UO‘K 811.512.133'32

O‘ZBEK TILIDAGI MATNLAR KOREFERENSIYASINI AVTOMATIK ANIQLASH TIZIMINING LINGVISTIK TA’MINOTINI SHAKLLANTIRISH XUSUSIDA

Shahlo Abdisalomova Abdimurod qizi
Tayanch doktorant
abdisalomovashahlo@gmail.com
ToshDO‘TAU

Annotatsiya. Tabiiy tilga ishlov berish uchun ishlab chiqilgan tizimlar, birinchi navbatda, ma’lumotlar bazasi bilan ta’minlanishi kerak. Ma’lumotlar bazasidagi ma’lumotlarni tez fursatda qayta ishlash, ulardan qulay tarzda foydalanish maqsadida lingvistik bazalar ma’lumotlar bazasini boshqarish tizimlari (MBBT)da saqlanadi. Ushbu maqolada ma’lumotlar bazasini yaratishda e’tibor qaratish zarur bo‘lgan jihatlar, o‘zbek tilidagi matnlar koreferensiyasini avtomatik aniqlash tizimining lingvistik ta’minotini tuzishda, ularni annotatsiyalashda tayanilgan tamoyillar, ma’lumotlar bazasini SQL Serverga kiritish jarayoni haqida so‘z yuritiladi. Ma’lumotlarni MBBTda saqlash nafaqat tahlilni soddalashtiradi, balki NLP modellari bilan ishlashda aniq va izchil ma’lumot almashinuvini ta’minlaydi. Tadqiqot o‘zbek tilida koreferensiyani modellashtirishga doir dastlabki amaliy yondashuvlardan biri bo‘lib, koreferensiyani avtomatik aniqlash tizimi, o‘zbek tilidagi matnlarni tahlil qiluvchi algoritmlar va ma’lumotlar bazasi asosida ishlovchi lingvistik vositalarni yaratishda muhim poydevor bo‘lib xizmat qiladi.

Abstract. Systems designed for Natural Language Processing must first be equipped with a database. In order to quickly process and conveniently use the data in the database, linguistic databases are stored in database management systems (DBMS). This article discusses the aspects that need to be paid attention to when creating a database, the principles used in creating the linguistic support of the automatic coreference detection system for Uzbek texts, their annotation, and the process of entering the database into SQL Server. Storing data in DBMS not only simplifies analysis, but also ensures accurate and consistent information exchange when working with NLP models. The study is one of the first practical approaches to modeling coreference in the Uzbek language and serves as an important foundation for creating an automatic coreference detection system, algorithms for analyzing Uzbek texts, and linguistic tools based on the database.

Абстракт. Системы, предназначенные для обработки естественного языка, должны быть в первую очередь оснащены базой данных. Лингвистические базы данных хранятся в системах управления базами данных

(СУБД) для быстрой обработки и удобного использования данных базы данных. В статье рассматриваются аспекты, которые необходимо учитывать при создании базы данных, принципы, используемые при создании лингвистического обеспечения системы автоматического определения корреспонденции узбекских текстов, их аннотирование, а также процесс ввода базы данных в SQL Server. Хранение данных в MBBT не только упрощает анализ, но и обеспечивает точный и последовательный обмен информацией при работе с моделями NLP. Исследование является одним из первых практических подходов к моделированию кореференции в узбекском языке и служит важной основой для создания системы автоматического определения кореференции, алгоритмов анализа узбекских текстов и лингвистических инструментов на основе баз данных.

Kalit so‘zlar: *koreferensiya, NLP, ma’lumotlar bazasi, annotatsiyalash, SQL Server, korpus, tizim, o‘zbek tili*

KIRISH

Hozirgi kunda sun’iy intellekt va tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sohasidagi rivojlanishlar inson tilini tushunish, uni tahlil qilish va undan samarali foydalanish imkoniyatlarini kengaytirmoqda. Ayniqsa, matnlarda uchraydigan **koreferensiya** hodisasi, ya’ni biror so‘z yoki ifodaning boshqa bir so‘z yoki ifodani nazarda tutishi – tabiiy til semantikasi va kontekstini tushunishda muhim ahamiyatga ega. Masalan, *“Kelajak uni kashf etganlarga nasib qiladi”* jumlasidagi *“u”* olmoshining *“kelajak”* so‘ziga tegishli ekanini aniqlash koreferensiya hodisasiga misoldir. Bu hodisani aniqlash nafaqat tilshunoslikda, balki mashina tarjimasi, savol-javob tizimlari, matnni umumlashtirish kabi ko‘plab NLP vazifalarida ham muhim o‘rin tutadi.

Koreferensiyani avtomatik aniqlashdagi birinchi qadam modelni o‘qitish va baholash uchun lingvistik ta’minotni shakllantirishdan iborat. Jahon tajribasiga nazar tashlasak, ko‘pgina tillarda koreferensiyani avtomatik aniqlash tizimlari ishlab chiqilgan va ular maxsus ma’lumot bazalarida sinovdan o‘tkazilgan [1, 2, 3]. O‘zbek tilida koreferensiyani avtomatik aniqlash tizimini yaratishdagi asosiy muammo o‘zbek tilida annotatsiyalangan katta hajmli ma’lumotlar bazasiga bo‘lgan ehtiyoj bilan bog‘liq. O‘zbek tilida turli maqsadga yo‘naltirilgan korpuslar mavjud bo‘lsa-da [4, 5, 6, 7] ularni koreferensiyani avtomatik aniqlash tizimiga to‘g‘ridan to‘g‘ri tatbiq etish kutilgan natijani bermaydi. Demak, o‘zbek tilida koreferensiya hodisasi kuzatiladigan matnlarni aniqlash va ularni strukturaviy ma’lumot shaklida bazaga joylashtirish bo‘yicha tadqiqotlarga ehtiyoj katta. O‘zbek tilidagi matnlar koreferensiyasini avtomatik aniqlash tizimi uchun lingvistik baza shakllantirilgan va № BGU 1914 raqami bilan ro‘yxatdan o‘tkazilgan. Maqolada ushbu jarayonda amalga oshirilgan ishlar ketma-ketligi batafsil yoritib beriladi.

O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasini shakllantirish

O‘zbek tilidagi matnlar koreferensiyasini yaratishda quyidagi bosqichlar amalga oshirilishi zarur:

1. Keng ko‘lamda koreferent birliklar ishtirok etgan o‘zbekcha matn fragmentlarini to‘plash.

Ushbu bosqichda **matnni tanlash mezonlari, uning hajmi va tadqiqot manbalarini** aniq belgilash maqsadga muvofiqdir.

a) Ma’lumki, koreferensiya istalgan matnda uchraydigan hodisa emas. Matnni tanlash mezonlari orqali ma’lumotlar bazasi uchun qanday xususiyatga ega bo‘lgan matnlar zarurligi oydinlashadi. Biz lingvistik bazani yaratishda amal qilgan matnni tanlash mezonlari 1-jadvalda havola qilinadi:

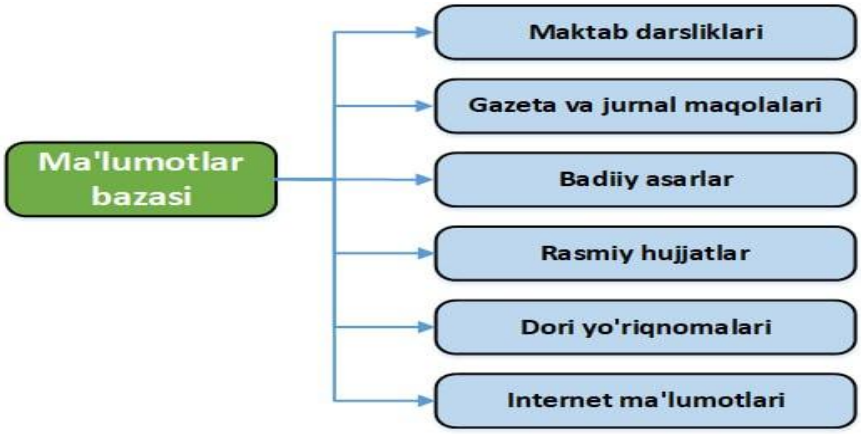
1-jadval. Ma’lumotlar bazasi uchun matnni tanlash mezonlari

Mezon	Tavsif
Turli nutq uslubiga oid matnlar	Badiiy, ilmiy-ommabop, publitsistik, so‘zlashuv va rasmiy uslubdagi matnlar qamrab olinishi maqsadga muvofiq.
Referentlar xilma-xilligi	Matnda koreferentlikni hosil qiluvchi turli leksik birliklar: olmoshlar, sinonimlar, murojaat birliklari, atoqli ot va otli birikmalar ishtiroki
Murakkab sintaktik tuzilmalar	Matnda katafora, anafora hodisalarining mavjudligi
O‘zaro bog‘liqlik	Matn qismlarining mazmunan yaxlit va izchil bo‘lishi muhim.

b) Matnlarni unda ifodalangan axborotning hajm belgisiga ko‘ra, minimal va maksimal matn tiplariga ajratish mumkin. Shuningdek, ma’lum bir matn tarkibidagi murakkab sintaktik butunlikka nisbatan mikromatn, yaxlit matnga nisbatan esa makromatn atamasi qo‘llaniladi [8]. Ayrim adabiyotlarda matn hajm jihatidan uch turga bo‘linadi: kichik, o‘rta va katta hajmli matnlar [9]. Ma’lumotlar bazasining mashinalar uchun xizmat qilishini va matn tasniflarini hisobga olgan holda maxsus koreferensiya korpusini tuzishda minimal, kichik va o‘rta hajmli matnlarga tayanish tavsiya qilinadi.

c) Ma’lumotlar bazasini tuzishda matnlarning qaysi manbaga oid ekanligini ham inobatga olish zarur. Koreferentlikning turli ko‘rinishlarini [10] aks ettirgan matnlarni yig‘ish turli shakldagi manbalarga murojaat qilishni talab qiladi. Mavjud koreferensiya korpuslari tahlili shuni ko‘rsatadiki, koreferensiyani avtomatik aniqlash tizimining ma’lumotlar bazasi uchun manbalardan foydalanish ixtiyoriydir. Ya’ni foydalaniladigan adabiyotlar ro‘yxati, ularning muayyan sohaga oid yoki sohalararo ekanligi, janri bo‘yicha cheklov o‘rnatilmagan. Jumladan, MUC

korpusida [11] “Wall Street” jurnalining 318 ta annotatsiyalangan maqolasi mavjud bo‘lsa, GUM korpusidan [12] muloqot, ta’lim va yangiliklarga oid matnlar o‘rin olgan. WikiCoref korpusi esa asosan 30 ta annotatsiyalangan Vikipediya maqolalaridan iborat [13]. O‘zbek tilidagi matnlar koreferensiyasini avtomatik aniqlash tizimining lingvistik ta’minoti quyidagi manbalar asosida shakllantirildi (1-rasm):



1-rasm. O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasi tarkibi

2. Ma’lumotlar bazasini annotatsiyalash.

Annotatsiyalash deganda korpusga izohlovchi lingvistik ma’lumotlarni qo‘shish [14] nazarda tutiladi. Koreferensiya korpuslarini tuzishda matnlar ichidagi referent birliklar va ularning koreferentlarini qo‘lda yoki yarim avtomatik usulda belgilash, ya’ni annotatsiyalash zarur. Bu jarayonda quyidagi maxsus terminlardan foydalaniladi (2-jadval):

2-jadval. Koreferensiya hodisasining asosiy tushunchalari

Termin	Izoh
Eslatma	Koreferentlikni hosil qiluvchi birlik
Antisedent	Asosiy referent; unga ishora qiluvchi birlikdan oldinda joylashadi.
Anafora	Asosiy referentdan keyin kelib, unga ishora qiluvchi birlik
Postsedent	Asosiy referent; unga ishora qiluvchi birlikdan keyin joylashadi.
Katafora	Asosiy referentdan oldin kelib, unga ishora qiluvchi birlik
Klaster	Bitta shaxs/narsa-hodisani anglatuvchi eslatmalar guruhi
Singleton	Koreferenti mavjud bo‘lmagan referent

Avvalo, koreferensiya korpusini annotatsiyalashda har bir tilning o‘z xususiyatlari asosida qoidalar tizimi ishlab chiqiladi va korpus shu qoidalar yordamida teglanadi. O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasini annotatsiyalashda quyidagi qoidalar va ko‘rsatmalarga rioya qilindi:

1. Faqat ot, otli birikma, harakat nomi va olmosh o‘rtasida yuzaga kelgan koreferentlik hodisasi inobatga olinsin. Boshqa holatlar e’tibordan chetda qoldirilsin:

a) Antisedent/postsedent gapga teng kelgan holatlar: *Yaxshi odatingiz **shuki,** barvaqt turasiz.*

b) To‘pdan ajratilgan shaxs yoki narsani ifodalovchi belgilash olmoshiga ishoralar: ***Har kim o‘zi** istagan axborotni izlash, olish va uni tarqatish huquqiga ega; **Har kim o‘z** og‘zining qorovuli bo‘lsa, **uning** nafasi hech qachon bo‘g‘ilmaydi.*

d) Noaniqlikni ifodalagan referentlar o‘rtasidagi munosabat: ***Kim** izlansa, **u**, albatta, maqsadiga yetadi.*

e) Son so‘z turkumiga ishoralar: *Kimyoviy bog‘ – **ikki** yoki **undan** ortiq atomlarning o‘zaro ta’sirlashuvi.*

f) Anafora hodisasining pronominal anaforadan boshqa turlari.

2. Koreferentlar bitta shaxs/narsa-hodisani bildirgan taqdirda bitta klasterga joylashtiriladi:

Matn: *Kitob bilim manbayi, shuning uchun biz uni sevamiz.*

Koreferent juftlik: *[Kitob, uni]*

3. Ega va ot-kesim munosabatini koreferentlik sifatida belgilamang. Bu munozarali holat bo‘lib, ma’lumotlar bazasida koreferentlar juftligi safiga kiritilmadi: ***Toshkent** – **O‘zbekistonning poytaxti.***

4. Izohlovchilarni koreferent juftligiga kiritish-kiritmaslik yuzasidan ikki xil qarash bor. Bu faqat **bitta gap tarkibida ro‘y bergan koreferentlik** hodisasiga taalluqlidir:

***Bobokalonimiz Amir Temur** haqida so‘z borganda uning buyuk davlatchilik asoschisi bo‘lganligi ko‘p bor ta’kidlanadi.*

***O‘g‘lim, qo‘zichog‘im,** orom olyapti.*

Ayrim tadqiqotlarda izohlovchi-izohlanmish munosabatining anaforik va kataforik holatlari koreferentlik sifatida aniqlangan bo‘lsa, ayrim olimlar bu fikrni rad etadilar. Biz Stenford CoreNLP platformasidagi [15] o‘rganishlarimiz natijasida izohlovchilarni izohlanmish bilan birgalikda **span** sifatida bitta birlik deb qabul qildik. Span – yaxlit ma’nosi bir nechta tokenlar orqali anglashiladigan birlik, otli birikma [16].

5. O‘z-o‘ziga havolalar koreferent guruhiga kiritilsin: *Seton-Tompson **tabiatni** ardoqladi, **tabiat** ham uni ardoqladi...*

Yuqori aniqlikka erishish maqsadida o‘zbek tilidagi matnlar koreferensiyasi ma’lumotlar bazasi keltirilgan qoidalar asosida murakkab jarayon hisoblangan usulda – qo‘lda teglab chiqildi. Namuna 3-jadvalda aks ettiriladi:

3-jadval. O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasini teglash

Matn	Koreferent juftligi	Klaster	Eslatma tegi	Izoh
Farg‘ona fojiasining sabablari, uni harakatga keltirgan kuchlar kim edi?	[Farg‘ona, uni]	1	[NER, olmosh]	[Antisedent, anafora]
Ona – Quyosh, uning mehri tafti hech qachon sovimaydi.	[Ona, uning]	1	[ot, olmosh]	[Antisedent, anafora]
Bobur she’riyatining o‘ziga xos xususiyati unda shoir shaxsiyatining bo‘rtib turishidir.	[Bobur she’riyatining, unda] [Bobur, shoir]	1 2	[NP, olmosh] [NER, ot]	[Antisedent, anafora] [Antisedent, anafora]
Seton-Tompson tabiatni ardoqladi, tabiat ham uni ardoqladi...	[Seton-Tompson, uni] [tabiatni, tabiat]	1 2	[NER, olmosh] [ot, ot]	[Antisedent, anafora] [Antisedent, anafora]
Ana u! Hamma narsa birdan unutildi: u yerda, oldinda, Issiqko‘lning ko‘m-ko‘k sathida oppoq kema paydo bo‘ldi.	[kema, u] [Issiqko‘lning ko‘m-ko‘k sathida, u yerda]	1 2	[ot, olmosh] [NP, ravish]	[postsedent, katafora] [postsedent, katafora]
“Nima deding, Fitna? Nega dilgir bo‘libdilar?” – so‘radi so‘fi Qurvonbibidan.	[Qurvonbibidan, Fitna]	1	[NER, NER]	[postsedent, katafora]

Ma’lumotlarni SQL Serverda saqlash

SQL Server – ma’lumotlarni saqlaydigan va ularga so‘rov qilish imkonini beruvchi ma’lumotlar bazasi mexanizmi [17]. Ushbu tizimda matnlar va koreferent birliklar alohida-alohida o‘zaro bog‘langan jadvallarda saqlanadi va so‘rov orqali tegishli matndagi koreferent bo‘lgan barcha birliklar taqdim qilinadi.

1. SQL Serverda matnlar uchun jadval yaratish.

```
CREATE TABLE documents (
    id INT PRIMARY KEY IDENTITY (1,1),
    title TEXT,   -- Matn sarlavhasi (ixtiyoriy)
    content TEXT  -- Asl matn
);
```

Yuqoridagi kod asosida to‘plangan matnlar ketma-ket tizimga kiritiladi va matnlardan iborat jadval yaratiladi:

id	title	content
1	1A	Har kim o'zi istagan axborotni izlash, olish va uni tarqatish huquqiga ega.
2	2A	Asarni tushunish uchun, avvalo, uning mazmuni bilan tanishib chiqish darkor.
3	3A	Bobur she'riyatining o'ziga xos xususiyati unda shoir shaxsiyatining bo'rtib turis...
4	4A	Kelajak uni kashf etganlarga nasib qiladi.
5	5A	Tuxumning eng qimmatli qismi uning sarig'idir.
6	6A	Qadimda rossiyaliklar marvaridni tozalash uchun avval uni tovuqqa yutdirishgan.
7	7A	Buyuk shaxslar go'zal nutqlar vositasida emas, o'z mehnatlari va ularning natijal...
8	8A	Biz muammoni yaratgan fikrlash tarzimiz bilan uni hal eta olmaymiz.
9	9A	Ommaga ergashgan odam undan o'zib ketolmaydi.
10	10A	Lava otilishidan keyin uning tarqalish tezligi itning yugurish tezligiga yaqindir.

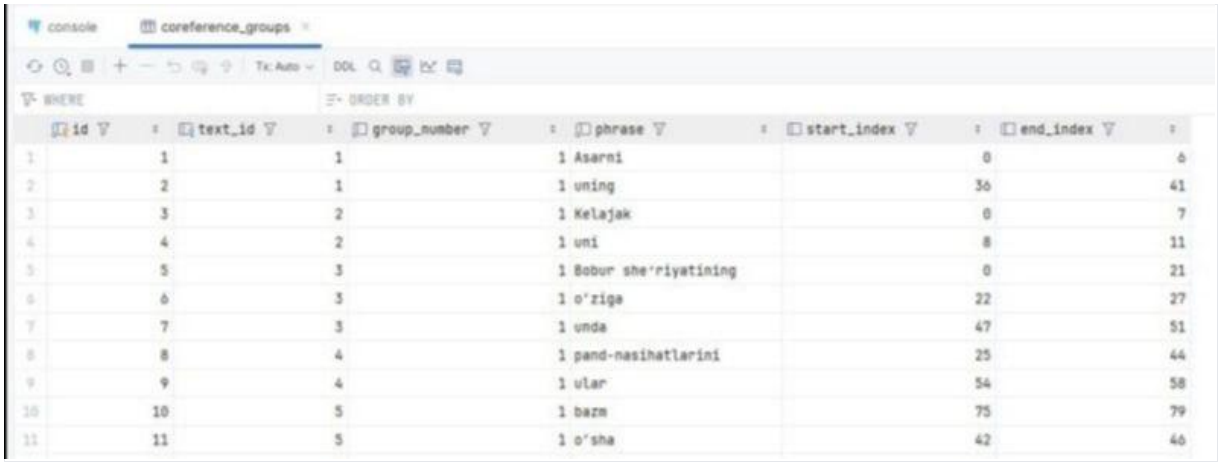
2-rasm. SQL Serverda matn ma’lumotlarining aks ettirilishi

2. Koreferent guruhlar jadvalini hosil qilish.

Koreferent guruhlar jadvalini tuzishda quyidagi kod qo‘llanildi:

```
CREATE TABLE coreference_groups (
    id SERIAL PRIMARY KEY,
    text_id INTEGER REFERENCES texts(id) ON DELETE CASCADE,
    group_number INTEGER NOT NULL,
    phrase TEXT NOT NULL,
    start_index INTEGER,
    end_index INTEGER
);
```

Koreferent klasterlari jadvalida har bir referent va u bilan bog‘liq koreferent ifodalar jadvalda aniq koordinatalar (boshlanish va tugash indekslari) bilan ko‘rsatiladi. Jadval namunasi bilan quyida tanishish mumkin (3-rasm):



	id	text_id	group_number	phrase	start_index	end_index
1	1	1	1	Asarni	0	6
2	2	1	1	uning	36	41
3	3	2	1	Kelajak	0	7
4	4	2	1	uni	8	11
5	5	3	1	Bobur she'riyatining	0	21
6	6	3	1	o'ziga	22	27
7	7	3	1	unda	47	51
8	8	4	1	pand-nasihatlarini	25	44
9	9	4	1	ular	54	58
10	10	5	1	bazm	75	79
11	11	5	1	o'sha	42	46

3-rasm. SQL Serverda koreferent klasterlari jadvali

3. SQL Serverga so‘rov yuborish va ma’lumot olish.

SQL Serverda so‘rov berish orqali o‘zaro bog‘langan jadvallardagi umumiy ID ga ega bo‘lgan ma’lumotlar birlashtiriladi. SQL Serverda so‘rov yuborish uchun kod:

```
SELECT
   t.id AS text_id,
   t.content AS text_content,
   cg.group_number,
   cg.phrase,
   cg.start_index,
   cg.end_index
FROM texts t
LEFT JOIN coreference_groups cg ON t.id = cg.text_id
WHERE t.id = 1;
```

So‘rov yuborilgach, matn va unga tegishli koreferent birliklar natija sifatida berilishi kerak. 4-rasmda *“Asarni tushunish uchun, avvalo, uning mazmuni bilan tanishib chiqish darkor”* gapi misolida so‘rov natijalari tasvirlangan:


Output Result 5

	text_id	text_content	group_number	phrase	start_index	end_index
1	1	Asarni tushunish uchun, avvalo, uning mazmuni bilan tanishi…	1	Asarni	0	6
2	1	Asarni tushunish uchun, avvalo, uning mazmuni bilan tanishi…	1	uning	36	41

4-rasm. SQL Serverda so‘rov natijalari

XULOSA

NLP da tabiiy tilga ishlov berishning so‘nggi bosqichi koreferensiyani avtomatik aniqlashdir. O‘zbek tilida bunday tizimni, uning lingvistik ta’minotini shakllantirishdagi kuzatishlar keying ishlarimiz uchun quyidagi xulosalarni berdi.

1. O‘zbek tilidagi matnlar koreferensiyasini aniqlovchi tizimning lingvistik ta’minoti hajmini yanada kengaytirish lozim. Chunki ma’lumot bazasining hajmi modelning aniqlik ko‘rsatkichiga ta’sir etadi.

2. O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasida gapga ishoralarni ham annotatsiyash foydalanuvchilar uchun yanada qulaylikni ta’minlaydi.

3. Lingvistik ta’minotni ma’lumotlarni boshqarish tizimida saqlash uchun aniq formatni (JSON, CoNLL) tanlash zarur.

Foydalanilgan adabiyotlar:

1. Büyüktekin F., Özge U. A coreference corpus of Turkish situated dialogs / Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024), August 15, 2024. – pages 42–52.

2. Dobrovolskii V., Michurina M., Ivoylova A. RuCoCo: a new Russian corpus with coreference annotation. / https://doi.org/10.48550/arXiv.2206.04925.

3. Poesio M., Camilleri M., Garcia1 P.C., Yul J., Müller M. The ARRAU 3.0. / CorpusProceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), March 21, 2024. – pages 127–138.

4. https://uznatcorpara.uz/

5. http://alishernavoicorpus.uz/uz/about

6. https://uzschoolcorpara.uz/

7. https://uzbekcorpus.uz/newIndex

8. Yo‘ldoshev M. Badiiy matnning lisoniy tahlili. / O‘quv qo‘llanma. – Toshkent, 2007. – 150 b.

9. Qilichev E. Matnning lingvistik tahlili. – Buxoro, 2000. – 36 b.

10. Abdisalomova Sh. NLPda koreferensiyani hal etish vazifasining o‘rni. / Guliston davlat universiteti axborotnomasi, Gumanitar-ijtimoiy fanlar seriyasi, № 4, 2024. – 166-169-betlar.

11. Hirshman, L. and Chinchor, N. MUC-7 coreference task definition. version 3.0. / In Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998. – pages 127-138.

12. https://gucorpling.org//gum/

13. Ghaddar A., Langlais. WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles. https://aclanthology.org/L16-1021.pdf; https://github.com/victoriasovereigne/WikiCoref-CoNLL

14. Primova M. Til korpuslarida matnlarni annotatsiyalash: afzallik va kamchiliklari. / “O‘zbekiston: Til va madaniyat” jurnali, Vol. 4 (6), 2023. – 6-18-betlar.

15. https://corenlp.run/

16. Elova D. O‘zbek tili matnlarida koreferensiyani hal qilish bosqichlari. / “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” Xalqaro ilmiy-amaliy konferensiya materiallari, Vol. 1, № 01 (2023). – 156-161-betlar.

17. https://docs.dot-net.uz/database/ms-sql-server