

MATNGA QAYTA ISHLOV BERISH TIZIMLARI

O‘ZBEK TILI SO‘ZLARINING STEMINI TOPISH UCHUN LINGVISTIK MANBA YARATISH

Sharipov Maqsud Siddiqovich

Texnika fanlari nomzodi, dotsent,

m.sharipov@urdu.uz

UrDU, Kompyuter ilmlari kafedrası dotsenti

Amanbayeva Raxima Xushnud qizi,

rahimaamanbayeva@gmail.com,

UrDU, Kompyuter ilmlari kafedrası magistranti

Jumanazarova Onajon Matnazarovna,

anaxon54@gmail.com,

UrDU, Kompyuter ilmlari kafedrası katta o'qituvchisi

Annotatsiya. Stemlarni (soʻzning asosiy qismi yoki oʻzak soʻz) aniqlash uchun lingvistik manbalar, soʻzning tarkibiy qismi, maʼnosi, va qoʻllanilishining tushunchalarini taʼkidlaydigan, bu soʻzlarga oid maʼlumotlarni taqdim etadigan manbalardir. Bu maqola oʻzbek tili soʻzlarining tarkibini va ularning stemini aniqlash boʻyicha lingvistik manbaning yaratilish jarayonini taʼriflaydi. Maqolada tahlil va yigʻilish bosqichlari, koʻplab manbalar bilan ishlash, lugʻatlar va tahlil kitoblardan foydalanish, izlanadigan standartlarga rioya qilish, tahlil va natijalarni koʻrib chiqish, natijalarni yigʻib chiqish va yuklash bosqichlari koʻrsatilgan. Oʻzbek tili soʻzlarining stemini aniqlash uchun lingvistik manba yaratishning metodologiyasini va jarayonning muhimligini belgilaydi.

Abstract. Linguistic resources for identifying stems (the main part of a word or root word) are resources that provide information about these words, emphasizing concepts of word structure, meaning, and usage. This article describes the process of creating a linguistic resource for determining the composition of Uzbek words and their stem. The article illustrates the steps of analysis and assembly, working with multiple sources, using dictionaries and analysis books, following the standards sought, reviewing the analysis and results, collating and uploading the results. Summarizes the methodology and importance of the process of creating a linguistic resource for determining the stem of Uzbek words.

Аннотация. Лингвистические ресурсы для определения основы (основной части слова или корневого слова) — это ресурсы, которые предоставляют информацию об этих словах, подчеркивая концепции структуры, значения и использования слова. В данной статье описан процесс создания лингвистического ресурса для определения состава узбекских слов и их основы. В статье показаны этапы анализа и сборки, работа с несколькими источниками, использование словарей и книг по анализу, следование искомым стандартам, просмотр анализа и результатов, сопоставление и загрузка

результатов. Обобщены методология и значение процесса создания лингвистического ресурса для определения основы узбекских слов.

Kalit so‘zlar. *Lingvistik manba, o‘zak, stem, stemming, lug‘at, qo‘shimcha, so‘z, o‘zbek tili.*

Lingvistik manba, tilning tarkibiy elementlarini, ularning o‘zaro aloqalarini va qonuniyatlarni izlash, aniqlash va tushuntirish uchun ishlatiladigan muhim ma’lumot manbasidir. Lingvistik manbalar tilining lug‘atlarini, tarixiy manbalari, akademik tadqiqotlar va maktab darsliklarini o‘z ichiga oladi. Lingvistik manbalar, tilning aloqador so‘zlari, izohlar va ta’riflari orqali tilning o‘zini tushunishda muhim vazifalarni o‘taydi. Stem – so‘zning qo‘shimchalarsiz qismi. Masalan, “kitob” so‘zi esa “kitoblar”, “kitobni”, “kitobdan” so‘zlarining stemidir. Stemlarni (so‘zning asosiy qismi yoki o‘zak so‘z) aniqlash uchun lingvistik manbalar, so‘zning tarkibiy qismi, ma’nosi, va qo‘llanilishining tushunchalarini ta’kidlaydigan, bu so‘zlarga oid ma’lumotlarni taqdim etadigan manbalardir. Ushbu maqola o‘zbek tili so‘zlarining tarkibini, strukturasi va tarixiy rivojlanishini tahlil qilish va ularning stemini aniqlash uchun lingvistik manba yaratish jarayonini ta’kidlaydi.

Stemming uchun lingvistik manba yaratish jarayonini amalga oshirishdan oldin, avvalgi ma’lumotlar yoki oldingi bajarilgan ishlarni ko‘rib chiqish, ushbu faoliyatni boshlash va maqsadlarni belgilash uchun muhim nazariy va amaliy yordam beradi. Ushbu maqolada qo‘shimchalarni guruhlariga ajratgan holda o‘zbek tili uchun qoidaga asoslangan stemming algoritmini chekli holatlar mashinasi (ChHM) yordamida ishlab chiqish haqida so‘z boradi [Sharipov, Yuldashov, 2022]. Ushbu maqolada tabiiy tilni qayta ishlash (NLP)ning muhim yo‘nalishlaridan biri hisoblangan stemming masalasi o‘zbek tilidagi fe’l so‘z turkumi uchun qo‘llanilishi ko‘rib chiqilgan bo‘lib, o‘zbek tilidagi fe’l so‘z turkumiga mansub so‘zlarni chekli avtomatlar asosida qo‘shimchalarini qirqib olish algoritmini yaratish masalasi tushuntirilgan [Sharipov, Salayev, Matlatipov, 2021]. So‘z turkumlarini belgilash (teglash) vositasi va belgilangan ma’lumotlar bazasi haqida ushbu maqolada so‘z borgan bo‘lib, bunda ular o‘zbek tilidagi barcha 12 ta so‘z turkumlariga belgi kiritib belgilab chiqishgan va 20 ta sohaga tegishli balanslangan matnlarda o‘zlarining bazasidan va teglash usulidan foydalanib testlab, muhim natijalar olishgan va ularni jadvallar shaklida e’lon qilishgan [Sharipov, Kuriyozov, Yuldashev, Sobirov, 2023]. O‘zbek tilidagi so‘zlarni lemmasini topish uchun lemmatizatsiya algoritmi ishlab chiqish, bunda qoidaga asoslangan usuldan foydalanilgan [Sharipov, Sobirov, 2022]. Ushbu maqolada matnlarni tahlil qilishda mazmunini o‘zgartirmagan holda matndan nomuhim so‘zlarni olib tashlash masalasi katta ahamiyatga ega ekanligi va o‘zbek tilida yozilgan matn uchun avtomatik aniqlash xususiyatiga ega bo‘lgan nomuhim so‘zlarni olib tashlash, kerak bo‘lganda uni asl matnga qaytarib qo‘yish masalasi ko‘rib chiqilgan [Madatov, Sharipov, Bekchanov, 2021]. Bu maqolada amaliy loyiha doirasida yaratilgan ta’limiy korpus haqida so‘z borgan bo‘lib, ushbu korpusni yaratish tamoyillari, omillari, ma’lumotlar bazasi, modellari, axborot

tizimi, arxitekturasi va imkoniyatlari haqida ham soʻz borgan va barcha jarayonli muhim chizmalar va jadvallar orqali keltirilgan [Elov, Raupova, Abjalova, 2021]. Ushbu maqolada matnlarni belgilash zarurati, muhimligi haqida soʻz borgan boʻlib, ushbu jarayonni amalga oshirish uchun Yashirin Markov Modeli yordamida soʻzlarni teglash haqida soʻz yuritilgan, bundan tashqari jarayon Atoqli otlarni ham tanish xususiyatiga ega, bu jarayonlarni amalga oshirishda esa oʻzbek tili korpusidan foydalanilgan [Elov, Sirojiddinov, Hamroyeva, Adali, Xusainova, 2023]. Ushbu maqolada soʻzlarni ikkita asosiy qism nominal va verbal kombinatsiyasining lingvistik modeli asosida oʻzbek tili korpusini sintaktik teglash haqida soʻz boradi, modelda soʻzlar bir-biriga uch xil holatda bogʻlangan boʻlib, maqolada yana quyidagilar haqida soʻz yuritilgan, sintaktik munosabatlarni tahlil qilish va NLP uchun meta til modelini qurish [Abduraxmanova, 2020] Bundan tashqari ushbu maqola ilmiy mazmun-mohiyati ostida yotgan lingvistik manba yaratish ushuni tadqiqot obyekti sifatida bir nechta manbalardan foydalanilgan [Tojiyev, Nazarova, Tojiyeva, 2012].

Lingvistik soʻzlar, ularning tarixi va tarkibiy qonuniyatlari, uslubi va tarixi haqida maʼlumotlarni oʻrganishda asosiy maʼlumot manbalarining roli juda katta. Bu maqolada, tilshunoslik sohasidagi muhim konseptlarni va lingvistik manbalar haqida tafsilotlar keltirilib, ularning qanday qilib ilmiy tadqiqotlarda va boshqa sohalarda qoʻllanilishini tahlil qilganmiz. Lingvistik manba tildagi maʼlumotlarni izlash va tushunish uchun foydalaniladigan resurslar va manbalarni anglatadi. Quyidagi koʻrinishdagi manbalardan lingvistik manbalar hosil qilishda foydalanish mumkin:

- **tilshunoslik manbalari:** “Oʻzbek tilining tuzilishi”, “Oʻzbek tilining morfologiyasi”, “Oʻzbek tilida sintaksis”; “Oʻzbek tilining frazeologiyasi”, “Oʻzbek tilining qoʻshma soʻzlar toʻplami”;

- **lugʻatlar;**
- **internet manbalari;**
- **elektron kutubxonalar;**
- **ilmiy maqolalar;**
- **korpuslar.**

Lugʻatlar. Tildagi soʻzlar va ularning maʼnolari, koʻrinishi va tarixi haqida maʼlumotlar taqdim etuvchi kitoblar. Bu lugʻatlar odatda oʻzbek tilining tarixi va zamonaviy lugʻatlari, etimologik lugʻatlar, sinonimlar va antonimlar lugʻatlari, shuningdek, atamalardan foydalanilgan jargon va soʻzlar uchun mahsulotlar boʻlishi mumkin, masalan: “Oʻzbek tilining izohli lugʻati”, “Oʻzbek tilining imlo lugʻati” va b.

Tilshunoslik manbalari. Tilshunoslik sohasidagi tadqiqotlarda foydalanuvchilarga tilning strukturasi, oʻzgarishlarini va rivojlanishini tushunishga yordam beradi. Bu manbalar, oʻzlarining turli xil xususiyatlari va

foydalanish sohalariga ega bo'lganlik tufayli, tilshunoslik sohasidagi tadqiqotchilar uchun juda muhimdir.

Internet manbalari. Bu kabi manbalar internetdagi bloglar, web-saytlar, akademik platformalar va elektron kutubxonalardan foydalanishni o'z ichiga oladi.

Ilmiy maqolalar. Tilshunoslik sohasidagi so'nggi ilmiy o'zgarishlar, nazariy va amaliyotdagi yangiliklar, ko'nikmalar va teoriyalar haqida ma'lumot taqdim etuvchi ilmiy jurnallar va kitoblar. Ular odatda lug'atlar va tadqiqotlarda ko'rsatilgan ma'lumotlarni amaliyotda sinovdan o'tkazish uchun foydalaniladi.

Stemming - tilshunoslik (lingvistika) sohasida amal qiladigan bir tahlil metodidir. Ushbu metod ma'lum bir so'zni o'zining asosiy qismini (stemini) aniqlash va uni ko'p sonli variantlardan, ya'ni qo'shimchalar (masalan, ko'plik, zamon va hokazo) qo'shilgan holatidan qo'shimchalarni olib tashlashni amalga oshiradi. So'zning stemini topish uchun lingvistik manba yaratish o'zgaruvchan so'zlar, ularning tarkibiy qonuniyatlari va tarixi bo'yicha ma'lumotlar taqdim etadi. Bu tilshunoslik sohasidagi ilmiy tadqiqotlarda foydalanish uchun muhimdir. Lingvistikada stemming uchun manba yaratishda avvalo, so'zlar tarkibining analizini amalga oshirish kerak. Bu muhim yoki kerakli so'zlar ro'yxatini yaratish va ularning tarkibiy elementlarini aniqlashga yordam beradi. Misol uchun, “*ishchi*” so'zda “*ish*” so'zining stemi, “*chi*” esa suffiksdir. Tilning ma'nolari va tarkibiy elementlari haqida ma'lumotlarni o'rganishda lug'atlar va leksikonlardan foydalanish muhimdir. Ular odatda so'zlarning stemini belgilovchi to'plamlar, so'zlar tarkibini tahlil qiladigan muhim manbalar sifatida xizmat qiladi. Tildagi so'zlar va ularning o'zgartirishlarini tahlil qilish uchun korpuslar va tilning monologlaridan foydalanish ham muhimdir. Bu amaliy ma'lumotlarga ega bo'lish va so'zning har xil kontekstlaridagi o'zgarishlarni tushunishga yordam beradi. Stemming uchun lingvistik manba yaratish jarayonida, muhim bo'lgan so'zlarni, ularning tarkibiy qonuniyatlari va tarixi bo'yicha tahlil qilish uchun tajribali tadqiqotchilarning yordamiga ega bo'lish ham juda muhimdir.

Ushbu maqola o'zbek tili so'zlarining stemini aniqlash uchun lingvistik manba yaratishning metodologiyasini va jarayonning muhimligini belgilaydi. Bundan tashqari ushbu maqola o'zbek tilining lingvistik tahlilida va izlanishida muhim ma'lumotlar bilan ta'minlangan va tilshunoslik sohasidagi ilmiy tadqiqotchilar uchun muhimdir.



Foydalanilgan adabiyotlar:

1. Sharipov M, Yuldashov O. UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language. CEUR Workshop Proceedings, CEUR-WS 2022, 137 – 144.
2. Sharipov M, Salayev U, Matlatipov G. O'zbek tili fe'l so'z turkumi uchun chekli avtomatlar asosidastemming algoritmini yaratish. Computer Linguistics: Problems, Solutions, Prospects 2021; 1.
3. Sharipov M, Kuriyozov E, Yuldashev O, Sobirov O. UzbekTagger: The rule-based POS tagger for Uzbek language. arxiv preprint arXiv:230112711 2023;
4. Sharipov M, Sobirov O. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language. CEUR Workshop Proceedings, CEUR-WS 2022, 154 – 159.
5. Madatov X, Sharipov M, Bekchanov S. O'zbek tili matnlaridagi nomuhim so'zlar. Computer linguistics: problems, solutions, prospect 2021; 1: 51–55.
6. Elov B, Raupova L, Abjalova M. O'zbek tilining ta'limiy korpusi va uning imkoniyatlari. 2021; 4.
7. Boltayevich EB, Samariddinovich SS, Mirdjonovna KS, Adali E, Yuldashevna XZ. Pos Taging of Uzbek Text Using Hidden Markov Model. UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering, 2023, 63 – 68.
8. Zaynobiddin AN. Syntactic Models For Parsing Of Uzbek Corpus. Вестник науки и образования 2020; 15–19.
9. Tojiyev Y, Nazarova N, Tojiyeva G. O'zbek tilidagi ergash morfemalarning semantik-stilistik xususiyatlari (affiksial ma'nodoshlik). - Toshkent, 2012.