O‘ZBEKISTON RESPUBLIKASI OLIY VA O‘RTA MAXSUS TA’LIM VAZIRLIGI
ALISHER NAVOIY NOMIDAGI
TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

“O‘ZBEK TILI TARAQQIYOTI VA XALQARO HAMKORLIK MASALALARI”

mavzusidagi xalqaro konfrensiya

“ZAMONAVIY LEKSIKOGRAFIYA, TIL KORPUSLARI VA TURKIY TILLAR PLATFORMALARINI YARATISH MUAMMOLARI”

nomdagi sho‘ba materiallari

(2021-yil 18-oktabr)

Toshkent – 2021

LINGVISTIK ONTOLOGIYALARNI TAKOMILLASHTIRISHDA TIL KORPUSLARIDAN FOYDALANISH OMILLARI

Abjalova Manzura Abdurashetovna*

Annotatsiya. Ko‘p holatda tadqiqot jarayonida bir necha masalalar ko‘ndalang bo‘ladi. Kompyuter lingvistikasining yo‘nalishi hisoblanmish kompyuter leksikografiyasida elektron lug‘at, virtual kutubxona kabi tushunchalar qatorida tezaurus, lingvistik ontologiya, Word.Net singari terminlar faol qo‘llaniladi. Mazkur maqolada lingvistik ontologiya konsepsiyasi yoritildi va lingvistik ontologiyani yaratishda til korpuslarining ahamiyati, ulardan foydalanish tamoyillari to‘g‘risida fikr yuritildi.

Kalit so‘zlar: *ontologiya, korpus, lingvodidaktika, taksonomiya, semantik tarmoq.*

Kirish

Zamonaviy axborot qidirish va axborot-tahlil tizimlari keng va cheklanmagan mavzularda, chegaralanmagan turdagi o‘zaro munosabatlarga kirishadigan tushunchalarni qamrab olgan minglab atributlariga ega bilim sohalaridagi matnli ma’lumotlar bilan bilan ishlaydi. Ammo axborot qidirish va matnlarni avtomatik qayta ishlash dasturlarida qo‘llaniladigan lingvistik va ontologik bilimlar (dunyo haqidagi bilim)ning yetishmasligi turli muammolarga olib keladi. Bilimlar yetishmasligi, yanayam aniqroq aytganda, ushbu tizimlarda leksik ma’lumotlar bazasidagi so‘zlararo munosabatlarning bilimli mutaxassislar tomonidan to‘liq shakllantirilmaganligi ahamiyatsiz yoki zaruriylik darajasi past qidiruv natijasini beradi. Ma’lum bo‘lganidek, ontologik baza yaratishdagi bilimlarning yetishmasligi yoxud tizimning mukammal shakllantirilmaganligi satri uzun so‘rov (kengaygan so‘z birikmasi, kengaygan gap, yoyiq nom)larni qayta ishlashda, savol-javob tizimlarida savollarga javob izlashda murakkablashadi.

“Ontologiya” atamasi ko‘plab sohalarda qo‘llaniladi va ikki xil ma’noga ega:

1) “borliq” va “mohiyat”ni o‘zida namoyon etuvchi falsafiy tushuncha;

2) elementlarning mazmunini tavsiflaydigan, ular o‘rtasida tarmoqli munosabat o‘rnatilgan tizim.

Ma’lumotlar manbalari

* Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti filologiya fanlari bo‘yicha falsafa doktori (PhD).manzura_ok@mail.ru. ORCID 0000-0002-1927-2669

Dunyo va til bilimlariga ega bo‘lish uchun va ularni tavsiflash uchun sohaning holati to‘g‘risida to‘liq tasavvur beradigan darsliklarga murojaat qilinadi. Darsliklar yaxshi mutaxassis tajribasi ifodasidir. Til korpuslari, ayniqsa, ta’limiy korpusning o‘ziga xos xususiyati shundaki, u nafaqat darsliklar, balki darsliklarning yaratilishiga zamin bo‘lgan manbalarni ham qamrab oladi. Milliy til korpusida esa tabiiy tilning barcha jabhasi qamrab olinadi, shu bois til korpuslari lingvistik ontologiyalarda semantik tarmoqlar va sinset (ma’nodosh so‘zlar tarmog‘i)ni yaratishda, so‘zlarga misollarni taqdim etishda, muayyan so‘zning sohalararo qo‘llanilish ko‘lamini aniqlashda muhim manba hisoblanadi.

Tilshunoslikda ontologiya masalasi. Ontologiya tushunchasi uzoq vaqtdan beri ma’lum, ammo qayta ko‘rib chiqilgan holda, u yaqin yillardan kompyuter texnologiyalarida faol qo‘llanilmoqda. Ontologiyaga semantik tarmoq sifatidagi qarashlar XX asrning 90-yillar oxirlarida boshlangan.

“Lingvistik ontologiya” yoxud “til ontologiyasi” atamalari tilshunoslik faniga qaraganda, axborot texnologiyalari sohasida ko‘proq qo‘llaniladi, asosan, matnlarni avtomatik tarzda qayta ishlash uchun ixtisoslashtirilgan axborot qidirish tezaurusi, ya’ni tilning lug‘at boyligini o‘zida jamlagan, so‘zlarning semantik munosabatlari o‘rnatilgan (yoxud so‘zlar tarmog‘iga ega) turli maxsus lingvistik dasturiy ta’minotlarni anglatadi. Lingvistik ontologiya (LO) til borligi va mohiyati haqidagi fan sifatida kamdan-kam tilga olinadi. LO lisoniy borliqni tahlil qilish orqali tilning mohiyatini ochib berishga xizmat qiladi. Tilshunos F. de Sossur ta’kidlaganidek, “... *tilshunoslikning maqsadi tilning mohiyatini hech qanday cheklovlarsiz anglash, inson tilini uning paydo bo‘lish tarixi va lingvistik xilma-xillik sabablari bilan birgalikda tilning barcha ko‘rinishlari va aloqalarida hamda nutqda namoyon bo‘lish shakllarida o‘rganish hisoblanadi*” [Sossur, 2000. 171]. Tilning mohiyatini esa, birinchi o‘rinda, nutq, so‘zlash tashkil etadi. Nemis faylasufi M.Xaydegger ta’biri bilan aytganda, “*Tilning ekzistensial-ontologik asoslarini nutq tashkil etadi*” [Xaydegger, 2003. 187].

Borliqning asosiy sohalariga tabiat, jamiyat va ong kiradi [To‘rayev, 2011. 5]. Lingvistik ontologiyalarda ham tabiiy til boyligi, undan foydalanish imkoniyati va lison qamrab olinadi. Hozirgi vaqtda bilim bazalarining eng keng tarqalgan shakli ontologik tipdagi bilimlar bazasi hisoblanadi. Bugungi raqamli texnologiyalar davrida ontologiya termini bir muncha ommalashdi. Ontologiyalar – bu dunyo haqidagi bilimlarning rasmiylashtirilgan tavsifini o‘z ichiga olgan kompyuter resurslari.

Ontologiya tushunchasiga ta’riflar.

Ontologiya – bu kontseptualizatsiya spetsifikatsiyasi, deydi rus tadqiqotchisi Gruber.

Kontseptualizatsiya esa predmet sohasining lug‘at va aniq vaziyatga bog‘liq bo‘lmagan holda ko‘rib chiqiladigan haqiqat tuzilishi. Masalan, stol ustidagi kubikning turishi mumkin bo‘lgan pozitsiyalari to‘plami – bu uning konseptualizatsiyasi, muhimi kubikning ayni vaqtdagi turgan holati emas, balki u turishi mumkin bo‘lgan holatlar to‘plamidir.

Lingvistik ontologiyalarni yaratish o‘zbek amaliy tilshunoslik va kompyuter lingvistikasi sohalarida yangi yo‘nalish bo‘lib, monografik planda hali chuqur tadqiq etilmagan. Ontologiyalarni yaratish va ulardan foydalanish bo‘yicha ishlarning aksariyati chet ellarda olib borilgan (kirish qismida sanab o‘tildi), shu jumladan, Rossiyada bu sohada bir qancha tadqiqot natijalari e’lon qilingan [4,5,6].

Shu o‘rinda ta’kidlash joizki, tildagi barcha so‘zlarning semantik va pragmatik xususiyatlarini yoritib berish faqat o‘lik tillar uchun to‘liq bajarilishi mumkin. Boisi o‘lik til statik holatda qolgan bo‘lib, unda taraqqiyot nolga teng bo‘ladi, ya’ni “*o‘lik til taraqqiyoti = 0*”. Natijada tilda o‘zgarish bo‘lmaydi, bunday til asosida qurilgan dasturiy ta’minot bazasini qayta yangilanishga ehtiyoj bo‘lmaydi.

Ko‘plab manbalarda tezaurus va ontologiya atamalari qiyosiy tahlil qilinmaganligi va har ikki terminga berilgan ta’rifning o‘xshashligi sababli bunday lug‘atlar imkoniyati hamda ularni yaratish mezonlari o‘z chegarasi va aniqligiga ega bo‘lmagan. Ushbu atamalarning kompyuter lingvistikasi hamda sun’iy intellekt kesishmasida parallel ravishda faol qo‘llanilishi ularning vazifalari va faoliyat yo‘nalishlarini yanada aniqroq taqsimlashni talab qiladi.

Ontologiyalar tarkibiy qismi.

Lingvistik ontologiya tushunchalar (taksonomik tarmoqlangan atamalar), ularning tavsiflari va qoidalardan iborat bo‘ladi.

Ontologiyalar ko‘plab kompyuter dasturlari uchun [Eiji Aramaki, 2005] ma’lumot manbalari sifatida qo‘llaniladi (axborot qidirish, matnni tahlil qilish, avtomatik tarjima, bilimlarni yig‘ish va boshqa axborot texnologiyalari uchun). Ontologiya murakkab va xilma-xil ma’lumotlarni samarali qayta ishlashga yordam beradi [Gladun, 2006]. Ma’lumotlar bilan ishlashning bunday usuli dasturlar uchun insonga tushunarli bo‘lgan, ammo kompyuterga ma’lum bo‘lmagan semantik farqlarni tanib olishga imkon beradi.

Ontologiyaning asosiy tarkibiy qismlarini quyidagilar tashkil etadi:

• tushunchalar;

• munosabatlar;

• vazifalar;
• aksiomalar;
• misollar.

Korpus asosida qo‘llaniladigan metod. Lingvistik ontologiyalar semantik tarmoqlarini boyitib borish uchun tilshunoslik tadqiqotining taniqli usullaridan biri matn birliklarining distributsiyasi (birikish usullari, qo‘llanish doirasi, joylashish o‘rni) va ularning sonli parametrlari haqidagi ma’lumotlarga tayanadigan distributiv-statistik tahlil usuliga asoslaniladi. Kompyuter lingvistikasining ilk davrlarida muayyan matnda leksik birliklarning uchrashi haqidagi chastotali ma’lumotlarga asoslanib, ma’lum bir formula bo‘yicha so‘z birikmalari va ko‘p so‘zli birliklar (qo‘shma so‘z, frazema)ni aniqlash uchun leksik birliklarlar bog‘lanishining miqdoriy xarakteristikasini olishga urinishlar bo‘lgan, keyinchalik bu distributiv-statistik usulda o‘z ifodasini topgan.

Lingvistik ontologiyani yaratishda til korpusining foydalanish.

Ontologiyada aynan tushunchalar izohini berishda, ontologiyani yaratish metodologiyasiga asosan ko‘p ma’noga ega so‘zning qo‘llanilish ko‘lami bo‘yicha izohlarini darajalashda, so‘zga misollar massivini taqdim etishda til korpuslarining o‘rni beqiyos. Shuningdek, til korpuslari lingvistik ontologiyalar uchun so‘zlarning paradigmatik yadrosini yoki boshqacha qilib aytganda, leksik-semantik maydonni aniqlashda muhim manba hisoblanadi.

Ma’lumki, lingvoprotsessorlar o‘z ishi jarayonida, birinchi galda, kompyuterlashtirilgan an’anaviy yohud elektron lug‘atlarga tayanadi. Bu jihatdan til korpuslarining leksikografik bazasi lingvistik ontologiyaning semantik tarmog‘ini to‘ldirishda qulay imkoniyatni beradi. Ko‘p hollarda, lug‘atlar lingvistik belgi qo‘llanishining ikki jihatini aks ettiradi – sintagmatika va paradigmatika [Zaxarov, 2015]. Leksik birliklar orasidagi paradigmatik va sintagmatik aloqalar har xil turdagi an’anaviy lug‘atlarda to‘liq bo‘lmasa ham aks ettirilgan. Bu jihatdan korpus tilshunosligi “ko‘p ma’noli birliklarni” birlashtirishtiruvchi soha hisoblanadi.

V.Zaxarov tadqiqotida matn korpuslari asosida “*dvigatel*” termini tezaurusi tuziladi va uning ko‘lami baholanadi [Zaxarov, 2015]. Dastlabki matn materiali sifatida ruTenTen 2011 korpusi (18,28 mlrd token, 14,55 mlrd so‘zshakllari) tanlangan. *Dvigatel* so‘zining korpusda uchrashi soni 2 milliondan oshadi (ppm = 113.05). Natijada quyidagi distributiv ontologiya yuzaga kelgan:

Lemma	Score	Freq	Lemma	Score	Freq
мотор	0,560	590 265	корпус	0,213	1 920 685
агрегат	0,341	415 228	датчик	0,211	647 667
движок	0,305	224 399	котел	0,211	480 105
автомобиль	0,295	5 415 679	фильтр	0,211	786 733
насос	0,288	567 423	компрессор	0,209	220 470
прибор	0,285	1 583 195	модуль	0,208	969 885
привод	0,285	472 927	деталь	0,206	1 811 720
генератор	0,285	391 057	узел	0,205	948 220
механизм	0,272	2 086 090	тормоз	0,204	317 220
устройство	0,271	3 545 701	подвеска	0,202	339 078
оборудование	0,270	4 428 495	авто	0,198	763 847
машина	0,267	5 899 922	батарея	0,198	570 385
аппарат	0,267	1 958 038	труба	0,196	1 423 759
дизель	0,265	139 042	кондиционер	0,193	584 881
блок	0,263	2 003 819	панель	0,193	1 169 684
система	0,251	18 251 572	самолет	0,193	1 720 844
колесо	0,247	1 047 174	радиатор	0,191	288 106
модель	0,244	4 919 587	турбина	0,189	124 670
установка	0,242	3 458 650	аккумулятор	0,189	400 288
электродвигатель	0,235	132 171	цилиндр	0,188	304 511
конструкция	0,234	2 308 028	часть	0,187	12 381 025
коробка	0,231	840 107	лампа	0,186	684 590
техника	0,229	3 193 906	мотоцикл	0,185	348 738
диск	0,225	1 647 056	кузов	0,185	467 488
элемент	0,223	3 537 867	тип	0,185	5 106 728
камера	0,222	1 550 599	изделие	0,184	2 161 851
инструмент	0,216	2 264 279	компонент	0,182	1 291 366
вентилятор	0,216	293 447	автомат	0,180	769 619
шина	0,215	595 074	схема	0,180	2 604 256
компьютер	0,214	2 497 874	продукт	0,180	4 639 169

1-jadval. ruTenTen 2011 korpusidagi “dvigatel” so‘zi bilan distributiv bog‘liq so‘zlar ro‘yxat

Ma’lumotlar uchta ustunda keltirilgan: *Lemma* – so‘z; Score – berilgan so‘zning kalit so‘zga semantik yaqinlik darajasini ko‘rsatuvchi statistik o‘lchov qiymati; *Frek* – korpusda berilgan so‘zning chastotasi. Jadvaldagi so‘zlar statistik o‘lchov (Score)ning qiymati bo‘yicha tartiblangan.

Ushbu jarayondan keyin ushbu ro‘yxatdan “dvigatel” so‘zi semantik yaqin bo‘lgan so‘zlarni saralashda mutaxassis ishi talab qilinadi.

V.Zaxarov til korpusining lingvistik ontologiya va tezauruslarni yaratishda korpuslar ahamiyatini aniqlash maqsadida 998 mln so‘zshaklga ega “ruTenTen 2011 sample” (ruTenTen 2011 korpusidan 15 marta kichik) korpusidan ham “dvigatel” so‘ziga semantik jihatdan yaqin yoki bog‘liq so‘zlar ro‘yxatini oladi va ikkala jadvalni solishtiradi. Natijada 1-jadval (ruTenTen 2011)dagi 60 ta so‘z bilan 52 tasi kesishadi. (2 -jadval).

ruTenTen 2011 sample	ruTenTen 2011

(998 mln so‘zshakl)	(14,55 mlrd so‘zshakl)
texnologiya	avtomat
stansiya	akkumulyator
tarmoq	radiator
protsessor	tormoz
obyekt	turbina
kompleks	ilmoq
manba	mototsikl
versiya	kuzov

2-jadval. RuTenTen 2011 korpusi va uning nisbatan kichik hajmli turi ruTenTen 2011 sample asosida tuzilgan ikkita ro‘yxatdagi leksik farqlar

Ma’lum bo‘lganidek, 2-jadvalning 2-ustunidagi so‘zlar “dvigatel” tushunchasiga semantik jihatdan muvofiq.

Xulosa

So‘nggi vaqtlarda tibbiy, ilmiy, bank-moliya, siyosiy qidiruv kabi axborot qidirishning ixtisoslashgan turlari tobora muhim ahamiyat kasb etmoqda va bunday axborot tizimlarining sifatini ta’minlashda fan sohasidagi bilimlarning o‘rni muhim. Umuman, matnga avtomatik ishlov berishning zamonaviy usullari yordamida dasturiy tizimlarga til va dunyo haqidagi bilimlarni kiritish qiyin vazifadir. Buning yechimi esa til va dunyo to‘g‘risidagi bilimlarning maxsus yaratilgan manbalar (tezaurus, ontologiyalar)da aks etishi bilan bog‘liq, bunday manbalarda o‘n minglab so‘zlar va iboralarning tavsifi, boshqa so‘z va birliklar bilan semantik munosabatga kirishish va mantiqiy xulosa chiqarish imkoniyatlari bo‘ladi. Ulardan foydalanilganda, odatda, so‘zlarning ko‘pma’nolilik, omonimlik va polifunksionallik xususiyatlari avtomatik tarzda hal qilinadi. Shuningdek, har qanday resursning bazaga kiritilishi yoxud istalgan manbaga tayanilishi fan sohasining rivojlanishiga to‘siq bo‘ladi, shu bois til va dunyo haqidagi bilimlarga asoslangan tarmoqlangan leksik ma’lumotlar bazasini yaratish, bu jarayonda bilimlarni ham, matnni qayta ishlashning eng yaxshi zamonaviy statistik usullarini ham hisobga oladigan kombinatsiyalangan usullarni ishlab chiqish zarur hisoblanadi.

Umuman, korpus hajmining oshirilishi lingvistik ontologiya, tezaurus sifatini yaxshilaydi.

Adabiyotlar:

1. Соссюр, Ф. де. Заметки по общей лингвистике / Ф. де Соссюр ; пер. с фр. ; общ. ред., вступ. ст. и коммент. Н. А. Слюсаревой. – М. : Прогресс, 2000. – С. 171.
2. Хайдеггер, М. Бытие и время / М. Хайдеггер ; пер. с нем. В. В. Бибихина. – Харьков : Фолио, 2003. – С. 187.
3. To‘rayev B.O. Borliq: mohiyati, shakllari, xususiyati: monografiya/ B.O.To‘rayev; maxs. muharrir M.N.Abdullayeva, O‘zR FA I.Mo‘minov nomidagi Falsafa va huquq instituti. – Toshkent: Falsafa va huquq instituti nashriyoti (FHIN), 2011. – B. 5.
4. Андреев А.М., Березкин Д.В., Рымарь В.С., Симаков К.В. Использование технологии Semantic Web в системе поиска несоответствий в текстах документов. //URL: http://fccl.ksu.ru/issue_spec/docs/oent-kgu.doc
5. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска. //URL: http://fccl.ksu.ru/issue_spec/docs/oent-kgu.doc.
6. Загоруйко Н.Г. и др. Система “Ontogrid” для построения онтологий // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. конференции Диалог'2005 . М., 2005. С. 146-152.
7. Захаров В. Корпусно-ориентированный подход к построению тезаурусов и онтологий // https://www.researchgate.net/publication/290820760
8. Eiji Aramaki, Takeshi Imai, Masayo Kashiwagi, Masayuki Kajino, Kengo Miyo and Kazuhiko Ohe. Toward medical ontology using Natural Language Processing. //URL: http://www.m.u-tokyo.ac.jp/medinfo/ont/paper/2005-aramaki-1.pdf
9. Гладун А.Я., Рогушина Ю.В. Онтологии в корпоративных системах, Часть II // Корпоративные системы №1 / 2006. //URL: http://www.management.com.ua/ims/ims116.html