# TIL KORPUSI MATNLARI CHASTOTASI, KONKORDANS VA KWIC

**Yuldashev Aziz**
yuldashevaziz@navoiy-uni.uz
ToshDO‘TAU o‘qituvchisi

**Annotatsiya.** NLPga asoslangan ilovalar(axborot tizimlari)ni ishlab chiqish uchun NLP tizimni mavjud ma’lumotlarni o‘rganishini ta’minlash kerak. Ushbu amalni til korpusi vositasida amalga oshirish mumkin. Korpus matnlarini annotatsiyalash yoki teglash orqali so‘z, ibora, so‘z birikmasi shu kabi leksik birliklarni izohlash kabi matnga teg (izoh) qo‘shish va ularni belgilashdan iborat. Til korpusi foydalanuvchilar uchun to‘liq foydali bo‘lishi uchun uni teglash kerak. Bugungi kunda dunyodagi mashur korpuslar turli mezonlar bo‘yicha teglangan. Lingvistik teglash – bu qaror qabul qilish maqsadida kompyuterda o‘qiladigan ma’lumotlarni uning ma’nosiga bog‘lash jarayoni. Texnik jihatdan, u tildagi murakkab naqshlarni aniqlash uchun hissiyotlarni tahlil qilish yoki NLP ilovalari tomonidan ishlatilishi mumkin bo‘lgan lingvistik metama’lumotlarga ega matnga izoh berishni o‘z ichiga oladi.

**Abstract.** To develop applications (information systems) based on NLP, it is necessary to ensure that the NLP system learns the available data. This can be done using a language corpus. Annotating or tagging corpus texts consists of adding a tag (explanation) to the text and defining them, such as explaining words, phrases, phrases, and similar lexical units. In order for a language corpus to be fully useful to users, it needs to be tagged. Today, the world's most famous corpora are tagged according to various criteria. Linguistic tagging is the process of linking computer-readable information to its meaning for decision-making purposes. Technically, it involves annotating text with linguistic metadata that can be used by sentiment analysis or NLP applications to identify complex patterns in language.

**Аннотация**. Для разработки приложений (информационных систем) на основе НЛП необходимо обеспечить, чтобы система НЛП обучалась имеющимся данным. Это можно сделать с помощью языкового корпуса. Аннотирование или разметка корпусных текстов заключается в добавлении к тексту тега (пояснения) и его определении, например пояснении слов, словосочетаний, словосочетаний и подобных лексических единиц. Чтобы языковой корпус был полностью полезен пользователям, его необходимо пометить. Сегодня самые известные корпорации в мире маркируются по различным критериям. Лингвистическая разметка — это процесс связывания читаемой компьютером информации с ее значением для целей принятия решений. Технически это включает в себя аннотирование текста лингвистическими метаданными, которые могут использоваться анализом

настроений или приложениями НЛП для выявления сложных языковых шаблонов.

**Kalit so‘zlar:** *Til korpusi, matnni qayta ishlash, lingvistik tahlil, mashinali o‘qitish modellari, korpus turlari, teglangan korpuslar, POS teg.*

## Kirish

So‘nggi bir necha yil ichida lingvistik tahlillar onlayn korpuslar orqali amalga oshirilmoqda. Masalan, Sketch Engine 38 milliard so‘zni o‘z ichiga olgan English Web 2020 (enTenTen20) kabi eng mashhur ochiq korpuslar xizmatini taqdim etadi. Shunisi e’tiborga loyiqki, ijtimoiy tarmoqlar va axborot tizimlaridagi katta hajmdagi strukturlanmagan ma’lumotlar faqat matn emas, balki turli formatlarda (audio, video va boshqalar) ham bo‘lishi mumkin. Natijada, lingvistik teglash matn tahlili bilan chegaralanib qolmaydi. Teglar turli formatlarga qo‘llanilishi mumkin: *transkripsiya*, *vaqt belgisi, nutq sharhi, ma’no teglari* va boshqalar [Bi, 2018].

Tabiiy tilni tushunishga (natural language understanding, NLU) asoslangan ko‘plab NLP ilovalarini ishlab chiqish uchun, til korpuslarini shakllantirish va ularni teglash lozim [Bender, Lascarides, 2019]. Shu sababli mualliflar tomonidan o‘zbek tili korpusini lingvistik teglash turlari hisoblangan, *fonetik, morfologik, sintaktik* va *semantik teglar*ning o‘zbek tiliga mos to‘plami shakllantirildi hamda korpusda qo‘llandi.

Strukturlanmagan matnlarni tozalash NLP vazifasi orqali matn tahlil qilish sifati va aniqligini oshirish uchun muhim qadamdir. Matnni tozlash jarayoni: imlo va formatlashdagi nomuvofiqliklarni bartaraf etish orqali matnni kichik harflarga aylantirish bilan birga maxsus belgi, raqam va nomuhim so‘zlar kabi ahamiyatsiz (yoki ortiqcha) ma’lumotlarni olib tashlashni o‘z ichiga oladi. Matnni tozalash, shuningdek, imlo xatolarni qayta ishlash, so‘zlarni o‘zak shakliga keltirish (lemmatizatsiya) va matnni kodlash muammolarini hal qilish yechimlarini taklif qiladi. Ijtimoiy tarmoqlar, onlayn servislarda hosil qilinadigan katta hajmdagi ma’lumotlar, matn terishdagi xatolik, nomuvofiq so‘zlar, sheva unsurlarini misol sifatida keltirish mumkin. Tozalanmagan matnlar NLP modeli ish jarayoniga salbiy ta’sir o‘tkazadi.

*Korpus chastotasi*

Tilshunoslik – bu statistika va matematika sohasi bilan uzviy bog‘liq bo‘lgan fan. Matematik tilshunoslik, kompyuter lingvistikasi, korpus lingvistikasi, amaliy tilshunoslik, sud lingvistikasi, stilometriya va boshqalar tabiiy til korpusidan olingan turli statistik natijalarni talab qiladi. Tilning turli xossalari haqidagi statistik ma’lumotlarni yetarli darajada bilmaslik natijasida lingvistik ma’lumotlar bilan ishlashda ham, kuzatishda ham xatoga yo‘l qo‘yilishi mumkin.

Til korpusi ham miqdoriy, ham sifat jihatidan tahlil qilinishi mumkin.

**Miqdoriy tahlilda** tilning turli lingvistik xususiyatlari tasniflanadi va murakkabroq statistik modellar ishlab chiqiladi. Statistik modellar asosida qaysi

hodisalar til yoki turli xil xatti-harakatlarning haqiqiy aks etishi mumkinligini va qaysilari shunchaki tasodifiy hodisalar ekanligini aniqlash imkonini beradi.

**Sifat tahlili** miqdoriy tahlil orqali korpusda kuzatilgan hodisalarning to‘liq va batafsil ta’rif berishga qaratilgan. Korpusning sifat tahlili aniq xulosalarni chiqarishga imkon beradi. Chunki bu holda ma’lumotlarni cheklangan tasniflar to‘plamiga kiritish shart emas. Ham miqdoriy, ham sifat tahlillarini amalga oshirish natijasida korpusni tahlil qilish samaradorligi oshadi. Biroq miqdoriy tahlil evaziga olingan natijalar sifat tahlilidan olingan natijalarga qaraganda samaradorligi pastroq hisoblanadi.

Bugungi kunda korpus ustida miqdoriy tahlilni amal oshirishning quyida keltirilgan statistik yondashuvlari mavjud:

– *oddiy tavsiflovchi statistik yondashuv;*
– *inferensial statistik yondashuv;*
– *baholovchi statistik yondashuv.*

*Oddiy tavsiflovchi statistik yondashuv* kuzatilgan ma’lumotlarning eng muhim xususiyatlarini umumlashtirish imkonini beradi. *Inferensial statistik yondashuv* savollarga javob berish yoki gipotezani shakllantirish uchun tavsiflovchi statistik yondashuvdan olingan ma’lumotlardan foydalanadi. *Baholovchi statistik yondashuv* gipoteza ma’lumotlardagi dalillar bilan tasdiqlanganligini va matematik model yoki ma’lumotlarning nazariy taqsimoti haqiqatga qanday bog‘liqligini tekshirishda ahamiyatli.

Korpus ustida amalga oshirilgan miqdoriy tahlil natijalarini taqqoslash va yashirin shablonlarni aniqlash uchun quyida keltirilgan ko‘p o‘lchovli statistik usullardan foydalaniladi:

– *omillar tahlili (Factor Analysis);*
– *ko‘p o‘lchovli masshtablash (Multidimensional Scaling);*
– *klaster tahlili (Cluster Analysis);*
– *log-linear modellari (Log-linear Models).*

*So‘zlarni tartiblash/saralash*

Korpusda saqlangan so‘zlarni ikki usulda tartiblash mumkin. **Raqamli saralash** jarayoni miqdoriy ma’lumotlar bilan ishlashning eng oddiy usuli hisoblanadi. Ushbu usul orqali saralashda obyektlar ma’lum bir sxema bo‘yicha tasniflanadi hamda sxemadagi har bir sinfga tegishli bo‘lgan *matnlar ichidagi elementlarning soni bo‘yicha arifmetik hisoblash* amalga oshiriladi. Oddiy chastotalarni hisoblashda mavjud bo‘lgan ma’lumotlar alifbo tartibida yoki raqamli tartibda foydalanuvchiga taqdim etiladi. Hosil qilingan ro‘yxat foydalanuvchi talabiga ko‘ra *o‘sish* yoki *kamayish* tartibida joylashtirilishi mumkin. Matn tahlili jarayonida har xil element (so‘z, token, so‘z birikmasi, lemma)ning necha martadan uchrashi aniqlanadi.

Tahlil natijasida **so‘zlarning chastota ro‘yxati** hosil qilinadi. Ro‘yxat o‘rganib chiqish matn tuzilishi haqida tasavvur hosil qiladi; shunga mos ravishda tahlil rejalashtiriladi. Shuningdek, alifbo tartibida tartiblangan so‘zlar ro‘yxati oddiy umumiy havolalar uchun ishlatilishi mumkin. Alfavit tartibidagi chastotalar ro‘yxati o‘rganish obyekti sifatida ahamiyatli, chunki u ko‘pincha tekshirilishi kerak bo‘lgan gipotezalarni shakllantirish va qaror qabul qilishga yordam beradi.

O‘zbek tili korpusida chastotani hisoblashdan oldin korpusda ishlatiladigan *belgi, so‘z, idioma, ibora, fraza* va *gaplar* bilan ishlash jarayoni haqida mulohaza yuritish kerak. Ushbu amal bizni o‘zbek tilining turli lingvistik xususiyatlari haqida noto‘g‘ri kuzatish va xulosalardan saqlaydi.

*Konkordans*

Konkordans tuzish jarayoni korpusda ishlatiladigan so‘zlarni indekslashni anglatadi. Ushbu jarayonda har bir so‘z matn(lar)da kelgan joyiga qarab indekslanadi. Konkordans tuzish jarayoni til modellarini qurishda muhim ahamiyat kasd etadi va tabiiy tildagi qoliplarni aniqlash imkonini beradi.

Bugungi kunda korpusni tahlil qilish uchun konkordans dasturlari ishlab chiqilgan. Masalan, korpusni tartiblash va chastotasini aniqlash uchun **MonoConc, Conc**; matnlarni parallel qayta ishlash uchun **ParaConc**, matnlarni qayta ishlash va saralash uchun **FreeText** kabi dasturiy ta’minotlar mavjud. Konkordans funksiyasi, asosan, leksikografik tadqiqotlar va til o‘rgatishda qo‘llanadi. Shuningdek, undan bir va ko‘p so‘zli satrlar, so‘z, ibora, idioma, maqol va boshqa leksik birliklarni qidirishda foydalanish mumkin.

Konkordans jarayoni **leksik, semantik, sintaktik**, **matn** modellarini tadqiq qilish va matnlarning uslubiy qonuniyatlarini o‘rganishda ishlatiladi. Bu ko‘p ma’noli, polifunksional va polisemantik so‘zlar va morfemalarni tekshirishda muhim vosita bo‘lib xizmat qiladi.

*Leksik kollokatsiya (Erkin birikma)*

So‘zlarni qo‘shish usuli matndagi so‘zlarning o‘rni va qurshovini tushunishga yordam beradi. Korpusdan aniqlangan so‘z birikmalari tilda mavjud ba’zi an’anaviy lingvistik tavsif va farazlardan farq qiladi. Leksikografiya bo‘yicha amalga oshirilgan tadqiqotlar shuni ko‘rsatadiki, ko‘p ishlatiladigan faol so‘zlar ma’nosi bizning ongimizga birinchi kelgan va lug‘atlarda keltirilgan ma’no bilan aynan emas. Demak, korpusdan aniqlangan so‘z birikmalari o‘rtasida *leksik kollokatsiya*ni aniqlashga yordam beradi. Bunda ikki so‘zning yonma-yon qo‘llanish ehtimoli aniqlanadi.

Har bir so‘z juftligiga ball beriladi. Ball qancha yuqori qimmatli bo‘lsa, so‘zlarning birikuvchanligi shunchalik yuqori sanaladi. Bu amal tilshunoslikdagi *leksikografiya* va *texnik tarjimada* foydalanish uchun korpusdan so‘z birliklarini ajratib olish imkonini beradi. Shuningdek, mazkur amal orqali ma’no o‘zgarishini aniqlash maqsadida o‘xshash so‘zlarni aniqlash mumkin. Masalan, **daryo qirg‘og‘i = landshaft**, **ustozi piri komil = muallim**.

Bu esa ma’nosi o‘xshash so‘zlar o‘rtasidagi foydalanishdagi farqlarni ajratish imkonini yaratadi. Til ta’limida bunday ma’lumotlar ikki o‘xshash so‘z o‘rtasidagi tafovutlar haqidagi bilimlarni olish, shu bilan birga, tilni yaxshiroq o‘rganishda muhim rol o‘ynaydi. Turli lingvistik elementlarning (masalan, so‘z, morf, idioma va boshqalar) birikmalari haqidagi ma’lumotlar til ta’limidan tashqari *lug‘atlarni ishlab chiqish, NLP vazifalari* va *mashina tarjimasi* uchun muhim. Biroq qaysi so‘z birikmasi muhim ekanligini aniqlash oson jarayon emas.

*Kontekstdagi kalit so‘z (Key-Word-In-Context, KWIC)*

KWIC funksiyasi korpusni qayta ishlashda mayyan so‘zning hujjatlarda mavjudligini qidirishda keng qo‘llaniladi. Tekshirilayotgan so‘z har bir satrning o‘rtasida paydo bo‘lib, ikkala tomonda qo‘shimcha qismlardan tashkil topadi. NLP vafizasiga ko‘ra, kontekstning uzunligi turli xil bo‘lishi mumkin. U markazda so‘zning har ikki tomonida ikki, uch yoki to‘rtta so‘zdan iborat qism (n-gram)ni ko‘rsatadi. So‘z, ibora va gaplarni tahlil qilishda yaxshiroq tushunish uchun qo‘shimcha kontekstga ehtijoz seziladi. KWICni o‘z-o‘zidan matn sifatida talqin qilish va **markaziy so‘z** qurshovidagi so‘zlarning chastotasini o‘rganish ahamiyatli.

Til korpusi KWIC vosidasida tahlil qilganidan so‘ng lingvistik tavsifda turli vazifalarni hal qilish maqsadida turli amallarni ishlab chiqish mumkin.

KWIC kontekstning ahamiyati, assotsiativ so‘zlarning roli, kontekstdagi so‘zlarning xarakteristikasi, qurshovdagi so‘zlar haqidagi ma’lumotlar va so‘zdan foydalanishda kontekstli cheklovlar mavjudligini tushunishga yordam beradi.

*So‘zlarni guruhlash (Local Word Grouping, LWG)*

LWG – bu matnlarni tahlil qilish yoki korpusni qayta ishlashning yana bir turi bo‘lib, u matnlardagi so‘zlar guruhidan foydalanish shablonlarini aniqlab beradi. Bu usul gapning ma’nosini aniqlashda so‘z tartibi muhim bo‘lgan NLP vazifalarini hal qilishda foydalaniladi. LWG usuli so‘z birikmalari va gaplar darajasida tahlil qilish vaqtida tarkibiy qismlarning funksional xatti-harakatlari haqida ma’lumot beradi. LWG usuli unchalik keng tarqalgan bo‘lmasa-da, refleksiv shakllarga o‘ziga xos yaqinlikka ega bo‘lgan fe’l shakllarining (masalan, o‘yin-kulgi, o‘z-o‘zini xursand qilish, qarzga berish, eslatish va h.) taqsimlanishini belgilaydi.

Bunday namuna(shablon)larni bilish til o‘rganuvchilarni o‘rta darajadan yuqori malaka darajasiga o‘tkazishda muhim ahamiyat kasb etadi.

LWG usuli o‘zbek tili korpusi matnlaridagi ma’lumotlar asosida fe’lli birikma va otli birikmalarini (yoki guruh) tahlil qilishga yordam beradi. LWG ma’lumotlari turli leksik elementlarning assotsiatsiyasidan kelib chiqadigan leksik noaniqlikni bartaraf etishga yordam beradi.

O‘zbek tili korpusi ustida olib borilgan tajribalar shuni ko‘rsatadiki, omonim so‘zlarning ma’nosini aniqlashda LWG usulidan foydalanish yuqori samaradorlikni ta’minladi. Ko‘pgina qo‘shma fe’l, sifat va otlar uchun ma’lum bir so‘z birikmasi bilan ifodalangan ma’noni alohida so‘zlarning ma’nolaridan olish mumkin emas.

## Xulosa

Bugun axborot asrida tabiiy tillarning tuzilishini o‘rganuvchi NLP yaqin kelajakda fizika, biologiya kabi juda muhim sohaga aylanishi kerak, chunki axborot tuzilmasi, asosan, tabiiy tillar birliklaridan iborat. Fizika dunyodagi jismoniy harakatlar qonunlarini o‘rgansa, NLP tabiiy tillar qoidalarini, axborot tashuvchilarni, axborot tarmoqlari dunyosini o‘rganadi. Tabiiy tilni kompyuter yordamida tadqiq qilish va qayta ishlash, odatda, to‘rt jihatni: formallashtirish, algoritmlash, dasturlash va amaliyotga qo‘llashni o‘z ichiga oladi. Korpus matnlarini lingvistik teglash dastlab lingvistik nazariyalarni yoki bugungi kunda ma’lum bo‘lganidek, til korpuslarni ishlab chiqish va tahlil qilish uchun ma’lumot berish uchun amalga oshirildi. Ushbu maqolada til korpusi matnlari chastotasi, konkordans va KWIC texnologiyasi haqida fikr mulohaza yuritildi.

## Foydalanilgan adabiyotlar:

1. Bi, P. (2018). Handbook of Linguistic Annotation. *Journal of Quantitative Linguistics*. https://doi.org/10.1080/09296174.2018.1424495

2. Bender, E. M., & Lascarides, A. (2019). Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics. In Synthesis Lectures on Human Language Technologies (Vol. 12, Issue 3). https://doi.org/10.2200/S00935ED1V02Y201907HLT043

3. Boltayevich, E. B., Mirdjonovna, H. S., & Ilxomovna, A. X. (2023). Methods for Creating a Morphological Analyzer. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13741 LNCS. https://doi.org/10.1007/978-3-031-27199-1_4

4. Elov B.B., Hamroyeva Sh.M., Xusainova Z.Y. The pipeline processing of NLP // E3S Web of Conferencesç INTERAGROMASH. Rossiya, Rostov, 2023. https://doi.org/10.1051/e3sconf/202341303011

5. Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash // “O‘zbek amaliy filologiyasi istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiyasi – Toshkent, 2022. №.1. – B.154-163.

6. Garside, R. G., Leech, G., & Mcenery, A. M. (2020). Introducing corpus annotation. In Corpus Annotation. https://doi.org/10.4324/9781315841366-7

7. Hovy, E., & Lavid, J. (2010). Towards a ‘ Science ’ of Corpus Annotation : A New Methodological Challenge for Corpus Linguistics. International Journal of Translation, 22(1).

8. Pustejovsky, J., & Stubbs, a. (2013). Natural language annotation for machine learning. In Vasa.

9. Sheng, D. (2023). Statistics in corpus linguistics: a practical guide. Social Semiotics, 33(4). https://doi.org/10.1080/10350330.2021.1969215