

TURK TILSHUNOSLIGIDA YARATILGAN LUG'ATLAR TIZIMI

Sharipov Elmurod Sa'dulla o'g'li
elmurodsharipov15@gmail.com

<https://doi.org/10.5281/zenodo.10021850>

Annotatsiya. Bugungi o'zbek tilshunosligining asosiy muammolaridan biri lug'atlar yaratish masalasidir. Ushbu maqolada turk tilshunosligida yaratilgan lug'atlar, ularning yaratilishiga doir fikr va mulohazalar, ontologiyalarning o'ziga xos xususiyatlari yoritilgan.

Kalit so'zlar: ontologiya, KeNet, sinset, PWN, morfologik tahlil, leksik xususiyat

Turk tili ontologiyasi **KeNet**¹ nomi bilan mashhur. KeNet ham boshqa WordNetlar kabi o'ziga xos semantik xususiyatlariga ega. U jami 76 757ta sinsetni o'zida qamrab olgan. PWN (Princeton WordNet) bilan tillar o'rtasidagi aloqa orqali bog'langan. KeNetda sinonimlarni aniqlash va ajratish uchun ikki metoddan:

- Birlashtiruvchi;
- Ajratuvchilardan foydalanilgan.

KeNet'da birlashish jarayonida birlashtirilishi kerak bo'lgan turli xil sinset to'plamlari aniqlangan, identifikatsiya qilingan va sinsetlar bir to'plam sifatida guruhlangan. Sinsetlar to'plamini birlashtirishda uchta narsa juda ahamiyatli sanaladi:²

- 1) har bir ma'lumotlar to'plami yagona va o'ziga xos izoh/tavsifga ega bo'lishi;
- 2) har bir ma'lumotlar to'plamida sinonim a'zosi sifatida original sinonimlarning mavjud bo'lishi;
- 3) har bir ma'lumotlar to'plamida birinchi tarkibiy qismi reprezentativlangan (o'zida lingvistik ma'lumotni ifodalovchi, ko'rsatuvchi, grammatik xarakterli) bo'lishi kerak. Noto'g'ri birlashtirishning uchta asosiy omili bor:

- 1) ma'no, izohlarning turli xilligi; 2) POS, ya'ni so'z turkumlarini teglash bilan bog'liq muammo; 3) morfologiyaga bog'liq muammolar.³

Mazmun bilan bog'liq bo'lgan muammo shundaki, semantik jihatdan bir-biriga yaqin bo'lgan, ammo aynan sinonim bo'lolmaydigan (aynan, yoki bir-biri o'rnida qo'llanilmaydigan) so'zlar sinsetlarning tarkibiy qismi sifatida berilgan. Masalan, *dere* (irmoq) va *irmak nehir* (daryo) kabi o'xshash ma'nolarga ega bo'lgan otlar ideografik jihatdan o'rinli joylashgan bo'lsa-da, bir-biriga semantik muvofiqligi jihatdan noto'g'ri birlashtirilgan. POS muammosi ma'no jihatdan bir-biriga muvofiq, ammo turli xil so'z turkumlariga mansub bo'lgan sinset tarkibiy qismlari uchraganda yuzaga keladi. Masalan, *güç* (kuch, quvvat) va *güç* (qarshilik) so'zlari ot turkumiga oid, *güç* (qiyin) sifat turkumiga oid so'z bir o'zakka birlashtirilib, unga izohlar umumiy holda berilgan. Bunday holat natijasi sinset qismlarining xato birlashtirilishiga olib keladi. Morfologiya bilan bog'liq muammo esa so'zning turli morfologik shakllari bitta sinset tarkibiy qismlari sifatida noto'g'ri birlashtirilganda yuzaga kelgan. Masalan, bir xil asosga ega *sopalamak* (savalamoq) – sof fe'l, aniq nisbat, fe'lning noaniq shaklida va *sopalanmak* (savalanmoq, kaltaklanmoq) – majhul nisbatdagi fe'l turli xil bajaruvchilariga egaligini bildirishi bilan ma'no farqlariga ega, morfologik ko'rsatkichlari ham farqli bo'lsa-da, bir xil

¹ KeNet qisqartma nomidagi Ke qismi turkcha "kelime" (kalom, so'z) so'zining birinchi bo'g'ini hisoblanadi.

² Özge Bakay and others. TurkishWordNet KeNet. Global Wordnet Virtual Conference. 2021. January. – P. 166. https://www.researchgate.net/publication/348264475_Turkish_WordNet_KeNet

³ Bakay O., Ergelen O., and Yildiz O.T. 2019. Problems caused by semantic drift in wordnet synset construction. In UBMK.

guruhda birlashtirilgan. KeNet`da bunday shakllar ham ajratilib, ular uchun alohida ma'lumotlar bazasi yaratilgan. KeNet ning yaratilishida turk olimlari: Ozge Bakay, Ozlem Ergelen, Elif Sarmis, Selin Yildirim, Atilla Kocabalcioglu, Bilge Nas Arıcan, Merve Ozcelik, Ezgi Saniyar, Oguzhan Kuyrukcu, Begum Avar, Olcay Taner Yildizlarning hissalarini beqiyos.⁴

Hozirda turk tili uchun ikkita WordNet mavjud: BalkaNet TR-wordnet va KeNet. Turk tili uchun kengroq so'z tarmog'i sifatida KeNet qabul qilingan. KeNet ham til ichidagi semantik munosabatlarga ega va tillararo munosabatlar orqali PWN ga ham bog'langan. KeNet yaratilgunga qadar undan kichikroq bo'lgan ikki tizim: **Turkish Estate WordNet va Turkish Tourism WordNet** yaratilgan edi. KeNet yaratilgunga qadar GoogleSheetsda ikki ro'yxat yaratildi. Ro'yxat asosida turk tili bazasi ham WordNet va KeNetda, ham Turkish Tourism WordNetda bir xil shakllantirildi.

Domenga bog'liq bo'lgan so'z tarmoqlarida ba'zi bir giperonimiya munosabatlarini o'rnatilgandan so'ng, keng qamrovli iyerarxik daraxt shakllantirila boshlandi. Bu ishni qilish uchun avvalo eng ko'p takrorlanadigan so'zlarni yig'ish kerak edi. Natijada, eng ko'p ishlatiladigan 700 so'z ajratib olindi. Boshqa semantik qurilishda bo'lgani kabi so'zlar o'rtasidagi mumkin bo'lgan giperonim munosabatlar asosida KeNetdagi sinsetlar ajratib olindi.

Ingliz tilidagi lug'at tarjimalari, PWN va PWN dagi sinsetlar o'rtasidagi munosabatlar ro'yxati Google Sheets va inson annotatorlari tomonidan birma-bir tekshirilib mumkin bo'lgan giperonomiya munosabatlari aniqlab chiqildi.

To'g'ri bo'lgan so'zlar munosabatlari KeNet ga qo'shildi. Kichik tadqiqotlarni birlashtirish katta guruhlashni tashkil etishning muhim bosqichi bo'ldi. Oldingi jarayonlardan giperonomiya munosabatiga ega bo'lgan "butun birlik" ostida joylashtirildi. KeNet yaratishda ko'plab til bilan bog'liq muammolarga duch kelindi. Sinonimiyani belgilovchi turkcha tizimli lug'at mavjud bo'lmaganligi bois sinsetlarni yaratishda baza bilan bog'liq muammoni bartaraf etish masalasini birinchi hal qilish lozim edi. KeNet`dagi ba'zi sinsetlar turli POS`ga ega sinset a'zolarini o'z ichiga oladi. Turk tilida o'xshash, ammo ma'nosi tubdan farq qiladigan so'zlar sinonimlarning noto'g'ri guruhlanishiga sabab bo'lishi mumkin edi. Shuning uchun yaratilgan sinsetlar bazasidagi har bir o'xshash so'zning alohida sinsetini yaratish uzoq mehnat va ko'p vaqt talab qildi. Turk tili ontologiyasining iyerarxiyasini yaratishda xuddi daraxtning ildizi va shoxini ulovchi ma'lum bir tizim yo'q edi. Shuning uchun bu tizim boshidan yaratildi va KeNet tizimiga ulandi. Bundan tashqari, KeNet qurilishida faqat birma-bir xaritalash to'g'ri bo'lmadi, bu hodisa tillararo munosabatlar bilan bevosita bog'liq masala edi. Turk va ingliz tilini solishtirib, birdaniga ko'plab xaritalash jarayoni amalga oshirildi. Boshqa WordNet lug'atlarning yaralishidan farqli o'laroq, KeNetning yaratilishidan avtomatik tahrir, mashina omilidan ko'ra inson omili, annotatorlari, ya'ni qo'lda qilingan mayda mehnatlar katta rol o'ynadi. Bu bevosita turk tilining lisoniy xususiyatlari bilan bog'liq edi. KeNetning yaratilishida turli qiyinchiliklar, boshqa tillar bilan tafovutlash, turk tilining grammatik, leksik xususiyatlaridan kelib chiqish sintez yaratish muhim omil bo'ldi. Bu tadqiqotlar turk tilining morfologik va sintaktik qurilish xususiyatlariga aniq yondashuvni talab qildi. Morfologik tahlillarda ingliz tilida mavjud xaritalash va so'zlarning leksik munosabatlariga yondashuv alohida ishlab chiqilib, bu ishni ingliz tilidan to'g'ridan-to'g'ri turkchaga ko'chirib bo'lmasdi.

⁴https://www.researchgate.net/publication/348264475_Turkish_WordNet_KeNet/link/6005e2da45851553a053b147/download

Bu muammolar tillarning ichki va tashqi semantik munosabatlarida ham namoyon bo'ldi. Ingliz tilida mavjud iyerarxiyani turk tiliga qo'lda bittalab moslab, turk tili tabiatidan kelib chiqib moslab chiqishni talab qildi. KeNet yaralishida kichik kichik leksik bazalar to'plami ahamiyatga molik bo'ldi: TRopBank, Turk SentiNet, HisNet, Turk FrameNet, va Estate (domen so'z tarmoqlari) va Turizm va turk olimlari tomonidan olib borilgan bir necha NLP tadqiqotlar.

Yuqorida ta'kidlab o'tganimizdek, bu soha bo'yicha jahon tilshunosligida turfa xil ishlar loyihalashtirilgan. Bundan tashqari, WordNet lug'atlarni yaratish, takomillashtirish bo'yicha xalqaro darajadagi an'anaviy konferensiya (International WordNet Conference) ham hozirgi kunda o'z faoliyatini qizg'in olib bormoqda. Ushbu konferensiya manbalaridan ma'lum bo'ladiki, WordNet lug'atlarini yaratish ma'lum bir sxemaga solingan bo'lib, xalqaro darajada turli davlatlar tomonidan, turli tillar bilan bog'liq ancha ishlar amalga oshirilgan. Hozirgi kunda pulli va tekin foydalanish huquqi berilgan turli ko'rinishdagi onlayn lug'atlar shular jumlasiga kiradi. Ushbu lug'atlar o'zaro bir sxema asosida yaratilgan bo'lsa-da, ammo dastur interfeysi, ishlash tizimi, so'z boyligi qamrovi, foydalanuchiga qulayligi bilan bir-biridan farq qiladi.

References:

1. Özge Bakay and others. TurkishWordNet KeNet. Global Wordnet Virtual Conference. 2021. January. – P. 166. https://www.researchgate.net/publication/348264475_Turkish_WordNet_KeNet
2. Bakay O., Ergelen O., and Yildiz O.T. 2019. Problems caused by semantic drift in wordnet synset construction. In UBMK.
3. https://www.researchgate.net/publication/348264475_Turkish_WordNet_KeNet/link/6005e2da45851553a053b147/download