

UDK: 811`322
KBK: 81.320'z

O'ZBEK TILI FE'L SO'Z TURKUMI UCHUN CHEKLI AVTOMATLAR ASOSIDA STEMMING ALGORITMINI YARATISH

Sharipov Maksud Siddiqovich,

Urganch davlat universiteti Axborot texnologiyalari
kafedrasi dotsenti, texnika fanlari nomzodi.

maqsbek72@gmail.com.

Salaev Ulugbek Ikramovich,

UrDU Axborot texnologiyalari kafedrasi tayanch doktoranti.

Matlatipov Gayrat Raximbayevich,

UrDU Axborot texnologiyalari kafedrasi dotsenti,
texnika fanlari nomzodi.

Annotatsiya. Ushbu maqolada tabiiy tillarni qayta ishlash jarayoni (NLP– Natural Language Processing) ning eng muhim yo'nalishlaridan biri hisoblangan stemming masalasini o'zbek tili fe'l so'z turkumi uchun qo'llash qaraladi. O'zbek tilidagi fe'l so'z turkumiga oid so'zlar uchun qo'shimchalarni qirqib olishga asoslangan stemming algoritmini chekli avtomatlar asosida yaratish masalasi yoritib berilgan. Fe'l so'z turkumining qo'shimchalari morfologik qoidalar asosida sinflarga ajratilgan va har bir sinf uchun qo'shimchalarni qirqishga mo'ljallangan chekli avtomatlar qurilgan. Ishlab chiqilgan chekli avtomatlar asosida stemming algoritmi ishlab chiqilgan.

Kalit so'zlar: *Tabiiy tillarni qayta ishlash, stemming, chekli avtomatlar, o'zbek tili, fe'l.*

IMPLEMENTED STEMMING ALGORITHMS BASED ON FINITE STATE MACHINE FOR UZBEK VERBS

Abstract. This article presents the application of stemming, one of the most important aspects of NLP-Natural Language Processing, of the Uzbek verbs. The methodology is proposed for doing the stemming of the Uzbek verb words with a affix stripping approach whereas not including any lexicon. Verb affixes are classified into three classes and designed the finite state machines (FSMs) for each class according morphological rules. A stemming application for Uzbek verb words has been developed based the FSMs.

Keywords: *Natural Language Processing, Stemming, Finite State Machines, Uzbek language, Verb.*

Stemming – soʻzning barcha affikslarini olib tashlash orqali soʻzni qoʻshimchalarsiz holatga oʻtkazish yoxud soʻzning asosini topish jarayonidir. Stemmingdan maʼlumotlarni qidirish, mashinali tarjima qilish va matnni umumlashtirish kabi masalalarni yechishda samaradorlikni oshirish uchun oldindan ishlov berish bosqichi sifatida foydalaniladi. Shu sababli stemming eng muhim NLP masalalaridan biri hisoblanadi. Hozirgi kungacha oʻzbek tili uchun stemming masalasini yechish boʻyicha yetarli darajada algoritmlar va dasturlar ishlab chiqilmaganligi sababli bu masalani oʻrganish va yechish dolzarb hisoblanadi.

Oʻzbek tili agglutinatив strukturaga asoslanganligi va turkiy tillar oilasiga mansubligi sababli, mazkur ishda turk tili uchun yaratilgan morfologik tahlil algoritmlari oʻrganib chiqildi. Turk tili uchun hech qanday leksikondan foydalanilmagan holda affikslarni olib tashlash yoʻli bilan morfologik tahlil qilish tizimi yaratilgan. Turk tilidagi soʻzlarning yasash qoidalariga binoan affikslar sinflarga ajratilgan, har bir sinf uchun chekli avtomatlar loyihalangan va ularni asosiy chekli avtomatga birlashtirish jarayoni keltirilgan [Eryigit, 2004].

Oʻzbek tilidagi matnlarni morfologik tahlil qilish [Matlatipov, 2009; Abjalova, 2021] ishida qoidaga asoslangan morfologik tahlil uchun lugʻat tuzish maqsadida gap grammatikasini aniqlash-(DCG-Define Clause Grammar) usulidan foydalanilgan. Chunki, bu usul Prolog tilida ifodalashda, hamda lugʻatni kodlashda tushunarli va oddiy hisoblangan. Bu ishda qoʻshimchalar ikkita sinfga – yasovchi (Derivational) va bogʻlovchi (Conjugational) sinflarga ajratilgan va morfologik qoidalar oʻrnatilgan. Ish natijasida Prolog tilida UZMORPP (Uzbek Morphological Parser) dasturi yaratilgan. Bu dastur 1000 ta elementdan iborat lugʻatdan hamda 108 ta qoʻshimchadan foydalanadi.

[Bakayev, 2021] ishida o'zbek tili uchun morfologik tahlillagich arxitekturasini taklif qilingan, IDEF0 modelidan foydalangan holda funksiyalar tasvirlangan, grammatik va morfologik qoidalarining yakuniy konvertori yaratilgan. O'zbek morfologik tahlillagichi uchun so'z birikmalari, morfotaktika va morfologiya qoidalari asosida ishlaydigan tuzilma taklif qilingan.

Yuqoridagi ko'rib chiqilgan ishlardan ushbu ishning farqi o'zbek tili fe'l so'z turkumi uchun affikslarni 3 ta guruhga ajratilishi va har bir guruh uchun chekli avtomat qurilishidir. O'zbek tili fe'l turkumi uchun qo'shimchalarni olib tashlagan holda tahlil qilish tizimi uchun qo'shimchalarni ularning roli va tartibiga ko'ra tasniflandi. 1-jadvalda qo'shimchalar sinflari, mos ravishda ularning qo'shimchalari hamda shakldoshlari soni keltirilgan.

1-jadval: Qo'shimchalar sinfi

Sinf ID	Sinf nomi	Qo'shimchalar soni	Qo'shimcha va uning shakldoshlari soni
1	Zamon, shaxs-son qo'shimchalari	25	28
2	Fe'l qo'shimchalari	27	45
3	Fe'lning nisbat qo'shimchalari	13	23
Jami		68	96

O'zbek tilida ayrim qo'shimchalar ko'p shaklga ega, masalan *–gancha*, *–kancha*, *–qancha* qo'shimchalari, shu uchta qo'shimcha uchun umumiy *–gancha* qo'shimchasini ishlatamiz. Birdan ortiq shaklga ega qo'shimchalarni belgilashda $G=\{g,k,q\}$, $Y=\{a,y\}$, $K=\{k,g\}$, $Q=\{k,g,g',q\}$, $T=\{t,d\}$ belgilashlarni kiritamiz va qavs ichidagi harflarning qatnashishi ixtiyoriy holda ekanligini eslatib o'tamiz.

Fe'llar morfologik jihatdan eng boy va juda murakkab so'z turkumi bo'lib, bo'lishli-bo'lishsizlik, o'timli-o'timsizlik, tuslanish, mayl, zamon va nisbat kabi grammatik kategoriyalarga ega. Fe'l asosiga qo'shimchalarning qo'shilish tartibi *Asos + so'z yasovchi + nisbat + bo'lishsizlik + zamon + shaxs-son* shaklida bo'ladi [Muhiddinova, 2006:112].

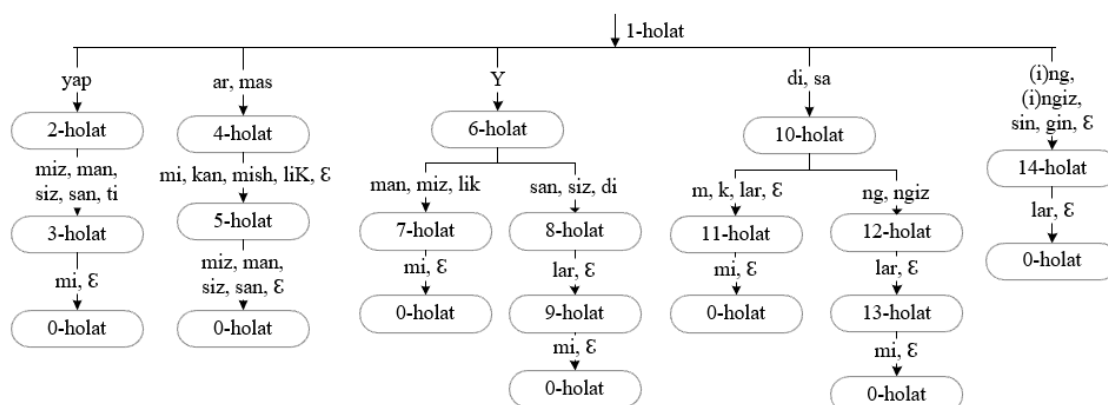
Quyida fe'l uchun uchta sinfnin chapdan o'ngga (to'g'ri tartibdagi) chekli avtomatlar loyihasi keltirilgan. Chekli avtomatdagi holatlar orasidagi ε belgisi oldingi holatdan keyingi holatga hech qanday qo'shimchani qirqmasdan, ya'ni bo'sh o'tish mumkinligini bildiradi. Quyidagi so'zlarni zamon, shaxs-son qo'shimchalarining chapdan o'ngga chekli avtomati (1-chizma) bo'yicha

yasalishiga doir misollarni qaraylik:

$o'qi(1) \rightarrow di(10) \rightarrow lar(11) \rightarrow mi(0);$

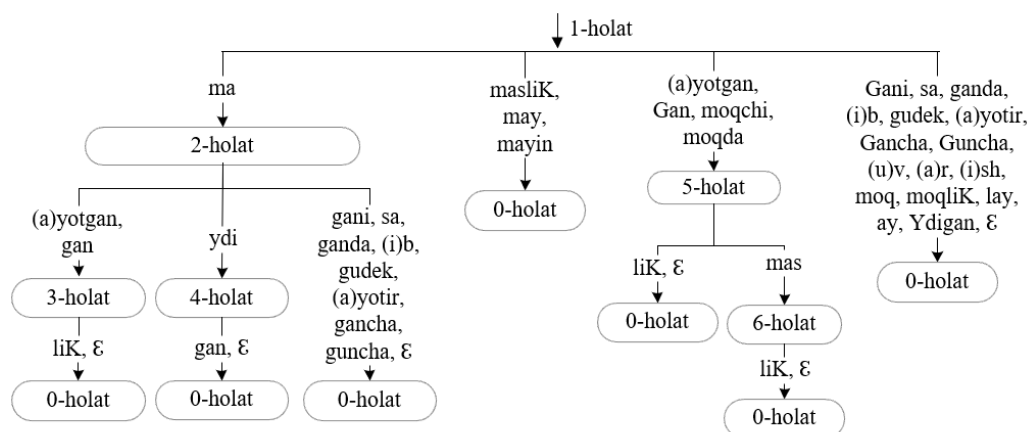
$ayt(1) \rightarrow a(6) \rightarrow siz(8) \rightarrow lar(9) \rightarrow mi(0);$

bu yerda qavs ichidagi sonlar 1-chizmadagi o'tish holatlarni bildiradi.



1-chizma: Zamon, shaxs-son qo'shimchalarining chapdan o'ngga chekli avtomati.

Fe'lning vazifaviy shakllari ravishdosh, sifatdosh va harakat nomiga ajraladi. Fe'l qo'shimchalari ravishdosh, sifatdosh va harakat nomini bildiruvchi, hamda bo'lishsizlik qo'shimchalarini o'z ichiga olgan [Muhiddinova, 2006:120]. Quyidagi 2-chizmada fe'l qo'shimchalarining chapdan o'ngga chekli avtomati keltirilgan:

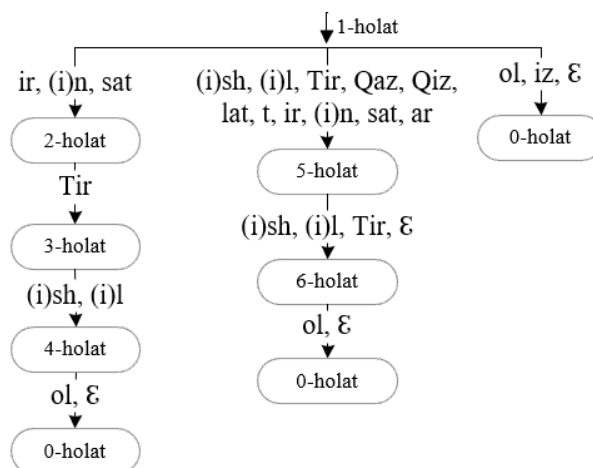


2-chizma: Fe'l qo'shimchalarining chapdan o'ngga chekli avtomati

O'zbek tilida nisbat qo'shimchalari juda faol qo'llaniladi. Bitta fe'lga ikkita yoki uchta nisbat qo'shimchasi qo'shib kelishi ham mumkin [Hojiyev, 2010]: *ko'rsattirmoq, undirtirmoq, gapirtirmoq, bezantirildi, yuvintirildi* kabi. 3-chizmada

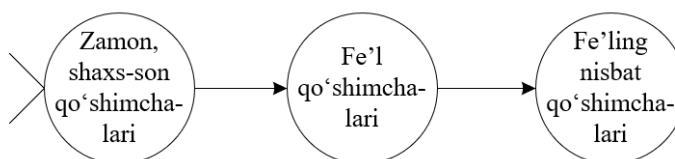
fe'lining nisbat qo'shimchalarining chapdan o'ngga chekli avtomati keltirilgan, bu avtomatning ishlashiga doir misol keltiramiz:

beza(1)→n(2)→tir(3)→il(4)→(0)
(bezantirildi so'zidagi –di qo'shimchasi 1-chizmada qoraladi);



3-chizma: Fe'lining nisbat qo'shimchalarining chapdan o'ngga chekli avtomati.

So'zni to'liq tahlil qilish uchun oldingi qurilgan chekli avtomatlarni bitta asosiy chekli avtomatga birlashtiramiz (4-chizma). Bunda tahlil qilinadigan so'z zamon, shaxs-son qo'shimchalarini aniqlovchi chekli avtomatdan kiradi va har bir chekli avtomatdan chiqqan natija keyingi chekli avtomatga kiruvchi ma'lumot bo'ladi. Bu holda har bir chekli avtomat o'ziga tegishli qo'shimchalarni qirqib boradi, agar so'zda sinfga taalluqli affiks bo'lmasa, u holda hech qanday qo'shimcha qirqilmasdan navbatdagi avtomatga o'tishi amalga oshadi.



4-chizma: Asosiy chekli avtomatni qurish

Namuna: Quyida “ko'rsatilmayaptimi” so'zini asosiy chekli avtomat orqali tahlil qilishdan so'ng chiqadigan natijani qaraymiz. Har bir qator so'ngida qo'shimchalar sinfi keltirilgan:

“*ko 'r-sat-il-ma-yap-ti-mi*”

ko 'r fe'l, o'zak

-sat	nisbat qo'shimchasi	Fe'ning nisbat qo'shimchalari
-il	nisbat qo'shimchasi	Fe'ning nisbat qo'shimchalari
-ma	bo'lishsizlik qo'shimchasi	Fe'l qo'shimchalari
-yap	zamon qo'shimchasi	Zamon, shaxs-son qo'shimchalari
-ti	3-shaxs birlik qo'shimchasi	Zamon, shaxs-son qo'shimchalari
-mi	so'roq yuklamasi	Zamon, shaxs-son qo'shimchalari

Izoh: Mazkur loyihalarga asosan stemming masalasi dasturiy ta'minotini yaratishda har bir sinf uchun affikslar jadvalini, tartiblanmagan chekli avtomatni (NFA – Non-deterministic finite automaton) va tartiblangan chekli avtomatni (DFA – Deterministic finite automaton) qurish hamda tartiblangan chekli avtomatlarni 4-chizmaga asosan birlashtirish talab qilinadi.

Morfologik tahlil imloni tuzatish vositalari, mashinali tarjima tizimlari va lug'at vositalari kabi tabiiy tilni qayta ishlash tizimlarining muhim komponentasi hisoblanadi. O'zbek tilidagi gaplarda so'zlar ko'p sonidagi qo'shimchalar qo'shish bilan yasaladi. Shu sababli, ma'lumotlarni ajratish tizimida tezkor bajariluvchi stemming algoritmi zarur bo'ladi. Odatda webga asoslangan tizimlarda tezkor bajarilish shartlari mavjud bo'lgan holda so'zning o'zagini topish uchun leksikondan (leksikon katta hajmda bo'lishi va uning ustida ish ko'rish algoritmi bajarilish samaradorligiga ta'sir ko'rsatishini inobatga olgan holda) foydalanish nomaqbul holat hisoblanadi.

Mazkur ishda o'zbek tilidagi fe'l turkumidagi so'zning o'zak qismini aniqlash maqsadida tilning morfologik qoidalari asosida qo'shimchalarni qirqish algoritmi ishlab chiqildi. Bu model yondoshuviga asosan, o'zak morfemani ajratib olishda leksikondan foydalanilmaydi. O'zakni ajratib olish maqsadida, fe'l so'z turkumiga doir asosiy qo'shimchalar uchta sinfga ajratildi va ularning o'zaro bog'liqligi o'rnatildi. Har bir sinf uchun chekli avtomatlar yaratildi. Chekli avtomatlar tilning morfologik qoidalari asosida qo'shimchalar qo'shilishi orqali yasaladigan so'zning strukturasi va qo'shimchalarni qirqishning teskari tartibdagi modelini tasvirlaydi. Mazkur sinflar bir-biri bilan o'zaro bog'langan

holda fe'l uchun asosiy chekli avtomatni tashkil qiladi va tahlil uchun kiritilgan so'zni o'zak va qo'shimchalarga ajratadi.

Mazkur model fe'l uchun chapdan o'ngga tartibidagi chekli avtomat loyihalarini o'z ichiga oladi va ishda yaratilgan modelni so'zlarni morfologik tahlil qilish orqali o'zbek tilidagi hujjatlardan ma'lumotlarni ajratib olish jarayonlarida, shuningdek stemming masalasining yechilishida qo'llash hamda rivojlantirish mumkin.

Foydalanilgan adabiyotlar:

- [1] Gülşen Eryiğit & Eşref Adalı, An affix stripping morphological analyzer for Turkish, Proceedings of the IASTED International Conference Artificial Intelligence and Applications, Innsbruck, Austria, 2004, 299-304
- [2] Gayrat Matlatipov and Zygmunt Vetulani. Representation of Uzbek Morphology in Prolog, Aspects of Natural Language Processing, 2009
- [3] Ismailov A, Abdul Jalil M M, Abdullah Z and Abd Rahim N H 2016 A comparative study of stemming algorithms for use with the Uzbek language Computer and Information Sciences (ICCOINS) (Kuala Lumpur: IEEE) pp 7-12
- [4] I.I.Bakaev, R.I.Bakaeva, Creation of a morphological analyzer based on finite-state techniques for the Uzbek language 2021 J. Phys.: Conf. Ser. 1791 012068
- [5] Ayla Kayabaş, Helmut Schmid, Ahmet E. Topcuz, Özkan Kiliç. TRMOR: a finite-state-based morphological analyzer for Turkish. Turk J Elec Eng & Comp Sci, (2019) 27: 3837 – 3851
- [6] Azim Hojiyev, O'zbek tili morfologiyasi, morfemikasi va so'z yasalishining nazariy masalalari. Toshkent-2010,
- [7] X.Muhiddinova, D.Xudoyberganova. Hozirgi o'zbek adabiy tili, Toshkent, "O'qituvchi", 2006.
- [8] Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. [Matn] : monografiya / M.A. Abjalova. – Toshkent: Nodirabegim, 2020. – B. 38-39.