O‘ZBEKISTON MILLIY UNIVERSITETI JIZZAX FILIALI

ISSN 2181-3884

ZAMONAVIY ILMIY-TEXNIK TADQIQOTLAR

xalqaro jurnali

Vol. 2 No. 4

2026

International Journal of Contemporary Scientific and Technical Research

Volume 2 Issue 4
2026

Zamonaviy ilmiy-texnik tadqiqotlar xalqaro jurnali

4-son
2026

Международный журнал современных научно-технических исследований

Том 2 Выпуск 4
2026

MIRZO ULUG‘BEK NOMIDAGI O‘ZBEKISTON MILLIY UNIVERSITETINING JIZZAX FILIALI

Xalqaro ilmiy-elektron jurnali

Tahrir kengashi raisi:

P.f.d., prof., O.X. Turakulov

Bosh muharrir:

A.A. Abdumalikov

Tahrir kengashi mas’ul kotibi:

D.M. Jomurodov

Tahririyat hay’ati:

1. B.f.d., prof., S. Murodova (O’zMU Jizzax filiali, UZB)
2. P.f.n., prof., S. Usmanov (O’zMU Jizzax filiali, UZB)
3. T.f.n., dos., S. Tavboev (O’zMU Jizzax filiali, UZB)
4. F.f.d., prof., N. Djusupov (O’zDJTU, UZB)
5. T.f.d., dos., S. Ruzimov (TTPU, UZB)
6. P.f.d., prof., G. Lekerova (SKU, KZH)
7. DSc., Ali Coşkan (ASUI, TR)
8. T.f.d., prof., D. Irgasheva (TATU, UZB)
9. T.f.d., prof., I. Siddikov (NRU TIIAME, UZB)
10. T.f.n., dos., S. Eshkabilov (NDU, US)

Zamonaviy ilmiy-texnik tadqiqotlar xalqaro elektron jurnali (International Journal of Contemporary Scientific and Technical Research) O’zMU Jizzax filiali Kengashining qaroriga asosan tashkil etilib, 2022-yil 9-sentabrda O‘zbekiston Respublikasi Prezidenti Administratsiyasi huzuridagi Axborot va ommaviy kommunikatsiyalar agentligi tomonidan №038200 raqami bilan ro‘yxatidan o‘tkazilgan, shuningdek davlatlararo standartlar talabi asosida O‘zbekiston Milliy kutubxonasidan jurnal uchun 2181-3884 ISSN raqami olingan.

© “O‘zMU Jizzax filiali” – 2023

Tahririyat manzili:

130100, Jizzax viloyati, Jizzax shahri,
Sh.Rashidov Shox ko‘chasi, 259-uy.

International Journal of Contemporary Scientific and Technical Research (IJCSTR)
VOLUME 2 ISSUE 4/ ISSN 2181-3884

TABIIY TILNI QAYTA ISHLASH (NLP) JARAYONIDA MATN TAHLILIGA DOIR NAZARIY QARASHLAR EVOLYUTSIYASI

Isroilov Jasurbek Baxodir o‘g‘li

University of Business and Science, Xalqaro aloqalar bo‘limi boshlig‘i.
e-mail: Jasurbek9109@gmail.com

Annotatsiya.

Maqolada tabiiy tilni qayta ishlash (NLP) sohasining nazariy-metodologik evolyutsiyasi, uning shakllanish bosqichlari hamda zamonaviy taraqqiyot tendensiyalari keng yoritilgan. Tadqiqotda matnni tahlil qilishda qo‘llaniladigan to‘rtta asosiy yondashuv — qoidaviy (generativ), statistik va korpusga asoslangan, mashinali o‘rganish hamda neyron tarmoqlarga asoslangan paradigmalar ilmiy-nazariy jihatdan chuqur tahlil qilingan. Shuningdek, xorijiy olimlarning ilmiy qarashlari bilan bir qatorda, rus, tatar va o‘zbek lingvistik maktablarining ushbu sohadagi yutuqlari, xususan, turkiy tillarning agglyutinativ xususiyatlarini hisobga olgan holda avtomatik qayta ishlash masalalari qiyosiy o‘rganilgan. Maqolada til birliklarini empirik va formal modellashtirishning o‘zaro uyg‘unligi asoslanib, zamonaviy sun’iy intellekt tizimlarida semantik noaniqlik va kontekstni hisobga olishning ilmiy ahamiyati ochib berilgan.

Kalit so‘zlar: tabiiy tilni qayta ishlash (NLP), kompyuter lingvistikasi, qoidaviy yondashuv, korpus lingvistikasi, statistik modellar, mashinali o‘rganish, neyron tarmoqlar, transformerlar, semantik tahlil, kontekst.

ЭВОЛЮЦИЯ ТЕОРЕТИЧЕСКИХ ПОДХОДОВ К АНАЛИЗУ ТЕКСТА В ОБЛАСТИ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА (NLP)

Исроилов Жасурбек Баходир угли

Университет бизнеса и науки, начальник отдела международных связей.

Аннотация.

В статье подробно рассматриваются теоретико-методологическая эволюция обработки естественного языка (NLP), этапы становления данной области и современные тенденции её развития. В исследовании проводится углублённый научно-теоретический анализ четырёх основных подходов к анализу текста: основанного на правилах (генеративного), статистического и корпусного, машинного обучения, а также парадигмы, основанной на нейронных сетях. Наряду с научными взглядами зарубежных исследователей сопоставительно изучаются достижения русской, татарской и узбекской лингвистических школ в данной области, в частности вопросы автоматической обработки тюркских языков с учётом их агглютинативных особенностей. В статье обосновывается взаимосвязь эмпирического и формального моделирования языковых единиц, раскрывается научная значимость учёта семантической неоднозначности и контекста в современных системах искусственного интеллекта.

Ключевые слова: обработка естественного языка (NLP), компьютерная лингвистика, подход на основе правил, корпусная лингвистика, статистические модели, машинное обучение, нейронные сети, трансформеры, семантический анализ, контекст.

THE EVOLUTION OF THEORETICAL APPROACHES TO TEXT ANALYSIS IN NATURAL LANGUAGE PROCESSING (NLP)

Isroilov Jasurbek Baxodir o‘g‘li

University of Business and Science, Head of the International Relations Department.

Abstract.

The article provides a comprehensive examination of the theoretical and methodological evolution of Natural Language Processing (NLP), the stages of its

formation, and contemporary trends in its development. The study presents an in-depth theoretical analysis of four major approaches to text analysis: rule-based (generative), statistical and corpus-based, machine learning, and neural network-based paradigms. Alongside the scholarly perspectives of international researchers, the achievements of Russian, Tatar, and Uzbek linguistic schools in this field are comparatively examined, particularly the challenges of automatic processing of Turkic languages in view of their agglutinative characteristics. The article substantiates the interrelationship between empirical and formal modeling of linguistic units and highlights the scientific significance of addressing semantic ambiguity and contextual information in modern artificial intelligence systems.

Keywords: Natural Language Processing (NLP), computational linguistics, rule-based approach, corpus linguistics, statistical models, machine learning, neural networks, transformers, semantic analysis, context.

Kirish. XXI asrda axborot-kommunikatsiya texnologiyalari jadal rivojlanayotgan bir davrda tabiiy tilni qayta ishlash (Natural Language Processing — NLP) sohasining ilmiy va amaliy ahamiyati oshib bormoqda. Ushbu soha inson tilini kompyuter orqali tahlil qilish, tushunish, generatsiya qilish va undan samarali foydalanish imkonini beruvchi sun'iy intellektning muhim yo'nalishidir. So'nggi o'n yilliklarda xorijda NLP bo'yicha olib borilgan tadqiqotlarda matematik modellar, chuqur o'rganish metodlari, katta hajmli korpuslar hamda real dunyo muammolariga yo'naltirilgan amaliy yechimlar orqali katta yutuqlarga erishilmoqda.

Bugungi raqamli texnologiyalar rivoji sharoitida o'zbek tilini xalqaro miqyosda targ'ib qilish, uni o'rganish va o'rgatish jarayonlarini takomillashtirishda kompyuter lingvistikasi alohida ahamiyat kasb etadi. Kompyuter lingvistikasi fanining asosiy maqsadi — lingvistik masalalarni yechishga xizmat qiluvchi kompyuter dasturlarini ishlab chiqishdan iborat. Ushbu fanning asosiy vazifalari qatoriga til o'qitish jarayonlarini avtomatlashtirish, bilimlar bazasini yaratish, matnlarni grammatik va uslubiy jihatdan tahrirlash, mashina tarjimasi hamda elektron lug'atlar uchun mo'ljallangan dasturiy vositalarni ishlab chiqish kabilarni kiritish mumkin.

Kompyuter lingvistikasi — tilning matematik algoritmlarini ishlab chiqishga ixtisoslashgan, til muammolarini kompyuter texnologiyalari vositasida bartaraf etishga mo'ljallangan sohalardan biri sanaladi. N. Abduraxmonova o'zining "Kompyuter lingvistikasi" darsligida [1] ushbu fanning maqsadi hamda boshqa sohalar bilan o'zaro aloqasi haqida quyidagi fikrlarni keltirib o'tadi: "Kompyuter lingvistikasi kognitiv fanlar va sun'iy intellekt o'rtasidagi soha hisoblanadi. Fanning nazariy jihatlari lingvistika va kognitiv sohalar bilan bog'liq. Dunyo tajribasida formal lingvistik nazariyalar uchun kompyuter modellarini rivojlantirishda kognitiv psixologiya yordamida lingvistlar dasturchilar bilan hamkorlik qiladilar. Amaliy komponentlar sun'iy intellektga bog'liq bo'lgan amaliy modellarni rivojlantiradi. Shu jihatdan "Til injeneriyasi" (Language Engineering) yoki "Til texnologiyasi" (Language Technology) kompyuter lingvistikasi terminiga sinonim sifatida qaraladi. Amaliy kompyuter lingvistikasining asosiy vazifalaridan biri inson va mashina o'rtasidagi muloqot jarayonini osonlashtirishga yordam berish va til texnologiyasini yaratish hisoblanadi".

Yuqoridagi fikrlardan kelib chiqib, zamonaviy tilshunoslik va axborot texnologiyalari kesishmasida shakllangan tabiiy tilni qayta ishlash sohasi bugungi kunda xorijiy ilmiy tadqiqotlarning ustuvor yo'nalishlaridan biri sifatida e'tirof etilmoqda. Ushbu sohada olib borilayotgan izlanishlar tabiiy til hodisalarini faqat nazariy-lingvistik nuqtai nazardan emas, balki ularni hisoblash jarayonlari orqali modellashtirish, algoritmik tahlil qilish va raqamli muhitda qayta ishlash masalalariga qaratilganligi bilan tavsiflanadi. Natijada til birliklari va ularning o'zaro munosabatlari an'anaviy tavsifiy yondashuvdan farqli ravishda, empirik ma'lumotlarga tayangan holda tizimli ravishda o'rganila boshlandi.

METODOLOGIYA VA YONDASHUVLAR TAHLILI

Xorijiy ilmiy adabiyotlar tahlili shuni ko'rsatadiki, tabiiy tilni qayta ishlashning dastlabki bosqichlarida til hodisalarini formallashtirish va qat'iy qoidalar asosida tavsiflashga ustuvor ahamiyat berilgan. 1950–1970-yillarda tabiiy tilni qayta ishlash masalalari, asosan, tilni qat'iy formallashtirilgan tizim sifatida talqin qiluvchi qoidaviy yondashuvlar doirasida rivojlandi. Bu jarayonda yaratilgan tizimlar uchta qoida asosida ishlaydi: avval berilgan matn so'z turkumlariga ajratiladi, ularning kompyuter dasturiga kiritilgan mos kodlari topiladi va foydalanuvchiga umumiy natijani taqdim etish uchun jumla tarkibi qidiruvga mos ravishda qayta tartibga solinadi. Bu jarayonda natijaning aniqlik darajasi hamda tezligini kamaytirishi mumkin bo'lgan tilning istisno holatlari mavjud bo'lib, iboralar, kollokatsiyalar, murakkab tuzilishga ega bo'lgan birikmalar shular jumlasidandir. Ularni dasturga kiritish mashaqqatli vazifalardan biri hisoblanadi. Ushbu yondashuvning nazariy asoslari N. Chomsky [2] tomonidan ishlab chiqilgan generativ grammatika nazariyasiga borib taqaladi. Olim tilni inson ongida mavjud bo'lgan abstrakt qoidalar majmui sifatida talqin qilib, lingvistik modellashtirishning markaziga sintaktik tuzilmani qo'yadi.

Biroq ushbu nazariy pozitsiya barcha tadqiqotchilar tomonidan bir xil talqin qilinmagan. Jumladan, Y. Wilk [3] generativ yondashuvning hisoblash nuqtai nazaridan haddan tashqari ideallashtirilganligini ta'kidlab, tabiiy tilni qayta ishlash jarayonida semantik noaniqlik va kontekst omilini hisobga olish zarurligini ilgari suradi. Uning fikriga ko'ra, tilni faqat qat'iy qoidalar orqali modellashtirish real nutq jarayonlarini yetarli darajada aks ettira olmaydi.

Shuningdek, T. Winograd [4] qoidaviy yondashuvning amaliy qo‘llanishi cheklanganligini ko‘rsatib, tilni tushunish jarayoni faqat sintaktik qoidalarga emas, balki kontekstual bilim va nutqiy vaziyatga bog‘liq ekanini asoslab beradi. Uning tadqiqotlarida tilni qayta ishlash muammosi lingvistik tizim emas, balki bilimlarni modellashtirish muammosi sifatida talqin qilinadi.

Yuqoridagi qarashlarga yaqin holda, R. Schank [5] tabiiy tilni qayta ishlashda qoidaviy grammatikani emas, balki inson tajribasi va ssenariylarga asoslangan bilim modellarini markaziy o‘ringa qo‘yadi. U tilni tushunish jarayonini faqat formal tuzilma emas, balki semantik va pragmatik bilimlar majmui sifatida izohlaydi.

Qoidaviy yondashuv tabiiy tilni qayta ishlash tadqiqotlarida til hodisalarini aniq belgilangan grammatik va strukturaviy qoidalar asosida tavsiflashga qaratilgan metodologik yo‘nalish sifatida qaraladi. Ushbu yondashuv jahon tilshunosligida shakllangan nazariy qarashlar bilan bir qatorda, turkologik va rus lingvistik an’analarida ham muhim o‘rin egallab kelgan. Ayniqsa, turkiy tillarning agglutinativ xususiyatlari ularni formal modellashtirish uchun qulay obyekt sifatida talqin qilishga sabab bo‘lgan.

Xususan, N. Abduraxmonova o‘z tadqiqotlarida o‘zbek tilini avtomatik qayta ishlash jarayonida grammatik qoidalarni formallashtirish muhim metodologik asos bo‘lib xizmat qilishini ta’kidlaydi. Olimaning fikriga ko‘ra, morfologik va sintaktik qoidalar til birliklarini dastlabki avtomatik tahlil qilish uchun zarur tayanch vazifasini bajaradi, biroq bu qoidalar real matnlarda uchraydigan semantik noaniqlik va kontekstual xilma-xillikni to‘liq qamrab ololmaydi. Ushbu qarash turkologiya va rus ilmiy maktablarida shakllangan pozitsiyalar bilan mushtarak bo‘lib, qoidaviy yondashuvning imkoniyatlari va chegaralarini aniq belgilab beradi. Abduraxmonova formal modellarni tilni avtomatik qayta ishlashning zarur, ammo yetarli bo‘lmagan bosqichi sifatida baholab, ularni korpusga asoslangan va empirik yondashuvlar bilan uyg‘unlashtirish zarurligini ilmiy jihatdan asoslaydi. Bu yondashuv o‘zbek tilini qayta ishlash masalalarini jahon va mintaqaviy ilmiy jarayonlar bilan uzviy bog‘lashga xizmat qiladi.

Shunday qilib, qoidaviy yondashuv doirasida olib borilgan tadqiqotlar tahlili shuni ko‘rsatadiki, ushbu nazariya yagona ilmiy pozitsiya sifatida emas, balki turli tadqiqotchilar tomonidan turlicha talqin qilingan konseptual yo‘nalishdir. Bu yondashuv sintaktik modellashtirishda muhim natijalar bergan bo‘lsa-da, kontekstual va semantik omillarni hisobga olish masalasida cheklangan bo‘lib, keyingi bosqichlarda statistik va korpusga asoslangan yondashuvlarning shakllanishiga ilmiy zamin yaratgan.

NATIJALAR

Qoidaviy yondashuvlar tabiiy tilni qayta ishlashning dastlabki bosqichlarida muayyan nazariy va amaliy natijalar bergan bo‘lsa-da, real nutq materiallari bilan ishlash jarayonida ularning imkoniyatlari cheklanganligi tobora yaqqol namoyon bo‘la boshladi. Xususan, tabiiy tillarga xos bo‘lgan ko‘pvariantlilik, kontekstga yuqori darajada bog‘liqlik, semantik noaniqlik va istisnolarning ko‘pligi til hodisalarini faqat qat’iy qoidalar orqali tavsiflash yetarli emasligini ko‘rsatdi. Natijada xorijiy ilmiy tadqiqotlarda tilni ideallashtirilgan formal tizim sifatida emas, balki real qo‘llanish jarayonida shakllanadigan dinamik hodisa sifatida o‘rganish zarurati yuzaga keldi.

XX asrning 1980–1990-yillariga kelib, tabiiy tilni qayta ishlash sohasida qoidaviy yondashuvlarning cheklanganligi ilmiy jihatdan tobora yaqqol namoyon bo‘la boshladi. Ayniqsa, real nutq materiallarida uchraydigan kontekstual o‘zgaruvchanlik, ko‘pma’nolilik va ehtimollik xususiyatiga ega hodisalar tilni qat’iy qoidalar orqali modellashtirish yetarli emasligini ko‘rsatdi. Shu sababli xorijiy tadqiqotlarda til birliklarini empirik ma’lumotlar, ya’ni katta hajmdagi matnlar majmuasi asosida o‘rganishga tayangan yangi paradigma shakllandi. Bu tabiiy inson tilidagi o‘zgarishlarni va uning o‘ziga xos murakkab jihatlarini nazoratga olish imkonini berdi [6]. Tabiiy tilni qayta ishlash vazifalari uchun Bayes klassifikatori [7], Markov zanjiri [8] va boshqalar ishlatila boshlandi. Ma’lumotlar asosida o‘rganish qobiliyati tufayli bu davrda tabiiy tilni qayta ishlash tizimlarining aniqligi va samaradorligi sezilarli darajada oshdi.

Bu yo‘nalishning nazariy asoslari dastlab Britaniya korpus lingvistikasi maktabida ishlab chiqildi. Xususan, J. Sinclair [9] til birliklarining ma’nosi va funksiyasi ularning real nutqiy qo‘llanishi orqali namoyon bo‘lishini ta’kidlab, lingvistik tahlilni katta hajmdagi elektron matnlar asosida olib borish zarurligini ilmiy jihatdan asoslab berdi. Uning fikriga ko‘ra, til hodisalarini izohlashda alohida misollarga emas, balki chastotaviy va kontekstual naqshlarga tayangan holda xulosa chiqarish maqsadga muvofiqdir. Bu qarash korpus lingvistikasi doirasida empirik yondashuvning ustuvorligini belgilab berdi.

Statistik yondashuvning shakllanishi esa asosan AQSH ilmiy maktabi bilan bog‘liq bo‘lib, bu yerda til hodisalarini ehtimollik va matematik modellar asosida tavsiflashga alohida e’tibor qaratildi. F. Jelinek [10] tomonidan ilgari surilgan qarashlarga ko‘ra, tabiiy tilni qayta ishlashda lingvistik qoidalarga tayanishdan ko‘ra, statistik qonuniyatlar va ma’lumotlarga asoslangan modellar yuqoriroq natija beradi. Ushbu yondashuv tilni deterministik tizim sifatida emas, balki ehtimollik hodisasi sifatida talqin qilishga olib keldi.

Statistik NLP paradigmasining nazariy jihatdan mukammal bayoni K. Manning va H. Schütze [11] tomonidan yaratilgan fundamental asarda berilgan. Ular tabiiy til hodisalarini korpus asosida o‘rganish, ehtimollik modellarini qurish va algoritmlarni real matnlar orqali baholashni statistik NLPning asosiy tamoyillari sifatida belgilaydilar. Ushbu asarda qoidaviy yondashuvlar tarixiy bosqich sifatida baholanib, statistik metodlar esa amaliy jihatdan samaraliroq paradigma sifatida asoslab beriladi.

Korpusga asoslangan yondashuvning metodologik jihatlari keyinchalik Tony McEnery va Andrew Hardie [12]

tadqiqotlarida tizimli ravishda rivojlantirildi. Ular elektron korpuslarni til hodisalarini qayta tekshirish, natijalarni takrorlash va empirik dalillar asosida umumlashtirish imkonini beruvchi ilmiy infratuzilma sifatida talqin qiladilar. Bu yondashuv tilshunoslik tadqiqotlarida subyektiv talqinlardan voz kechib, obyektiv va tekshiriladigan xulosalar chiqarish imkonini berdi.

Statistik va korpusga asoslangan yondashuvlar rus ilmiy maktabida o‘ziga xos nazariy asoslar bilan boyitildi. Rus tilshunosligida statistik yondashuv dastlab matematik lingvistika va struktur-semantik tahlil doirasida rivojlanib, til birliklarini empirik ma’lumotlar asosida o‘rganish zarurati ilgari surildi. Bu jarayonda til hodisalarini faqat formal qoidalar orqali emas, balki real matnlarda kuzatiladigan taqsimot va chastota qonuniyatlari asosida tavsiflash mumkinligi ilmiy jihatdan asoslandi. Rus ilmiy maktabida statistik yondashuvning shakllanishida A. Markov [13] tomonidan ishlab chiqilgan ehtimollik zanjirlari nazariyasi muhim metodologik ahamiyat kasb etdi. Garchi Markov bevosita tabiiy tilni qayta ishlash bilan shug‘ullanmagan bo‘lsa-da, uning ehtimollik modellari keyinchalik til birliklarining ketma-ketlik asosida statistik tavsiflanishi uchun nazariy poydevor vazifasini bajardi. Ushbu yondashuv til birliklarini deterministik emas, balki ehtimollik xususiyatiga ega hodisalar sifatida talqin qilish imkonini berdi.

Statistik va empirik yondashuvlar rus lingvistik maktabida I. Melchuk [14] tomonidan ishlab chiqilgan Mazmun–Matn (Meaning–Text) nazariyasi bilan uyg‘un holda rivojlandi. Mazkur nazariyada til birliklari chuqur semantik darajadan yuzaki matn darajasigacha izchil modellar orqali tavsiflanadi. Melchuk formal modellashtirishga tayanar ekan, real nutq materiallarini hisobga olish zarurligini ham tan olib, empirik kuzatuvlar bilan boyitilgan nazariy tavsifni ilgari suradi. Bu holat rus maktabida qoidaviy va statistik yondashuvlarning qarama-qarshi emas, balki o‘zaro to‘ldiruvchi metodologiyalar sifatida talqin qilinishiga olib keldi.

Korpusga asoslangan yondashuv rus tilshunosligida Yu. Apresjan [15] tadqiqotlarida muhim o‘rin tutadi. Olim integral leksik tavsif nazariyasida so‘z ma’nosini real nutq kontekstida namoyon bo‘ladigan hodisa sifatida talqin qilib, empirik ma’lumotlarga tayangan tahlilning ahamiyatini ta’kidlaydi. Apresjan qarashlarida leksik birliklarni korpus materiallari asosida o‘rganish semantik noaniqlik va ko‘pma’nolilikni aniqlashda samarali vosita sifatida baholanadi. Shu tariqa, rus ilmiy maktabida statistik va korpusga asoslangan yondashuvlar formal modellashtirishga qarama-qarshi qo‘yilmay, balki uni empirik ma’lumotlar bilan boyituvchi metodologiya sifatida shakllandi. Ushbu yondashuv til hodisalarini real nutq materiallari asosida o‘rganish imkonini kengaytirib, keyingi bosqichlarda murakkab semantik hodisalarni, jumladan, omonimlikni avtomatik tahlil qilish uchun zarur bo‘lgan nazariy va metodologik asosni yaratdi.

Tatar tilshunosligida statistik va korpusga asoslangan yondashuvlarning shakllanishi elektron korpuslarni yaratish hamda ularni lingvistik tadqiqotlarning asosiy empirik manbai sifatida qo‘llash jarayoni bilan uzviy bog‘liq holda kechdi. Ushbu metodologik yo‘nalish doirasida olib borilgan izlanishlar tatar tilini real nutq materiallari asosida o‘rganish zaruratini ilgari surib, empirik tahlilga tayangan lingvistik tadqiqotlar uchun mustahkam ilmiy zamin yaratdi.

A. Gatiatullin [16] o‘z tadqiqotlarida tatar tilining elektron korpusini lingvistik tahlilning markaziy empirik manbai sifatida talqin qilib, til birliklarini real nutqiy qo‘llanish asosida o‘rganish statistik qonuniyatlarni aniqlashda muhim metodologik ahamiyatga ega ekanini ta’kidlaydi. Olimning fikriga ko‘ra, aynan korpus materiallariga tayangan holda til birliklarining chastotaviy taqsimoti, kontekstual xususiyatlari hamda funksional imkoniyatlarini aniqlash mumkin. Shu bilan birga, Gatiatullin agglutinativ tillarga xos bo‘lgan morfologik murakkablikni hisobga olgan holda, sof statistik yondashuvlar til hodisalarini to‘liq tavsiflash uchun yetarli emasligini ko‘rsatib, ularni grammatik va morfologik modellashtirish bilan uyg‘unlashtirish zarurligini ilmiy jihatdan asoslaydi.

Gatiatullin tadqiqotlarida elektron korpus nafaqat lingvistik ma’lumotlar bazasi, balki nazariy xulosalar chiqarish imkonini beruvchi metodologik vosita sifatida talqin qilinadi. Olim tatar tilida kuzatiladigan ko‘pvariantlilik va kontekstga kuchli bog‘liqlik kabi xususiyatlarni aniqlashda korpusga asoslangan statistik tahlilning muhim ahamiyatga ega ekanini asoslab beradi. Mazkur ilmiy qarashlar tatar tilshunosligida qoidaviy va empirik yondashuvlarni qarama-qarshi qo‘ymasdan, balki ularni o‘zaro to‘ldiruvchi metodologiyalar sifatida uyg‘un holda qo‘llash tamoyilining shakllanishiga xizmat qiladi.

Statistik va korpusga asoslangan yondashuvlar tabiiy tilni qayta ishlash sohasida jahon ilmiy paradigmasi sifatida shakllangan bo‘lsa-da, ularni turkiy tillarga tatbiq etish jarayonida tilning tipologik va strukturaviy xususiyatlarini inobatga olgan holda maxsus metodologik yondashuvlarni ishlab chiqish zarurati yuzaga keldi. Turkiy tillar doirasida statistik va korpusga asoslangan metodlarni tizimli ravishda qo‘llashga oid muhim tadqiqotlar Kemal Oflazer [17] tomonidan olib borilgan ilmiy ishlarda o‘z ifodasini topgan. Olim turkiy tillar uchun avtomatik morfologik tahlil masalalarini korpus materiallariga tayangan holda ko‘rib chiqib, til birliklarini faqat qat’iy grammatik qoidalar asosida tavsiflash real matnlarda uchraydigan noaniqliklarni to‘liq qamrab olish imkonini bermasligini asoslab beradi.

Turkiy tillar bo‘yicha korpus lingvistikasi masalalari keyinchalik Ch. Cho‘ltekin [18] ishlarida yanada rivojlantirildi. Uning tadqiqotlarida korpusga asoslangan statistik tahlil metodlari turkiy tillarning morfologik murakkabligi bilan bevosita bog‘liq holda ko‘rib chiqiladi. Cho‘ltekin korpus ma’lumotlariga tayangan holda til birliklarini avtomatik tahlil qilishda faqat chastotaviy ko‘rsatkichlarga tayanish yetarli emasligini, balki morfologik va sintaktik belgilarni ham hisobga olish zarurligini ta’kidlaydi.

MUHOKAMA

Statistik va korpusga asoslangan yondashuvlar tabiiy tilni qayta ishlashda empirik tahlilning muhim bosqichini

tashkil etgan bo‘lsa-da, ushbu metodologiyalar asosan oldindan belgilangan belgilar va soddalashtirilgan ehtimollik modellariga tayanishi bilan tavsiflanadi. Ushbu holat real nutqda uchraydigan murakkab semantik bog‘lanishlar va kontekstual munosabatlarni to‘liq modellashtirish imkonini cheklab qo‘ydi. Shu sababli XXI asr boshlarida tabiiy tilni qayta ishlash sohasida mashinali o‘rganishga asoslangan yondashuvlar shakllandi.

Mashinali o‘rganish paradigmasining asosiy g‘oyasi til birliklarini qo‘lda ishlab chiqilgan qoidalar yoki qat’iy statistik formulalar asosida emas, balki katta hajmdagi ma’lumotlar asosida avtomatik tarzda o‘rganiladigan modellar orqali tavsiflashdan iborat. Ushbu yondashuvda model til hodisalarini qat’iy oldindan belgilangan qoidalarga emas, balki elektron korpus materiallari asosida shakllantirilgan, lingvistik belgilarga ega o‘quv tanlamalari (training dataset) orqali o‘rganadi. Mazkur yondashuv tilni avtomatik qayta ishlashda empirik modellashtirishning ustuvor bosqichga chiqishini anglatadi.

Mashinali o‘rganishga asoslangan yondashuvlarning nazariy poydevori V. Vapnik [19] tomonidan ishlab chiqilgan statistik o‘rganish nazariyasi bilan bog‘liq. Vapnikning qarashlariga ko‘ra, modelning asosiy vazifasi faqat mavjud ma’lumotlarni yodlab olish emas, balki umumlashtirish qobiliyatiga ega bo‘lishdir. Ushbu nazariya asosida ishlab chiqilgan tayanch vektor mashinalari (SVM) tabiiy tilni qayta ishlashda matnlarni tasniflash, so‘z ma’nosini farqlash va kontekstual bog‘lanishlarni aniqlashda keng qo‘llanildi.

Mashinali o‘rganish yondashuvlarining amaliy samaradorligi T. Mitchell [20] tomonidan ilgari surilgan konseptual ta’riflarda ham o‘z ifodasini topadi. Mitchell mashinali o‘rganishni tajriba asosida modelning ishlash sifati yaxshilanib boradigan jarayon sifatida talqin qilib, ushbu yondashuvni qat’iy qoidaviy tizimlardan tubdan farqlaydi. Bu qarash tabiiy tilni qayta ishlashda til hodisalarining dinamik va o‘zgaruvchan tabiatini hisobga olish imkonini berdi.

Mashinali o‘rganishga asoslangan yondashuvlarning asosiy yutuqlari quyidagilardan iborat:

1. Katta hajmdagi korpuslar bilan ishlash imkonini beradi.
2. Til birliklari o‘rtasidagi yashirin bog‘lanishlarni avtomatik aniqlashga qodir.
3. Kontekstga sezgir modellarni yaratish uchun qulay sharoit yaratadi.

Natijada so‘z ma’nosini aniqlash, ko‘pma’nolilik va omonimlikni avtomatik farqlash kabi masalalarda oldingi statistik yondashuvlarga nisbatan yuqoriroq aniqlik ko‘rsatkichlariga erishildi. Shu bilan birga, mashinali o‘rganishga asoslangan yondashuvlar qator cheklovlarga ham ega. Xususan, bunday modellar katta hajmdagi belgilangan (annotatsiyalangan) ma’lumotlarga kuchli darajada bog‘liq bo‘lib, kam resursli tillar uchun ularni qo‘llash murakkablik tug‘diradi. Bundan tashqari, mashinali o‘rganish modellarining ichki qaror qabul qilish mexanizmi ko‘pincha ochiq talqin etilmasligi sababli, ularni lingvistik jihatdan izohlash qiyinlashadi.

Mashinali o‘rganishga asoslangan yondashuvlar statistik va korpus paradigmasining mantiqiy davomi bo‘lib, tabiiy tilni qayta ishlashda empirik modellashtirish imkoniyatlarini sezilarli darajada kengaytirdi. Mashinali o‘rganishga asoslangan yondashuvlar til hodisalarini o‘quv tanlamalari asosida modellashtirish imkonini kengaytirgan bo‘lsa-da, ular ko‘pincha qo‘lda tanlangan belgilar (features)ga tayanishi va kontekstni cheklangan darajada hisobga olishi bilan tavsiflanadi. Ushbu cheklovlar tabiiy tilni qayta ishlashda semantik munosabatlar va uzoq masofali bog‘lanishlarni yanada chuqurroq modellashtirish zaruratini yuzaga keltirdi.

Natijada neyron tarmoqlarga asoslangan yondashuvlar shakllanib, til birliklarini ko‘p o‘lchovli vektorli ifodalar orqali avtomatik o‘rganish imkonini berdi. Ushbu yondashuvlarning nazariy poydevori sun’iy neyron tarmoqlar konsepsiyasiga borib taqaladi. Neyron tarmoqlar g‘oyasi dastlab F. Rosenblatt [21] tomonidan taklif etilgan perseptron modeli bilan bog‘liq bo‘lib, u kiruvchi signallar asosida qaror qabul qiluvchi elementar hisoblash birliklarini modellashtirishni nazarda tutadi. Keyingi tadqiqotlarda bu model ko‘p qatlamli arxitekturalar orqali rivojlantirildi.

Tabiiy tilni qayta ishlashda neyron yondashuvlarning amaliy jihatdan samarali qo‘llanishi Y. Bengio, G. Hinton va Y. LeCun [22] tadqiqotlari bilan bevosita bog‘liq. Ushbu olimlar chuqur neyron tarmoqlarni o‘qitish mexanizmlarini ishlab chiqib, katta hajmdagi ma’lumotlar asosida yashirin semantik munosabatlarni avtomatik o‘rganish mumkinligini ilmiy jihatdan asoslab berdilar. Aynan shu yondashuv til birliklarini qo‘lda belgilangan qoidalar yoki belgilar orqali emas, balki ma’lumotlardan avtomatik hosil bo‘ladigan ichki ifodalar orqali modellashtirish imkonini berdi.

Neyron yondashuvlarning muhim metodlaridan biri so‘zlarning taqsimotga asoslangan vektorli ifodalarini yaratishdir. Bu yo‘nalish T. Mikolov [23] tomonidan ishlab chiqilgan word2vec modeli orqali keng tarqaldi. Ushbu yondashuvda so‘zlar o‘zlari ishtirok etadigan kontekstlar asosida ko‘p o‘lchovli vektorlar orqali ifodalanadi. Natijada semantik yaqinlik va ma’no munosabatlarini avtomatik aniqlash imkoniyati yuzaga keldi. Bu metod oldingi statistik yondashuvlardan farqli ravishda til birliklari o‘rtasidagi yashirin semantik aloqalarni aniqroq modellashtirish imkonini berdi.

Keyingi bosqichda ketma-ketlikni modellashtirishga mo‘ljallangan neyron arxitekturalar, xususan, qayta bog‘lanadigan neyron tarmoqlar va ularning takomillashtirilgan ko‘rinishlari qo‘llanila boshlandi. Ushbu yo‘nalishda S. Hochreiter va Yu. Schmidhuber [24] tomonidan taklif etilgan uzoq-qisqa muddatli xotira (LSTM) modellari matn tarkibidagi uzoq masofali bog‘lanishlarni hisobga olish imkonini berdi. Bu holat til birliklarining kontekstga bog‘liq ma’nolarini aniqlashda muhim ahamiyat kasb etdi.

Neyron yondashuvlarning eng so‘nggi va samarali ko‘rinishlaridan biri transformerlarga asoslangan modellashtirish hisoblanadi. Ushbu yondashuv A. Vasvani [25] va hammualliflari tomonidan taklif etilgan self-attention

mexanizmi asosida ishlab chiqilgan. Transformerlarda til birliklari butun matn konteksti doirasida bir vaqtning o‘zida modellashtiriladi, bu esa omonimlik va ko‘pma’nolilik kabi kontekstga kuchli bog‘liq hodisalarni aniqlashda yuqori natijalarni ta’minlaydi.

FOYDALANILGAN ADABIYOTLAR

1. Abduraxmonova N. Z. Kompyuter lingvistikasi: darslik. – Toshkent: Nodirabegim, 2021. – 394 b.
2. Chomsky N. Syntactic Structures. – The Hague: Mouton, 1957. – URL: https://www.scribd.com/document/522339308/Chomsky-N-1957-Syntactic-Structures.
3. Wilks Y. A Preferential, Pattern-Seeking, Semantics for Natural Language Inference // Artificial Intelligence. – 1975. – Vol. 6, № 1. – P. 53–74. – URL: https://aclanthology.org/J75-3003.pdf.
4. Winograd T. Understanding Natural Language // Cognitive Psychology. – 1972. – Vol. 3, № 1. – P. 1–191. – URL: https://sites.google.com/view/terrywinogradhomepage/home.
5. Schank R. C., Abelson R. P. Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. – Hillsdale, NJ: Lawrence Erlbaum Associates, 1977. – P. 35.
6. History and Evolution of Natural Language Processing (NLP) // GeeksforGeeks. – URL: https://www.geeksforgeeks.org/history-and-evolution-of-nlp/.
7. Langley P., Iba W., Thompson K. An Analysis of Bayesian Classifiers. – AI Research Branch, USA, 1992. – P. 46–58.
8. Norris J. R. Markov Chains. – Cambridge: Cambridge University Press, 1997.
9. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991.
10. Jelinek F. Statistical Methods for Speech Recognition. – Cambridge, MA: MIT Press, 1997.
11. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge, MA: MIT Press, 1999. – URL: https://nlp.stanford.edu/fsnlp/.
12. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012. – URL: https://api.pageplace.de/preview/DT0400.9781139153225_A23866597/preview-9781139153225_A23866597.pdf.
13. Markov A. A. An Example of Statistical Investigation of the Text Eugene Onegin. – URL: https://web.archive.org/web/20230324143335/http://www.alpha60.de/research/markov/DavidLink_AnExampleOfStatistical_MarkovTrans_2007.pdf.
14. Mel'čuk I. A. Dependency Syntax: Theory and Practice. – Albany: State University of New York Press, 1988. – URL: https://books.google.co.uz/books?id=diq29vrjAa4C.
15. Apresjan Yu. D. Lexical Semantics: User’s Guide to Contemporary Russian Vocabulary. – Ann Arbor: Karoma Publishers, 1992.
16. Gatiatullin A. The Tatar National Corpus: Principles of Creation and Linguistic Use.
17. Oflazer K. Two-Level Description of Turkish Morphology. – 1994. – URL: https://aclanthology.org/E93-1066.pdf.
18. Çöltekin Ç. A Set of Open Source Tools for Turkish Natural Language Processing // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). – Reykjavik, 2014. – URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/437_Paper.pdf.
19. Vapnik V. N. Statistical Learning Theory. – New York: Wiley, 1998.
20. Mitchell T. M. Machine Learning. – New York: McGraw-Hill, 1997.
21. Rosenblatt F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain // Psychological Review. – 1958. – Vol. 65, № 6. – P. 386–408.
22. LeCun Y., Bengio Y., Hinton G. Deep Learning // Nature. – 2015. – Vol. 521. – P. 436–444. – DOI: https://doi.org/10.1038/nature14539.
23. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. – 2013. – URL: https://arxiv.org/pdf/1301.3781.
24. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. – 1997. – Vol. 9, № 8. – P. 1735–1780. – DOI: https://doi.org/10.1162/neco.1997.9.8.1735.
25. Vaswani A. et al. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). – 2017. – Vol. 30.