GAZETA MATNLARI KORPUSINI YARATISH BO‘YICHA XALQARO TADQIQOTLAR TAHLILI

Gulchehraxon Arabboyeva
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti,
kompyuter lingvistikasi II bosqich magistratura talabasi

ANNOTATSIYA

Hozirgi kunda tilni rivojlantirishga e’tibor va talab har qachongidanda yuqoriroq bo‘lganligi sababli nafaqat Jahonda shu bilan birga O‘zbekistonda ham kompyuter lingvistikasi sohasi jadal rivojlanmoqda va NLP (Natural language prossesing) doirasida turli tadqiqotlar olib borilmoqda. Tabiiy til jarayonini rivojlantirishga mo‘ljallangan izlanishlar natijasida katta va tizimlangan ma’lumotlar bazasi bo‘lgan korpuslar ishlab chiqilib, turli tadqiqotlar obyektiga aylanmoqda. Ushbu maqolada jahonda gazeta matnlari korpusini yaratish bo‘yicha olib borilgan xalqaro tadqiqotlar tahlil qilingan.

Kalit so‘zlar: *korpus; gazeta matnlari korpusi; bir tilli korpus; ko‘p tilli korpus.*

ABSTRACT

Currently, the attention and demand for language development is higher than ever, not only in the world, but also in Uzbekistan, the field of computational linguistics is developing rapidly, and various researches are being conducted within the framework of NLP (Natural language processing). As a result of research aimed at the Natural language processing, corpora with a large and systematic database have been developed and are becoming the object of various studies. This paper analyzes the international research conducted on the creation of newspaper corpora in the world.

Keywords: *corpus; newspaper corpora; monolingual corpus; multilingual corpus.*

KIRISH

Jahonda korpus lingvistikasida qilingan ishlarning ko‘lami keng. O‘zbek korpus lingvistikasida esa endigina dastlabki qadam qo‘yilgan bo‘lib, hali yechimini topmagan masalalar talaygina [1]. Xorijda bir necha o‘n yillar davomida korpus lingvistikasi ustida ko‘plab izlanishlar olib borildi. Ayni vaqtda korpus lingvistikasi kompyuter lingvistikasidan ayro holda, katta bir soha sifatida o‘rganilmoqda. O‘zbek korpus lingvistikasida xorijiy korpus yaratish tamoyillari andozasi asosida turli maqsadlarga yo‘naltirilgan bir qancha turdagi korpuslar yaratildi. Parallel korpus, ta’limiy korpus, mualliflik korpuslari shular jumlasidandir. Ushbu korpuslar turli

manbalarga asoslangan. Hozirgi vaqtda jahonda gazeta matnlariga asoslangan korpuslar yaratish ancha keng tus olgan.

MUHOKAMA VA NATIJALAR

Raqamli ko‘rinishdagi gazetalar to‘plamlari gumanitar va ijtimoiy fanlar va bir qator boshqa fanlar bo‘yicha tadqiqotchilar uchun boy ma’lumot manbasi bo‘lib, tarix, media va aloqa tadqiqotlaridan tortib, leksikografiyaga qadar xronologik tadqiqotlar uchun ayniqsa qimmatli. Gazeta korpuslari neologizmlar va boshqa leksikografik hodisalarning boy manbasidir [2]. Hozirda jahonda ko‘plab gazeta korpuslari yaratilgan, va ular ko‘plab tadqiqotlar manbayi bo‘lib xizmat qilmoqda. Ayni shu sabab biz gazeta matnlariga asoslangan xronologik korpus yaratishni maqsad qilib olganmiz. Korpusning boshqa manbalar yoki janrlarga emas aynan gazeta matnlariga asoslanganining sababi gazetalarda chiqadigan maqolalar sanasi aniqligi va izchilligidadir.

Umumiy til resurslari va texnologiyasi infratuzilmasi (Common Language Resources and Technology Infrastructure (CLARIN)) gazeta korpuslarini o‘zida jamlagan asosiy ko‘zga ko‘ringan manbalardan biri hisoblanadi. Ushbu infratuzilma keng qamrovli bo‘lib, unda 34 ta gazeta korpusiga kirish imkoniyati mavjud. CLARIN infratuzilmasiga kirgan maqolalarning katta qismi tarixiy bo‘lib, ular XVIII asrdagi eng qadimgi nashriyot matnlaridan tashkil topgan. Infratuzilmadagi gazeta korpuslarining 7 tasi ko‘p tilli, ya’ni multilingual, qolgan qismi esa monolingual (bir tilli) korpuslardir. CLARIN chex, fin, arab, nemis, fransuz, polyak, norveg, shved, yunon va italyan tillaridagi gazeta korpuslarini o‘z ichiga olgan.

Tilshunoslikda korpus (ko‘plikda corpora) yoki matn korpusi katta va tizimlangan matnlar to‘plamidan (hozirgi kunda odatda elektron saqlanadi va qayta ishlanadi) iborat til manbayidir [3]. Korpus tilshunosligida ular muayyan til doirasida statistik tahlillarni amalga oshirish va gipotezani tekshirish, tildagi hodisalarni kuzatish yoki nazariy lingvistik qoidalarni tekshirish uchun foydalaniladi. Korpusda bir tildagi (bir tilli korpus yoki monolingvistik korpus) yoki bir nechta tildagi (ko‘p tilli korpus yoki multilingvistik korpus) matnli ma’lumotlar bo‘lishi mumkin. CLARIN infratuzilmasida keltirilgan korpuslarning aksariyati mana shunday ko‘p tillidir [4]. Quyidagi jadvalda bir tilli gazeta matnlari korpuslariga misollar keltirilgan.

1-jadval. CLARIN infratuzilmasiga kirgan bir tilli gazeta korpuslariga misollar

	Korpus nomi	Korpus hajmi	Til	Korpus tavsifi	Korpusdan foydalanish litsenziyasi
1.	An-Nahar gazeta matnlari korpusi [5]	24 million token	Arab	Ushbu korpusda 1995 yildan 2000 yilgacha arabcha “An-Nahar” gazetasidan olingan maqolalar mavjud	ELRA END USER
2.	The Karelian Finnish gazeta matnlari korpusi [6]	500 000 token	Fin	Ushbu korpus Finlyandiyaning Karjalan Sanomat gazetasining 2012 yildan 2014 yilgacha bo‘lgan maqolalarini o‘z ichiga oladi	CLARIN ACA
3.	TIGER gazeta korpusi [7]	900 000 token	Nemis	Ushbu korpusda Germaniyaning Frankfurter Rundschau gazetasi maqolalari mavjud	CLARIN PUB
4.	Zamonaviy grekcha matnlar korpusi: “Makedoniya” gazetasi korpusi [8]	3 million token	Grek	Ushbu korpusda turli mavzularda (siyosat, iqtisodiyot, sport) gazeta maqolalari mavjud	CC-BY-NC-SA
5.	GP 1994 and 2001-2011 [9]	271 million token	Shved	Ushbu korpus Shvetsiyaning Göteborgsposten gazetasidan 1994 yildan va 2001 yildan 2011 yilgacha bo‘lgan maqolalarni o‘z ichiga oladi.	CC-BY

Bir tilli korpus korpusning eng keng tarqalgan turi hisoblanadi. Unda faqat bitta tildagi matnlar mavjud.

Yuqorida keltirilgan monolingual gazeta korpuslaridan bir nechtasi haqida umumiy ma’lumot beradiga bo‘lsak, An-Nahar Livan gazetasining matn korpusi 1995 yildan 2000 yilgacha (6 yil) CDrom muhitida HTML fayllari sifatida saqlangan standart arab tilidagi maqolalarni o‘z ichiga oladi. Har yili 45 000 ta maqola va 24 million so‘z kiritiladi. Har bir maqola sarlavha, gazetaning nomi, sanasi, mamlakati, turi, sahifasi va boshqalar kabi ma’lumotlarni o‘z ichiga oladi

TIGER Corpus (2.1 va 2.2 versiyalari) ilovadan iborat. Frankfurter Rundschaudan olingan 900 000 token (50 000 jumla) nemis gazetasi matni. Korpus yarim avtomatik ravishda POS-teglangan va sintaktik tuzilish bilan izohlangan. Bundan tashqari, u terminal tugunlari uchun morfologik va lemma ma’lumotlarini o‘z ichiga oladi.

Ko‘p tilli gazeta matnlari korpusi bir tilli gazeta korpusidan farqli ravishda ikki yoki undan ortiq tildagi alohida nashr etilgan gazetalarni o‘z ichiga oladi. Bu esa ushbu gazetalar ustida olib boriladigan tadqiqotlarning yanada kengroq o‘rganilishiga imkon yaratadi.

2-jadval. CLARIN infratuzilmasiga kirgan multilingual gazeta korpuslari

	Korpus nomi	Korpus hajmi	Til	Korpus tavsifi	Korpusdan foydalanish litsenziyasi
1.	MLCC Ko‘p tilli va Parallel Korpus [10]	100 million token	golland, ingliz, fransuz, nemis, italyan, ispan	Ushbu korpusda 1986 yildan 1994 yilgacha golland, ingliz, frantsuz, nemis, italyan va ispan tillaridagi gazetalardagi maqolalar mavjud	ELRA END USER
2.	Gazeta matnlari korpusi [11]	435 million token	Shved, ingliz, fin	Ushbu korpusda turli shved, ingliz va fin gazetalaridagi maqolalar mavjud	-
3.	SETIMES - Bolqon tillarining parallel korpusi [12]	341.83 million token	Rumin, turk, serb, ingliz, bolgar, makedon, xorvat, yunon, alban	Ushbu parallel korpusda SETimes veb-sahifasidan olingan onlayn yangiliklar maqolalari mavjud.	Cheklovlar bilan qayta foydalanish uchun ochiq
4.	Parallel Global Voices [13]	8 million	40 ta til	Ushbu korpus https://globalvoices.org/ veb-saytidagi maqolalarni o‘z ichiga oladi, bu yerda ko‘ngillilar 40 dan ortiq tillarda yangiliklarni nashr etadilar va tarjima qiladilar	CC BY

Multilingual korpus parallel korpus ham deyiladi. Bunda korpus ikki yoki undan ortiq tildagi matnli ma’lumotlardan iborat bo‘ladi. Korpusdagi matnlar bir-birining aynan tarjimasi bo‘ladi. Masalan, biror mashhur asar va uning boshqa tildagi tarjimasi kabi. Ikkala til ham bir-biriga moslashtirilishi kerak, ya’ni segmentlar, jumlalar yoki mavzular mos kelishi kerak. Bunda foydalanuvchi bir tildagi so‘z yoki iboraning barcha misollarini qidirishi mumkin va natijalar boshqa tildagi tegishli jumlalar bilan birga ko‘rsatiladi. Keyin foydalanuvchi qidiruv so‘zi yoki iborasi qanday tarjima qilinganligini kuzatishi mumkin. 2-jadvalda nomi keltirilgan gazeta korpuslari aynan shunday bo‘lib, ikki va undan ortiq tildagi matnlarni o‘z ichiga olgan. Bu esa korpus qamrovini, izlanishlar imkoniyatini kengaytiradi.

XULOSA

Xulosa sifatida shuni aytish mumkinki, gazeta korpusining yaratilishi minglab maqolalarni bir joyga umumlashtirish imkoniyatini yaratgan. Bu esa tadqiqotchilar tomonidan ilmiy izlanishlar olib borilishini yengillashtirgan. Ayniqsa, multilingual gazeta korpuslari bu izlanishlarni yanada kengroq olib borish imkoniyatini ta'minlagan.

REFERENCES

1. Abduraxmonova, N. Z. Q., & Arabboyeva, G. S. Q. (2022). XRONOLOGIK LINGVISTIK KORPUS YARATISHDA GAZETA MATNLARI O 'RGANISH OBYEKTI SIFATIDA. *Academic research in educational sciences*, *3*(4), 663-667.
2. S. Kübler, H. Zinsmeister. "Corpus Linguistics and Linguistically Annotated Corpora." *Bloomsbury*. 312 pp, 2015. ISBN: 978-1-4411-6447-6
3. R. Nordquist. "Definition and Examples of Corpus Linguistics," ThoughtCo., 2019. [Online]. Havola: https://www.thoughtco.com/what-is-corpus-linguistics-1689936
4. S. Wallis, G. Nelson. "Knowledge discovery in grammatically analysed corpora." *Data Mining and Knowledge Discovery*, 5: 305–335. 2001.
5. "An-Nahar Newspaper Text Corpus," ELRA, 2010. [Online]. Havola: https://catalog.elra.info/en-us/repository/browse/ELRA-W0027/
6. H. Kemppanen, Saikonen, A., & Mäkisalo, J. (2016). The Karelian Finnish Newspaper Corpus [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2014092601
7. S. Brants, S. Dipper, P. Eisenberg, S. Hansen, E. König, W. Lezius, Ch. Rohrer, G. Smith, H. Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, 2004 (2), 597-620.
8. CGL Modern Greek Texts Corpora: newspaper corpus "Makedonia" (2015). [Dataset (Text corpus)]. CLARIN:EL. http://hdl.handle.net/11500/KEG-0000-0000-24FB-D
9. "GP 1994 and 2001-2011," [Online]. Havola: https://spraakbanken.gu.se/en/resources
10. "MLCC Multilingual and Parallel Corpora," ELRA, 2012. [Online]. Havola: http://catalog.elra.info/en-us/repository/browse/ELRA-W0023/
11. T. Jauhiainen, S. Pöyhönen (2012). Corpora of Newspaper Texts [text corpus]. [Online]. Havola: http://urn.fi/urn:nbn:fi:lb-20140730175
12. SETIMES - A parallel corpus of the Balkan languages (2015). [Dataset (Text corpus)]. CLARIN:EL. http://hdl.handle.net/11500/ATHENA-0000-0000-2591-2