

TABIIY TILNI QAYTA ISHLASH(NLP)DA KORPUS TEXNOLOGIYASINING O'RNI

Raximov Xasanboy Komiljonovich

ADU 3-kurs tayanch doktoranti

Namangan davlat chet tillar instituti stajyor o'qituvchisi

xasanraximov@adu.uz

<https://doi.org/10.5281/zenodo.14718241>

Annotatsiya: Maqolada NLP va korpus texnologiyalari o'rtasidagi o'zaro bog'liqlik ham tahlil qilingan. Korpus texnologiyalari NLP uchun zarur bo'lgan ma'lumot va vositalarni taqdim etadi, NLP esa ushbu ma'lumotlarni real amaliyotlarda qo'llaydi. Ushbu integratsiya lingvistika va texnologiya sohalarida yangi imkoniyatlarni ochib beradi. Korpus texnologiyalarining lisoniy qoliplar, grammatik qoidalar va lug'atlarni rivojlantirishdagi rolini yoritib, uning ontologik bilimlar bazasi, semantik tarmoqlar va sun'iy intellekt tizimlarida qo'llanilishiga urg'u beradi. Shuningdek, tilshunoslik tadqiqotlarida maxsus korpuslardan foydalanish statistik natijalarga erishishning kvantitativ yondashuv sifatida taqdim etiladi. Korpus texnologiyalarining asosiy vositalari, masalan, AntConc, konkordansiya vositalari, kollokatsiya grafiklari va n-gram tahlillari matnning yuzaki xususiyatlarini tahlil qilish uchun samarali hisoblanadi.

Kalit so'zlar: tabiiy tilni qayta ishlash (NLP), korpus texnologiyalari, lingvistik modellash, grammatik tahlil, sun'iy intellekt, semantik tahlil, statistik tahlil, konkordansiya vositalari, avtomatik tarjima, sentiment tahlil.

Annotation: The article also analyzes the interrelation between NLP and corpus technologies. Corpus technologies provide essential data and tools for NLP, while NLP applies this data in real-world practices. This integration opens new opportunities in the fields of linguistics and technology. The article highlights the role of corpus technologies in the development of linguistic patterns, grammatical rules, and dictionaries, emphasizing their application in ontological knowledge bases, semantic networks, and artificial intelligence systems. Additionally, the use of specialized corpora in linguistic research is presented as a quantitative approach to achieving statistical results. Core tools of corpus technologies, such as AntConc, concordance tools, collocation graphs, and n-gram analyses, are considered effective for analyzing the surface features of texts.

Keywords: Natural Language Processing (NLP), corpus technologies, linguistic modeling, grammatical analysis, artificial intelligence, semantic analysis, statistical analysis, concordance tools, machine translation, sentiment analysis.

Tabiiy tilni qayta ishlash (NLP)da korpus texnologiyasi zamonaviy lingvistika va kompyuter texnologiyalarining kesishgan nuqtasida joylashgan muhim sohadir. Korpus texnologiyasi kompyuter lingvistikasining asosiy obyekti va predmeti sifatida matnlarni qayta ishlashda modeli vazifasini bajaradi. Korpus texnologiyasi tilshunoslikda muhim ahamiyatga ega bo'lib, turli olimlar tomonidan ta'riflangan. Masalan, O'zMU professori Nilufar Abdurahmonova o'zining "Korpusga asoslanib o'zbek tilini xorijiy til sifatida o'qitish" va "O'zbek tilida sinonimiya hodisasini o'qitishda korpusdan foydalanish" nomli maqolalarida korpusni elektron matnlarning lingvistik ma'lumotlar bazasi sifatida ta'riflaydi. Ushbu ta'rifga ko'ra, korpus turli uslubga tegishli og'zaki va yozma matnlarni o'z ichiga oladi. Elektron

korpusning tizimlashtirilgan va ma'lum me'yorlarga asoslangan tasnifiy matnlari yordamida lisoniy qoliqlar, grammatik qoidalar, lug'atlarni yanada takomillashtirishga hissa qo'shadi. Tilning pragmatik va kognitiv tasviri aks etgan og'zaki va yozma matnlar majmuasi (korpus) ontologik bilimlar bazasi, semantik va neyro to'rlar va sun'iy intellekt texnologiyasi hamda lingvoprotsessor uchun tilning lisoniy modellari va nutqiy aktlarni o'rganishda katta rol o'ynaydi. Filologiyaning muayyan sohasida ilmiy tadqiqotlar olib borishda maxsus korpuslar yaratishda lingvistik instrumentlaridan foydalanish statistik natijalarga erishishning kvantativ usuli bo'lib xizmat qiladi. Matnlarni tahlil qilishda muayyan turdagi instrumentariylar korpus foydalanuvchilarning maqsad va vazifalaridan kelib chiqib individual foydalanuvchilar interfeysi va korpus menedjeriga ega bo'ladi. Parallel matnlar uchun segmentlash va lingvistik tahlil jarayonida Wordfast kabi instrumentlar samarali texnologik vosita sanaladi. Parallel matnlardagi konkordanslarni aniqlashda so'z, so'z birikmasi yoki barqaror birikmalarning u yoki bu tildagi muqobil ekvivalentligini aniqlashda tayyor lingvistik instrumentariylarga tayanish muhim. Parallel matnlardagi kalit so'zlar uchun kontekstda tez-tez qo'llanadigan birliklarni tarjimon xotiraga yuklash orqali parallel matnlarning qidiruv tizimini yaratishda foydalaniladi¹. Korpus lingvistikasida hisoblash vositalari ularning funksiyalari va qo'llanilishiga qarab bir necha turga bo'linadi. Birinchi navbatda, asosiy korpus vositalari, masalan, AntConc² va kontekstdagi kalit so'zlar (KWIC)³ foydalanuvchilarga so'z yoki iboralarni qidirish va ularning yaqin kontekstda har bir holatini ko'rish imkonini beradi. Ular so'zlar va iboralar ro'yxatlarini, chastotali ro'yxatlarni va boshqa vizualizatsiya usullarini yaratadi. Ushbu dasturiy ta'minotning maqsadi matn ma'lumotlarini korpus statistikasi standartlariga asoslangan holda chuqur tahlil qilishdan iborat. Misollarga konkordansiya vositalari, kollokatsiya grafiklari, tarmoq vositalari, tavsif vositalari, so'zlar ro'yxati vositalari, kalit so'zlar vositalari, n-gram vositalari, matn vositalari, GraphColl va Wizard kiradi⁴. Shunga qaramay, ushbu vositalar tabiiy tilni qayta ishlash tahlilini (NLP) qo'llab-quvvatlay olmaydi. Ular matnning yuzaki xususiyatlari bilan cheklanadi, masalan, konkordansiya, so'z va iboralarni qidirish, so'z chastotasi tahlili. Shunday bo'lsa-da, ular matnlarni qo'lda batafsil tahlil qilish uchun juda samarali. Shu sababli, tabiiy tilni qayta ishlash tahlili uchun muqobil hisoblanadi. Bu yanada xilma-xildir va kengroq qo'llash doirasiga ega bo'lib, avtomatik matnni qisqartirish, sentimental tahlil, tilni tarjima qilish va nutqni tanib olishgacha bo'lgan sohalarda ishlatiladi. NLP vositalari ko'pincha tuzilmasiz matn ma'lumotlarini qayta ishlash uchun murakkab algoritmlar va Microsoft Excel kabi dasturlarni ishlatadi. Bu vositalar yirik matnlarni qisqartirish, asosiy nuqtalar va ma'lumotlarni ajratib olish, va mashinaviy o'qitish hamda sun'iy intellekt (AI) orqali vaqt o'tishi bilan takomillashish imkonini beradi⁵. Shuningdek, ular vizualizatsiya ma'lumotlarini tahlil qilish, ma'lumotlar bazasi boshqaruv tizimlari

¹Abduraxmonova Nilufar Zaynobiddin qizi, Korpus lingvistikasi(darslik), Toshkent-2023,8-bet.

² D. Vojnović, "Key noun+noun collocations in the language of tourism: a corpus-based study of English and Serbian," ELOPE:English Language Overseas Perspectives and Enquiries, vol. 18, no. 2, pp. 51–68, 2021, doi: 10.4312/ELOPE.18.2.51-68.

³ J. Buts and H. Jones, "From text to data: mediality in corpus-based translation studies," Monografias de Traducción e Interpretación (MonTI), no. 13, pp. 301–329, 2021, doi: 10.6035/MonTI.2021.13.10.

⁴ A. Sarudin, H. F. M. Redzwan, and A. A. Tan, "The development of a medicinal nadir glossary through the # lancsbox 6.0 of wizard software," Malaysia Journal of Invention and Innovation, vol. 1, no. 1, pp. 73–84, 2022.

⁵Abhay A. Dande and Dr. M. A. Pund, "A review study on applications of natural language processing," International Journal of Scientific Research in Science, Engineering and Technology, vol. 10, no. 2, pp. 122–126, 2023, doi: 10.32628/ijrsrset2310214

(DBMS)⁶(masalan, Python'ning NLTK yoki spaCy)⁷, integratsiyalangan dasturlash muhitlari (IDEs), veb ma'lumotlarini yig'ish va chiqarish vositalari, nutqni qayta ishlash vositalari va statistik tahlil dasturlari orqali matnni chuqur tahlil qilish va tushunish imkonini beradi. NLPning eng mashhur vositalaridan ba'zilari quyidagilardir: Wordify, Mlphon, morfologik tahlilchi, Runyakitara vositasi, LexTutor, Coh-Metric, Linguistic Inquiry and Word Count (LIWC), UAM Corpus Tool, SketchEngine (SkE) , Wmatrix, MultiAzterTest, sublang korpusini tahlil qilish to'plami (SubCAT), EnvText, InLang, Berri Corpus Manager, UCREL Semantic Analysis System (USAS), PyMongo (MongoDB, Python texnologiyasi, Flask), LanksBox 5.1.2, LanksBox 4.5, NooJ platformasi va Bi-LSTM⁸. Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) texnologiyalari tilni chuqur tahlil qilishda sintaksis, semantika, pragmatika va diskurs kabi tilshunoslikning asosiy yo'nalishlarini qamrab oladi. Ushbu yo'nalishlar NLP tizimlarining funktsionallik doirasi va ularning qo'llanilish sohalari belgilovchi muhim komponentlardir. Sintaksis til birligi sifatida jumlaning grammatik tuzilishini tahlil qilib, so'zlarning bir-biri bilan qanday bog'liq ekanligini aniqlash imkonini beradi. Semantika, o'z navbatida, jumlaning ma'nosini tushunish va undagi so'zlar va iboralar o'rtasidagi semantik bog'lanishlarni ifodalashga qaratilgan. Pragmatika tilning kontekstual ma'nosini, ya'ni gapiruvchi yoki yozuvchi tomonidan maqsad qilinadigan xabarni tushunishni ta'minlaydi. Diskurs tahlili esa kengroq ko'lamda bir nechta jumalardan tashkil topgan matnni yoki nutqni izchil tahlil qilib, ular o'rtasidagi bog'lanishlarni aniqlaydi. NLP vositalari ushbu yo'nalishlarni birlashtirib, matnlarni avtomatlashtirilgan tarzda tahlil qilish, ma'nosini anglash va ularga javob berish imkoniyatlarini taqdim etadi. Buning natijasida NLP texnologiyalari dialog tizimlari, avtomatik tarjima, matnni toifalash, til modellashtirish va tavsiyalarni ishlab chiqish kabi murakkab va dinamik ilovalarda muvaffaqiyatli qo'llaniladi. Ushbu texnologiyalarning rivojlanishi sun'iy intellekt va tilshunoslik o'rtasidagi o'zaro bog'liqlikni kuchaytirib, inson-mashina o'zaro aloqasini yanada intuitiv va samarali qilmoqda. NLP modellari va til nazariyalarining birlashuvi ushbu sohada yangi kashfiyotlarga olib keldi. Ushbu yutuqlar tabiiy til bilan bog'liq vazifalar, til modellari, sintaktik tahlil, mashinaviy tarjima, sentimental tahlil va savollarga javob berish kabi yo'nalishlarda kuzatildi⁹. NLP vositalari bugungi ma'lumotlarga asoslangan dunyoda murakkab va dinamik ilovalar uchun kengroq va yanada rivojlangan imkoniyatlarni taqdim etadi¹⁰. O'z o'rnida har qanday yangi imkoniyat yangi izlanishlar orqali muvaffaqiyatlarga erishish kaliti bo'lib xizmat qiladi.

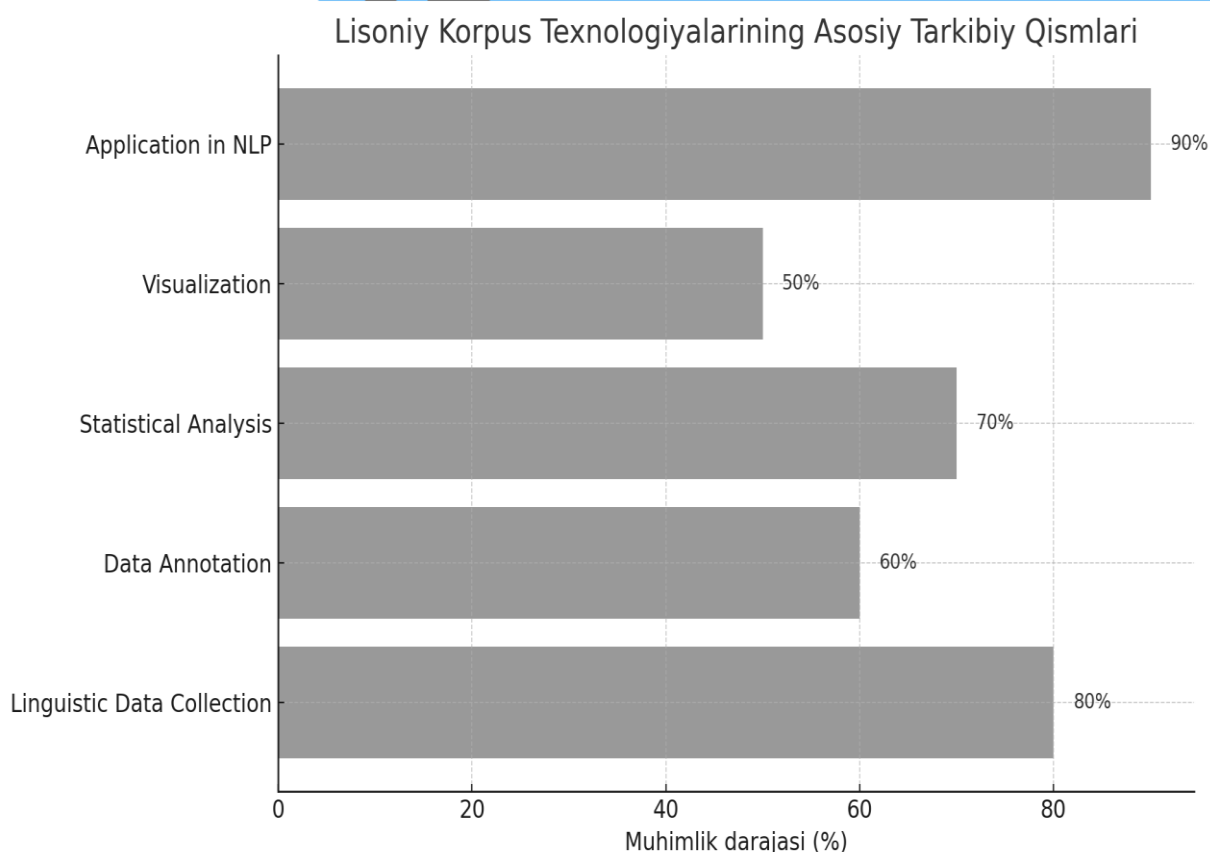
⁶ T. Gaillat et al., "A data repository for the management of dynamic linguistic datasets," in A data repository for the management of dynamic linguistic datasets, 2021.

⁷ X. Q. Xia, M. McClelland, and Y. Wang, "TabSQL: a MySQL tool to facilitate mapping user data to public databases," BMC Bioinformatics, vol. 11, pp. 2–5, 2010, doi: 10.1186/1471-2105-11-342.

⁸ (PDF) *Exploring corpus linguistics via computational tool analysis: key finding review*. Available from: https://www.researchgate.net/publication/380253006_Exploring_corpus_linguistics_via_computational_tool_analysis_key_finding_review#fullTextFileContent

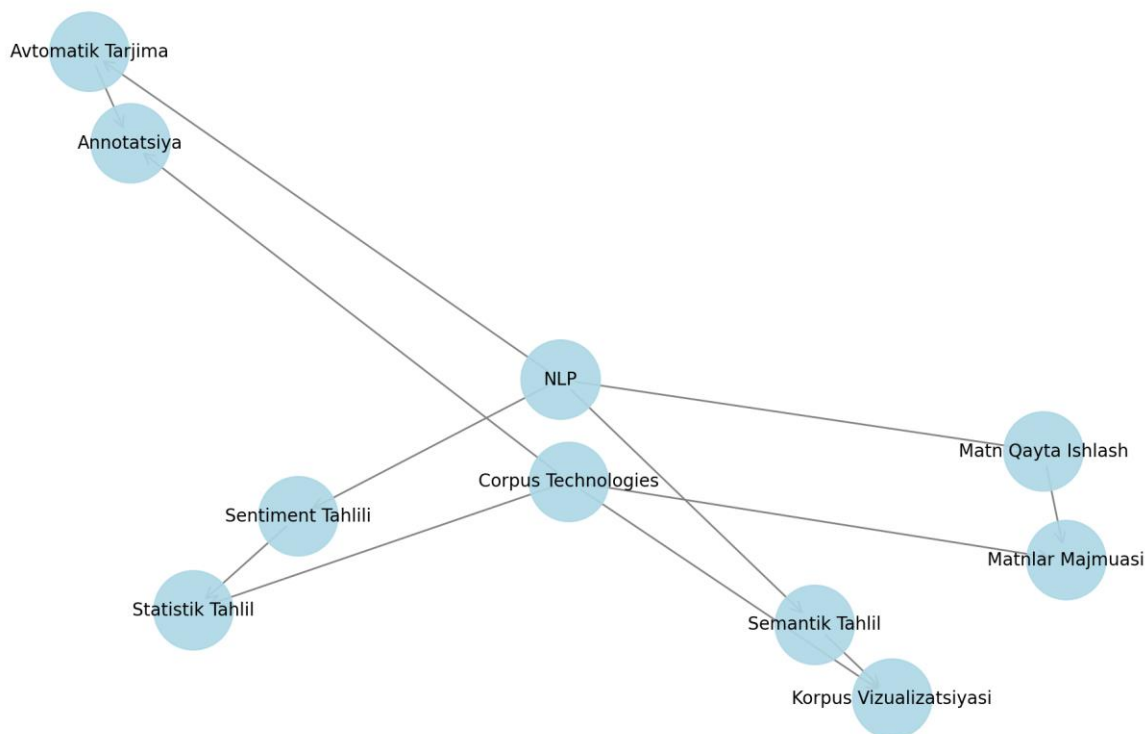
⁹ B. Li, "Integrating linguistic theory and neural language models," ArXiv Preprint, 2022. [Online]. Available: <http://arxiv.org/abs/2207.09643v1>.

¹⁰ P. Yao, M. Kosmajac, A. Waheed, K. Guzhva, N. Hervieux, and D. Barbosa, "NLP workbench: efficient and extensible integration of state-of-the-art text mining tools," EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations, pp. 18–26, 2023, doi: 10.18653/v1/2023.eacl-demo.3.



Yuqoridagi diagramma korpus texnologiyalarining asosiy tarkibiy qismlarini ko'rsatadi. Ushbu qismlar quyidagilardan iborat: Linguistic Data Collection - Korpus yaratish uchun lisoniy ma'lumotlarni yig'ish jarayoni. Data Annotation - Lisoniy ma'lumotlarni tasniflash va belgilash. Statistical Analysis - Korpusdagi ma'lumotlarni statistik jihatdan tahlil qilish. Visualization - Olingan natijalarni diagrammalar yoki grafiklar orqali ko'rsatish. Application in NLP - Olingan bilimlarni tabiiy tilni qayta ishlashda qo'llash. Bu diagramma korpus texnologiyalarining jarayonlarini bosqichma-bosqich tushunish uchun vizual ko'rsatkich sifatida xizmat qiladi.

NLP va Korpus Texnologiyalari o'rtasidagi Bog'liqlik



Yuqoridagi chizma NLP (Tabiiy Tilni Qayta Ishlash) va korpus texnologiyalari o'rtasidagi bog'liqlikni ko'rsatadi.

NLP yo'nalishlari: Matnni qayta ishlash, semantik tahlil, avtomatik tarjima, va sentiment tahlili kabi vazifalarni bajaradi.

Korpus texnologiyalari: Matnlar majmuasini yig'ish, statistik tahlil, annotatsiya, va vizualizatsiya orqali NLP uchun zarur bo'lgan ma'lumotlarni taqdim etadi.

O'zaro bog'liqlik: NLP vazifalari korpus texnologiyalari tomonidan yaratilgan ma'lumot va vositalarga asoslanadi, masalan, sentiment tahlili uchun statistik tahlildan foydalanish.

Tabiiy tilni qayta ishlash (NLP) va korpus texnologiyalari bir-birini to'ldiruvchi sohalar bo'lib, tilshunoslik va texnologiya kesishmasida muhim rol o'ynaydi. Korpus texnologiyalari statistik tahlil, annotatsiya va vizualizatsiya orqali NLP uchun asosiy bilim va ma'lumotlarni taqdim etsa, NLP korpuslardan olingan bilimlarni real vazifalar — matnni qayta ishlash, tarjima va sentiment tahlili kabi yo'nalishlarda amaliyotga tatbiq etadi. Ushbu integratsiya lingvistik tadqiqotlarda chuqur tahlil va yangi imkoniyatlarni yaratish uchun poydevor bo'lib xizmat qiladi.

References:

1. Abduraxmonova Nilufar Zaynobiddin qizi, Korpus lingvistikasi(darslik), Toshkent-2023,8-bet
2. D. Vojnović, "Key noun+noun collocations in the language of tourism: a corpus-based study of English and Serbian," ELOPE: English Language Overseas Perspectives and Enquiries, vol. 18, no. 2, pp. 51–68, 2021, doi: 10.4312/ELOPE.18.2.51-68.
3. J. Buts and H. Jones, "From text to data: mediality in corpus-based translation studies,"

Monografias de Traduccion eInterpretacion (MonTI), no. 13, pp. 301–329, 2021, doi: 10.6035/MonTI.2021.13.10.

4. A. Sarudin, H. F. M. Redzwan, and A. A. Tan, “The development of a medicinal nadir glossary through the # lancsbox 6 . 0 ofwizard software,” Malaysia Journal of Invention and Innovation, vol. 1, no. 1, pp. 73–84, 2022.

5. Abhay A. Dande and Dr. M. A. Pund, “A review study on applications of natural language processing,” International Journal ofScientific Research in Science, Engineering and Technology, vol. 10, no. 2, pp. 122–126, 2023, doi: 10.32628/ijrsrset2310214

6. T. Gaillat et al., “A data repository for the management of dynamic linguistic datasets,” in A data repository for the managementof dynamic linguistic datasets, 2021.

7. X. Q. Xia, M. McClelland, and Y. Wang, “TabSQL: a MySQL tool to facilitate mapping user data to public databases,” BMCBioinformatics, vol. 11, pp. 2–5, 2010, doi: 10.1186/1471-2105-11-342.

8. (PDF) *Exploring corpus linguistics via computational tool analysis: key finding review*. Available from:

https://www.researchgate.net/publication/380253006_Exploring_corpus_linguistics_via_computational_tool_analysis_key_finding_review#fullTextFileContent

9. B. Li, “Integrating linguistic theory and neural language models,” ArXiv Preprint, 2022. [Online]. Available: <http://arxiv.org/abs/2207.09643v1>.

10. P. Yao, M. Kosmajac, A. Waheed, K. Guzhva, N. Hervieux, and D. Barbosa, “NLP workbench: efficient and extensibleintegration of state-of-the-art text mining tools,” EACL 2023 - 17th Conference of the European Chapter of the Association forComputational Linguistics, Proceedings of System Demonstrations, pp. 18–26, 2023, doi: 10.18653/v1/2023.eacl-demo.3.