Academic Research in Educational Sciences
ISSN: 2181-1385
DOI: 10.24412/2181-1385-2022-4-663-667

Volume 3 | Issue 4 | 2022
Cite-Factor: 0,89 | SIS: 1,12
SJIF: 5,7 | UIF: 6,1

XRONOLOGIK LINGVISTIK KORPUS YARATISHDA GAZETA MATNLARI O‘RGANISH OBYEKTI SIFATIDA

Nilufar Zaynobiddin qizi Abduraxmonova
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti dotsenti

Gulchehra Shuhratjon qizi Arabboyeva
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti magistranti

ANNOTATSIYA

Ushbu maqolada korpus turlaridan biri bo‘lgan xronologik korpus haqida umumiy ma’lumotlar keltirilgan. Jahonda xronologik korpus yuzasidan olib borilgan tadqiqotlar, korpus turlari va ularning tilshunoslikda tutgan ahamiyati atroflicha tahlil qilingan. Xususan, Polshaning ChronoPress nomli bosma nashrlardan tashkil topgan xronologik korpusi, uning o‘ziga xos xususiyatlari, bajaradigan funksiyalari, afzalliklari va yaratilish asoslari izohlangan. Bundan tashqari gazeta matnlari asosida xronologik korpus yaratish va uning afzalliklari haqida ma’lumotlar yoritib berilgan.

Kalit so‘zlar: korpus lingvistikasi, korpus, xronologik korpus, sinxroniya, diaxroniya.

ABSTRACT

This article presents an overview of a chronological corpus, one of the types of the corpus. The study of a chronological corpus in the world, the types of corpus and their importance in linguistics are analyzed in detail. In particular, the chronological corpus of Polish publications ChronoPress, its peculiarities, functions, advantages and basics of creation are explained. There is also information about the creation of a chronological corpus based on newspaper articles and its advantages.

Keywords: corpus linguistics, corpus, chronological corpus, synchrony, diachrony.

KIRISH

Texnologiyalarning rivojlanishi natijasida XX asrga kelib Jahonda tabiiy tillar jarayoni (NLP) ni tadqiq etish bir muncha yuqori pog‘onaga ko‘chdi. Shu qatorida kompyuter lingvistikasi sohasida ham jadal izlanishlar olib borildi. Ko‘plab tadqiqotchilar tomonidan

tilshunoslikka oid turli izlanishlarni samarali va sifatli amalga oshirish uchun katta hajmli va tizimga solingan matnlar to‘plamidan iborat korpuslar yaratildi. Ushbu korpuslarning yaratilishi nafaqat tilshunoslikda balki boshqa sohalarda ham keng izlanishlar olib borishga yo‘l ochib berdi. Shu tariqa kompyuter lingvistikasi sohasida korpus lingvistikasi deb atalgan katta bir yo‘nalish vujudga keldi.

Hozirgi kunda Jahonda bir qancha turdagi korpuslar yaratilib, alohida soha sifatida o‘rganilmoqda. Mana shunday korpus turlaridan biri bu xronologik korpusdir.

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Xronologik korpus tushunchasi korpus lingvistikasi uchun nisbatan yangi bo‘lib, hozirgi vaqtda ayrim nashrlardagina paydo bo‘lgan. Xorijdagi korpus lingvistikasi va boshqa fanlarga oid adabiyotlarda ham bu tushuncha deyarli mavjud emas. O‘zbek kopus lingvistikasida ham bu turdagi korpus bo‘yicha ma’lumotlar, ilmiy ishlar hali o‘rganilmagan.

Demakki, oldimizda “Xronologik korpus o‘zi nima?” degan savol paydo bo‘ladi. Adam Pavlovskiy o‘zining “Chronological corpora: Challenges and opportunities of sequential analysis. The example of ChronoPress corpus of Polish” maqolasida xronologik korpus nima ekanligini tushunish uchun, avvalo, strukturaviy tilshunoslikning markazida joylashgan sinxroniya va diaxroniya o‘rtasidagi asosiy farqni tushunish kerakligini aytib o‘tadi. Sinxroniya - bu tilni ma’lum bir vaqtda o‘rganish bo‘lib, unda grammatika, lug‘at va talaffuzda aniq o‘zgarishlar bo‘lmasa, “lahza”, hatto bir necha o‘n yillar davom etishi mumkin. Diaxroniyaga kelsak, u tilning rivojlanish jarayonida, odatda, uzoq davrlarda, hatto bir necha asrlarni qamrab olgan evolyutsiyasini ochib beradi. Biroq, Adamning fikricha, NLP vositalari tomonidan qo‘llab-quvvatlanadigan korpus tadqiqotlari vaqt o‘zgaruvchisiga nisbatan ancha moslashuvchan yondashuvni ta’minlaydi, chunki ularning tipografiyasi, imlosi va grammatikasiga mos keladigan matn namunalari ularning aniq nashr sanalari bilan izohlanishi mumkin. Keyinchalik ilmiy tavsif so‘z shakllarining faraziy proto-tildan boshlab hozirgi holatgacha bo‘lgan evolyutsiyasiga emas, balki muayyan leksemalarning (yoki boshqa segmentlarning) vaqt bo‘yicha chastotasi o‘zgarishi dinamikasiga qaratiladi.

Binobarin, xronologik korpus deganda vaqt o‘qi bo‘yicha keyingi nuqtalarga (masalan, haftalar, oylar va h.k.) to‘g‘ri keladigan imlo va grammatika nuqtai nazaridan mos keladigan matn namunalari ketma-ketligi tushunilishi kerak. Bunday korpus vaqt qatorlari tahlili usuli yordamida leksema chastotalarining ketma-ket davrlarda

o‘zgarish dinamikasini o‘rganish imkonini beradi. Xronologik korpus diaxronikdan farq qiladi, chunki oldingi matnlarda vaqt bir tekis tarqaladi va so‘z shakllari o‘zgarishsiz qoladi, ikkinchisida esa buning aksi so‘z shakllari qiziqish obyektiga aylanish uchun rivojlanishi kerak va o‘lchovlar orasidagi vaqt oralig‘i har qanday uzunlikda bo‘lishi mumkin.

Hozirgi vaqtda xronologik jihatdan ichki tartibga solingan bir qancha xorijiy korpuslar mavjud. Shekspir korpusi (Shakespeare Corpus), Tomas korpusi (Korpus Thomisticum) va Platon korpusi (Corpus Platonicum) shular jumlasidan.

Shekspir korpusida uning 37 ta pyesasi, shuningdek, qahramonlarning barcha nutqlari mavjud. Ya’ni, biz korpusdan har qanday qahramon nutqini alohida olishimiz mumkin. Shuningdek, spektakllar ro‘yxati va ularning sanalari ham mavjud.

Korpus Thomisticum da esa Tomas Akviniskiyning asarlari, XIII asrdan hozirgi kungacha bo‘lgan Tomas va uning ta’limotiga oid barcha tadqiqotlar indeksi va nashrlari bilan birga keltirilgan. Ma’lumotlar bazasini boshqarish tizimi so‘zlar, iboralar, kotirovkalar, o‘xshashliklar va statistik ma’lumotlarni qidirish, solishtirish va saralash uchun amalga oshiriladi. Korpusning asosiy tili lotin tili hisoblanadi. Bu orqali izlanuvchilar Tomas Akviniskiyning asarlarini lotin tilida yozilgan asl matnlarini o‘qishi mumkin.

Agar korpusdagi har bir ma’lumotning yaratilgan sanasi ma’lum bo‘lsa, matnning uslubiy xususiyatlari statistikasini vaqt o‘tishi bilan rivojlanishini aniqlash mumkin. Agar matn xronologiyasi qisman noma’lum bo‘lib qolsa, stilometrik tadqiqotlar sanasi ko‘rsatilmagan asarlarni to‘g‘ri tartibda joylashtirishga yordam beradi.

Hozirgi kunda xronologik korpusning ko‘zga ko‘ringan namunalaridan yana biri bu Polsha ChronoPress (ChronoPress corpus of Polish) korpusidir. Adam Pavlovskiyning keltirgan ma’lumotlariga ko‘ra ushbu korpusda Polshaning 1945-yildan 1962-yilgacha bo‘lgan matbuot matnlari jamlangan. Ma’lumotlar Polsha kundalik va haftalik nashrlaridan olinib, metadatada gazeta sarlavhalari, maqolalar nomi, mualliflari va shu kabi ma’lumotlar keltirilgan. Polsha milliy korpusi uchun mo‘ljallangan Morfeusz vositasi yordamida korpus morfosintaktik tarzda izohlangan.

Vaqt o‘zgaruvchisining kiritilishi foydalanuvchilarga uzoq vaqt davomida kundalik matbuotda aks ettirilgan voqea va hodisalar dinamikasini kashf qilish va o‘rganish imkonini berdi. ChronoPress veb-xizmati, shuningdek, “ommaviy” va “satrda” (Gustav Xerdan terminologiyasidan foydalangan holda) matn tahlilining samarali statistik vositalari

bilan ta’minlangan. Ma’lumotlardan bilim olishni osonlashtirish uchun Korpus CLARIN standartlariga muvofiq morfosintaktik tarzda izohlangan va foydalanuvchi interfeysi bilan ta’minlangan.

MUHOKAMA VA NATIJALAR

Ayni damda biz o‘z oldimizga yuqorida keltirilgan xorijiy korpuslarning yaratilish tamoyillari asosida O‘zbek tili korpusi takibida gazeta matnlariga asoslangan xronologik korpus yaratishni maqsad qilib olganmiz. Korpusning boshqa adabiyotlar yoki janrga emas aynan gazetalarga asoslanganining asosiy sababi undagi maqolalar sanasi aniqligi va nashrdan chiqarilishi davomiyligidadir. Bu bizga maqolalarda qo‘llanilayotgan so‘zlarning stilistik jihatini davr nuqtai nazaridan o‘rganish imkoniyatini beradi.

Ishimizning obyekti sifatida biz “Yangi O‘zbekiston” gazetasini tanladik. Hozirda ushbu gazetada chop etilgan maqolalarni skanerlab, txt formatga o‘tkazmoqdamiz. Gazeta metama’lumotiga gazeta soni, chop etilgan sanasi, maqolaning sarlavhasi, muallifi, qaysi rukn ostida chiqqanligi va shu kabi ma’lumotlar kiritilgan.

Olib borgan tadqiqotlarimiz natijasida uzbekcorpus.uz bazasida xronologik matnlar korpusi yaratiladi. Ushbu tadqiqot kelgusida terminologik bilimlar bazasi, amaliy tilshunoslikning o‘rganish obyekti vazifasini bajaradi.

XULOSA

Xulosa sifatida shuni aytish mumkinki, o‘zbek korpus lingvistikasida gazeta matnlariga asoslangan xronologik lingvistik korpusning yaratilishi NLP (tabiiy tillar jarayoni) ning rivojlanishiga o‘z ta’sirini ko‘rsatmay qolmaydi. Xususan, ushbu korpus izlanuvchilarga tilimizdagi so‘zlarning ma’lum vaqt oralig‘idagi qo‘llanilish darajasi va uslubiy jihatlarini o‘rganish imkonini beradi. Yuqorida keltirilgan tadqiqot ishlari natijalari kelajakda xronologik lingvistik korpuslar yaratish yuzasidan olib boriladigan tadqiqotlarga asos vazifasini ham bajarishi mumkin.

REFERENCES

1. A. Pawlowski, “Chronological corpora: Challenges and opportunities of sequential analysis. The example of ChronoPress corpus of Polish,” in *Conference: Digital Humanities 2016*, 2016, no. July, pp. 311–313.
2. J. M. Gottman (1981). Time-series analysis: a comprehensive introduction for social scientists. Cambridge, London etc.: Cambridge University Press.

3. J. Cryer (1986). Time series analysis. Boston: Duxbury Press.
4. A. Pawłowski (2001). Metody kwantytatywne w sekwencyjnej analizie tekstu [Quantitative methods in sequential text analysis]. Warszawa: Uniwersytet Warszawski, Katedra Lingwistyki Formalnej
5. W. Lutosławski (1897). The origin and growth of Plato's logic. London, New York, Bombay: Longmans, Green and Co
6. M. Piasecki (2007), Polish Tagger TaKIPI: Rule Based Construction and Optimisation. TASK Quarterly 11, 151-167.
7. Aripov, M., Sharipbay, A., Abdurakhmonova, N., Razakhova B.: Ontology of grammar rules as example of noun of Uzbek and Kazakh languages. In: Abstract of the VI International Conference "Modern Problems of Applied Mathematics and Information Technology - Al-Khorezmiy 2018", pp. 37–38, Tashkent, Uzbekistan (2018)
8. Abduraxmonova, N. Z. "Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref." (2018).
9. Abdurakhmonova N. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta. 2016;2 (38):12-7.
10. Abdurakhmonova N, Tuliyev U. Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language. Literature, Education. 2018(3):68.
11. Abdurakhmonova N, Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL). 2019;6(1-2019):131-7.
12. Abdurakhmonov N. Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation. Journal of Social Sciences and Humanities Research. 2017;5(03):89-100.
13. Kubedinova L. Khusainov A., Suleymanov D., Gilmullin R., Abdurakhmonova N. First Results of the TurkLang-7 Project: Creating Russian-Turkic Parallel Corpora and MT Systems. Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020) .2020/11: 90-101
14. Abdurakhmonova N. Dependency parsing based on Uzbek Corpus. InProceedings of the International Conference on Language Technologies for All (LT4All) 2019.
15. https://www.corpusthomisticum.org/