ISSN 2181-922X

O‘ZBEKISTON TIL VA MADANIYAT

UZBEKISTAN LANGUAGE & CULTURE

2023 Vol. 2

www.navoiy-uni.uz
www.uzlc.navoiy-uni.uz

ISSN 2181-922X

O‘ZBEKISTON:

TIL VA MADANIYAT

UZBEKISTAN:

LANGUAGE AND CULTURE

2023 Vol. 2

www.navoiy-uni.uz
www.uzlc.navoiy-uni.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Bosh muharrir: Shuhrat Sirojiddinov

Bosh muharrir o‘rinbosarlari: Nodir Jo'raqo'ziev
Ziyoda Teshaboyeva

Mas’ul kotiblar: Ozoda Tojiboyeva
Sanjar Mavlyanov

“O‘zbekiston: til va madaniyat” jurnali – lingvistika, tarix, adabiyot, tarjimashunoslik, san’at, etnografiya, falsafa, antropologiya va ijtimoiy tadqiqotlarni o‘rganish kabi sohalarni qamrab olgan akademik jurnal.

Jurnal bir yilda to‘rt marta chop etiladi.

Jurnalning maqsadi – ko‘rsatilgan sohalarga oid dolzarb mavzulardagi bahs-munozaraga undaydigan, yangi, innovatsion g‘oyalarga boy, o‘z konsepsiyasiga ega bo‘lgan tadqiqotlarni nashr etishdir.

Ingliz, rus va o‘zbek tillaridagi, shuningdek, boshqa turkiy tillarda yozilgan maqolalar qabul qilinadi. Iqtisodiy tahlillar hamda siyosatga oid maqolalar e’lon qilinmaydi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e’lon qilinadi. Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103.

Email: uzlangcult@gmail.com
Website: www.uzlc.navoiy-uni.uz

"Uzbekistan: Language and Culture" is an academic journal that publishes works in the field of linguistics, history, literature, translation studies, arts, ethnography, philosophy, anthropology and social studies.

The journal is published four times a year.

The purpose of the journal is to publish the results of the latest research that are rich in new, innovative ideas and has its own concept, which stimulates debate on topical issues in these areas.

The language of articles can be English, Russian and Uzbek. Other Turkic languages are also welcome. We do not publish economic analyses or political articles.

In addition to research articles, the journal announces book and literary work reviews, conference reports and research project results.

The authors' ideas may differ from those of the editors'.

Alisher Navo'i Tashkent State University of the Uzbek Language and Literature.
103, Yusuf Khos Hojib, Yakkasaray, Tashkent, Uzbekistan.

Email: uzlangcult@gmail.com
Website: www.uzlc.navoiy-uni.uz

MUNDARIJA

Lingvistika

Adabiyotshunoslik

CONTENT

Linguistics

Literature

ISSN 2181-922X 6—39

LINGVISTIKA
LINGUSTICS

Agglutinativ tillar uchun pos teglash va stemming masalasi (turk, uyg‘ur, o‘zbek tillari misolida)

Botir Elov[1],
Shahlo Hamroyeva[2],
Oqila Abdullayeva[3],
Zilola Husainova[4],
Nizomaddin Xudayberganov[5]

Abstrakt

Agglutinativ tillarda mumkin bo‘lgan so‘z shakllari soni nazariy jihatdan cheksiz hisoblanadi. Bu o‘z navbatida agglutinativ tillarda lug‘atdan tashqari (out-of-vocabulary, OOV) so‘zlarni POS teglash (part-of-speech) muammosini yuzaga keltiradi. Agglutinativ tillarda o‘zak va qo‘shimchalarni birlashtirib so‘z hosil qilinadi. O‘zakka qo‘shimchalar qo‘shilganda fonetik uyg‘unlik va disgarmoniya yuzaga kelgani uchun

[1]*Elov Botir Boltayevich* – texnika fanlari bo'yicha falsafa doktori (PhD), dotsent, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2]*Hamroyeva Shahlo Mirdjonovna* – filologiya fanlari doktori (DSc), dotsent, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: shaxlo.xamrayeva@navoiy-uni.uz
ORCID: 0000-0002-5429-4708
[3]*Abdullayeva Oqila Xolmo‘minovna* – filologiya fanlari bo‘yicha falsafa doktori (PhD), Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: abdullayeva.oqila@navoiy-uni.uz
ORCID: 0000-0002-2524-4832
[4]*Husainova Zilola Yuldashevna* – tayanch doktorant, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: xusainovazilola@navoiy-uni.uz
ORCID: 0000-0003-4357-7515
[5] *Xudayberganov Nizomaddin Uktamboy o‘g‘li* – o‘qituvchi, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: nizomaddin@navoiy-uni.uz
ORCID: 0000-0002-6213-3015
Iqtibos uchun: Elov, B.B., Hamroyeva, Sh.M., Abdullayeva, O.X., Husainova, Z.Y., Xudoyberganov, N.U. 2023. “Agglutinativ tillar uchun pos teglash va stemming masalasi (turk, uyg‘ur, o‘zbek tillari misolida)”. *O‘zbekiston: til va madaniyat* 2: 6–39.

ham fonetik, ham morfologik o‘zgarishlarni tahlil qilish zarur.

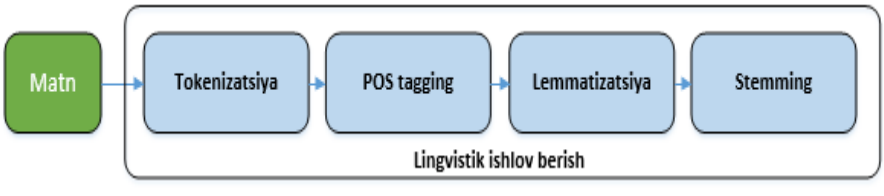
Ko‘pgina NLP vazifalarni hal qilishda so‘z shakllarini ularni o‘zakkacha qisqartirish (stemlash)ga to‘g‘ri keladi. So‘zdan barcha flektiv affikslarni olib tashlash va so‘zning qolgan qismini lemmatizatsiya qilish tabiiy tilni qayta ishlash (NLP)ning muhim vazifalaridan biri hisoblanib, ushbu jarayon **stemming** deb yuritiladi. Stemming jarayoni axborot qidirish (IR, Information Retrieval) tizimlarida muhim ahamiyat kasb etadi.

Kalit so‘zlar: *nutq qismlarini belgilash, POS teglash, stemming, axborot qidirish, IR, stemming algoritmlari*.

Kirish

Axborot qidirish tizimlarida foydalanuvchi so‘roviga mos natijani qaytarish tezligini oshirish eng muhim masala hisoblanadi. Buni amalga oshirishning eng oson va qulay usuli stemming jarayonidir. NLPda so‘zning turli morfologik variantlarini ularning umumiy shakl (o‘zak, stem)ini aniqlaydigan metod **stemming algoritmi** deyiladi [Paice, 1994]. Axborot qidiruv tizimlarida o‘zakni aniqlash uchun uning suffiks va prefiks (qo‘shimcha)larini olib tashlash lozim [Anjali, Jivani, 2011].

POS teglash – berilgan gapdagi har bir so‘zshaklga uning turkum (*ot, fe’l, sifat, son, ravish* yoki *olmosh)*ga mansubligini belgilash (teglash) vazifasidir. POS teglash tabiiy tilni qayta ishlash (Natural Language Processing, NLP)ning asosiy vazifalaridan bo‘lib, pipeline konveyerining muhim bosqichi hisoblanadi (1-rasm).



1-rasm. Matnga boshlang‘ich ishlov berish bosqichlari.

Mashina tarjimasi, matnni umumlashtirish, savol-javob va hissiyotlarni tahlil qilish kabi NLP ilovalari uchun POS teglash muhim qadam hisoblanadi. Masalan, “olma” so‘zini boshqa tilga tarjima qilish uning POS tegidan foydalaniladi. “Olma” (apple) ot so‘z turkumiga mansub bo‘lsa predmet bo‘ladi, “ol-ma” (don’t take) fe’l so‘z turkumiga mansub bo‘lsa, u harakatni bildiradi.

POS teglash lug‘at asosida yoki lug‘atsiz amalga oshirilishi mumkin. POS teglash bo‘yicha amalga oshirilgan ilmiy tadqiqotlarning aksariyati [Gao, Johnson, 2008; Goldwater, Griffiths, 2007; van Gael, Vlachos, 2009] so‘zlarga asoslangan (word-based) bo‘lib,

so‘zlarning morfologik segmentatsiyasini (morphological segmentation) amalga oshirmaydi.

Agglutinativ tillarning ba‘zilarida POS teglash jarayonini amalga oshirish uchun so‘zlarning stemlaridan foydalaniladi [Dincer, Karaog’lan, 2003]. O‘zbek, turk va uyg‘ur tillaridagi so‘zlar va uning stemi turli POS tegga mansub bo‘lishi mumkin.

1-jadval. Agglutinativ tillaridagi so‘zshaklning lemmasi, stemi va POS tegi.

№	so‘z	lemma	POS	Stem	POS	Root	POS
O'zbek tili							
1	muzladi	muzlamoq	VB	muz	N	muz	N
2	issiqroq	issiq	JJ	isi	VB	isi	VB
3	sodda-lashtiriladi	sodda-lashtirmoq	VB	sodda	JJ	sodda	JJ
4	ixtiyoriy	ixtiyoriy	JJ	ixtiyor	N	ixtiyor	N
5	qo‘llaniladi-gan	qo‘llamoq	VB	qo‘l	N	qo‘l	N
6	yo‘lakda	yo‘lak	N	yo‘l	N	yo‘l	N
7	qishlog‘im	qishloq	N	qishlog‘	?	qishloq	N
Turk tili							
7	yetkili	yetkili	ADJ	yetkili	ADJ	yetki	N
8	kurullarim-izla	kurul	N	kurul	N	kurul	N
9	teşkilatlari-mizla	teşkilat	N	teşkilat	N	teşkil	N
10	seçimlere	seçim	N	seçim	N	seç	VB
11	futbolcularin	futbolcu	N	futbol-cu	N	futbol	N
12	kullandi	kullanmak	F	kulla	F	kulla	F
13	bilgi	bilgi	N	bilgi	N	bil	F
Uyg'ur tili							
14	tarazichi	tarazichi	N	tarazi-chi	N	tarazi	N
15	yashaptu	yashamaq	VB	yasha	VB	yash	N
16	yegizligi	yegizlik	N	yegizlig	N	yegiz	VB
17	og’urluqqa	og’urluq	N	og‘ur-luq	N	og‘ur	N
18	chyshkən-ligini	chyshkənliq	N	chysh-kənlig	N	chysh	?

Agglutinativ tillar uchun pos teglash va stemmingni amalga oshirish uchun ba‘zi terminlar izohini keltiramiz:

O‘zak (root) – so‘zning asl ma‘nosini bildirib, boshqa ma‘noli qismlarga bo‘linmaydigan, mustaqil holda leksik ma‘no bildiradigan eng kichik qism. So‘zga turlicha affikslar qo‘shilib kelganda ham, o‘zakning ma‘nosi yo‘qolmaydi, undan yasalgan so‘zlarning ma‘nosi ana shu ma‘no bilan bog‘langan bo‘ladi. Shuningdek, o‘zak boshqa ma‘noviy qism, ya‘ni morfemaga bo‘linmaydigan qism.

Lemma (leksema) – faqat o‘zakdan yoki o‘zak+so‘z yasovchi qo‘shimcha shaklidan iborat bo‘ladi. **Leksema** (yun. lexis – so‘z, ifoda) – til qurilishining leksik ma‘no anglatuvchi lug‘aviy birligi. Leksema bildiradigan ma‘no so‘zning material qismi: ma‘lum tovush kompleksini ma‘lum obyektiv voqelikka bog‘lash bilan kishi ongida yuzaga keladigan mazmun-mundarija.

Stem – so‘zshaklning qo‘shimchalarini kesib tashlashdan hosil bo‘luvchi qism bo‘lib, ba‘zi hollarda ma‘no anglatmasligi mumkin. Shuningdek, stem so‘zning morfologik o‘zagi bilan aynan mos bo‘lmasligi yoki mos tushishi mumkin.

Turk va uyg‘ur tillarida stemming jarayoniga quyidagicha ta‘rif berilgan:

Stemming (turk va uyg‘ur) – bu so‘zga qo‘shilgan *flektiv* qo‘shimchalarni olib tashlash orqali uning o‘zagigacha qisqartirish vazifasidir. Quyidagi 2-jadvalda o‘zbek, turk hamda uyg‘ur tillaridagi so‘zlar, ularning stemi va o‘zagiga qo‘shilgan so‘z yasovchi hamda shakl yasovchi qo‘shimchalarning namunalari keltirilgan.

2-jadval. Agglyutinativ tillaridagi so‘zshaklning stemi va qo‘shimchalari.

Til	**So‘zshakl**	**Stem**	**So‘z yasovchi qo‘shimcha**	**Shakl yasovchi qo‘shimcha**
UZ	ko‘zlagan = ko‘z + la + gan tinchimiz = tin + ch + imiz bilimdon = bil + im + don birlik = bir + lik moyladim = moy + la + di	ko‘z tin bil bir moy	la ch im lik la	gan imiz don bir di +m
TR	oyuncularin = oyun+cu+lar+in futbolcularin = fut-bol+cu+lar+in karşilaşmalar = karşi+laş+ma+lar değerlendirilip = değer+len+dir+il+ip açikladi = açik+la+di	oyun futbol karşi değer açik	cu cu laş len la	lar+in lar+in ma+lar dir+il+ip di

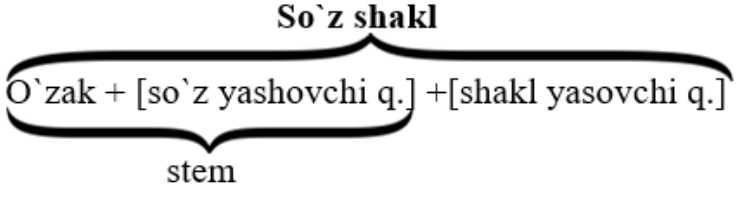
UY	tarazichi = tarazi+chi yashaptu = yasha+p+tu yegizligi = yegiz+lig+i og'urluqqa = og'ur+luq+qa chyshkənligini = chyshkən+lig+i+ni	tarazi yash yegiz og'ur chyshkən	chi a lig luq lig	- p+tu i qa i+ni

Biroq o'zbek tili uchun stemming jarayoni quyidagicha ta'riflanadi:

Stemming (o'zbek) – bu so'zga qo'shilgan *derivatsion* va *flektiv* qo'shimchalarni olib tashlash orqali uning o'zagigacha qisqartirish vazifasidir.

O'zbek, turk va uyg'ur tillarida gaplar alohida so'zlardan tashkil topadi. Morfologik jihatdan bu uch tildagi so'zlar o'zakka ba'zi qo'shimchalar qo'shish orqali hosil qilinadi. Bu jarayonda so'zda fonetik o'zgarishlar (phonetic harmony) yuzaga kelishi mumkin va bu bevosita matnda o'z aksini topadi. O'zakning o'zi ham so'zning o'ziga xos ma'nosini ifodalovchi so'z bo'lishi mumkin. Affikslar gapda muhim rol o'ynasa-da, mustaqil ma'noga ega emas.

Affikslar *so'z yasovchi (derivational suffixes)* va *shakl yasovchi (inflectional suffixes)* turga ajratiladi [Hojiyev 2005]. Turk va uyg'ur tillarida so'z yasovchi qo'shimchalar yangi stem hosil qilishi mumkin (2-rasm). Shakl yasovchi qo'shimchalar esa so'zning faqat grammatik vazifasini o'zgartiradi. So'z yasovchi qo'shimchalarni o'zakka qo'shish orqali so'zda semantik o'zgarish yuzaga kelishi mumkin. Shakl yasovchi qo'shimchalar so'zda sintaktik o'zgarishlarni keltirib chiqaradi. O'zakka avval so'z yasovchi qo'shimchalar, so'ngra shakl yasovchi qo'shimchalar biriktiriladi. Biroq o'zakka to'g'ridan-to'g'ri shakl yasovchi qo'shimchalar biriktirilishi ham mumkin.



2-rasm. Turk va uyg'ur tillaridagi so'zning umumiy morfologik tuzilishi.

Ingliz tili kabi ba'zi flektiv tillarda bu atributlar **predloglar** kabi alohida so'zlar bilan belgilanadi va tabiatan juda oddiy [Porter 2006]. Ingliz tilida so'z cheklangan miqdordagi qo'shimchalarni (odatda bitta) olishi mumkin. Shu sababli, ingliz tili uchun aso-

siy stemming algoritmlari juda oddiy. Biroq, agglutinativ tillarda bir ildizdan son-sanoqsiz leksik shakllarining hosil bo‘lishi, tabiiy tilni tushunishda muhim va murakkab masala hisoblanadi.

Turk va uyg‘ur tillarida ildizlar so‘z yasovchi qo‘shimchalari bilan birgalikda stem (o‘zak)larga aylanadi. Agglutinativ tillarda shakl yasovchi qo‘shimchalar odatda so‘z yasovchi qo‘shimchalardan keyin keladi. Biroq ba‘zi hollarda ***-gil, -siz*** kabi shakl yasovchi qo‘shimchalar avval kelishi mumkin.

O‘zbek tilida so‘z yasovchidan oldin lug‘aviy shakl yasovchi kelishiga misol sifatida ***o‘chirg‘ich, muzlatkich*** so‘zi misol bo‘la oladi.

– o‘chirg‘ich = ***o‘ch*** *(o‘zak)+ir (lug‘aviy shakl yasovchi) +g‘ich (so‘z yasovchi);*

– muzlatkich = ***muz*** *(o‘zak) +la (so‘z yasovchi) +t (lug‘aviy shakl yasovchi) +kich (so‘z yasovchi).*

Turk tilida o‘zakdan keyin so‘z yasovchi+lug‘aviy shakl yasovchi+so‘z yasovchi shakli uchraydi:

*– baş+la+n+****gıç;***

Shuningdek, o‘zak+sintaktik shakl yasovchi+so‘z yasovchi shakli ham uchraydi:

– aşağıdaki (sorular), aşağıdakiler, sınıftaki (öğrenciler), sınıftakiler, raftaki (eşyalar), yuvadaki [www.turkedebiyati].

Uyg‘ur tilida ham bu holat, ya‘ni tarkibi o‘zak + so‘z yasovchi + shakl yasovchi tartibiga mos tushmaydigan, ya‘ni o‘zak + shakl yasovchi + so‘z yasovchi tartibida bo‘ladigan so‘zlar uchraydi:

– oqu+t-quchi; qolla+n-ma [www.tilachar.ru].

So‘zga qo‘shila oladigan qo‘shimchalar soni va ularning ko‘p sonli birikmalari agglutinativ tillarda o‘zakni aniqlash jarayonini murakkab muammoga aylantiradi. Chunki ko‘pchilik agglutinativ tillarda qo‘shimchalar kombinatsiyasi murakkab so‘z shakllarini hosil qiladi. Yangi so‘z yasovchi yoki so‘z shakllarini hosil qiluvchi qo‘shimchalar soni bo‘yicha ko‘rsatkichlar ham har xil hisoblanadi. O‘zbek tilida 228 ta so‘z yasovchi, 69 ta lug‘aviy shakl hosil qiluvchi va 41 ta sintaktik shakl hosil qiluvchi qo‘shimchalar mavjudligi darsliklarda ko‘rsatiladi. Agar ularning variantlarini ham qo‘shsak, aslida bu son yana oshib ketadi [https://github.com/KhZilola/Python-Codes].

Turk tilida so‘z yasovchi qo‘shimchalar soni 24 ta, shakl hosil qiluvchi qo‘shimchalar soni 303 ta ekanligi manbalarda ko‘rsatiladi [https://en.wiktionary.org/wiki/Appendix:Turkish_suffixes]. Uyg‘ur tilida 11 ta so‘z yasovchi, 100 dan oshiq shakl hosil

qiluvchi qo‘shimchalar mavjud [https://en.wiktionary.org/wiki/Category:Uyghur_suffixes].

Yuqoridagi 2-jadvaldan ko‘rinib turganidek, o‘zbek, turk va uyg‘ur tillarida stem va lemmaga qarash turlicha. O‘zbek tilida lemma tub yoki yasama so‘z shaklida bo‘ladi: *kitob, kitobxon, bilim, bilimdon*. Demak, o‘zbek tilida lemma lug‘atda mavjud leksemaga teng keladi. O‘zbek tilida o‘zakdosh (asosdosh) so‘zlar alohida-alohida lemma sanaladi.

O‘zbek tilida stemmingni amalga oshirish uchun so‘z-shakldagi o‘zakkacha bo‘lgan barcha qo‘shimchalar kesib tashlanadi. ***Maktab****+dosh+lar+imiz* so‘zshaklida so‘z yasovchi va shakl yasovchi qo‘shimcha mavjud. O‘zbek tilida stemming jarayonida shu qo‘shimchalarning barchasi kesib tashlanadi:

So‘zshakl: ***maktab****+{dosh}+(lar)+(imiz)*

Lemma: ***maktabdosh***

Stem: ***maktab***

O‘zak (root): ***maktab***

Turk tilida stemlash jarayonida so‘zshakldagi faqat sintaktik va lug‘aviy shakl yasovchi qo‘shimchalar kesiladi, ammo so‘z yasovchilar qoldiriladi. Masalan:

So‘zshakl: ***seçim****+(ler)+(e)*

Stem: ***seçim***

Ko‘rinadiki, turk tilida stem tarkibida so‘z yasovchi qo‘shimcha qoladi, o‘zak bilan stemning farqi so‘z yasovchi qo‘shimchaning mavjudligidadir.

So‘zshakl: ***seçim+(****ler)+(e)*

Lemma: ***seçim***

Stem: ***seçim***

O‘zak (root): ***seç***

Uyg‘ur tilida stemlash jarayonida so‘zshakldagi sintaktik va lug‘aviy shakl yasovchi qo‘shimchalar kesiladi, ammo so‘z yasovchilar qoldiriladi.

oqutquchi

So‘zshakl: ***oqut*** *+ (qu) + (chi)*

Lemma: ***oqut***

Stem: ***oqut***

O‘zak (root): ***o***

Stemming jarayonidagi muammolar

Stemming jarayonida quyidagi muammolar yuzaga kelishi mumkin:

1) o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi;
2) so‘zning tovush o‘zgarishiga uchrashi;
3) neologizm va NERlarni stemmlash.

O‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi

Bugungi kunda tabiiy tildagi so‘zlar uchun turli stemming usullari ishlab chiqilgan. Zamonaviy stemming algoritmlari hech qanday sintaktik ma‘lumotlardan foydalanmagan holda yaratilmoqda [Polus, Abbas 2021; Memon, Mallah, Shaikh 2020; Khyani 2021].

Shuningdek, an’anaviy stemming usul (algoritm)lari – bu qo‘shimchalar va ba‘zi morfologik qoidalarga asoslangan bo‘lib, stemming jarayoni natijasida stemdagi noaniqlik yuzaga kelishi mumkin. Ko‘p ma‘noli o‘zakni aniqlash ancha murakkab jarayon bo‘lib, stemming jarayonida gap darajasidagi semantik ma‘lumotlar e’tiborga olinmaydi. Ba‘zida so‘zning POS tegi uning o‘zagining POS tegi bilan bir xil bo‘lmasligi mumkin.

O‘zbek tilida so‘z yasovchi va shakl hosil qiluvchi qo‘shimchalar o‘rtasida omonimiya hodisasi ham uchraydi. Bu esa stemming jarayonida muammoli vaziyatni yuzaga keltiradi. Quyidagi 3-jadvalda omonim qo‘shimchalar ro‘yxatini ko‘rish mumkin:

3-jadval. So‘z yasovchi va shakl hosil qiluvchi qo‘shimchalar o‘rtasida omonimiya.

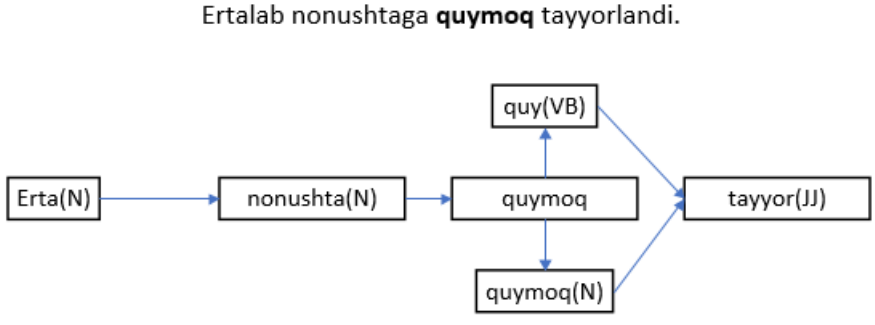
Shakl yasovchi qo‘shimcha	So‘z yasovchi qo‘shimcha
- ay (lug‘aviy shakl yas.) boray	kuchay (fe’l)
- gi (lug‘aviy shakl yas.) borgim	supurgi (ot), yozgi (sifat)
- da (sintaktik shakl yas.) uyda	undamoq (fe’l)
- i (sintaktik shakl yas.) do‘sti	jannati (sifat), boyi (fe’l)
- in (lug‘aviy shakl yas.) ko‘rin	ekin (ot), sog‘in (sifat)
- im (sintaktik shakl yas.) uyim	bilim (ot), ayrim (sifat)
- ir (lug‘aviy shakl yas.) o‘chir	gapir (fe’l)
- iq (lug‘aviy shakl yas.) siniqmoq	yo‘liq (fe’l), ochiq (sifat), chiziq (ot)
- y (lug‘aviy shakl yas.) o‘qiy	qoray (sifat)
- k (sintaktik shakl yas.) bordik	to‘shak (ot), chirik (sifat)
- ka (lug‘aviy shakl yas.) surka	iska (fe’l)
- kin ((lug‘aviy shakl yas.) to‘kkin	epkin (ot), keskin (sifat)
- la (lug‘aviy shakl yas.) quvla	so‘zla (fe’l)
- lab (lug‘aviy shakl yas.) yuzlab	haftalab (ravish)

- m (sintaktik shakl yas.) otam, ko‘rdim	to‘plam (ot)
- ma (lug‘aviy shakl yas.) gapirma	qatlama (ot), bo‘g‘ma (sifat)
- moq (lug‘aviy shakl yas.) ichmoq	quymoq (ot)
- sa (lug‘aviy shakl yas.) kelsa	suvsa (fe’l)
- siz (sintaktik shakl yas.) yozasiz	yuzsiz (sifat), to‘xtovsiz (ravish)
- xon (lug‘aviy shakl yas.) otaxon	kitobxon (ot)
- cha (lug‘aviy shakl yas.) qizcha	farg‘onacha (sifat), tushuncha (ot), erkakcha (ravish)
- chak (lug‘aviy shakl yas.) kelinchak	kuyunchak (sifat), belanchak (ot)
- chiq (lug‘aviy shakl yas.) qopchiq	sirpanchiq (sifat)
- choq (lug‘aviy shakl yas.) toychoq	o‘yinchoq (ot), maqtanchoq (sifat)
- qa (sintaktik shakl yas.)	
- qa (lug‘aviy shakl yas.) chayqa	qisqa (sifat)
- qin (lug‘aviy shakl yas.) boqqin	toshqin (ot), jo‘shqin (sifat)

4-jadval. Agglutinativ tillaridagi so‘zshaklning stemi va qo‘shimchalari.

Til	So‘zshakl	1-ma‘nosi	2-ma‘nosi
Turk	gelecek	keladi (will come)	kelajak (future)
Uyg‘ur	alma	ol+ma (don’t take)	olma (apple)
O‘zbek	quymoq	quy+moq (pour)	quymoq (panke)

O‘zbek tilidagi gaplarda stemdagi noaniqlikni quyidagi 3-rasmda ko‘rish mumkin:



3-rasm. O‘zbek tilidagi gaplarda stemdagi noaniqlik.

Turk tilida:
1-ma‘noda: Kış yine **gelecek**.
2-ma‘noda: **Gelecek** hakkında ne düşünüyorsunuz?
Uyg‘ur tilida:
1-ma‘noda: Qalamni qolunga **alma**.

2-ma‘noda: U bazardin **alma** setiwaldi.

O‘zbek tilida: (quymoq)

1-ma‘noda: Zarifa mehmonlarga choy **quymoqchi** bo‘ldi.

2-ma‘noda: Ertalab nonushtaga **quymoq** tayyorlandi.

Misol uchun, koyun (turkcha) so‘zini agar gapda fe’l sifatida kelsa **koy-(mak)** ko‘rinishida stemmlash mumkin. Agar gapda ot sifatida kelsa **koyun (qo‘y)** ko‘rinishida stemmlash lozim. POS teglash jarayonida turli muammolar yuzaga kelishi mumkin. Ulardan biri POS teglashdagi noaniqlikdir. So‘zlar gapdagi sintaktik roliga qarab turli so‘z turkumlarga mansub bo‘lishi mumkin. So‘zning aniq/to‘g‘ri POS tegi uning o‘zagini ham topishga yordam beradi.

Misol uchun,

1. Aydınlık gelecek günler bizi bekliyor. (Kelajakda bizni yorqin kunlar kutmoqda).

2. Ahmet birazdan gelecek. (Ahmad tez orada keladi);

Birinchi gapdagi *gelecek* – *sifatdosh*, o‘zak esa **gelecek (kelajak)** bo‘ladi. Ikkinchi gapda *gelecek* – *fe’l*, o‘zak esa **gel-(mek) (kelmoq)**. Yuqoridagi fikr mulohazalardan, POS teglash jarayoni stemmingda muhim rol ekanligini qayd etish mumkin.

Uyg‘ur tilida ham xuddi shunga o‘xshash holatni kuzatishimiz mumkin. Masalan *alma* so‘zi olma mevasi ma‘nosida olma shaklida stemmlash, *ol-ma* fe’l sifatida esa olmoq shaklida stemmlanadi. Stemmlashda so‘z shakllardagi POS teglashdagi farqni kelgüsi so‘zida ham kuzatish mumkin.

1. Kelgüsi ishimni planladim. (Kelajak ishlarimni reja qildim.)

2. Bala ete kelgüsi. (Bola ertaga keladi.)

Birinchi gapda *kelgüsi* – sifat, o‘zak esa kelgüsi (kelajak) bo‘ladi. Ikkinchi gapda *kelgüsi* – kelasi zamon shaklidagi fe’l, o‘zak esa kel-(mek) (kelmoq) shaklidadir.

O‘zbek tilida o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi va POS teglash, stemmini aniqlashdagi murakkabliklarni ko‘plab misollarda ko'rish mumkin. Misol uchun, *tortma, olma, yozma, o‘sma* va hokazo so‘zshakllarda. Bu so‘zlar tortma – tort-(moq), olma-ol-(moq), yozma-yoz-(moq), o‘sma-o‘s-(moq) stemmlari shaklida bo‘lib, POS tegi ot va fe’l deb belgilanadi. Masalan:

1. *Sen bozordan kitob olma.*

2. Akbar kecha olma yedi.

Bu yerda birinchi gapda *olma* – inkor ma‘nosidagi fe’l, o‘zak ol-(moq) shaklida bo‘lsa, ikkinchi gapda *olma* – ot, o‘zak ham olma

bo‘ladi.

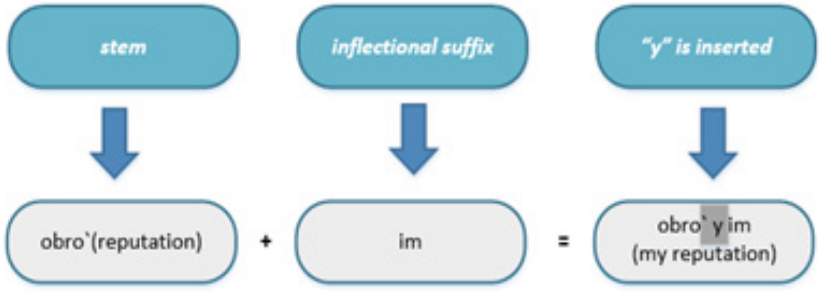
Yuqoridagi fikr-mulohazalardan bilish mumkinki, uchala turkiy tilda ham o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi holati uchraydi va bu vaziyatda POS teglash jarayoni stemmingda muhim rol ekanligini qayd etish mumkin.

So‘zning tovush o‘zgarishiga uchrashi

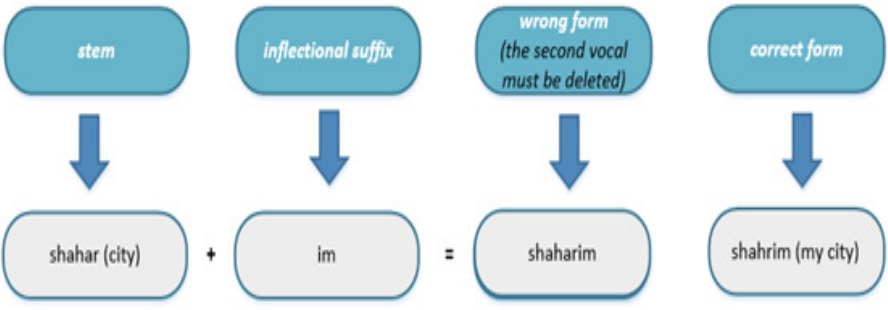
Shakl yasovchi qo‘shimchalarni o‘zakning oxirgi harflariga qo‘shish naijasida ba‘zi hollarda so‘zda fonetik o‘zgarishlar (insertion, deletion, phonetic harmony, and assimilation) yuzaga kelishi mumkin [Tsygankin, Ivanova 2019; Mirtojiyev 2013]. Agglutinativ tillarda so‘zda *tovush ortishi, tushishi* va *almashinishi* (weaking, assimilation) kabi uch xil fonetik o‘zgarishlar amalga oshirilishi mumkin *(5-jadval)*.

5-jadval. Agglyutinativ tillarida stemming jarayonidagi kamchiliklar.

O'zbek		Turk		Uyg'ur	
to‘g‘ri	**xato**	**to‘g‘ri**	**xato**	**to‘g‘ri**	**xato**
lavozim+ida	boshlig‘+i	yara+landi-g'ini	ögre-nlere	yoğ+an	binay+im
ish+lagan	san+aydi	belirt+ti	jandar+ma-lig'ini	eshek+-medeği	oghl+um
hafta+larida	tarog‘+ini	ara+sinda	rastla+diği	chaplish+ivalidu	yot+im
bo‘lim+i	me+ning	koşul+lar-da	iznin+e	bash+lapti	yurag+im
hokim+ining	obro'y+imiz	gösteri+ci-nin	geti+riliyor	dep+ti	shahr+im
ish+lagan	achch+iq	belir+li		ini+sinin	



4-rasm. So‘zda tovush ortishi.



5-rasm. So‘zda tovush tushishi.

6-rasm. So‘zda tovush almashinishi.

Stemni aniqlashdagi tovush o‘zgarishiga uchrashi muammosini hal qilish uchun, birinchi bosqichda o‘zak va qo‘shimchalarning chegaralari aniqlanadi, ikkinchi bosqichda esa lemmatizatsiya amalga oshiriladi. Lemmatizatsiya natijasida xato hosil qilingan stemlar lug‘atda mavjud root (o‘zak)ga o‘zgartiriladi.

Neologizm va NERlarni stemmlash

NERlarni stemlash muammolari

-lik qo‘shimchasi, asosan ot, sifat va ravish turkumiga oid so‘zlardan ot yasaydi. Ot turkumiga oid leksemalardan yasalgan so‘zlar yasovchi asos anglatgan narsa-predmetning xususiyati bilan bog‘liq holda turli ma’nolarni bildiradi:

1) shaxs (inson) bildiradigan so‘zlardan yasalganda: qarindoshlik, (*otalik, onalik, tog‘alik, o‘g‘illik, farzandlik, erlik, xotinlik*); umrning ma’lum davrini bildiradigan so‘zdan yasalgan otlar (*bolalik, yigitlik, qizlik, o‘smirlik, kelinlik, kuyovlik*); kasb, amal-unvon egasini anglatuvchi so‘zlardan yasalgan otlar (*mudirlik, o‘qituvchilik, qassoblik, chorvadorlik, tabiblik, suvchilik, sartaroshlik, savdogarlik, rassomlik, shofyorlik, aktyorlik*);

2) asosdan anglashilgan narsa ishg‘ol etgan obyektni bildiruvchi ot (*botqoqlik, qumlik, muzlik*);

3) yer sathining yasovchi asosdan anglashilgan qismini bildiruvchi ot (*jarlik, do‘nglik, qiyalik, pastlik, ichkarilik, yalanglik*).

Sifat va ravishlarga qo‘shilib, belgi oti yasaydi: *qizillik, semizlik, xursandlik, aniqlik.* Bu kabi holatlarda ularning tarkibidagi so‘z yasovchi qo‘shimchalar kesiladi, qolgan qism stem sanaladi.

Ammo joy nomini bildiruvchi atoqli otlarga –***lik*** qo‘shimchasi qo‘shilganda, ular turdosh otga aylanadi va kichik harf bilan yoziladi: *samarqandlik, buxorolik, amerikalik, o‘zbekistonlik, turkiyalik, arabistonlik.* Bunday holatda –***lik*** qo‘shimchasi kesiladi, qolgan qism stem deb tushuniladi, bosh harfga aylantirilib, NER sifatida tan olinadi.

samarqandlik = Samarqand~~lik~~

amerikalik = Amerika~~lik~~

kanadalik = Kanada~~lik~~

NERlarning stemini topish muammosi yuzaga kelganda shakl yasovchilar kesiladi, so‘z yasovchi shaklidagi qo‘shimcha yoki so‘zning qismi qoldiriladi, shu qism NER sanaladi: O‘*zbekistondan* so‘zshaklining stemi O‘*zbekiston*.

Shunday qo‘shimchalar borki, ular so‘z yasovchi va shakl yasovchi vazifasida keladi (6-jadval).

6-jadval. So‘z yasovchi va shakl yasovchi omonim qo‘shimchalar.

Shakl yasovchi va so‘z yasovchi qo‘shimchalar		
-ay	-k	-chak
-gi	-ka	-chiq
-da	-kin	-choq
-i	-la	-qa
-in	-lab	-qin
-im	-m	-sa
-ir	-ma	-siz
-iq	-moq	-xon
-y	-cha	

Bunday qo‘shimchalar bosh harf bilan yozilgan so‘zlarning tarkibida kelganda, shakl yasovchilar va so‘z yasovchilar tarkibida bo‘lsa, so‘z shakl tarkibida qoldiriladi va shu shaklida stem deb olinadi. Masalan, *Jon Kennedi* so‘zning tarkibida ***–i*** harfi bor. Dastur o‘zakni bilmaganligi sababli, ya’ni mazkur so‘z o‘zbek tili lug‘atida mavjud emasligi sababli o‘zakni ajratolmay qoladi, natijada **–i** qo‘shimchasini kesib, ***Kenned*** so‘zini o‘zak deb olishi

mumkin. Bunday holatdan qochish maqsadida shakl yasovchi va so‘z yasovchilar orasida omonimiya hosil qiluvchi qo‘shimcha bilan shakldosh bo‘lgan har qanday birlik so‘zshakl tarkibida qoldiriladi.

Neologizmlarni stemlash muammolari

Neologizm yun. “neos” — yangi, “logos” — so‘z — jamiyat taraqqiyoti, hayotning talab-ehtiyoji bilan paydo bo‘lgan yangi narsa va tushunchalarni ifodalovchi so‘zlar. Neologizmlarning yangiligi dastlab paydo bo‘lgan vaqtlardagina sezilib turadi: vaqt o‘tgach, ular “yangilik” xususiyatini yo‘qotib, odatda, faol so‘zlar qatoriga o‘tadi.

Neologizmning shakliy neologizm, semantik neologizm, funksional neologizm, ijtimoy neologizm, texnologik neologizm, stilistik neologizm kabi turlari mavjud. Ularning turi haqida quyidagi 7-jadvalda batafsil ma’lumot beriladi [Qo‘ziboyeva 2022].

7-jadval. Neologizm turlari.

Misol	Ta’rif	Turi
Popemobil	Rim papasini tashib yurish uchun maxsus ishlab chiqilgan avtomobilning norasmiy nomi	Shakliy
Kopirayting	Reklama yoki marketing maqsadida matnni yozish jarayoni.	Semantik
Qidiruv tizimi	Kompyuterda, kompyuter tarmog‘ida yoki butunjahon web tarmog‘ida World Wide Web saqlanayotgan ma’lumotlarni qidirishga mo‘ljallangan dastur	Semantik
Tırtıl	Kunduzi va tungi kapalaklarning lichinkasi	Funksional
Shizofrenik	Shizofreniya kasalligi bilan og‘rigan bemor	Stilistik
Kibermakon	Kompyuter tarmoqlari orqali amalga oshiriladigan muloqot maydoni	Texnologik

Neologizmlarning paydo bo‘lish yo‘llari xilma-xil bo‘lib, ular tilning mavjud lug‘aviy tarkibi va grammatik qonun-qoidalari asosida yangi so‘z yasash yo‘li, shuningdek, mavjud so‘zning lug‘aviy ma’nolaridan birini yangi ma’noda qo‘llash yo‘li bilan va boshqa tildan so‘z qabul qilish orqali hosil qilinadi.

Neologizmlar tarkibida –izm (neologism**),** -ik (dalton**ik),** -la (gugl**la**sh) kabi qo‘shimchalar uchraydi.

Neologizmlar lug‘atda mavjud bo‘lmaganligi sababli ularni stemlashda muammolar yuzaga chiqadi. Ularning tarkibidagi qo‘shimchalar, so‘zning bir qismining qo‘shimchaga o‘xshab qolishi muammolari shular shumlasidan. Bunday holatda shakl yasovchi qo‘shimchalar bazasida mavjud qo‘shimchalar kesiladi. Qolgan qism stemga teng keladi. O‘zbek tilidagi neologizmlar va NERlarga mos

stem yuqoridagi 2-rasmda keltirilgan turk va uyg‘ur tillaridagi stem ta’rifiga mos keladi.

Neologizmlar keng jamoatchilik tomonidan ma’lum vaqt oralig‘ida faol qo‘llanib og‘zaki va yozma nutqqa ko‘chganda tilning leksik boyligi sifatida rasman e’tirof etilishi, ya’ni lug‘atga kiritilishi mumkin. Bu jarayondan so‘ng ularni stemlash lug‘atdagi so‘zlarni stemlash qoidasi asosida amalga oshiriladi.

Muammoning o‘rganilishi

Avvalo POS teglash va stemming vazifasini ikki alohida vazifa sifatida mustaqil ravishda ko‘rib chiqamiz. So‘ngra, POS teglash vazifasini morfologik segmentatsiya kabi boshqa vazifalar bilan birlashtirgan POS tegining qo‘shma modellarini ko‘rib chiqamiz.

POS teglashga oid tadqiqotlar

Korpus matnlarini POS teglash NLPdagi klasterlash muammosi sifatida keng tarqalgan. Brown so‘zlarning sintaktik sinflarini o‘rganish uchun *murakkab iyerarxik klasterlash* algoritmi asosida klasslarga asoslangan **n-gram modelini** taqdim etgan [Brown, Della Pietre, de Souza, 1992]. Tadqiqotda kontekstli ma‘lumotlar n-gramm shaklida kiritilgan bo‘lib, boshlang‘ich holatda har bir so‘z bitta sinfga mansub bo‘ladi. So‘ngra, o‘rtacha minimal yo‘qotishni beradigan har bir klaster juftligi barcha klasterlar bitta klaster ostida birlashtirilgunga qadar umumlashtiriladi. Keyingi qadamda, sintaktik kategoriyalar orasidagi iyerarxiyani ifodalovchi binar daraxt shakllantiriladi.

Schutze har bir so‘zning ikkita chap va ikkita o‘ng qo‘shni so‘zidan olingan so‘z vektorlari tomonidan tuzilgan kontekst matritsasining o‘lchamini kamaytirish uchun **Singular Value Decomposition (SVD)**dan foydalangan [Schütze, 1993]. Keyingi qadamda kontekstli ma‘lumotlardan foydalangan holda so‘zlarni klasterlash uchun Buckshot klasteri qo‘llanilgan [Cutting, Kupiec, Pedersen, Sibun 1992].

Biemann to‘rtta so‘zli kontekst oynalari va eng ko‘p uchraydigan so‘zlarni ishlatib, “Chinese Whispers” grafli klasterlash algoritmini qo‘llagan [Biemann 2006].

Ba‘zi boshqa yondashuvlarda POS teglash amali gapdagi so‘zlarni ketma-ketlikdagi belgilash/teglash muammosi sifatida qaraladi. Bunday turdagi yondashuvlar asosidagi algoritmlarda ko‘p hollarda Yashirin Markov Modeli (Hidden Markov Models, HMMs)dan foydalaniladi.

Merialdo uch sinfdan iborat Markov modelini taqdim et-

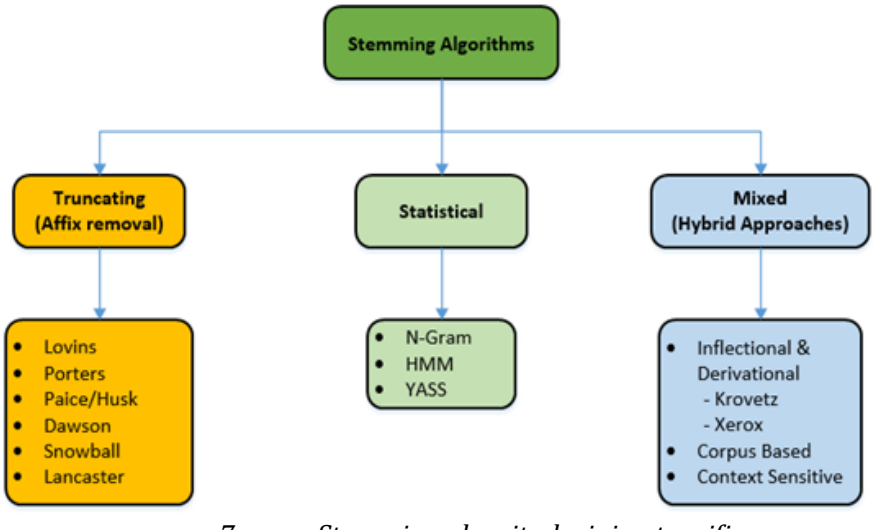
gan bo'lib, tadqiqotda korpusdagi o'quv ma'lumotlarining turli o'lchamlari uchun turli parametrlarni baholash usullari taqqoslangan [Merialdo 1994]. Korpus ma'lumotlari uchun nisbiy chastotali trening ishlatilgan va tegsiz ma'lumotlar uchun ***Maksimal ehtimollik (Maximum Likelihood)*** usuli qo'llanilgan.

Banko va Moore har bir so'zni faqat joriy so'z "teg"idan emas, balki oldingi va keyingi so'z "teg"lari (valentlik)ni o'z ichiga olgan uchta qo'shni teg asosida kontekstli HMM teggerini taqdim etadilar [Banko, Moore, 2004]. Ushbu model asosiy HMM bilan solishtirganda ko'proq kontekstli ma'lumotlarni o'z ichiga olgan va samarali natija qaytargan.

Jonson **HMMga asoslangan POS teglash**da ishlatiladigan turli parametrlarni solishtirgan. Shu maqsadda **Expectation Maximization (EM), Variatsion Bayes** va **Gibbs** namunalaridan foydalangan [Haghighi, Klein 2006]. Tadqiqot Gibbs namunasi va variatsion Bayes baholovchisi bilan solishtirganda **EM** algoritmining past samaradorligini ko'rsatib bergan.

Stemmingga oid tadqiqotlar

Zamonaviy stemming algoritmlari odatda uchta sinfga bo'linadi: *qoidaga asoslangan*, *statistik* va *gibrid* algoritmlar (7-rasm). Qoidalarga asoslangan stemmerlar avtomatik bo'lmagan qoidalardan foydalangan holda stemmlarni aniqlashga qaratilgan. *Qoidalarga asoslangan* ommabop stemmerlar sifatida Lovins [Lovins, 1968], Porter [Porter 2006; 2001] va Krovets [Krovetz 2000]larni keltirish mumkin. Qoidalarga asoslangan stemmlash algoritmlari odatda nazorat qilinadi.



7-rasm. Stemming algoritmlarining tasnifi.

Statistik stemmlash algoritmlari stemmlarni o‘rganish uchun statistik usullardan foydalanadi. Xu va Croft [Xu, Croft, 1998], Porter stemmerining [Porter, 2006] kamchiliklarini bartaraf etish uchun tasodifiy yuzaga keladigan statistik so‘zdan foydalanadigan usulni taqdim etishgan. Tasodifiy statistik ma‘lumotlariga asoslanib, ular Porter stemmeri tomonidan yaratilgan sinflar sonini kamaytirish uchun *grafni qismlarga ajratish algoritmini* amalda qo‘llashgan [Porter 2006].

Gibrid stemmlash algoritmlari qoidalarga asoslangan va statistik usullarni yagona tizimga birlashtiradi. Ba‘zi gibrid stemmlash algoritmlari Shrivastava [Manish, Nitin, Bibhuti], Goweder [Goweder Alhami] va Adam [Adam 2010] tomonlaridan ishlab chiqilgan.

Goldsmith [Goldsmith 2001; 2006], minimal tavsif uzunligi (Minimum Description Length, MDL) tamoyiliga asoslangan, nazoratsiz stemming modelini taklif qilgan. Model, asosan morfologik segmentatsiya uchun mo‘ljallangan bo‘lib, undan stemmer sifatida ham foydalanish mumkin. Har bir so‘zdagi segmentatsiya nuqtalari, korpusning umumiy hajmini qisqartirish uchun mo‘ljallangan.

Graflarga asoslangan stemming algoritmi Bacchin tomonidan taklif qilingan [Bacchin, Ferro, Melucci 2002]. Algoritm satr ostilar to‘plamini aniqlash uchun birinchi bosqichda har bir so‘zni barcha mumkin bo‘lgan bo‘linish nuqtalariga ajratadi. Ikkinchi bosqichda satr ostilar to‘plamidan foydalangan holda yo‘naltirilgan graf hosil qilinadi. Nihoyat grafdagi satr ostidagilar chastotasiga qarab prefiks va suffiks ballari hisoblanib, stem aniqlanadi.

Melucci va Orio **HMM asosidagi stemmer**ni taqdim etadilar [Melucci, Orio 2003]. HMM holatlari prefiks va suffikslarga muvofiq kelib, holatlar orasidagi o‘tishlar grammatik qoidalarga mos keladi. Parametrlarni baholash uchun **Expectation Maximization (EM)** algoritmidan foydalanilgan. Parametrlar baholangandan so‘ng, maksimal ehtimollikka ega yo‘nalish bo‘yicha segmentatsiya amalga oshirilgan va stem aniqlangan.

McNamee va Mayfield **n-grammlarga asoslangan muqobil stemmlash algoritmi**ni taqdim etishgan [McNamee, Mayfield 2004]. Har bir so‘z uchun korpus asosida barcha bigramma va trigrammalar generatsiya qilingan bo‘lib, o‘xshash so‘zlar n-grammning asosiy qismini tashkil etgan.

Bacchin graflar asosidagi stemmer modelini kengaytirib

[Baschin, Ferro, Melucci 2005], dastlabki qadamda har bir so‘zni barcha pozitsiyalarga bo‘lish orqali mumkin bo‘lgan satr ostilar to‘plami aniqlagan. So‘ngra satr ostilarni ifodalaydigan yo‘nalishli graf shakllantirilgan. Agar *z* so‘zi, *z = xy* ni qanoatlantirsa, *x* va *y* tugunlar orasiga yo‘nalishli qirra hosil qilingan. Affiks ballarini baholash **HITS** algoritmi [Kleinberg, Kumar, Raghavan 1999] asosida hisoblangan. Prefiks va sufiks ballari asosida so‘zlarga tegishli bo‘lgan prefiks va qo‘shimchalar juftlarining ehtimolini maksimal darajaga oshirish orqali eng katta ehtimoliy bo‘linish nuqtasi aniqlangan.

Majumder tomonidan **YASS** (Yet Another Suffix Striper) [Majumder, Mitra, Parui 2007] deb nomlangan stemmlash algoritmi taqdim etilgan bo‘lib, u satrlar orasidagi masofa o‘lchovidan foydalanadigan klasterlash algoritmiga asoslangan. Satrlar orasidagi masofa o‘lchovi so‘zlar orasidagi morfologik o‘xshashlikni baholash uchun ishlatilgan.

So‘zlar orasidagi o‘xshashlikni aniqlashga asoslangan stemmer Peng tomonidan ishlab chiqilgan [Peng, Ahmed, Li, Lu 2007]. Ushbu stemmer qidiruv tizimi natijalarini yaxshilash uchun IR vazifalari uchun qo‘llanilgan va yuqori samaradorlikni bergan.

Paik tomonidan graflarga asoslangan **GRAS** (GRAph-based Stemmer) stemmeri ishlab chiqilgan bo‘lib, u so‘zlarni guruhlash uchun leksik ma‘lumotlardan foydalanadigan statistik stemmer hisoblanadi. Ushbu algoritmda so‘zlar grafning tugunlari sifatida ifodalangan. Algoritm grafni parchalash orqali so‘zlar o‘rtasidagi bog‘lanishni aniqlaydi.

Brychcin va Konopik tomonidan taklif etilgan yuqori aniqlikdagi stemmer (High Precision Stemmer, HPS)da imlo va semantik ma‘lumotlardan so‘zlarni o‘zak va qo‘shimchalarga bo‘lish xususiyati sifatida foydalanilgan. Usul ikki bosqichdan iborat:

— *orfografik va semantik jihatdan o‘xshash so‘zlar maksimal o‘zaro ma‘lumot (Maximum Mutual Information, MMI) yordamida klasterlash*;

— *birinchi bosqichdan olingan klasterlar yordamida maksimal entropiya tasniflagichini amalga oshirish.*

Turk tilidagi stemmingni amalga oshirish usuli Köksal tomonidan kiritilgan [Brown, Della Pietra, 1992]. Ushbu usul dastlabki 5-6 harfni o‘zak deb hisoblashga asoslangan. Kut va boshqalar o‘z tadqiqotlarida L-M (Longest Match) nomli usulni ishlab chiqqanlar [Schütze 1993]. So‘z o‘zaklari va ularning mumkin bo‘lgan shakllarini o‘z ichiga olgan lug‘atdan foydalangan holda, usul chap-

dan o‘ngga harflarni lug‘atda joylashgan so‘zlar bilan o‘zak so‘zni taqqoslaydi. Eng uzun mos kelgan so‘z o‘zak hisoblanadi.

Solak va Can [Paice 1994] stemlarni aniqlashda ildizlar lug‘atidan foydalanganlar. Har bir ildiz chapdan o‘ngga stem hosil qilish usullariga mos keladigan 64 xususiyatga ega deb qayd etilgan. Harf birliklari ildiz leksikasiga chapdan o‘ngga tartibda moslashtirilgan va agar mos stem aniqlansa, tizim qo‘shimcha qoidalar asosida mumkin bo‘lgan stemlarni aniqlaydi. **AF algoritmi** deb ataladigan ushbu tadqiqot, asosan, Oflazer [Banko, Moore 2004] tomonidan ishlab chiqilgan morfologik tahlil usulining varianti hisoblanadi.

FindStem – bu Sever va Bitirim [Goldwater, Griffiths 2007] tomonidan ishlab chiqilgan stemming usuli bo‘lib, asosan uchta amalni o‘z ichiga oladi: *ildizni aniqlash, o‘zakni morfologik tahlil qilish* va *aniqlash*. Usul so‘zlarning morfologik va POS xususiyatlarini, sintaktik qoidalarini o‘z ichiga olgan lug‘atdan foydalanadi. Sever va Bitirimning ta‘kidlashicha, FindStem algoritmi AF va L-M algoritmlariga qaraganda yaxshiroq va samaraliroq ishlaydi.

Turk tilidagi so‘zlarning o‘zagini aniqlashga oid boshqa tahliliy usullarga Akin [Cutting, Kupiec, Pedersen, Sibun 1992] tomonidan ishlab chiqilgan **“zemberek”** va Childen [Haghighe, Klein 2006] tomonidan ishlab chiqilgan **“snowball”** algoritmlarini keltirish mumkin. Shuningdek, Dincher [Merialdo, 1994] o‘zak va qo‘shimchalar orasidagi chegarani n-gramm statistikasidan foydalangan holda hal qilish usulini taklif qilgan. Ushbu tadqiqot amaliyotga qo‘llanilishi natijasida samaradorlik 95,8% ni tashkil qilgan.

Batuer Aisha va Maoshistal Sun statistik yondashuvga asoslangan tokenizatsiya ishlab chiqdilar [Aisha, Sun 2009]. Ushbu tokenizatsiya **UC uyg‘ur tili denoted korpusi** ustiga qurilgan. UC korpusi **594172 ta** uyg‘ur tili so‘zlari va UC ikkita korpusdan tashkil topgan, ya‘ni UCS denoted qo‘lda **stemlangan** va **lemmalangan** korpusi. Keyinchalik uyg‘ur tili tokenizatsiyasi uchun ikki bosqichli jarayondan foydalandilar.

Aishan Vumaier va boshqa tadqiqotchilar 2009-yilda yangi uyg‘ur tili ot so‘z turkumi stemming usulni ishlab chiqdilar [Maimaiti, Wumair 2017]. Uyg‘ur ot so‘z turkumi stemming usulini 2 bosqichda amalga oshirdilar:

— Uyg‘ur tili **FSM** qo‘shimchalari yordamida;

— Uyg‘ur tili **FSM** qo‘shimchalari bilan yuzaga kelgan noaniqliklarni bartaraf qilish uchun **CRF** metodi yordamida.

Birinchi bosqichda, uyg‘ur tili ot so‘z turkumi stemming jarayonini FSM ot qo‘shimchalari yordamida ishlab chiqdilar. Stemming jarayoni 55625 kiritilgan so‘z ustida amalga oshirilib, natijada 6239 noto‘g‘ri over stemming keltiruvchi so‘zlar aniqlandi. Ikkinchi bosqichda, stemming jarayonida yuzaga kelgan noaniqliklarni CRF metodi yordamida aniqlab, 55125 ta so‘zli korpus qurildi. Korpus 17317 ta noaniq qo‘shimchali so‘zlar, 6239 ta correctsiz qo‘shimchalar va 11078 ta correct qo‘shim- chali so‘zlardan tashkil topgan. Algoritm natijasi shuni ko‘rsatadiku, jarayonda FSM qo‘shimchalaridan foydalanilsa eslatma (recall rate) 88.78%ni, FSM va CRFdan foydalanilsa eslatma (recall rate) 94,04% aniqlikka erishildi. Xulosa shuki, CRF usulidan foydalanish qayta tiklash stavkasini 5,26% yaxshilaydi.

2012-yilda Azragul, Qixiangjwei va Yusupulla Uyg‘ur tili stemmerini ishlab chiqdilar [Azragul, Qi, Yusup 2012]. Ular lug‘atga asoslangan usuldan foydalanganlar. Algoritm ishlash jarayonida kiritilgan so‘z stem lug‘atidan qidiriladi. Bunda, qo‘shimchalar lug‘ati yordamida so‘z ajratiladi va qo‘shimchalarni o‘chirish bilan ajratilgan so‘z nomzod so‘zi lug‘atdan izlanadi.

Amalga oshirilgan tadqiqotlar shuni ko‘rsatadiki, oldingi tadqiqotlarda to‘liq bo‘lmagan lug‘atdan (ochiq lug‘at) foydalanilgan va stemming natijasida hosil bo‘lgan noaniqliklar keyinchalik boshqa usullar orqali hal qilingan.

Pos teglash va stemmlarni aniqlash modellari

NLP sohasi olimlarini agglyutinativ tillarda lug‘at cheklanmagan degan fikrga kelishgan. Ya‘ni, *yangi so‘zlar*, *yangi atamalar* yoki *neologizmlar* madaniy almashinuv orqali tilga doimiy kirib keladi. Yangi so‘zlar mavjud ildizga tegishli qo‘shimchalarni birlashtirish orqali hosil bo‘ladi. Tildagi ildiz va qo‘shimchalar vaqt o‘tishi bilan juda sekin o‘zgaradi va ularni turg‘un deb hisoblash mumkin. Ildiz va qo‘shimchalarning cheksiz ko‘p mumkin bo‘lgan birikmalari lug‘atning doimiy kengayishiga olib keladi. Xulosa sifatida lug‘atdan foydalanadigan usullardan foydalanish bir qator murakkabliklarni yuzaga keltirishini qayd etish lozim.

O‘zbek tilidagi so‘zlarni stemmlash va POS teglash uchun N.Bölücü va B.Can modelidan foydalanamiz [Bölücü, Can 2019]. Tadqiqotda ushbu modelni asosiy model sifatida qabul qilamiz va ba‘zi o‘zgartirishlarni amalga oshiramiz. Maqolada ishlatiladigan atamalarning qisqacha tavsifi:

– **w = s + m** – *so‘zning segmentatsiyasi;*

bu yerda ***s*** *– stem va* ***m*** *–* ***w*** *so‘zining qo‘shimcha(lar)si;*

– $\mathbf{t_i}$, $\mathbf{w_i}$, $\mathbf{s_i}$, $\mathbf{m_i}$ – *korpusdagi* ***i*** *ning mos ravishda teg‘i, so‘zi, stemi va qo‘shimchasi;*

– **I(.)** – *identifikator funksiyasi bo‘lib, agar argumenti* ***true*** *bo‘lsa* ***1*** *ni,* ***false*** *bo‘lsa* ***0*** *ni qaytaradi;*

– $\boldsymbol{n(t_i, w_i)}$ – (t_i, w_i) *so‘zlar juftligining chastotasi;*

– $\boldsymbol{n(t_i, s_i)}$– (t_i, s_i) *stemlar juftligining chastotasi;*

– $\boldsymbol{n(t_i, m_i)}$–(t_i, m_i) *qo‘shimchalar juftligining chastotasi;*

– $\boldsymbol{cos(s_i, w_i)}$ –(s_i, w_i)*so‘zlar o‘rtasidagi kosinus o‘xshashlik;*

– $\boldsymbol{n(t_{i-2}, t_{i-1})}$ – $<t_{i-2}, t_{i-1}>$ *teglar bigrammasi chastotasi;*

– $\boldsymbol{n(m_{i-2}, m_{i-1})}$ –$<m_{i-2}, m_{i-1}>$*qo‘shimchalar bigrammasi chastotasi;*

– $\boldsymbol{n(t_{i-2}, t_{i-1}, t_i)}$ –$<t_{i-2}, t_{i-1}, t_i>$ *teglar trigrammasi chastotasi;*

– $\boldsymbol{n(m_{i-2}, m_{i-1}, m_i)}$ – $<m_{i-2}, m_{i-1}, m_i>$ *qo‘shimchalar trigrammasi chastotasi;*

– $\boldsymbol{t_{-i}}$ – t_{-i} *dan tashqari barcha teglarning joriy qiymatlari;*

– $\boldsymbol{w_{-i}}$) – w_{-i} *dan tashqari barcha so‘zlarning joriy qiymatlari;*

– $\boldsymbol{s_{-i}}$– s_{-i} *dan tashqari barcha stemlarning joriy qiymatlari;*

– $\boldsymbol{m_{-i}}$– m_{-i} *dan tashqari barcha qo‘shimchalarning joriy qiymatlari,*

– $\mathbf{W_{ti}}$, $\boldsymbol{S_{ti}}$ va $\mathbf{M_{ti}}$ – *mos ravishda* t_i *dan hosil bo‘lgan so‘z, stem va qo‘shimcha turlarining umumiy soni;*

– ***T*** *– teg to‘plamining hajmi;*

– $\boldsymbol{M(\varpi^t)}$ – $\boldsymbol{\varpi^t}$ *parametrlarni o‘z ichiga olgan Multinomial taqsimoti shaklidagi emissiya taqsimoti;*

– $\mathbf{Mult(\tau^{(t,t')})}$ – $\boldsymbol{\tau^{(t,t')}}$ *parametrli tranzit taqsimot;*

– $\mathbf{Mult(\psi^{(t)})}$– $\boldsymbol{\psi^{(t)}}$ *parametrlarni o‘z ichiga olgan Multinomial taqsimoti shaklidagi qo‘shimchalar taqsimoti;*

– $\boldsymbol{\varpi^t}$ – $\boldsymbol{\beta}$ giperparametrdan iborat *;* ***Dirichlet(β);***

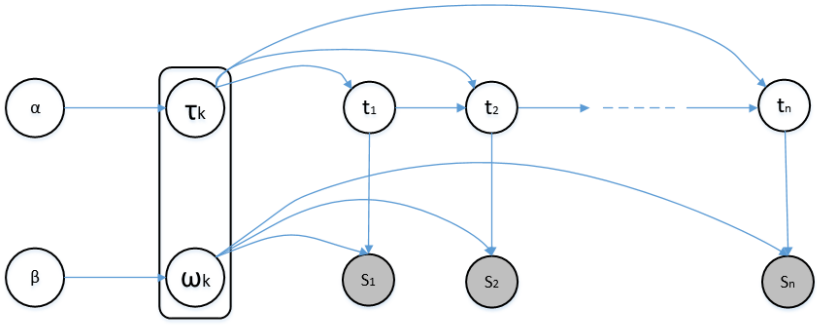
– $\boldsymbol{\varpi^t}$ –$\boldsymbol{\beta}$ giperparametrdan iborat *;* ***Dirichlet(β);***

– $\boldsymbol{\tau^{(t,t')}}$ –$\boldsymbol{\alpha}$ giperparametrdan iborat *;* ***Dirichlet(α);***

– $\boldsymbol{\psi^{(t)}}$– $\boldsymbol{\psi}$ giperparametrdan iborat *.* ***Dirichlet(ψ).***

Matematik model tavsifi

Goldwater va Griffiths so‘zga asoslangan Bayes HMM modelini taklif qiladilar. Ular standart HMM modeli parametrlariga avvalgi taqsimotlarni qo‘shib, parametrlarning nuqtaviy baholarini emas, balki Bayes sozlamalarida parametrlarning taqsimlanishini o‘rganish orqali kengaytirishgan. POS teglar va stemmlarni birgalikda aniqlash uchun Bayes HMM bazaviy modelini kengaytirish lozim. Shu maqsadda biz ma‘lumotlardagi turli bog‘liqliklarni qabul qilib, turli versiyalarni taklif qilamiz.



8-rasm. Stemga asoslangan Bayes HMM diagrammasi

Stemga asoslangan Bayes HMM (Bayesian S-HMM)

Agglyutinativ (turk) tildagi so‘z tarkibi 2-rasmda ko‘rsatilgan bo‘lib, so‘z stemi so‘zdan flektiv qo‘shimchalarni olib tashlash orqali hosil qilinadi. Bu esa so‘z va uning stemi bir xil POS tegga ega bo‘lishi kerak degan qoidadir. So‘zning POS tegi uning stemi uchun ahamiyatli ko‘rsatkichdir. Misol uchun, *gelecek* ot so‘z turkumiga mansub bo‘lsa, *-ecek* qo‘shimchasini olib tashlash xato hisoblanadi. Agar *gelecek* fe’l so‘z turkumiga mansub bo‘lsa, *-ecek* qo‘shimchasini olib tashlash to‘g‘ri bo‘ladi, chunki bu holda *-ecek* fleksiyon qo‘shimcha hisoblanadi. Shuningdek, agglyutinativ tillarda stemlardan foydalanish emissiya siyrakligini kamaytiradi. Shunday qilib, stem emissiyalari bilan so‘zga asoslangan HMM modelini kengaytiramiz. Buning matematik modeli quyidagicha shakllantiriladi:

$$t_i|t_{(i-1)},t_{(i-2)} = t', \tau^{(t,t')} \propto Mult(\tau^{(t,t')})$$

$$s_i|t_i = t, \varpi^t \propto Mult(\varpi^{(t)})$$

$$\tau^{(t,t')}|\alpha \propto Dirichlet(\alpha)$$

$$\varpi^{(t)}|\beta \propto Dirichlet(\beta)$$

Bizning modeldagi $Mult(\varpi^{(t)})$ asosiy modeldan biroz farq qiladi. Bunda parametrga ega Multinomial taqsimot ko‘rinishidagi stem emissiyasi taqsimoti hisoblanadi. Ya‘ni, har bir HMM holatiga mos stem aniqlanadi. Modelning diagrammasi yuqoridagi 8-rasmda keltirilgan. Matematik modelga asoslanib, teg va stemning shartli ehtimoli quyidagicha aniqlanadi:

$$P(t_i|t_{i-1},\alpha) = \frac{n(t_{i-2},t_{i-1},t_i)+\alpha}{n(t_{i-2},t_{i-1})+T\alpha}$$

$$P(s_i|t_{-i},s_{-i},\beta) = \frac{n(t_i,s_i)+\beta}{n(t_i)+S_{t_i}\beta}$$

Quyidagi taqsimot orqali aprior baho hosil qilinadi:

$$P(t,s|\alpha,\beta) \propto P(s|t,\beta)\ P(t,\alpha)$$

Mos stemni aniqlash uchun Gibbs namunalari (Gibbs sampling)dan foydalanamiz. Bunda barcha teglar tasodifiy inisializatsiya qilinadi va barcha so‘zlar stem va qo‘shimchalar sifatida tasodifiy ikkita segmentga ajratiladi. Algoritmning har bir iteratsiyasida quyidagi namunaviy taqsimotdan har bir so‘z uchun teg va stem namunasi olinadi:

$$P(t_i,s_i|t_{-i},s_{-i},\alpha,\beta) = \frac{n(t_i,s_i)\beta}{n(t_i)+S_{t_i}\beta}\cdot\frac{n(t_{i-2},t_{i-1},t_i)+\alpha}{n(t_{i-2},t_{i-1})+T\alpha}$$

$$\cdot\frac{n(t_{i-1},t_i,t_{i+1})+I(t_{i-2}=t_{i-1}=t_i=t_{i+1})+\alpha}{n(t_{i-1},t_i)++I(t_{i-2}=t_{i-1}=t_i)+T\alpha}$$

$$\cdot\frac{n(t_i,t_{i+1},t_{i+2})+I(t_{i-2}=t_i=t_{i+2},t_{i-1}=t_{i+1})+I(t_{i-1}=t_i=t_{i+1}=t_{i+2})+\alpha}{n(t_{i1},t_{i+1})++I(t_{i-2}=t_i,t_{i-1}=t_{i+1})+I(t_{i-1}=t_i=t_{i+1})+T\alpha}$$

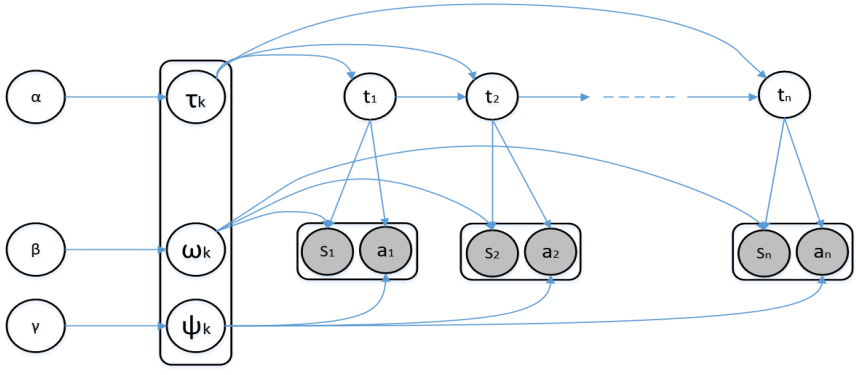
Tegni tanlash bir vaqtning o‘zida modeldagi uchta teg trigrammasiga ta’sir qiladi. Chunki har bir teg ketma-ket uchta teg trigrammalarida uchraydi. Shu sababli, o‘zgarishlar identifikatsiya funksiyasi bilan hisobga olinadi. Ushbu jarayon tizim birlashmaguncha takrorlanadi.

Stem va qo‘shimchaga asoslangan Bayes HMM (Bayesian SM-HMM)

So‘zning qo‘shimchasi odatda uning sintaktik kategoriyasi haqida ma’lumot beradi. Misol uchun, ingliz tillarida **-ly** qo‘shimchasi bilan tugagan so‘zlar ravishlar (adverbs) hisoblanadi. Stemga asoslangan (stem-based) Bayes HMM modeliga stem emissiyalariga qo‘shimcha ravishda suffiks emissiyalarini qo‘shamiz:

$$m_i|t_i = t,\psi^{(t)} \propto Mult(\psi^{(t)})$$

$$\psi^{(t)}|\gamma \propto Dirichlet(\gamma)$$



9-rasm. Stemm va qo‘shimchaga asoslangan Bayes HMMning diagrammasi.

Matematik modelga asoslangan holda, suffiksning shartli ehtimolligi quyidagicha aniqlanadi:

$$P(m_i|t_{-i}, m_{-i}, \gamma) = \frac{n(t_i, m_i) + \gamma}{n(t_i) + M_{t_i}\gamma}$$

Stem va teg uchun shartli ehtimollik stemga asoslangan modelda berilgani bilan bir xil hisoblanadi. Quyidagi taqsimotni orqali aprior baho hosil qilinadi:

$$P(t,s,m|\alpha,\beta,\gamma) \propto P(s|t,\beta)\ P(m|t,\gamma)P(t,\alpha)$$

Bu yerda stem va qo‘shimchalar bir-biridan mustaqil deb faraz qilinadi. Mos stemni aniqlash uchun bu holda ham Gibbs namunalari (Gibbs sampling)dan foydalanamiz. So‘z teglari tasodifiy inisializatsiya qilinadi va barcha so‘zlar amal boshida ikkita segmentga tekis ajratiladi. Algoritmning har bir iteratsiyasida quyidagi namunaviy taqsimotdan har bir so‘z uchun *teg, stem* va *qo‘shimchalar* tanlanadi:

$$P(t_i, s_i|t_{-i}, s_{-i}, m_{-i}\alpha, \beta, \gamma) = \frac{n(t_i, s_i)\beta}{n(t_i) + S_{t_i}\beta} \cdot \frac{n(t_{i-2}, t_{i-1}, t_i) + \alpha}{n(t_{i-2}, t_{i-1}) + T\alpha}$$

$$\cdot \frac{n(t_{i-1}, t_i, t_{i+1}) + I(t_{i-2} = t_{i-1} = t_i = t_{i+1}) + \alpha}{n(t_{i-1}, t_i) + +I(t_{i-2} = t_{i-1} = t_i) + T\alpha}$$

$$\cdot \frac{n(t_i, t_{i+1}, t_{i+2}) + I(t_{i-2} = t_i = t_{i+2}, t_{i-1} = t_{i+1}) + I(t_{i-1} = t_i = t_{i+1} = t_{i+2}) + \alpha}{n(t_{i1}, t_{i+1}) + +I(t_{i-2} = t_i, t_{i-1} = t_{i+1}) + I(t_{i-1} = t_i = t_{i+1}) + T\alpha}$$

$$\cdot \frac{n(t_i, m_i)\gamma}{n(t_i) + M_{t_i}\gamma}$$

Neural Word Embeddings (NWE) metodiga asoslangan Bayes HMM (Bayesian CS-HMM)

So‘z yasovchi qo‘shimchalardan farqli o‘laroq, flektiv qo‘shimchalar so‘zning ma‘nosini o‘zgartirmaydi. Flektiv qo‘shimchalar faqat fe’llarda jins va zamon (asgender and tense) kabi ba‘zi sintaktik vazifalarni bajaradi. Stemga asoslangan Bayes modelini mukammalroq qilish uchun biz ba’zi semantik ma’lumotlarni qo‘shamiz. Shu maqsadda stem va so‘z shakli o‘rtasidagi o‘xshashlikni baholash uchun **word2vec** usulidan olingan neyron so‘zlarni qo‘shamiz [Mikolov, Chen, Corrado, Dean 2006]. Stem va so‘z o‘rni qanchalik o‘xshash bo‘lsa, so‘z shakli ko‘proq so‘z yasovchi emas, balki fleksiyaga uchragan bo‘ladi. Modeldagi dastlabki ma‘lumot sifatida stem va so‘z shakllari o‘rtasidagi kosinus o‘xshashligidan foydalaniladi. Ushbu modeldagi $\mathbf{t}_i$ va $\mathbf{s}_i$ ning namunaviy taqsimoti 5-tenglama orqali hisoblanadi. **{stem, tag}** juftligi ehtimolini kosinus o‘xshashligiga mutanosib ravishda oshirish/kamaytirish uchun omil sifatida kosinus o‘xshashligidan foydalanamiz:

$$\sum_{t_i,s_i} P(t_i, s_i | t_{-i}, s_{-i}, \alpha, \beta) \cos(s_i, w_i) = 1$$

Stem va qo‘shimchaga asoslangan NEW Bayes HMM (Bayesian CSM-HMM)

Ushbu modelda oldingi modelga o‘xshash NEW modelidan olingan semantik ma‘lumotlarni qo‘shish orqali stem va qo‘shimchaga asoslangan Bayes HMM modelini kengaytiramiz. Shuning uchun matematik model stem-qo‘shimchaga asoslangan Bayes HMM modeli bilan bir xil. $\mathbf{t}_i$, $\mathbf{s}_i$ va $\mathbf{m}_i$ ning yangi shartli taqsimoti (9) misol bilan bir xil bo‘ladi. **{stem, suffix, tag}** ketma-ketligining ehtimolini kosinus o‘xshashligiga proporsional ravishda oshirish/kamaytirish uchun omil sifatida yana kosinus o‘xshashligidan foydalanamiz:

$$\sum_{t_i,s_i} P(t_i, s_i, m_i | t_{-i}, s_{-i}, m_{-i} \alpha, \beta, \gamma) \cos(s_i, w_i) = 1$$

Xulosa

Lug‘at orqali POS teglash va stemmingni amalga oshirish

tabiiy tilni qayta ishlashning ko‘plab vazifalariga qiyinchilik tug‘diradi. POS teglash va stemlash uchun til korpusidan foydalanish orqali lug‘at bilan bo‘ladigan muammolar hal qilinadi. Til korpusi ustida o‘tkazilgan turli tajribalar shuni ko‘rsatadiki, o‘zak ma‘lumotlarini sintaktik vazifa bilan birlashtirish morfologik jihatdan boy til uchun POS teglash natijasini yaxshilaydi, bu esa NLP vazifasining hal qilish samaradorligini oshirishga xizmat qiladi. Maqolada korpusda turli xil bog‘liqliklarni qabul qiladigan bir nechta turli xil qo‘shma modellar taqdim etildi. Umumiy eksperimental natijalar shuni ko‘rsatadiki, neural word embeddingsdan foydalanadigan Bayes HMM modeli POS teglash vazifasi uchun boshqa modellardan ustundir. Shuningdek, flektiv morfologiyani aniqlash uchun o‘zak va so‘zlar o‘rtasidagi semantik o‘xshashlikdan foydalanishda flektiv qo‘shimchalar so‘zning ma‘nosini o‘zgartirmaydi. Shu maqsadda word2vecdan olingan neural word embeddings metodidan foydalanish lozim. Natijalar shuni ko‘rsatadiki, semantik ma‘lumotlardan foydalanish stemlash va POS teglashni sezilarli darajada yaxshilaydi.

Adabiyotlar

Paice, C. D. 1994. An evaluation method for stemming algorithms. Proceedings of the 17th Annual International ACM SIGIR *Conference on Research and Development in Information Retrieval,* SIGIR. https://doi.org/10.1007/978-1-4471-2099-5_5

Anjali, M., & Jivani, G. 2011. “A Comparative Study of Stemming Algorithms”. *Int. J. Comp. Tech. Appl. 2*(6).

Gao, J., & Johnson, M. 2008. “A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers”. *EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT*, a Special Interest Group of the ACL. https://doi.org/10.3115/1613715.1613761

Goldwater, S., & Griffiths, T. L. 2007. “A fully Bayesian approach to unsupervised part-of-speech tagging”. *ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

van Gael, J., Vlachos, A., & Ghahramani, Z. 2009. “The infinite HMM for unsupervised PoS tagging”. *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009.* https://doi.org/10.3115/1699571.1699601

Taner Dinçer, B., & Karaoğlan, B. 2003. *Stemming in agglutinative languages: A probabilistic stemmer for Turkish.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Arti-

ficial Intelligence and Lecture Notes in Bioinformatics), 2869. https://doi.org/10.1007/978-3-540-39737-3_31

Ҳожиев, А. 2005. Ўзбек тилида сўз ясалиши. Ташкент.

Porter, M. F. 2006. "An algorithm for suffix stripping". *Program, 40(3).* https://doi.org/10.1108/00330330610681286

https://www.turkedebiyati.org/yapim-ekleri/

https://tilachar.ru/ru/grammar/24-grammar

Polus, M. E., & Abbas, T. 2021. "Development For Performance Of Porter Stemmer Algorithm". *Eastern-European Journal of Enterprise Technologies*, *1*(2(109)). https://doi.org/10.15587/1729-4061.2021.225362

Memon, S., Mallah, G. A., Memon, K. N., Shaikh, A., Aasoori, S. K., & Dehraj, F. U. H. 2020. "Comparative study of truncating and statistical stemming algorithms". *International Journal of Advanced Computer Science and Applications*, *2*. https://doi.org/10.14569/ijacsa.2020.0110272

Khyani, D., S, S. B., M, N. N., & M, D. B. 2021. "An Interpretation of Lemmatization and Stemming in Natural Language Processing". *Journal of University of Shanghai for Science and Technology*, *22*(10).

Tsygankin, D. v., & Ivanova, G. S. 2019. "Assimilative potential of vowel harmony in languages of agglutinative type". *Bulletin of Ugric Studies*, *9*(3). https://doi.org/10.30624/2220-4156-2019-9-3-510-518

Миртожиев, М.М. 2013. *Ўзбек тили фонетикаси.* Тошкент: Фан.

P.~Brown, V.~Della Pietra, de Souza, P., J.~Lai, & R.~Mercer. 1992. "Class-based n-gram models of natural language". *Computational Linguistics*, *18*.

Schütze, H. 1993. *Part-of-speech induction from scratch.* https://doi.org/10.3115/981574.981608

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. 1992. A practical part-of-speech tagger. https://doi.org/10.3115/974499.974523

Biemann, C. 2006. "Unsupervised part-of-speech tagging employing efficient graph clustering". *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop.* https://doi.org/10.3115/1557856.1557859

Merialdo, B. 1994. "Tagging English text with a probabilistic model". *Computational Linguistics*, *20*(2).

Banko, M., & Moore, R. C. 2004. "Part of speech tagging in context". *COLING 2004 - Proceedings of the 20th International Conference on Computational Linguistics.* https://doi.org/10.3115/1220355.1220435

Haghighi, A., & Klein, D. 2006. "Prototype-driven learning for sequence models". *HLT-NAACL 2006 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference.* https://doi.org/10.3115/1220835.1220876

Lovins, J. B. 1968. "Development of a stemming algorithm". *Mechanical Translation and Computational Linguistics*, *11*(June).

Porter, M.F. 2001. *Snowball: A language for stemming algorithms.*

Krovetz, R. (2000). "Viewing morphology as an inference process". *Artificial Intelligence*, 118 (1–2). https://doi.org/10.1016/S0004-3702(99)00101-0

Xu, J., & Croft, W. B. 1998. "Corpus-based stemming using cooccurrence of word variants". *ACM Transactions on Information Systems*, 16(1). https://doi.org/10.1145/267954.267957

Manish, Sh., Nitin, A., Bibhuti, M. "Morphology based natural language processing tools for indian languages". *In Proceedings of the 4th Annual Inter Research Institute Student Seminar in Computer Science*, IIT, Kanpur, India, April. Citeseer.

A Goweder, H Alhami, Tarik Rashed, and A Al-Musrati. "A hybrid method for stemming Arabic text". *Journal of computer Science.*

Adam, G., Asimakis, K., Bouras, C., & Poulopoulos, V. 2010. *An efficient mechanism for stemming and tagging: The case of Greek language.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6278 LNAI (PART 3). https://doi.org/10.1007/978-3-642-15393-8_44

Goldsmith, J. 2001. "Unsupervised learning of the morphology of a natural language". *Computational Linguistics*, *27*(2).
https://doi.org/10.1162/089120101750300490

Goldsmith, J. 2006. "An algorithm for the unsupervised learning of morphology". *Natural Language Engineering*, *12*(4).
https://doi.org/10.1017/S1351324905004055

Bacchin, M., Ferro, N., & Melucci, M. 2002. *The effectiveness of a graph-based algorithm for stemming.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2555. https://doi.org/10.1007/3-540-36227-4_12

Melucci, M., & Orio, N. 2003. "A novel method for stemmer generation based on Hidden Markov Models". *International Conference on Information and Knowledge Management, Proceedings.* https://doi.org/10.1145/956863.956889

McNamee, P., & Mayfield, J. 2004. "Character n-gram tokenization for European language text retrieval". *Information Retrieval*, 7(1–2). https://doi.org/10.1023/b:inrt.0000009441.78971.be

Bacchin, M., Ferro, N., & Melucci, M. 2005. "A probabilistic model for stemmer generation". *Information Processing and Management*, 41(1). https://doi.org/10.1016/j.ipm.2004.04.006

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. 1999. The web as a graph: Measurements, models, and methods. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1627. https://doi.org/10.1007/3-540-48686-0_1

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., & Datta, K. 2007. "YASS: Yet another suffix stripper". *ACM Transactions on Information Systems*, *25*(4). https://doi.org/10.1145/1281485.1281489

Peng, F., Ahmed, N., Li, X., & Lu, Y. 2007. "Context sensitive stemming for web search". *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*. https://doi.org/10.1145/1277741.1277851

Aisha, B., & Sun, M. 2009. "A statistical method for Uyghur Tokenization". *2009 International Conference on Natural Language Processing and Knowledge Engineering*, NLP-KE 2009.

https://doi.org/10.1109/NLPKE.2009.5313764

Maimaiti, M., Wumaier, A., Abiderexiti, K., & Yibulayin, T. 2017. "Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging". *Information (Switzerland)*, 8(4). https://doi.org/10.3390/info8040157

Azragul, X. Qi and A. Yusup, *2012.* "Website Phrasal Survey Based Modern Uighur Stem Extraction and Application Study", *Computer Applications and Software* 29 (3): 32-34.

Bölücü, N., & Can, B. 2019. "Unsupervised joint PoS tagging and stemming for agglutinative languages". *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18 (3). https://doi.org/10.1145/3292398

Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2006. "Distributed Representations of Words and Phrases and their Compositionality". *Neural Information Processing Systems*, 1.

https://uz.wikipedia.org/wiki/Neologizmlar

Qo'ziboyeva, G. 2022. "Tilimizga kirib kelgan neologizmlar va ularning tahlili". *International scientific journal* 3: 78-83.

The Problem of pos Tagging and Stemming for Agglutinative Languages (turkish, uyghur, uzbek languages)

Botir Elov[1]
Shahlo Hamroeva[2]
Oqila Abdullaeva[3]
Zilola Khusainova[4]
Nizomaddin Khudayberganov[5]

Abstract

The number of possible word forms in agglutinative languages is theoretically unlimited. This, in turn, creates the problem of POS tagging (part-of-speech) of out-of-vocabulary (OOV) words in agglutinative languages. In agglutinative languages, words are formed by combining stems and suffixes. Because phonetic harmony and disharmony occur when suffixes are added to the root, it is necessary to analyze both phonetic and morphological changes. When solving many NLP tasks, it is necessary to reduce word forms to their root (stemming). Removing all inflectional affixes from a word and lemmatizing the rest of the word is considered

[1]*Botir B. Elov* – doctor of philosophy of technical sciences (PhD), associate professor, Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2]*Shahlo M. Hamroeva* – doctor of philological sciences, associate professor, etc. Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
Email: shaxlo.xamrayeva@navoiy-uni.uz
ORCID: 0000-0002-5429-4708
[3]*Oqila K. Abdullayeva* – doctor of philosophy in philology, Senior teacher, Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
Email: abdullayeva.oqila@navoiy-uni.uz
ORCID: 0000-0002-2524-4832
[4]*Zilola Y. Khusainova* – PhD student of Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: xusainovazilola@navoiy-uni.uz
ORCID: 0000-0003-4357-7515
[5] *Nizomaddin U. Khudayberganov* – Teacher of Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail:nizomaddin@navoiy-uni.uz
ORCID: 0000-0002-6213-3015
For citation: Elov, B.B., Hamroeva, Sh.M., Abdullaeva, O.K., Khusainova, Z.Y., Khudoyberganov, N.U. 2023. "The Problem of pos Tagging and Stemming for Agglutinative Languages (turkish, uyghur, uzbek languages)". *Uzbekistan: Language and Culture* 2: 6-39.

one of the important tasks of natural language processing (NLP), and this process is called stemming. The stemming process is important in information retrieval (IR) systems.

Key words: *part-of-speech, POS tagging, stemming, information retrieval, IR, stemming algorithms.*

References

Paice, C. D. 1994. An evaluation method for stemming algorithms. Proceedings of the 17th Annual International ACM SIGIR *Conference on Research and Development in Information Retrieval,* SIGIR. https://doi.org/10.1007/978-1-4471-2099-5_5

Anjali, M., & Jivani, G. 2011. "A Comparative Study of Stemming Algorithms". *Int. J. Comp. Tech. Appl. 2*(6).

Gao, J., & Johnson, M. 2008. "A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers". *EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT*, a Special Interest Group of the ACL. https://doi.org/10.3115/1613715.1613761

Goldwater, S., & Griffiths, T. L. 2007. "A fully Bayesian approach to unsupervised part-of-speech tagging". *ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

van Gael, J., Vlachos, A., & Ghahramani, Z. 2009. "The infinite HMM for unsupervised PoS tagging". *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009.* https://doi.org/10.3115/1699571.1699601

Taner Dinçer, B., & Karaoğlan, B. 2003. *Stemming in agglutinative languages: A probabilistic stemmer for Turkish.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2869. https://doi.org/10.1007/978-3-540-39737-3_31

Ҳожиев, А. 2005. Ўзбек тилида сўз ясалиши. Ташкент.

Porter, M. F. 2006. "An algorithm for suffix stripping". *Program, 40(3).* https://doi.org/10.1108/00330330610681286

https://www.turkedebiyati.org/yapim-ekleri/

https://tilachar.ru/ru/grammar/24-grammar

Polus, M. E., & Abbas, T. 2021. "Development For Performance Of Porter Stemmer Algorithm". *Eastern-European Journal of Enterprise Technologies*, *1*(2(109)). https://doi.org/10.15587/1729-4061.2021.225362

Memon, S., Mallah, G. A., Memon, K. N., Shaikh, A., Aasoori, S. K., & Dehraj, F. U. H. 2020. "Comparative study of truncating and statistical stemming algorithms". *International Journal of Advanced Com-*

puter Science and Applications, *2*. https://doi.org/10.14569/ijacsa.2020.0110272

Khyani, D., S, S. B., M, N. N., & M, D. B. 2021. "An Interpretation of Lemmatization and Stemming in Natural Language Processing". *Journal of University of Shanghai for Science and Technology*, *22*(10).

Tsygankin, D. v., & Ivanova, G. S. 2019. "Assimilative potential of vowel harmony in languages of agglutinative type". *Bulletin of Ugric Studies*, *9*(3). https://doi.org/10.30624/2220-4156-2019-9-3-510-518

Миртожиев, М.М. 2013. *Ўзбек тили фонетикаси.* Тошкент: Фан.

P.~Brown, V.~Della Pietra, de Souza, P., J.~Lai, & R.~Mercer. 1992. "Class-based n-gram models of natural language". *Computational Linguistics*, *18*.

Schütze, H. 1993. *Part-of-speech induction from scratch.* https://doi.org/10.3115/981574.981608

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. 1992. A practical part-of-speech tagger. https://doi.org/10.3115/974499.974523

Biemann, C. 2006. "Unsupervised part-of-speech tagging employing efficient graph clustering". *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop.* https://doi.org/10.3115/1557856.1557859

Merialdo, B. 1994. "Tagging English text with a probabilistic model". *Computational Linguistics*, *20*(2).

Banko, M., & Moore, R. C. 2004. "Part of speech tagging in context". *COLING 2004 - Proceedings of the 20th International Conference on Computational Linguistics.* https://doi.org/10.3115/1220355.1220435

Haghighi, A., & Klein, D. 2006. "Prototype-driven learning for sequence models". *HLT-NAACL 2006 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference.* https://doi.org/10.3115/1220835.1220876

Lovins, J. B. 1968. "Development of a stemming algorithm". *Mechanical Translation and Computational Linguistics*, *11*(June).

Porter, M.F. 2001. *Snowball: A language for stemming algorithms.*

Krovetz, R. (2000). "Viewing morphology as an inference process". *Artificial Intelligence*, 118 (1–2). https://doi.org/10.1016/S0004-3702(99)00101-0

Xu, J., & Croft, W. B. 1998. "Corpus-based stemming using cooccurrence of word variants". *ACM Transactions on Information Systems*, 16(1). https://doi.org/10.1145/267954.267957

Manish, Sh., Nitin, A., Bibhuti, M. "Morphology based natural language processing tools for indian languages". *In Proceedings of the 4th Annual Inter Research Institute Student Seminar in Computer Science*, IIT, Kanpur, India, April. Citeseer.

A Goweder, H Alhami, Tarik Rashed, and A Al-Musrati. "A hybrid method for stemming Arabic text". *Journal of computer Science.*

Adam, G., Asimakis, K., Bouras, C., & Poulopoulos, V. 2010. *An efficient mechanism for stemming and tagging: The case of Greek language.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6278 LNAI (PART 3). https://doi.org/10.1007/978-3-642-15393-8_44

Goldsmith, J. 2001. "Unsupervised learning of the morphology of a natural language". *Computational Linguistics*, *27*(2).
https://doi.org/10.1162/089120101750300490

Goldsmith, J. 2006. "An algorithm for the unsupervised learning of morphology". *Natural Language Engineering*, *12*(4).
https://doi.org/10.1017/S1351324905004055

Bacchin, M., Ferro, N., & Melucci, M. 2002. *The effectiveness of a graph-based algorithm for stemming.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2555. https://doi.org/10.1007/3-540-36227-4_12

Melucci, M., & Orio, N. 2003. "A novel method for stemmer generation based on Hidden Markov Models". *International Conference on Information and Knowledge Management, Proceedings.* https://doi.org/10.1145/956863.956889

McNamee, P., & Mayfield, J. 2004. "Character n-gram tokenization for European language text retrieval". *Information Retrieval*, 7(1–2). https://doi.org/10.1023/b:inrt.0000009441.78971.be

Bacchin, M., Ferro, N., & Melucci, M. 2005. "A probabilistic model for stemmer generation". *Information Processing and Management*, 41(1). https://doi.org/10.1016/j.ipm.2004.04.006

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. 1999. The web as a graph: Measurements, models, and methods. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1627. https://doi.org/10.1007/3-540-48686-0_1

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., & Datta, K. 2007. "YASS: Yet another suffix stripper". *ACM Transactions on Information Systems*, *25*(4). https://doi.org/10.1145/1281485.1281489

Peng, F., Ahmed, N., Li, X., & Lu, Y. 2007. "Context sensitive stemming for web search". *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*. https://doi.org/10.1145/1277741.1277851

Aisha, B., & Sun, M. 2009. "A statistical method for Uyghur Tokenization". *2009 International Conference on Natural Language Processing and Knowledge Engineering*, NLP-KE 2009.
https://doi.org/10.1109/NLPKE.2009.5313764

Maimaiti, M., Wumaier, A., Abiderexiti, K., & Yibulayin, T. 2017. "Bidirectional long short-term memory network with a condition-

al random field layer for Uyghur part-of-speech tagging". *Information (Switzerland)*, 8(4). https://doi.org/10.3390/info8040157

Azragul, X. Qi and A. Yusup, *2012.* "Website Phrasal Survey Based Modern Uighur Stem Extraction and Application Study", *Computer Applications and Software* 29 (3): 32-34.

Bölücü, N., & Can, B. 2019. "Unsupervised joint PoS tagging and stemming for agglutinative languages". *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18 (3). https://doi.org/10.1145/3292398

Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2006. "Distributed Representations of Words and Phrases and their Compositionality". *Neural Information Processing Systems*, 1.

https://uz.wikipedia.org/wiki/Neologizmlar

Qo'ziboyeva, G. 2022. "Tilimizga kirib kelgan neologizmlar va ularning tahlili". *International scientific journal* 3: 78-83.

MAQOLA TAQDIM QILISH TALABLARI

O‘zbekiston: til va madaniyat (O‘zTM) – zamonaviy O‘zbekiston (sobiq Turkiston) bilan bog‘liq bevosita Markaziy Osiyo mintaqasini birlashtiradigan til, tarix, san’at, etnografiya, madaniyat va ijtimoiy fanlar sohalarini qamrab olgan ilmiy jurnaldir. O‘zTM munozarali, zamonaviy, innovatsion, konseptual jihatdan qiziqarli, original mavzudagi ilmiy tadqiqotlarni nashr qiladi. Jurnal lingvistika, adabiyotshunoslik, tarjimashunoslik, din, falsafa, ilohiyot, fan, ta’lim, metodika, sotsiologiya, psixologiya, tarix, madaniyat, san’at, etnologiya, etnografiya, antropologiyaga oid ilmiy yo‘nalishdagi maqolalar va taqrizlar hamda konferensiya hisobotlarini qabul qiladi.

I. Maqola taqdim etish uchun umumiy talablar

Qo‘lyozmalar o‘zbek, ingliz, rus, fors, shuningdek, boshqa turkiy tillarda ham qabul qilinadi. Agar muallif o‘z maqolasini jurnalning muayyan sonida nashr ettirmoqchi bo‘lsa, unda qo‘lyozma jurnal nashridan kamida besh oy oldin taqdim etilishi lozim.

Qo‘lyozmalar MS Word (.doc) formatida (uzlangcult@gmail.com) elektron pochtasiga yuboriladi. Iqtiboslar va izohlar uchun MS Word menejerini qo‘llash mumkin.

Barcha qo‘lyozmalar tahririyatga muallif (mualliflar) haqidagi qisqacha ma’lumot bilan taqdim etiladi.

Asosiy matn *Times New Roman* shrifti, 14 hajm, satr oralig‘i 1 interval, hoshiyalar chapdan 3 sm, o‘ngdan 1,5 sm, yuqori va pastdan 2 sm bo‘lishi kerak.

Maqolalar *The Chicago Manual of Style, 16th Edition* formatida shakllantiriladi. Maqola matni 3 000–5 000 so‘zdan iborat bo‘lishi kerak.

O‘zbek va ingliz tillarida 100–150 so‘zdan iborat abstrakt (annotatsiya) va 5–10 so‘zdan kam bo‘lmagan kalit so‘zlar (o‘zbek va ingliz tillarida). Abstraktda maqolaning qisqacha mazmuni va dolzarbligi, tadqiqot natijalari aks etishi lozim.

Adabiyotlar ro‘yxati 5 sahifadan oshmasligi kerak.

Kitobga taqriz (ingliz yoki boshqa tillarda bo‘lishi mumkin) 1500 so‘zdan oshmasligi talab etiladi.

Taqriz formati: 1) sarlavha: kitob nomi, muallif (mualliflar), nashr qilingan shahar: nashriyot nomi, nashr yili, sahifasi soni. Narxi, ISBN raqami, (qattiq/yumshoq muqova); 2) taqriz so‘ngida: taqrizchining F.I.O., ish joyi, pochta manzili.

II. Maqola bo‘limlarini rasmiylashtirish

Maqola nomi – normal harflarda, to‘q bo‘yoqda, 16 hajm.

Maqola nomi o‘zbek va ingliz tillarida (agar maqola boshqa tilda yozilgan bo‘lsa, maqola yozilgan til va ingliz tilida) beriladi.

Maqola kirish, asosiy qism bo‘limlari va xulosadan tashkil topadi.

Maqola bo‘limlari sarlavhasi – to‘q bo‘yoqda, 14 hajm.

III. Maqolada tarjimalardan foydalanish

Boshqa tillardagi matn yoki boshqa manbalar tarjimoni aniq ko‘rsatilishi kerak. Agar matn maqola muallifi tomonidan tarjima qilingan bo‘lsa, u holda “tarjima muallifniki”

shaklida beriladi.

Rasmiy nashrdan olingan tarjima-matn tahrir qilinmaydi.

Zarur holatda tarjima matnga sana, turli diakritik belgilar va boshqa elementlar kiritilishi mumkin.

Tarjima qilingan matn olingan manba nomi asl holicha beriladi. Zarur deb topilsa, uning nomi qavs ichida berilishi mumkin.

Geografik nomlar tarjima qilinmaydi va asl shaklida beriladi.

Tashkilotlar nomi tarjima qilinmaydi va asl shaklida beriladi.

Davr nomi rasmiy qabul qilingan shaklda beriladi.

IV. Ko‘chirma va tarjima parchaning berilishi

Manbadan olingan ko‘chirma parcha asosiy matndan 1 qator tashlab ajratiladi, satr oralig‘i 1 interval, markazda, 12 hajmda yoziladi.

Ko‘chirmaning tarjimasi qavs ichida ( ) satr boshidan yozilishi kerak. Bunday ko‘chirma *Times New Roman* shrift, 12 hajm, normal yozuvda beriladi.

V. Havola va izohlar berish

Manbaga havola matn ichida to‘rtburchak qavsda [ ] beriladi. Havola qilingan manbalar bir nechta bo‘lsa, ular nuqtali vergul (;) bilan ajratiladi.

Izohlar tegishli sahifa pastida, tartib raqami bilan joylashtiriladi.

VI. Qo‘lyozma (toshbosma) manbalar va nashr etilgan asarlar bibliografiyasi

Bibliografiyada muallif yoki asar nomi satr boshidan, boshqa barcha qatorlari xatboshidan yoziladi. Adabiyotlar *bibliografiya*da o‘zbek lotin alifbosi tartibida ko‘rsatiladi.

VII.Qo‘lyozma va toshbosma manbalar bibliografiyasi

Qo‘lyozma yoki toshbosma manbalarni bibliografiyada o‘zi yozilgan grafikada berish maqsadga muvofiq. Lotin alifbosidagi transliteratsiyasini berish ham mumkin. Ba’zan qo‘lyozma asarning nomi muallif ismidan oldin yozilishi ham mumkin.

Muallif nomi. Ko‘chirilgan asr (agar mavjud bo‘lsa). Asar nomi. Qo‘lyozma (toshbosma): saqlanayotgan joy, inventar raqam.

Xondamir. XV asr. Makorim ul-axloq. Qo‘lyozma: O‘zFASHI, № 742.

VIII.1. Kitoblar uchun

Bibliografiyada:

Familiya, ism. Nashr yili. *Kitob nomi,* Shahar: Nashriyot nomi.

Qudratullayev, Hasan. 2018. *Boburning adabiy-estetik olami.* Toshkent: Ma’naviyat.

Matnda kitobga havola:

[Familiya kitob nashr yili, sahifa raqami]

[Qudratullayev 2018, 99]

Agar bir muallifning bir yilda nashr qilingan kitoblaridan foydalanilgan bo‘lsa, bibliografiyada kitobning nashr yili o‘zbek lotin alifbosi harflari bilan ajratilib ko‘rsatiladi.

Sirojiddinov, Shuhrat. 2011 (a). *Alisher Navoiy: manbalarning qiyosiy-tipologik, tekstologik tahlili.* Toshkent: Akademnashr.

Sirojiddinov, Shuhrat. 2011 (b). *O‘zbek adabiyotining falsafiy sarchashmalari*. Toshkent: Akademnashr.

Matnda kitobga havola:
[Familiya, kitob nashr yili, sahifa raqami]
[Sirojiddinov 2011 (a), 99]
[Sirojiddinov 2011 (b), 67]

Ikki muallif tomonidan yozilgan kitobni bibliografiyada berish tartibi:

Familiya, Ism va Ism Familiya. Nashr yili. *Kitobning nomi*. Shahar: Nashriyot nomi.

Abdurahmonov, G‘anijon, Alibek Rustamov. 1984. *Navoiy tilining grammatik xususiyatlari.* Toshkent: Fan.

Matnda kitobga havola:
[Familiya va Familiya nashr yili, sahifa raqami]
[Abdurahmonov, Rustamov 1984, 52]

Agar kitobning uch va undan ortiq mualliflari bo‘lsa, bibliografiyada barcha mualliflarning ismlari to‘liq yoziladi. Bunday kitobga havola qilinganda birinchi muallif ismi yoziladi va “boshqalar” deb ko‘rsatiladi.
[Familiya va boshqalar kitob nashr yili, sahifa raqami]
[Vohidov va boshqalar 2010, 847]

Kitob yoki to‘plam maqolasini bibliografiyada berish tartibi:
Familiya, ism. Nashr yili. “Maqola nomi.” *Kitob yoki to‘plam nomi*, Ism Familiya, Ism Familiya muharrirligida, maqola sahifasi raqamlari. Shahar: Nashriyot.

Abdug‘afurov, Abdurashid. 2016. “Badoye’ ul-bidoya”ning tuzilish sanasi”. *XX asr o‘zbek mumtoz adabiyotshunosligi*, Olim To‘laboyev muharrirligida, 174–184. Toshkent: O‘zbekiston milliy ensiklopediyasi.

Matnda kitob yoki to‘plam maqolasiga havola:
[Familiya nashr yili, sahifa raqami]
[Abdug‘afurov 2016, 176]

Elektron shaklda nashr qilingan kitoblar uchun:
Elektron kitobning bir nechta formati bo‘lsa, bibliografiyada foydalanilgan format ko‘rsatiladi. Elektron kitobning internet manzili (URL) hamda shu manba olingan sana ko‘rsatilishi lozim.

Elektron kitobni bibliografiyada berish:
Familiya, Ism. Nashr yili. *Kitob nomi.* Shahar: Nashriyot nomi. URL. Foydalanilgan sana.

Mamatov, Ulug‘bek. 2018. *O‘zbekiston madaniyatida tarixiy janrdagi tasviriy san’at asarlari.* Toshkent: Mumtoz so‘z. https://kitobxon.com/uz/catalog/sanat/. 12.03.2019.

Matnda elektron kitobga havola:
[Familiya nashr yili, sahifa raqami]
[Маматов 2018, 11]

Ikki muallif tomonidan yozilgan elektron kitobni bibliografiyada berish tartibi:

Familiya, Ism va Ism Familiya. Nashr yili. *Kitobning nomi.* Shahar: Nashriyot nomi. Internet adres (URL).

Sirojiddinov, Shuhrat va Sohiba Umarova. 2017. *O‘zbek matnshunosligi qirralari.* Chikago: Chikago universiteti nashriyoti. http://press-pubs.uchicago.edu/founders/.

Matnda elektron kitobga havola:
[Familiya nashr yili, sahifa raqami]
[Sirojiddinov 2017, 19-hujjat]

VIII.2. Jurnal maqolasi uchun

Chop etilgan jurnal maqolasini bibliografiyada berish tartibi:

Familiya, Ism. Nashr yili. “Maqola nomi”. *Jurnal nomi* jurnal soni: maqola sahifalari.

Mahmudov, Nizomiddin. 2013. “Termin, badiiy so‘z va metafora”. *O‘zbek tili va adabiyoti* 4: 3 – 8. Toshkent.

Matnda jurnal maqolasiga havola:
[Familiya nashr yili, sahifa raqami]
[Mahmudov, 2013, 5]

Elektron jurnal uchun:

Elektron jurnal uchun jurnalning DOI manzili ko‘rsatiladi. Agar DOI manzili mavjud bo‘lmasa, internet adresi ko‘rsatilishi kerak (URL). DOI – bu o‘zgarmas ID bo‘lib, internet tarmoqlarining elektron adreslari tizimiga ulangan, ya’ni manbani boshqaruvchi http://dx.doi.org/ manzildir.

Elektron jurnal maqolasini bibliografiyada berish:

Familiya, Ism. Nashr yili. “Maqola nomi.” *Jurnal nomi* jurnal soni: maqola sahifalari. DOI adres (yoki URL).

Aminov, Hasan. 2018. “O‘zbekiston san’atida temuriylar siymosi”. *O‘zbekistonda xorijiy tillar* 2: 246 – 253. doi: 10.36078/1596780051.

Matnda maqolaga havola:
[Familiya nashr yili, sahifa raqami]

VIII.3. Gazeta yoki ilmiy-ommabop jurnal uchun

Gazeta yoki ilmiy-ommabop jurnal maqolasiga havola matn shaklida beriladi (masalan, Muhammadjon Imomnazarovning 27.02.2005dagi "O‘zbekiston adabiyoti va san'ati" gazetasida chop etilgan maqolasida aytilganidek…); odatda, bunday manbalar umumiy adabiyotlar ro‘yxatida keltirilmaydi. Agar keltirilsa, kitoblarga qo‘yiladigan talablarga asosan beriladi.

Agar onlayn maqolaga havola berilayotgan bo‘lsa, uning internet manzili (URL), maqola olingan sana ko‘rsatilishi kerak.

Gazeta yoki ilmiy-ommabop jurnal maqolasini bibliografiyada berish:

Familiya, Ism. Nashr yili. "Maqola nomi." *Gazeta-Jurnal nomi*, nashr sanasi.

Imomnazarov, Muhammadjon. 2005. "Jomiy "Xamsa" yozganmi?." *O‘zbekiston adabiyoti va san'ati*, January 25.

Matnda maqolaga havola:

[Familiya nashr yili, sahifa raqami]

[Imomnazarov 2005, 4]

Elektron gazeta yoki ilmiy-ommabop jurnal maqolasini bibliografiyada berish:

Familiya, Ism. Nashr yili. "Maqola nomi." *Jurnal nomi*, nashr sanasi. Internet adres.

Jabborov, Rustam. 2019. "Navoiyning Tabrizda yashagan xorazmlik kotibi". *UZA: O‘zbekiston Milliy axborot agentligi*, 08.12. https://uza.uz/uz.

Matnda maqolaga havola:

[Familiya nashr yili, sahifa raqami]

[Jabborov 2010, 17]

Maqola so‘ngida foydalanilgan adabiyotlar o‘zbek lotin alifbosi tartibida beriladi. Adabiyotlar ro‘yxati ikki qismdan iborat bo‘lishi, birinchi qismda foydalanilgan adabiyot chop etilgan grafikada yuqorida ko‘rsatilgan shaklda rasmiylashtirilishi, ikkinchi qismda esa barcha foydalanilgan adabiyotlar o‘zbek lotin alifbosida berilishi talab qilinadi. Misol uchun:

Adabiyotlar

Баранов, Х.К. 1958. Арабско – русский словарь. Москва: Наука.

Adabiyotlar

Baranov, X.K. 1958. Arabsko – russkiy slovar. Moskva: Nauka.

Maqolani rasmiylashtirish talablarining ingliz tilidagi variantini "The Chicago Manual of Style, 16th Edition" qo‘llanmasi yoki https://www.chicagomanualofstyle. havolasidan ko‘rib olishingiz mumkin.

GUIDELINES FOR CONTRIBUTORS

Uzbekistan: language and culture is an academic journal, publishing research in linguistics, history, literature, translation studies, arts, ethnography, philosophy, anthropology and social studies. We aim to publish cutting edge, innovative, conceptually interesting, original case studies and new research, which shape and lead debates in multifaceted studies. We do not publish economic analyses or policy papers. Any opinions and views expressed in publications are the opinions and views of the authors, and the publishers are not responsible for the views/ reviews of the contributors.

The journal is published four times a year. The language of articles can be English, Russian and Uzbek. Other Turkic languages are also welcomed. In addition to research articles, the journal welcomes book reviews, literature overviews, conference reports and research project announcements.

1. General

- Submission Guideline

1. Manuscripts may be submitted at any time during the year. However, if the author wishes to have his/her manuscript published in a certain issue of the journal, the submission should be made at least five months in advance of the proposed publication date.

2) Manuscripts should be submitted by email (uzlangcult@gmail.com) as an attachment in MS Word document (.doc) format and use MS Word Source.

3) All manuscripts should be submitted with a cover page including an email address, a mailing address and a short introduction about the author(s) /contributor(s)'.

2. Manuscript format

1) The main texts should be written in Times New Roman font, 12 point, and single-spaced in 44 pagination with 1-inch margins.

2) Submissions must follow the author-date system of *The Chicago Manual of Style*, 16th Edition.

3) Quotations are given in brackets in the text.

4) A research article should normally be no more than 9,000 words in length, including the following contents:

- an abstract of 150-200 words (in English, Russian, and Uzbek) and seven to ten keywords;
- a list of references of no more than five (5) pages;
- tables and figures, if any.

5) A book review should generally be about 1,500 English words (or other languages) in length, and must include the heading and closing in the following format:

- Heading: *Title of the Book*. By Author's Name(s). City of Publication: Publisher Name, Year. pp. Price, ISBN:, (hardcover/paperback).
- Closing: Book reviewer's name, affiliation and postal address at the end.

6) Style Points Headings. Limit: Four levels.

Level 1. Title Style (e.g. the first letter of each word upper case, except prepositions), Bold, and 14 point.

Level 2. Title Style, Italics, 14 Point.

Level 3. Modified "down" style (first letter upper case, or first letter of first two words if the first word is an article), Bold, and l2 point.

Level 4. Modified down style, Bold, 11 point.

3. Style and Usage

1) Translation

- Translated excerpts from classical texts or non-English sources should be annotated with clarification of its original/published language and translator. Likewise, "Author's own" translations of quoted texts should be noted as such.

- The author is expected to provide an English translation of key terms in the work, rather than a translator without expertise in the subject.

- Excerpts or quoted texts from published translation will not be edited. However, UzLC editors may query or modify translations of key terms or texts provided by the author.

- Where necessary, short supplementary information such as dates, an item in its original characters, or the Romanized form of a non-English item, may be included.

- Names of foreign publishers, and titles of sources published in a foreign language should primarily appear in Romanized form without translation. However, if necessary, a translation may be added in brackets ([ ]).

2) Names and Terms

- Place Names (foreign):

Designation for division of areas should be either translated or hyphenated after the given area name.

Designation for geographical/structure names are not hyphenated, and appear without the equivalent English term.

Institutional names are considered proper nouns. Their names should appear following the preference of the individual institutions.

3) The descriptive designation of a period is usually lowercase, except for proper names or traditionally capitalized terms.

4. Quotation

1) Block Quotations:

- A block quotation should start with double line spacing and an indentation from the left margin. From the second paragraph of the block quotation, additional paragraph indentation is needed.

Texts in block quotation should be written in Times New Roman 10 pts., and not be entirely italicized.

5. Others

1) There is one space after sentence punctuation and not two.

2) The end parenthesis, closing quotation mark, and footnote numbers come after the sentence punctuation.

3) For parentheses within parentheses, use brackets ([ ]).

6. Basic Citation Format

The following examples illustrate citations using the **author-date** system. Each example of a reference list entry is accompanied by an example of a corresponding parenthetical citation in the text. For more details and many more examples, see chapter 15 of The Chicago Manual of Style.

BOOK

Reference List (hanging indent):

Pollan, Michael. 2006. *The Omnivore's Dilemma: A Natural History of How Eating Has Evolved*. New York: Penguin.

In Text Cite:

[Pollan 2006, 99–100]

Reference List (hanging indent):

Ward, Geoffrey C., and Ken Burns. 2007. *The War: An Intimate History, 1941–1945*. New York: Knopf.

In Text Cite:

[Ward and Burns 2007, 52]

For four or more authors, list all of the authors in the reference list; in the text, list only the first author, followed by et al. ("and others"):

[Barnes et al. 2010, 847]

Reference List (hanging indent) book chapter:

Kelly, John D. 2010. "Seeing Red: Mao Fetishism, Pax Americana, and the Moral Economy of War." In *Anthropology and Global Counterinsurgency*, edited by John D. Kelly, Beatrice Jauregui, Sean T. Mitchell, and Jeremy Walton, 67–83. Chicago: University of Chicago Press.

In Text Cite:

[Kelly 2010, 77]

Chapter of an edited volume originally published elsewhere (as in primary sources):

Reference List (hanging indent) book originally published elswehre:

Cicero, Quintus Tullius. 1986. "Handbook on Canvassing for the Consulship." In *Rome: Late Republic and Principate*, edited by Walter Emil Kaegi Jr. and Peter White. Vol. 2 of University of Chicago Readings in Western Civilization, edited by John Boyer and Julius Kirshner, 33–46. Chicago: University of Chicago Press. Originally published in Evelyn S. Shuckburgh, trans., *The Letters of Cicero*, vol. 1 (London: George Bell & Sons, 1908).

In Text Cite:

[Cicero 1986, 35]

BOOK PUBLISHED ELECTRONICALLY

If a book is available in more than one format, cite the version you consulted. For books consulted online, list a URL; include an access date only if one is required by your discipline. If no fixed page numbers are available, you can include a section title or a

chapter or other number.

Reference List (hanging indent):

Austen, Jane. 2007. *Pride and Prejudice: A Novel in Five Books*. New York: Penguin Classics. Kindle edition.

In Text Cite:

[Austen 2007, 101]

Reference List (hanging indent):

Kurland, Philip B., and Ralph Lerner, eds. 1987. *The Founders' Constitution*. Chicago: University of Chicago Press. htpp: //press-pubs.uchicago.edu/founders

In Text Cite:

[Kurland and Lerner, chap. 10, doc. 19]

JOURNAL ARTICLE

Article in a print journal

In the text, list the specific page numbers consulted, if any. In the reference list entry, list the page range for the whole article.

Reference List (hanging indent):

Weinstein, Joshua I. 2009. "The Market in Plato's Republic." *Classical Philology* 104:439–58.

In text cite:

[Weinstein 2009, 440]

Article in an online journal

Include a DOI if the journal lists one. A DOI is a permanent ID that, when appended to http://dx.doi.org / in the address bar of an Internet browser, will lead to the source. If no DOI is available, list a URL. Include an access date only if one is required by your discipline.

Reference List (hanging indent):

Kossinets, Gueorgi, and Duncan J. Watts. 2009. "Origins of Homophily in an Evolving Social Network."*American Journal of Sociology* 115:405–50. doi:10.1086/599247.

In text cite:

[Kossinets and Watts 2009, 411]

Article in a newspaper or popular magazine

Newspaper and magazine articles may be cited in running text ("As Sheryl Stolberg and Robert Pear noted in a New York Times article on February 27, 2010…"); they are commonly omitted from a reference list. The following examples show more formal versions of the citations. If you consulted the article online, include a URL; include an access date only if your discipline requires one. If no author is identified, begin the citation with the article title.

Reference List (hanging indent):

Mendelsohn, Daniel. 2010. "But Enough about Me." *New Yorker*, January 25.

In text cite:

[Mendelsohn 2010, 68]

Reference List (hanging indent):

Stolberg, Sheryl Gay, and Robert Pear. 2010. "Wary Centrists Posing Challenge in Health Care Vote." New York Times, February 27. http://www.nytimes.com/2010/02/28/us/politics/28health.html.

In text cite:

[Stolberg and Pear 2010, 12]

WEBSITE

A citation to website content can often be limited to a mention in the text ("As of July 19, 2008, the McDonald's Corporation listed on its website . . ."). If a more formal citation is desired, it may be cited as in the examples below. Because such content is subject to change, include an access date or, if available, a date that the site was last modified. In the absence of a date of publication, use the access date or last-modified date as the basis of the citation.

Bibliography (hanging indent):

Google. 2009. "Google Privacy Policy." Last modified March 11. http://www.google.com/intl/en/privacypolicy.html.

In text cite:

[Google 2009]

Reference List (hanging indent):

McDonald's Corporation. 2008. "McDonald's Happy Meal Toy Safety Facts." http://www.mcdonalds.com/corp/about/factsheets.html.

In text cite:

[McDonald's 2008]

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: www.uzlc.navoiy-uni.uz
E-mail: uzlangcult@gmail.com

Bosishga 30.06.2023-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.