi. Diplomatik matnlar bazasini XML va JSON formatlariga o‘tkazish algoritmi yaratildi. Matnlarga ishlov berish quyidagi uch bosqichli model asosida amalga oshirildi:

- **Tokenizatsiya (Tokenization):** Matndagi so‘z shakllari va tinish belgilarini alohida birliklarga ajratish. Diplomatik matnlardagi ko‘p so‘zli terminlar (masalan, *"favqulodda va muxtor elchi"*) yaxlit token sifatida belgilanishi texnik jihatdan asoslandi.
- **Lemmatizatsiya (Lemmatization):** So‘z shakllarini ularning boshlang‘ich (lug‘at) shakliga keltirish.
- **Morfolojik belgilash (POS-tagging):** Har bir so‘zning so‘z turkumi va diplomatik terminologik funksiyasini aniqlovchi teglarni (tags) biriktirish.

Discussion (Munozara)

Diplomatik matnlar korpusini yaratishda boshqa sohaviy (masalan, tibbiy yoki iqtisodiy) korpuslardan farqli o‘laroq, bir qancha o‘ziga xos muammolar yuzaga keladi. Birinchidan, diplomatik matnlarda ochiqlik va maxfiylik muammosi mavjud. Shu sababli, korpusga faqat ochiq manbalardagi rasmiy hujjatlarni kiritish texnik mezon sifatida belgilandi.

Ikkinchidan, o‘zbek tilidagi diplomatik terminlarning aksariyati ingliz va rus tillaridan kalkalash (so‘zma-so‘z tarjima) qilish orqali kirib kelgan. Masalan, *"persona non grata"*, *"status-kvo"*, *"demash"* kabi terminlar korpusda maxsus teglashni talab qiladi. Biz taklif etayotgan TEI-kodlash tizimida chet tildan kirgan xalqaro terminlar uchun <term type="international"> tegi, milliy ekvivalentlar uchun esa <term type="national"> tegi qo‘llanishi diplomatik diskursni aniqroq tahlil qilish imkonini beradi. Bu esa xorijiy tillar (ingliz tili) kafedralarida talabalarga diplomatik tarjimani o‘rgatishda lingvistik xatolarning oldini olishga xizmat qiladi.

Conclusion (Xulosa)

Diplomatik matnlar korpusini shakllantirish shunchaki matnlar to‘plamini yaratish emas, balki tilning xalqaro munosabatlardagi funksional tabiatini raqamlashtirishdir. Tadqiqot davomida aniqlangan lingvistik (janriy muvozanat, etiket munosabatlari) va texnik (XML standartlari, ko‘p bosqichli teglash) mezonlar kelajakda mukammal **Inglizcha-O‘zbekcha parallel diplomatik korpus** yaratishning dasturiy asosi bo‘ladi. Ushbu model asosida shakllantirilgan korpus nafaqat kompyuter tilshunosligi, balki ingliz tili leksikologiyasi, tarjima nazariyasi va xalqaro munosabatlar mutaxassislarini tayyorlashda muhim amaliy manba bo‘lib xizmat qiladi.

References (Foydalanilgan adabiyotlar)

1. O‘zbekiston Respublikasi Vazirlar Mahkamasining "O‘zbek tili milliy korpusini yaratish chora-tadbirlari to‘g‘risida"gi qarori. – Toshkent, 2022.
2. Sinclair, J. (2004). *Developing Linguistic Corpora: a Guide to Good Practice*. Lexicography and Corpus Linguistics.
3. Ide, N., & Pustejovsky, J. (2010). *Handbook of Linguistic Annotation*. Springer Science & Business Media.
4. Mamedov, A. (2023). *Kompyuter tilshunosligi va raqamli leksikografiya asoslari*. Toshkent: Fan.