

**O‘ZBEKISTON RESPUBLIKASI OLIY TA’LIM, FAN VA
INNOVATSIYALAR VAZIRLIGI
ALISHER NAVOIY NOMIDAGI
TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI
UNIVERSITETI**

**“KOMPYUTER LINGVISTIKASI: MUAMMOLAR,
YECHIM VA ISTIQBOLLAR”**

mavzusidagi III an’anaviy xalqaro ilmiy-amaliy konferensiya

(2023-yil 28-aprel)

Toshkent – 2023



Kompyuter lingvistikasi: muammolar, yechim va istiqbollar / Xalqaro an'anaviy ilmiy-amaliy konferensiya to'plami. Elektron nashr / ebook. – Toshkent: ToshDO'TAU, 28.04.2023. – 380b.

Mazkur to'plamdan kompyuter lingvistikasi sohasiga doir tadqiqot natijalari, filologik dasturiy ta'minotlar, ularning imkoniyatlari va afzalliklari, tabiiy tilni qayta ishlashga oid masalalar, semantik tizim bazalari, til va adabiy korpuslar, lingvistik ta'minot manbalari, mashina tarjimasida lingvistik muammolar, lingvodidaktik dasturlar va platformalar, terminologik lug'atlarni yaratish masalalariga oid materiallar joy olgan.

Ushbu ilmiy to'plamdan kompyuter lingvistikasi sohasi vakillari, tadqiqotchilar, mustaqil izlanuvchilar, magistrantlar va kelajak sohalariga qiziquvchi talabalar foydalanishlari mumkin.

Mas'ul muharrir:

Sirojiddinov Sh.S. Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti rektori, f.f.d., professor

Taqrizchilar:

Normurodova N. Toshkent davlat o'zbek tili va adabiyoti universiteti, dotsent, f.f.d.

Eshref Adali. Istanbul texnika universiteti, professor, t.f.d.

Karimov S. Samarqand davlat universiteti, professor, f.f.d.

TASHKILIY QO'MITA

Sirojiddinov Shuhrat Samariddinovich (Toshkent davlat o'zbek tili va adabiyoti universiteti)

Jamoliddinova Odinaxon Rustamovna (Toshkent davlat o'zbek tili va adabiyoti universiteti)

Normurodova Nozliya Zarilovna (Toshkent davlat o'zbek tili va adabiyoti universiteti)

Eshref Adali (Turkiya, Istanbul texnika universiteti)

Zaxarov Viktor Pavlovich (Sankt-Peterburg davlat universiteti)

Suleymanov Djavdet Shevketovich (Tatariston Fanlar akademiyasi Amaliy semiotika instituti)

Gilmullin Rinat Abrekovich (Tatariston Fanlar akademiyasi Amaliy semiotika instituti)

Gatiatullin Ayrat Rafizovich (Tatariston Fanlar akademiyasi Amaliy semiotika instituti)

Mengliyev Baxtiyor Rajabovich (Toshkent davlat o'zbek tili va adabiyoti universiteti)

Elov Botir Boltayevich (Toshkent davlat o'zbek tili va adabiyoti universiteti)

Jo'raqo'ziyev Nodir Imomqo'ziyevich (Toshkent davlat o'zbek tili va adabiyoti universiteti)

Murat Orxun (Istanbul Bilgi universiteti)

Hamroyeva Shahlo Mirdjonovna (Toshkent davlat o'zbek tili va adabiyoti universiteti)

Abdullayeva Oqila Xolmo'minovna (Toshkent davlat o'zbek tili va adabiyoti universiteti)

TAHRIR HAY'ATI

Normurodova Nozliya Zarilovna

Eshref Adali

Mengliyev Baxtiyor Rajabovich

Elov Botir Boltayevich

Hamroyeva Shahlo Mirdjonovna

Abdullayeva Oqila Xolmo'minovna

Xusainova Zilola Yuldashevna

Xudayberganov Nizomaddin Uktambay o'g'li

KOMPYUTER LINGVISTIKASI – MILLIY TILSHUNOSLIGIMIZ ISTIQBOLI

Assalomu alaykum, hurmatli konferensiya ishtirokchilari, qadrlı mehmonlar!
Mamlakatimizda amalga oshirilayotgan keng ko'lamli islohotlar jarayonida davlat tilining hayotimizdagi o'rni va nufuzi tobora oshib bormoqda. O'zbek tili siyosiy, ijtimoiy-iqtisodiy, ma'naviy-ma'rifiy jabhalarda faol qo'llanib, xalqaro minbarlarda baralla yangramoqda. Xorijiy mamlakatlarda tilimizga e'tibor va uni o'rganishga qiziqish kuchaymoqda.

O'zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi “O'zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to'g'risida”gi PF-5850-sonli Farmoni, 2020-yil 20-oktabrdagi “Mamlakatimizda o'zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to'g'risida”gi PF-6084-sonli Farmonlari bu vazifalarni amalga oshirishni yanada jadallashtirdi.

Mamlakatimiz ijtimoiy-siyosiy hayotining barcha sohalarida davlat tili imkoniyatlaridan to'liq va to'g'ri foydalanishga erishish; ta'lim tashkilotlarida davlat tilini o'qitish tizimini yanada takomillashtirish, uning ilm-fan tili sifatidagi nufuzini oshirish; davlat tilining sofligini saqlash, uni boyitib borish va aholining nutq madaniyatini oshirish; davlat tilining zamonaviy axborot texnologiyalari va kommunikatsiyalariga faol integratsiyalashuvini ta'minlash; davlat tilining xalqaro miqyosdagi o'rni va nufuzini oshirish, bu borada xorijiy hamkorlik aloqalarini rivojlantirish bo'yicha bosqichma-bosqich tizimli ishlar amalga oshirilmoqda.

Davlat tilining zamonaviy axborot texnologiyalari va kommunikatsiyalariga integratsiyalashuvini ta'minlash sohasida 2025-yilga qadar davlat tilida sohaviy ish yuritish hujjatlarining yagona elektron namunalarini ishlab chiqish va ulardan foydalanish uchun 14 ta onlayn dastur ishlab chiqish, 2030-yilgacha esa ushbu ko'rsatkichni 25 taga yetkazish; dasturiy mahsulotlarning o'zbekcha ilovalari va elektron lug'at dasturlarini yaratish; xorijliklar uchun o'zbek tilini o'rgatuvchi dasturlar yaratish vazifalari belgilab qo'yilgan.

Bugun universitetimizda tashkil etilgan “Kompyuter lingvistikasi: muammolar, yechim va istiqbollar” mavzusidagi xalqaro konferensiya ham davlat tilining axborot va kommunikatsiya texnologiyalari, xususan, Internet jahon axborot tarmog'ida munosib o'rin egallashini ta'minlash, o'zbek tilining kompyuter dasturlarini yaratish kabi vazifalar ijrosiga bag'ishlangan.

Konferensiyaning barcha ishtirokchilarini tabriklayman, konferensiya ishiga omad tilayman!

**Sirojiddinov Shuhrat Samariddinovich,
Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti
rektori, f.f.d, professor**

KOMPYUTER LINGVISTIKASI TADRIJI XUSUSIDA

Dunyo kompyuter lingvistikasida oxirgi yillarda milliy korpuslar, avtomatik tarjima platformalari, tezaurus, elektron lug'atlardan foydalanish, ularni ilmiy-nazariy jihatdan o'rganish harakati avj oldi. Shu jarayonda korpusni tuzishning dasturiy va lingvistik tamoyillarini asoslash, tabiiy tilni qayta ishlash (NLP), mashina tarjimasi, matnga qayta ishlov berish tizimlari, kompyuter lingvodidaktikasi, kompyuter leksikografiyasi, semantik tizimlar va lingvistik ontologiyalar ustida ishlash va korpus grammatikalarini yaratishga keng imkon yaratadi.

Mamlakatimizda global ilmiy-texnik islohotlar amalga oshirilayotgan bugungi kunda kompyuter lingvistikasi doirasida avtomatik tarjima, o'zbek tilini tushunish va qayta ishlash, uni boshqa tillar bilan chog'ishtirishda sun'iy intellektning rolini belgilashga kuchli ehtiyoj sezilmoqda. Shu ma'noda "...davlat tilining sofligini saqlash, uni boyitib borish va aholining nutq madaniyatini oshirish; davlat tilining zamonaviy axborot texnologiyalari va kommunikatsiyalariga faol integratsiyalashuvini ta'minlash" kabi ustuvor yo'nalishlar tahlil qilinayotgan til faktlarini tizimlashtirish, akademik maqsadlarda milliy va xorijiy til ko'nikmalarini egallash, tilni real tasvirlash, shuningdek, amaliy tilshunoslikda til qo'llanishi etalonlarini optimallashtirishga keng yo'l ochadi.

Kompyuter lingvistikasi istiqbolli ilmiy yo'nalish ekanligini inobatga olgan holda o'zbek tili milliy korpusini yaratish, lingvistik modellarni tuzish singari masalalarni zamonaviy ilmiy tamoyillar asosida tadqiq etish fanimiz oldida turgan dolzarb vazifa hisoblanadi. Shu nuqtayi nazardan o'zbek tilining formal grammatikasini yaratish, korpus lingvistikasi, tabiiy tilni qayta ishlash (NLP), mashina tarjimasi, matnga qayta ishlov berish tizimlari, kompyuter lingvodidaktikasi, kompyuter leksikografiyasi, semantik tizimlar va lingvistik ontologiyalar, kompyuter lingvistikasida terminologiya va terminografiya, kompyuter lingvistikasida sun'iy intellekt kabi sohalarni rivojlantirish nihoyatda muhim.

Bugun universitetimizda tashkil etilgan "Kompyuter lingvistikasi: muammolar, yechim va istiqbollar" mavzusidagi xalqaro konferensiya mazkur masalalarni mutaxassislar va yosh tadqiqotchilar bilan muhokama qilish, bu sohalarda qilingan ishlar bilan tanishishga imkon beradi.

Isajon Sulton,
Oliy ta'lim, fan va innovatsiyalar vaziri maslahatchisi



KOMPYUTER LINGVISTIKASI ISTIQBOLI YOSHLAR QO‘LIDA

O‘zbek tilining xalqimiz ijtimoiy hayotida va xalqaro miqyosdagi obro‘-e‘tiborini tubdan oshirish, unib-o‘tib kelayotgan yoshlarimizni vatanparvarlik, milliy an‘ana va qadriyatlarga sadoqat, ulug‘ ajdodlarimizning boy merosiga vorislik ruhida tarbiyalash, mamlakatimizda davlat tilini to‘laqonli joriy etishni ta‘minlash borasida qator ishlar amalga oshirilmoqda. Xususan, O‘zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to‘g‘risida”gi PF-5850-sonli Farmoni, 2020-yil 20-oktabrdagi “Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida”gi PF-6084-sonli Farmonlari bu vazifalarni amalga oshirishni yanada jadallashtirdi.

O‘zbekiston Respublikasi Prezidentining “2022-2023-yillarda axborot-kommunikatsiya texnologiyalari sohasini yangi bosqichga olib chiqish chora-tadbirlari to‘g‘risida”gi PQ-357-son qaroriga asosan axborot-kommunikatsiya texnologiyalari sohasida IT-ta‘lim xizmatlarini ko‘rsatish, dasturiy mahsulotlar ishlab chiqish hamda eksportga yo‘naltirish bo‘yicha start up loyihalarni qo‘llab-quvvatlash uchun Investitsiyalar va tashqi savdo vazirligi huzuridagi Yosh tadbirkorlarni qo‘llab-quvvatlash jamg‘armasi hisobidan Innovatsion rivojlanish vazirligi tomonidan o‘tkaziladigan tanlov asosida 1 milliard so‘mgacha bo‘lgan grantlar ajratish tartibi joriy qilindi.

Jahon zamonaviy tilshunosligida kompyuter fanlari va lingvistik ta‘limotlar kesishuvida yuzaga kelgan kompyuter lingvistikasiga e‘tibor kuchayib bormoqda. Lisoniy bilimlarni amalda qo‘llashning referativ usullaridan elektron formatlariga o‘tish parallel korpuslar yaratish ko‘lamini va ularning lingvistik-dasturiy funksionalini kengaytirish zaruriyatini belgilaydi. Bu, o‘z navbatida, vaqt tejamini ta‘minlagan holda til tizimi va nutqiy-fikriy faoliyatning tarkibiy qismi bo‘lgan matn va uning turli ko‘rinishlarini lingvistik haqiqat rakursida idrok etish, tarjima matnlarini chog‘ishtirish orqali tarjimon strategiyalari asosini tashkil etuvchi tamoyillarni belgilash, ularning turli kontekstlarda voqelanadigan diskursiv va krosslingvistik qiyofasini ilmiy baholashga xizmat qiladi.

Bugun universitetimizda tashkil etilgan “Kompyuter lingvistikasi: muammolar, yechim va istiqbollar” mavzusidagi xalqaro konferensiya mazkur masalalar doirasida tadqiqot olib borayotgan olimlar, yosh mutaxassislar uchun fikr almashish maydoni bo‘lib xizmat qilishiga ishonaman.

Jamoldinova Odinaxon Rasulovna,
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti
ilmiy ishlar va innovatsiyalar bo‘yicha prorektori

YALPI YIG'ILISH MA'RUZALARI ALFABE (ABECE) VE BİLİŞİM

Eşref ADALI

İstanbul Teknik Üniversitesi
Bilgisayar ve Bilişim Fakültesi
Türkiye
adali@itu.edu.tr

Öz Bir dilin en önemli varlıklarından birisi abecesidir. İnsanlar arası iletişimin en önemli bileşeni olan abece, bilişim dünyasındaki gelişmelerin sonucu daha önemli olmuştur. Günümüzde bilişim olanaklarından yararlanmak isteyen bir ulusun abecesinin, ölçünlü bilişim abecesine uyumlu olması gerekir. Uyumlu abecesi olmayanlar, bilişimin sağladığı olanaklardan yararlanmakta zorlanırlar. Bu bildiride abecenin önemi ve yeni abece geliştirmeye çalışan Özbekler için dikkat etmeleri gereken konular anlatılmıştır.

Anahtar sözcükler: *Alfabe, abece, IPA, Sesçil abece, tuş takımı, UTF*

Abstract. One of the most important assets of a language is its alphabet. The most important component of interpersonal communication, alphabet, has become more important as a result of developments in the world of informatics. Today, the alphabet of a nation that wants to benefit from the information technologies should be compatible with the standard information alphabet. Those who do not have a harmonious alphabet have difficulty in taking advantage of the opportunities provided by informatics. In this paper, the importance of the alphabet and the issues that should be considered by the Uzbeks who are trying to develop a new alphabet are explained.

Keywords: *Alphabet, alphabet, IPA, Phonetic alphabet, keypad, UTF.*

Giriş

Dilin en temel bileşeni, ses ve bu seslere karşılık düşen harflerdir. Ses üretiminde ciğerimiz körük, gırtlak ve dilimiz ses üretim aygıtı olarak çalışır. Dillerin yazıya dönüşebilmesi için bir abecenin geliştirilmesi gerekmektedir. Bugünkü anlamda ilk abecenin Sami ırkından olan denizci Fenikeliler (kendi dillerine göre Kenaniler) tarafından geliştirildiği tarihî belgelerde yazılmaktadır. MÖ 1000 yılında kullanılmaya başlandığı sanılan bu abecede 22 harf bulunmaktadır ve bunların hepsi ünsüz harflerdir.

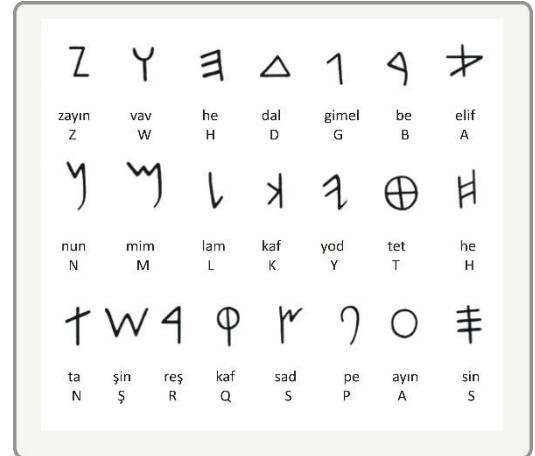
Diğer bir deyişle Fenike abecesinde ünlü harf yoktur. Şekil-1'de Fenike abecesi gösterilmiştir. Fenike abecesi, denizci ulus tarafından Akdeniz'e kıyısı olan uluslara aktarılmıştır. Bu aktarmalar sonrasında Arap, Yunan ve daha sonra Latin abeceleri oluşmuştur. II-IV. yüzyıllar arasında oluşan Arap abecesinin harflerinin adları Fenike harflerinin adları ile aynıdır. Ayrıca Fenike yazısı gibi Arap yazısı da sağdan sola doğru yazılır. Musevilerin kullandıkları abece de Fenike abecesinden üretilmiştir ve harf adları aynıdır, yazma yönü sağdan sola doğrudur. Fenike abecesinde büyük ve küçük harf ayrımı yoktur [1,2,3]. Bu nedenle Arap abecesinde de büyük ve küçük harf ayrımı yoktur ancak harfin biçimi sözcük başında, ortasında ve sonunda değişir.

Herodot yazı ve uygarlığı Yunanlara öğreten ulusun Fenikeliler olduğunu yazmaktadır. Yunanlar Fenike abecesinin şekillerini ve yazma yönünü değiştirmişler; ayrıca Yunancada karşılığı olmayan ünsüzleri, Yunancanın ünlüleri için kullanmaya başlamışlardır. Böylece elif'ten A, he'den E, yod'dan İ, ayin'dan O ve vaw'dan U ünlülerini türetmişlerdir [4]. Bu çalışmaların MÖ IX. yüzyılda yapıldığını tarihî kayıtlardan öğreniyoruz. Yunan abecesi zaman içinde değişime uğramış, bu sırada küçük ve büyük harf kavramı oluşmuş ve yazma yönü soldan sağa değiştirilmiştir. Eski Latin abecesinin MÖ VI. yüzyılda Etrüks abecesinden türetildiği söylenmektedir. Bu abeceye Yunan abecesinden Y ve Z harfleri alınmıştır. Daha sonra diğer abecelerden harfler eklenerek bugünkü biçimini kazanmıştır [1].

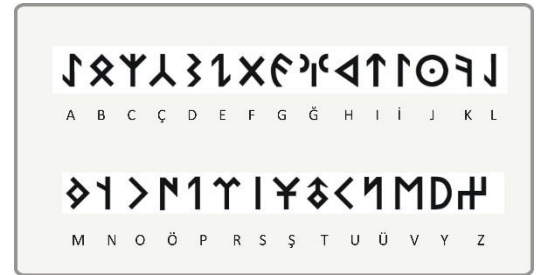
Orhun anıtlarında Göktürk abecesi kullanılmıştır. Bu abecede 26 ünsüz, 4 ünlü ve 8 birleşik harf (bunlara seslem diyebiliriz) bulunmaktadır. Harfler ayırık olarak sağdan sola doğru yazılır. Sözcükler, aralarına " : " konarak ayrılır. Harflerin biçimleri Runik abecesine benzediği için Göktürk abecesine Runik abece diyenler vardır. Runik abecesinde simgeler bir harf veya bir sesleme karşı düşmektedir. Örneğin "ko", "ku" seslemi için aşağı doğru ok biçimine bir simge, a ve e sesleri için iki ucu çengelli simgeler kullanılmıştır. Göktürk abecesinin, Fenike abecesinden türetilmiş olan Arami abecesinden türetildiği de söylenmektedir. Göktürk abecesi Şekil-2'de gösterilmiştir.

Rusya ve komşu ülkelerinde kullanılan Kiril abecesi IX. yüzyılda Kiril ve Methodius kardeşler tarafından geliştirilmiştir.

Yazılı metinler insanlar arasındaki iletişimde kullanılan en güçlü aktarım araçlarından biridir. Bu yüzden yönettikleri toplumun diğer toplumlar ile iletişimini

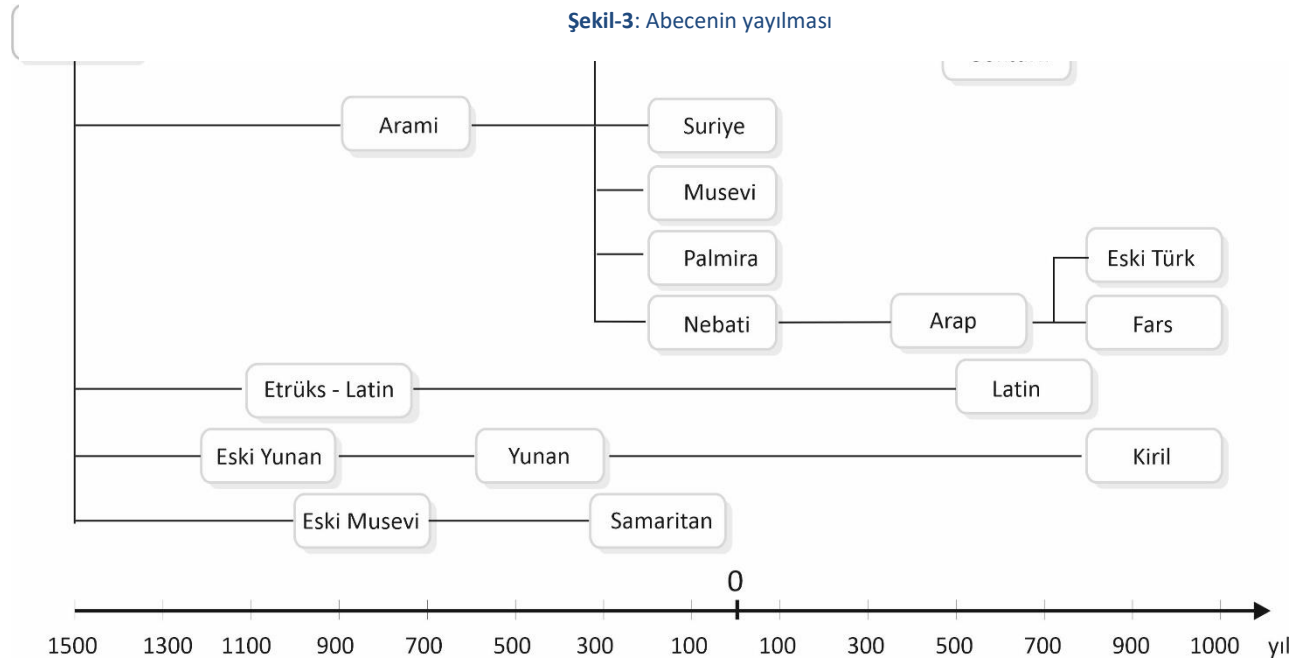


Şekil-1: Fenike abecesi



Şekil-2: Göktürk abecesi

engellemek isteyen bazı yöneticiler abece değişikliğine gitmişlerdir. Şekil-3'te Fenike abecesinden türetilen abeceler gösterilmiştir:



Anadolu'ya gelen Türkler, bu sırada kabul ettikleri İslamiyet'in etkisiyle Arap abecesini kullanmaya başlamışlardır. Arap abecesinde olmayan bazı Türkçe sesleri (ç, p gibi) Arap abecesine eklemişlerdir. Arap abecesinin üç ünlüsü, ünlü zengini olan Türkçe için yeterli olmamıştır. Arap abecesinin bu eksikliği yazıda önemli yanlışlara neden olmuştur. Örneğin gel, kel ve gül sözcükleri aynı harflerle yazılmak zorunda kalmıştır. Her üç sözcük için K ve L harfleri kullanılmıştır. K yerine göre K yerine göre G sesi için kullanılmıştır. Yazıda ünlü olmadığı için K ve L arasına her ünlü yakıştırılabilmekteydi diğer bir deyişle gel, gül, kel aynı biçimde yazılıyordu.

2 Sesler ve Abece

Temel olarak dillerdeki sesler ünlü ve ünsüz sesler olarak iki kümeye ayrılırlar. Ünlüler ciğerden gelen hava akımının ses yolunda ve ağızda bir engel ile karşılaşmadan çıkarılan seslerdir. Bu nedenle ünlülere özgür sesler de diyebiliriz. Ciğerden gelen hava akımının ses telleri süzgecinden geçerek oluşturduğu seslere ünsüz sesler denir. Ünsüzlere engelli sesler de denilmektedir.

Bir dilde anlam ayırma görevini yapan asıl seslere fonem (phonème) veya sesbirim denilmektedir. Türkçede her sese karşılık bir harf bulunduğundan Türkçenin her harfi bir sesbirimdir. Bir solukta çıkan bir tek ya da birleşik sese hece ya da seslem denir. İnsanlar düşünce, duygu ve konuşmalarını kalıcı kılmak için yazıyı icat etmişlerdir. İlk yazı biçimi Sümerlerin bulduğu çivi yazısı ve daha sonra Mısırlıların kullandığı şekil yazısıdır (Hiyeroglif). Fenikeliler bir adım daha ileri gidip her ses için bir simge daha doğrusu bir harf kullanmayı başarmışlardır. Fenikeliler abeceyi geliştirirken, bir

sese qarşı bir harf ilkesini benimsemişlerdir. Bu abecedan yararlanarak kendi dilleri için abece oluşturan uluslar dillerine özgü seslerin karşılığı harfleri bulamayınca birleşik harfler kullanmaya başlamışlardır. Örneğin Fransızlar ş sesi için CH kullanmışlardır. Bazı uluslar, kendilerine öğretilen dili basitleştirmeye çalışmışlar ve çok sayıda sözcüğü başka dillerden almışlardır. Bunun sonucu olarak İngilizcedeki gibi kuralsız bir yazım biçimi oluşmuştur.

İngilizcenin yazım kurallarındaki belirsizlikleri gidermek üzere Fransız dil bilimci Paul Passy'nin önderliğinde Fransız ve İngiliz dil öğretmenleri 1887 yılında Uluslararası Fonetik Derneğini kurarlar (International Phonetic Association) [4]. Amaçları Romic abece [5] ile önerilen, "her ses için bir harf kullanmak" kuralına uygun olarak İngilizce için bir abece geliştirmektir. Bu düşünce ile geliştirilen IPA abecesi, diğer bir deyişle sesçil abece daha sonra diğer diller için de uyarlanmıştır. Günümüzde Türk ulusları çeşitli abeceler kullanmaktadırlar. Bunun çeşitli nedenleri vardır. Türk dilleri sondan eklemeli dillerdir ve ünlü uyumu gibi ilginç özellikleri vardır. 7 Türk ulusunun dillerinde görülen ünlüler Çizelge-1'de görülmektedir. Çizelge-2'de 7 Türk ulusunun kullandığı abeceler IPA abecesi ile birlikte görülmektedir.

Çizelge-1: Türk dillerindeki ünlüler

Dil	Ünlü sayısı	Ünlüler
Türkçe	8	a, e, ı, i, o, ö, u, ü
Azerbaycan	9	a, e, é, ı, i, o, ö, u, ü
Türkmen	9	a, ä, e, ı, i, o, ö, u, ü
Özbek	10	a-a, ı-i, o-o, u-u , ı-i, o-o, u-u
Uygur	8	a, e, é, i, o, ö, u, ü
Kazak	9	a, e, a, ı, i, o, o, u, u
Kırgız	8	a, e, ı, i, o, o, u, u
Tatar	9	a, e, ä, ı, i, o, o, u, u

Çizelge-2: Türk Uluslarının Kullandıkları Abeceler

Türk		Türkmen		Azerbaycan		Özbek		Uygur			Kazak				Tatar				IPA
L	L	L	L	L	L	L	L	L	L	A	L	L	K	K	L	L	K	K	
a	a	a	a	a	a	a	a	a	a	r	a	a	r	r	a	a	r	r	
A	a	Ä	a	Ä	a	Ä	a	Ä	a	ı	Ä	a	Ä	a	Ä	a	Ä	a	/a/
		Ä	ä	E	e	Ä	ä	E	e	o/	Ä	Ä	Ə	ə	Ä	ä	Ə	ə	/æ/
B	b	B	b	B	b	B	b	B	b	б	B	B	Б	б	B	b	Б	б	/b/



C	c	J	j	C	c	J	j	J	j	ج					C	c	Д	д	/dʒ/
Ç	ç	Ç	ç	Ç	ç	Ch	ch	Ch	ch	چ	Ç	Ç	Ч	ч	Ç	ç	Ч	ч	/tʃ/
D	d	D	d	D	d	D	d	D	d	د	D	D	Д	д	D	d	Д	д	/d/
E	e	E	e	Ə	ə	E	e	Ė	ė	ع /	E	e	E	e	E	e	E	e	/e/, /æ/
				E	e						É	é	Э	э					/e/
											Y	y	Ë	ë					/jo/
F	f	F	F	F	f	F	f	F	f	ف	F	f	Ф	ф	F	f	Ф	ф	/f/
G	g	G	G	G	g	G	g	G	g	گ	G	g	Г	г	G	g	Г	г	/g/, /ʝ/
Ğ	ğ			Ğ	ğ	G	g	G _h	g _h	غ	Ğ	ğ	Ғ	ғ	Ğ	ğ	Г	г	/uɣ/
H	h	H	H	H	h	H	h	H	h	ه	H	h	Һ	һ	H	h	Һ	һ	/h/
				X	x	X	X	X	x	خ	X	x	X	x	X	x	X	x	/x/
I	ı	Y	y	I	ı						I	ı	Ы	ы	I	ı	Ы	ы	/u/
İ	i	İ	İ	İ	i	İ	İ	İ	i	ي /	İ	İ	И	и	İ	i	И	и	/i/
J	j	Ž	ž	J	j			Z _h	z _h	ژ	J	j	Ж	ж	J	j	Ж	ж	/ʒ/
K	k	K	k	K	k	K	K	K	k	ك	K	k	K	k	K	k	K	k	/k/, /c/
L	l	L	l	L	l	L	L	L	l	ل	L	l	Л	л	L	l	Л	л	/l/, /l/
M	m	M	m	M	m	M	M	M	m	م	M	m	М	м	M	m	М	м	/m/
N	n	N	n	N	n	N	N	N	n	ن	N	n	Н	н	N	n	Н	н	/n/, /ŋ/
		Ñ	ñ					N _g	n _g	نڭ	Ñ	ñ	Ң	ң	Ñ	ñ	Ң	ң	Nasal n
O	o	O	o	O	o	O	O	O	o	و /	O	o	O	o	O	o	O	o	/o/
Ö	ö	Ö	ö	Ö	ö	O	o	Ö	ö	و /	Ö	ö	Ө	ө	Ö	ö	Ө	ө	/œ/
P	p	P	p	P	p	P	P	P	p	پ	P	p	П	п	P	p	П	п	/p/
R	r	R	r	R	r	R	R	R	r	ر	R	r	Р	р	R	r	Р	р	/r/
				Q	q	Q	Q	Q	q	ق	Q	q	Қ	қ	Q	q	Қ	қ	/q/
S	s	S	s	S	s	S	S	S	s	س	S	s	С	с	S	s	С	с	/s/
Ş	ş	Ş	ş	Ş	ş	S _h	s _h	S _h	s _h	ش	Ş	ş	Ш	ш	Ş	ş	Ш	ш	/ʃ/
											Ş	şş	Щ	щ					/ʃtʃ/, / ʃ:/

T	t	T	t	T	t	T	T	T	t	ت	T	t	T	т	T	t	T	т	/t/
U	u	U	u	U	u	U	U	U	u	و	W	w	Y	y	W	w	Y	y	/u/
											U	u	Y	y	U	u	Y	y	/o/
Ü	ü	Ü	ü	Ü	ü			Ü	ü	و	Ü	ü	Y	y	Ü	ü	Y	y	/y/
V	v	W	w	V	v	V	V	W	w	و	V	v	B	b	W	w	B	b	/v/
Y	y	Y	y	Y	y	Y	Y	Y	y	ي	Y	y	Y	y	Y	y	Y	y	/j/
Z	z	Z	z	Z	z	Z	z	Z	z	ز	Z	z	З	з	Z	z	З	з	/z/
											Y	y	Ю	ю					/ju/, /jy/
											Y	y	Я	я					/ja/, / ja/

2 Abece ve Bilişim

Günümüz bilgi veya bilişim çağı olarak adlandırılıyor, bilgisayara hayatımızın her alanında yer buluyor. Bir anlamda bilgisayar özelliği gösteren cep telefonları hayatımızın bir parçası. Bilişimdeki gelişmeler, bilgisayarlı dil bilimi çalışmalarını da başlatmıştır. Doğal Dil İşleme (DDİ) olarak kısıltığımız bu yeni bilim alanında şu çalışmaları sürdürüyoruz:

- Ses bilimi
- Bütün bilgisi
- Biçim bilimi
- Söz dizimi
- Anlam bilimi
- Söylem bilgisi
- Edim bilimi

DDİ bilim dalının uygulamaları ise şöyle sıralanabilir:

- Metin seslendirme
- Konuşmayı yazıya dökme
- Yazım yanlışlarını giderme
- Bilgiye erişim
- Bilgi çıkarma
- Özetleme
- Anlama
- Soru yanıtlama
- Diller arası çeviri

DDİ uygulamaları üzerinde çalışırken karşımıza çıkan ilk konu dilin abecesidir. Çünkü bilişim ortamındaki çalışmalar abecenin harflerine karşılık düşen kodlardır. Tüm dillerin abecelerinin aynı olmadığı bilinmektedir ve bilişim ortamında kullanılan tuş takımı zorlama sonucu bugünkü biçimine gelmiştir. Konuya açıklık getirmek amacıyla önce konuyu tuş takımı açısından ele alalım. Günümüzde farklı tuş takımları kullanılmaktadır ancak bilgisayar dünyası 105 tuşlu tuş takımı konusunda anlaşmış gibidir. Bu ölçünlü tuş takımında 35 tuş ana kısım (abece, nokta, virgül vb.), 13 tuş sayı ve noktalama işaretleri için ayrılmıştır. Harflerin yerleşimi ülkeden ülkeye değişebilmektedir. QWRT ile başlayan tuş takımı insanları yavaş yazmaya zorlayan tuş takımı olarak bilinir. Bir yüzyıl önceki mekanik daktilolarda yazma hızını düşürmek gerektiği için bu yola gidilmiştir. Şekil-4'te QWRT tuş takımı, Şekil-5'te Q-Türk tuş takımı ve Şekil-6'da F Türk tuş takımı görülmektedir. On parmak kör yazmaya alışık olan Türkler hızlı yazmayı sağladığı için F-Türk tuş takımını kullanırlar. Bilgisayarcılar Q-Türk tuş takımını kullanırlar.

QWRT tuş takımında 35'lik ana kısım içine 26 harf yer almaktadır. Geri kalan 9 tuş takımı çok kullanılan noktalama işaretlerine ayrılmıştır. Türkçe abece 29 harf olduğu için bazı noktalama işareti tuşları 13 tuşluk üst sıraya kaydırılmıştır. Özbek abecesinde de 29 harf görülmektedir. Ancak bu harflerin bazıları iki harf ile bazıları iki vuruşla yazılabilmektedir. **Yeni bir Özbek abecesine geçilecek ise sesçil bir abece kullanılması ve çift vuruşlu harflerden vazgeçilmesi önemli olacaktır.** Bu önerimizin nedenini şöyle açıklayabiliriz:

Sesçil abecede bir sese karşılık bir harf kullanılır, dolayısıyla bir harfe karşılık bilgisayarda bir kot karşılık düşürülür. Örneğin Türkçe abecede görülen “Ş” harfi bilgisayarda bir kot karşılık gelirken, aynı sesi “SH” ikilisi ile gösterirsek, bilgisayarda iki kot kullanmamız gerekir.

Metin üzerinde çalışırken, örneğin sözcükleri abece sırasına koyarken çift harf ile gösterilen sesler zorluk çıkarmaktadır.

Bir ses için iki tuşa basmak yazma hızını da düşürür.



Şekil 5: QRT tuş takımı



Şekil 4: Q-Türk tuş takımı



Şekil 6: F Türk Tuş takımı

Özbek dili için ölçünlü tuş takımının yeterli olacağı açıktır. Harfler tuş takımında istenen yere yerleştirilebilirler. Özbeklerin alışık olduğu bir harf yerleşim kuralı varsa bu kullanılabilir. Önemli olan harflere karşılık düşecek kotların ölçünlü bilişim kodlarına denk getirilmesidir. Konuyu Türkçe abecesinden örnek vererek anlatmak uygun olacaktır. Türkçe abecesinde bulunan harfler şunlardır: a, b, c, ç, d, e, f, g, ğ, h, ı, i, j, k, l, m, n, o, ö, p, r, s, ş, t, u, ü, v, y, z. İngiliz abecesinde olan q, w, x harfleri Türkçede yoktur. Buna karşın ek olarak ç, ğ, ı, ö, ş, ü harfleri bulunmaktadır. Yabancı sözcükleri yazabilmek amacıyla Türkçe tuş takımlarında q, w ve x harflerine yer verilmiştir. 6 Türk harfine 35'lik alanda yer bulunmuştur. Türkçenin bu özel harfleri Bilişim dünyasının kod abecesinde yer almıştır. ç, ö ve ü harfleri ölçünlü bilişim abecesinde olduğu için yalnızca ğ, ı, İ ve ş bilişim abecesine konmuştur. Dolayısıyla bugün Türkçenin bitin harfleri ölçünlü bilişim abecesinde yer almaktadır.

Özbek abecesi için iki “a”, iki “e” geniz n si önerildiğini duyuyorum é, ä, ñ. Türkçede olan ı, j, ü harfleri yok. Buna göre yeni düşündüğünüz abecedeki Özbekçe'ye özgü harfleri ölçünlü bilişim abecesi UTF-8 e Çizelge-3'teki gibi yerleştirebilirsiniz.

Sonuç ve Öneriler:

- *Her ses için bir harf kullanılması,*
- *Çift vuruşlu harflerin kullanılmaması*
- *Özbekçeye özgü harflerin mutlaka ölçünlü bilişim abecesinden seçilmesi*

önerilir.

Çizelge-3: Özbekçe harfler için öneri

	Büyük harf	Küçük harf
ä	&#196	&#228
ç	&#199	&#231
é	&#201	&#223
ğ	&#1x11E	&#1x11F
ñ	&#209	&#231
ö	&#214	&#246
ü	&#220	&#241

Kaynakça

- [1] Phoenicians Alphabet, <http://www.phoenician.org/alphabet.htm>
- [2] Phoenician/Canaanite, <http://www.omniglot.com/writing/phoenician.htm>
- [3] The Phoenician Alphabet, <http://phoenicia.org/alphabet.html>
- [4] C. Violatti, Greek Alphabet, http://www.ancient.eu/Greek_Alphabet/
- [5] M. A. Kılıç, Türkiye Türkçesindeki Ünlülerin Sesbilgisel Özellikleri, 10. Uluslararası Türk Dilbilim Kurultayı (ICTL), 16)18 Ağus.2000 İstanbul

SO‘ZLARNING BIRIKUVCHANLIGI LUG‘ATI

Baxtiyor Mengliyev Rajabovich

ToshDO‘TAU, Amaliy tilshunoslik kafedrası mudiri,
filologiya fanlari doktori, professor

Javlon Ibragimov (QarDU)

Zamonaviy tilshunoslikning eng dolzarb muammolaridan biri so‘zlarning birikuvchanligi masalasidir. So‘zlarning birikuvchanlik xususiyati deganda uning gap, so‘z birikmasi yoki boshqa sintaktik qurilmada boshqa so‘zlar qurshovida qo‘llanishidir. Bu xususiyat ma‘lum bir tilning leksik, grammatik va boshqa lingvistik qoidalari bilan tartibga solinadi. Ya‘ni bunda tilning bir nechta bo‘limlari (leksikologiya, sintakasis, frazeologiya, morfologiya)ning o‘zaro hamkorligi zarurdir. Chunki birikmada so‘zlarning ma‘naviy muvofiqligi leksikologiyaga aloqador bo‘lsa, ularni biriktiruvchi morfologik vositalar va ularning nutqda qo‘llanishi morfologiya, sintaksis va uslubiyat kabilar bilan bog‘liqlikni ko‘rsatadi. Iboralar ma‘nosining bir so‘zga teng kelishi esa, ularning gap qurilishida bir so‘z kabi maqomga ega bo‘lishini anglatadi.

So‘zning birikuvchanlik qobiliyati uning lug‘aviy ma‘nosiga ham bevosita bog‘liq. Leksik-semantik birliklar anglatadigan ma‘nosi bilan ma‘lum bir leksik paradigmalariga kiradi va boshqa birliklar bilan birikadi. Lug‘atlarda lingvistik nuqtayi nazardan so‘zlarni aniq tavsiflash tildagi so‘zlar alohida holda emas, balki boshqa so‘zlar bilan bog‘liqlikda mavjud bo‘lishiga asoslanadi. Lug‘atlashtirishda bu xususiyatlarni hisobga olish leksik birliklarning turli xususiyatlarini, shu jumladan, birikish xususiyatlarini yanada izchil va tizimli aks ettirishga imkon yaratadi. Lug‘at maqolasidagi bosh so‘z birikuvchanlik xususiyatlarining berilishi uning semantik tuzilishini to‘liq ifodalashga yordam beradi, chunki leksik birlik birikuvchanligi tadqiqi, nazarimizda, uni tahlil qilishning asosiy vositasidir. Shu o‘rinda so‘z ma‘nosi va uning birikuvchanligi XX asr o‘rtalaridan boshlab dunyo tilshunosligida eng dolzarb masala bo‘lib kelayotganligi ta‘kidlash joiz. Bu lug‘aviy birliklarning matndagi o‘zaro bog‘liqligining o‘ziga xosligi va muqarrarligi ularning deyarli barcha paradigmatic munosabatlarida aks etishi bilan bog‘liq. Darhaqiqat, birikuvchanlikning leksikologiya, sintaksis kabi bo‘limlar bilan bog‘liqligi uning muayyan tilda batafsil tadqiq qilinishi sanalgan tilshunoslik sohalari haqida ham juda qimmatli ilmiy xulosalar chiqarishga ko‘maklashadi. So‘z ma‘nosi tahliliga bag‘ishlangan tadqiqotlarda so‘zlar birikuvchanligidan foydalanish ham aynan shunga asoslanadi. Shunday qilib, so‘zni semantik tahlil qilish vositalaridan biri sifatida so‘zlar birikuvchanligining ahamiyati juda katta. Ammo hozirgi kunda bu sohada amaliy va nazariy tadqiqotlarning kamligi birikuvchanlik lug‘atlari va ularga oid tadqiqotlarning zarurligini ko‘rsatadi.

Hozirgi vaqtda so‘zlar birikuvi masalasi tadqiqi (shu bilan birga birikuvchanlik lug‘atlarini yaratishning ham) dolzarbligini bu soha zamonaviy

tilshunoslikning amaliy-nutqiy yo'nalishlari, bevosita muloqot, so'z ma'nosi taraqqiyoti, shakllanishi, nutqda qo'llanishi kabi muammolarni o'rganishga e'tiborni qaratishi bilan asoslanishi mumkin. Til birliklarining birikuvchanlik xususiyatlarini umumlashtirish va tasniflash zamonaviy tilshunoslik zarurati hisoblanadi. Shuningdek, chet tillarini o'qitish ko'lamining kengayishi so'zlar birikuvchanligi yuzasidan qiyosiy-tipologik tadqiqotlar o'tkazish talabini keltirib chiqaradi. Keyingi bosqichlarda esa tildan foydalanishning turli sohalarida maxsus birikuvchanlik lug'atlari yaratish zaruratini paydo qiladi.

So'zlar birikuvchanligiga doir tadqiqotlar o'tkazish va birikuvchanlik lug'atlarini yaratishning yana bir zarurati hozirgacha bu sohada boy manba to'planganligi, ularni tizimlashtirish lozimligi bilan belgilanadi. Ammo bu faoliyat (so'zlar birikuvchanligi tadqiqi) so'zlar birikuvi masalasining murakkabligi, vaqt o'tishi bilan so'zlar birikuviga ta'sir etishi mumkin bo'lgan omillarning tadqiq qilinmaganligi, so'zlarning lug'aviy birikuvchanligi tamoyillari qanday shakllanganligi aniqlanmaganligi kabi muammolarga duch keladi va ularni hal etish bilan amalga oshadi. Albatta, bu kabi muammolarni ham yaratilajak mukammal birikuvchanlik lug'atlari va birikuvchanlikka doir olib boriladigan tadqiqotlar yechadi.

So'zlar birikuvchanligini aks ettiradigan lug'atlarga ehtiyoj sezadigan sohalardan biri tilshunoslikka (yoki umuman tilga) oid ma'lumotlardan foydalanadigan dastur yaratuvchilaridir. Ular ana shu ma'lumotlarning yetishmasligi, ba'zan umuman mavjud emasligi muammosiga duch keladilar. Ma'lumotlar deganda turli tipdagi lug'atlar, korpus ma'lumotlari, grammatik qo'llanmalar, umuman, yaratilajak dastur uchun so'z haqidagi har qanday ma'lumot nazarda tutilmoqda. Aynan so'z haqidagi ma'lumotlar boshqa sohalarda til, uning o'ziga xosligi haqida tasavvur hosil qilishda muhim hisoblanadi. Tilning har bir sathida so'z o'z ma'lumotini tashiydi (fonetik, morfologik, sintaktik, semantik va h.k.z.). Semantik ma'lumot deganda so'zning izohli lug'atlarida so'z izohi sifatida berilgan ma'lumotlar to'plamini tushunishimiz mumkin. Bunda tarjima lug'atlari ham turli tillardagi so'zlarning izohini beruvchi avtomatik qidiruvlarda zarur bo'lgan muhim ma'lumot manbayi hisoblanadi. Sintaktik ma'lumotlar esa so'zlarning birikuvchanlik imkoniyatlari, xususiyatlari va ularning voqelanishi natijasida hosil bo'ladigan qurilmalar haqidagi axborot jamlanmasidir. Birikuvchanlik lug'atlari aloqa qoliplari (masalan, qaratqich kelishigidagi so'zning egalik shaklidagi so'zga birikishi; nisbat shakllarining fe'llar birikish xususiyatlariga ta'sir qilishi kabi ma'lumotlar birikuvchanlik lug'atlarida beriladi), lug'aviy va sintaktik valentliklar mosligi, iboralar, shuningdek boshqa ko'chma ma'nodagi so'zlarning birikish-biriktirish xususiyatlarini aks ettiruvchi lug'atlardir (ya'ni, maxsus birikuvchanlik lug'atlari). Birikuvchanlik lug'atlarining muhimligi shundaki, ular asosida sintaktik xususiyatlarni aks ettiruvchi boshqa lug'atlarni ham yaratish mumkin.

Birikuvchanlik lug'atlaridan olinadigan sintaktik ma'lumotlar tabiiy tillar elektron dasturlarda muhim (sun'iy tillar yaratilsa ham, avvalo, elektron dasturlarga mos kelishi hisobga olinishi shubhasiz) hisoblanadi. Chunki birikuvchanlik haqidagi maxsus qo'shimcha ma'lumotlar birikma ma'nosidagi polisemantiklik bilan bog'liq noaniqliklarni bartaraf etadi, aynan mana shu tipdagi noaniqliklar avtomatik qidiruv bilan bog'liq asosiy muammolardan biridir. Quyida avtomatik qidiruvda tushunmovchilik keltirib chiqaradigan bir nechta misollarni keltiramiz:

Karim stoldan nonni oldi va tushirib qo'ydi.

Ushbu misoldagi *u* olmoshi katta ehtimol bilan *non* so'zini almashtirmoqda. Aksincha bo'lishiga, ya'ni *stolni* tushirib qo'yishiga ishonish qiyin. Yana bir gapni tahlil qilamiz:

Karim nonni stoldan oldi va uni artidi.

Endi *u* olmoshi *stol* so'zi o'rnida qo'llangan ekanligini shu tilda so'zlashuvchilar, umuman dunyo haqida, ya'ni odatda, stolning tushib ketmasligi, nonning artilmasligi haqida tasavvurga ega insonlar tushunadi. (Stolning, masalan, tashilayotganda tushib ketishi, nonning artilishi (masalan, tushirib qo'yilganda g'uborni ketkazish maqsadida) kamdan-kam hollar bu yerda hisobga olinmadi). Ammo kompyuter, smartfon kabi elektron jihozlar bunday bilimga ega emas. Biroq bu ma'lumotlarni birikuvchanlik lug'atlari yordamida bemalol modellashtirish mumkin. Chunki birikuvchanlik lug'atlarida *nonni tushirib qo'ymoq*, *stolni artmoq* kabi birikuvlar qayd etilgan bo'ladi. Shuningdek, dasturchilar *mebelni artmoq*, *yegulikni tushirib qo'ymoq* birikmalarini ham umumlashtirgan holda avtomatik qidiruv tizimiga kiritishlari mumkin. Bunda stol mebelning bir turi, non ham umuman olganda ovqat ekanligi haqidagi ma'lumotlar qo'l kelishi tabiiy.

Birikuvchanlik lug'atlarini muvaffaqiyat bilan qo'llash mumkin bo'lgan yana bir soha – avtomatik tarjima sohasidir. Aslida ikkita (tarjima amalga oshirilayotgan tillardagi lug'atlar) katta birikuvchanlik lug'atlari va ularning muvofiqligi bexato tarjima uchun yetarli bo'lishi lozim. Hatto unchalik katta bo'lmagan lug'atlar ham ba'zan mukammal tarjima uchun yetarli bo'ladi. Masalan, *to pay attention* – (so'zma-so'z “*diqqatni to'lamoq*” tarzida o'giriladi) rus tiliga *обратить внимание*, ispan tiliga *prestar atención* (so'zma-so'z *e'tiborni jalb qilmoq*) shaklida o'giriladi. O'zbekchaga *to pay attention* birikmasi avtomatik tarjimada “*e'tibor bermoq*” tarzida o'giriladi. Bu birikma mazkur tillarga so'zma-so'z o'girilganda tushunmovchiliklar yuzaga chiqar edi. Bunda, albatta, birikuvchanlik haqidagi ma'lumotlar muhim o'rin tutadi.

ЎЗБЕК ТИЛИ МАТНЛАРИНИНГ РАЗМЕТКАЛАНГАН КОРПУСИНИ МОДЕЛЛАШТИРИШ

Қаршиев Абдували Беркинович

Физика-математика фанлари номзоди,
ТАТУ Самарқанд филиали профессори
abduvalikarshiyev@gmail.com

Аннотация. Ушбу ишда, Ўзбек тили миллий корпусини қурилишида пайдо бўлган назарий ва амалий масалалар келтирилган. Корпусни қуриш вақтни ва меҳнатни талаб қилувчи жараёндир. Мавжуд корпусларни таҳлил қилиш натижасида қўриш мумкинлиги, корпусларни қуришда барча тиллар учун ягона бўлган методология мавжуд эмас. Ушбу мақолада корпусга қўйиладиган талаблар, корпус қуриш модели ва босқичлари таҳлил қилинган.

Калит сўзлар: *корпус, модел, маълумотлар базаси, матн, разметка, тег, JSON, сўз.*

Abstract. In this work, the theoretical and practical issues arising in the construction of the national corpus of the Uzbek language are presented. Building a hull is a time-consuming and labor-intensive process. As a result of the analysis of existing corpora, it can be seen that there is no single methodology for building corpora for all languages. This article analyzes the hull requirements, hull construction model and steps.

Keywords: *corpus, model, database, text, markup, tag, JSON, word.*

Абстракт. В данной работе представлены теоретические и практические вопросы, возникающие при построении национального корпуса узбекского языка. Строительство корпуса — трудоемкий и трудоемкий процесс. В результате анализа существующих корпусов видно, что единой методики построения корпусов для всех языков не существует. В этой статье анализируются требования к корпусу, модель конструкции корпуса и этапы.

Ключевые слова: *корпус, модель, база данных, текст, разметка, тег, JSON, слово.*

Кириш. Мавжуд корпусларни таҳлил қилиш натижасида қўриш мумкинлиги, корпусларни қуришда барча тиллар учун ягона бўлган методология мавжуд эмас. Бунинг сабаби шундаки, турли тилларда турли хил қоидалар, технологик жараёнлар амал қилади. Морфологик тавсифлар тўпламлари турлича бўлишига қарамасдан, юқорида қараб чиқилган корпусларнинг барчасида морфологик (ёки грамматик) разметка таъминланган. Синтаксис разметка эса баъзи корпусларда бор, лекин турлича ёндошувлар асосида амалга оширилган бўлса, баъзиларида умуман йўқ. Корпусларни тилни қамраб олиш даражаси бўйича солиштирганда, рус тили миллий корпусини ҳам хронология, ҳам жанрлараро энг мувозанатлашган ва матнлар хилма-хиллиги таъминланган корпус деб ҳисоблаш мумкин. Лекин рус тили миллий корпусининг бу ютуғи

унинг асосий камчилигини ҳам юзага келтирган. Ҳар қандай йирик корпусда тўлиқ ва аниқ разметкаланишнинг имконияти чекланган бўлади. Ноаниқликлар, асосан, автоматик разметкалашда омоним сўзлар туфайли пайдо бўлади.

Мавжуд корпуслар таҳлилидан қўриш мумкинки, миллий корпус тилнинг қонуният ва хусусиятларини иложи борида қўп тақдим эта олиши ва таркиби бўйича мувозанатлашган бўлиши керак. Матнларнинг бир тури тилдаги жами матнлар ичида қандай ҳажмий улушга эга бўлса, ана шундай миқдор миллий корпусда ҳам сақланиб қолиши керак [А.Б.Қаршиев, 2021].

Лингвистик корпусларга талаблар. Корпус қўлланилишига доир қуйидаги талаблар қўйилиши мумкин:

1. Корпус ўқув ва маълумот қўлланмалар ёзишда, луғатлар тузишда ҳамда лингвистик экспертизалар ва таҳлилларни ишлаб чиқишда, тил маълумотлари сифатида ишончли ва ваколатли манба ҳисобланади;

2. Корпус турли хил лингвистик фанларни ўқитишда худди тилнинг кўргазмали (иллюстрация) манбаси, ҳамда таҳлилларни тақдимотлилик кўриниши сифатида ишлатиш мумкин;

3. Корпус илмий тадқиқот ишларида, шу ўринда илмий фаразларни текширувчи сифатида қўлланилиши мумкин;

4. Корпус тил фаолиятининг универсал маълумотлар қўлланмаси сифатида бўлиши мумкин.

Корпус яратишда қуйидаги саволларга асосланган ҳолда талабларни белгилаб олиш муҳим:

1. Корпус фойдаланувчиси ким? (индивидуал, гуруҳ, лингвистик жамият).

2. Корпуснинг мантиқий ғояси нимада?

3. Корпусни қуришда қанча ҳажмдаги маълумотлар билан ишламоқчисиз?

4. Нима ишлатилади: матн парчалари, тўлиқ матнлар ёки иккаласи ҳамми?

5. Корпусдаги матнларни танлаш тартиби қандай?

Корпус яратишда турли хил разметканинг бўлиш ёки бўлмаслигини ва агар бўлса, унинг аниқлик даражасини ҳисоб олиш мақсадга мувофиқ. Корпус дастурий таъминоти таркибида автоматик разметкалаш дастурлари албатта бўлиши керак. Разметканинг юқори аниқликда бўлишини фақат филолог мутахассис таъминлаши мумкин [Kennedy G., 1998]. Филолог мутахассис ҳар бир матннинг разметкаланишини сўзма-сўз синчиклаб текшириши, автоматик разметка хатоларини тузатиши ва омоним ҳолатларни бартараф қилиши керак. Ишнинг асосий қисми қўл меҳнати орқали бажарилиши мумкин. Шунинг учун филолог мутахассиснинг компьютердаги ишига мўлжалланган махсус разметкалаш дастурлари яратилиши зарур. Бу дастурлар қуйидаги сифатларга эга бўлиши керак:

- фойдаланувчи сўзлар ёки гапларнинг грамматик ёки синтаксис характеристикаларини етарли даражада тўлиқ аниқлаш ва ўзгартириш имкониятига эга бўлсин;

- таҳлил максимал даражада қулай ва содда бўлиб, тез бажарилсин ва бу таҳлилнинг тўлиқлигини камайтирмасин [M.S.Tursunov, 2022].

Корпус моделининг схематик шакли 1-расмда тасвирланган [Седов А.В., 2013].

1-расм. Корпус моделининг схематик тасвири



Корпус модели. Корпусни T_i матнларнинг жамланмаси деб қараймиз. Матн эса p_{ik} тиниш белгилари, пробел ёки сатр охири белгилари билан ўзора ажратилган w_{ij} сўзлар кетма-кетлигидан иборат [Седов А.В., 2013]:

$$Corpus = \{T_1, T_2, \dots, T_n\},$$

$$T_i = \{w_{ij}\} U \{p_{ik}\},$$

ерда j - w_{ij} сўзнинг i -матндаги ўрнини, k - p_{ik} тиниш белгининг ўрнини билдирувчи қиймат. Шунингдек, ҳар бир сўзга унинг матндаги жойлашув ўрни мос қўйилади:

$$Disposition: w \rightarrow positions \in Positions, Positions = Pos1, Pos2, \dots, Posq,$$

бу ерда w сўз учун унинг матндаги координатаси q аниқланади. Матндаги боблар, параграфлар, абзацлар, гаплар, гап бўлаклари - унинг қўшимча

тузилмавий бирликларидир. Модел матннинг энг кичик тузилмавий бирликларидан (сўзлардан) сўз бирикмалари ҳосил бўлишини инобатга олади. Гапни сўзлар тўплами деб қаралса, ҳар бир сўз бирикмани қисмий тўплам деб ҳисоблаш мумкин:

$$SSj = wi1, wi2, \dots, wik \subset S, \text{ бунда } UjSSj = S.$$

Ҳар бир сўз фақат битта сўз бирикмаси таркибига кириши мумкин, шунинг учун иккита бирикманинг кесишмаси бўш тўпламдан иборат бўлади: $SSk \cap SSj = \emptyset$, бунда $k \neq j$.

Ҳар бир сўз учун матндаги ўрнидан ташқари унинг морфологик параметрлари тўплами ҳам аниқланиши керак. От сўз туркумига тегишли сўз учун бу параметрлар – келишик, сон (бирлик/кўплик), жонли/жонсиз.

Морфологик параметрларнинг ўзи яна бир векторни ташкил қилади:

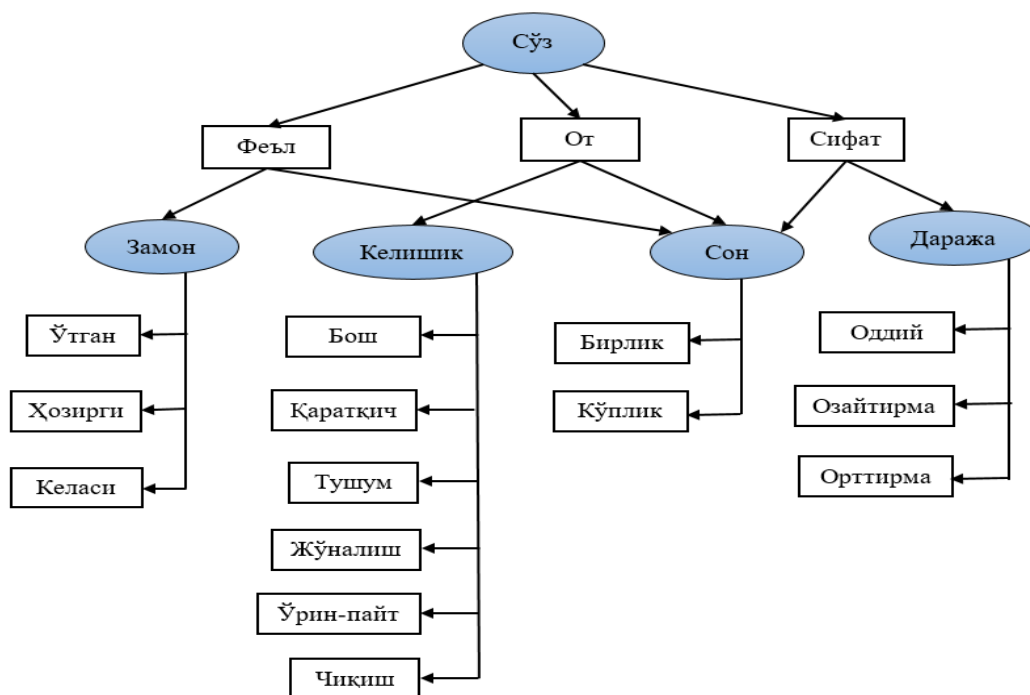
$Morfology: w \rightarrow intParams \in Nm, intParams = Param1, Param2, \dots, Paramm$, бу ерда m – сўзнинг грамматик белгиларининг сони. Морфологик белгилар натурал сонлар билан кодланган. Ушбу белгиларнинг декодланиши алоҳида конфигурация файлда сақланади ва дастур администратори ёки фойдаланувчининг ўзи ҳам белгиларни ўзгартириши мумкин.

Грамматик белгиларнинг ўзаро боғлиқлигини график кўринишда тасвирлаш мумкин:

$$MorfAttr = (W, E), W = MorfParam \cup MorfItems, MorfParam, MorfItems \neq \emptyset, MorfParam \cap MorfItems = \emptyset, E = E1 \cup E2 \quad E1 = u, v: u \in MorfParam, v \in MorfItems, \quad E2 = v, u: u \in MorfParam, v \in MorfItems,$$

бу ерда $MorfParam$ – морфологик параметрлар тўплами, $MorfItems$ – морфологик параметрлар қийматлари тўплами. Морфологик параметр, масалан, от сўз туркумига тегишли сўз учун – бу келишик бўлса, морфологик қийматлар – бош, қаратқич, тушум, жўналиш, ўрин-пайт, чиқиш.

Мисол сифатида ўзбек тили корпусига киритилган матнни грамматик разметкалашда (А иловага қаранг) сўз таҳлилининг бир қисмини схема шаклида келтирамиз (2-расм).



2-расм. Сўзнинг морфологик таҳлили иерархик дарахтини қуришга доир
мисол

Ҳар бир сўз учун морфологик параметрлардан ташқари сатрли параметрлар деб аталадиган қўшимча параметрлар ҳам киритилган. Бу параметрлар, хусусан, сўз ўзаги ва леммасидан иборат. Сўзларни разметкалашнинг универсал дастурини яратишда бундай параметрлар сонини ва таркибини олдиндан аниқлаб олиш имконияти назарда тутилиши зарур. Формал ҳолда $WStrAttr:w \rightarrow sParams=sParam_i, i=1, l$ деб ёзсак, $sParams$ тузилма фойдаланувчи томонидан алоҳида конфигурация файлида берилиши керак.

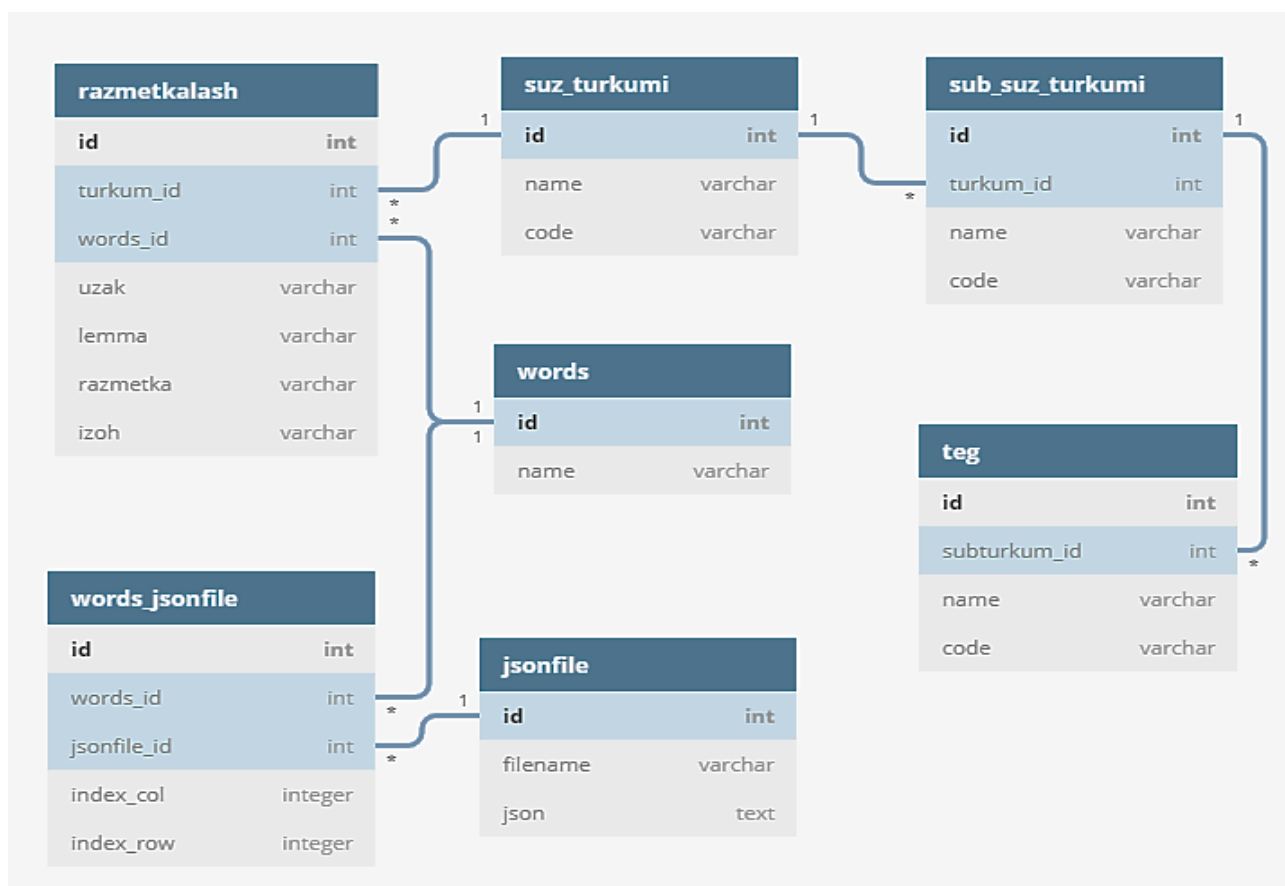
Шундай қилиб, ҳар бир сўзга қуйидаги тузилма мос қўйилади:

$$w \rightarrow \langle Spelling, Disposition, intParams, sParams \rangle$$

(сўз ўрни, морфологик параметрлар, сатрли параметрлар). Бу тузилмани сўзнинг грамматик разметкаси деб ҳисоблаш мумкин.

Маълумотларни компютерлаштиришда турли хил файл форматлари билан танланган матнлар, УТФ-8 кодланган матн форматларига ўтказилади [Гвишиани Н.Б., 2018, Грудева Е.В. 2012]. Мавжуд ўзбекча матнларни қайта ишлаш ва мониторинг қилиш учун вебга асосланган қидирув ва бошқарув корпуси ишлаб чиқилган. Ўзбек тили миллий корпусикорпус бошқарув тизимида компютерлаштирилган файллар биринчи метаразметка ҳолатида базага (PostgreSQL) киритилади. Сўнгра матнли маълумотлар ва транскрипцион сўзлар тадқиқот гуруҳининг эксперт тилшунослари томонидан кузатилади ва таҳрир қилинади. Ўзбек тили миллий корпусига матн маълумотлари киритилаётганда батафсил текширилади.

Разметкаланган корпус сақланадиган маълумотлар базасининг тузилмаси 3-расмда тасвирланган.



3-расм. Маълумотлар базасининг тузилиши

Words жадвал корпусдаги сўзларни сақлайди. *Razmetkalash* жадвали корпусдаги сўзларнинг грамматик маълумотларни (ўзак, лемма ва разметкасини) ўз ичига олади. *Suz_turkumi* жадвали ўзбек тили сўз туркумларини ўз ичига олади ва сўзни разметкалаш жараёнида, унинг туркуми шу жадвалдан чақирилади. *Sub_suz_turkumi* жадвалида тилнинг грамматик маълумотларини ўз ичига олади. *Teg* жадвалида *Sub_suz_turkumi* грамматик маълумотларини ички грамматикасини ўз ичига олади. Сўзни разметкалаш жараёнида *razmetkalash* жадвалининг *razmetka* майдонига грамматик маълумотларни ёзиш жараёнида, даставвал *Suz_turkumi* жадвалига боғланиб, сўзнинг туркуми белгиланади. Ундан сўнг *Sub_suz_turkumi* жадвали билан боғланиб, бу жадвал *Teg* жадвалини чақиради ва *Sub_suz_turkumi* ҳамда *Teg* жадваллари фойдаланувчи учун танланган сўз туркумига мос равишда грамматик маълумотларни тақдим этади. Фойдаланувчи томонидан танланган грамматик маълумотлар *Razmetkalash* жадвалининг *razmetka* майдонига ёзилади. *Jsonfile* жалвали корпусдаги матнларни сақлайди. *Words_jsonfile* жадвали корпусдаги сўзларни, *Jsonfile* жалвалида жойлашган матнлардаги адреслар (ўрганилаётган сўз жойлашган матн ва ушбу матндаги учраш нуқтаси) харитаси ҳисобланади.

Корпус яратиш босқичлари. Корпусни яратиш жараёни бир неча босқичлардан иборат:

маълумотларни тўплаш: корпусга киритиладиган матнлар рўйхатини тузиш, уларнинг манбасини аниқлаш;

маълумотларни компьютерлаштириш: танланган маълумотларни (электрон матнлар, сканерланган матнлар, клавиатурадан терилган матн ва ҳ.к.) юклаш;

метамаълумотларни кодлаштириш: танланган матн ҳақидаги метамаълумотларни ёзиш;

маълумотларни разметкаш: матннинг ҳамма бирликларини теглаш;

корпусда қидирув интерфейсини ишлаб чиқиш: корпусдан маълумотлар олишни қулайлаштирадиган ва енгиллаштирадиган график веб-интерфейс яратиш;

корпусни ишга тушириш: дастлаб синов тариқасидаги версияни, сўнгра миллий ва халқора даражада фойдаланишга мўлжалланган биринчи версия ишини йўлга қўйиш.

тил корпуси нафақат қидирув тизими ёрдамида конкорданс тузиш, балки сўзларнинг морфологик маълумотларини тақдим этиши лозим. Разметкаланган матнлар JSON форматда корпус таркибига киритилади, бунда матн абзацлар, гаплар ва сўзлар каби таркибий бўлақларга бўлинади. Бунда ҳар бир сўзнинг ёнига, унинг морфологик маълумотлари ҳам бириктирилади.

Хулоса. Мавжуд корпуслар таҳлили асосида қуйидаги хулосалар олинди: корпусларни қуришда барча тиллар учун ягона бўлган методология мавжуд эмас;

морфологик характеристикалар тўплamlари турлича бўлишига қарамасдан, юқорида қараб чиқилган корпусларнинг барчасида морфологик (ёки грамматик) разметка таъминланган.

ҳар қандай йирик корпусда тўлиқ ва аниқ разметкалашнинг имконияти чекланган бўлиб, ноаниқликлар, асосан, автоматик разметкалашда омоним сўзлар туфайли пайдо бўлади;

миллий корпус тилнинг қонуният ва хусусиятларини иложи борида қўп тақдим эта олиши ва таркиби бўйича мувозанатлашган бўлиши керак;

матнларнинг бир тури тилдаги жами матнлар ичида қандай ҳажмий улушга эга бўлса, ана шундай миқдор миллий корпусда ҳам сақланиб қолиши керак.

Корпус яратишда турли хил разметканинг бўлиш ёки бўлмаслигини ва агар бўлса, унинг аниқлик даражасини ҳисоб олиш мақсадга мувофиқ. Корпус дастурий таъминоти таркибида разметкалаш дастурлари албатта бўлиши керак. Бу дастурлар қуйидаги сифатларга эга бўлиши керак:

- фойдаланувчи сўзлар ёки гапларнинг грамматик ёки синтаксис характеристикаларини етарли даражада тўлиқ аниқлаш ва ўзгартириш имкониятига эга бўлсин;

- таҳлил максимал даражада қулай ва содда бўлиб, тез бажарилсин ва бу таҳлилнинг тўлиқлигини камайтирмасин.



Фойдаланилган адабиётлар

1. Седов А.В., Математические модели, методы и алгоритмы построения размеченных корпусов текстов, дис. кан. наук, Петрозаводск, 2013.
2. А.Б.Қаршиев, С.А.Каримов, М.С.Турсунов, Development of a Modern Corpus of Computational Linguistics // Conference: 2020 International Conference on Information Science and Communications Technologies (ICISCT), DOI: 10.1109/ICISCT50599.2020.9351376, 2021.
3. M.S.Tursunov, Description of the management system programs of the national corpus of the uzbek language // International journal of engineering mathematics, <https://iejemta.com/index.php/em/article/view/9/9>, volume 4 ISSUE 1, 2022, 32-42 p.
4. Гвишиани Н.Б. Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. М.: 2008.
5. Грудева Е.В. Корпусная лингвистика: Учебное пособие. 2-е изд., стереотип. М.: 2012.
6. Kennedy G., An introduction to corpus linguistics. London: Longman, 1998.

SUN'IY INTELLEKT: BORIS

Xolmanova Zulxumor

Filologiya fanlari doktori, professor,
ToshDOTAU professori
E-mail:zulxumor-uzmu@mail.ru

Annotatsiya. Maqolada sun'iy intellekt va uning fan sohalarida, xususan, kompyuter lingvistikasida qo'llanishi haqida so'z yuritilgan. Matnni avtomatik tahrirlash va matn ustida ishlash amallarining sun'iy intellekt asosidagi tizimlari, xususan, darak gaplardan iborat matnlarni avtomatik tahlil qilish tizimi (BORIS)ning ishlash prinsiplari bayon qilingan. Matn mazmunini o'zlashtirishda, matn osti ma'nolarini anglashda, presuppozitsiyalarni aniqlashda sun'iy intellekt asosidagi BORIS tizimining o'rni ko'rsatib berilgan. O'zbek tilidagi matnlar asosida ishlaydigan sun'iy intellektlarning yaratilishi bu tilni kompyuter dasturlari tiliga aylantirishda muhim o'rin tutishi ta'kidlangan.

Kalit so'zlar: *kompyuter lingvistikasi, sun'iy intellekt, BORIS tizimi, kompyuter dasturlari ishlash prinsipi, savol-javob tizimi, START, METAKOP (sabab zanjiri), PTRANS (ma'lum bir obyektning bir joyda ikkinchi joyga ko'chishi), ATRANS (ma'lum bir shaxsning muayyan narsaga ega bo'lishi), ATTEND (voqeaga turtki bo'lgan holat), PROPEL (ma'lum narsaga erishish uchun harakat qilish), ATB (affektiv mazmunli bloklar, insonning psixologik holati).*

Аннотация. В статье обсуждается искусственный интеллект и его применение в отраслях науки, в частности в компьютерной лингвистике. Описаны принципы автоматического редактирования текста и обработки текста на основе искусственного интеллекта, в частности система автоматического анализа текста (BORIS). Показана роль системы BORIS на основе искусственного интеллекта в освоении текстового содержания, понимании подтекстовых смыслов, определении пресуппозиций.

Подчеркивается, что создание искусственных интеллектов, работающих на основе текстов на узбекском языке, сыграет важную роль в превращении этого языка в язык программирования.

Ключевые слова: компьютерная лингвистика, искусственный интеллект, система BORIS, принципы работы компьютерных программ, вопросно-ответная система, START, METAKOP (причинно-следственная цепь), PTRANS (перемещение определенного объекта из одного места в другое), ATRANS (иметь что-то, ATTEND (состояние, мотивирующее событие), PROPEL (стремление к достижению чего-то), ATB (блоки аффективного содержания, психологическое состояние человека).

Abstract. The article discusses artificial intelligence and its application in the fields of science, in particular in computational linguistics. The principles of automatic text editing and text processing based artificial intelligence, in particular,

the automatic text analysis system (BORIS), are described. The role of the BORIS system based on artificial intelligence in mastering textual content, understanding subtextual meanings, and determining presuppositions is shown. It is emphasized that the creation of artificial intelligences based on texts in the Uzbek language will play an important role in turning this language into a programming language.

Keywords: computational linguistics, artificial intelligence, BORIS sistem, principles of computer programs, question-answer system, START, METAKOP (causal chain), PTRANS (moving a certain object from one place to another), ATRANS (to have something), ATTEND (State, motivating event), PROPEL (desire to achieve something), ATB (blocks of affective content, psychological state of a person).

Sun'iy intellekt – XX asr oxiri va XXI asr boshlarida turli fan sohalarida, jumladan, kompyuter lingvistikasida faol qo'llanayotgan tushunchalardan biridir. Sun'iy intellekt fanlararo soha hisoblanadi, keyingi yillarda bu tushuncha kompyuter lingvistikasi bilan yonma-yon istifoda etilmoqda. Sun'iy intellekt integrativ xususiyatga ega bo'lib, uning turli ko'rinish va usullaridan bir qator sohalarda foydalaniladi. Kompyuter lingvistikasi nutqni va matnlarni tadqiq qilishning kompyuter usullarini taqdim etar ekan, psixologiya, mantiq, sotsiolingvistika kabi fanlar qatorida sun'iy intellekt ko'rinishlarini ham qamrab oladi.

Kompyuter lingvistikasi taraqqiyoti natijasida sun'iy intellekt qo'llanadigan yo'nalishlar doirasi ham kengaydi. Shu bois kompyuter lingvistikasiga doir adabiyotlarda sun'iy intellekt kompyuter lingvistikasining bir yo'nalishi sifatida talqin qilingan. Fanda hali-hanuz sun'iy intellektning aniq va mukammal ta'rifi mavjud emas, chunki fanda bu tizimga hali to'liq erishilgani yo'q. Mavjud ta'rif va tavsiflarda mazkur sohada erishilgan yutuqlarga asosanlanilgan.

Virtual qomus hisoblangan Wikipedia sahifalarida sun'iy intellektga (artificial intelligence) shunday ta'rif berilgan: «Sun'iy intellekt – insonning aqliy faoliyati bilan bog'liq bo'lgan masalalarni, xususan, fikrlash, nutq mazmunini anglash, ma'lumotlarni umumlashtirish kabilarni yechishga yo'naltirilgan kompyuter yoki robot imkoniyatidir. Bu termin mazkur xususiyatlarga ega sistemalarni ishlab chiqish bilan bog'liq kompyuter texnologiyalarining bir bo'limiga nisbatan ham ishlatilmoqda».[www.wikipedia]. Sun'iy intellekt bo'yicha tadqiqotlar olib borgan olim Jon Makkarti shunday yozadi: «Sun'iy intellekt – intellektual mashinalar, intellektual kompyuter dasturlarini yaratish texnologiyasi va u haqidagi fan. Intellektual sistema strukturasi uch asosiy blokni o'z ichiga oladi: bilimlar bazasi, masalani yechuvchi tizim va intellektual interfeys» [www-formal.stanford]. Bu ta'riflardan kelib chiqib, sun'iy intellektni fan, soha yoki yo'nalish emas, “kompyuter texnologiyasi tizimi”, “intellektual faoliyat”, “imkoniyat” sifatida baholash to'g'riroq bo'ladi.

Sun'iy intellekt insondagi intellekt tushunchasining aynan nusxasi emas, balki inson intellektual faoliyatining ko'plab qirralarini o'z ichiga olgan, modellashtirilgan tizimlar texnologiyasi demakdir. «Artificial intelligence» – sun'iy intellekt termini

birinchi marta Jon Makkarti tomonidan 1956-yilda Darmut universitetidagi konferensiyada qo'landi va ommalashdi. Jahon miqyosida sun'iy intellekt bo'yicha assotsiatsiyalar tuzilgan. Mazkur assotsiatsiyalar hududiy bo'limlarni o'z ichiga oladi. Masalan, Rossiya sun'iy intellekt assotsiatsiyasi Peterburg bo'limi bir necha yillardan beri faoliyat olib bormoqda.

XX asrning 70-yillariga qadar sun'iy intellekt bo'yicha tadqiqotlar kibernetika hamda informatika doirasida olib borilgan. XX asrning 80-90-yillaridan boshlab sun'iy intellekt juda ko'plab fanlarning o'rganish obyektiga aylandi. Jumladan, neyrolingvistika, psixologiya, informatika, neyrofiziologiya, gnoseologiya, kognitologiya, kognitiv lingvistika, kompyuter lingvistikasi kabi fanlar ham o'z doirasida sun'iy intellekt masalasiga murojaat qiladi [Rassel, Norvig, 2006:8; Po'latov, 2009:7-9].

Hozirda dunyo axborot tizimida Sage, GPT-4, Claude, Claude-instant, ChatGPT, NeevaAI, Dragonfly nomlaridagi sun'iy intellektlardan foydalanilmoqda.

Sun'iy intellekt tizimiga doir amaliy ishlardan biri ekspert tizimlari hisoblanadi. Ekspert tizimi – muayyan sohadaagi ekspertlik yoki mavjud bilimlar yig'indisi sifatida yaratilgan kompyuter dasturi. Ekspert tizimi maxsus sohaga oid barcha bilimlar jamlangan qomus yoki professional ekspertlarning kompyuter modelidir. Ekspert tizimlari odatiy kompyuter dasturlaridan farq qiladi. Odatiy kompyuter dasturlari qidiruv tizimi, hisoblash va manipulyatsiya tizimigagina ega bo'ladi. Ekspert tizimlari mazkur jihatlardan tashqari xuddi insonlarda bo'lgani kabi dalillarga tayanib muhokama yuritish imkoniyatiga ham egadir. Ekspert tizimlarining asosiy ikkita tarkibiy qismi mavjud: ma'lumotlar ombori va mantiqiy mashina [Xolmanova, 2019:148].

Ko'pchilik mutaxassislarning fikricha, ekspert tizimlari va neyron tarmoqlari sun'iy intellektga asoslangan ikki mustaqil yo'nalishdir. Har ikkala yo'nalishning o'ziga xos afzalliklari hamda cheklangan tomonlari mavjud. Shuni e'tiborga olgan holda aytish mumkinki, neyron tarmoqlari va ekspert tizimlarining birgalikda ishlashini ta'minlovchi gibrid tizimlar ishlab chiqilishi lozim [Xolmanova, 2019:148].

Ekspert tizimlariga ekspertlar hamda bir qator yordamchi dasturlar kiradi. Ekspert tizimining bir ko'rinishi bo'lgan savol-javob tizimi sun'iy intellekt imkoniyatlarini kengroq namoyon qilishga xizmat qiladi. Savol-javob asosidagi dastur foydalanuvchi bilan tabiiy tilda munosabatga kirishishni ta'minlab beradi. Masalan, START dunyodagi birinchi savol-javob tarzida ishlaydigan on-line boshqaruv tizimi hisoblanadi.

Axborot uzatishda faqat xabar olish emas, balki xabar atrofidagi tegishli ma'lumotlarni ham o'zlashtirish, bilish lozim bo'ladi. Shu ma'noda, jahon tilshunosligida kompyuter va sun'iy intellekt imkoniyatlari asosida bir qator dasturiy tizimlar yaratilganini ta'kidlash joiz. Ana shunday dasturiy tizimlardan biri ingliz tili asosidagi BORIS tizimidir.

BORIS – “darak gaplardan iborat matnlarni avtomatik tahlil qilish tizimi” degan mazmundagi ifodaning qisqartmasi bo‘lib, XX asrning II yarmida ingliz kompyuter lingvistikasida keng yoyilgan. BORIS matnni tushunadigan va savollarga javob beradigan eksperimental tizim, u bir qator bilim manbasiga asoslanadi. Matnni uning asosiy mazmuninigina o‘zlashtirib, qolgan jihatlarini e’tibordan chetda qoldiradigan yuzaki tizimlarga zid o‘laroq, BORIS tizimi matnni chuqur tushunishga qaratilgan. Matn qismlarini BORIS dasturi asosida qayta ishlash natijasida mazmunni chuqur o‘zlashtirish imkoni yaratiladi.

BORIS savol-javob asosida amalga oshadigan dasturiy tizim bo‘lib, matnni ma’lum bir algoritm asosida tahlil qiladi. BORIS tizimida mashinaning o‘zi dastur asosida savol qo‘yadi, javob beradi va matnni tahlil qiladi. Tizim javob berishi mumkin bo‘lgan savollar botiniy matn talqiniga aloqador bo‘lgan, fikr asosida turuvchi tasavvurdagi sabab-oqibat munosabatlarini aks ettiradi.

BORIS tizimi istiqbolli tadqiqotlar loyihalari agentligi (ARPA), AQSH Milliy ilmiy fondi tomonidan moliyalashtirilgan. Sun’iy intellekt asosidagi BORIS tizimining ishlash prinsipi, protsedurasi, bosqichlari Vendi Lenert, M.G.Dayer, P.N.Djonson, K.Dj.Yang, C.Xarlilar tomonidan ko‘rsatib berilgan [Wendi, 1983:15-62]. Maqolaning ruscha variantidan foydalandik va BORIS tizimini tavsiflashda rus tilidagi tarjima matnni aynan keltirishni ma’qul topdik. BORIS tizimi “RAZVOD” (Ajrim) sarlavhali matn asosida ko‘rsatilgan [Vendi, 1989:107-159]. Dastur tizimi asosidagi savollar kichik harflarda, javoblar katta harflarda ifodalangan.

(БОРИС. РАЗВОД).

ОБРАБАТЫВАЕТСЯ АБЗАЦ 1

Ричард несколько лет не имел никаких известий о Поле, своем товарище по комнате во время учебы в университете. В то время Ричард одолжил у Поля некоторую сумму, которая так и не была возвращена. Теперь же он совершенно не представлял себе, где он может найти своего старого друга. Когда вдруг из Сан-Франциско пришло письмо, Ричард загорелся желанием узнат, как живет Поль.

СИСТЕМА ГОТОВА К РАБОТЕ В РЕЖИМУ В/О

Какое событие произошло у Ричарда дома?

РИЧАРД ПОЛУЧИЛ ПИСЬМО ОТ ПОЛЯ.

Кто такой Поль?

ДРУГ РИЧАРДА.

Хотел ли Ричард видеть Поля?

ДА, РИЧАРД ХОТЕЛ ЗНАТ, КАК ЖИВЕТ ПОЛЬ.

Помогал ли когда –либо Пол Ричарду?

ДА, ПОЛЬ ДАЛ РИЧАРДУ ДЕНГИ В ДОЛГ.

Почему Ричард не вернул долг Полю?

РИЧАРД НЕ ЗНАЛ, ГДЕ НАХОДИТСЯ ПОЛЬ.

Что почувствовал Ричард, когда пришло письмо?

РИЧАРД БЫЛ ОБРАДОВАН, ПОТОМУ ЧТО ПОЛЬ И ОН БЫЛИ ДРУЗЬЯМИ.

Savol-javob asosida sun'iy intellekt bir qator ma'lumotlarni oydinlashtiradi. Matn boshlanishida keltirilgan Richardning bir necha yillardan buyon Pol haqida hech qanday ma'lumotga ega bo'lmagani haqidagi axborot Richardning Polni qidirayotgani va uning qayerdaligini bila olmayotganini ham ifoda etgan. Uning Polni nima uchun qidirayotgani esa keyingi jumladan ma'lum bo'ladi: o'sha yillarda (talabalik davrida) Richard Poldan ma'lum miqdorda pul qarz olib, haligacha qaytara olmayotgan edi.

Matnda Poldan xat kelgani aytilmagan. Lekin sun'iy intellekt “Richard Poldan xat oldi” degan hukmni ma'lum qilgan. Richardning ruhiy holati ham matnda aytilmagan. Tizim savol-javob asosida “Richard xursand bo'ldi, chunki u Pol bilan do'st edi” degan ma'lumot bergan.

ОБРАБАТЫВАЕТСЯ АБЗАЦ 2

К сожалению, новости не были хорошими. Жена Поля Сара настаивала на разводе. Она хотела также получить машину, дом, взять детей и добиться алиментов. Поль хотел развода, но он не хотел, чтобы к Саре перешло все их имущество. Его жалованье в государственной школе было очень мало. Не зная, к кому обратиться, он надеялся на содействие единственного знакомого ему адвоката. Поль указал номер своего домашнего телефона на тот случай, если Ричард сочтет, что может быть ему полезным.

СИСТЕМА ГОТОВО К РАБОТЕ В РЕЖИМЕ В/О

Что сообщалось в письме?

ПОЛЬ И САРА ЗАТЕЯЛИ ДЕЛО О РАЗВОДЕ.

Написал ли Поль письмо Саре?

НЕТ, НЕ САРЕ. ПИСЬМО ПОЛУЧИЛ РИЧАРД.

Почему Поль был расстроен в связи с разводом?

ПОЛЬ И САРА СПОРИЛИ ПО ПОВОДУ СЕМЕЙНОГО ИМУЩЕСТВА.

Чем Поль зарабатывал себе на жизнь?

ПОЛЬ БЫЛ УЧИТЕЛЕМ.

Чего хотела Сара?

САРА ХОТЕЛА ПОЛУЧИТЬ МАШИНУ И ДОМ, ВЗЯТЬ МАЛЫШЕЙ И ДОБИТЬСЯ АЛИМЕНТОВ.

Почему Поль написал Ричарду?

ПОЛЬ ХОТЕЛ, ЧТОБЫ РИЧАРД БЫЛ ЕГО АДВОКАТОМ.

Ikkinchi abzas tahlilida Richardning xatni o'qigani haqida ma'lumot yo'q. Ammo butun boshli abzas xat mazmuniga bag'ishlangan. Xatda yoqimsiz xabar yozilganiga ishora qilingan. Savol-javobda BORIS “yoqimsiz xabar”ni oydinlashtiradi: “Pol va Sara munosabati ajralishgacha borib yetgan edi”. Orada tizim sinovchi savolni ham o'rta tashlaydi: “Pol Saraga xat yozdimi?”. Bu savolning qo'yilishi ikki omilga bog'lanadi: 1) maktub yozishning aksariyat hollarda sevishganlar yoki er-xotinlar o'rtasidagi munosabatlarda kuzatilishi; 2) so'z

borayotgan ajrim va uning natijasi sifatidagi mol-mulk musodarasida Polning saraga maktub yozishi mumkinligi. Tizimning o'zi bunga aniqlik kiritadi: “Yo‘q. Saraga emas. Xatni Richard oldi”. Bu javob orqali tizim “xat Richardga yozilgan edi” degan mazmunni ifoda etgan. Matnning tahlil qilingan abzasida Polning ajrimdan qoniqmayotganligi aytilmagan. Bu jihat ham savol-javob orqali ochilgan: “Pol va Sara oilaviy mol-mulk masalasida kelisha olmayotgan edi”. Tizim muammo nimadaligini aniq topgan.

“Uning davlat maktabida topayotgani juda kam edi” degan axborot asosida BORIS “Pol o‘qituvchi edi” deya hukm qilgan. Oddiy o‘quvchi bu o‘rinda Polning ish haqi kamligi bilan advokat qidirish o‘rtasidagi aloqani tezda ilg‘aydi: advokat yollash uchun katta miqdordagi mablag‘ kerak. BORIS bu aloqadorlikni qahramonning “kimga murojaat qilishni bilmasdan, o‘ziga tanish bo‘lgan yagona advokatga umid bog‘lagani”dan topadi. Shu asosda “Pol Richard uning advokati bo‘lishini xohlagan edi” degan hukmni ifoda etadi. Qizig‘i shundaki, tahlil qilinayotgan matnda Richardning advokatligi aytilmagan.

ОБРАБАТЫВАЕТСЯ АБЗАЦ 3

Ричард нетерпеливо схватил трубку телефона и набрал номер. После короткого разговора Поль договорился с ним позавтракать вместе на следующий день. В его голосе слышались явное облегчение и благодарность.

СИСТЕМА ГОТОВО К РАБОТЕ В РЕЖИМЕ В/О

Почему Ричард позвонил Полю?

РИЧАРД ХОТЕЛ ДОГОВОРИТЬСЯ С ПОЛЕМ О ВСТРЕЧЕ.

Установил ли Ричард связь с Полем?

ДА, РИЧАРД ПОЗВОНИЛ ПОЛЮ.

Как Ричард установил связь с Полем?

ПО ТЕЛЕФОНУ.

Что почувствовал Поль, когда позвонил Ричард?

ПОЛЬ БЫЛ ДОВОЛЕН, ТАК КАК РИЧАРД СОГЛАСИЛСЯ БЫТЬ АДВОКАТОМ ПОЛЯ.

Uchinchi abzasda keltirilgan “Richardning sabrsizlik bilan telefon trubkasini olgani” semiotik xarakterdagi axborot hisoblanadi. Bu xabar asosida BORIS “Richard uning advokati bo‘lishga rozilik berganidan Pol mamnun edi” degan xulosaga kelgan. Bu hukmning zanjiri ikkinchi abzasga bog‘lanadi. Ikkinchi abzas oxirida “Richard advokatlikka rozi bo‘lsagina telefon qilishi mumkinligi ta’kidlanib, telefon nomeri ko‘rsatilgani” aytilgan edi. Tizim shu jumladagi mantiqni anglagan holda tegishli xulosaga kelgan.

BORIS tizim savol-javob asosida quyidagi qo‘shiicha bilimlarni hosil qiladi:

KOP (aloqadorlik): qarzdorlik;

METAKOP (sabab zanjiri): ajrim;

PTRANS (ma’lum bir obyektning bir joydan ikkinchi joyga ko‘chishi): restoranga, restorandan uyga borish;

ATRANS (ma'lum bir shaxsning muayyan narsaga ega bo'lishi, nazorat qilishi): Polning mol-mulkka ega bo'lishi;

ATTEND (voqealarga turtki bo'lgan holat): qahfaning to'kilib ketishi;

PROPEL (ma'lum narsaga erishish uchun harakat qilish): Polning uy-joy, mashina, mol-mulkning ma'lum qismiga da'vo qilishi);

ATB (affektiv mazmunli bloklar, insonning psixologik holati):

1. Mamnunlik.
2. Asabiylashish.
3. G'azablanish.
4. To'satdan anglash.

BORIS tizimiga aloqador mazkur tushunchalar ingliz tilidagi so'z va qisqartmalar bo'lib, terminologik mazmun ifodalaydi.

BORIS tizimi sun'iy intellekt imkoniyatlarini namoyon qilishda ahamiyatlidir. Tizim quyidagilarda ham o'ziga xos o'rin tutadi:

1. Matn ustida ishlash ko'nikmasini shakllantiradi.
2. Savol-javob asosida mantiqiy xulosalar chiqarish ko'nikmasini takomillashtiradi
3. Matn tahlilini amalga oshirgan moderator funksiyasini bajaradi.
4. Matn mazmunini to'liq o'zlashtirishga, matn osti hukmlarni anglashga, presuppozitsiyani yoritishga xizmat qiladi.
5. Matnda ishtirok etgan shaxs ruhiyatini, ichki olamini aks ettiradi.
6. Matnning ekspozitsiyasi, tuguni, voqealar rivoji, yechim nuqtalarini ajratib tahlil qilish imkonini beradi.
7. Matn qismlari o'rtasidagi aloqadorlikni ko'rsatishga xizmat qiladi.

Sun'iy intellekt inson tomonidan yaratiladi. Shu bois uning cheklangan tomonlari ham mavjud. Masalan, yuqorida qayd etilgan Sage, GPT-4, Claude, Claude-instant, ChatGPT, NeevaAI, Dragonfly kabi sun'iy intellekt ingliz tilidagi savollarga batafsil, to'liq javob beradi. Boshqa tillarda operatsiyalarning nisbatan cheklanishi namoyon bo'ladi. Bu holat sun'iy intellektning qaysi tildagi ma'lumotlarni ko'p o'zlashtirgani va qaysi tilda ko'p bora mantiqiy amallarni bajarishi bilan belgiladi.

Sun'iy intellekt o'zbek tilidagi matnlarni to'g'ri tushunishi va talqini qilishi uchun shu tildagi matn ustida ko'p bora amallar bajarilgan bo'lishi kerak. Shu tildagi ma'lumotlar bazasi o'zlashtirilgan bo'lishi lozim. Inson muloqot qilishi, so'zlashi uchun lison zarur bo'lganidek, sun'iy intellekt uchun ham ma'lumotlar bazasi asosidagi lisoniy ong zarur bo'ladi.

O'zbek kompyuter lingvistikasi hali shakllanish bosqichida. Yaqin kunlarda rivojlanish bosqichiga o'tib, o'zbekcha matn ustida turli operatsiyalarni bajarishga qaratilgan sun'iy itellekt tizimlarini yaratish dastlabki vazifalardan biridir.

Xulosa qilib aytganda, BORIS tizimini takomillashtirish kompyuter lingvistikasining istiqbollaridan biri bo'lib, matn mazmunini o'zlashtirishda yuqori samaradorligi bilan ahamiyatlidir.



Foydalanilgan adabiyotlar:

1. Wendi G. Lehnert, Michael G. Dyer, Peter N. Johnson, C.J. Yang, Steve Harley (Department of Computer Science, Yale University. New Haven., USA). BORIS – An Experiment in-Depth Understanding of Narratives. - “Artificial Intelligence”. Vol. 20. –№1, 1983. –P.15-62.
2. Венди Ленерт, М.Г.Дайер, П.Н.Джонсон, К.Дж.Яна, С.Харли. BORIS – экспериментальная система глубинного понимания повествовательных текстов. «Новое в зарубежной лингвистике». – М., 1989. – С. 107-159.
3. Рассел С. Норвиг П. Искусственный интеллект: современный подход / Artificial Intelligence: a Modern Approach /Пер. с англ. и ред. К.А.Птицына. 2-е изд. - М.: Вильямс, 2006. – С.8-17.
4. Po‘latov A., Muhamedova S. Kompyuter lingvistikasi. –Toshkent, 2009.
5. Rahimov A. Kompyuter lingvistikasi asoslari. –Toshkent, 2011.
6. Xolmanova Z. Kompyuter lingvistikasi. –Toshkent, 2019.
7. http://www.wikipedia.org/wiki/artificial_intelligence.html
8. <http://www-formal.stanford.edu/jmc/whatisai.html>



TABIIY TILNI QAYTA ISHLASH (NLP)

ПРОЕКТИРОВАНИЕ ГЕОИНФОРМАЦИОННОЙ СИСТЕМЫ "ДИАЛЕКТОЛОГ" С ИСПОЛЬЗОВАНИЕМ НЕЧЕТКОЙ ЛОГИКИ

Галимов Мансур Радикович

Студент по компьютерным наукам, студент

magl.galimov@gmail.com

Институт Прикладной Семиотики Академии Наук Республики Татарстан

Казанский Федеральный Университет, Казань, Россия

Бурнашев Рустам Арифович

Кандидат технических наук, к.т.н

r.burnashev@inbox.ru

Институт Прикладной Семиотики Академии Наук Республики Татарстан

Казанский Федеральный Университет, Казань, Россия

Гатиатуллин Айрат Рафизович

Кандидат технических наук, к.т.н

agat1972@mail.ru

Институт Прикладной Семиотики Академии Наук Республики Татарстан

Казанский Федеральный Университет, Казань, Россия

Прокопьев Николай Аркадиевич

Магистр по компьютерным наукам, магистр

nikolai.prokopyev@gmail.com

Институт Прикладной Семиотики Академии Наук Республики Татарстан

Казанский Федеральный Университет, Казань, Россия

Аннотация. В данном докладе рассматриваются результаты научных исследований по проектированию прототипа нечеткой экспертной системы для обработки знаний в области диалектологии, а также последующей визуализации данных с использованием геоинформационных систем и технологий. Своими исследованиями мы стремимся внести свой вклад в развитие интеллектуальных языковых информационных ресурсов с использованием геоинформационных систем и технологий. Для разработки веб-интерфейса программной оболочки был использован фреймворк Django языка программирования Python. Для анализа, обработки и визуализации данных использовались программные библиотеки Pandas и Folium. Для повышения эффективности обработки знаний была использована графовая база данных neo4j.

Ключевые слова: нечеткие экспертные системы, база знаний, Folium, Pandas, ГИС.

Annotation. This report discusses the results of scientific research on designing a prototype fuzzy expert system for processing knowledge in the field of dialectology, as well as subsequent data visualization using geographic information systems and technologies. With our research, we strive to contribute to the development of intelligent language information resources using geographic

information systems and technologies. The Django framework of the Python programming language was used to develop the shell web interface. Pandas and Folium software libraries were used for data analysis, processing and visualization. To increase the efficiency of knowledge processing, the neo4j graph database was used.

Keywords: fuzzy expert systems, knowledge base, Folium, Pandas, GIS.

ВВЕДЕНИЕ

В настоящее время диалектолог взаимодействует с большим набором языковых информационных ресурсов и баз данных. Из-за большого объема данных специалисту трудно найти необходимую информацию в процессе профессиональной деятельности. В связи с этим одним из главных приоритетов является создание современных программных приложений с элементами искусственного интеллекта, связанных геопространственными данными.

В профессиональной деятельности диалектолога возникает трудность в сборе и обработке входной информации, получаемой в режиме реального времени. Использование на практике современных средств геоинформационных технологий для считывания и обработки различий диалектов народов с последующим картографированием является актуальной задачей. Использование лингвогеографического метода повысит эффективность принятия решений во время научных экспедиций и исследований.

Мы также провели эксперимент по внедрению графической базы данных neo4j для работы с геоинформационной системой.

В качестве одной из последних разработок в области создания лингвистической ГИС можно выделить работу Дж. Мороза из Лаборатории языковой конвергенции и Школы лингвистики, программа *lingtypology* [1]. Эта программа представляет собой бесплатный R-пакет для картографирования и лингвистической типологии.

Лингвистическое картографирование [3] является относительно устоявшейся областью исследований в лингвистике, но только недавно лингвогеография начала использовать геоинформационные системы (ГИС) в качестве инструмента [4]. Сочетание географических характеристик с типологическими исследованиями и новейшими достижениями в области информационных технологий позволяет создавать современные инструменты, такие как автоматизированное рабочее место диалектолога, типолога, этимолога и т.д. Использование таких инструментов позволяет осуществлять поиск корреляций между языковыми параметрами и географическим распределением языков. Инструментарий расширяется за счет использования статистических технологий, что, например, позволяет усовершенствовать методы оценки географических расстояний между различными языковыми сообществами. Несмотря на появление работ, использующих методы

лингвогеографии, не существует единого технологического инструментария, в частности, для интеграции баз данных предыдущих исследований, и это приводит к тому, что каждая созданная карта создается один раз для определенных целей и не может быть использована для других исследований.

Основой для создания диалектов разных языков являются теоретические положения лингвистической географии [3]. Картографирование диалектных явлений основано на следующих концепциях:

- диалектный язык - это совокупность диалектов языка, включающая как черты, общие для всего языка, так и черты, которые имеют различные варианты в диалектах, но в то же время являются элементами общей языковой системы.;
- диалектное различие, формирующее соответствующее явление (в более поздних работах также междиалектное соответствие), является основным объектом отображения в атласе, лингвистическим явлением, которое в диалектах проявляется в различных вариантах. Graphic information is perceived several times faster [4] than text information.

В настоящее время ГИС-технологии активно используются системами искусственного интеллекта, в частности, экспертными системами [5-8]. Например:

- MAPEX [9, 10] представляет собой экспертную систему, основанную на правилах автоматического обобщения картографических ресурсов.
- AUTONAP [9, 10] представляет собой экспертную систему для автоматического размещения названий объектов на географической карте. Эта система имитирует действия опытного картографа при выполнении задачи по нанесению названий объектов на карту. AUTONAP использует эвристические знания о размещении имен на основе установленных процедур и соглашений в базе знаний.
- CES [11, 12] является прототипом картографической экспертной системы и используется в качестве картографического консультанта в области энергетики, шахт и т.д. Используя таблицы принятия решений, он предоставляет набор картографических спецификаций, которые оптимально соответствуют набору требований к карте, предъявляемых в системе.

Для разработки нашего программного приложения, использовались библиотеки языка программирования Python: Django, Pandas, Folium, os, NumPy, Neomodel, и т.д.

Библиотека Folium использовалась для работы с интерактивными картами [13].

Folium это библиотека в Python, которая помогает разрабатывать карты Leaflet.js (JavaScript язык программирования). Позволяет нам выполнять обработку данных в Python и визуализировать свои объекты на карте.

Folium позволяет легко визуализировать данные, которые были обработаны в Python на интерактивной карте Leaflet. Он позволяет отображать данные на карте, а также передавать векторные / растровые / HTML-элементы в качестве маркеров на карте.

В библиотеке есть несколько встроенных наборов карт из OpenStreetMap, Mapbox и Stamen (Рис. 1.). Библиотека поддерживает обработку файлов в формате представления различных географических структур данных (GeoJSON и TopoJSON), изображений и видео.



Fig. 1. Добавление географических координат на карту.

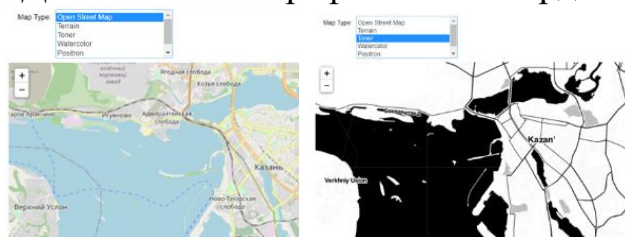


Fig. 2. Добавление интерактивных стилей (OpenStreetMap, Toner)

Обработка и анализ данных. Современные библиотеки программного обеспечения вносят значительный вклад в разработку программных средств и средств обработки знаний. Одним из наиболее важных свойств таких инструментов является общая доступность и полезность их использования. Такие инструменты могут решать интеллектуальные задачи, характерные для людей – классифицировать различные объекты, распознавать закономерности в изображениях, предсказывать события и т.д. Даже сейчас интеллектуальные информационные системы обрабатывают данные и решают проблемы гораздо быстрее и точнее, чем люди.

Для обработки и анализа данных использовалась библиотека pandas [14]. Pandas - это программная библиотека Python для обработки и анализа данных. Функциональность программной библиотеки pandas с данными построена поверх библиотеки NumPy, которая является инструментом более низкого уровня. Предоставляет специальные структуры данных и операции для работы с числовыми таблицами, временными рядами и др.

Библиотека включает в себя объект DataFrame (двумерный массив), аналогичный электронной таблице Excel (данные из файла можно загрузить с

помощью команды `pandas read_csv` ('имя файла')).

С помощью библиотеки программного обеспечения вы можете выполнять те же манипуляции с данными: объединять список, сортировать по требуемому атрибуту, выполнять анализ, генерировать отчеты и т.д.

Ниже (Рис. 3) приведен пример данных, полученных из файла для последующей интеграции в географическую информационную систему диалектолога.

```
In [322]: df
```

```
Out[322]:
```

	word	latitude	longitude	country	city	icon_num
0	жан	55.78	49.13	Russia	Kazan	1
1	жылы	55.90	49.13	Russia	Kazan	2
2	жирпе	55.95	49.13	Russia	Kazan	3
3	жухаз	55.78	48.40	Russia	Innopolis	4
4	божра	56.40	49.13	Russia	Kazan	5
5	жөмле	55.12	49.13	Russia	Kazan	6

Рис. 3. Фрагмент DataFrame после считывания и обработки геопространственных данных.

Ниже приведен программный код для интеграции данных с файловой системы на графическую карту программной библиотеки Folium (Рис. 4.):

```
In [313]: # plot dialects locations
for (index, row) in df.iterrows():
    folium.Marker(location=[row.loc['latitude'], row.loc['longitude']],
                  popup=row.loc['word'] + ' ' + row.loc['city'] + ' ' + row.loc['country'],
                  tooltip='click').add_to(folium_map)

# display map
folium_map
```

Рис. 4. Фрагмент кода для добавления данных на карту

Для работы с графиками и отчетами использовалась программная библиотека Matplotlib. Matplotlib — это библиотека на языке программирования Python для визуализации данных с помощью двумерной (2D) графики. Обработанные графики могут быть использованы в качестве иллюстраций при формировании отчета (статистики).

Формат файла GeoJSON предназначен для хранения географовых структур данных, основанных на JSON.

GeoJSON поддерживает следующие типы объектов:

- Точка (включая адреса и местоположения).
- Строка линий (включая улицы, шоссе, границы).
- Полигон (включая страны, провинции и земельные участки).
- Составные объекты типов point, linestring или polygon.

Ниже приведен листинг программного кода файла GeoJSON:

```
{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "properties": {},
      "geometry": {
        "type": "Polygon",
        "coordinates": [
          [[46.120605, 57.01192],
            [48.537598, 57.569959],
            [51.61377, 56.856072],
```

[52.536621,54.782834],
.....,
[44.736328,55.110943],
[46.120605,57.01192]]]]}}}}

Ниже приведен прототип системы (Рис. 4.), созданный в результате исследований по обработке знаний и последующей визуализации географических структур данных.

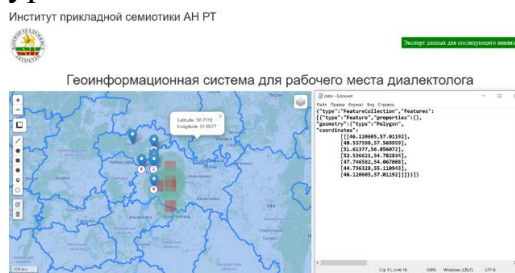


Рис. 4. Основная форма с интерактивными элементами на карте

Программная библиотека Neomodel была использована при разработке для реализации следующей архитектуры. Каждый язык может иметь связи с регионами (Рис. 5.), в свою очередь, язык может содержать слова, которые могут использоваться сразу в нескольких диалектах.

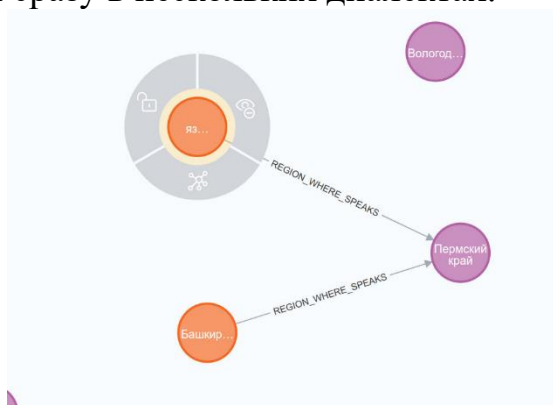


Рис. 5. Пример взаимосвязи между регионом и диалектом

Таким образом, мы получаем набор узлов (Рис. 6.), которые можно выводить в геоинформационную систему и взаимодействовать с ней для получения дополнительной информации об узле.

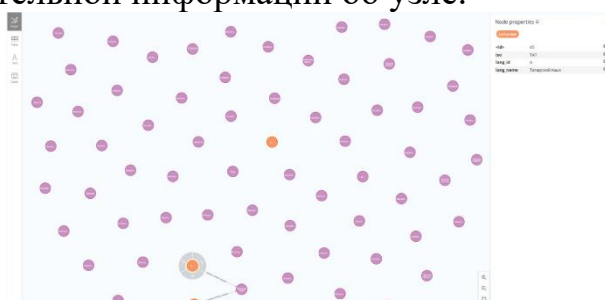


Рис. 6. Набор регионов и языковых узлов.

В качестве примера разработанной модели возьмём Language. Модель Neo позволяет связать модели Django со спецификой Neo4j, где соединения раскрываются с помощью необычного внешнего ключа, а также RelationshipTo и RelationshipFrom.

```
Class Language(StructuredNode,NodeUtils:
lang_code = IntegerProperty()
iso = StringProperty(max_length=5)
lang_id=IntegerProperty(index=True)
lang_name = StringProperty(max_length=100, required=True)
regions = RelationshipTo(".region.Region",
"REGION_WHERE_SPEAKS")
```

Экспертные системы - это информационные системы, которые частично или полностью заменяют эксперта в определенной предметной области. Основной целью ЭС является разработка программных средств, которые при решении сложных для человека задач получают результаты, не уступающие по качеству и эффективности решениям, получаемым человеком-экспертом. Главным преимуществом экспертных систем является способность накапливать знания, сохранять их в течение длительного времени, обновлять и тем самым обеспечивать относительную независимость конкретной организации от наличия в ней квалифицированных сотрудников. Накопление знаний позволяет повышать квалификацию специалистов, работающих на предприятиях, используя лучшие, проверенные решения. Практическое применение систем искусственного интеллекта в организациях и в экономике основано на экспертных системах, которые повышают качество и сокращают время принятия решений, а также способствуют увеличению объема работы и профессиональному развитию специалистов.

Основной группой методов, используемых в нечеткой экспертной системе (ЭС) [15-17], предлагаемой в проекте, являются методы нечеткой логики. Теория нечетких множеств была предложена американским математиком Лотфи Заде [18]. Применение теории нечетких множеств позволяет описывать нечеткие понятия и знания, оперировать этими знаниями и делать нечеткие выводы. Неопределенные данные и ситуации обрабатываются с использованием нечеткой логики с использованием лингвистических переменных [19, 20].

Нечеткая логика – это логическая система или система управления n-разрядной логической системой, которая использует степени состояния ("степени истинности") входных данных и генерирует выходные данные, которые зависят от состояний входных данных и скорости изменения этих состояний.

Нечеткая операция предполагает использование нечетких множеств и функций принадлежности. Каждый нечеткий набор является представлением лингвистической переменной, которая определяет возможное состояние

выходного сигнала. Функция принадлежности - это функция общего значения в нечетком множестве, так что и общее значение, и нечеткое множество принадлежат универсальному множеству.

Степени принадлежности к этому общему значению в нечетком множестве определяют выходные данные на основе принципа "Если-То". Принадлежность присваивается на основе предположения о выходе с использованием входных данных и скорости изменения входных данных. Функция принадлежности - это, по сути, графическое представление нечеткого множества. Структура нечеткой экспертной системы показана ниже (Рис. 7).

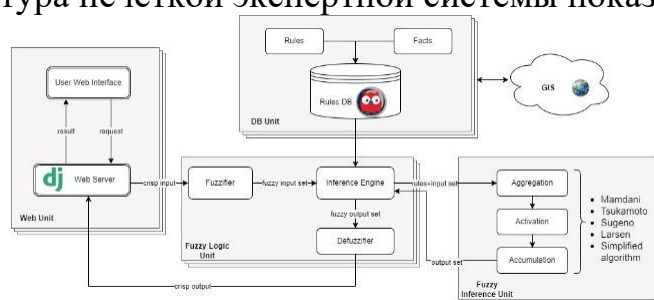


Рис. 7. Структура прототипа нечеткой экспертной системы

База знаний нечеткой экспертной системы включает в себя полную информацию о данных исследований в области диалектологии тюркских языков. Экспертная система включает в себя следующие программные модули: модуль сбора знаний, решатель, пояснительный и диалоговый компоненты.

Модуль получения знаний автоматизирует процесс наполнения новыми знаниями с использованием геоинформационных технологий (ГИС).

Решатель, используя исходные данные из рабочей памяти и знания из баз знаний, формирует последовательность правил логического вывода, которые впоследствии приводят к решению задач в области диалектологии.

Пояснительный и диалоговый компоненты предназначены для получения количественных и качественных показателей для выбранного на карте диалектологического атласа и организации дружественного интерфейса, что повышает эффективность консультаций в области диалектологии.

Логическая программа состоит из предикатов, которые представляют собой функции, генерирующие логические значения. База знаний формируется с использованием логического слова-сказуемого, которое включает в себя факты о предметной области.

Таким образом, разработка нечеткой экспертной системы будет осуществляться на основе представленных этапов разработки, с промежуточной проверкой используемых методов нечеткой логики и лингвистических переменных для конкретных идиом (языков и диалектов). Интеллектуальный интерфейс ЭС позволит специалистам по тюркским языкам и диалектам работать при принятии решений в условиях неполных и противоречивых данных.

База знаний нечеткой экспертной системы будет взаимосвязана с научной работой [21], которая проводится в Институте прикладной семиотики Академии наук Республики Татарстан. Инструменты этого портала позволяют вам описывать как языки, так и диалекты на разных лингвистических уровнях и в разных проявлениях.

Акцент на диалектах делается на том, что разные диалекты одного и того же языка могут быть представлены в базе данных портала как отдельные языки. А такие концептуальные семантические структуры, как тезаурусы типа WordNet [22] и Frameset frame base, позволяют описывать семантику языковых единиц в разных диалектах и проводить сравнительные исследования между диалектами.

Благодарность

Исследование выполнено при финансовой поддержке Республики Татарстан в рамках проекта молодежного научного гранта «Разработка программного приложения для визуализации данных о диалектах с использованием геоинформационных технологий»

Заключение

С использованием современных средств визуализации данных удастся представить информацию максимально доступной для восприятия специалистом в предметной области.

Для интеллектуализации системы планируется дальнейшая интеграция экспертной системы в оболочку геоинформационной системы.

Разработанный проект планируется в будущем интегрировать в структуру портала "Турецкая морфема". ([http://modmorph.turklang.net /](http://modmorph.turklang.net/)).

Для разработки веб-интерфейса программной оболочки был использован фреймворк Django языка программирования Python, для анализа, обработки и визуализации данных использовались программные библиотеки Pandas и Folium.

Список литературы:

1. lingtypology: introduction and installation, Electronic resource: ropensci.github.io/lingtypology
2. Egorov, I.M., Dybo, A.V. and Kassian, A.S., 2022. Phylogeny of the Turkic Languages Inferred from Basic Vocabulary: Limitations of the Lexicostatistical Methods in an Intensive Contact Situation. *Journal of Language Evolution*, 7(1), pp.16-39.
3. Salimov F. I. i dr. Jelektronnaja versija atlasa tatarskih narodnyh govorov // *Filologija i kul'tura*. – 2011. – №. 26. – S. 205-210.
4. Zhdanova E. A., Belyh A. A. Geograficheskie informacionnye sistemy v lingvisticheskix issledovaniyax // *Intellectual'nye sistemy v proizvodstve*. Izhevsk, 2014. № 2 (24). S. 169–174.
5. Frazer, T.C., Breton, R.L., & Schiffman, H.F. (1991). *Geolinguistics: Language Dynamics and Ethnolinguistic Geography*.

6. Briscoe, Ulla. (2009). Geolinguistic GIS Applications: Aspect of Data Quality in Mapping Lesser-Used Languages, Centre for Geoinformatics, Salzburg University
7. R. Gibadullin and N. Marushkai, "Development of Predictive CNN Based Model for Vital Signs Alerts," in International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 2021, pp. 404-409.
8. R. Burnashev, I. Enikeev and A. Enikeev, "Design and Implementation of Integrated Development Environment for Building Rule-Based Expert Systems," 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 2020, pp. 1-4.
9. Choi, Jinmu & Usery, E.. (2004). System Integration of GIS and a Rule-Based Expert System for Urban Mapping. Photogrammetric Engineering & Remote Sensing. 70. 217-224.
10. Eldrandaly, Khalid. (2008). Integrating Expert Systems and GIS for Spatial Decision Making: Current Practices and New Trends.
11. Fisher, P. F. (1989). Expert System Applications in Geography. Area, 21(3), 279–287.
12. Rodriguez-Bachiller, Agustin & Glasson, John. (2004). Expert Systems and Geographic Information Systems for Impact Assessment.
13. K.A. Fedutinov. Structuring environmental information using geographic information technologies. Modeling, optimization and information technology. 2019;7(4). Available from: https://moit.vivt.ru/wp-content/uploads/2019/11/Fedutinov_4_19_1.pdf DOI: 10.26102/2310-6018/2019.27.4.044
14. Il'ichev V.Ju., Jurik E.A. Analiz massivov dannyh s ispol'zovaniem biblioteki Pandas dlja Python // Nauchnoe obozrenie. Tehnicheskie nauki. – 2020. – № 4. – S. 41-45
15. Pospelov D.A. (Red.). Jekspertnye sistemy: sostojanie i perspektivy. 1989. 152 s.
16. Popov Je. V. Osobennosti razrabotki i ispol'zovaniya jekspertnyh sistem //Iskusstvennyj intellekt. Kn. – 1990. – T. 1.
17. Averkin A. Hybrid intelligent systems based on fuzzy logic and deep learning //Artificial Intelligence. – Springer, Cham, 2019. – S. 3-12.
18. Zadeh, L.A., 1988. Fuzzy logic. Computer, 21(4), pp.83-93
19. Zadeh L.A. From computing with numbers to computing with words – from manipulation of measurement to manipulation of perceptions – IEEE Trans. On Circuits and Systems -1: Fundamental Theory and Applications, 45, 1, 1999, 105–119.
20. Demenkov N.P. Jazyk nechetkogo upravleniya // Promyshlennyye ASU i kontrollery. – 2005. – № 5. – S. 30–36.
21. Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B. (2020) About Turkic Morpheme Portal. CEUR Workshop Proceedings Institute for history,



language and literature, Ufa scientific center, Russian Academy of Sciences Proceedings of TurkLang 2020, pp. 226-243.

22. Gatiatullin, A., Kubedinova, L. Multilingual thesaurus for turkic languages (2020) 5th International Conference on Computer Science and Engineering, UBMK 2020, 2020, pp. 393–398, 9219378

O'ZBEK TILI UCHUN TABIIY TILNI QAYTA ISHLASHDA (NLP) MASHINALI O'RGANISHNING (MACHINE LEARNING) O'RNI

Sharipov Maksud Siddiqovich,

Urganch davlat universiteti dotsenti

maqsbek72@gmail.com

Adinayev Xushnudbek Saylboyevich,

TATU Urganch filiali assistant o'qituvchi

hushnud8414@mail.ru

Raximov Raximjon Rajapboyevich,

Urganch davlat universiteti katta o'qituvchisi

rahimjon03@mail.ru

Annotatsiya. Globalizatsiya va raqamli axborotlashtirish jarayonida o'zbek tilining rivojida axborot texnologiyalari o'ta muhim o'ringa ega. Kompyuter lingvistikasining tabiiy tilni qayta ishlash (Natural Language Processing – NLP) jarayoni va tabiiy tilning kompyuter tushunadigan til (formal) shakli yaratilishi natijasida tilga oid masalalar (tahrir, tahlil, tarjima, elektron matnni ovozlashtirish, og'zaki nutqni elektron matnga aylantirish, robot bilan muloqot qilish, yirik matnni ixcham matnga aylantirish kabi) bo'yicha kompyuter dasturlari yaratilmoqda.

Machine Learning va tabiiy tilni qayta ishlash (Natural Language Processing– NLP) so'nggi paytlarda mashhurlikka erishgan sun'iy intellektning muhim sohalaridir. Machine Learning va tabiiy tilni qayta ishlash sun'iy agentni, sun'iy aqlli agentga aylantirishda juda muhim rol o'ynaydi. Sun'iy ravishda aqlli tizim atrof-muhitdan yaxshiroq ma'lumotlarni qabul qilishi va tabiiy tilni qayta ishlashning rivojlanishi tufayli atrof-muhitga foydalanuvchilar uchun qulay tarzda ta'sir qilishi mumkin.

Kalit so'zlar: *tabiiy tilni qayta ishlash, NLP, Machine Learning, Deep Learning, NLTK.*

Abstract. In the process of globalization and digital informatization, information technologies play a very important role in the development of the Uzbek language. As a result of the process of natural language processing (NLP) of computer linguistics and the creation of a language (formal) form that can be understood by a computer, language-related issues (editing, analysis, translation, voicing of electronic text, converting spoken speech into electronic text, using robot computer programs are being created for communicating, converting large text into compact text).

Machine Learning and Natural Language Processing (NLP) are important areas of artificial intelligence that have gained popularity recently. Machine learning and natural language processing play a very important role in transforming an artificial agent into an artificially intelligent agent. An artificially intelligent system can receive better information from the environment and interact with the

environment in a user-friendly way thanks to the development of natural language processing.

Keywords: *natural language processing, NLP, Machine Learning, Deep Learning, NLTK.*

Аннотация: В процессе глобализации и цифровой информатизации информационные технологии играют очень важную роль в развитии узбекского языка. В результате процесса обработки естественного языка (НЛП) компьютерной лингвистики и создания языковой (формальной) формы, понятной компьютеру, даны ответы на вопросы, связанные с языком (редактирование, анализ, перевод, озвучивание электронного текста, преобразование устной речи в электронный текст, с помощью роботов создаются компьютерные программы для общения, преобразования большого текста в компактный текст).

Машинное обучение и обработка естественного языка (NLP) — важные области искусственного интеллекта, которые в последнее время приобрели популярность. Машинное обучение и обработка естественного языка играют очень важную роль в превращении искусственного агента в агента с искусственным интеллектом. Система с искусственным интеллектом может получать более качественную информацию из окружающей среды и взаимодействовать с окружающей средой удобным для пользователя способом благодаря развитию обработки естественного языка.

Ключевые слова: *обработка естественного языка, NLP, Machine Learning, Deep Learning, NLTK.*

Jahon tilshunosligida XX asrning 50-yillaridan til va matnni avtomatik tahlil qilish muammolari bilan bog'liq masalalar kompyuter texnologiyalari yordamida hal qilinmoqda. Natijada matnni avtomatik tushunish, mashina tarjimasini, matnni referatlash (gipermatnni asosiy mazmunni o'zida saqlagan minimatnga keltirish, ya'ni matn hajmini qisqartirish, uni qisqa bayon holiga aylantirish), tasniflash mavzu, uslub va janr jihatidan o'zaro yaqin matnlarni guruhlash), matnni tahrir va tahlil qilish, matnni generatsiyalash (bir nechta tabiiy tilga oid hujjatli matnlardan formula, texnik ishlanmalar, dasturiy tizimlarni yig'ish), matndan ma'lumotni olish, og'zaki nutqni raqamli ma'lumotda berish va aksincha, yozma nutqni og'zaki nutqqa aylantirish, tabiiy tilning Milliy korpusi va boshqa tur lingvistik korpuslarini yaratish, ontologik lug'atlar bazasini shakllantirish kabi dolzarb masalalar kompyuter lingvistikasining mundarijasini egallagan va tabiiy tilni qayta ishlash (Natural Language Processing / NLP) sohasida asosiy vazifa hamda yo'nalishlariga aylangan

Tabiiy tilni qayta ishlash (NLP) nima?

Kompyuterlar kundalik turmushning ajralmas qismi bo'lib, jadvalli / elektron ma'lumotlar bilan ishlashda juda qulay texnik vositasi hisoblanadi. Biroq, odamlar odatda jadvallar shaklida emas, balki so'zlar va jumlar bilan muloqot qilishadi. Og'zaki va yozma nutq matnlari xususiy bo'lib, til umumiylik kasb etib, muayyan

tuzilish (struktura)ga ega [1]. Shu bois kompyuterlar uchun ushbu turdagi ma'lumotlarni qayta ishlash usullarini yaratish taqozo etiladi. Tabiiy tilni qayta ishlash (NLP)dan maqsad kompyuterlarga strukturlangan matnni tushuntirish va mazmun olishni o'rgatishdan iborat. Tabiiy tilni qayta ishlash (NLP) sun'iy intellektning kichik sohasi bo'lib, uning maqsadi kompyuterlar va odamlar o'rtasidagi o'zaro aloqalarni o'rnatishdan iborat. Tabiiy tilni qayta ishlashga oid masalalarni yechish uchun Python dasturlash tilida yaratilgan NLTK kutubxonasi yuklab olinishi talab etiladi [2].

NLP yo'nalishlari

- Mashina tarjimasi (Machine Translation)
- Nutqni tanish (Speech Recognition)
- Kayfiyat / tuyg'ularni tahlil qilish (Sentiment Analysis)
- Savollarga javob berish (Question Answering)
- Matnni referatlash (Summarization of Text)
- Chatbot (Chatbot)
- Intellektual tizimlar (Intelligent Systems)
- Matn tasniflari (Text Classifications)
- Belgilarni aniqlash (Character Recognition)
- Imlo tekshiruvi (Spell Checking)
- Spamni aniqlash (Spam Detection)
- (Matnni) Avtomatik to'ldirish (Autocomplete)
- Nomga ega obyektning aniqlash (Named Entity Recognition)
- Taklifli terish (Predictive Typing)

Sun'iy intellekt tizimi qabul qilingan ma'lumotlarni qayta ishlashi va Machine Learning usullarini qo'llash tufayli uning harakatlari uchun yaxshiroq bashoratlarni amalga oshirishi mumkin.

Machine Learning tizimga o'tgan tajriba va misollardan o'rganish qobiliyatini beradi. Umumiy algoritmlar dasturlashtirilgan narsaga ko'ra qat'iy belgilangan ijro to'plamini bajaradi va ular noma'lum muammolarni hal qilish imkoniyatiga ega emas. Haqiqiy dunyoda duch kelinadigan muammolarning aksariyati an'anaviy algoritmlarning samaradorligini pasaytiradigan ko'plab noma'lum o'zgaruvchilarni o'z ichiga oladi. Bu yerda Machine Learning birinchi o'ringa chiqadi. Yuqorida sanab o'tilgan yo'nalishlarning barchasi machine learning yordamida modellarni qurish uchun o'z ma'lumotlar bazasi – datasetlarni talab qiladi. O'zbek tili uchun ham bunday datasetlardan ancha qurilgan, masalan Sentiment tahlil datasetlari [3], [4], so'zlar o'xshashlik va bog'liqlik darajalarini o'z ichiga olgan semantik tahlil dataseti [5], hamda o'zbek tilidagi nomuhim so'zlar datasetlari [6] shular qatoriga kiradi.

Keltirilgan klassik misollardan ba'zilar spam pochta ni aniqlashni o'z ichiga oladi. Elektron pochta qonuniy yoki spam ekanligini aniqlash va tasniflash ko'plab noma'lumlarni o'z ichiga oladi. Spam filtrlarini chetlab o'tishning ko'plab usullari mavjud. An'anaviy algoritm ishlashi uchun har bir xususiyat va o'zgaruvchi

qattiq kodlangan bo'lishi kerak, bu juda qiyin, agar iloji bo'lsa. Holbuki, mashinali o'qitish algoritmi umumiy qoidani o'rganish va shakllantirish qobiliyati tufayli bunday muhitda ishlashi mumkin. O'zbek tili uchun ham machine learning yordamida ancha ishlar qilingan, ulardan gap bo'laklari(parts-of-speech)ni ajratib beruvchi dataset va dasturiy vositasi [7], O'zbek tilidagi nomuhim so'zlarni topish metodi [8], hamda gaplar o'zagini aniqlashda ishlatiladigan stemming dasturiy vositalari [9] shular jumlasidandir.

Deep Learning - bu Machine Learning algoritmlarining ixtisoslashuvi, sun'iy neyron tarmog'idir. So'nggi paytlarda Deep Learning usullari keng qo'llanilib, yaxshi natijalarga erishayotgani kuzatilmoqda.

Arxitektura bo'yicha qaror qabul qilishda Deep Learning texnikasi tomonidan taqdim etilgan moslashuvchanlik ushbu texnikalar muvaffaqiyatining muhim sabablaridan biridir. Deep Learning usullari tabiiy tilni qayta ishlash bo'yicha tadqiqotlar uchun ishlatiladigan Machine Learning usullari orasida birinchi o'rinda turadi.

Tabiiy tilni qayta ishlash foydalanuvchilarning murakkab elektronika bilan o'zaro aloqasini osonlashtirgan bo'lsa ham, boshqa tomondan, bu aloqani amalga oshirishga imkon beradigan sahna ortida ko'plab qayta ishlash jarayoni sodir bo'ladi. Mashinali o'qitish tilni qayta ishlashda juda muhim rol o'ynadi.

Machine Learning tabiiy tilni to'g'ri qayta ishlashda rol o'ynashdan tashqari, tabiiy tilni qayta ishlashning muhim ilovalarida ham juda konstruktiv rol o'ynaydi. Hissiy tahlil, Chatbot tizimlari, Savollarga javob berish tizimlari, Axborot qidirish tizimlari, Mashina tarjimai va Elektron pochta tasniflash kabi muhim tabiiy tilni qayta ishlash ilovalari yaxshi ishlashi uchun Machine Learning usullarini o'z ichiga oladi.

Tabiiy tilni qayta ishlashda mashinani o'rganishning o'rni

Mashina tabiiy tilni tushunishi uchun tabiiy tilni qayta ishlash ko'p bosqichlarni qamrab oladi. Ushbu bosqichlar morfologik tahlil, sintaktik tahlil, semantik tahlil, nutq tahlili va pragmatik tahlilni o'z ichiga oladi, odatda, bu tahlil vazifalari ketma-ket qo'llaniladi. Machine Learning deyarli barcha jarayonlarda u yoki bu shaklda muhim qo'shimcha qiymat sifatida ishlaydi. Hozirgi kunda o'zbek tilida NLP masalalarini yechish bo'yicha ko'plab ilmiy ishlar olib borilmoqda.

1. Morfologik tahlil:

Yuqorida aytib o'tilganidek, hisoblash tizimi tomonidan olingan ma'lumotlar 0 va 1 lar shaklida bo'ladi. Bu 0 va 1 larni ASCII kodi yordamida alifboga aylantirish mumkin. Shunday qilib, aytish mumkinki, mashina unga jumla yoki paragraf berilganda bir nechta belgilarni oladi. Morfologik tahlil darajasida birinchi vazifa so'z va gaplarni aniqlashdir. Ushbu identifikatsiya tokenizatsiya deb ataladi. Tokenizatsiya uchun ko'plab turli xil Machine learning va Deep Learning algoritmlari, shu jumladan Yordam vektor mashinasi (Support Vector Machine) va takroriy neyron tarmog'i (Recurrent Neural Network) qo'llanilgan.

Tokenizatsiya tugallangandan so'ng mashinada bir nechta so'zlar va jumlar mavjud bo'ladi. Tuzilgan gaplarning aksariyati affikslardan iborat. Ushbu affikslar mashinalar ishini murakkablashtiradi, chunki barcha so'zlarni barcha mumkin bo'lgan affikslari bilan o'z ichiga olgan so'z ma'nosi lug'atiga ega bo'lish deyarli mumkin emas. Demak, morfologik tahlil darajasining navbatdagi vazifasi bu affikslarni olib tashlashdir. Bu affikslarni o'zaklash yoki lemmatizatsiya yordamida olib tashlash mumkin. Tasodifiy o'rmon (random forest) va qaror daraxti (decision tree) kabi Machine Learning algoritmlari stemming vazifasini bajarishda juda muvaffaqiyatli bo'ldi.

2. Sintaktik tahlil

Tabiiy tilni qayta ishlashning navbatdagi vazifasi berilgan gap tilning grammatik qoidasiga amal qilishini tekshirishdan iborat. Buning uchun so'zlar birinchi navbatda so'z turkumlariga ajratiladi. Bu sintaktik tahlilchilarga grammatika qoidalarini tekshirishda yordam beradi. Tasodifiy o'rmon va takroriy neyron tarmog'i kabi mashinali o'qitish va chuqur o'qitish algoritmlari ushbu vazifa uchun muvaffaqiyatli qo'llanildi. Sintaktik parserlarni amalga oshirish uchun K- eng yaqin qo'shni (K- nearest neighbor) kabi mashinani o'rganish algoritmlari ham ishlatilgan. Bu sohada o'zbek tili uchun ham qilingan eng katta ishlardan biri bu sintaktik tahlil qilingan datasetning yaratilganidir [10].

3. Semantik tahlil

Bu bosqichda so'z ma'nolari izohli lug'atlar yordamida aniqlanadi. Bu yerda duch kelinadigan muammo shundaki, bir xil so'z jumla kontekstiga ko'ra turli xil ma'nolarga ega bo'lishi mumkin. Masalan, “o't” so'zi olov yoki “o't” o'simlik ma'nosini anglatishi mumkin, bu noaniqlikni keltirib chiqaradi. Shunday qilib, ushbu noaniqlikni bartaraf qilish so'z ma'nosini ajratish deb ataladigan tabiiy tilni qayta ishlashning ushbu bosqichidagi muhim vazifalardan biridir.

So'z ma'nosini aniqlash - bu turli darajadagi muvaffaqiyatlar bilan tadqiq qilingan klassik tasniflash muammolaridan biridir. Tasodifiy o'rmon, gradientni kuchaytirish va qarorlar daraxtlari kabi mashinali o'qitish usullari muvaffaqiyatli qo'llanildi. Biroq, so'nggi paytlarda takroriy neyron tarmog'i, uzoq qisqa muddatli xotiraga asoslangan takroriy neyron tarmoq, eshikli takrorlanuvchi birlikka asoslangan takroriy neyron tarmoq va konvolyutsiya neyron tarmog'i kabi chuqur o'qitish algoritmlari o'rganilgan va juda yaxshi natijalar bergan.

4. Diskurs tahlili

Tahlilning joriy ko'rinishidan tashqarida olmoshlar qo'llanilgan yoki ma'lum mavzular/obyektlarga havola qilingan holatlar mavjud. Bunday hollarda semantik tahlil gapga kerakli ma'no bera olmaydi. Bu mashinali o'qitish va chuqur o'qitish algoritmlari tomonidan hal qilingan mos yozuvlar yechimining yana bir klassik muammosidir.

5. Pragmatik tahlil

Ko'p vaqtlarda jumlar so'zlar tasvirlagandan ko'ra chuqurroq ma'noni anglatadi. Ya'ni, mashina semantik tahlildan so'ng tushunilgan so'z ma'nosidan voz

kechib, mo'ljallangan yoki nazarda tutilgan ma'noni egallashi kerak. Buni qilishdan ko'ra aytish osonroq. Ko'p yillar davomida bu tabiiy til jarayoni tadqiqotchilarni qiziqtirib kelmoqda. Pragmatik tahlilning klassik misollaridan biri kinoyani aniqlashdir.

Ko'pchilik, aslida deyarli barcha turli xil Machine learning va Deep learning algoritmlari kinoyani aniqlash yoki umuman pragmatik tahlilni amalga oshirish uchun turli xil muvaffaqiyatlar bilan qo'llanilgan.

Xulosa. Shunday qilib, Machine Learning va Deep Learning usullari tabiiy tilni qayta ishlash sohasida foydalanish uchun keng qamrovli tadqiq qilinayotganini kuzatish mumkin. Ko'rinib turibdiki, bu o'rganish usullari deyarli barcha tabiiy til vazifalarini qayta ishlashda, shuningdek, tabiiy tilni qayta ishlashning deyarli barcha ilovalarida muhim rol o'ynaydi.

Tabiiy til vazifalarini har xil qayta ishlash va tabiiy tilni qayta ishlashning turli xil ilovalari o'z-o'zidan tadqiqotning turli sohalaridir. Va hozirda ushbu tadqiqot sohalarida Machine Learning va Deep Learning usullari yuqori darajadagi muvaffaqiyat bilan keng ko'lamda o'rganilmoqda. Xulosa qilib aytish mumkinki, Machine Learning va Deep Learning texnikasi tabiiy tilni qayta ishlash va uning ilovalarida juda ijobiy rol o'ynamoqda.

Foydalanilgan adabiyotlar:

1. V. Garousi, S. Bauer, and M. Felderer, “NLP-assisted software testing: A systematic mapping of the literature,” *Information and Software Technology*, vol. 126. 2020. doi: 10.1016/j.infsof.2020.106321.
2. DataCamp, “spaCy,” Python Cheat Sheet, 2020.
3. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodriguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers, 2022*, pp. 232–243.
4. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings, 2022*, vol. 3315, pp. 126–136. [Online]. Available: www.scopus.com
5. U. Salaev, E. Kuriyozov, and C. Gómez-Rodriguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 2022*, pp. 199–206. [Online]. Available: www.scopus.com
6. K. Madatov, S. Bekchanov, and J. Vičič, “Lists of uzbek stopwords,” *Univerza na Primorskem, Inštitut Andrej Marušič*, 2021.



7. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, "UzbekTagger: The rule-based POS tagger for Uzbek language," arXiv preprint arXiv:2301.12711, 2023.
8. X. Madatov, M. Sharipov, and S. Bekchanov, "O'ZBEK TILI MATNLARIDAGI NOMUHIM SO'ZLAR," COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS, vol. 1, no. 1, 2021.
9. M. Sharipov and U. Salaev, "Uzbek affix finite state machine for stemming," arXiv preprint arXiv:2205.10078, 2022.
10. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, "Creating a morphological and syntactic tagged corpus for the Uzbek language," arXiv preprint arXiv:2210.15234, 2022.

KOREFERENSIYA VA ANAFORA TALQINI

Elova Dilrabo Qudratullayevna

PHD, ToshDO'TAU katta o'qituvchisi

elova@navoiy-uni.uz

Abdisalomova Shahlo Abdimurod qizi

ToshDO'TAU tayanch doktoranti

abdisalomovashahlo@gmail.com

Annotatsiya. Ushbu maqola NLPning muhim komponenti bo'lgan koreferensiya tadqiqiga bag'ishlangan. Unda koreferensiyaning mohiyati, o'ziga xos xususiyatlari tavsiflanadi, koreferensiya va anaforaning o'xshash va farqli jihatlari yoritiladi, koreferensiya tadqiqiga doir adabiyotlar tahlil qilinadi.

Kalit so'zlar: *NLP, kompyuter lingvistikasi, koreferensiya, anafora, referent, antisident, havola bo'lak.*

Abstract. This article is devoted to the study of coreference, an important component of NLP. The essence and specific features of coreference are described, the similar and different aspects of coreference and anaphora are accentuated, and the literature on the study of coreference is analyzed.

Keywords: *NLP, computational linguistics, coreference, anaphora, referent, antecedent, reference*

Абстракт. Данная статья посвящена изучению кореференции, важного компонента НЛП. Описаны сущность и особенности кореференции, выделены сходные и различные стороны кореференции и анафоры, проанализирована литература по изучению кореференции.

Ключевые слова: НЛП, компьютерная лингвистика, кореференция, анафора, референт, антисидент, референциальный фрагмент.

Tabiiy tilni qayta ishlash (NLP) – tabiiy til va nutqni hisoblash texnikasi yordamida tahlil va sintez qilish sohasidir. Tilning noaniqlik xususiyati tabiiy tilni qayta ishlash jarayonida qiyinchiliklarni yuzaga keltiradi. NLPning bir qator vazifalari ushbu muammolarni bartaraf etishga qaratilgan.

Koreferensiya tabiiy tilni qayta ishlash jarayonining tarkibiy qismi, matnni semantik tahlil qilishda muhim bo'lgan ishlov berish bosqichidir. Koreferensiya jahon tilshunosligi va kompyuter lingvistikasida keng tadqiq etilgan. Tilshunoslik nuqtayi nazaridan, B.Russell, J.Searle va K.Donnellan koreferensiyaga “bir xil referent haqida turli yo'llar bilan gapirish”, G.Frege “murojaat qilish usuli”, N.D.Arutyunova, V.Quine, Sh.A.Kripke, Y.V.Paduchevalar “referentga ishoraning aniqlik turlari” deya ta'rif bergan. J.Layonz, R.Bogrand asarlaridagi matn semantikasida koreferensiyaning ikki turi ko'rib chiqilgan: anaforik (anafora referentdan keyin keladi) va kataforik (anafora referentdan oldin keladi). Halliday va Hasanning “Cohesion in English” asarida koreferensiyaga matn uzviyligi –

kogeziyaning mustaqil vositasi sifatida qaraladi. Kognitiv lingvistikada esa “suhbatdoshlarning diqqat-e’tiborini rivojlantirish” deb tushuniladi [1].

Kompyuter lingvistikasida **koreferensiya** obyekt haqidagi matn ma’lumotlarini birlashtirishning asosiy texnologiyasi [2], matndagi bir xil obyektga ishora qiluvchi barcha havola bo’laklarni aniqlash jarayonidir [3]. U anafora tushunchasi bilan chambarchas bog’langan. Elango P. [2], Sukthanker R. [4], Singh S. [5] tadqiqotlarida koreferensiya va anafora terminlari ko’pincha bir-birining o’rnida ishlatilishi, koreferensiyaning qat’iy ta’rifiga amal qilinmaganligi ta’kidlanadi hamda koreferensiya va anaforani hal etish vazifasining chegarasini belgilashga urg’u beriladi. Dmitrey Kiselyovning [6] anafora va koreferensiya munosabatlariga bag’ishlangan maqolasida ushbu hodisalarning tafovutlari ochib berilgan. Unga ko’ra, anafora til hodisasi, matn qismlari o’rtasidagi sinsemantik, assimetriya, qaytmaslik munosabatlari bilan tavsiflanadi, aynan korrelyatsiya jarayoni; koreferensiya tildan tashqaridagi hodisa, avtosemantik, simmetriya munosabatlarini qamrab oladi. Yuqoridagi fikrlarga tayanib aytish mumkinki, koreferensiya *matn tarkibidagi barcha havola bo’laklarni ekvivalent sinflarga guruhlash vazifasidir*. Bunda ma’lum bir sinf ichidagi havola bo’laklar bitta nutq obyektiga tegishli bo’ladi. Quyidagi namuna asosida koreferensiya hodisasini tahlil qilamiz:

Hoshimni₁ taniysizmi? Ro’zivoy akaning sartaresh o’g’lini₂ tanimaysizmi? Yo’q taniysiz. Hozir esingizga tushiraman. Uning₃ to’g’risida To’xtaboyev degan yozuvchi kitob ham yozgan.

Yuqorida indekslangan so’zlardan *Hoshim* – nutq obyekti (referent), *Ro’zivoy akaning sartaresh o’g’li, u* birliklari referentga turli ko’rinishda murojaat qilayotgan havola bo’laklar (koreferentlar)dir. Bunda barcha koreferentlar bitta ekvivalentlik guruhini tashkil etmoqda.

Anafora – etimologiyasi *ana* (yun. “orqa”) va *pheri* (yun. “olib ketmoq”) bo’lib, “takrorlash” degan ma’noni bildiradi. Kompyuter tilshunosligida anafora nutqda ilgari aytib o’tilgan narsalarga havola yoki Mitkov ta’riflaganidek [7], “orqaga ishora” sifatida izohlanadi. Anafora – mohiyati o’zidan oldingi boshqa so’zga bog’liq bo’lgan ifoda va antisidentni topish vazifasi. Bunda **anafora** – antisidentga ishora qiluvchi so’z, **antisident** – anafora nazarda tutilgan shaxsdir. Koreferensiya esa matndagi qaysi (yo’naltiruvchi) ot iboralari bitta nutq obyektiga ishora qilishini aniqlashga qaratilgan va shu asosida asosiy referent zanjirlarini yaratadi. Demak, anafora natijalari antisident-anafora juftliklari bo’lsa, koreferensiya natijalari bir xil nutq obyektiga tegishli bo’lgan har xil turdagi (yo’naltiruvchi olmoshlar va ularning oldingi so’zlari, qavm-qarindoshlik otlari, murojaat birliklari, kasb-hunarni atovchi so’zlar va b.) birliklar to’plamidir. Koreferensiya anaforaga nisbatan umumiy, kengroq tushunchani anglatadi, ya’ni koreferensiyaning hal etish jarayoni anaforani hal etishni ham o’z ichiga oladi.

Aravakash zerikdi, shekilli, astagina ashula boshladi. **Uning** (1) ashulasi ham xirgoyidan nariga o'tmaganidan keyin **Salti** ko'tarildi: – O'lmasxon, bormisiz? Durustroq aytsangiz-chi! **Qumri** **uning** (2) so'zini quvvatladi.

Ushbu parchada anafora jarayonini aniqlaymiz: bu yerda anafora natijasi **aravakash-uning** (1), **Salti-uning** (2) juftliklari bo'lib, **aravakash** va **Salti** so'zlari **antisidentlar**, **uning**(1) va **uning** (2) birliklari bu so'zlarga ishora qilgani bois **anaforalardir**. Muammo shundaki, matndagi **O'lmasxon** birligi ham **aravakashga** tegishli havola bo'lak, biroq anafora jarayoni uni aniqlashni e'tibordan chetda qoldiradi. Koreferensiya uchun esa matndagi referentga tegishli bo'lgan barcha havola bo'laklarni aniqlash muhimdir. Бодрова tadqiqotida [8] koreferensiyaning aniqlashning ikki bosqichi keltiriladi:

1. obyektga ishora qiluvchi barcha nomzodlarni belgilash;
2. klasterlash.

Shu asosida keltirilgan nutqiy parchadagi koreferensiyaning aniqlaymiz:

1. Obyektga ishora qiluvchi barcha nomzodlarni belgilash.

Ushbu bosqichda nomzodlar: olmoshlar, otlar, otli birikmalar aniqlab olinadi: **aravakash, uning** (1), **Salti, O'lmasxon, Qumri, uning** (2)

2. klasterlash. Bu bosqichda nomzodlar ekvivalentlik sinflariga ajratiladi:

aravakash-uning (1)-**O'lmasxon**

Salti – uning (2)

Qumri

Ko'rish mumkinki, nomzodlar 2 ta ekvivalentlik guruhini tashkil etdi. **Qumri** referenti esa klasterlash uchun mos nomzodlarga ega bo'lmagan sababli **singlton** hisoblanadi. **Singlton** – bu koreferenti mavjud bo'lmagan referentdir [9].

Ingliz, nemis, fransuz, ispan, arab, italyan, rus, turk va boshqa ko'plab tillarda koreferensiyaning aniqlash masalasi tadqiq etilgan. Bugungi kunda koreferensiyaning aniqlash tizimlari qo'lda yaratishga asoslangan usullardan neyron tarmoq tizimlariga [8, 9] o'tdi. Mitkov [7] anaforani hal qilish vazifasi uchun ushbu sohadagi eng dolzarb usullar va ta'riflarni bayon qilgan. Vincent Ng [2] koreferensiyaning hal qilish bo'yicha 20 yil davomida amalga oshirilgan tadqiqotlarni tahlil qilgan, yutuq va kamchiliklar muhokama qilingan. Hassan Haji Mohammadi [10] maqolasida mavjud korpuslar va baholash ko'rsatkichlari ko'rib chiqildi, asosiy algoritmlarning umumiy ko'rinishi taqdim etildi, fors tilidagi koreferensiya va olmoshlarni aniqlash tizimlari tekshirildi. Lata [11] koreferensiya tizimlari uchun barcha turdagi koreferentlikni aniqlash usullarini batafsil ko'rib chiqqan. Zhekovaning [12] ilmiy ishida ko'p tilli koreferensiyaning aniqlash tizimi taklif etildi. Grobol og'zaki fransuz tilining koreferensiyaning hal etish tizimini yaratdi. Tolpegin rus tilidagi matnlarda uchinchi shaxs olmoshlari asosida koreferensiyaning aniqlash bo'yicha nomzodlik ishini himoya qilgan [13]. Tüfekci [14] ilmiy izlanishlarida turk tilida uchinchi shaxs olmoshlarini hal qilish uchun Hobbsning sodda yondashuvi va Xomskiyning tobelik nazariyasiga tayanilgan tizimni ishlab chiqqan.

Koreferensiya mashina tarjimasida, savollarga javob berish, matnning umumlashtirish, hissiyotlarni tahlil qilish, matnning tasniflash, nutqni tanish, nomlangan obyektni aniqlash, chatbot kabi NLP yo'nalishlariga tatbiq etiladi. Koreferensiyani hal etish uchun quyidagi bosqichlar bajarilgan bo'lishi lozim:



O'zbek tilshunosligida koreferensiya va anafora masalasiga oid tadqiqotlar ozchilikni tashkil etadi. Matn tilshunosligi va pragmatik lingvistik sohalarida koreferensiya, anafora xususida nazariy ma'lumotlar uchraydi. "Tilshunoslik terminlarining izohli lug'ati"da anaforik so'z haqida shunday deyilgan: "Oldingi so'z bildirgan shaxs yoki narsaga, oldin aytilganga ishora qiluvchi so'z" [15]. Bu yerda A.Hojiyev asosan olmoshlarni anaforik so'z sifatida talqin qilgan. Sh.Safarov monografiyasining [16] "Deyksis va nutqiy muloqot matni yoxud matnda mazmunning ifoda topishi" qismida anafora mavzusiga to'xtalgan. Olim anafora matndagi bog'lanishning bir turi, o'rin almashtirish vazifasi, kontekst deyksisi sifatida yondashadi, koreferentlikka nisbatan hamreferentlik atamasini qo'llaydi. Olimning fikricha, "o'rin almashtirish vazifasining bajarilishi nutqiy aktda yoki matnda koreferentlik (hamreferentlik) hodisasi mavjud bo'lishini talab qiladi", "anafora nutqiy tuzilma bo'laklarining mazmunan bog'lanishini ta'minlaydi, u koreferentlik haqida ma'lumot beruvchi lisoniy hodisadir. Tilshunoslar anafora hodisasiga tavsif berayotib, an'anaga binoan anaforik olmoshlarning matnning oldingi qatorlarida eslatilgan referentini izlashgan". Keltirilgan fikrlar Sh.Safarovning anafora masalasiga lingvistik yondashuvi jahon tilshunoslarining qarashlari bilan hamohang ekanligini ko'rsatmoqda. Koreferensiya masalasi **o'zbek kompyuter tilshunosligida, NLPda tadqiqot obyekti sifatida o'rganilmagan**. Demak, koreferensiyani hal etish masalasi bugungi o'zbek kompyuter tilshunosligining, NLPning dolzarb vazifasidir.

Foydalanilgan adabiyotlar:

1. <https://tproger.ru/articles/natural-language-processing/>
2. Ng V. Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research. / Vol.31 No.1 (2017): Thirty-First AAAI Conference on Artificial Intelligence, pp.4877-4884. – San Francisco, California USA / <https://doi.org/10.1609/aaai.v31i1.11149>.
3. Elango P. Coreference Resolution: A Survey. – University of Wisconsin, Madison, WI, pp.1-8, 2005.

4. Sukthanker R., Poria S., Cambria E., Thirunavukarasu R. Anaphora and coreference resolution: A review. / *Inf Fusion* 59, pp.139–162, 2020. / <https://doi.org/10.1016/j.inffus.2020.01.010>
5. Singh S., Bhattacharjee K., Darbari H., Verma S. Analyzing Coreference Tools for NLP Application. / *IJCSE Vol. 7, Issue 5*, pp.608-615, 2019.
6. Киселёв Д.А. К проблеме соотношения анафоры и (ко) референции. / *So‘z san’ati xalqaro jurnali*, № 4, pp.34-38, 2019.
7. Mitkov R. Anaphora resolution. – Routledge, 2014.
8. Бодрова Анастасия Александровна. Разрешение кореференции методом кластеризации / маг. дисс. – Санкт-Петербург, 2016. – 41 с.
9. Recasens M.P. Coreference: Theory, Annotation, Resolution and Evaluation. / PhD thesis. – Universitat de Barcelona, 2010. – 259 p.
10. Zhou M., Duan N., Liu S., Shum H.Y. Progress in Neural NLP: Modeling, Learning and Reasoning. In *Engineering*, 2020. / <https://doi.org/10.1016/j.eng.2019.12.014>
11. Aralikatte R. Towards Better Neural Coreference Resolution. / PhD Thesis. – University of Copenhagen, 2021. – 165 p.
12. Mohammadi Hassan Haji, Talebpour Alireza, Aznavah Ahmad Mahmoudi, Yazdani Samaneh. Review of coreference resolution in English and Persian. / *Computation and Language; Artificial Intelligence*, pp.1-44, 2022. / <https://doi.org/1048550/arXiv.2211.04428>
13. Lata K., Singh P. & Dutta K. Mention detection in coreference resolution: survey. / *Applied Intelligence* 52, pp.9816-9860 (2022). / <https://doi.org/10.1007/s10489-021-02878-2>
14. Zhekova D. Towards Multilingual Coreference Resolution. / PhD dissertation. – Languages and Literary Studies University of Bremen, 2013. – 333 p.
15. Толпегин П.В. Автоматическое разрешение кореференции местоимений третьего лица русскоязычных текстов. / дисс.кандидат технических наук. – Москва, 2008. – 241 с.
16. Tüfekci P. Türkçe için sözdizim tabanlı bir adıl çözümleme sistemi. / Doktora tezi. – Trakya Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Ana Bilim Dalı, 2009. – 137 s.
17. Ҳоҷиев А. Тилшунослик терминларининг изохли луғати. – Тошкент: Ўзбекистон миллий энциклопедияси, 2002. – 168 б.
18. Сафаров Ш. Прагмалингвистика. / Монография. – Тошкент, 2008. – 366 б.

DATA MINING TECHNIQUES IN NATURAL LANGUAGE PROCESSING: A REVIEW

Qurbonova Ro'zikajon Ulug'bek qizi

ruzikajonqurbanova0809@gmail.com,

Ro'zimboyeva Sevara Nurmat qizi

rozimboyevasevara@gmail.com

Master students at Urgench State University

Annotatsiya. So'nggi yillarda ijtimoiy tarmoqlar, bloglar va yangilik maqolalari kabi turli manbalardan olingan matnli ma'lumotlar miqdorida katta o'sish sodir bo'ldi. Ma'lumotlarning ko'payishi bilan mazmunli tushunchalarni olish uchun samarali va tez tabiiy tilni qayta ishlash (NLP) usullariga ehtiyoj paydo bo'ladi. Katta hajmdagi ma'lumotlar to'plamlaridan namunalar va bilimlarni olishni o'z ichiga olgan ma'lumotlar ilmi texnikasi NLP ilovalarida foydali ekanligini isbotladi. Ushbu maqolada biz NLP vazifalariga qo'llaniladigan turli xil ma'lumotlarni yig'ish usullarini ko'rib chiqamiz, jumladan matn tasnifi (klassifikatsiya), hissiyotlarni tahlil qilish (sentiment), obyektni tanib olish (NER) va mavzuni modellashtirish. Biz har bir texnikaning afzalliklari va cheklovlarini muhokama qilamiz va NLPda ma'lumotlar qazib olishdan foydalanish bilan bog'liq ba'zi qiyinchiliklarni ta'kidlaymiz. Va nihoyat, biz ushbu sohadagi kelajakdagi tadqiqot yo'nalishlari haqida ba'zi fikrlarni taqdim etamiz.

Kalit so'zlar: *Tabiiy tilni qayta ishlash, ma'lumotlar ilmi, katta hajmdagi ma'lumotlar.*

Аннотация. В последние годы произошел взрывной рост количества текстовых данных, генерируемых из различных источников, таких как социальные сети, блоги и новостные статьи. С этим увеличением данных возникает потребность в эффективных и действенных методах обработки естественного языка (NLP) для извлечения осмысленной информации. Методы интеллектуального анализа данных, которые включают извлечение шаблонов и знаний из больших наборов данных, оказались полезными в приложениях НЛП. В этой статье мы рассмотрим различные методы интеллектуального анализа данных, которые применялись к задачам НЛП, включая классификацию текста, анализ настроений, распознавание сущностей и тематическое моделирование. Мы обсудим преимущества и ограничения каждого метода и выделим некоторые проблемы, связанные с использованием интеллектуального анализа данных в НЛП. Наконец, мы даем некоторое представление о будущих направлениях исследований в этой области.

Ключевые слова: *Обработка естественного языка, интеллектуальный анализ данных, большие данные.*

Abstract. In recent years, there has been an explosion in the amount of textual data generated from various sources, such as social media, blogs, and news articles.



With this increase in data comes the need for efficient and effective natural language processing (NLP) techniques to extract meaningful insights. Data mining techniques, which involve the extraction of patterns and knowledge from large data sets, have proven to be useful in NLP applications. In this paper, we review various data mining techniques that have been applied to NLP tasks, including text classification, sentiment analysis, entity recognition, and topic modeling. We discuss the advantages and limitations of each technique, and highlight some of the challenges associated with using data mining in NLP. Finally, we provide some insights into future research directions in this area.

Keywords: *Natural Language Processing, Data Mining, Big Data*

1. Introduction.

The vast amounts of textual data available in today's world present both an opportunity and a challenge for researchers and practitioners in the field of natural language processing (NLP) [1]. On one hand, this data can be leveraged to gain valuable insights into human behavior, sentiment, and trends. On the other hand, the sheer volume of data can make it difficult to extract meaningful information from it. Data mining techniques, which have been used extensively in other domains such as marketing and finance, offer a promising approach for tackling this challenge in NLP [2]–[4].

In recent years, NLP has become an increasingly important field due to the explosion of textual data generated from various sources, such as social media, blogs, and news articles [5]. The challenge for NLP researchers and practitioners is how to efficiently and effectively extract meaningful insights from this vast amount of data. Data mining techniques offer a promising approach to tackling this challenge [6]. Data mining is the process of extracting useful and relevant information from large datasets. It involves using advanced computational techniques and algorithms to identify patterns, trends, and relationships within data. The goal of data mining is to extract insights and knowledge from data that can be used for decision-making, prediction, and optimization. Data mining techniques can be applied to various domains such as finance, healthcare, marketing, and engineering [7], [8]. The widespread availability of data and advancements in computing power have led to the growth of data mining as a field, and it has become an essential tool for businesses and researchers alike.

2. Literature Review.

In this section, we review various data mining techniques that have been applied to NLP tasks. There has been a significant amount of research in the area of data mining techniques in NLP. In this section, we review some of the relevant work in this area.

Text classification is one of the most extensively studied NLP tasks. A study conducted by Pang and Lee [9] compared various machine learning algorithms for

text classification and found that SVM performed the best. Another study by Wang et al. [10] proposed a novel feature selection method for text classification based on mutual information, which improved the performance of the classification algorithm.

Sentiment analysis has also received significant attention from researchers. A study conducted by Jurek et al. [11] proposed a lexicon-based approach for sentiment analysis, which achieved high accuracy on several benchmark datasets. Another study by Matlatipov et al. [12], [13] proposed a sentiment analysis data and models specific for Uzbek language, which achieved high accuracy for identifying the sentiment polarity of different aspects of a product.

Entity recognition is another important NLP task that has been extensively studied. A study by Finkel et al. [14] proposed a conditional random field model for named entity recognition, which achieved state-of-the-art performance on several benchmark datasets.

Furthermore, there has been a recent rapid increase in the development of NLP resources and tools specifically for Uzbek language. These include a tool for removal of stopwords [15], [16], Latin-Cyrillic transliteration [17] and text classification [18], as well as tools for stemming [19], [20], lemmatization [21].

3. NLP tasks and Data Mining Techniques.

Text classification is the process of categorizing documents into predefined classes based on their content. This technique has been widely used in many NLP applications such as email spam filtering, sentiment analysis, and language identification. Machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and Random Forests have been applied to text classification tasks with promising results.

Sentiment analysis is the process of identifying the sentiment expressed in a piece of text, such as positive or negative. Sentiment analysis has many applications such as product reviews analysis, customer feedback analysis, and political sentiment analysis. Machine learning algorithms such as Logistic Regression, SVM, and Neural Networks have been applied to sentiment analysis tasks with high accuracy.

Named Entity recognition (NER) is the process of identifying and extracting named entities from text, such as people, places, and organizations. Entity recognition has many applications such as social network analysis, news article analysis, and customer feedback analysis. Machine learning algorithms such as Conditional Random Fields (CRF), Hidden Markov Models (HMM), and Deep Learning models have been applied to entity recognition tasks with high accuracy.

Topic modeling is the process of identifying the underlying themes or topics in a corpus of text. Topic modeling has many applications such as text summarization, document clustering, and trend analysis. Machine learning algorithms such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) have been applied to topic modeling tasks with high accuracy.

Despite the promising results of data mining techniques in NLP, there are some challenges associated with using these techniques in practice. One of the main challenges is the need for large amounts of labeled data for training machine learning algorithms. Another challenge is the difficulty of interpreting the results of some techniques, such as neural networks and deep learning models.

4. Conclusion and Future Work.

In conclusion, data mining techniques offer a promising approach for addressing the challenges of extracting meaningful insights from large amounts of textual data in NLP. These techniques can be applied to various NLP tasks such as text classification, sentiment analysis, entity recognition, and topic modeling. However, there is still much work to be done in terms of developing more effective and efficient techniques, as well as addressing some of the challenges associated with using these techniques in practice.

In the future, we expect to see continued growth and development in the field of data mining techniques in NLP. One area that is likely to see significant advancement is deep learning, which has shown promise in many NLP tasks. Furthermore, with the increasing amount of data being generated in various domains, there will be a need for more efficient and scalable data mining techniques. Additionally, we expect to see more research on combining multiple NLP tasks and applying data mining techniques to solve complex problems. Finally, the ethical implications of data mining in NLP will need to be carefully considered, including issues such as privacy, bias, and fairness. Overall, the future of data mining techniques in NLP looks promising, and we anticipate many exciting advancements in this field.

References

1. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings*, 2022, pp. 199–206. [Online]. Available: www.scopus.com
2. M. Jamolbek Maqsudovich, “Clustering Class Association Rules to form a Meaningful and Accurate Classifier: doctoral dissertation,” Univerza na Primorskem, Fakulteta za matematiko, naravoslovje in..., 2020.
3. M. Jamolbek and M. Sanatbek, “Extracting the hidden regularities on latent features by using interval methods in pattern recognition problems,” *European science review*, no. 5–6, pp. 22–23, 2016.
4. J. Mattiev and B. Kavšek, “CMAC: clustering class association rules to form a compact and meaningful associative classifier,” in *Machine Learning*,

Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part I 6, 2020, pp. 372–384.

5. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.

6. J. Mattiev and B. Kavšek, “Distance based clustering of class association rules to build a compact, accurate and descriptive classifier,” *Computer Science and Information Systems*, vol. 18, no. 3, pp. 791–811, 2021.

7. J. Mattiev and B. Kavšek, “ACHC: Associative Classifier Based on Hierarchical Clustering,” in *Intelligent Data Engineering and Automated Learning–IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings* 22, 2021, pp. 560–571.

8. J. Mattiev and B. Kavšek, “Coverage-based classification using association rule mining,” *Applied Sciences*, vol. 10, no. 20, p. 7013, 2020.

9. B. Pang, L. Lee, and others, “Opinion mining and sentiment analysis,” *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

10. B. Wang, Y. Huang, W. Yang, and X. Li, “Short text classification based on strong feature thesaurus,” *Journal of Zhejiang University SCIENCE C*, vol. 13, no. 9, pp. 649–659, 2012.

11. A. Jurek, M. D. Mulvenna, and Y. Bi, “Improved lexicon-based sentiment analysis for social media analytics,” *Secur Inform*, vol. 4, no. 1, pp. 1–13, 2015.

12. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com

13. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243.

14. J. R. Finkel and C. D. Manning, “Nested named entity recognition,” in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 141–150.

15. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.

16. K. Madatov, S. Bekchanov, and J. Vičič, “Lists of uzbek stopwords,” Univerza na Primorskem, Inštitut Andrej Marušič, 2021.

17. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com



18. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.
19. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” *arXiv preprint arXiv:2205.10078*, 2022.
20. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.
21. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” *arXiv preprint arXiv:2210.16006*, 2022

SO‘Z MA’NOSINI ANIQLASHDA MASHINALI O‘QITISH YONDASHUVLARI

Axmedova Xolisxon Ilxomovna

Alisher Navoiy nomidagi Toshkent davlat
o'zbek tili va adabiyoti universiteti tayanch doktoranti
xolisa0629@gmail.com

Mirzaxmedova Parizoda

Alisher Navoiy nomidagi Toshkent davlat
o'zbek tili va adabiyoti universiteti bakalavr talabasi
mpari0621@gmail.com

Annotatsiya. Tabiiy tilni qayta ishlash jarayonidagi ko'plab masalalar bugungi kunda o'z yechimini topmoqda. So'z ma'nosini aniqlashning muhim vazifalaridan biri omonimiyani aniqlash masalasidir. O'zbek tilidagi omonim so'zlarning gapdagi aynan ma'nosini aniqlash usullari bo'yicha ma'lumot berish maqolamizning asosiy vazifasi. Ushbu maqolada bir so'z turkumi doirasida omonimlik hosil qiluvchi so'zlarning gapdagi ma'nosini aniqlashda mashinali o'qitish usulidan foydalanish va uning turlari, vazifalari haqida ma'lumot berilgan.

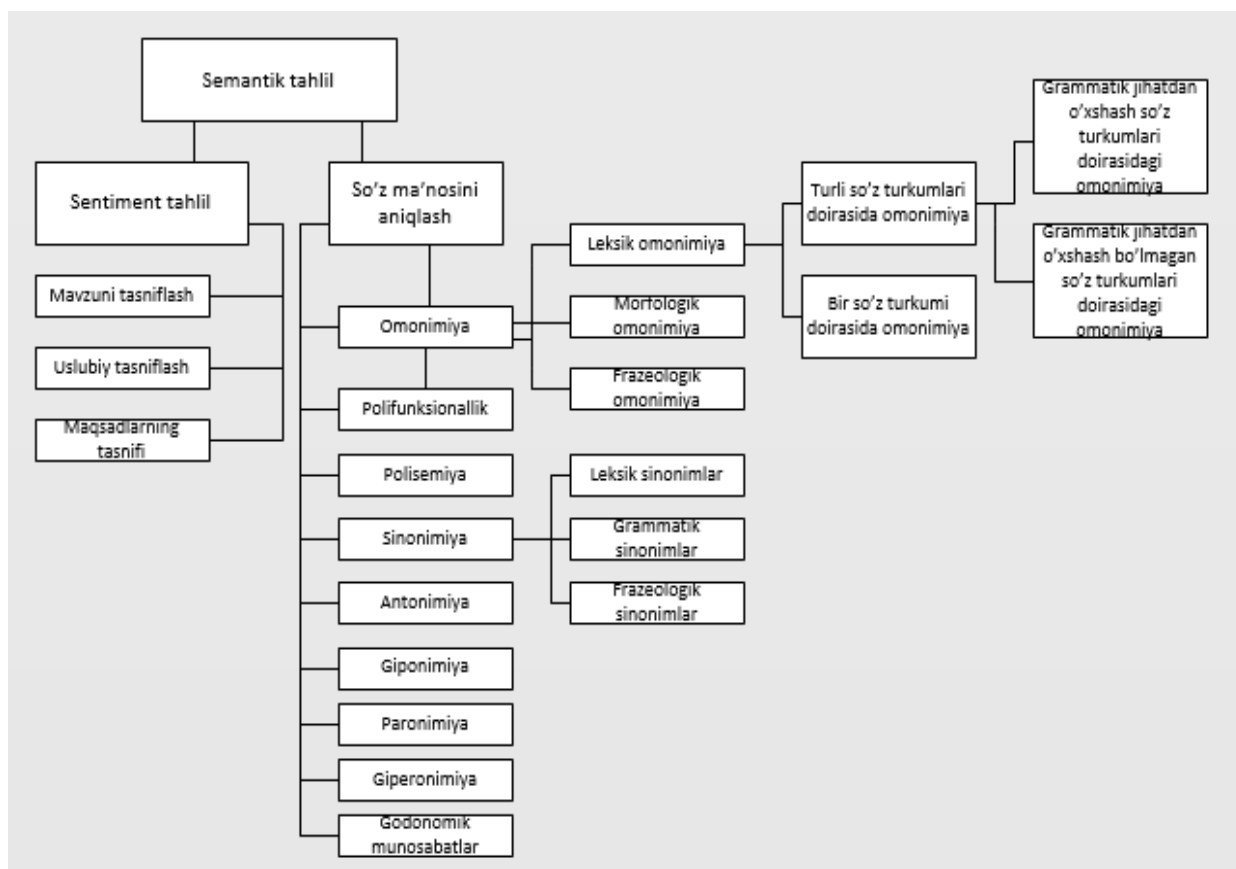
Kalit so'zlar. *NLP, semantik tahlil, omonimiya, WSD, mashinali o'qitish usuli, nazorat ostidagi va nazoratsiz usullar.*

Abstract. Many problems in natural language processing are being solved today. One of the important tasks of determining the meaning of a word is the issue of determining homonymy. The main task of our article is to provide information on the methods of determining the exact meaning of homonyms in the Uzbek language. This article provides information on the use of the machine learning method and its types and tasks in determining the meaning of homonymous words in a sentence.

Key words. *NLP, semantic analysis, homonymy, WSD, machine learning method, supervised and unsupervised methods*

Аннотация: В настоящее время решается множество задач обработки естественного языка. Одной из важных задач определения значения слова является вопрос определения омонимии. Основная задача нашей статьи – предоставить информацию о способах определения точного значения омонимов в узбекском языке. В данной статье представлена информация об использовании метода машинного обучения и его видах и задачах при определении значения омонимичных слов в предложении.

Ключевые слова. *НЛП, семантический анализ, омонимия, WSD, метод машинного обучения, контролируемые и неконтролируемые методы*



Tabiiy tilni qayta ishlash (Natural Language Processing (NLP)) – bu sun’iy intellekt va matematik tilshunoslikning umumiy yo‘nalishi bo‘lib, tabiiy tillardagi gaplarni kompyuterda analiz va sintez qilish muammolarini o‘rganadi. Ushbu muammoni hal qilish inson va kompyuter o‘rtasida o‘zaro aloqaning yanada qulay shaklini yaratishni anglatadi. NLP sohasining muhim vazifalaridan biri bu gaplarni semantik tahlil qilish bo‘lib, semantik tahlil ham o‘z navbatida 2 ta guruhga bo‘linadi.

So‘z ma’nosini aniqlash. So'z ma'nosini aniqlash (WSD) usullari kirishning semantik talqinini talab qiladigan ko'plab NLP vazifalari uchun foydalidir. Bundan tashqari, bunday usullar leksikografik tadqiqotlar va tilni o'rganish resurslari uchun muhim bo'lgan turli xil korpuslardagi so'zlarni turli ma'nolarini aniqlash chastotalarini baholashga yordam beradi. Ma'lumotlarga ko'ra, xorijiy tabiiy tillardagi omonimiyani aniqlashda *yarim nazorat ostida* (semi-supervised) so'z ma'nosini aniqlashni amalga oshiradigan har tomonlama baholangan usul yo'q.

Sentiment tahlil qilish- qayta ishlash texnologiyalari(NLP) va Informationsion Texnologiyalar(IT)sohasida yaxshi natijalarga erishilmoqda. Ushbu ilmiy sohaning eng qiziqarli va samarali usullari orasida sentiment tahliliajralib turibdi. Sentiment tahlil qilish kompyuter lingvistikasi uchun samarali tadqiqot yo‘nalishlaridan hisoblanadi.Dastlab, sentiment nima o‘zi degan savolga javob bersak. Bu-matndagi

emotsional rangdagi soʻz boyligini avtomatlashtirilgan tarzda aniqlash uchun ishlatiladi. Sentiment tahlil yana quyidagicha nomlanadi:

- Fikrni izlash (Opinion extraction)
- Fikrni oʻrganish (Opinion mining)
- Hissiy tahlil qilish (Sentiment analysis)
- Subyektiv tahlili (Subjectivity analysis)

Xissiyotlarni tahlil qilish va oʻrganish – bu yozma tildan odamlarning fikrlari, baholari, munosabati va hissiyotlarni tahlil qiladigan soha boʻlib dunyo boʻylab sentiment tahlil deb nomlanadi.

Ushbu maqolamizda soʻz manosi aniqlash boʻyicha jahon tajribasida qoʻllanilgan usullar haqida soʻz boradi. Jahon tajribasi shuni koʻrsatadiki, soʻz maʼnosini aniqlashda usullarini quyidagi guruhlariga ajratish mumkin.

- Qoidalarga asoslangan usul
- Statistik maʼlumotlarga asoslangan usul
- Mashinali oʻqitishga asoslangan usul
- Neyron tarmoq

Usullardan foydalanish samarali ekanligi tajribalardan maʼlum [1],[2], [3],[4]. Navbatdagi dozarb masala gapda uchragan bir soʻz turkumi doirasidagi omonim soʻzlarni semantik farqlash.

1-jadval. Bir va turli soʻz turkumlari doirasidagi omonim soʻzlar va ularning maʼnolari

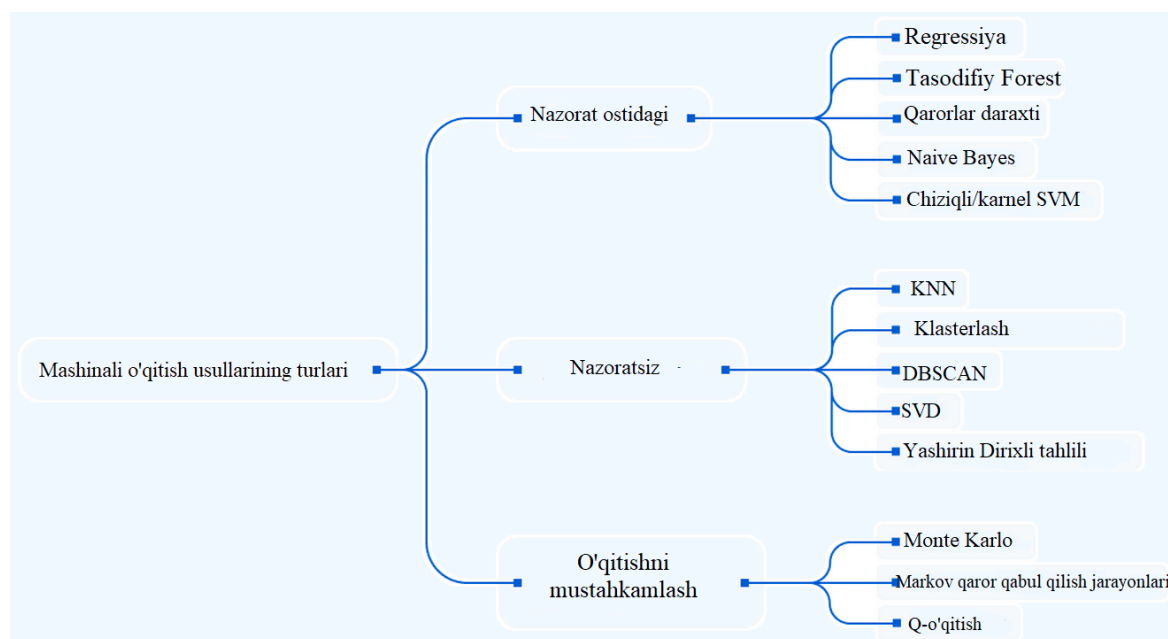
Soʻz	Soʻz turkumi	Maʼnosi
Oʻt	Feʼl	Oʻtmoq feʼli
	Ot	Maysa, oʻt-oʻlan
	Ot	Inson oʻrgani
	Ot	Olov
Oz	Ravish	Kam, miqdori nisbatan koʻp boʻlmagan.
	Feʼl	Oriqlamoq, etidan yoʻqotmoq.
	Feʼl	Noxush boʻlmoq, kuchsizlanmoq, holsizlanmoq.
	Feʼl	Adashmoq, toʻgʻri yoʻldan chetga chiqmoq
Boʻy	Ot	Uzunlik oʻlchovi.
	Ot	Hid, is
...	...	...

1-jadvaldagi omonim soʻzlarning soʻz turkumlarini aniqlash masalasini *qoidalarga asoslangan* hamda *statistik maʼlumotlarga asoslangan usullar* yordamida aniqlash tajribalaridan maʼlum. Navbatdagi masala bir soʻz turkumidagi omonim soʻzning maʼnosini aniqlash.

Masalan:

1. *Bemorning oʻtida tosh borligi aniqlandi*
2. *Qirdagi oʻtlarning yam-yashilligidan insonni quvontiradi.*
3. *Bolaxonadagi somonga ketgan oʻt oʻchirildi.*

Ushbu keltirilgan gaplardagi oʻt soʻzi har bir gapda turli maʼnolarni anglatmoqda. Bu omonim soʻzning gapdagi aynan maʼnosini aniqlash masalasi kunning dolzarb masalalaridan biridir. Bir soʻz turkumidagi omonim soʻzlarni maʼnolarini aniqlash masalasini yechishda mashinali oʻqitish usuli va uning algoritmlaridan foydalanish haqida soʻz yuritsak. Mashinali oʻqitish usullarini 1-rasmda keltirilgani kabi turlarga ajratish mumkin.



1-rasm: Mashinali oʻqitish usulining turlari

1-rasmda keltirilgan usullarga taʼrif beramiz.

Nazorat ostida oʻrganish - bu maʼlumotlarni tasniflash va natijalarni bashorat qilish algoritmlarini oʻrgatish uchun belgilangan maʼlumotlar toʻplamidan foydalanish orqali aniqlanadigan mashinali oʻqitish usuli. Oldindan belgialb qoʻyilgan maʼlumotlar toʻplamidan nimani qidirish kerakligini tushunish uchun mashina kirish maʼlumotlariga mos keladigan natija chiqadi. Nazorat ostidagi oʻqitishda foydali boʻlgan ikkita asosiy yoʻnalish mavjud: *tasniflash muammolari* va *regressiya muammolari*.

Tasniflash–kirish qiymatini olish va uni diskret qiymatga aylantirishni anglatadi. Tasniflash parametrlarini aniqlash orqali omonimiyani aniqlashda Naïve Bayes usulidan foydalanish masalasini [7] da ko‘rib chiqqan edik.

Regressiya–uzluksiz ma'lumotlar (qiymat funktsiyalari) bilan bog'liq. Regressiyada bashorat qilingan so‘zning ma’nosining miqdor ko‘rinishi haqiqiy sonlardan iborat bo‘ladi.

Nazoratsiz o‘qitish modellari–nazorat ostidagi o‘qitish modellariga qaraganda murakkabroq vazifalarni bajarishi mumkin. Ushbu yondashuvdan foydalanadigan asosiy vazifalar: klasterlash, uyushgan (association), hajmlarni qisqartirishdan iborat.

Klasterlash –bu nazoratsiz o‘qitish turi bo‘lib, ma'lumotlarda ularning o'xshashliklari yoki farqlari asosida yashirin farqlarni aniqlaydi. Ushbu farqlar omonim so‘zning ma’nolariga qarab turlicha bo‘lishi mumkin va ma'lumotlar elementlarini guruhlash yoki klasterlar yaratish uchun ishlatiladi.

Uyushgan yoki assotsiatsiya –bu ham nazoratsiz o‘qitish turi bo‘lib, unda so‘zning bir ma'nosi bilan uning boshqa ma’nolari orasidagi aloqasini, qisqa qilib aytganda semantik o‘xshashlinigi topish mumkin.

O‘qitishni mustahkamlash– nazorat ostidagi va nazoratsiz o‘qitish usullari yordamida olingan ma'lumotlarning aniqlik darajasini qayta tekshirishdan iborat.

Tajribalar 23 ta noaniq so'zlar, har biri oltita kontekstda (138 juft so'z) nisbatan nozik so'zlardan foydalangan holda to'g'ri ma'noni aniqladi

Xulosa o‘rnida shuni aytish mumkinki, omonim so‘zlarni aniqlashda qoidalarga asoslangan usullar, statistik ma'lumotlarga asoslangan usullardan foydanish natijasida olinga natijalarning aniqlik darajasi mos ravishda 75% , 82.3% ni tashkil etgan bo‘lsa, bu aniqlikni yanada yuqori darajaga yetkazish mashinali o‘qitish usullaridan foydalanish samarali natija berishiga ishonamiz.

Foydalanilgan adabiyotlar:

1. Elov B.B, Axmedova X.I. Uchta so‘z turkumi doirasidagi omonimiyani farqlovchi biznes jarayonni modellashtirish//O‘zbekiston respublikasi innovatsion rivojlanish vazirligining, ilm-fan va innovatsion rivojlanish ilmiy jurnal 2022 / 1, 150-161-b.
2. Ш.К.Гулямова, Х.И. Ахмедова. Омоним сўзларнинг лингвистик асоси, математик модели ва алгоритмлари (от ва феъл ҳамда сифат ва феъл туркуми доирасида)// Qo‘qonDPI. Ilmiy xabarlar,2021/1
3. Axmedova X. Ravish/fe‘l va ravish/yuklama so‘z turkumlari orasidagi omonimlikni farqlovchi matematik modellar // “Zamonaviy leksikografiya, til korpuslari va turkiy tillar platformalarini yaratish muammolari” nomdagi sho‘ba materiallari. Elektron nashr / ebook. – Toshkent: ToshDO‘TAU, 18.10.2021. – 452 b.
4. Axmedova X., Xusainova Z. Fe‘l/ko‘makchi va sifat/son so‘z turkumlari orasidagi omonimlikni farqlovchi matematik modellar //“Zamonaviy leksikografiya,



til korpuslari va turkiy tillar platformalarini yaratish muammolari” nomdagi sho‘ba materiallari. Elektron nashr / ebook. – Toshkent: ToshDO‘TAU, 18.10.2021. – 452 b.

5. Aarts, “Corpus linguistics: An appraisal.” In Hammesse, Jacqueline and Zampolli, Antonio (Eds.), Computers in Literary and Linguistic Research, Champion Slatkine, Paris-Genève: 1990/January, 13-28

6. Adriaens, Geert (1987). “WEP (word expert parsing) revised and applied to Dutch.” Proceedings of the 7th European Conference on Artificial Intelligence, ECAI’86, July 1986.

7. Elov B.B, Axmedova X.I. Polisemantik so‘zlarni semantik tahlil qilishning matematik modellar. O‘zbekiston: til va madaniyat (Amaliy filologiya), 2021, 1(4). 77-85-b.

8. Elov B.B, Axmedova X.I. Business Process Modeling That Distinguishes Homonymy Within Three Parts of Speeches in Uzbek Language // 7th International Conference on Computer Science and Engineering, 14-16 September 2022, Turkey. 289-294 -b.

O'ZBEK TILIDA STEMMINGNI AMALGA OSHIRISHNING GIBRID STATISTIK YONDASHUVI

Xusainova Zilola Yuldashevna,
ToshDO'TAU tayanch doktoranti,
xusainovazilola@navoiy-uni.uz.

Annotatsiya. Bir leksemadan nazariy jihatdan cheksiz ko'p so'z shakllarni yasash mumkin bo'lgan turk tili kabi aglutinativ tillar uchun o'zakni aniqlash murakkab va muhim masaladir. Bugungi kunda o'zbek tilidagi so'zlarning o'zagini aniqlashda ham analitik, ham statistik yondashuvlar qo'llanilgan. Ushbu yondashuvlar asosida o'zakni aniqlashda yaqqol namoyon bo'ladigan ikkita asosiy muammo – bu *yopiq lug'at* va *ko'plab nomzodlar orasidan haqiqiy o'zakning aniqlanish*. Maqolada statistik usullardan foydalangan holda so'zlardan qo'shimchalarni olib tashlash masalasi ko'rib chiqiladi. Maqolada keltirilgan algortim orqali amalga oshirilgan stemming usulning samaradorligi 97,5% ga teng.

Kalit so'zlar: *Stemming, NLP, affiks, agglyutinativ til,NER, neologizm, POS teglash.*

Abstract. For agglutinative languages, such as Turkish, where a single lexeme can theoretically form an infinite number of word forms, determining the root is a complex and important issue. Today, both analytical and statistical approaches are used to determine the root of words in the Uzbek language. The two main problems that are evident in stem detection based on these approaches are the closed vocabulary and the identification of the true stem from among many candidates. The article examines the problem of removing suffixes from words using statistical methods. The effectiveness of the stemming method implemented by the algorithm presented in the article is 97.5%.

Keywords: *Stemming, NLP, affixation, agglutinative language, NER, neologism, POS tagging.*

Аннотация. Для агглютинативных языков, таких как турецкий, где одна лексема теоретически может образовывать бесконечное количество словоформ, определение корня является сложной и важной проблемой. Сегодня для определения корня слов в узбекском языке используются как аналитический, так и статистический подходы. Две основные проблемы, которые очевидны при обнаружении основы на основе этих подходов, - это закрытый словарь и идентификация истинной основы среди множества кандидатов. В статье рассматривается проблема удаления суффиксов из слов с помощью статистических методов. Эффективность метода стемминга, реализованного по представленному в статье алгоритму, составляет 97,5%.

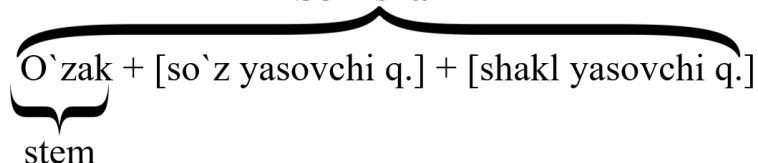
Ключевые слова: *Стемминг, НЛП, аффиксация, агглютинативный язык, NER, неологизм, POS-тегирование.*

Kirish

O'zbek tilida so'zning eng kichik ma'noli qismi ildiz (**root**) sifatida ta'riflangan bo'lsa, **stem** so'zga ma'no beruvchi eng katta qismidir. Shuning uchun so'zni ikki qismdan iborat deyishimiz mumkin: **ma'noni anglatuvchi stem** va **qo'shimchalar**.

Ingliz tili kabi ba'zi analitik tillarda bu atributlar **predloglar** kabi alohida so'zlar bilan belgilanadi va tabiatan juda oddiy. Ingliz tilida so'z cheklangan miqdordagi qo'shimchalarni (odatda bitta) olishi mumkin. Shu sababli, ingliz tili uchun asosiy stemming algoritmlari juda oddiy. Biroq, agglyutinativ tillarda bir ildizdan *son-sanoqsiz leksik shakllarining hosil bo'lishi*, tabiiy tilni tushunishda muhim va murakkab masala hisoblanadi. Agglyutinativ tillardagi ma'lumotlar bazasidan axborotni qidirish va tabiiy tilni tushunish kabi NLP vazifalari samaradorligini oshirish uchun stemming jarayoni muhim bosqich hisoblanadi[1].

So'z shakl



1-rasm. O'zbek tilidagi so'zning umumiy morfologik tuzilishi

O'zbek tilida gaplar alohida so'zlardan tashkil topadi. Morfologik jihatdan 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 [2]. Shakl yasovchi qo'shimchalar 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 (1-rasm).

So'zga qo'shila oladigan qo'shimchalar soni va ularning ko'p sonli birikmalari agglutinativ tillarda o'zakni aniqlash jarayonini murakkab muammoga aylantiradi. O'zakni aniqlash kontekstga bog'liq bo'lib, POS jarayoni noaniqlikni bartaraf qilishga yordam beradi.

Ushbu maqolada NLPdagi qoidalarga asoslangan va statistik yondashuvlardan foydalan holda "stem disambiguation" muammosini hal qilishga va oldingi ilmiy tadqiqotlarda yuzaga kelgan o'zaklarni aniqlash murakkabligini

kamaytirish haqida fikr mulohaza yuritiladi. Taklif qilinayotgan usul lug'atga bog'liq bo'lmaganligi sababli ishonchliligi va aniqligi yuqori hisoblanadi.

O'zbek tilida stemming usullari

Bugungi kunda Porter tomonidan taqdim etilgan algoritmnining turli xil ilovalari va modernizatsiyalangan variantlari ishlab chiqilgan bo'lib, ingliz tilidagi stemming uchun de-fakto standart algoritmgaga aylangan. Ushbu algoritmgaga qo'shimcha ravishda, tadqiqotchilar turli yondashuv va aniqlik darajasiga ega bo'lgan algoritmlarni ishlab chiqdilar:

- *Brute Force Algorithms;*
- *Suffix Stripping Algorithms;*
- *Lemmatization Algorithms;*
- *Stochastic Algorithms;*
- *Matching Algorithms.*

Shuningdek, hozirgi kunda boshqa tillarda stemming bo'yicha ko'plab tadqiqotlar olib borilgan [3,4,5,6].

Stemlash jarayonida analitik yoki statistik yondashuvlar qo'llanilsa, ushbu tadqiqotlarning metodologiyasi to'g'ridan-to'g'ri (labeled) deb nomlanadi. Morfologik tahlil yoki grammatik xususiyatlarni o'z ichiga olgan tadqiqotlar **analitik** deb tasniflanadi. Shuningdek, statistika/ehtimolga asoslangan usullar **statistik** deb ataladi.

NLP sohasi olimlarini agglyutinativ tillarda lug'at cheklanmagan degan fikrga kelishgan [7]. 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.

Ushbu maqolada o'zbek tilidagi so'zlardagi mumkin bo'lgan qo'shimchalar jadvalidan foydalangan holda stemming jarayonini amalga oshiruvchi yondashuv ko'rib chiqiladi. Stemming jarayoni natijasida xato hosil qilingan stem POS amali asosida to'g'ri shakl (root)ga aylantiriladi. O'zbek tilida POS amalini bajarishda 6 xil so'z turkumlari (*ot, fe'l, sifat, ravish, son, olmosh*) ko'rib chiqiladi.

O'zbek tilida so'z turkumiga mansub so'zlar turli qo'shimchalar bilan birikadi [7]. Shu sababli, fe'l va ot so'z turkumiga mansub so'zlardan qo'shimchalarni olib tashlash uchun ikkita chekli holatlar mashinasi (finite state machines, FSM) ishlab chiqiqish mumkin [8]. Ushbu FSMlarni tegishli qo'shimchalar shablonlari bilan qayta ishlash bir nechta mumkin bo'lgan ildizlarga olib kelishi mumkin.

O'zak va qo'shimcha chegaralarini aniqlash uchun o'zbek tili korpusidagi n-gramm statistikasi usulidan foydalanish mumkin [9]. Shunday qilib, taklif etilayotgan yondashuv analitik (qo'shimchalarni olib tashlash) va statistik (o'zakni aniqlash) usullarini qo'llaydigan gibrid usuldir (2-rasm).



2-rasm. Stemni aniqlash bosqichlari

Qo'shimchalarni olib tashlash

Morfologik tahlil sohasida hozirgacha bajarilgan ishlarni ikki sinfga birlashtirish mumkin: **stemga** va **affikslarni olib tashlashga asoslangan**. Stemga asoslangan morfologik tahlilni amalga oshirish uchun dastlab so'zning stemi, so'ngra affikslari aniqlanadi. Affikslarni olib tashlashga asoslangan yondashuvda dastlab so'zdagi affikslar aniqlanadi. Affikslar olib tashlangandan so'ng, so'zning qolgan qismini **stem** deb hisoblash mumkin yoki bu taxminni tasdiqlash uchun lug'atdan foydalanish mumkin.

So'zga qo'shila oladigan qo'shimchalar soni va ularning ko'p sonli birikmalari agglyutinativ tillarda o'zakni aniqlash jarayonini murakkab muammoga aylantiradi. Chunki ko'pchilik agglyutinativ 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 ketadi¹.

O'zbek tilida so'z yasovchi va shakl hosil qiluvchi qo'shimchalar o'rtasida omonimiya hodisasi uchraydi [10]. Bu esa stemming jarayonida muammoli vaziyatni yuzaga keltiradi. Quyidagi 1-jadvalda omonim qo'shimchalar ro'yxatidan namuna keltirilgan[11]:

1-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.)	kuchay (fe'l)
boray	
- gi (lug'aviy shakl yas.)	supurgi (ot), yozgi (sifat)
borgim	
- da (sintaktik shakl yas.)	undamoq (fe'l)
uyda	

1 <https://github.com/KhZilola/Python-Codes>

do'sti	- i (sintaktik shakl yas.)	jannati (sifat), boyi (fe'l)
ko'rin	- in (lug'aviy shakl yas.)	ekin (ot), sog'in (sifat)
uyim	- im (sintaktik shakl yas.)	bilim (ot), ayrim (sifat)
o'chir	- ir (lug'aviy shakl yas.)	gapir (fe'l)
siniqmoq	- iq (lug'aviy shakl yas.)	yo'liq (fe'l), ochiq (sifat), chiziq (ot)
o'qiy	- y (lug'aviy shakl yas.)	qoray (sifat)

O'zbek tilida o'zakka qo'shilishi mumkin bo'lgan **prefiks va suffiks qo'shimchalar** quyidagi 2,3-jadvallarda keltirilgan.

2-jadval. O'zbek tilidagi prefiks qo'shimchalari

Prefix qo'shimchalar		
ba-	be-	ham-
bo-	no-	bar-
bad-	ser-	be-
bar-	g'ayri-	xush-

3-jadval. O'zbek tilidagi suffiks qo'shimchalar (POS bo'yicha)

POS	so'z yasovchi (derevatsion)	shakl yasovchi (flektiv)
Ot (N)	51	17
Son (Num)	1	5
Fe'l (Vb)	30	22+10 (nisbat)
Sifat (JJ)	97	9
Olmosh (P)	-	-
Ravish (R)	18	5

O'zbek tili korpusidan har bir so'z turkumiga oid bir nechtdan (faol) so'z shakllari olinib, ularning stemmini so'z shaklidan o'chirib tashlash orqali, berilgan so'z turkumidagi stendlarga qo'shilishi mumkin bo'lgan **prefiks va suffiks**

qo'shimchalar satrlar to'plami aniqlandi. Suffiks qo'shimchalar satri bu so'z shaklini hosil qilish uchun stemga qo'shilgan o'zbek tilidagi bir nechta **qo'shimchalar birikmasi** hisoblanadi (4-jadval).

4-jadval. Til korpusidan aniqlangan o'zbek tilidagi prefiks va suffikslar birikmalari to'plami

	Ot (N)	Son (Num)	Fe'l (Vb)	Sifat (JJ)	Olmosh (P)	Ravish (R)
1.	chiligimizning	inchi	dilarki	almashtir gichda	nga	aydiki
2.	dagilarning	larcha	nganlaridan	bardoshli gini	ngacha	aydimi
3.	doshlarimizni kidan	larda	yapsizlarmi	imizning	ngagina	chiligim izning
4.	korlarimizning	ovi	ilmayotgani dan	lamagani dan	nikilarn ing	dangina
5.	laganlaridagina	ovimiz	ishimizdan	lashayotg anligini	dek	gilarini ng
...	...	...	...	...	...	...
Jami:	1773	229	15870	725	509	1376

Affikslarni olib tashlashga asoslangan yuqorida keltirilgan algoritim o'zbek tili ta'limiy korpusi (<http://uzschoolcorpara.uz/uz/Search>)dagi 100000 dan iborat gaplarga qo'llash natijasida **97.5%** aniqlikdagi natijani berdi.

Xulosa

POS teglash va stemlash uchun o'zbek 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 o'zbek tili korpusi asosida stemmingni amalga oshirishning gibrid statistik yondasuvi taqdim etildi.

Foydalanilgan adabiyotlar:

1. Б.Б. Элов, Ш.М. Хамроева, З.Ю. Хусайнова. NLPning zamonaviy algoritmlari va konsepsiyalari. - Международный журнал искусство слова, 42-576, 2023(1). <https://doi.org/10.5281/zenodo.7679911>
2. Ҳожиёв А. Ўзбек тилида сўз ясаши. – Ташкент, 2005.

3. Sharma, A., Kumar, R., & Mansotra, V. (2016). Proposed Stemming Algorithm for Hindi Information Retrieval. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization)*, 3297(6). <https://doi.org/10.15680/IJIRCCE.2016>
4. 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>
5. Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2).
6. 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>
7. B.Elov, Sh.Hamroyeva, O.Abdullayeva, Z.Xusainova, N.Xudayberganov. (2023). Agglyutinativ tillar uchun POS teglash va stemming masalasi (turk, uyg'ur, o'zbek tillari misolida). *Til va madaniyat, Kompyuter lingvistikasi. Vol.1(6)*. – B. 35-50.
8. Kışla, T., & Karaoğlu, B. (2016). A hybrid Statistical Approach to Stemming in Turkish: An Agglutinative Language. *Anadolu University Journal of Science and Technology-A Applied Sciences and Engineering*, 17(2). <https://doi.org/10.18038/btda.31812>
9. B.Elov, Sh.Hamroyeva, Z.Xusainova, M.Samatboyeva. *Tabiiy tilni qayta ishlash (NLP)da stemming jarayoni tavsifi (2023)*.
10. B.Elov, K.Akhmedova. A Mathematical Model That Semantically Analyzes Polysemantic Words. (2021). *Journal of Pedagogical Inventions and Practices. P. V3. 119-122*
11. B.Elov, M.Samatboyeva. Identifying NER (Named Entity Recognition) objects in Uzbek language texts. *Science and Innovation International Scientific Journal Volume 2 Issue 4 April 2023 UIF-2022: 8.2 | ISSN: 2181-3337. P.44-57*
12. Abdullayeva O.X. Korpus lingvistikasida lemmalashtirish, stemming va tokenlashtirish jarayoni / “O‘zMU xabarlari”, № 1/4/1 – Toshkent, 2021. –b. 246-249.
13. Abdullayeva O.X. Morphological annotation system in the corpus of internet information texts in Uzbek language // 7th International Conference on Computer Science and Engineering, 14-16 September 2022, Turkey160-164-b.

SO‘ZLARNI SEMANTIK TAHLIL QILISH ALGORITMLARI

Axmedova Xolisxon Ilomovna

Alisher Navoiy nomidagi Toshkent davlat
O‘zbek tili va adabiyoti universiteti
tayanch doktoranti

xolisa0629@gmail.com

Xushmuratova Madina

Alisher Navoiy nomidagi Toshkent davlat
O‘zbek tili va adabiyoti universiteti talabasi

madinaxushmurodova60@gmail.com

Annotatsiya: So‘z ma’nosini aniqlash (Word Sense Disambiguation (WSD))- tabiiy tilni qayta ishlash (NLP) sohasidagi muhim, ammo qiyin vazifadir. Yuzlab WSD algoritmlari va tizimlari mavjud, ammo optimal WSD algoritmlarini tanlash bo‘yicha kamroq ish qilingan. Ushbu maqola WSD uchun ishlatiladigan turli yondashuvlarni umumlashtiradi va mavjud WSD algoritmlarini ularning texnikasiga ko‘ra tasniflaydi. Undan tashqari nazorat ostida o‘qitish, nazoratsiz o‘qitish va yarim nazorat ostidagi o‘qitish algoritmlari muhokama qilindi.

Kalit so‘zlar: *So‘z ma’nosini aniqlash, Tabiiy tilni qayta ishlash (NLP), nazorat ostidagi yondoshuvlar, nazoratsiz yondoshuvlar, kontekst, noaniqlik, ko‘p ma’noli so‘zlar, omonim, Mashinada o‘qiladigan lug‘at va WordNet.*

Abstract: Word Sense Disambiguation is an important but challenging task in the field of Natural Language Processing (NLP). There are hundreds of WSD algorithms and systems, but less work has been done on selecting optimal WSD algorithms. This article summarizes the different approaches used for WSD and classifies existing WSD algorithms according to their techniques. In addition, supervised learning, unsupervised learning and semi-supervised learning algorithms were discussed

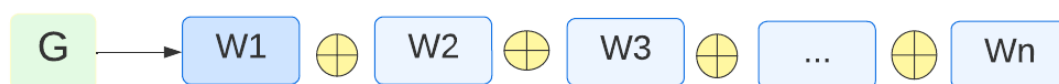
Keywords: *Word sense disambiguation, Natural language processing, supervised approaches, unsupervised approaches, semisupervised approaches, polysemantic words, homonymy, machine learning dictionary and WordNet.*

Аннотация: Устранение неоднозначности смысла слов является важной, но сложной задачей в области обработки естественного языка (NLP). Существуют сотни алгоритмов и систем WSD, но меньше работы было сделано по выбору оптимальных алгоритмов WSD. В этой статье обобщаются различные подходы, используемые для WSD, и классифицируются существующие алгоритмы WSD в соответствии с их методами. Кроме того, обсуждались алгоритмы обучения с учителем, обучения без учителя и полуправляемого обучения.

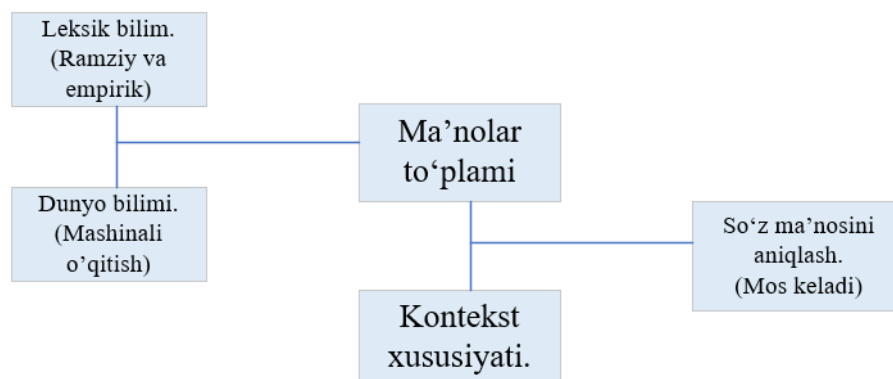
Ключевые слова: *устранение неоднозначности смысла слова, обработка естественного языка, контролируемые подходы,*

неконтролируемые подходы, полуконтролируемые подходы, многозначные слова, гомонимия, словарь машинного обучения и WordNet.

Kirish: Tabiiy tillarda turli kontekst uchun har xil ma'noga ega bo'lgan so'zlar mavjud, ammo ular bir xil talaffuz qilinadi. Bu so'zlar *ko'p ma'noli so'zlar* deyiladi. So'z ma'nosini aniqlash muammoning yechimidir [3]. So'z ma'nosini aniqlash - bu so'zlarning to'g'ri ma'nosini topish va ma'lum bir kontekstda ko'p ma'noli so'zlarga avtomatik ravishda uning to'g'ri manosi berish vazifasi. So'z ma'nosini aniqlash tabiiy tilni qayta ishlash sohasidagi muhim, ammo qiyin vazifadir[9]. Bu *mashina tarjimasi, semantik xaritalash, semantik izoh va ontologiyani o'rganish* kabi ko'plab ilovalari uchun zarur. Bundan tashqari *ma'lumot qidirish, ma'lumot olish va nutqni aniqlash* kabi ko'plab ilovalarning ishlashini yaxshilashda foydali bo'ladi. So'z ma'nosini aniqlashdagi muhim elementlar sifatida omonim so'zlar hamda polisemantik so'zlarni olish mumkin. Omonim so'zlar – *suzmoq, oy, ot, o't*, polisemantik so'zlar–*ko'z, etak, qo'l* kabi so'zlar ko'p ma'noli so'zlarga misol bo'la oladi. So'zning kontekstdagi ma'nosini to'g'ri aniqlash inson uchun juda oson. Ammo kompyuter uchun bu juda qiyin ish. Shunday qilib, biz odamlarga o'xshab ishlaydigan avtomatik tizimni ishlab chiqishimiz kerak, ya'ni ma'lum bir kontekstda so'zning to'g'ri ma'nosini topadigan va joriy so'zga avtomatik ravishda optimal ma'noni belgilaydigan tizimni ishlab chiqishimiz kerak. *Kontekst* - bu ko'p ma'noli so'zning birikuvchilaridan iborat bo'lgan matn yoki so'zlar. Kontekstdan foydalanib ushbu so'zning to'g'ri ma'nosini aniqlash inson uchun oson bo'lishi mumkin. Shunday qilib, biz kompyuterga ba'zi qoidalarni joylab chiqishimiz kerak, ular yordamida tizim so'zning bir nechta ma'nolaridan mutloq ma'noni baholashi mumkin. Agar kontekst W so'zlar ketma-ketligi deb hisoblasak, ya'ni $W_1, W_2, W_3, \dots, W_n$.



So'z ma'nosini aniqlash G matnidagi barcha yoki ba'zi so'zlar uchun to'g'ri ma'noni belgilash vazifasidir. So'z ma'nosini aniqlash uchun kontseptual model [4] 1-rasmda ko'rsatilgan.



1-rasm: So'z ma'nosini ajratib ko'rsatish kontseptual modeli.

So'z ma'nosini aniqlashga doir yondashuvlar: So'z ma'nosini aniqlash uchun ikkita yondashuv qo'llaniladi: *Mashinali o'qitishga asoslangan yondashuv* va *bilimga asoslangan yondashuv*. Mashinali o'qitishga asoslangan yondashuvda tizimlar so'z ma'nosini aniqlash vazifasini bajarishga o'rgatiladi. Bilimga asoslangan yondashuvda WordNet, lug'at, tezaurus va boshqalar kabi tashqi leksik resurslarni talab qiladi.

Mashinali o'qitishga asoslangan yondashuv: Klassifikator xususiyatlarni o'rganish uchun ishlatiladi va misollardagi noaniq ma'nolarni aniqlaydi. Mashinali o'qitish usuli yordamida *noaniqliklarni* aniqlash uchun so'zning har bir ma'nosini anglatuvchi kontekstlar zarur. Xususiyatning qiymati bu so'zning maqsadli so'zlar miqyosida necha marta sodir bo'lishidir. Mashinali o'qitishga asoslangan yondashuvlarning uchta turi mavjud:

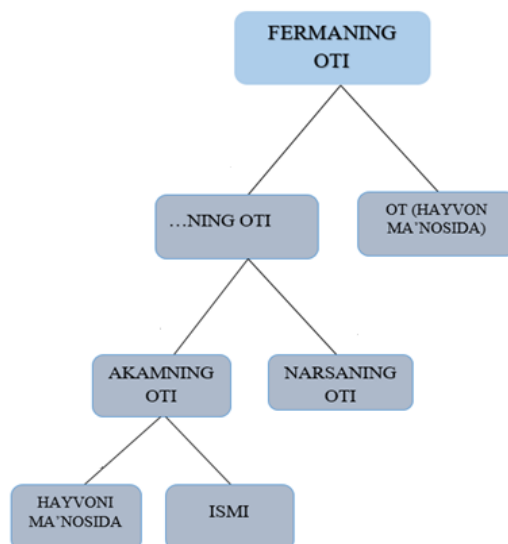
- nazorat ostidagi yondashuvlar
- nazoratsiz yondashuvlar
- yarim nazorat ostidagi yondashuvlar.

Nazorat ostidagi yondashuvlar: Ushbu yondashuv mutaxassislar tomonidan so'z ma'nolarini tasniflovchi parametrlardan foydalaniladi [6]. Tasniflagichni o'rganish uchun foydalaniladigan katta hajmdagi ma'lumotlar majmuasi odatda ma'lum bir maqsadli so'zga mos yozuvlar lug'atining ma'no inventarizatsiyasidan olingan ma'no bilan qo'lda teglangan misollar to'plamini o'z ichiga oladi. So'z ma'nosini aniqlovchi axborot tizimi ishlab chiqish uchun semantik teglangan misollar to'plamidan parametrlar o'rganiladi. Bizda misollar to'plami sifatida o'zbek tili milliy korpusini olishimiz mumkin. Asosan bu so'z ma'nosini aniqlash algoritmi boshqa yondashuvlarga qaraganda yaxshi natija beradi. So'z ma'nosini aniqlash nazorat ostidagi usullari quyidagilardan iborat:

Qaror ro'yxatlari: O'rganish uchun olingan korpus ma'lumotlari misollarini toifalarga ajratish uchun tartiblangan qoidalar to'plamidir. Buni vaznli [if-then-else] qoidalari ro'yxati sifatida ko'rish mumkin. Korpus ma'lumotlari majmuasi bir qator xususiyatlarni keltirib chiqarish uchun ishlatiladi. Har qanday so'z ko'rib chiqilganda, birinchi navbatda uning paydo bo'lishi hisoblab chiqiladi va uning xususiyat vektorini nuqtai nazaridan qarorlar ro'yxatini yaratish uchun ishlatiladi.

Vektordagi uchun eng yuqori ko'rsatkichga ega bo'lgan so'z ma'nosining vazni ma'noni ifodalaydi.

Qaror daraxti: Qaror daraxti trening ma'lumotlarini rekursiv usulda va qoidalarni ifodalaydigan daraxt tuzilishidagi tasnifini ajratadi[8]. 2-rasmda tasvirlangan "ot" omonim so'zining ma'nolari tasnifidan iborat qaror daraxtini keltirish mumkin.



2-rasm: Qarorlar daraxtiga misol

Naive Bayes: Naive Bayes klassifikatori Bayes teoremasi asoslangan ehtimollik tasniflagichi. U berilgan har bir w ko'p ma'noli so'zning har bir ma'nosini shartli ehtimolini hisoblashga tayanadi. f_n – kontekstdagi xususiyatlarni bildiradi. Quyidagi formulani maksimal darajaga keltiradigan S ma'nosi kontekstga eng mos ma'no sifatida tanlanadi.

$$\begin{aligned} \hat{S} &= \operatorname{argmax}_{S \in \text{ma'no}_D(w)} P(S_i | f_1, \dots, f_m) = \operatorname{argmax}_{S \in \text{ma'no}_D(w)} \frac{P(f_1, \dots, f_m | S_i)}{P(f_1, \dots, f_m)} = \\ &= \operatorname{argmax}_{S \in \text{ma'no}_D(w)} P(S_i) \neq P(f_j | S_i) \end{aligned}$$

Naive Bayes klassifikatoridan foydalanish uchun tasniflash parametrlarini aniqlab olish talab qilinadi. [13] da batafsil ko'rish mumkin.

Nazoratsiz yondashuvlar: Nazoratsiz yondashuvlarda nazorat ostidagi yondashuvdan farqli o'laroq, qo'l kerak emas. katta hajmdagi ma'no ma'lumotlarining izohlash uchun manbalarga asoslanadi. O'xshash ma'noga ega so'zlar o'xshash bo'ladi. So'z ma'nolari so'zlarning semantik klasterlarini hosil qilish orqali hosil bo'ladi va uning vazifasi olingan klasterlarga yangi hodisani tasniflashdan iboratdir. Ushbu yondashuv ma'no belgilarini belgilash o'rniga klasterlarni aniqlaydi.

Kontekstni klasterlash: Bu usulda birinchi kontekst vektorlari bo'lgan klasterlash usullariga asoslanadi [12]. Klasterlarga guruhlanadi va shu tariqa

so'zning ma'nosi aniqlanadi. Bu usuldan foydalanish so'z fazosi kabi vektor fazo va uning o'lchamlari faqat so'zlar.

So'zlarni klasterlash: Ushbu algoritmda o'xshash ma'noga ega so'zlar bir xil klasterga birlashtiriladi. Bittasi [10] da tilga olingan yondashuvlar maqsadli so'z bilan bir xil so'zlar qatorini topish edi. So'zlar orasidagi o'xshashlik sintaktik bog'liqlik orqali beriladi. Agar W w_m ga o'xshash so'zlardan iborat bo'lsa, dastlab faqat bitta tugun w_m bilan hosil bo'ladi va w_i tugunida w_i bo'lganda w_m tuguniga ega bo'ladi, w_m ga o'xshash ma'noga ega so'z ekanligi aniqlandi. Yordamchi ma'lumotlar va klasterdan muhim xususiyatlarni o'rganadigan turli xil usullar va yondashuvlarni klasterlash deb ataladigan yana bir yondashuv yoki olingan ma'lumotlardan foydalangan holda ma'lumotlarni izohlash. Maqsadli so'zlarga duch kelganda o'xshashlik matritsasi deb ataladigan matritsa S_{mn} tuziladi, uning har bir elementi ikkita w_m va w_n so'z o'rtasidagi o'xshashlikdir. Yakuniy bosqichda W ga tegishli har bir maqsadli so'z bo'ladi.

Yarim nazorat ostida o'rganish yondashuvi: Yarim nazorat ostida o'rganish yondashuvlari, ma'lumotlar kabi nazorat qilinadi, lekin kamroq ma'lumot berilishi mumkin. Yarim nazoratli yoki minimal nazorat ostidagi usullar qobiliyatlari tufayli mashhurlik kasb etmoqda. Ko'pincha katta ma'lumotlar to'plamlarida mutlaqo nazoratsiz usullardan ustun turadigan kichik miqdordagi izohli ma'lumotnoma malumotlariga ega bo'ling. Yordamchi ma'lumotlar va klasterdan muhim xususiyatlarni o'rganadigan turli xil usullar va yondashuvlar mavjud yoki olingan ma'lumotlardan foydalangan holda ma'lumotlarni izohlanadi.

Xulosa. Xulosa o'rnida shuni aytish mumkinki so'z ma'nosini aniqlashda mashinali o'qitish yondoshuvlaridan jahon miqyosida keng qo'llanilmoqda. Nazorat ostidagi, nazoratsiz hamda yarim nazorat ostidagi yondoshuvlardan foydalanish uchun katta hajmdagi ma'lumotlar bazasiga ehtiyoj seziladi. Bunday ma'lumotlar bazasi – har bir tilning milliy korpusidir.

Foydalanilgan adabiyotlar:

1. Tacoa, F., Bollegala, D. and Ishizuka, M. 2012. A Context Expansion Method for Supervised Word Sense Disambiguation. In IEEE Sixth International Conference on Semantic Computing
2. Sreedhar, J. Viswanadha, S., Raju, A., Babu, V., Shaik, A. and Kumar, P. 2012. Word Sense Disambiguation: An Empirical Survey. International Journal of Soft Computing and Engineering (IJSCE). 2.
3. Agirre, E. and Edmonds, P. 2006. Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology). Springer-Verlag New York, Inc. Secaucus, NJ, USA.
4. Navigli, R. 2009. Word Sense Disambiguation: A Survey. Universita di Roma La Sapienza, ACM Computing Surveys. 41.

5. Jurafsky, D. and Martin, J. H. 2008. An Introduction to Natural Language processing, Computational Linguistics, and Speech Recognition. Pearson Education.
6. Kolte, S. G. and Bhurud, S. G. 2008. Word Sense Disambiguation using Word Net Domains. In Proceedings of ICETET'08.
7. Bala, P. 2013. Word Sense Disambiguation Using Selectional Restriction. In International Journal of Scientific and Research Publications.
8. Zheng, Z. and Shu, Z. 2009. A New Approach to Word Sense Disambiguation in MT System. In World Congress on Computer Science and Information Engineering.
9. Lin, D. and Pantel, P. 2002. Discovering word senses from text”, In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Edmonton, Alta., Canada). 613–619.
10. V'Eronis, J. 2004. Hyperlex: Lexical cartography for information retrieval. Comput. Speech Lang. 18: 223–252.
11. Palta, E. 2006. Word Sense Disambiguation, M. Tech. dissertation, dept. CSE Indian Institute of Technology, Mumbai.
12. Collins, M. 2004. Parameter estimation for statistical parsing models: Theory and practice of distribution free methods. In New Developments in Parsing Technology, H. Bunt, J. Carroll, and G.Satta, Eds. Kluwer, Dordrecht, The Netherlands, 19–55.
13. B.B.Elov, X.I.Axmedova. “Homonymy detection using a naïve bayes classifier” Journal of Computer Science Engineering and Information Technology Research (JCSEITR) ISSN(P): 2250-2416; ISSN(E): Applied Vol. 13, Issue 1, March 2023, 5–20.
14. B.B.Elov, X.I.Axmedova. Business Process Modeling That Distinguishes Homonymy Within Three Parts of Speeches in Uzbek Language // 7th International Conference on Computer Science and Engineering, 14-16 September 2022, Turkey. 289-294.
15. B.B.Elov, X.I.Axmedova. O'zbek tilidagi polifunksional so'zlarni semantik farqlovchi biznes-jarayonlarni modellashtirish. O'zbekiston: til va madaniyat (Amaliy filologiya), 2021, 5(1). 87-10-b.
https://scholar.google.com/citations?view_op=view_citation&hl=ru&user=wYCKIls oAAAAJ&citation_for_view=wYCKIls oAAAAJ:Wp0gIr-vW9MC

BPE ALGORITMI ASOSIDA O‘ZBEK TILIDAGI MATNLARNI TOKENLASH

Xusainova Zilola Yuldashevna

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili
va adabiyoti universiteti tayanch doktoranti.

xusainovazilola@navoiy-uni.uz

Qosimova Sarvinoz Furqat qizi

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili
va adabiyoti universiteti Kompyuter
lingvistikasi mutaxassisligi 1-kurs magistranti.

sarvinozq54@gmail.com

Baxtiyorova Nazokat Doniyor qizi

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili
va adabiyoti universiteti, Kompyuter
lingvistikasi yo'nalishi 1-kurs talabasi

nazokatbaxtiyorova04@gmail.com

Turayeva Farzona Sherzod qizi

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili
va adabiyoti universiteti ToshDO‘TAU Kompyuter
lingvistikasi yo'nalishi 1-kurs talabasi

turayevafarzona@gmail.com

Annotatsiya. Tokenizatsiya matnli ma'lumotlar bilan ishlashning asosiy va boshlang'ich boshqichi hisoblanadi. Ushbu maqolada bayt juftligini kodlash (BPE) tokenizatsiya algoritmi tavsiflanadi va o'zbek tili korpusidagi matnlarga qo'llaniladi.

Kalit so'zlar: BPE, bayt juft kodlash, belgilar tokenizatsiyasi, Count Vectorization, OOV so'zlari, lug'atdan tashqari so'zlar, so'z ostilar tokenizatsiyasi.

Abstract. Tokenization is the primary and primary driver of textual data processing. This article describes the Byte Pair Encoding (BPE) tokenization algorithm and applies it to texts in the Uzbek language corpus.

Keywords: BPE, Byte Pair Encoding, Character Tokenization, Count Vectorization, OOV Words, Non-Vocabulary Words, Subtext Tokenization.

Абстракт. Токенизация является основным и основным фактором обработки текстовых данных. В этой статье описывается алгоритм токенизации Byte Pair Encoding (BPE) и применяется к текстам в корпусе на узбекском языке.

Ключевые слова: BPE, кодирование пар байтов, токенизация символов, векторизация подсчета, слова OOV, нелексические слова, токенизация подслов.

Kirish

Tabiiy tilni qayta ishlash (NLP) – sun'iy intellekt sohasining mashinali o'rganish metodlari orqali tabiiy tildagi ma'lumotlarni qayta ishlashga mo'ljallangan

yo'nalishidir. Tabiiy tilini qayta ishlash mashinalar uchun murakkab amal hisoblanadi. Chunki mashinalar matn bilan emas, raqamlar bilan ishlaganlari sababli, ularni raqamli shaklga olib kelish lozim. Tabiiy tilni tushunishda asosiy, boshlang'ich amallardan biri tokenizatorlar tomonidan bajariladi. Tokenizatorning har bir turi mashinalarga matnni boshqa usulda qayta ishlashga imkon beradi [1]. Tokenizatorlar qo'yilgan masalaga qarab turlicha yondashuv asosida ishlab chiqilishi mumkin.

Tokenizatsiya – dastlabki ishlov berish bosqichida strukturlanmagan tabiiy til matni yaxshiroq strukturlangan ko'rinishga aylantirishni anglatadi [2,3]. Tokenizatsiya jarayonida matn tokenlar deb ataladigan foydali ma'lumotlarni o'z ichiga olgan qismlarga ajratiladi. O'zbek tili morfologik analizatorini ishlab chiqishda ushbu maqolada keltirilgan tokenizatsiya algortimi qo'llanilgan [4,5].

Tokenizatsiya orqali tokenlar orasidagi chegaralar belgilanadi. Eng oddiy ajratuvchi sifatida bo'shliq yoki tinish belgilari orqali ajratishdan iborat. Biroq, yapon va xitoy tillarida *boshqa turdagi bo'shliqdan* foydalanadilar. Bigrammalar va idiomalar (masalan, “kundan kunga” yoki “ega bo'lgan”) kabi maxsus birikmalarni alohida tokenlarga bo'lish yoki bo'lmaslik to'g'risida **qaror qabul qilish holatlari** qo'yilgan NLP masalasiga bog'liq.

NLP ilovalarini ishlab chiqishda, ba'zi vositalar tokenizatsiya va sintaktik tahlildan so'ng matndagi gaplar aniqlanadi. Shunday qilib, agar ikkita tinish belgisi orasidagi so'zlar bir nechta gaplarni o'z ichiga olsa ham, ular ushlanib qoladi.

Tokenizatsiya haqida qisqacha ma'lumot

Tokenizatsiya **Count Vectorizer** kabi an'anaviy NLP usullarida va Transformers kabi Deep Learning arxitekturalaridagi asosiy qadamdir. Tokenlar tabiiy tilning qurilish bloklari hisoblanadi.

Tokenizatsiya – bu matn qismini tokenlar deb ataladigan kichikroq birliklarga ajratish usuli. Bu yerda tokenlar sifatida *so'zlar, belgilar* yoki *so'zostilar* bo'lishi mumkin. Demak, tokenizatsiyani keng ma'noda 3 turga – so'z, belgi va so'zostilar (n-gramm belgilar)ga ajratish mumkin.

Misol uchun, quyidagi gapni ko'rib chiqamiz: *"Hech qachon taslim bo'lmang"*. Tokenlarni shakllantirishning eng keng tarqalgan usuli bo'shliqqa (probel) asoslangan. Bo'sh joyni ajratuvchi sifatida qabul qilsak, gapning tokenizatsiyasi 4 ta tokenni hosil qiladi – *Hech-qachon-taslim-bo'lmang*. Har bir token so'zdan iborat bo'lgani uchun u **Word tokenizatsiyasining** namunasiga aylanadi [6,7].

Xuddi shunday, tokenlar sifatida belgilar va so'z ostilar hosil qilinishi mumkin. Misol uchun, keling, *"aqlliroq"* so'zini ko'rib chiqaylik:

- belgilar tokeni: *a-q-l-l-i-r-o-q;*
- so'z osti tokeni: *aqlli-roq.*

Tokenizatsiya jarayoni

Tokenlar tabiiy tilning qurilish bloklari bo'lganligi sababli, strukturlanmagan matnni qayta ishlashning eng keng tarqalgan usuli token darajasida sodir bo'ladi.

Shuningdek, **Transformer**ga asoslangan modellar – NLPdagi chuqur o'rganish arxitekturalari (State of The Art, SOTA) – strukturlanmagan matnni token darajasida qayta ishlaydi. Xuddi shunday, **RNN**, **GRU** va **LSTM** kabi **NLP** uchun eng mashhur chuqur o'rganish arxitekturalari ham token darajasida strukturlanmagan matnni qayta ishlaydi [8,9].

Shunday qilib, tokenizatsiya matnli ma'lumotlarni modellashtirishda eng muhim qadamdir. Tokenlarni olish uchun korpusda tokenizatsiya amalga oshiriladi. Keyingi qadamda tokenlardan lug'atni shakllantirishda foydalanilamiz. Lug'at korpusdagi o'ziga xos belgilar to'plamini anglatadi. Lug'at korpusdagi har bir noyob (unikal) *tokenni hisobga olgan holda* yoki *maksimal k martagacha uchraydigan so'zlarni hisobga olgan holda* tuzilishi mumkin. **Lug'at yaratish** - bu tokenizatsiyaning yakuniy natijasi.

NLP modeli samaradorligini oshirishning eng oddiy amallaridan biri *k martagacha uchraydigan so'zlardan iborat* lug'at yaratishdir. An'anaviy va ilg'or chuqur o'rganishga asoslangan NLP usullari orqali lug'atdan foydalanishni ko'rib chiqamiz.

Count Vectorizer va **TF-IDF** kabi an'anaviy NLP yondashuvlari xususiyatlar (mezonlar) sifatida lug'atdan foydalanadi. Lug'atdagi har bir so'z o'ziga xos xususiyat sifatida qoraladi:

	w1	w2	w3	w4	w5	w6
1- hujjat	1	1	1	1	1	1
2- hujjat	1	1	0	1	0	1
3- hujjat	1	0	1	0	1	0
4- hujjat	0	0	1	1	1	1

1-rasm. An'anaviy NLP: Count vektorizator

Ilg'or chuqur o'rganishga asoslangan NLP arxitekturalarida lug'atdan berilgan gaplarini tokenlash uchun foydalaniladi va modelga kirish parametri sifatida uzatiladi.

Tokenizatsiya usulini tanlash

Yuqorida aytib o'tilganidek, tokenizatsiya *so'z*, *belgi* yoki *so'z osti* darajasida amalga oshirilishi mumkin. Berilgan NLP vazifasini hal qilishda qaysi tokenizatsiyadan foydalanishimiz kerak?

So'zlarni tokenizatsiya qilish usuli

So'zlarni tokenizatsiya qilish eng ko'p ishlatiladigan algoritmidir. Tokenizatsiya algoritmidagi ma'lum bir ajratuvchiga asoslanib, matn alohida so'zlarga (qism)larga ajratiladi. Ajratuvchilarga qarab, turli xil so'z darajasidagi tokenlar hosil

bo'ladi. **Word2Vec** va **GloVe** kabi usullarda oldindan tayyorlangan so'zlar tokenizatsiya orqali aniqlanadi. Biroq, buning bir nechta kamchiliklari bor.

Tokenlar bilan bog'liq asosiy muammolardan biri lug'atdan tashqaridagi so'zlar (**Out Of Vocabulary, OOV**) bilan bog'liq. OOV so'zlari test paytida uchragan yangi so'zlarni anglatadi. Bu yangi so'zlar lug'atda mavjud emas. Shunday qilib, bu usullar OOV so'zlarini boshqarishda muvaffaqiyatsizlikka uchraydi. Quyida OOV so'zlar bilan yuzaga keladigan muammolarni bartaraf qilish usullari keltirilgan:

– *Lug'atni k marotabadan ko'p uchraydigan so'zlar bilan shakllantirish va o'quv ma'lumotlaridagi unikal (noyob) so'zlarni noma'lum tokenlar (UNK) bilan almashtirish lozim. Bu modelga OOV so'zlarining UNK tokenlari nuqtai nazaridan ifodalanishini o'rganishga yordam beradi.*

– *Test vaqtida lug'atda mavjud bo'lmagan har qanday so'z UNK tokeniga joylashtiriladi.*

Ushbu yondashuv bilan bog'liq muammo shundaki, OOVni UNK tokenlari bilan moslashtirilganda, so'z haqidagi barcha ma'lumotlar yo'qoladi. So'zning tuzilishi uni to'g'ri ifodalashda yordam beradi. Shuningdek, har bir OOV so'zi bir xil ko'rinishga ega bo'ladi.

So'zlarni tokenizatsiya qilish bilan bog'liq yana bir muammo lug'at hajmi bilan bog'liq. Odatda, oldindan tayyorlangan modellar katta hajmida matnlardan iborat korpus asosida o'qitiladi. Katta hajmdagi korpusdagi barcha noyob so'zlar bilan lug'at yaratishni tasavvur qiling. Bu holda katta hajmdagi so'zlardan iborat lug'at hosil qilinadi!

Belgilarni tokenizatsiya qilish usuli

Belgilar tokenizatsiyasi matnning qismini belgilar to'plamiga ajratadi. Ushbu usul *so'zlarni tokenizatsiya qilish*da keltirilgan kamchiliklarni bartaraf etadi. Belgilar tokenizatorlari so'z haqidagi ma'lumotlarini saqlab, OOV so'zlarini qayta ishlaydi. Bunda OOV so'zi belgilarga ajratiladi va so'zni shu belgilar asosida ifodalaydi. Shuningdek, ushbu usulda lug'at hajmini cheklanadi ($n=26$).

Belgilar asosida tokenlash OOV muammosini hal qilsa-da, bunda kiruvchi va chiquvchi gaplarining uzunligi oshishi kuzatiladi. Chunki gaplarni belgilar ketma-ketligi sifatida ifodalaymiz. Natijada, ma'47741

noli so'zlarni shakllantirish uchun belgilar o'rtasidagi munosabatlarni o'rganish qiyin bo'ladi. Ushbu kamchilikni so'z va belgilar tokenizatsiyasi o'rtasida joylashgan **Subword Tokenization** deb nomlanuvchi boshqa tokenizatsiya bartaraf qiladi.

So'z ostilarni tokenlash usuli

So'z ostilarni tokenlash usulida matn qismini so'z ostilarga (yoki n-gramm belgilarga) ajratadi. Misol uchun, **lower** so'zini **low-er** ga, **smartest** so'zini **smart-est** ga bo'lish mumkin. NLPdagi SOTA (Transformed based models) kabi modellar tokenizator algoritmlaridan lug'atni tayyorlash uchun foydalanadi.

BPE usuli

Byte Pair Encoding (BPE) – transformerga asoslangan modellar (*transformer-based models*) orasida keng qo'llaniladigan tokenizatsiya usuli. BPE so'zlar va belgilar tokenizatorlari muammolarini hal qiladi [10-12]:

– BPE OOV muammosini samarali hal qiladi. U OOVni so'z ostilar sifatida ajratadi va qayta ishlaydi;

– BPEdan keyin kirish va chiqish gaplarining uzunligi belgilar tokenizatsiyasiga nisbatan qisqaroq.

BPE – so'zlarni segmentlash algoritmi bo'lib, u eng ko'p uchraydigan belgilar yoki belgilar ketma-ketligini iterativ tarzda birlashtiradi. BPE algoritmini amalga oshirish qadamlari:

1. Korpusdagi so'zlarni $\langle /w \rangle$ qo'shimchasidan keyin belgilarga ajratish;

2. Korpusdagi noyob so'zlar bilan lug'atni shakllantirish;

3. Korpusdagi bir juft belgilar yoki belgilar ketma-ketligi chastotasini hisoblash;

4. Korpusdagi eng ko'p uchraydigan juftlikni birlashtirish;

5. Eng yaxshi juftlikni lug'atga saqlash;

6. Muayyan miqdordagi iteratsiya uchun 3-5-bosqichlarni takrorlash

Quyidagi korpus berilgan bo'lsin:

Corpus		
kam	kamroq	kitob
kam	kamroq	kitob
kam	betob	kitob
kam	betob	kitob
kam	betob	kitob

1a) Korpusdagi har bir so'zning so'z oxiriga ($\langle /w \rangle$) belgisini qo'shish:

Corpus		
kam $\langle /w \rangle$	kamroq $\langle /w \rangle$	kitob $\langle /w \rangle$
kam $\langle /w \rangle$	kamroq $\langle /w \rangle$	kitob $\langle /w \rangle$
kam $\langle /w \rangle$	betob $\langle /w \rangle$	kitob $\langle /w \rangle$
kam $\langle /w \rangle$	betob $\langle /w \rangle$	kitob $\langle /w \rangle$
kam $\langle /w \rangle$	betob $\langle /w \rangle$	kitob $\langle /w \rangle$

1b) Korpusdagi so'zlarni belgilarga ajratish:

Corpus		
k a	k a m r o	k i t o b
m $\langle /w \rangle$	q $\langle /w \rangle$	$\langle /w \rangle$
k a	k a m r o	k i t o b
m $\langle /w \rangle$	q $\langle /w \rangle$	$\langle /w \rangle$

k a m </w>	b e t o b </w>	k i t o b </w>
k a m </w>	b e t o b </w>	k i t o b </w>
k a m </w>	b e t o b </w>	k i t o b </w>

2. Lugʻatning boshlangʻich holati:

Lugʻat									

1-iteratsiya:

3. Chastotani hisoblash:

Chastota		
k- a (7)	o-q (2)	o- b (8)
a- m (7)	q- </w> (2)	b- </w> (8)
m- </w> (5)	k-i (5)	b- e (3)
m- r (2)	i-t (5)	e-t (3)
r- o (2)	t-o (8)	

4. Eng koʻp uchraydigan juftlikni birlashtirish:

Corpus		
k a m </w>	k a m r o q </w>	k I to b </w>
k a m </w>	k a m r o q </w>	k I to b </w>
k a m </w>	b e to b </w>	k I to b </w>
k a m </w>	b e to b </w>	k I to b </w>
k a m </w>	b e to b </w>	k I to b </w>

5. Eng yaxshi juftlikni saqlash:

Lugʻat

o									

Endi har bir iteratsiya uchun 3-5-bosqichlarni takrorlash lozim. Yana bir takrorlash uchun tasvirlab beramiz.

2-iteratsiya:

3. Chastotani hisoblash:

Chastota		
k- a (7)	o-q (2)	b- </w> (8)
a- m (7)	q- </w> (2)	b- e (3)
m- </w> (5)	k-i (5)	e- to (3)
m- r (2)	i-to (5)	
r- o (2)	to b (8)	

4. Eng ko'p uchraydigan juftlikni birlashtirish:

Corpus		
k a m </w>	k a m r o q </w>	k i tob </w>
k a m </w>	k a m r o q </w>	k i tob </w>
k a m </w>	b e tob </w>	k i tob </w>
k a m </w>	b e tob </w>	k i tob </w>
k a m </w>	b e tob </w>	k i tob </w>

5. Eng yaxshi juftlikni saqlash:

Lug'at									
	r								
	t								
o	ob								

10 ta iteratsiyadan so'ng, BPE birlashma operatsiyalari quyidagicha ko'rinadi:

Lug'at									
		m			e		i		b
		t			k		k		b
o	ob	ob</w>	a	am	am</w>	i	itob</w>	e	etob</w>

Xulosa

Korpus matnlarini samarali tarzda qayta ishlash uchun BPE har bir iteratsiyada so'zning chastotasiga qarab birlashishning har qanday potensial variantidan o'tadi va eng yaxshi yechimni optimallashtirish uchun o'ziga xos yondashuvga amal qiladi. BPE eng ko'p ishlatiladigan subword-tokenizatsiya algoritmlaridan biri bo'lib, samarali natijalarni taqdim etadi. Maqolada keltirilgan usullar yordamida o'zbek tilidagi korpus (matnlar)ga qo'llash mumkin.

Foydalanilgan adabiyotlar:

1. Webster, J. J., & Kit, C. (1992). *Tokenization as the initial phase in NLP*. <https://doi.org/10.3115/992424.992434>
2. Rai, A., & Borah, S. (2021). Study of various methods for tokenization. In *Lecture Notes in Networks and Systems* (Vol. 137). https://doi.org/10.1007/978-981-15-6198-6_18
3. ARAVIND PAI. (2020). *What is Tokenization in NLP? Here's All You Need To Know*.
4. B.Elov, Sh.Hamroyeva, X.Axmedova. (2022). Methods for creating a morphological analyzer, *14th International Conference on Intelligent Human Computer Interaction, IHCI 2022, 19-23 October 2022, Tashkent*.
5. B.Elov, Sh.Hamroyeva, D.Elova. (2022). Morfologik analizatorni yaratish usullari, *O'zbekiston: til va madaniyat (Amaliy filologiya)*, 2022, 5(1).
6. Tailor, C., & Patel, B. (2019). Sentence Tokenization Using Statistical Unsupervised Machine Learning and Rule-Based Approach for Running Text in Gujarati Language. *Advances in Intelligent Systems and Computing*, 841. https://doi.org/10.1007/978-981-13-2285-3_38
7. Song, X., Salcianu, A., Song, Y., Dopson, D., & Zhou, D. (2021). Fast WordPiece Tokenization. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*. <https://doi.org/10.18653/v1/2021.emnlp-main.160>
8. Ngo Ho, A. K., & Yvon, F. (2021). Optimizing Word Alignments with Better Subword Tokenization. *Proceedings of Machine Translation Summit XVIII: Research Track*.
9. https://huggingface.co/docs/transformers/tokenizer_summary



10. Vilar, D., & Federico, M. (2021). A Statistical Extension of Byte-Pair Encoding. *IWSLT 2021 - 18th International Conference on Spoken Language Translation, Proceedings*. <https://doi.org/10.18653/v1/2021.iwslt-1.31>
11. Bostrom, K., & Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*. <https://doi.org/10.18653/v1/2020.findings-emnlp.414>
12. Heinzerling, B., & Strube, M. (2019). BPEMB: Tokenization-free pre-trained subword embeddings in 275 languages. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*.
13. Gutierrez-Vasques, X., Bentz, C., Sozinova, O., & Samardžić, T. (2021). From characters to words: The turning point of BPE merges. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. <https://doi.org/10.18653/v1/2021.eacl-main.302>
14. B.Elov, Sh.Hamroyeva, X.Axmedova. O'zbek-ingliz tillarining teglangan parallel korpusini yaratish bosqichlari. / O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 90-102-b.

O'ZBEK TILI MATNLARINI QAYTA ISHLASHDA BOW ALGORITMIDAN FOYDALANISH

Xusainova Zilola Yuldashevna

ToshDO'TAU tayanch doktoranti.

xusainovazilola@navoiy-uni.uz

Xusanboyeva Jamila Ixtiyor qizi

ToshDO'TAU

Kompyuter lingvistikasi yo'nalishi 1-kurs talabasi

jamilaixtiyarovna04@gmail.com

Bekqulova Sevinch Furqatovna

ToshDO'TAU

Kompyuter lingvistikasi yo'nalishi 1-kurs talabasi

bekqulovasevinch2310@gmail.com

Bekchanova Munira Oybek qizi

ToshDO'TAU

Kompyuter lingvistikasi yo'nalishi 1-kurs talabasi

bekchanovamunir@gmail.com

Annotatsiya. Bag Of Words (BoW) modellash algoritmidan foydalanib, matnni raqamli matritsalariga aylantirish va qayta ishlash mumkin. So'zlar sumkasi (BoW) – so'zning hujjatdagi statistikasini hisoblash algoritmi hisoblanadi. BoW algoritmidan foydalanib hujjatlarni o'zaro solishtirish, qidiruv tizimlarida ma'lumotlarni izlash, hujjatlarni tasniflash va tematik modellash mumkin. Ushbu maqolada o'zbek tilidagi matnlarni BoW algoritmidan foydalanib raqamli shaklga o'tkazish usullari keltiriladi.

Kalit so'zlar: *BoW, Bag of words, so'zlar jamlanmasi, so'z vektori, token, BoW algoritmi, TF-IDF usuli.*

Abstract. Using the Bag Of Words (BoW) modeling algorithm, text can be converted and processed into digital matrices. Bag of Words (BoW) is an algorithm for calculating word statistics in a document. Using the BoW algorithm, it is possible to compare documents, search for information in search engines, classify documents and perform thematic modeling. This article presents the methods of digitalization of Uzbek texts using the BoW algorithm.

Keywords: *BoW, Bag of words, set of words, word vector, token, BoW algorithm, TF-IDF method.*

Абстракт. Используя алгоритм моделирования Bag Of Words (BoW), текст можно преобразовать и обработать в цифровые матрицы. Bag of Words (BoW) - это алгоритм подсчета статистики слов в документе. С помощью алгоритма BoW можно сравнивать документы, искать информацию в поисковых системах, классифицировать документы и выполнять тематическое моделирование. В данной статье представлены методы оцифровки узбекских текстов с использованием алгоритма BoW.

Ключевые слова: *BoW, мешок слов, набор слов, вектор слов, токен, BoW-алгоритм, метод TF-IDF.*

Kirish

Soʻzlar jamlanmasi (“sumkasi”) (**Bag of words, BOW**) – tabiiy tilni qayta ishlashda matnni raqamlashtirish usuli. Ushbu maqolada BoW algoritmini qoʻllanilish sohasi, kontseptsiyasi haqida fikr mulohaza yuritiladi va oʻzbek tilidagi matnlar uchun Python tilidagi tadbirgʻi keltiriladi. Axborot tizimlaridagi matnli maʼlumotlarni *tushunish* va muayyan *qarorlarni qabul qilish* uchun ularni **raqamli shaklga olib kelish** kerak [1,2].

BoW algoritmi

BoW algoritmi yordamida matnli maʼlumotlardan zarur *xususiyatlar ajratib olinadi*. BoW usulida matndagi soʻzlar soni (statistikasi) aniqlanadi. Lekin, grammatik tafsilotlar va soʻz tartibi eʼtiborsiz qoldiriladi. Usulning soʻzlar “sumkasi” deb atalishi hujjatdagi soʻzlarning *tartibi yoki tuzilishi* haqidagi har qanday maʼlumotning tashlab yuborilishiga asoslangan [3].

Koʻp hollarda maʼlumotlar tartibsiz va strukturlanmagan shaklda beriladi. Mashinali oʻrganish (ML) algoritmlari esa strukturlangan va aniq belgilangan qatʼiy uzunlikdagi maʼlumotlarni qayta ishlashga asoslangan. Shu sababli, BoW algoritmidan foydalanib, oʻzgaruvchan uzunlikdagi matnlarni fikslangan uzunlikdagi matn (**vektor**)ga aylantirish orqali ulardan ML algoritmlarida foydalanish mumkin [4,5]. BoW algoritmidan foydalanib, matnga ekvivalent raqamlar vektori hosil qilinadi. BoW usuli orqali matnni vektorga aylantirish jarayonini oʻzbek tilidagi gaplar misolida koʻrib chiqamiz:

1-gap: “*Adirlar ham bahorda lola bilan goʻzal, chunki lola – bahorning erka guli*”

2-gap: “*Lola ham shifokorlik kasbini tanladi*”

1-jadval. Gaplarni tokenlarga ajratish

1-gap	2-gap
Adirlar	Lola
Ham	ham
bahorda	shifokorlik
lola	kasbini
bilan	tanladi
goʻzal	
,	
chunki	
lola	
–	
bahorning	
erka	
guli	

1-qadam: Yuqoridagi matndagi barcha soʻzlarni koʻrib chiqib, umumiy soʻzlardan iborat lugʻatni shakllantiramiz.

- Adirlar
- ham
- bahorda
- lola
- bilan
- goʻzal
- ,
- chunki
- —
- bahorning
- erka
- guli
- Lola
- shifokorlik
- kasbini
- tanladi

Hosil qilingan lugʻatda **“Lola”** va **“lola”** soʻzlari bir xil emas. Chunki ular gaplarda turli shaklda kelgan va shuning uchun takrorlanmaydi. Shuningdek, soʻzlar roʻyxatidan vergul: “,” ham oʻrin olgan. Berilgan ikkita gapdan hosil qilingan lugʻatda **14 ta** token mavjud boʻlib, har bir tokenni baholash uchun vektorda bitta pozitsiyaga ega boʻlgan 14 uzunlikdagi hujjat vektoridan foydalanish mumkin. Matnga mos vektorni hosil qilishda har bir soʻzning mavjudligini qayd etishda **1** yoki **0** qiymatlaridan (chastota) foydalaniladi. 1-gapga mos vektor quyidagicha hosil qilinadi:

2-jadval. 1-gapga mos soʻzlar chastotasi

Soʻz	Chastota
Adirlar	1
ham	1
bahorda	1
lola	1
bilan	1
goʻzal	1
,	1
chunki	1
—	1
bahorning	1
erka	1
guli	1

Lola	0
shifokorlik	0
kasbini	0
tanladi	0

1-gapdagi chastotalardan foydalanib quyidagi vektorni hosil qilamiz:

[1,1,1,1,1,1,1,1,1,0,0,0]

2-gapga mos vektor quyidagicha hosil qilinadi:

3-jadval. 2-gapga mos so'zlar chastotasi

So'z	Chastota
Adirlar	0
ham	1
bahorda	0
bola	0
bilan	0
go'zal	0
,	0
chunki	0
—	0
bahorning	0
erka	0
guli	0
Lola	1
shifokor	1
kasbini	1
tanladi	1

Xuddi shunday, yuqoridagi chastotalardan 2-gapga mos vektorni hosil qilamiz:

[0,1,0,0,0,0,0,0,0,0,0,1,1,1,1]

Berilgan ikkita so'zga mos vektorni quyidagicha aniqlaymiz:

	adir lar	ham	bahord a	lola	bilan	go'zal		hun i		bahorni ng	erka	guli	lola	shifo korlik	kas bini	tan adi
-gap																
-gap																

Yuqoridagi misolda BoW usulining eng yaxshi jihatlari o'z aksini topmagan. Chunki, “Lola” va “lola” so'zlari bir xil ma'noga ega bo'lsa-da, ikki marta qaytd etilgan. Shuningdek, hech qanday ma'lumotni bildirmaydigan vergul “,” , “—” belgisi

ham lug‘at tarkibiga kiritilgan va tahlil jarayonida ishtirok etgan. BoW usuliga ba’zi o‘zgarishlarni amalga oshirish orqali samaradorlikni oshirishni ko‘rib chiqamiz.

Boshlang‘ich ishlov berish

1-gap: *“Adirlar ham bahorda lola bilan go‘zal, chunki lola – bahorning erka guli”*

2-gap: *“Lola ham shifokorlik kasbini tanladi”*

1-qadam: Berilgan gaplarni kichik harf (registr)ga aylantirish.

2-qadam: Matndan maxsus belgilar va nomuhim so‘zlar(stopwords)ni olib tashlash. Stopwords - "albatta", "ba'zi", "bilan", "ham", "chunki" kabi matnda muhim ahamiyat kasb etmagan so‘zlar. Yuqoridagi amallar bajarilganidan so‘ng, berilgan gaplarda quyidagi o‘zgartirishlar amalga oshiriladi:

1-gap: *“adirlar bahorda lola go‘zal lola bahorning erka guli”*

2-gap: *“lola shifokorlik kasbini tanladi”*

Yuqoridagi gaplar ma'noga ega bo‘lmasa-da, gap mazmuniga ta'sir qiluvchi maksimal ma'lumotlarga ega so‘zlardan iborat.

3-qadam: Berilgan gaplardagi barcha so‘zlardan iborat lug‘atni shakllantirish.

- adirlar
- bahorda
- lola
- go‘zal
- bahorning
- erka
- guli
- shifokorlik
- kasbini
- tanladi

Hosil qilingan lug‘atda endilikda **10** ta so‘z mavjudligi sababli, har bir so‘zni baholash uchun vektorda bitta o‘ringa ega bo‘lgan 10 ta qat'iy uzunlikdagi hujjat vektoridan foydalanish mumkin. 1-gapga mos vektor quyidagicha hosil qilinadi:

4-jadval. 1-gapga mos so‘zlar chastotasi

So‘z	Chastota
Adirlar	1
bahorda	1
lola	2
go‘zal	1
bahorning	1
erka	1
guli	1
shifokorlik	0

kasbini		0
tanladi		0

1-gapdagi chastotalardan foydalanib quyidagi vektorni hosil qilamiz:

[1,1,2,1,1,1,0,0,0]

2-gapga mos vektor quyidagicha hosil qilinadi:

5-jadval. 2-gapga mos so'zlar chastotasi

So'z	Chastota
adirlar	0
bahorda	0
lola	1
go'zal	0
bahorning	0
erka	0
guli	0
shifokorlik	1
kasbini	1
tanladi	1

2-gapdagi chastotalardan foydalanib quyidagi vektorni hosil qilamiz:

[0,0,1,0,0,0,1,1,1]

Berilgan ikkita so'zga mos vektorni quyidagicha aniqlaymiz:

	Adirlar	bahord a	lola	go'zal	bahorni ng	erka	guli	shifokorlik	kasbi ni	tanladi
-gap								0		
-gap								1		

Mashinali o'rganishda ishlatiladigan ma'lumotlar to'plamlari juda katta hajmda bo'lib, bir necha ming yoki hatto millionlab so'zlardan iborat lug'atni o'z ichiga olishi mumkin. Shunday qilib, so'zlar jamlanmasidan foydalanishdan oldin matnni boshlang'ich qayta ishlash lozim. BoW usuli samaradorligini oshirishi mumkin bo'lgan turli xil qayta ishlash bosqichlari mavjud bo'lib, ularning ba'zilari ushbu maqolada batafsil keltiriladi.

Yuqoridagi misollarda BoW vektorini hosil qilish uchun lug'atdagi barcha so'zlardan foydalandik. Bu BoW modelini amalga tadbqiq qilishda bir qator murakkabliklarni yuzaga keltiradi. Amalda vektorni yaratish uchun lug'atdan faqat bir nechta so'zlar (eng ko'p uchragan so'zlar) ishlatiladi.

BoW usulini Pythonda amalga oshirish

BoW algoritmini Python tilida quyidagicha tadbiq qilamiz:

```
def vectorize(tokens):
    vector=[]
    for w in filtered_vocab:
        vector.append(tokens.count(w))
    return vector
def unique(sequence):
    seen = set()
    return [x for x in sequence if not (x in seen or seen.add(x))]
# Nomuhim so'zlar ro'yxatini yaratish
stopwords=["ham","bilan","chunki",...]
# Maxsus belgilar ro'yxati
special_char=["",":"," ",";",".", "?", "-"]
# Korpusdagi gaplarni aniqlash
string1="Adirlar ham bahorda lola bilan go'zal , chunki lola - bahorning
erka guli"
string2="Lola ham shifokor bo'lib ishlaydi"
# so'zni kichik harfga aylantirish
string1=string1.lower()
string2=string2.lower()
# gaplarni tokenlarga ajratish
tokens1=string1.split()
tokens2=string2.split()
print(tokens1)
print(tokens2)

# lug'at ro'yxatini tuzish
vocab=unique(tokens1+tokens2)
print(vocab)
# lug'atlar ro'yxatini filtrlash
filtered_vocab=[]
for w in vocab:
    if w not in stopwords and w not in special_char:
        filtered_vocab.append(w)
print(filtered_vocab)
# gaplarni vektorlarga aylantirish
vector1=vectorize(tokens1)
print(vector1)
vector2=vectorize(tokens2)
print(vector2)
```


['adirlar', 'ham', 'bahorda', 'lola', 'bilan', 'gozal', ',', 'chunki', 'lola', '-',
'bahorning', 'erka', 'guli']
['lola', 'ham', 'shifokor', 'bo'lib', 'ishlaydi']
['adirlar', 'ham', 'bahorda', 'lola', 'bilan', 'gozal', ',', 'chunki', '-', 'bahorning',
'erka', 'guli', 'shifokor', 'bo'lib', 'ishlaydi']
['adirlar', 'bahorda', 'lola', 'gozal', 'bahorning', 'erka', 'guli', 'shifokor',
'bo'lib', 'ishlaydi']
[1, 1, 2, 1, 1, 1, 1, 0, 0, 0]
[0, 0, 1, 0, 0, 0, 0, 1, 1, 1]

Xulosa

BoW – hujjatlarni bir butun sifatida tasniflashning oddiy va qulay usuli hisoblanadi. BoW modeli orqali mashinali o'rganish algoritmlariga ma'lumotlarni raqamli ko'rinishda taqdim etish uchun matn modellashiriladi. BoW modelning aniqligini oshirishning bir necha yo'llari mavjud:

- BoW modelida unigramlarni (so'zlar) ishlatish o'rniga, bigramm yoki trigrammdan foydalanish model samaradorligini oshirishga xizmat qiladi.
- TF-IDF modelidan matndagi muhim va nomuhim so'zlarni aniqlashda foydalaniladi. Odatda TF-IDF modelidan mashinali o'rganishga asoslangan NLP vazifalarda qo'llaniladi.

Nisbatan oddiy model bo'lishiga qaramay, BoW ko'pincha matn tasnifi NLP vazifalari uchun ishlatiladi.

Foydalanilgan adabiyotlar:

1. Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, 12(2–3). <https://doi.org/10.1080/19312458.2018.1455817>
2. Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1–4). <https://doi.org/10.1007/s13042-010-0001-0>
3. Yan, D., Li, K., Gu, S., & Yang, L. (2020). Network-Based Bag-of-Words Model for Text Classification. *IEEE Access*, 8. <https://doi.org/10.1109/ACCESS.2020.2991074>
4. Qiu, D., Jiang, H., & Chen, S. (2020). Fuzzy information retrieval based on continuous bag-of-words model. *Symmetry*, 12(2). <https://doi.org/10.3390/sym12020225>
5. B.Elov, Sh.Hamraeva, Axmedova X. Methods for creating a morphological analyze // 14th International Conference on Intellegent Human Computer Interaction. 19-23 October 2022, Tashkent.



SYNTACTIC PARSING APPROACHES FOR ENGLISH, TURKISH AND UZBEK

Rakhmonova Mohira Abdusattor qizi

State University of Uzbek Language and
Literature, Tashkent, Uzbekiston
mokhirata@gmail.com

Abstract. This article discusses about syntactic parsing methods and related works on dependency Parsing. There are transition-based and graph-based approaches applied to dependency parsing problem in the literature. As well as the constituency parse tree is based on the formalism of context-free grammars. In this type of tree, the sentence is divided into constituents, that is, sub-phrases that belong to a specific category in the grammar.

Keywords: *parsing, dependency parsing, constituency parsing, transition-based approaches, graph-based approaches, parse tree, context-free grammars.*

Syntactic parsing is the automatic analysis of syntactic structure of natural language and the task of assigning a syntactic structure to a sentence, especially syntactic relations (in dependency grammar) and labelling spans of constituents (in constituency grammar). [1] It is motivated by the problem of structural ambiguity in natural language: a sentence can be assigned multiple grammatical parses, so some kind of knowledge beyond computational grammar rules are need to tell which parse is intended. Syntactic parsing is one of the important tasks in computational linguistics and natural language processing, and has been a subject of research since the mid-20th century with the advent of computers. [18]

There are two types of parsing methods:

1. Dependency Parsing
2. Phrase Structure Parsing (Constituency Parsing) [3]

Dependency Parsing is the task of finding the grammatical structure of a sentence by identifying the syntactic and semantic relationships between words. Dependency parsing has been utilized in many other NLP tasks such as machine translation [4, 5], relation extraction [6, 7], named entity recognition [8, 9], information extraction [10, 11], all of which involve natural language understanding to an extent. Each dependency relation is identified between a head word and a dependent word that modifies the head word in a sentence. Although such relations are considered as syntactic, they are naturally built upon semantic relationships between words. For example, each dependent has a role of modifying its head word, which is a result of a completely semantic influence. Dependency structures are represented either by hierarchical structures which are called dependency trees, or represented in the form of directed graphs, which are called dependency graphs. An

example dependency graph for the sentence Thank you, Mr. Pottering. is given below:

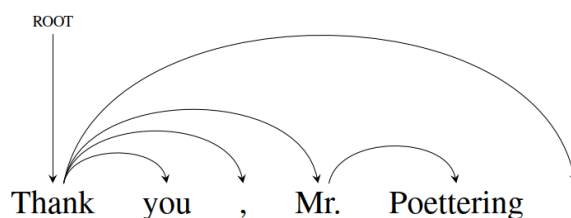


Figure 1. An example of dependency parsing for the sentence “Thank you, Mr. Pottering.” [12]

Syntactic relations between words are generally figured out with an arrow in a dependency tree, which connects each head word to a dependent. In other words, in a relation such as $I \rightarrow am$; “I” becomes the head and “am” becomes the dependent and the arrow between them states a dependency between the two words. The ROOT token represents the root of the dependency tree (i.e. the starting point of dependency parsing or the head of the complete sentence). Even if the rules of dependency parsing will be discussed later, it is good to state here that every sentence must contain a ROOT token in its dependency tree. Dependency parsing is a task that finds the lexical dependencies between words in a sentence, and thereby extracts the grammatical structure of a sentence. Dependency is a head-dependent relation between the words. The head is the one that affects the dependents, and in other words each dependent is affected by its head. The arrows in dependency trees are always from the head to the dependents. There are two main approaches applied to dependency parsing problem in the literature: transition-based and graph-based.[12]

Transition-based approaches are generally based on transition commands and a two-stack structure that contains a dependency stack and a word buffer. Word buffer contains the words in a sentence. Words are drawn from the word buffer and pushed into the dependency stack. If there is a transition between the top two words of the dependency stack, then a dependency is created between them and this operation continues until there are no words in the dependency stack. The last word in the dependency stack would be the ROOT, which is the root of the dependency tree; starting point of the whole dependency parsing process.[12]

Graph-based approaches are generally based on performing the entire parsing process as graph operations where the nodes in the graph represent the words in a sentence. For the sentence, “John saw Mary”, imagine a weighted graph G with four vertices where each of them refers to a word including the ROOT. Edges store the dependency scores between the words.

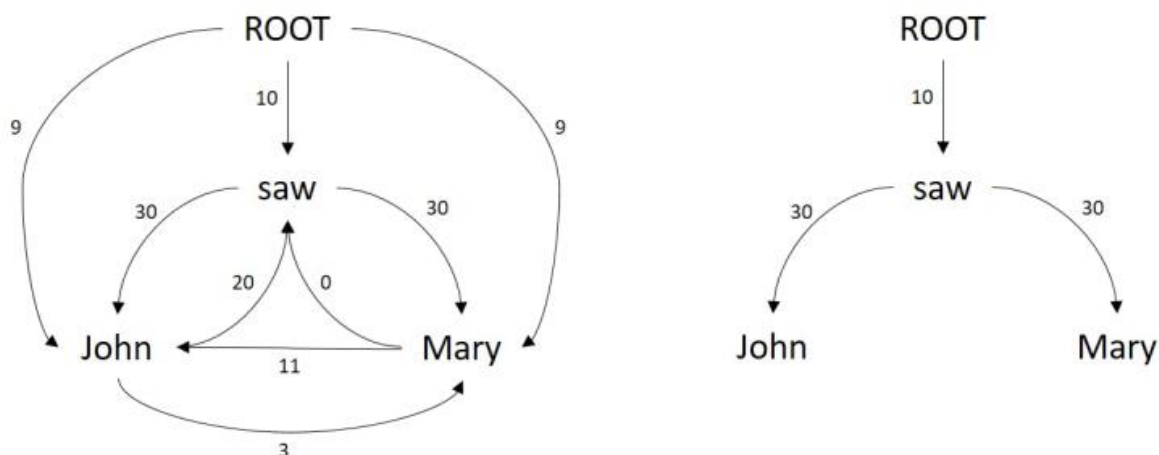


Figure 2. Graph-Based Dependency Parsing [12]

Related Work on Dependency Parsing

Eryigit and Oflazer (2006) [14] come with the idea of using inflectional groups (IGs) for dependency parsing. In their study, the authors use a statistical parser that firstly computes unit-unit relations where the units are words or IGs and then finds the maximum spanning tree from these computed relations. They have three baseline models: Word-based, IG-Based, and IG-Based with word-final IG contexts which is an IG-Based model with strict outputs. As expected, IG-Based models give the best results.

Eryigit, Oflazer and Nivre [13] show that the morphological structure plays a crucial role in Turkish dependency parsing. The authors show that parsing a sentence considering the IGs, which are sublexical units of a word, outperforms dependency parsing based on word tokens of sentences.

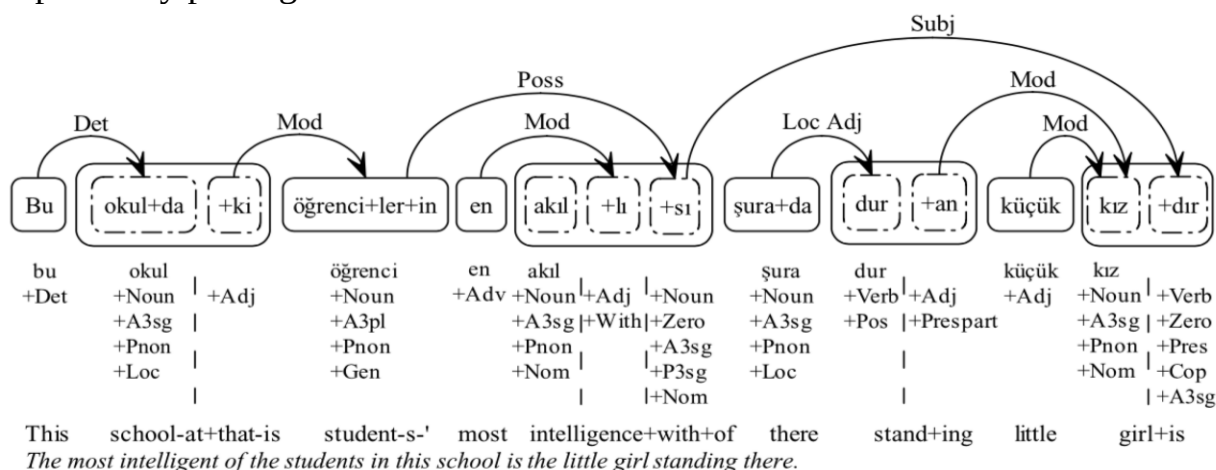


Figure 3. Inflection Groups (IGs) used in dependency parsing [13]

Oflazer (2014) [15] analyzes different NLP tasks on Turkish. In the dependency parsing task, the author underlines the importance of IGs and morphological units in dependency parsing.

Eryigit (2012) [16] makes an analysis on parsing in raw datasets in Turkish and shows that the locations of words in a sentence plays a crucial role in parsing.

Constituency Parsing

The constituency parse tree is based on the formalism of context-free grammars. In this type of tree, the sentence is divided into constituents, that is, sub-phrases that belong to a specific category in the grammar. In English, for example, the phrases “a dog”, “a computer on the table” and “the nice sunset” are all noun phrases, while “eat a pizza” and “go to the beach” are verb phrases. The grammar provides a specification of how to build valid sentences, using a set of rules. As an example, the rule $VP \rightarrow V NP$ means that we can form a verb phrase (VP) using a verb (V) and then a noun phrase (NP). While we can use these rules to generate valid sentences, we can also apply them the other way around, in order to extract the syntactical structure of a given sentence according to the grammar. Let's dive straight into an example of a constituency parse tree for the simple sentence, “I saw a fox”:[17]

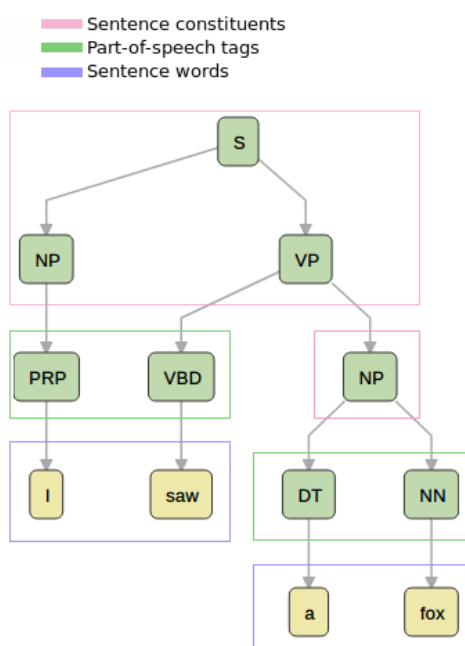


Figure 4. Constituency parse tree for the simple sentence [17]

Context free Chomsky grammar (CFG) is the most widely used formal system for modeling constituent structure in natural languages. CFG consists of a set of rules or productions, each of which expresses the ways that symbols of the language can be grouped and ordered together, and a lexicon of words and symbols. CFG G is defined by four parameters [2]:

$$G = \langle Ns, S, Ts, R \rangle,$$

where Ns – a set of nonterminal symbols; $S \in N$ – a start nonterminal symbol; Ts – a set of terminal symbols; R – a set of rules of the form $A \rightarrow \alpha$, $A \in N$ – a nonterminal symbol, $\alpha \in (Ns \cup Ts,)^*$ – a string of symbols from the infinite set of strings $(Ns \cup Ts,)^*$.

Constituency parsing and Dependency parsing

Constituency parsing focuses on identifying the constituent structure of a sentence, such as noun phrases and verb phrases.	Dependency parsing focuses on identifying the grammatical relationships between words in a sentence, such as subject-verb relationships.
Constituency parsing uses phrase structure grammar, such as context-free grammar or dependency grammar.	Dependency parsing uses dependency grammar, which represents the relationships between words as labeled directed arcs.
Constituency parsing is based on a top-down approach, where the parse tree is built from the root node down to the leaves.	Dependency parsing is based on a bottom-up approach, where the parse tree is built from the leaves up to the root.
Constituency parsing represents a sentence as a tree structure with non-overlapping constituents.	Dependency parsing represents a sentence as a directed graph, where words are represented as nodes and grammatical relationships are represented as edges.
Constituency parsing is more suitable for natural language understanding tasks.	Dependency parsing is more suitable for natural language generation tasks and dependency-based machine learning models.
Constituency parsing is more expressive and captures more syntactic information, but can be more complex to compute and interpret.	Dependency parsing is simpler and more efficient, but may not capture as much syntactic information as constituency parsing.

Constituency parsing is more appropriate for languages with rich morphology such as agglutinative languages.	Dependency parsing is more appropriate for languages with less morphological inflection like English and Chinese.
Constituency parsing is used for more traditional NLP tasks like Named Entity Recognition, Text classification, and Sentiment analysis.	Dependency parsing is used for more advanced NLP tasks like Machine Translation, Language Modeling, and Text summarization.
Constituency parsing is more suitable for languages with rich syntactic structures.	Dependency parsing is more suitable for languages with less complex syntactic structures.

Figure 5. Constituency parsing and Dependency parsing differences [20]

Conclusion

Uzbek is a [Turkic language](#) spoken by [Uzbeks](#). Uzbek is spoken as either native or second language by 44 million people around the world making it the second-most widely spoken Turkic language after [Turkish](#). [19] As an agglutinative language, Uzbek makes excessive use of morphological concatenation. With respect to syntactic properties, Uzbek has a relatively free word order. Even though SOV is the base word order, other permutations are highly utilized. For example:

Men maktabga onam bilan kecha bordim.

Kecha men onam bilan maktabga bordim

Men onam bilan kecha maktabga bordim

Maktabga kecha men onam bilan bordim

Uzbek has a free-order grammar and rich morphology, which makes dependency parsing even harder for Uzbek language. But, constituency parsing is more appropriate for languages with rich morphology such as agglutinative languages such as Uzbek.

References

1. Jurafsky, Dan; Martin, James H. (2021). [Speech and Language Processing](#)
2. Chomsky, N. 2002. Syntactic Structures. Mouton de Gruyter.
3. Ersin Ihsan Unkar (2007). PARSING TURKISH SENTENCES FOR NATURAL LANGUAGE WATERMARKING



4. Xavier Carreras and Michael Collins. Non-projective parsing for statistical machine translation. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 200–209. Association for Computational Linguistics, Singapore, 2009.
5. Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. Improved neural machine translation with a syntax-aware encoder and decoder. pages 1936–1945. 2017. P17-1177.
6. Katrin Fundel-Clemens, Robert Kuffner, and Ralf Zimmer. Relex – relation extraction using dependency parse trees. *Bioinformatics* (Oxford, England), 23:365–71, 2007.
7. Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2205–2215. Association for Computational Linguistics, Brussels, Belgium, 2018. D18-1244.
8. Zhanming Jie, Aldrian Obaja Muis, and Wei Lu. Efficient dependency-guided named entity recognition. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17, P3457–3465. AAAI Press, 2017.
9. Jenny Rose Finkel and Christopher D. Manning. Joint parsing and named entity recognition. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 326–334. Association for Computational Linguistics, Boulder, Colorado, 2009.
10. Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 344–354. Association for Computational Linguistics, Beijing, China, 2015. P15-1034.
11. Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5, 2017.
12. Salih Tuch. NEURAL DEPENDENCY PARSING FOR TURKISH. 2020. P. – 66.
13. Gulshen Eryigit, Joakim Nivre, and Kemal Oflazer. Dependency parsing of turkish. *Computational Linguistics*, 34(3):357–389, 2008.
14. Gulsen Eryigit and Kemal Oflazer. Statistical dependency parsing of turkish. 2006.
15. Kemal Oflazer. Turkish and its challenges for language processing. *Lang. Resour. Eval.*, 48(4):639–653, 2014.

16. Gulshen Eryigit. The impact of automatic morphological analysis & disambiguation on dependency parsing of Turkish. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 1960–1965. European Language Resources Association (ELRA), Istanbul, Turkey, 2012.
17. B.Elov, Sh.Hamraeva, D.Elova. Morfologik analizatorni yaratish usullari. O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(1). 67-87-b.
18. O.Abdullayeva. Til korpuslarida lingvistik annotatsiya va uning prinsiplari. O'zbek tilining milliy korpusi: muammo va vazifalar mavzusidagi xalqaro ilmiy-amaliy anjumani materiallari. – Toshkent, 31-may 2022-yil. – B. 130-136.
19. O.Abdullayeva. O'zbek tili korpusida matnlarni sintaktik annotatsiyalash masalasi. // O'zbek amaliy filologiyasi istiqbollari respublika ilmiy-amaliy anjumani materiallari. – Toshkent, 26-oktabr 2022-yil. – B. 122-126.
20. <https://www.baeldung.com/cs/constituency-vs-dependency-parsing>
21. [https://en.wikipedia.org/wiki/Syntactic_parsing_\(computational_linguistics\)#CITEREFJurafskyMartin2021](https://en.wikipedia.org/wiki/Syntactic_parsing_(computational_linguistics)#CITEREFJurafskyMartin2021)
22. https://en.wikipedia.org/wiki/Uzbek_language
23. <https://www.geeksforgeeks.org/constituency-parsing-and-dependency-parsing/>

O'ZBEK TILIDAGI SODDA GAPLAR UCHUN SINTAKTIK PARSING MODELI

Raxmonova Mohira Abdusattor qizi
ToshDO'TAU Kompyuter lingvistikasi
mutaxassisligi 2-kurs magistranti
mokhirata@gmail.com

Ilmiy rahbar: f.f.d. (DSc) Sh.Hamroyeva

Annotatsiya. Ushbu maqolada sintaktik analizatorning tahlil qilish algoritmi va tahlil qilish bosqichlari, o'zbek tilidagi sodda gaplarning o'zbekcha va inglizcha modellari ishlab chiqilgan. Buning uchun, avvalo, gaplarning sintaktik qoidalari formal grammatikadan foydalangan holda tavsiflanadi. Rasmiy grammatika sifatida Xomskiyning Context free grammatikasidan foydalanilgan.

Kalit so'zlar: *sintaktik analizator algoritmi, sodda gap modellari, Context free grammatikasi, POS tagging.*

Аннотация. В данной статье разработаны алгоритм и этапы анализа синтаксического анализатора, узбекская и английская модели простых узбекских предложений. Для этого, прежде всего, с помощью формальной грамматики описываются синтаксические правила предложений. Context free грамматика Хомского использовалась в качестве официальной грамматики.

Ключевые слова: *алгоритм синтаксического анализатора, модели простых предложений, Контекст-фрии грамматика, POS tagging.*

Abstract. In this article, an algorithm and stages of analysis of the Syntactic parser, Uzbek and English models of simple Uzbek sentences are developed. To do this, first of all, with the help of formal grammar, the syntactic rules of sentences are described. The Context free grammar of Chomsky was used as the official grammar.

Keywords: *Syntactic parser algorithm, simple sentence models, context-free grammar, POS tagging.*

Formal tillar va formal grammatika nazariyasi XX asrning 50-yillarida N.Xomskiyning generativ nazariyasidan kelib chiqdi. Ushbu nazariya sintaktik tuzilmalarini tahlil qilish uchun asos bo'ldi [3]. Avtomatik sintaktik tahlil (parsing) kompyuter lingvistikasi nazariyasining eng ko'p o'rganilayotgan sohasidir. Sintaktik analizator ma'lumotlar bazasi interfeyslarini loyihalashtirish, sun'iy intellekt, matnlarga avtomatik ishlov berish, avtomatlashtirilgan ma'lumot-qidiruv tizimi sohasi, mashina tarjimasida keng qo'llanilmoqda. Ta'kidlash kerakki, avtomatik sintaktik tahlil (ST) usullari ishi tabiiy tilning formal sintaktik nazariyalaridan kelib chiqadi [1; 78].

Sintaksis deganda tilshunoslikda asosiy birligi gap va eng kichik birligi so'z birikmalari bo'lgan sath nazarda tutilgan. Unda gap tarzida muayyan hukmni ifoda qiladigan grammatik izchil bog'langan so'zlar zanjiri tushuniladi. Grammatik izchil

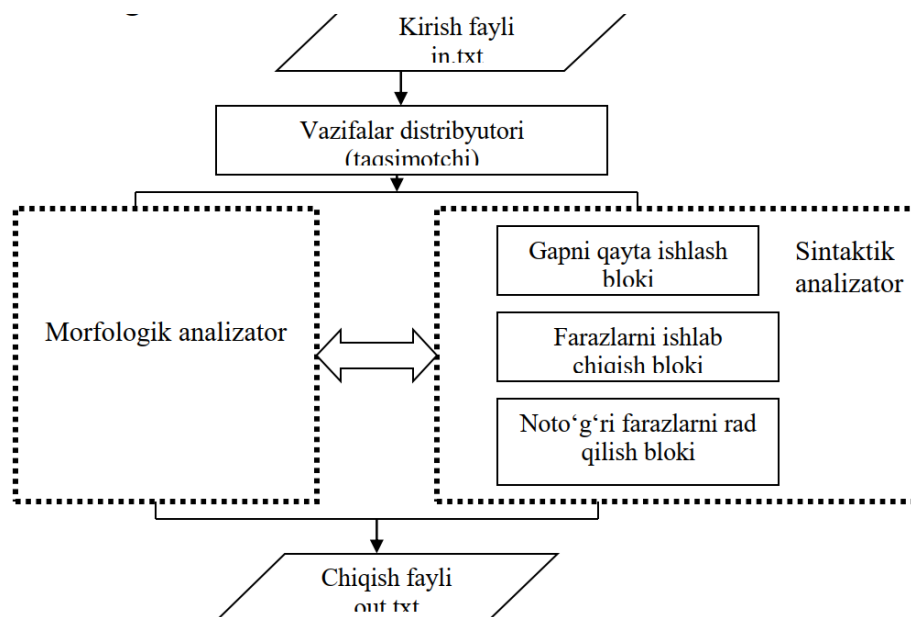


zanjir – so‘zshakllarning o‘zaro ma’lum bir grammatik munosabatlarda bog‘lanishidir. So‘zshakl yoki grammatik so‘z leksemaning o‘ziga biriktirishi mumkin bo‘lgan grammatik shaklda kelishidir. Shunday qilib, gapning sintaktik tuzilishi so‘z shakllarning o‘zaro sintaktik bog‘langan zanjiridan tashkil topadi [1; 78].

M.Abjalovanning ko‘rsatishicha, sintaktik tahlil moduli (STM) sintaktik tahlil qilish dasturi yoki lingvistik dasturning sintaktik tahlilni amalga oshirishga mo‘ljallangan qismi. Bugungi kunda avtomatik STMlarini yaratish kompyuter lingvistikasining dolzarb masalalaridan hisoblanib, buning yechimi turli amaliy maqsadlarni yuzaga keltirish uchun til sistemasini formallashtirishda yuqori sifatli natijalar imkonini beradi: til va nutq birliklarini avtomatik tekshirish sistemasini yaratish, Internetning qidiruv sistemasini mukammallashtirish shular jumlasidandir. Sintaktik tahlildan maqsad sintaktik strukturalar va sintaktik bog‘lanish me‘yorlarini kuzatishdir, biroq barcha tillarda ham STM yaratish murakkab ish hisoblanadi, chunki amaliy tilshunoslikda yetarlicha nazariy va amaliy tadqiqotlar mavjud emas; til va nutq birliklarining sintaktik bog‘lanishi individual xilma-xillikka ega [1; 82].

Har bir tilning sintaktik analizatori (parser) o‘sha til tabiati va qonuniyatlariga binoan yaratiladi. Parser ishi so‘zlar kategoriyasi (turkumlar) ma’lumotlari bazasiga asoslanadi va asosiy vazifasi ayrim tillarda jumlada gap bo‘laklarini aniqlash (mas., rus tilida) hisoblansa, muayyan tillarda, masalan, ingliz tilida N.Xomskiy nazariyasiga asosan otli guruh va fe‘lli guruh tarzida generativ tahlil qilinadi [1; 88-89].

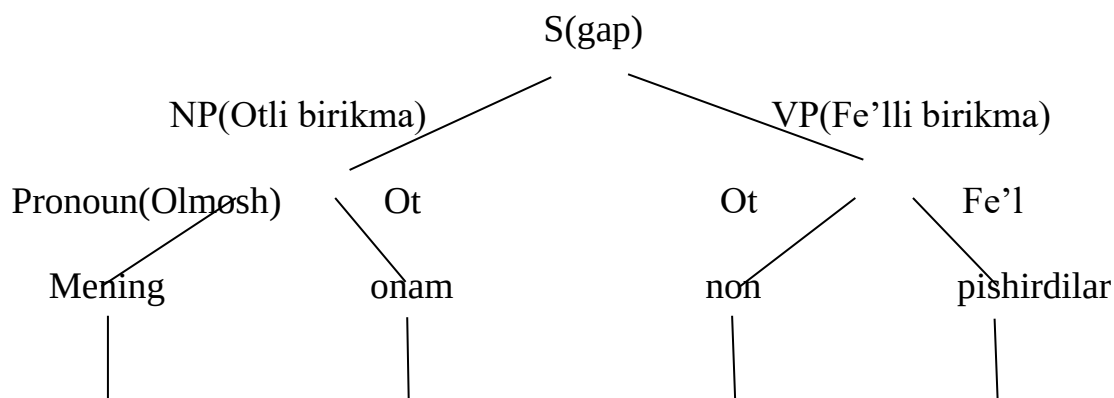
Quyida A.Leonteva va I.Kagirovlar tomonidan ishlab chiqilgan analizator dasturi blok-sxemasi berilgan. Bunda gaplarni sintaktik tahlil qilish bosqichlari keltirilgan.



1-chizma. A.Leonteva va I.Kagirovlar tomonidan analizator dasturi blok-sxemasi[6]

Kirish faylni oddiy gaplar tashkil etadi. Vazifalarni taqsimlovchi ma'lumotlarni tahlil qiluvchiga uzatadi. Ishlov berish blokida gaplar so'zma-so'z qayta ishlanadi. Asl so'z shakli morfologik tahlil blokiga topshiriladi, unda o'zaklarning barcha mumkin bo'lgan variantlari va unga mos keladigan grammatik ko'rsatkichlar tanlanadi. Har bir gap so'zma-so'z o'qiladi, shundan so'ng so'z shakli morfoanalizatorga kiritiladi. Natijada barcha mumkin bo'lgan o'zaklar va tegishli grammatik ko'rsatkichlar aniqlanadi. Agar berilgan so'z shakliga faqat bitta asos mos kelsa, u gipotezalarni qurish tartibiga kiradi. Gap qismiga va grammatik ko'rsatkichlarga ko'ra tegishli sintaktik guruh ajratiladi. Dastlabki ishlovdan so'ng so'z shakli gipoteza yaratish birligiga kiradi. Ushbu blok asosiy hisoblanadi. Uning kiritilishi so'z shaklidir. Bu gapdagi birinchi so'z bo'lsa, nutqqa mos ravishda sintaktik guruh aniqlanadi. Shunday qilib, morfologik va sintaktik noaniqlik tufayli bir gap uchun bir nechta farazlar tuzilishi mumkin. Bundan tashqari, bu farazlar noto'g'ri gipotezalarni tekshirish blokiga kiradi. Ushbu blokda tekshirishning ikki darajasi mavjud. Birinchi bosqichda sintaktik guruhlarning bir gap doirasidagi kelishiklari tekshiriladi. Chiqarish fayli gaplar ro'yxati bo'lib, ularning har biri sintaktik guruhlarga bo'linadi. Agar gapda lug'atda mavjud bo'lmagan so'z bo'lsa, u to'g'ridan-to'g'ri ko'rsatiladi [3; 399].

Sintaktik tahlilda jumla elementlarining sintaktik munosabatlari aniq belgilab qo'yilgan sintaktik daraxt, ya'ni tobelik daraxtiga asoslaniladi. Masalan: *Mening onam non pishirdilar* jumlasida quyidagicha tobeli qismlarga ajratilib, turkumlari bo'yicha tahlil qilinadi:



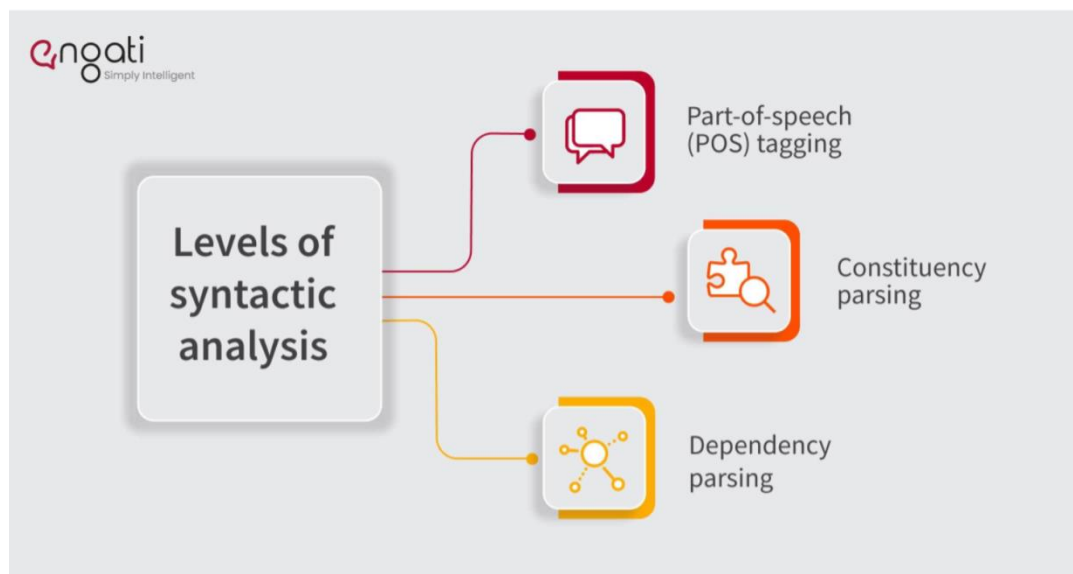
Bevosita ishtirokchilar (BI) metodiga ko'ra tayanch nuqta konstruksiya (qurilma) deb ataladi. Konstruksiya tarkibidan bevosita ishtirokchilar (BeI) va bilvosita ishtirokchilar (BiI) ajratiladi [1; 24-25].

O'zbek tili sintaktik analizatorining sintaktik tahlil algoritmi

An'anaviy grammatik vositalar yordamida gapni sintaktik tahlil qilishda quyidagi qadamlarni bajarish mumkin:

- 1) gap bo'laklari chegaralari va so'z chegaralarini aniqlash
- 2) bo'laklarini aniqlash;
- 3) tarkibiy qismlarni aniqlash;
- 4) tarkibiy qismlar uchun sintaktik kategoriyalarni aniqlash;
- 5) tarkibiy qismlarning grammatik vazifalarini aniqlash;
- 6) sintaktik strukturani chizish;
- 7) noto'g'ri tuzilgan gaplarni to'g'ri tartibga solish.

Sintaktik tahlilning maqsadi. Uning maqsadi kiritilgan matnining tuzilishini eng kichik asosiy bo'laklardan tortib gaplarga tushunish va undan mantiqiy ma'no chiqarishdir. Sintaktik tahlil tabiiy tilni qayta ishlashning (Natural Language Processing) juda muhim jihati hisoblanadi, chunki u har qanday jumlaning grammatik ma'nosini aniqlashga yordam beradi [4].



3-chizma. Sintaktik tahlil bosqichlari[2]

Sintaktik tahlil bosqichlari Part-of-speech tagging, Constituency parsing va Dependency parsing lardan iborat.

Gap bo‘laklarini (Part-of-speech) teglash. Bu sintaktik tahlilning birinchi bosqichidir. Gap bo‘lagini teglash sintaktik tahlilning muhim qismi bo‘lib, gapdagi so‘z (fe‘l, ot, sifat va boshqalar) va qo‘shimchalar sifatida belgilashdan iborat. Gap bo‘laklarini teglash gapning ma‘nosini tushunishimizga yordam beradi. Boshqa barcha tahlil usullari gap bo‘laklari teglaridan foydalanadi.

Guruhlarga ajratib tahlil qilish (Constituency Parsing). Guruhlarga ajratib tahlil qilish so‘zlarning gapdagi grammatik maqomiga ko‘ra, gapdan guruhlarga, ya‘ni kichik iboralarga ajratishni o‘z ichiga oladi. Otli birikmalar (Noun Phrases) va fe‘lli birikmalar (Verb Phrases) eng keng tarqalgan birikmalar hisoblanadi.

Tobe bog‘lanish (Dependency Parsing). Tobe bog‘lanish erkin so‘z tartibidagi tillarda keng qo‘llaniladi. Tobelikni tahlil qilishda so‘zlarning o‘rtasida bog‘lanish hosil bo‘ladi. Ikki so‘z o‘rtasida bog‘liqlik mavjud bo‘lsa, bir so‘z hokim, ikkinchisi esa tobe bo‘lak bo‘ladi [2].

O‘zbek tilida sodda gap

Sodda gap ma‘lum fikr, maqsadni ifodalovchi, grammatik va intonatsion tomondan shakllangan sintaktik birlikdir. Sodda gap tuzilishiga ko‘ra yig‘iq yoki yoyiq bo‘lishi mumkin. Bosh bo‘laklardan tarkib topgan gaplar – yig‘iq gap, agar bosh bo‘laklardan tashqari ikkinchi darajali bo‘laklar ham ishtirok etsa, yoyiq gap sanaladi. O‘zbek tilida so‘zlar tartibi asosan erkin bo‘lsa ham, sodda gapda har bir gap bo‘lagining odatda qo‘llanadigan o‘z o‘rni bor. Bu gapning strukturasi, gap bo‘lagining sintaktik ifodalanish usuli, u bilan bevosita bog‘liq bo‘lgan boshqa so‘zlarning o‘rni kabi hodisalar bilan belgilanadi. Shu asosda so‘zlarning to‘g‘ri (odatdagi) tartibi va teskari (odatdagidan chekinish) tartibi farq qilinadi [5; 179].

O'zbek tilida sodda gap modellari



Ushbu jadvalda 32 ta sodda gaplarning modellari keltirilgan. Sodda gap modellari o'zbek va ingliz tilida berilgan. Bundan tashqari, har bir gap modeli uchun Context free grammatikasi ishlab chiqilgan.

Context free Chomskiy grammatikasi (CFG) NLP da tarkibiy tuzilmalarni modellashtirish uchun eng keng tarqalgan rasmiy tizimdir. CFG bir qator qoidalar iboratdir. CFG G to'rtta parametr bilan belgilanadi. Context free grammatika - bu ma'lum bir rasmiy tilda barcha mumkin bo'lgan gaplarni yaratish uchun foydalaniladigan rasmiy grammatika.

Context free grammatikasida G to'rtta belgidan foydalaniladi: $G = (V, T, P, S)$

G – grammatika

T – terminal belgilarning cheklangan to'plamini

V – terminal bo'lmagan belgilarning cheklangan to'plami

P – ishlab chiqilgan qoidalari to'plamini

S – boshlanish belgisi.

E – ega, K – kesim, A – aniqlovchi, H – hol, T – to'ldiruvchi, S – subject, O – object, Adj – adjective, Adv – adverb, V – verb

№	Sodda gaplar	O'zbek tilidagi model	Ingliz tilidagi model	Context Free Grammar
1	Poyezd keldi	E+K	S+V	$S \rightarrow NP VP$ $NP \rightarrow S$ $VP \rightarrow V$ $S \rightarrow$ $Poyezd$ $V \rightarrow keldi$
2	Dalaga chiqmaysizmi?	H+K	Adv+V	$S \rightarrow VP$ $VP \rightarrow Adv V$ $Adv \rightarrow$ $dalaga$ $V \rightarrow chiqmaysizmi$



3	Kitobni o'qidi	T +K	O+V	S->VP VP->O V O->kitobni V->o'qidi
4	Anna kecha omadli edi	E+H+K	S+Adv+V	S->NP VP NP->S Adv S VP->Adv V V S->Anna Adv->kecha V->omadli edi
5	Men taqdimotga kechikdim	E+H+K	S+Adv+V	S->NP VP NP->S VP->Adv V S->Men Adv->taqdimotga V->kechikdim
6	Sabrina kecha to'pni dribling qildi	E+H+T+K	S+Adv+O+V	S->NP VP NP->S Adv S VP->Adv O V O V S- >Sabrina Adv->kecha O- >to'pni V->dribling qildi
7	Ali kimyo kitobini o'qiyapti	E+A+T+K	S+Adj+Adv+V	S->NP VP NP->S VP->Adj O V S->Ali Adj->kimyo O- >kitobini V->o'qiyapti
8	Men uchun quyosh chiqishi o'ziga xosdir	T+E+K	O+S+V	S->NP VP NP->O S VP-> S V O V S->quyosh chiqishi O->men uchun V- >o'ziga xosdir
9	Ali bu yomon qo'shiqni kuyladi	E+A+A+T+K	S+Adj+Adj+O+V	S->NP VP NP->S VP->Adj Adj O V S->Ali Adj->bu yomon O->qo'shiqni V- >kuyladi
10	Ali qo'shiqni kuyladi	E+T+K	S+O+V	S->NP VP NP->S VP->O V S->Ali O->qo'shiqni V- >kuyladi
11	Talabalar iborani bir oyda bitirdilar	E+T+H+K	S+O+Adv+V	S->NP VP NP->S VP->O Adv V S->Talabalar O- >iborani Adv->bir oyda V- >bitirdilar
12	Import qilingan mashinalar qimmat	A+E+K	Adj+S+V	S->NP VP NP->Adj S VP- >V Adj->Import qilingan S- >mashinalar V->qimmat
13	Bu muzqaymoq brendi ajoyib	A+E+K	Adj+S+V	S->NP VP NP->Adj S VP- >V Adj->bu S->muzqaymoq brendi V->ajoyib



14	Oq bulutlar bilan qoplangan	A+T+K	Adj+O+V	S->VP VP->Adj O V Adj- >Oq O->bulutlar bilan V- >qoplangan
15	Murat Orhun ertaga mashg'ulot o'tkazadi	E+H+T+K	S+Adv+O+V	S->NP VP NP->S VP->Adv O V S->Murat Orhun Adv- >ertaga O->mashg'ulot V- >o'tkazadi
16	Bu gullar ko'p suvga muhtoj emas	A+E+A+T+K	Adj+S+Adj+O+V	S->NP VP NP->Adj S VP- >Adj O V Adj->bu ko'p S- >gullar O->suvga V- >muhtoj emas
17	Bir o'rdak daryoni kesib o'tdi.	A+E+T+K	Adj+S+O+V	S->NP VP NP-> Adj S VP-> O V Adj-> bir S->o'rdak O- >daryoni V->kesib o'tdi
18	Ali kun bo'yi qo'shiq kuyladi	E+H+T+K	S+Adv+O+V	S->NP VP NP->S VP->Adv O V S->Ali Adv->kun bo'yi O->qo'shiq V->kuyladi
19	Karla baxtsiz xodisaga uchradi	E+A+T+K	S+Adj+O+V	S->NP VP NP->S VP->Adj O V S->Karla Adj->baxtsiz O->xodisaga V->uchradi
20	Men kinoga dugonam bilan bordim	E+H+T+K	A+Adv+O+V	S->NP VP NP->S VP->Adv O V S->men Adv->kinoga O->dugonam bilan V- >bordim
21	Gulnora kecha ishga kelmadi	E+H+H+K	S+Adv+Adv+V	S->NP VP NP->S VP->Adv Adv V S->Gulnora Adv- >kecha ishga V->kelmadi
22	Nilufar onasi uchun chiroyli gullar sotib oldi	E+T+A+T+K	S+O+Adj+O+V	S->NP VP NP->S O VP- >Adj O V S->Nilufar O- >onasi uchun gullar Adj- >chiroyli V->sotib oldi
23	Ushbu ko'cha doim gavjum bo'ladi	A+E+H+K	Adj+S+Adv+V	S->NP VP NP->Adj S VP- >Adv V Adj->ushbu S- >gullar Adv->doim V- >gavjum bo'ladi
24	Men kecha universitetga Hamid bilan keldim	E+H+H+T+K	S+Adv+Adv+O+V	S->NP VP NP->S VP->Adv Adv O V S->men Adv- >kecha universitetga O- >Hamid bilan V->keldim



25	Men kecha yaqin do'stim bilan universitetga keldim	E+H+A+T+H +K	S+Adv+Adj+O+Adv +V	S->NP VP NP->S Adv VP- >Adj O Adv V S->men Adv- > kecha universitetga Adj- >yaqin O->do'stim bilan V- >keldim
26	Biz Hamidning uyida imtihonlarga tayyorlandik	E+A+H+T+K	S+Adj+Adv+O+V	S->NP VP NP->S Adj Adv VP->O V S->biz Adj- >Hamidning Adv->uyida O- >imtihonlarga V- >tayyorlandik
27	Men sendan kitobimni kecha olmoqchi edim.	E+T+T+H+K	S+O+O+Adv+V	S->NP VP NP-> S O VP->O Adv V S->men O->sendan kitobimni Adv->kecha V- >olmoqchi edim
28	Men hech qachon hech kimdan hech narsa kutmayman	E+H+T+T+K	S+Adv+O+O+V	S->NP VP NP-> S Adv VP- >O O V S->men Adv->hech qachon O->hech kimdan hech narsa V->kutmaymanm
29	Men kecha universitetga Hamidning oyisi bilan keldim	E+H+H+A+T +K	S+Adv+Adv+Adj+O +V	S->NP VP NP->S Adv Adv VP->Adj O V S->men Adv- >kecha universitetga Adj- >Hamidning O->oyisi bilan V->bordimm
30	Biz do'stimning akasi bilan universitetga birga vaqtliroq keldik	E+A+T+H+H +H+K	S+Adj+O+Adv+Adv +Adv+V	S->NP VP NP->S Adj O VP->Adv Adv Adv V S->biz Adj->do'stimning O->akasi bilan Adv- >universitetga birga vaqtliroq V->keldik
31	Men senga bu buyumga qo'llaring bilan tegmasligingni bir necha bor aytgandim	E+T+A+T+T+ T+H+K	S+O+Adj+O+O+O+ Adv+V	S->NP VP NP->S O Adj O VP->O O Adv V S->Men O- >senga buyumga qo'llaring bilan Adj->bu Adv-> bir necha bor V->aytganman

Xulosa sifatida aytish joizki, o'zbek tili uchun yaratilgan sintaktik analizator sintaktik tahlil va parsing algoritmi uchun, tilshunoslarning ilmiy ishlari hamda boshqa analizatorlar yaratishda asos bo'lib xizmat qiladi. Sodda gaplar uchun tuzilgan o'zbekcha va inglizcha modellar asosida sintaktik analizator yaratish uchun algoritmlar ishlab chiqiladi.

Foydalanilgan adabiyotlar:

1. B.Elov, Sh.Hamroyeva, Z.Xusainova. NLP (tabiiy tilga ishlov berish)ning vazifalari va zamonaviy yondashuvlar. TerDU, filologik tadqiqotlar: til, adabiyot, ta'lim. 2022, 5-6.
2. O.Abdullayeva. Til korpuslarida lingvistik annotatsiya va uning prinsiplari. O'zbek tilining milliy korpusi: muammo va vazifalar mavzusidagi xalqaro ilmiy-amaliy anjumani materiallari. – Toshkent, 31-may 2022-yil. – B. 130-136.
3. O.Abdullayeva. O'zbek tili korpusida matnlarni sintaktik annotatsiyalash masalasi. // O'zbek amaliy filologiyasi istiqbollari respublika ilmiy-amaliy anjumani materiallari. – Toshkent, 26-oktabr 2022-yil. – B. 122-126.
4. <https://www.engati.com/glossary/syntactic-analysis>
5. https://www.hse.ru/data/2017/07/24//КЛ_задачи_подходы_ресурсы.
6. [https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/#:~:text=Tokenization%20is%20used%20in%20natural,into%20understandable%20parts%20\(words\)](https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/#:~:text=Tokenization%20is%20used%20in%20natural,into%20understandable%20parts%20(words)).
7. O'zbek tili grammatikasi. II t. Sintaksis. Mas'ul muharrirlar G'.A.Abdurahmonov, Sh.Shoabdurahmonov, A.Hojiyev. – Toshkent. 1976. – B. 179.
8. Леонтьева А., Кагиров И. Автоматический синтаксический анализ русских текстов // Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008. – С. 399.

NLP VAZIFALARI VA YO'NALISHLARI

Adilova Munojot

Mustaqil tadqiqotchi

adilovamunojot@gmail.com

Annotatsiya. Kompyuter lingvistikasi tilga oid muammolarni yechish bo'yicha dasturlarni shakllantirish bilan shug'ullanadi. Tabiiy tilni qayta ishlash (NLP) ancha murakkab jarayon hisoblanib, unda ijtimoiy tabiatga ega tilning barcha hodisalari, umumiy va xususiy jihatlari, istisnoli holatlari, fonetik, morfologik, leksik, grammatik, semantik va hatto orfoepik xususiyatlarini e'tiborga olish muhim sanaladi. Mazkur maqolada matnlarni kompyuter tahlili jarayonidagi ahamiyati, NLPning vazifalari va muammolari haqida so'z yuritiladi, shuningdek, NLPga to'siq bo'luvchi omillar va bajarilishi lozim bo'lgan vazifalar haqida to'xtalib o'tilgan. Tabiiy til matni yoki ovozli ma'lumotlarga mo'ljallangan, ma'nosini aniq belgilaydigan dasturiy ta'minotni yaratishdagi tildagi qiyinchiliklarni bartaraf etish vazifalari ham sanab o'tilgan.

Kalit so'zlar: *tabiiy tilni qayta ishlash, NLP, NER, PoS-tagger, teglash, modellashtirish, tabiiy til generatsiyasi.*

Annotation. Computational linguistics deals with the development of programs for solving language problems. Natural language processing (NLP) is a very complex process in which it is essential to consider all phenomena, general and particular aspects, exceptional cases, phonetic, morphological, lexical, grammatical, semantic, and even orthographic features of the language's social nature. This article talks about the importance of computer analysis of texts, the tasks, and problems of NLP, the factors hindering NLP, and the tasks that need to be performed. It also lists the challenges of overcoming language difficulties in creating meaningful software for natural language text or voice data.

Keywords: *Natural language processing, NLP, NER, PoS-tagger, tagging, modeling, natural language generation.*

Аннотация. Компьютерная лингвистика занимается разработкой программ для решения языковых задач. Обработка естественного языка (NLP) — очень сложный процесс, в котором важно учитывать все явления, общие и частные стороны, исключительные случаи, фонетические, морфологические, лексические, грамматические, семантические и даже орфографические особенности социальной природы языка. В данной статье говорится о важности компьютерного анализа текстов, сложностях и проблемах NLP, а также о факторах, препятствующих работе NLP и задачах, которые необходимо выполнить. Также перечислены задачи по преодолению языковых барьеров при создании значимого программного обеспечения для текстовых или голосовых данных на естественном языке.



Ключевые слова: *обработка естественного языка, NLP, NER, PoS-tagger, тегирование, моделирование, генерация естественного языка.*

Hozirgi raqamli texnologiyalar muhiti rivojlanayotgan damda sun'iy intellekt tizimini yaratishda matn birliklarini raqamli texnologiyalar orqali qayta ishlanishiga erishish muhim vazifa hisoblanadi. Bu jarayonda eng muhim vazifalar sirasiga so'zlar turkumini teglash kiradi. Jahon kompyuter lingvistikasidan ushbu lingvotexnik tahlilning so'zlar turkumini aniqlash – PoS-tagger, ya'ni so'z turkumlarini teglash, shuningdek, uning matnlarni avtomatik qayta ishlash jarayoni bosqichi ekani ma'lum. Til korpuslarini yaratish uchun boshlangan dastlabki teglash harakatlari bugungi kunga kelib, matn bilan bog'liq dolzarb masalalar yechimini kutmoqda. NLP, ya'ni tabiiy tilni qayta ishlash jarayonida ham so'z turkumlarini teglash birlamchi vazifa hisoblanadi.[1.13-14-b., 3.7-9-b.] Buning natijasida, avvalambor, omonimlikni, omofonlar, istehzolar (sarkazm), idiomalar, metaforalar, grammatika va foydalanishdagi istisnolar, shuningdek, ko'p ma'noli so'zlar semantikasini aniqlash kabi turli lingvistik noaniqliklar matn tarkibida tahlil qilinishiga erishiladi.

Tabiiy tilni qayta ishlash (NLP – *Natural language processing*) kompyuter fanlari sohasiga, aniqlash sun'iy intellekt (SI) yoki SI sohasiga tegishli bo'lib, kompyuterlarga matn va og'zaki (so'zlashuvdagi) so'zlarni xuddi insonlar kabi tushunish imkonini beruvchi tizim hisoblanadi [6.3-6-b., 14.40-48-b.]. NLP kompyuter lingvistikasi – sun'iy intellekt tizimlarini yaratishda inson va hayvonlarning intellektual jarayonlarini matematik va kompyuter modellashtirish sohasidagi ilmiy yo'nalish bo'lib, u tabiiy tillarni tavsiflashda matematik modellardan foydalanish tushuniladi. Inson tilini qoidalarga asoslangan holda modellashtirish – statistik tahlil, mashina tarjima va ma'lumotlarni chuqur o'rganish modellari bilan birlashtiradi. Bu texnologiyalar birgalikda kompyuterlarga inson tilini matn yoki ovozli ma'lumot sifatida qayta ishlash va uning to'liq ma'nosini, jumladan, ma'ruzachi yoki yozuvchining maqsadi va his-tuyg'ularini “tushunish” imkonini beradi.

NLP matnni bir tildan ikkinchi tilga tarjima qiladigan, og'zaki buyruqlarga javob beradigan va katta hajmdagi matnni tez, hatto real vaqtda umumlashtiradigan kompyuter dasturlarini boshqaradi. Ijtimoiy hayotda bu holatlarga juda ko'p murojaat qilinadi: ovozli GPS tizimlari, raqamli yordamchilar, nutqiy sintezator, mijozlarga xizmat ko'rsatish bo'yicha chatbotlar va boshqa iste'molchilar uchun qulayliklar ko'rinishida NLP bilan o'zaro aloqada bo'lishingiz uchun yaxshi imkoniyatlar bor. Ammo NLP, shuningdek, biznes operatsiyalarini tartibga solishga, bank, shifoxona, davlat xizmatlari kabi muassasalarda xodimlarning mahsuldorligini oshirishga va muhim biznes jarayonlarini soddalashtirishga yordam beradigan, korporativ yechimlarda muhim rol o'ynaydi. Bu holat insonga aniqlik, tezlik va navbatlardagi muammolarning yuzaga kelishining oldini oladi. Ba'zi hollarda inson omili qatnashmagan jarayonlar aniq bajariladigan jarayonlar ham uchraydi. Masalan, *bankomatdan pul yechib olasiz yoki aksincha valyutani o'z plastik kartangizga joylaysiz, shuningdek, to'lovlar jadvalida ham shu jarayon aks etadi. Siz kommunal*

yoki davlat xizmatlariga mobil ilova yoki bankomatlar orqali to'lashingiz va sanoqli daqiqalarda elektr yoki gaz avtomatik tarzda ishga tushishi mumkin. Ushbu jarayonlarning amalga oshishi uchun ham texnologiya va inson muloqoti zarur hisoblanadi.

Quyida NLP vazifalari va unda uchraydigan muammolar xususida qisqacha tushuntirib o'tildi.[19.15-25-b., 20.10-18-b.,]

NLPga to'siq bo'luvchi omillar va bajarilishi lozim bo'lgan vazifalari. Tabiiy tili matni yoki ovozli ma'lumotlarning mo'ljallangan ma'nosini aniq belgilaydigan dasturiy ta'minotni yozishni nihoyatda qiyinlashtiradigan noaniqliklarga to'la. Birinchi bo'lib uchraydigan muammolar: omonimlar, omofonlar, istehzolar (sarkazm), idiomalar, metaforalar, grammatika va foydalanishdagi istisnalar, jumlar tuzilishidagi o'zgarishlar inson tilidagi “qonunbuzarlik”larning faqat bir qismi bo'lib, odamlarga yillar davomida o'rganilishi kerak bo'ladi.[29.10-16-b.,] Ammo dasturchilar Tabiiy tildagi ilovalarni tanib olish va tanib olishni o'rgatishlari kerak. Ular ushbu ilovalar foydali bo'ladimi yoki yo'qligini aniq bilishi lozim, va buni bir nechta tajribalar isbotlab beradi. Tabiiy tilni qayta ishlash jarayoni kompyuter lingvistikasidagi muhim vazifalardan biri hisoblanib, til mezonlari kompyuter tushunadigan, raqamli ko'rinishga keltiriladi. Tildagi ijtimoiy hodisalar kompyuter tushunadigan tilda beriladi. Ba'zi NLP vazifalari inson matni va ovozli ma'lumotlarini kompyuterga nimani o'zlashtirayotganini tushunishga yordam bermaydi. Tilimizda shunday omillar borki ular NLP'ga xalaqit beradi. Ushbu vazifalarning ba'zilar quyidagilarni o'z ichiga oladi:

1. **Nutqni tanish** (ingl. *Speech to text*; rus. *Распознавание речи*;) – nutqni matnga aylantirish ham deyiladi, ovozli ma'lumotlarni matn ma'lumotlariga leksik, grammatik xatolarsiz aylantirish vazifasidir. [5.30-40-b., 14.10-16-b.,] Ovozli buyruqlarni bajaradigan yoki ovozli savollarga javob beradigan har qanday ilova uchun nutqni aniqlash talab qilinadi. Nutqni tanib olishni ayniqsa qiyinlashtiradigan narsa – bu odamlarning gapirish usuli: shevalardan foydalanishlari, tez gapirishlari, noaniq talaffuz, so'zlardagi urg'uning noto'g'ri qo'yib aytilishi va intonatsiyalar, orfoepik xatolarga yo'l qo'yilishi kabilar nutqni tanish jarayonida ko'p muammolarni yuzaga keltiradi.

2. **So'z turkumlarini teglash** (ing. *Part of speech tagging*; rus. *Тегирование частей речи*;) – shuningdek, grammatik teglash ham deb yuritiladigan jarayondir. Nutq belgilarining bir qismi ma'lum bir so'z yoki matnning nutq qismini uning qo'llanilishi va kontekstiga qarab aniqlash jarayonidir, ya'ni so'z yoki tekst fragmentining qaysi so'z turkimiga tegishli ekani qo'llanadigan manba yoki kontekstga qarab aniqlanadigan jarayon hisoblanadi.[20.14-15-b., 21.19-29-b.,] So'z turkumlarining identifikatsiyalashishi bo'yicha quyidagi misollari keltirish mumkin:

Kez1 – “ravish, payt holi” – M: Kezi kelganda aytish kerak deyishadi (kezi-palla, payt);

Kez2 – “fe'l” – M: Bog'ni kezayotib uni ko'rib qoldi (kezmoq – sayr qilmoq, aylanib yurmoq) shaklida kechadi. Ushbu so'z qaysi so'z turkimiga mansubligini

faqat kontekstda bilish mumkin. Yoki leksik polisemiyaning asosiy xususiyati shuki, soʻz qanchalik koʻp maʼnoli boʻlmasin, leksik maʼnolar orasida maʼlum munosabat, bogʻlanish boʻladi. Buni “joy” soʻzining leksik maʼnolari misolida koʻrishimiz mumkin.

Joy1. Aravalar joyidan quzgʻalishdi (S.Ahmad). – “Shaxs yoki narsa-predmet ishgʻol qilgan, egallagan, band qilgan yuza, sath, makon” maʼnosida.

Joy2. Nizomjon qurilish boʻlayotgan joyda mashina gurillaganini eshitgan edi. (S.Ahmad) – “Voqea yoki hodisa sodir boʻladigan, ish-harakat bajariladigan yuza, sath, punkt” maʼnosida.

Joy3. Yunusobod degan joydan kelayapman. (Oybek) – “Qandaydir ism bilan ataladigan geografik maydon, yuza, sath, territoriya” maʼnosida.

Joy4. Boylar, boyvachchalar anoyi emas, unga oʻz mehmonxonalaridan joy berib. (Oybek) – “Yashash uchun moʻljallangan yuza, makon” maʼnosida.

Joy5. Choʻlda yigirma sotix joyga oʻzim uchun sholi ekdim. (S.Ahmad) – “Ekin ekiladigan sath, yuza, makon” maʼnosida.

Joy6. Kifoyatxon joy soldi va boshini roʻyjaga burkab yotdi. (A.Qahhor) – “Yotish uchun moʻljallab biror sathga, yuzaga toʻshalgan koʻrpa, yostiqlik va shu kabilarning yigʻindisi” maʼnosida. Bu misollar asosida “joy” soʻzining 6 ta leksik maʼnosi oʻzaro yuza, sath, makon maʼno komponentlarda bir-biri bilan bogʻliq ekanini sezish qiyin emas. Lekin ularning barchasini teglash zarur chunki mashina ularni bir-biridan ajrata olmaydi.

3. Soʻz maʼnosini ajratish (ing. Word Sense Disambiguation (WSD); rus. Устранение неоднозначности смысла слова;) — maʼlum kontekstda koʻp maʼnoli soʻzni aniqlaydigan semantik tahlil jarayoni orqali koʻp maʼnoli soʻzning maʼnosini tanlash.[10.10-12-b.] Masalan, soʻzning maʼnosini ajratib koʻrsatish “surma” (koʻz boʻyogʻi) va “surma” (buyruq feʼl-biror narsani surma)dagi, “surma stol” (sifat-qavatlar surib ochiladigan buyum) soʻzning maʼnosini farqlashga yordam beradi. Bu jihat teglashda eng muhim jarayon hisoblanadi.

4. Nomga ega birliklarni aniqlash (ing. Named Entity Recognition (NER); rus. Распознавание именованных сущностей;) — Bu inson tilini qayta ishlash texnologiyasining yoʻnalishi boʻlib, dasturiy taʼminotni amalga oshirish nutq va matndagi soʻzlar va iboralarning obyektiv toifalarini topishga imkon beradi. [15.2-3-b.] Namuna tariqasida, Farhod bozorini – savdo rastalari joylashgan makon, Farhod esa atoqli ot, ism sifatida soʻzlar bazasiga kiritilishi kerak. Bunday soʻzlar jamlanmasi minglab topiladi va shunday soʻzlar NERni belgilash zarurati eng muhim vazifalardan biri ekanini eslatib turadi.

5. Anoforik tahlil – (ing. Anaphoric analysis; rus. Разрешение совместной ссылки;) — Oʻzidan oldingi soʻz, qaysi birlikka ishora koʻrsatilayotganini aniqlaydi. Bu tahlil ikki soʻz bir xil obyektga tegishli yoki yoʻqligini va qachon ikki xil maʼnoda kelishini aniqlash vazifasi hisoblanadi.[16.2-3-b.] Eng keng tarqalgan misol, maʼlum bir olmosh ishora qiladigan shaxs yoki obyektning aniqlashidir (masalan, “u” = “Abdugʻani”, “u”=“bank yoki bino”), baʼzan bitta soʻz u matndagi metafora yoki

idioma ta'rifini ham o'z ichiga olishi mumkin (masalan, “tulki” bu misol hayvon emas, balki ayyor odam tasviri bo'lishi mumkin).

6. His-hayajon so'zlarni aniqlashtirish va ularning tasnifi (ing. *Sentiment Analysis*; rus. *Анализ настроений, Анализ тональности*);). Tuyg'u va his-hayajon so'zlar va brikmalarni shuningdek, tuyg'uni anglatuvchi so'zlarni tahlil qilish, matndan subyektiv sifatlarni – munosabat, hissiyot, istehzo, chalkashlik, shubhalarni ajratib olishga harakat qilinadi.[13.2-4-b] Tabiiy til makonida his-hayajonni anglatuvchi so'zlarni qo'llamay turib ham his-hayajonni izhor qiluvchi, anglatuvchi so'zlar qo'shilmasi, gaplar mavjud. Buni faqat kontekstni o'qib anglash mumkin, lekin buni mashina tarjima qilayotganda ularni ajratib olishi bu so'zlarni yoki so'zlar qo'shilmasini tasniflash teglash lozim. Masalan, *Bu yilning eng ajoyib va qiziqarli filmlaridan biri deb hisoblayman*; – ing. *One of the most amazing and enjoyable films of the year*; rus. *Один из самых удивительных и приятных фильмов года*; (Muallif bu gapi bilan unga bu film yoqqanini bildirmoqda va bu gap ijobiy his-tuyg'uni anglatyapti). *Men ushbu suv o'tkazmaydigan kamera korpusini sotib oldim, chunki u odatiy kameradan ko'ra bardoshli bo'lishi kerak*; – ing. *I bought this waterproof camera case because it needs to be more durable than the stock one*; rus. *Я купил этот водонепроницаемый чехол для камеры, потому что он должен быть более надежным, чем стандартный*; Kontekstdan ko'rinib turibdiki, bu odam kutgan narsa emas. Gap salbiy ohangga ega, lekin u yashirin. Kayfiyat subyektivlik yoki obyektivlikka bog'liq emas, bu tahlilni murakkablashtirishi mumkin. Ammo fikr-mulohazalar ma'lumotlaridan qimmatli tushunchaga ega bo'lish uchun biz hali ham ifodalangan his-tuyg'ular, baholar yoki munosabatlarga ega bo'lmagan jumalarni ajrata olishimiz kerak.

7. Tabiiy til generatsiyasi – (ing. *Natural Language Generation*, rus. *Генерация естественного языка*);) – ba'zan nutqni tanib olish yoki nutqni matnga aylantirishga qarama-qarshi sifatida tavsiflanadi; tuzilgan ma'lumotni inson tiliga tarjima qilish vazifasidir. Matn va nutqni yaratish. Qarama-qarshi aniqlab olish vazifasi – generatsiya yoki sintez. Algoritm foydalanuvchi matnini yoki nutqiga javob berishi kerak. Bu savolga javob, foydali ma'lumot yoki kulgili ibora bo'lishi mumkin, ammo izoh berilgan mavzuda taaluqli bo'lishi kerak. Nutqni aniqlash tizimlari gaplarni qismlarga ajratadi. Bundan tashqari, ma'lum bir iborani talaffuz qilish uchun kompyuter ularni saqlaydi, o'zgartiradi va ko'paytiradi. Albatta, so'zlarning birikish chegaralarida buzilish bo'lishi mumkin, shuning uchun ovoz ko'pincha g'ayritabiiy eshitiladi. Matn yaratish algoritmgaga kiritilgan shablon javoblari bilan cheklanmaydi. U mashinani o'rganish algoritmlaridan foydalanadi. “So'zlashuvchi” dasturlar haqiqiy ma'lumotlarga tayangan holda o'rganishi mumkin. Mantiqiy tuzilishga ega she'rlar yoki hikoyalar yozish algoritmini olish mumkin, lekin ular odatda juda mazmunga boy bo'lmaydi.[17.45-48-b., 18.17-19-b.] Algoritm ma'lumotlar bazasidagi mavjud faktga tayangan holda savollarga javob beradi. NLG ma'lum kalit so'zlar yoki mavzular asosida odamlar kabi og'zaki matn yaratishga qaratilgan. Hozirda butun dunyoda mashina tarjimasiga alohida urg'u berilmoqda. Agarda

algoritmlar ishlashi uchun ma'lumotlar bazasini yaratsak (meshok slova), so'zlarning barchasi teglansa, tarjima samaradorligi oshadi, shuningdek, tarjima sifatiga ham ijobiy tasir qiladi. Hozirgi kunda ijtimoiy soha obyektlarida sun'iy intellektdan faol qo'llash ommalashdi. Bu chalkashlik va tartibsizliklarning oldini olmoqda. Misol qilib hozirgi kundagi faol jarayonlarni aytishimiz mumkin. NLP imkoniyatlariga ega aqlli chatbot mijozlar bilan mijozlarga xizmat ko'rsatish xodimlari kabi muloqot qilishi mumkin. Face ID tizimi orqali biron binoga kirish yoki biron programmani ishga tushurish mumkin. Barmoq izlari orqali bir nechta operatsiyalarni parol termasdan ochish, yopish, ishga tushurish va aksincha faoliyatini to'xtatish imkoni mavjud. Va bu, albatta, inson omili bilan birga bajariladigan jarayonlar hisoblanib, ijtimoiy hayotning yengillashishiga, qog'ozbozlikning kamayishiga va aniqlikka yordam beradi.

Xulosa

Ushbu tadqiqotdan ko'zlangan maqsad imkoniyati cheklangan insonlar uchun, shuningdek, ta'lim jarayonida, kundalik va rasmiy faoliyatda kompyuterdan foydalanish imkoniyatini oshirish, anjumanlardagi ilmiy munozaralar nutqining sifatli ko'chirilishi, yetkazib berilishi va to'g'ridan to'g'ri yozib olinishi, tarjima samaradorligini oshirish, o'zbek tilidan boshqa tillarga va aksincha, boshqa tillardan o'zbek tiliga matnlarni tarjima qilishda aniqlik va sifatning muhimligini oshirishga qaratilgan dasturiy ta'minotlar uchun tabiiy tilni qayta ishlash zaruratini aniqlash hisoblanadi. Tarjima dasturlarini takomillashtirish, shuningdek, so'z turkumlarini teglash algoritmlarni yaratish avtomatik tahlil, morfoanaliz va tarjimon dasturlari uchun birlamchi shart amalga oshirilsa, matnlarning grammatik jihatdan sifatli tahlil qilinishiga erishiladi. Lingvistik bilimlar kompyuter dasturlari va til korpuslari bazasi uchun eng zarur ma'lumot manbayi hisoblanadi. Lingvistik protsessorda lisoniy ma'lumotlar bazasini shakllantirishda, kundalik hayotda til bilimlaridan foydalanish ko'nikmalariga asoslanilsa, bunday vaziyatlar adabiy til me'yorlariga aylantirilsa, kompyuter dasturlarining tahlil imkoniyati mutaxassis darajasida mukammallikka yetadi. Aynan so'z (asosan, polifunksional, ko'p ma'noli, omonim so'zlar)ning turkumini belgilashda ham maqola avvalida “joy” so'zi misolidagi kabi so'zning pragmatik va kontekstual ma'nolari hamda qo'shimcha semalariga asoslanilsa, raqamli texnologiya dasturlari va tizimlarining amaliy ahamiyati yanada oshadi. Umuman olganda, bugungi kunda lingvistik bilimlar tildan amalda foydalanish imkoniyatiga asoslanib mukammallashtirilsa, nazariy manbalarga tayanilib, kompyuter dasturlari va tizimlari uchun formal o'zbek tili yaratilsa, kelajakda barcha turdagi (o'zbek tarjimon dasturi, nutqni tanish, o'zbek tili morfoanalizatori, kompyuter muloqoti kabi) dastur hamda elektron tizimlarining yuzaga kelishiga zamin yaratiladi.

Foydalanilgan adabiyotlar:

1. Abjalova, M. 2020. Tahrir va tahlil dasturlarining lingvistik modullari. Toshkent: Nodirabegim.

2. Abjalova, M., Yuldashev, A. 2021. "Methods for Determining Homonyms in Linguistic Systems". ACADEMICIA: An International Multidisciplinary 17
3. Abjalova, M., Tabiiy tilni qayta ishlash (NLP)da so'z turkumlarini teglash masalasi. Research Journal 11 (2): 700-715. DOI: 10.5958/2249-7137.2021.00522.
4. Abjalova, M. 2021. "O'zbek tili Milliy korpusida so'zshakllarni leksikografik baza asosida qidiruv imkoniyatlari". Kompyuter lingvistikasi: muammo, yechim, istiqbollar, 12-17. Toshkent: ToshDO'TAU.
5. Abjalova, M. 2021. "O'zbek tili milliy korpusida sinonimayzer yoxud sinonimizatorni yaratish masalasi". O'zbek Milliy va ta'limiy korpuslarining yaratishning nazariy hamda amaliy masalalari, 38-40. Toshkent: ToshDO'TAU. O'zbek tilining izohli lug'ati. 2006. 80 000 dan ortiq so'z va so'z birikmasi. A. Madvaliyev tahriri ostida. 5 jildli. Toshkent.
6. B.Brill, E. 1995. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging". Computational Linguistics 21: 543-565. <http://acl.ldc.upenn.edu/J/J95/J95-4004.pdf>. Brill, E. 1992. "A simple rule-based part of speech tagger". Proceedings of ANLC 154.
7. Baum, L. E.; Sell, G. R. 1968. "Growth transformations for functions on manifolds". Pacific Journal of Mathematics 27 (2): 211-227.
8. Elov B., Tabiiy tilni qayta ishlash (NLP)da SpaCY modulidan foydalanish, 42-43. Toshkent: ToshDO'TAU.
9. Ernst, Jason; Kellis, Manolis. 2012. "ChromHMM: automating chromatinstate discovery and characterization". Nature Methods 9 (3): 215– 216. doi:10.1038/nmeth.1906. PMC 3577932. PMID 22373907.
10. Pal A.R., Saha D.J Word sense disambiguation: A survey, 2015.
11. Rizayev, S. 2006. O'zbek tilshunosligida lingvostatistika asoslari. Toshkent: Fan
12. Sayfullayeva R., Mengliyev B., Boqiyeva G., Qurbonova M., Yunusova Z., Abuzalova M. Hozirgi o'zbek adabiy tili. – T.: Fan va texnologiya, 2010.
13. S.Sorokin. B.N Sintez rechi. -M.: Nauka, 1992, 390-392
14. Stanford Log-linear Part-Of-Speech Tagger. – 2016. – Access mode: <https://nlp.stanford.edu/software/tagger.html>
15. Stanford Named Entity Recognizer (NER). – 2016. – Access mode: <https://nlp.stanford.edu/software/CRF-NER.shtml>.
16. Stanford Named Entity Recognizer (NER). –2016. – Access mode: <https://nlp.stanford.edu/software/CRF-NER.shtml>.
17. Starner, Thad, Pentland, Alex. 1995. Real-Time American Sign Language Visual Recognition From Video Using Hidden Markov Models. Master's Thesis, MIT, Program in Media Arts. Li, N;



18. Stephens, M. 2003. “Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data”. *Genetics* 165 (4): 2213-2233. doi:10.1093/ genetics/165.4.2213.
19. К Кравченко Ю.А. Мансур, А.М. Мохаммад, Ж.Х. Модифицированный метод устранения неоднозначности смысла слова, основанный на методах распределенного представления DOI 10.18522/2311-3103-3-92-101 УДК 04489
20. K.Toutanova Chr. D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. – Access mode: <https://nlp.stanford.edu/~manning/papers/emnlp2000.pdf>.
21. Natural Language Toolkit. – Access mode: <http://www.nltk.org/>.
22. <https://asu-analitika.ru/analiz-nastroenij-tipy-instrumenty-i-scenarii-ispolzovaniya>
23. https://en.wikipedia.org/wiki/Hidden_Markov_model.
24. <https://www.freecodecamp.org/news/an-introduction-to-part-ofspeech-tagging-and-the-hidden-markov-model-953d45338f24>.
25. <https://coderlessons.com/tutorials/akademicheskii/obrabotkaestestvennogo-iazyka/pometka-chasti-rechi-pos>.
26. <https://habr.com/ru/post/125988>.
27. https://ru.wikipedia.org/wiki/Частеречная_разметка.
28. https://en.wikipedia.org/wiki/Part-of-speech_tagging
29. https://ru.wikipedia.org/wiki/Часть_речи

KORPUS LINGVISTIKASI

O'ZBEK-INGLIZ TILLARI PARALLEL KORPUSIGA QO'YILADIGAN LINGVISTIK VA EKSTRALINGVISTIK TALABLAR

Elov Botir Boltayevich

ToshDO'TAU Ijtimoiy-gumanitar fanlar va axborot
texnologiyalari fakulteti dekani, t.f.f.d. (PhD), dotsent,

Amirqulov Ma'rufjon Aliqul o'g'li

ToshDO'TAU Kompyuter lingvistikasi mutaxassisligi
2-kurs magistranti

Annotatsiya: Maqolada korpus tilshunosligida muhim rol o'ynovchi parallel korpuslar, xususan, o'zbek korpusshunosligida yaratilishi zaruriyatga aylangan o'zbek-ingliz parallel korpusiga qo'yiladigan talablar xususida so'z yuritiladi. Albatta korpusga, ayniqsa, parallel korpuslarga qo'yiladigan talablar kundan-kunga ortib bormoqda. Quyida o'zbek-ingliz parallel korpusiga qo'yiladigan lingvistik va ekstralingvistik talablarga bafurja to'xtalib o'tilib, jahon korpusshunosligi tajribasiga tayangan holda ushbu talablarga qanday javob berilishi, bu jarayonlarni qanday amalga oshirilishi zarurligi haqidagi fikrlar beriladi.

Kalit so'zlar: *Korpus, teglash, tenglashtirish, parallel korpus, Pos teglash, XML kodlash.*

Abstract: The article talks about parallel corpora, which play an important role in corpus linguistics, in particular, the requirements for the Uzbek-English parallel corpus, the creation of which has become a necessity in Uzbek corpus linguistics. Of course, the requirements for the corpus or corpora, especially for parallel corpora, are increasing day by day. Below, the linguistic and extra-linguistic requirements for the Uzbek-English parallel corpus will be discussed in detail, and based on the experience of world corpus studies, opinions will be given about how these requirements should be met and how these processes should be implemented.

Key words: *Corpus, tagging, alignment, parallel corpora, POS tagging, XML encoding.*

Аннотация. В статье говорится о параллельных корпусах, играющих важную роль в корпусной лингвистике, в частности о требованиях к узбекско-английскому параллельному корпусу, создание которого стало необходимостью в узбекской корпусной лингвистике. Конечно, требования к корпусу или корпусам, особенно к параллельным корпусам, растут день ото дня. Ниже будут подробно рассмотрены лингвистические и экстралингвистические требования к узбекско-английскому параллельному корпусу, и на основе опыта изучения мировых корпусов будут даны мнения о том, как эти требования должны быть выполнены и как эти процессы должны быть реализованы.

Ключевые слова: *Корпус, тегирование, выравнивание, параллельные корпуса, POS-тегирование, XML-кодирование.*

Kompyuter lingvistikasining eng yuqori imkoniyatlariga ega yo‘nalishlaridan biri hisolanmish **korpus lingvistikasi** so‘nggi yillarda ko‘plab tatqiqotlar, izlanishlar markaziga aylanib ulgurdi. Ma‘lumotlarning yozma shakldan elektron ko‘rinishga o‘tishi lisoniy birliklarni ham chetlab o‘tgani yo‘q, albatta. Bu esa til birliklari ustida bajariladigan vazifalarning bir qadar yengillashuviga olib keldi. Xususan, sarflanadigan vaqtning tejami bu afzalliklarning yorqin misollaridan biri bo‘lsa, birdan ortiq tillardagi lisoniy birliklarning o‘zaro elektron shaklda tenglashtirilishi tarjimashunoslik, qiyosiy lingvistika, lug‘atshunoslik hamda bir necha turli sohalardagi yangi yo‘nalishlarga keng yo‘l ochdi.

Korpuslar, ayniqsa, parallel korpuslar til va tarjima sohasining mashina tarjimasi, krosslingval ma‘lumotlarni qayta ishlash, qiyosiy va tilga doir tadqiqotlar, tilni o‘rganish va o‘rgatish hamda bilingval leksikografiyada beqiyos o‘ringa ega hisoblanadi. O‘zining tilini yanada tanitish, uning xususiyatlarini barcha uchun qulay ko‘rinishga olib kelishda ham parallel korpuslar muhim ahamiyat kasb etadi. Shu jihatlarni hisobga olgan holda korpus lingvistikasiga bo‘lgan e‘tibor so‘nggi o‘n yilliklarda, ayniqsa, ortib bormoqda.

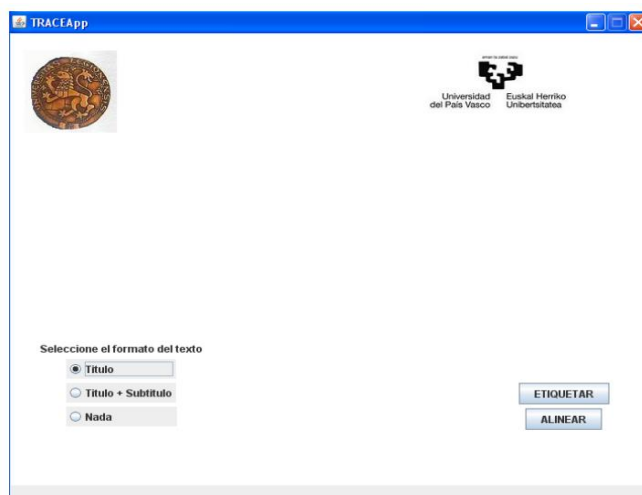
Yurtimizda kompyuter lingvistikasi sohasidagi izlanishlar, yutuqlar ko‘lami kundan-kunga ortib bormoqda va *“mamlakatimizda global ilmiy-texnik islohotlar amalga oshirilayotgan bugungi kunda kompyuter lingvistikasi doirasida avtomatik tarjima, o‘zbek tilini tushunish va qayta ishlash, uni boshqa tillar bilan chog‘ishtirishda sun‘iy intellektning rolini belgilash, ikki va ko‘p tilli parallel korpuslar yaratishga kuchli ehtiyoj sezilmoqda.”*[1] Ushbu korpus va uning yaratilishi *“davlat tilining sofliqini saqlash, uni boyitib borish va aholining nutq madaniyatini oshirish; davlat tilining zamonaviy axborot texnologiyalari va kommunikasiyalariga faol integratsiyalashuvini ta‘minlash”*[2]da o‘zining munosib hissasini qo‘shadi.

Parallel korpuslarning ilk elektron namunalari bundan 30 yil oldin yaratilishni boshlagan bo‘lsa, har bir til o‘zining lingvistik hamda dasturiy imkoniyatlaridan kelib chiqib parallel korpuslarni yaratish xususida fikrlar yuritishgan. Ba‘zi tillar juftligidagi parallel korpuslar gap kesimida tenglashtirilgan bo‘lsa, ba‘zilari hattoki so‘zlar kesimida ham tenglashtirilganiga guvoh bo‘lamiz. Ayrimlari lingvistik va ekstralingvistik teglashni maromiga yetkazgan bo‘lsa, qolganlari esa bu jihatlarga qisman e‘tibor qaratishgan. Korpusning foydalilik koeffitsenti nafaqat uning qay darajada tenglashtirilgani bilan, balki uning qay darajada ekstralingvistik va lingvistik teglanganligiga ham bog‘liq.

Lingvistik teglash – matndagi ma‘lumotlarning tilning grammatik, sintaktik, morfologik, semantik qoidalariga asosan izohlanishi.

Ekstralingvistik teglash – parallel korpusdagi matnlar haqida qo‘shimcha ma‘lumotlar biriktirish, ularni izohlash. Masalan, matnning kimga tegishli ekanligi, qachon va qayerda yozilganligi, kim tomonidan tarjima qilinganligi, janri, so‘zlar soni va hkz.

Parallel korpuslarning ilk prototiplari faqatgina matnni abzaslar yoki gaplar kesimida tenglashtirishga qaratilgan bo'lsa, hozirgi kunda yaratilayotgan korpuslarning imkoniyatlari bundan ancha keng ko'lamda ekanligini kuzatishimiz mumkin. Quyida nisbatan oddiy darajada lingvistik va ekstralingvistik teglangan Aleuska(nemiz-istan-bask) uch tilli parallel korpusi bilan tanishamiz. Naroa Zubillaga, Zurine Sanz va Ibon Uribarrilar[3] tomonidan yaratilgan korpus uchta tilni o'z ichiga qamrab oladi. Korpusning maqsadi nemis tilidan bask tiliga qilingan bevosita tarjimalarni o'rganish maqsadida yaratilgan. Bask tiliga ispan tilidan bilvosita qilingan tarjimalar ham o'rin olgan. Korpusni yaratish jarayonida olimlar Trace-Aligner dasturini ham yaratishgan bo'lib, ushbu dastur orqali matnlar abzas va gaplar kesimida tenglashtirilishi bilan birga, ular ekstralingvistik ma'lumotlar bilan ham boyitilganiga guvoh bo'lishimiz mumkin:



1-rasm. Trace Aligner dasturi interfeysi.

Dastlab dastur orqali ma'lumotlarga ekstralingvistik izoh beriladi: avval matnlar avtomatik XML formatda annotatsiyalanadi, "header" qismga matn haqidagi metatavsiflar beriladi; keyingi jarayonda gaplar avtomatik ravishda tenglashtiriladi; so'nggi bosqichda esa tenglashtirilgan avtomatik ma'lumotlarning to'g'rilik darajasi inson omili orqali tekshiriladi. Metatavsiflarni berish korpus uchun muhim omil deb sanalgan, ularning maqsadi korpus menejeri orqali so'rovlar bajarilganda foydalanuvchilarga qulaylik yaratish bo'lgan. Quyidagi rasmda Trace Aligner dasturi orqali tenglashtirilgan hamda ekstralingvistik teglangan ma'lumotlarni kuzatishimiz mumkin:


```

1 <?xml version="1.0" encoding="iso-8859-1"?>
2 <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="apa">
3 <teiHeader>
4 <fileDesc>
5 <titleStmt>
6 <title>Die Verwandlung</title>
7 <author>F. Kafka</author>
8 <translator>R. Iraola</translator>
9 <language>Deutsch</language>
10 <tradmode>Original</tradmode>
11 </titleStmt>
12 <publicationStmt>
13 <date>2011/08/08</date>
14 </publicationStmt>
15 <sourceDesc>
16 <p>Editorial_AGA: Coleccion: </p>
17 </sourceDesc>
18 </fileDesc>
19 </teiHeader>
20 <text>
21 <body>
22 <p n="1"><s n="1">Die Verwandlung.</s></p>
23 <p n="2"><s n="2">Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren
Versteifungen geteilten Bauch, auf dessen Höhe sich die Bettdecke, zum gänzlichen Niedergleiten bereit, kaum noch erhalten konnte
vielen, im Vergleich zu seinem sonstigen Umfang kläglich dünnen Beine flimmerten ihm hilflos vor den Augen.</s></p>
24 <p n="3"><s n="3">Was ist mit mir geschehen?, dachte er.</s></p><s n="4">Es war kein Traum.</s><s n="5">Sein Zimmer,

```

2-rasm. Trace Aligner dasturi orqali XML formatda annotatsiyalangan matn.

Ko‘rib turganimizdek, matnlar abzas va gaplar kesimida tenglashtirilgan. Bundan tashqari matn haqidagi ekstralingvistik ma’lumotlar ham biriktirilgan: matn nomi, muallif, tarjimon, qaysi tilda ekanligi (original yoki tarjima), tarjima vaqti kabi.

Korpusning gaplar kesimida tenglashtirilganligi, unga ekstralingvistik metatavsiflar berilganligiga qaramay, lingvistik teglash jarayonining amalga oshirilmaganligi mukammal parallel korpusning shakllanishiga to’sqin bo‘luvchi omil bo‘lgan. Zamonaviy korpusshunoslik rivojlanib borar ekan, korpuslarga, xususan, parallel korpuslariga qo‘yiladigan talablar ham ko‘payib boraveradi. Bunday talablarga mos keluvchi korpuslarni tuzish esa tadqiqotchilardan ko‘plab mashaqqatlarni, sabrni, ekstralingvistik, lingvistik hamda dasturiy bilimlarni talab qiladi. Bu esa mukammal parallel korpuslarning ko‘plab nusxalari yaratilishiga to‘siq bo‘lishi mumkin. Lekin bu kabi mashaqqatlar izlanuvchilarni harakatlardan to‘xtatib qo‘ygani yo‘q, albatta. Yillar o‘tgani sari parallel korpuslarga qo‘yiladigan talablarning ko‘lami ham ortib bordi. Yuqoridagi ta’kidlangan Aleuska parallel korpusidan farqli o‘laroq, 2004-yilda CHANG Baobao boshchiligida yaratilgan xitoy-ingliz parallel korpusi[4] so‘zlarning lingvistik izohlari sirasiga kiruvchi Pos teglash(so‘z turkumlari tegi) bilan boyitildi. Quyida ushbu teglanish qanday amalga oshirilganligini kuzatishimiz mumkin:

```

<TEXT>
<TEXT_HEAD>
<MODE>书面语</MODE><FIELD>工商</FIELD><STYLE>新闻</STYLE>
<PERIOD>当代</PERIOD><CH_TITLE>今年中国经济和社会发展八项任务</CH_TITLE>
<TEXT_HEAD>
<TEXT_BODY>
<p id="1">
<a id="1" no="1">
<s id="1">
<CH_TITLE><w pos="t">今年</w>
<w pos="ns">中国</w>
<w pos="n">经济</w>
<w pos="c">和</w>
<w pos="n">社会</w>
<w pos="v">发展</w>
<w pos="m">八</w>
<w pos="q">项</w>
<w pos="n">任务</w></CH_TITLE></s></a></p>
<p id="2">

```

3-rasm. Xitoy-ingliz parallel korpusining ekstralingvistik va lingvistik teglanishi.

Ko‘rib turganimizdek, korpusda so‘zlarning qaysi so‘z turkumiga tegishlilikini bildiruvchi teglar mavjud. Korpusni qurishda avtomatik dasturlardan foydalanilgan. Korpus tuzish jarayonida xitoy-ingliz tillaridagi matnlarni abzas va gaplar doirasida tenglashtiruvchi dastur, XML formatlovchi kodlash dasturi, xitoycha matnlarni segmentlarga bo‘lib Pos teglashni amalga oshiruvchi dasturlar yaratildi.

Alia Al-Sayed Ahmad, Bassam Hammo and Sane Yagilar boshchiligidagi yaratilgan ingliz-arab siyosiy parallel korpusini yanada mukammalroq yechimga ega korpuslar sirasiga kiritishimiz mumkin. Ushbu korpus[5] qirol Abdulloh II ning nutqlari, yozishmalari va bitta asarining arabcha hamda inglizcha talqinlaridan tashkil topgan. Ekstralingvistik teglanishning mukammal darajada amalga oshirilganligi korpusning ustunlik jihati bo‘lsa, yana biri uning yuqori darajada lingvistik izohlanishidadir. Korpus ustida metatavsif berish, matn segmentatsiyasi, tokenizatsiya, tenglashtirish, stemming va Pos teglash kabi jarayonlar olib borilgan. Quyida ushbu korpusdagi qirol Abdulloh II ning kitobiga biriktirilgan metatavsif bilan tanishamiz:

Title	Our Last Best Chance: The Pursuit of Peace in a Time of Peril
Publisher	Viking Press
Date of publication	2011
Place of publication	New York
No. of chapters	27
Chapter length (words)	2300-8194
English words	109491
Arabic title	فرصتنا الأخيرة: السعي نحو السلام في وقت الخطر
Publisher	Daralsaqi
Date of publication	2011
Place of publication	Beirut
Translator	Shukri Rahim
No. of chapters	27
Chapter length (words)	2300-8194
Arabic words	107634

1-jadval. Qirol Abdulloh II ning kitobiga biriktirilgan metatavsif.

Nashr etuvchi, nashr vaqti, nashr joyi, boblar soni, boblardagi so‘zlar soni, inglizcha so‘zlar soni, arabcha nomi, tarjimon, arabcha so‘zlar soni kabilar ekstralingvistik izohlar sifatida tanlab olingan.

Lingvistik teglashda esa stemming, Pos teglash, so‘z shakllarini, so‘zlarning asosini biriktirishdan foydalanilgan. Bu esa korpus menejeri qidiruv tizimini mukammal ko‘rinishga olib kelgan. Chunki foydalanuvchi o‘zi xohlagan so‘zni stemlar, so‘z shakllar, asoslar, so‘z turkumlari va tokenlar doirasida qidirishi va natijada ko‘proq ma’lumotlarga ega bo‘lishi mumkin.

Biroq bundanda mukammalroq parallel korpuslarni ham uchratishimiz mumkin, albatta. Rus tili milliy korpusi hozirgi kundagi eng ulkan parallel

korpuslardan biri bo'lib, nafaqat bu son jihatdan, balki sifatli (ekstra)lingvistik teglanganligi, foydalanuvchi uchun qulay interfeysga ega ekanligi bilan ham e'tiborga molik korpuslardan biridir. Parallel korpuslarining so'zlar soni 168 millionga teng. Barcha parallel korpuslarda ham uchramaydigan sifatlarning biri shundaki, rus tili milliy korpusi va uning tarkibidagi parallel korpuslar mukammal lingvistik va ekstralingvistik izohlangan. Lingvistik teglash morfologik, sintaktik, semantik jihatdan amalga oshirilgan. Shu sabab korpus mukammal qidiruv tizimiga ega.

Guvohi bo'lishimiz mumkinki, qidiruv tizimi kodlashdan umuman xabari bo'lmagan foydalanuvchilar uchun ham juda qulay qilib ishlangan. Kiritilgan ma'lumot so'z, so'z shakli, leksema, jumla, semantik, grammatik birliklar yoki qo'shimcha belgilar asosida qidirilishi mumkin. Quyida “несчастной” so'ziga biriktirilgan teglarni ko'ramiz:

forlorn that, while staring at my open book , I caught myself struggling to come up with something to sa		
the very end of the car. [André Aciman. Find Me (2019) Андре Асман. Найди меня (Наталья Рашковская, 2020)] [ом		
е настолько несчастной , что я, продолжая глядеть в открытую книгу, невольно стал придумывать, I		
André Aciman. Find Me (2019) Андре Асман. Найди меня (Наталья Рашковская, 2020)] [ом		
несчастной		
Лемма	несчастный (см. в словаре)	
Грамматика	прил, ж, предл, полн, ед прил, дат, ж, полн, ед прил, ж, род, полн, ед прил, ж, твор, полн, ед	Найди :
Семантика основная	r.qual, ev	я (Нате
Семантика дополнительная	r.qual, ev.neg	d Me (2
Лемма	несчастливая (см. в словаре)	d Me (20
Грамматика	сущ, одуш, ж, предл, ед сущ, одуш, дат, ж, ед сущ, одуш, ж, род, ед сущ, одуш, ж, твор, ед	ковская
Семантика основная	r.concr, d.fem, der.subst, t.hum, der:a	Рашко
Доп. признаки	bmark, bcomma, numred, norm	
be with my book , a	Сообщить об ошибке...	André /

4-rasm. Rus tili milliy korpusi tarkibidagi rus-ingliz parallel korpusidagi
“несчастной” so'ziga biriktirilgan lingvistik ma'lumotlar.

Xulosa o'rnida aytish mumkinki, parallel korpuslarni yaratishda barcha uchun mos keluvchi, ma'lum qolipga tayanuvchi, bir xil lingvistik va ekstralingvistik teglanishni talab qiluvchi metod yoki yo'l yo'q. Biroq o'zbek-ingliz tillari parallel korpusini yaratishda jahondagi eng mashhur va sifatli parallel korpuslar tajribasiga tayanish o'zbek foydalanuvchilarini qoniqarli parallel korpus bilan ta'minlashda muhim rol o'ynaydi. Ayniqsa, rus tili milliy korpusi tarkibidagi parallel korpuslar andozasiga tayanish ushbu maqsadlarga yetishda samaraliroq deb hisoblashni lozim ko'ramiz.



Foydalanilgan adabiyotlar

1. R.Karimov. O'zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. PhD ilmiy darajasini olish uchun yozilgan dissertatsiya. Buxoro-2022
2. O'zbekiston Respublikasi Prezidenti Shavkat Mirziyoevning 2020 yil 20 oktyabrdagi «Mamlakatimizda o'zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to'g'risida»gi PF-6084-son farmoni // <https://lex.uz/docs/5058351>
3. Naroa Zubillaga, Zuriñe Sanz & Ibon Uribarri. 2015. Building a trilingual parallel corpus to analyse literary translations from german into basque. In Claudio Fantinuoli & Federico Zanettin (eds.), New directions in corpus-based translation studies, 71–93. Berlin: Language Science Press
4. CHANG Baobao. Chinese-English Parallel Corpus Construction and its Application. PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo
5. Alia Al-Sayed Ahmad, Bassam Hammo and Sane Yagi. ENGLISH-ARABIC POLITICAL PARALLEL CORPUS: CONSTRUCTION, ANALYSIS AND A CASE STUDY IN TRANSLATION STRATEGIES. Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.
6. B.Elov, M.Amirqulov. O'zbek-ingliz tillarining teglangan parallel korpusini yaratish bosqichlari. / O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 90-102.
7. B.Elov, Sh.Hamroyeva, O.Abdullayeva, M.Uzoqova. O'zbek tilida pos tegging masalasi: muammo va takliflar. O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 45-63.
8. O.Abdullayeva. Programs used to create the language corpus and their principles. ACADEMICIA: An International Multidisciplinary Research Journal. Vol. 10, Issue 6, June 2020. – B. 1778-1783.
9. <https://ruscorpora.ru/>

O‘ZBEK TILI MILLIY KORPUSI UCHUN LITOSFERAGA OID TERMINLAR MIKROTIZIMI TASNIFI

Islomov Ikrom Xushboqovich,
Filologiya fanlari doktori (DSc),
ToshDO‘TAU dotsenti,
ikrom75islom@gmail.com

Annotatsiya. Ushbu maqolada o‘zbek tili geografik terminlari tizimining ideografik strukturasi tasnifiga doir tahlil va tasniflar amalga oshirilgan. Xususan, geografik terminologiyani tashkil etuvchi mikrotizimlardan biri bo‘lgan litosferaga oid terminlar guruhi alohida mikrotizim sifatida qayd qilingan holda uni tashkil qiluvchi terminologik birliklarning ichki guruhlarga mansubligi tasnif qilingan. Geografik terminlar tizimining litosferaga oid terminlar mikrotizimining ideografik strukturasi shu taxlitda ochib berilishi va tasniflanishi kompyuter lingvistikasi, kompyuter terminologiyasi va terminografiyasi uchun muhim manba sifatida xizmat qilishi nazarda tutilgan.

Kalit so‘zlar: *termin, geografik terminologiya, mikrotizim, iyerarxik munosabat, kompyuter lingvistikasi, kompyuter leksikografiyasi, korpus, terminografiya.*

Abstract. In this article, the analysis and classification of the ideographic structure of the system of geographical terms of the Uzbek language was carried out. In particular, the group of terms related to the lithosphere, which is one of the microsystems that make up geographical terminology, is recorded as a separate microsystem, and the terminological units that make up it are classified as belonging to internal groups. It is intended that the ideographic structure of the microsystem of terms related to the lithosphere and the classification of the ideographic structure of the system of geographical terms will serve as an important source for computer linguistics, computer terminology and terminography.

Key words: *term, geographic terminology, microsystem, hierarchical relationship, computer linguistics, computer lexicography, corpus, terminography.*

Абстракт. В данной статье проведен анализ и классификация идеографической структуры системы географических терминов узбекского языка. В частности, группа терминов, относящихся к литосфере, которая является одной из микросистем, составляющих географическую терминологию, фиксируется как отдельная микросистема, а составляющие ее терминологические единицы отнесены к внутренним группам. Предполагается, что идеографическая структура микросистемы терминов, связанных с литосферой, и классификация идеографической структуры системы географических терминов послужат важным источником для компьютерной лингвистики, компьютерной терминологии и терминографии.

Ключевые слова: *термин, географическая терминология, микросистема, иерархическая связь, компьютерная лингвистика, компьютерная лексикография, корпус, терминография.*

Jahon tilshunosligida fan va sohalarning rivojlanishi, o'zaro integratsiyasi, yangi tarmoq va yo'nalishlarning yuzaga kelishi terminologik tadqiqotlar bilan birga axborot texnologiyalariga asoslangan ilmiy izlanishlarga ham jiddiy e'tibor qaratilmoqda. Fan va sohalar rivoji, yangilanishlar jarayonida soha terminologiyasining tartibga solinishi nafaqat soha uchun, balki tilshunoslik, kompyuter lingvistikasi uchun ham muhim amaliy ahamiyat kasb etadi. Shu kungacha terminologiya va terminologik birliklar yuzasidan amalga oshirilgan ilmiy izlanishlar natijasi o'laroq qator lingvistik, ensiklopedik lug'atlar yuzaga kelgan bo'lsa, endigi vazifa ularni axborot texnologiyalariga asoslangan elektron lug'atlar, korpuslar uchun tadqiq qilish bilan bog'liq bo'lmoqda. Sohaga doir so'nggi yutuqlar sifatida elektron lug'atlar, til korpuslari, onlayn tarjima lug'atlarini alohida ta'kidlash lozim. Sohaviy leksikani, terminologik birliklarni tizimiy tadqiq qilish, ideografik qurilishi, ma'noviy xususiyatlarini aniqlash tilning ijtimoiy vazifalari va tarixiy taraqqiyotini yoritish, terminlarning nominativ, kommunikativ, akkumulyativ funksiyalari mohiyatini oydinlashtirish, kompyuter texnologiyalariga asoslangan elektron leksikografik mahsulotlarning mukammallik va samaradorlik darajasini ta'minlashda ilmiy-amaliy ahamiyat kasb etadi.

Dunyo tilshunosligida terminologiya, xususan, geografik terminologiyaning shakllanish va taraqqiyot manbalari, termin tushunchasi va terminologik tadqiqotlar tarixi, terminosistemalarning makro va mikrotizimlari, sistemaviy qurilishi, ideografik strukturasi, terminlar yasallishining asos va usullari, tarjimasi, terminlararo sinonimiya, polisemiya, omonimiya, giponimiya hodisalari, korpuslarda til birliklarining lingvistik va dasturiy ta'minotini yaratish kabi masalalar yuzasidan tadqiqotlar olib borilmoqda.

Zamonaviy o'zbek geografik terminologiyasi qisman o'zbek umumxalq atamalari va umumiste'mol so'z-terminlarni qamrab olgan, shu bilan birga, baynalmilal terminlarga ham boy terminologik tizimdir. Bu holat esa ularni ham milliy til, ham xorijiy til asosida tadqiq qilishni taqozo qiladi. Rus tilshunosi A.Reformatskiy tomonidan ta'kidlanganidek, "terminologiya – muayyan tilning umumlug'aviy tizimi doirasida aniq ajralib turuvchi kichik bir tizimdir, ... shu bilan birga, mavzuiy belgilarga ko'ra yanada kichik tizimlarga ajraladi"[Реформатский, 1961:121-122]. Bu holat o'zbek tili geografik terminlari tizimida ham namoyon bo'ladi.

"Tabiiy va gumanitar fanlarga oid atamalar bilan bog'liq semantik maydonlarni ideografik planda tadqiq etish orqali muayyan fanlar doirasida amalda bo'lgan hodisalarning asl mohiyatini ochishda, ularning tarkibidagi elementlarning bog'lanish qonuniyatlarini aniqlashda bizga yaqindan yordamlashadi"[Собиров, 2004:123-124].

Demak, tilning lugʻaviy birliklarini sistemaviy tarzda mavzuviy, semantik maydon nuqtai nazaridan tadqiq etish ideografik tahlilning asosini tashkil qiladi. Lugʻaviy birliklarni ideografik tarzda oʻrganish orqali esa ularning lisoniy qiymatini baholash, lingvistik tabiatini chuqur anglash imkoniyatini taʼminlash bilan birga kompyuter lingvistikasi, korpus tilshunosligi, kompyuter leksikografiyasi uchun ham muhim materiallar taqdim qilishga erishiladi.

Oʻzbek tili geografik terminlar tizimiga mansub barcha birliklar uchun arxisema vazifasini “geografiyaga oid, tegishli” semasi bajaradi.

Mazkur markaziy guruhni tashkil qilgan har bir tizim oʻziga xos semantik belgi asosida yuzlab, minglab mahalliy va baynalmilal terminlarni oʻz tarkibiga qamrab olgan. Oʻzbek tili geografik terminlar tizimiga mansub barcha birliklar uchun arxisema vazifasini “geografiyaga oid, tegishli” semasi bajaradi. Mazkur markaziy guruhni tashkil qilgan har bir tizim oʻziga xos semantik belgi asosida yuzlab, minglab mahalliy va baynalmilal terminlarni oʻz tarkibiga qamrab olgan.

I. Litosferaga oid terminlar mikrotizimi:

Bu mikrotizim asosida *litosfera* termini turadi. Bu termin asosida togʻ jinslari, relyef tushunchalari bilan bogʻliq narsa-hodisalar anglashiladi. Shuning uchun mazkur terminlar tizimi boshqa mikrotizimlarga nisbatan juda koʻp miqdordagi terminlarni oʻziga birlashtirgan mikrotizim hisoblanadi. Litosfera terminining etimologik maʼnosi asosida “tosh, qattiq” maʼno boʻlakchasi yotadi, u relyef va togʻ jinslari tushunchasi bilan bogʻliq barcha narsa-hodisalarni nazarda tutadi. Shu oʻrinda relyef terminiga oid leksikografik maʼlumotlarni ham qayd etib oʻtish joiz boʻladi. “**Relyef** (lat. relevo – koʻtaraman) geografiyada – yer yuzasi, okean va dengiz tubidagi notekislik (past-balandlik)lar yigʻindisi”[Ўзбекистон миллий энциклопедияси. 7-жилд, 2004:302]. Demak, litosferaga oid terminlar mikrotizimida relyef tushunchasi bilan bogʻliq terminologik birliklar salmoqli oʻrin egallaydi. Litosfera bilan bogʻliq barcha terminlarning semantik tabiatida denotativ belgi-xususiyatlar yetakchilik qiladi.

Litosferaga oid terminlar mikrotizimining tarkibida togʻ-tosh, qattiq modda va jinslar, asosiysi, quruqlik tushunchasi bilan bogʻliq juda koʻplab terminologik birliklar mavjud boʻlib, ularni semantik belgisidagi umumiy va farqlovchi belgilari asosida quyidagicha tasniflash mumkin boʻladi:

I.1. Relyefni ifodalovchi terminlar guruhi;

I.2. Togʻ jinslarini ifodalovchi terminlar guruhi.

Yer yuzasi relyefi shakl va hajm nuqtai nazaridan geotektura (Yer poʻstidagi yirik relyef shakllari), morfostruktura (yirik togʻ tizmalari va tekisliklar) va morfoskulptura (daryo vodiylari, allyuvial tekisliklar, muz relyef shakllari) turlarga boʻlinadi, shunga koʻra, relyefni ifodalovchi terminlar guruhini yana ichki guruhlariga ajratish mumkin:

I.1.1. Geotekturaga oid terminlar: *materik koʻtarilmalari, okean botiqlari, geosinklina, platforma, rift zonasi, tektonik jarayon, choʻkish* kabilar.

I.1.1.1. Tekisliklar va platformalar: *tokemberiy platformasi, arxey platforma, epiproterozoy platforma, epibaykal platforma, epikaledon platforma, epigersin platforma, qalqonlar, plita, tekislik, past-tekislik, yalanglik, do'ngli tekislik (peneplen)* kabilar;

I.1.1.2. Botiqlar: *Popigay botig'i, krater, botiqlar (depressiyalar), alaslar (botiq joy), okean Yer po'sti, kontinental riftlar (materik yer po'sti yemirilishi va uning okean yer po'stiga aylanishi)* kabilar.

I.1.2. Morfostrukturaga oid terminlar: *yirik tog' tizmasi, tekislik, baland tog', o'rtacha tog', past tog', Kordilyera tog'lari, Buyuk tekisliklar, Sharqiy Yevropa tekisligi, Turon tekisligi, Sharqiy Avstraliya tog'lari, tektonik jarayon, magmatizm, geosinklinal, nurash, yemirilish, ko'chish, burmalanish* va b.;

I.1.2.1. Tog' va balandliklar: *tog', baland tog', o'rtacha tog', past tog', cho'qqi, dovon, yonbag'ir, tepalik, balandlik, do'nglik, qir, adir, g'or, o'ngir, tog' tizmasi, sinklinallar, antiklinallar (tog'larning burmalanishi), yoriqlar, darzlar, Efiya rifti, Reyn rift zonasi, yassi tog'lar, And tog'lari, Braziliya yassi tog'ligi, Gipsografik egri chiziq, tog' qirrasi, tog' yo'laklari, dara, tangi, tog'li o'lka, tog' massivi, tog' tuguni, tog' zanjiri, yassi tog', tog'lik, burmali tog'lar, burmali-palaxsali tog'lar, vulkan tog'lari* kabilar.

I.1.2.2. Vulqonlar: *magmatizm, vulkanik jinslar, lava, metamorfizm, Kilimanjaro vulkani, magma, qalqonli vulkanlar, qatlamsimon vulkanlar, cho'qqi, qor chizig'i, Chimborasi vulkani, Elburs vulkani, Fudziyama vulkani* kabilar.

I.1.3. Morfoskulpturaga oid terminlar: *daryo vodiylari, allyuvial tekislik, muz relyef shakli, jar, kirg'oq, barxan, yemirilish, eroziya, yotqiziq, ariqcha, jar, ostona, qayir, terrasa, qirg'oq, karr (tog' jinslari yuzasida hosil bo'ladigan chuqur ariqlar)* kabilar.

I.2. Tog' jinslarini ifodalovchi terminlar guruhi ham, o'z navbatida, quyidagi ichki guruhlariga bo'linadi:

I.2.1. Noruda qazilmalarni ifodalovchi terminlar: *tuz, toshko'mir, neft, tabiiy gaz, yonuvchi slanets, ohaktosh, qum, gillar* kabilar.

I.2.2. Metall qazilmalarni ifodalovchi terminlar: *boksitlar, fosforit cho'kindi, marganets, temir rudalari; a) rangli metallar: oltin, kumush, mis, ruh, qo'rg'oshin, alyuminiy, qalay; b) qora metallar: magnitli temirtosh, qo'ng'ir temirtosh, qizil temirtosh* kabilar.

I.2.3. Noruda foydali qazilmalarni ifodalovchi terminlar: *olmos, yoqut, zumrad, sapfir* kabilar.

I.2.4. Magmatik tog' jinslari:

I.2.4.1. Intruziv magmatik tog' jinslari: *granit, gabbro;*

I.2.4.2. Effuziv magmatik tog' jinslari: *bazalt, liparit, vulkan shishasi* kabilar.

I.2.5. Cho'kindi jinslar:

I.2.5.1. Chaqiq (klastik) jinslar: a) yirik chaqiq jinslar: *psefit, xarsang tosh, mayda tosh, dresva, g'o'la tosh, shag'al tosh;* b) o'rtacha chaqiq jinslar: *qum, qumtoshlar (psammitlar), dag'al donali qumtosh, mayda donali qumtosh, mayin*

donali qumtosh; d) mayda yoki changisimon jinslar (alevrolitlar): *lyosslar, soz tuproqlar, gillar, qumoqlar, flish*; mineralli tog' jinslari: a) monomineralli tog' jinslari: *kvars, kvars minerali*; b) polimineral tog' jinslari: *granit, kvars, ortoklaz, slyuda, dala shpati* kabilar;

I.2.5.2. Kimyoviy tog' jinslari: *tuzlar, gips* kabi;

I.2.5.3. Organik (biogen) tog' jinslari: *marjonlar, ko'mir, bo'r, ohaktosh* kabilar.

I.2.6. Metamorfik tog' jinslari: *gneys, kvarsit, marmar* kabilar.

I.3. Landshaftga oid terminlar guruhi: *tog' landshafti, cho'l landshafti, plato landshafti, tog' oldi landshaft, glyatsialnival landshaft, tog' yonbag'irlari landshafti, madaniy landshaft* kabilar.

Litosferaga oid terminlar mikrotizimining markaziy termini sifatida *litosferani*, uning ilk qurshov terminlari deb *relyef* va *tog' jinslari* terminini qayd etamiz. Ana shu ilk qurshov terminlari o'zlariga uyadosh bo'lgan barcha geografik terminlarni mazkur mikrotizimga olib kiradi. Ushbu mikrotizim hozircha shartli ravishda 3 bosqichli tizim tarzida baholanishi mumkin. Har bir bosqichdagi mikrotizimlar o'zidan oldingi va keyingi bosqich tizimlar bilan iyerarxik (pog'onaviy) munosabat asosida o'zining barqaror o'rniga ega bo'ladi.

Yuqorida qayd qilingan pog'onaviy bo'linishli geografik terminlar tizimi o'ziga xos ideografik strukturani hosil qilgan. Bu ideografik strukturaning shakllanishi va shu tarzda tasniflanishida asosan tabiiy geografiya sohasi nuqtai nazariga monandlik yetakchilik qiladi, chunki terminosistema asosida soha, fan tushuncha va hodisalarining sohaviy strukturasi hal qiluvchi ahamiyatga ega. Ushbu terminologik tizimning ideografik strukturasi ochib berilishi kompyuter terminologiyasi va terminografiyasi uchun lingvistik jihatdan ahamiyatli natijalar taqdim qiladi deb hisoblaymiz.

Foydalanilgan adabiyotlar:

1. Реформатский А.А. Что такое термин и терминология? // Вопросы терминологии. – М., 1961. – С. 121-122.
2. Собиров А. Ўзбек тилининг лексик сатҳини системалар системаси тамойили асосида тадқиқ этиш. Монография. – Тошкент: Маънавият, 2004. – 160 б. – Б. 123-124.
3. Ўзбекистон миллий энциклопедияси. 7-жилд. Парчин – Солиқ. Таҳрир ҳайъати А.Абдувоҳитов, А.Азизхўжаев, М.Аминов, Т.Даминов ва б. – Тошкент: Ўзбекистон миллий энциклопедияси, 2004. – 704 б. – Б. 302.

UDK: 811.111.111`11

DATA SCIENCE IN CORPUS LINGUISTICS

Babayev Saidmuxammadjon Saidkamolovich

Master student at Urgench State University

saidmuhammadbabayev@gmail.com

Ruzimboyev Xusniddin Raximberganovich

Master student at Urgench State University

khusniddindatascientistn1@gmail.com

Annotatsiya. Ma'lumotlar ilmi (Data science) va korpus lingvistikasi so'nggi yillarda tobora o'zaro bog'langan ikkita sohadir. Ushbu maqola til ma'lumotlarining tahlil qilish uchun ma'lumotlar ilmi va tabiiy tillarni qayta ishlash (NLP) texnologiyasidan foydalangan so'nggi tadqiqotlarni ko'rib chiqish orqali ushbu sohalarning kesishishini o'rganadi. Biz ushbu usullardan Korpus lingvistikasida an'anaviy ravishda qo'lda bajariladigan ko'plab vazifalarni avtomatlashtirish uchun qanday foydalanish mumkinligini va ulardan ma'lumotlardagi qolip va tendensiyalarni aniqlash uchun qanday foydalanish mumkinligini muhokama qilamiz. Shuningdek, his-tuyg'ularni vaqt o'tishi bilan til o'zgarishini tahlil qilish va til ma'lumotlariga asoslangan bashoratli modellarni yaratish uchun ma'lumotlar ilmi va NLP texnikasining imkoniyatlarini o'rganamiz. Va nihoyat, biz ushbu sohalarning kelajagi va NLP va Data Science texnikasidan foydalanish bilan bog'liq axloqiy va maxfiylik muammolarini hal qilish zarurligini muhokama qilamiz.

Kalit so'zlar: *NLP, ma'lumotlar ilmi, Korpus lingvistikasi, Data mining.*

Abstract. Data Science and Corpus Linguistics are two fields that have become increasingly intertwined in recent years. This paper explores the intersection of these fields by reviewing recent studies that have used Data Science and NLP techniques to analyze large corpora of language data. We discuss how these techniques can be used to automate many of the tasks traditionally done manually in Corpus Linguistics, and how they can be used to identify patterns and trends in the data. We also explore the potential of Data Science and NLP techniques for analyzing sentiment, language change over time, and for building predictive models based on language data. Finally, we discuss the future of these fields and the need to address ethical and privacy concerns related to the use of NLP and Data Science techniques.

Keywords: *NLP, Data Science, Corpus Linguistics, Data Mining.*

Аннотация. Наука данных и корпусная лингвистика — две области, которые в последние годы все больше переплетаются. В этой статье рассматривается пересечение этих областей путем обзора недавних исследований, в которых использовались методы науки данных и NLP для анализа больших корпусов языковых данных. Мы обсудим, как эти методы

можно использовать для автоматизации многих задач, традиционно выполняемых вручную в корпусной лингвистике, и как их можно использовать для выявления закономерностей и тенденций в данных. Мы также изучаем потенциал методов науки данных и NLP для анализа настроений, языковых изменений с течением времени и для построения прогностических моделей на основе языковых данных. Наконец, мы обсуждаем будущее этих областей и необходимость решения проблем этики и конфиденциальности, связанных с использованием методов NLP и науки данных.

Ключевые слова: *NLP, наука данных, корпусная лингвистика, анализ данных.*

Introduction.

Corpus Linguistics is a subfield of linguistics that involves the analysis of large collections of language data, known as corpora. Data Science, on the other hand, involves using statistical and computational methods to extract insights and knowledge from data. Together, these two fields have the potential to revolutionize the way that we analyze and understand language. Corpus Linguistics has traditionally relied on manual methods for analyzing language data. However, with the advent of Data Science techniques, it is now possible to automate many of these tasks, which can save time and improve accuracy. For example, Data Science techniques can be used to analyze the frequency of words and phrases in a corpus, which can help researchers to identify patterns and trends in the data.

One area where Data Science has been particularly impactful in Corpus Linguistics is in the analysis of sentiment. Sentiment analysis involves the use of computational methods to identify the emotional tone of a piece of text. This can be useful for researchers who are interested in analyzing the attitudes and opinions expressed in a corpus. Data Science techniques can be used to analyze large corpora of text and identify patterns of sentiment, which can then be used to draw conclusions about the attitudes and opinions of the speakers or writers.

Another area where Data Science has been used in Corpus Linguistics is in the analysis of language change over time. By analyzing large corpora of text from different time periods, researchers can identify changes in the frequency of words and phrases, as well as changes in the structure and grammar of the language. Data Science techniques can be used to automate many of these tasks, which can save time and improve the accuracy of the analysis.

Data Science can also be used to build predictive models based on language data. For example, machine learning algorithms can be used to identify patterns in a corpus of text and make predictions about future language use. This has the potential to be useful in a variety of fields, such as marketing, where predictive models can be used to anticipate changes in consumer behavior.

Related Work.

Several researchers have explored the intersection of Data Science and Corpus Linguistics in recent years. For example, Li et. al. [1] used a combination of machine learning algorithms and natural language processing techniques to analyze a corpus of Chinese microblog data. Their study focused on identifying patterns of sentiment and emotion in the data, and they found that their approach was able to accurately predict the emotional tone of the microblogs with high accuracy.

In another study, Tadesse et. al. [2] used machine learning algorithms to analyze a corpus of English-language tweets. They focused on identifying patterns of opinion and emotion in the data, and found that their approach was able to accurately classify tweets into different sentiment categories.

These studies demonstrate the potential of Data Science and NLP techniques for analyzing language data in new and innovative ways. They also highlight the importance of developing accurate and efficient algorithms for processing and analyzing large corpora of text. As the field of Data Science and NLP continues to evolve, we can expect to see even more exciting applications of these techniques in the field of Corpus Linguistics.

Regarding the advances in Uzbek NLP resources that make use of Data science algorithms, a number of works have been done in such manner such as sentiment analysis [3], [4], morphological tagged corpus [5], as well as text summarization [6], [7]. Furthermore, there has been a recent rapid increase in the development of NLP resources and tools specifically for Uzbek language. These include a tool for removal of stopwords [8], [9], Latin-Cyrillic transliteration [10] and text classification [11], as well as tools for stemming [12], [13], lemmatization [14].

Methodology.

NLP tasks that can benefit from Data Science techniques include many of the traditional tasks of NLP, such as sentiment analysis, named entity recognition, and part-of-speech tagging. Machine learning algorithms can be used to automate many of these tasks, making it possible to process large volumes of text data quickly and efficiently. In addition, deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown great promise for tasks such as sentiment analysis and text generation. Clustering algorithms can be used to identify patterns and groups within large corpora of text, while topic modeling algorithms such as Latent Dirichlet Allocation (LDA) can be used to identify latent themes and topics within a corpus. Finally, word embedding techniques such as Word2Vec and GloVe can be used to represent words as vectors in a high-dimensional space, enabling semantic similarity and distance calculations, which are essential for many NLP tasks.

Following are the algorithms in the field of Data science that can be beneficial for NLP tasks:

- Naive Bayes: A probabilistic algorithm commonly used for text classification tasks, such as sentiment analysis or spam detection.

- Support Vector Machines (SVM): A supervised learning algorithm that can be used for text classification, named entity recognition, and other tasks.
- Random Forests: An ensemble learning algorithm that can be used for text classification, topic modeling, and other tasks.
- Convolutional Neural Networks (CNNs): A deep learning algorithm that can be used for sentiment analysis, text classification, and other tasks.
- Recurrent Neural Networks (RNNs): A deep learning algorithm that can be used for language modeling, text generation, and other tasks.
- Latent Dirichlet Allocation (LDA): A topic modeling algorithm that can be used to discover latent themes or topics in a corpus of text.
- Hierarchical clustering: An unsupervised learning algorithm that can be used for text clustering, topic modeling, and other tasks.
- k-means clustering: An unsupervised learning algorithm that can be used for text clustering and topic modeling.
- Principal Component Analysis (PCA): A dimensionality reduction technique that can be used for text visualization and clustering.
- t-SNE: A dimensionality reduction technique that can be used for text visualization and clustering.

Conclusion and Future work.

In conclusion, Data Science has the potential to revolutionize the field of Corpus Linguistics by providing new methods for analyzing language data. By automating many of the tasks that were traditionally done manually, researchers can save time and improve the accuracy of their analysis. Additionally, Data Science techniques such as sentiment analysis and predictive modeling can provide new insights into language use and change over time. As these fields continue to evolve, we can expect to see even more exciting applications of Data Science in Corpus Linguistics.

The future work in the field of Data Science and NLP will likely focus on advancing the accuracy and efficiency of these techniques. As the amount of unstructured data continues to grow, it will become increasingly important to develop methods for processing and analyzing this data in real-time. This will require the development of more sophisticated machine learning algorithms and the use of distributed computing frameworks to process large datasets. Additionally, there will be a need to address ethical and privacy concerns related to the use of NLP and Data Science techniques, such as the potential for bias and the need to protect sensitive information. As these fields continue to evolve, it will be important to stay up-to-date with the latest research and developments in order to take advantage of the full potential of NLP and Data Science.

References

1. X. Li, J. Zhang, Y. Du, J. Zhu, Y. Fan, and X. Chen, “A Novel Deep Learning-based Sentiment Analysis Method Enhanced with Emojis in Microblog Social Networks,” *Enterp Inf Syst*, pp. 1–22, 2022.
2. M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression-related posts in reddit social media forum,” *IEEE Access*, vol. 7, pp. 44883–44893, 2019.
3. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com
4. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243.
5. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a morphological and syntactic tagged corpus for the Uzbek language,” *arXiv preprint arXiv:2210.15234*, 2022.
6. K. Madatov, S. Bekchanov, and J. Vičič, “Uzbek text summarization based on TF-IDF,” *arXiv preprint arXiv:2303.00461*, 2023.
7. K. Madatov, S. Matlatipov, and M. Aripov, “Uzbek text’s correspondence with the educational potential of pupils: a case study of the School corpus,” *arXiv preprint arXiv:2303.00465*, 2023.
8. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.
9. K. Madatov, S. Bekchanov, and J. Vičič, “Lists of uzbek stopwords,” Univerza na Primorskem, Inštitut Andrej Marušič, 2021.
10. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com
11. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.
12. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” *arXiv preprint arXiv:2205.10078*, 2022.
13. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.
14. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” *arXiv preprint arXiv:2210.16006*, 2022.

TURK-O'ZBEK PARALLEL KORPUSIGA MATN TANLASHDA MUHIM XUSUSIYATLAR

Xolmonova Iqbola Alisher qizi

ToshDO'TAU, kompyuter
lingvistikasi yo'nalishi 1-kurs magistranti
iqbolabintualisher@gmail.com

Ilmiy rahbar: Sh.Xamroyeva,

ToshDO'TAU dotsenti, filologiya fanlari
doktori

Annotatsiya: Ushbu maqolada Turk-o'zbek parallel korpusiga matn tanlashda muhim xususiyatlar haqida so'z boradi. Matn tanlashda kompilyatorning vazifalari sanab o'tiladi. Korpusga matn tanlashda qanday uslublardan foydalanish mumkinligi aytiladi. Bundan tashqari, maqolada korpusga matn tanlashda qadamlar ketma-ketligi sanab o'tilgan.

Kalit so'z: *matn, parallel korpus, avtomatik ekstraktsiya, gibrid yondashuv, matn janri, parallel matn.*

Annotation: This article discusses important features in text selection for the Turkish-Uzbek parallel corpus. The tasks of the compiler are listed in the text selection. It tells you what styles can be used when selecting text for the corpus. In addition, the article lists the sequence of steps in selecting text for the Corpus.

Keywords: *text, parallel corpus, automatic extraction, hybrid approach, text genre, parallel text.*

Аннотация: В данной статье рассматриваются важные особенности выделения текста для турецко-узбекского параллельного корпуса. Задачи компилятора перечислены в текстовом выделении. Он сообщает вам, какие стили можно использовать при выборе текста для корпуса. Кроме того, в статье указана последовательность действий при выборе текста для Корпуса.

Ключевые слова: *текст, параллельный корпус, автоматическое извлечение, гибридный подход, жанр текста, параллельный текст.*

Kirish. Korpus ma'lum prinsiplarga muvofiq yaratilgan katta matnlar to'plami hisoblanadi. Lekin bu har qanday matnlar to'plami degani emas, ya'ni korpusga kiritiladigan matnlar ma'lum talablarga javob berishi kerak.

Korpus matnlari kompilyatorning korpusni yaratish maqsadidan kelib chiqib, aniq mezonlar asosida tanlanadi. Kompilyator korpus uchun matnlarning hajmi, janri, turi va boshqa parametrlarini o'zi belgilaydi. Korpus tarkibidagi matnlarning mazmun xususiyati va turlari jihatidan bir-biridan farq qilishi davrning butun lug'at boyligini ochib berish nuqtayi nazaridan muhim ahamiyatga ega. Korpus tarkibiga

kiruvchi matnlar turli janrlardan tashkil topgan bo'lsa, undan turli soha vakillari hamda keng omma foydalana oladi.

Ma'lumki, har qanday korpusning asosini uning birliklari tashkil qiladi. Odatda, korpusning reprezentativligini ta'minlash uchun korpus birligi sifatida turli janrdagi matnlar tanlanadi. Shuning uchun har qanday til korpusini yaratishda, birinchi navbatda, uning tarkibiy qismi bo'lgan matnlarni tanlash tamoyillarini aniqlash kerak.

Korpus tarkibiga kiritilayotgan matnlar kompilyator tomonidan erkin tanlanadi. Misol uchun, kompilyator badiiy asarlar korpusini tuzmoqchi bo'lsa, unda korpus tarkibiga hikoya, roman, qissa janridagi asarlarni kiritishi mumkin. Yoki ilmiy yo'nalishdagi korpus tuzmoqchi bo'lsa, korpus tarkibiga maqola, tezis, dissertatsiya, ilmiy asarlar kiritilishi mumkin. Nazmga qiziquvchi kompilyator (korpus tuzuvchi shaxs) faqat nazmiy asarlardan tarkib topgan korpus tuzishi ham mumkin.

Matn janri parallel korpus uchun muhim xususiyatdir, chunki u matnlarda ishlatiladigan tilga ta'sir qiladi. Turli janrlarda turlicha konvensiya va uslublar mavjud bo'lib, ular tilning so'z boyligi, grammatikasi va tuzilishiga o'z ta'sirini o'tkazadi. Misol uchun, tasviriy til va adabiy vositalardan foydalanish mumkin bo'lgan romandan ko'ra ko'proq rasmiy til va texnik atamalardan foydalanish parallel matn uchun qulayroq. Parallel korpusga turli janrlarni kiritish orqali tadqiqotchilar tilning turli kontekst va janrlarda qanday farqlanishini va matn janriga qarab tarjimalar qanday farqlanishini tahlil qilishlari mumkin. Bu turli soha va kontekstlarda tildan foydalanish va tarjima amaliyotlari haqida qimmatli tushunchalar berishga xizmat qiladi [1. 111-122].

Asosiy qism. Matn tanlashda korpusning turi ham muhim ahamiyatga ega. Tadqiqotimizning maqsadi turk va o'zbek badiiy asarlarining parallel korpusi uchun matn tanlash muammosi bo'lganligi sababli ushbu maqolada asosan bu haqida so'z boradi. Aslida, parallel korpus nima uchun kerak? Bu savolga javoblar quyidagicha:

- 1) mavjud dunyo parallel korpusi tayyor tarjimalarga asoslanadi;
- 2) tarjimani parallel dasturga birlashtirish asl asarni va butun asar tarjimasini yonma-yon o'qish imkonini beradi;

- 3) eng muhimi, parallel korpus u yoki bu birlik (sodda gap, qo'shma gap, ibora, frazeologik birlik, jargon, metafora) asl nusxada va tarjimada qanday ifodalanganligini kuzatishning eng maqbul usuli hisoblanadi [2. 348].

Parallel korpus uchun matn tanlayotganda asl matn va uning tarjimasini kerak bo'lishini hisobga olish zarur. Shuning uchun kompilyatorning vazifasi matn tanlash jarayonida korpusga kiritmoqchi bo'lgan matning tarjimasini mavjud ekanligiga amin bo'lgandagina uni korpusga kiritishi mumkin bo'ladi.

Parallel korpus uchun matn tanlashda quyidagi usullardan foydalanish mumkin:

1. **Avtomatik ekstraktsiya.** Bu veb-sahifalar kabi yirik onlayn resurslardan parallel matnlarni olish uchun dasturiy vositalardan foydalanishni o'z ichiga oladi.

2. **Qo'lda tanlash.** Bu mutaxassisdan uzunlik, mazmun va janr kabi talablarga mos keladigan matnlarni qo'lda tanlashni talab qiladi.

3. **Gibrid yondashuv.** Bu zarur mezonlarga javob beradigan parallel korpusni yaratish uchun avtomatlashtirilgan vositalardan foydalanish va qo'lda tanlashni birlashtiradi [3. 154].

Parallel korpus uchun matn tanlash jarayonida bir qancha amallarni bajarish kerak bo'ladi:

1. **Tillarni aniqlash.** Parallel korpus uchun manba va maqsadli tillarni belgilash.

2. **Maqsadni aniqlash.** Parallel korpusdan maqsadli foydalanishni aniqlash kerak bo'ladi, jumladan, mashina tarjimasini, tilni modellashtirish va ikki tilli lug'at yaratish yoki chet tilini parallel ravishda o'rganish kabi.

3. **Mavjudligini ko'rib chiqish.** Tanlangan tillar uchun matnlar mavjudligini baholash va mavjud matnlardan foydalanish yoki yangilarini yaratish haqida qaror qabul qilish lozim.

4. **Janrlarni tanlash.** Badiiy matn, ilmiy matn, rasmiy matn va ijtimoiy media postlari kabi parallel korpus maqsadiga eng mos keladigan janrlarni aniqlash va tanlash amalini bajarish zarur.

5. **Matnlarni tanlash.** Uzunlik, izchillik va dolzarblik talablariga javob beradigan matnlarni tanlash uchun avtomatlashtirilgan dasturiy vositalar yoki ekspert inson baholovchilaridan foydalanish kerak. Matnlar tarjima sifatida bir-biriga mos kelishi, o'xshash mazmun va tuzilishga ega bo'lishi hamda muhim madaniy yoki lingvistik farqlarga yo'l qo'ymasligi kerak.

6. **Sifat nazoratidan o'tkazish.** Olingan parallel korpusni mo'ljallangan maqsad va sifat standartlariga javob berishini tekshirish lozim.[4. 238-264]

Parallel korpus birliklari asl nusxadan bir yoki bir necha marta tarjima qilingan matnlardan iborat bo'ladi. Parallel korpuslar bir tomonlama tarjimini (faqat o'zbek tilidan turk tiliga) yoki ikki tomonlama tarjimini (o'zbek tilidan turk tiliga, turk tilidan o'zbek tiliga) o'z ichiga oladi. Shundan kelib chiqib, o'zbek-turk parallel korpusi ikkita blogdan iborat bo'ladi:

1. O'zbek-turk tilidagi blog;

2. Turk-o'zbek tilidagi blog.

Bu bloglar uchun materiallar bir xil emas: birinchi blogda o'zbek tilidan turk tiliga tarjima qilingan badiiy asarlar mavjud bo'lsa, ikkinchi blog turk tilidan o'zbek tiliga tarjima qilingan asarlardan iborat. Albatta, qidiruv tizimida janr, yaratilish davri, asar muallifi va tarjimoni bo'yicha qidiruvni amalga oshirish mumkin bo'ladi.

Turk va o'zbek tilidagi matnlarning parallel korpusga kiritilishi bir qancha sabablarga ko'ra muhim ahamiyatga ega. Birinchidan, turk va o'zbek tillari o'z mintaqalarida keng tarqalgan tildir: turk tili Turkiyaning rasmiy tili, o'zbek tili esa O'zbekistonning davlat tilidir. Bundan kelib chiqadiki, ushbu tillarda juda ko'p yozma materiallar mavjud, ular parallel korpus qurish uchun qimmatli manbalarga aylanadi. Ikkinchidan, turk va o'zbek tilidagi matnlarning kiritilishi tadqiqotchilarga

tilning turli til oilalari bo'yicha o'zgarishini tahlil qilish imkonini beradi. Qarindosh tillar o'rtasidagi tarjimalarni solishtirish orqali tadqiqotchilar lingvistik tuzilmalar va xususiyatlarning turli til guruhlari bo'ylab qanday taqsimlanishi yoki farqlanishi haqida tushunchaga ega bo'lishlari mumkin.

Nihoyat, turk va o'zbek tilidagi matnlarning parallel korpusga kiritilishi ushbu tillardagi tarjima amaliyoti haqida qimmatli fikrlarni beradi. Turk va o'zbek tilining o'ziga xos grammatik tuzilishi, so'z boyligi va idiomatik iboralari mavjud bo'lib, bu tarjimonlar uchun qiyinchilik tug'dirishi mumkin. Ushbu tillar va korpusdagi boshqa tillar o'rtasidagi tarjimalarni tahlil qilish orqali tadqiqotchi, tarjimonlar bu qiyinchiliklarni qanday hal qilishlari va o'z tarjimalarini turli janr, kontekstlarga qanday moslashtirishlari haqida tushunchaga ega bo'lishlari mumkin bo'ladi.

Xulosa qilib aytganda, turk va o'zbek tilidagi matnlarning parallel korpusga kiritilishi korpusni yaratish uchun qimmatli manbalarni taqdim etishi, qarindosh tillararo tahlil qilish imkonini beradi, mazkur tillardagi tarjima amaliyoti haqida tushuncha berishi bilan ahamiyatli.

Foydalanilgan adabiyotlar:

1. Trosterud, Trond 2002: Parallel corpora as tools for investigating and developing minority languages. In: Borin, Lars. Parallel Corpora, Parallel Worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April 1999. Amsterdam New York, NY: Rodopi. Pp. 111-122.
2. Михайлов М. Параллельные корпуса художественных текстов: диссертация. Финландия. ун-та, 2003. 348с
3. McEnery, Tony and Wilson, Andrew 2001: Corpus Linguistics: An Introduction. 2nd edition. Edinburgh: Edinburgh University Press.
4. Salkie, Raphael 2002: how can linguists profit from parallel corpora? In: Borin, Lars (ed.). Parallel Corpora, Parallel Worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April 1999. Amsterdam New York, NY: Rodopi. Pp. 93-10
5. Teubert, Wolfgang 1996: Comparable or Parallel Corpora. International Journal of Lexicography. Oxford University Press. 9 (3), 238-264.
6. Elov B., Amirqulov M. O'zbek-ingliz tillarining teglangan parallel korpusini yaratish bosqichlari. O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 90-102.
7. Hamroyeva Sh. O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Filol. Fan. bo'yicha falsafa dokt. (PhD) diss. - Buxoro, 2018. B-73.
8. Abdullayeva O. O'zbek tili korpusini yaratish bosqichlari va muammoli jihatlari. / "So'z san'ati", 4-jild, 1-son – Toshkent, 2021. –b. 153-158.
9. Abdullayeva O., Norqulova M. Internet matnlarining til korpusi rivojida ahamiyati. Science and innovation. Vol. 1, Issue 6. – Toshkent, 2022. (doi.org/10.5281/zenodo.718910).

KORPUSDA O‘ZBEK TILIDAGI IBORALARNI SEMANTIK TEGLASH MASALASI

Bo‘riyev Islom Eshmirzayevich,
ToshDO‘TAU 2-kurs tayanch doktoranti,
islomboriyev122@gmail.com

Annotatsiya: Mazkur maqola o‘zbek tilidagi iboralarni korpusda semantik teglashga bag‘ishlangan, shuningdek, maqolada ko‘p ma’noli iboralarni semantik teglashda e’tibor qilinishi kerak bo‘lgan masalalar ham yoritilgan.

Kalit so‘zlar: *semantik teg, semantik annotatsiya, elektron lug‘atlar, semantik maydon.*

Annotation: This article is devoted to the semantic tagging of phrases in the Uzbek language in the corpus, and the issues that should be considered in the semantic tagging of ambiguous phrases are also covered in the article.

Key words: *semantic tag, semantic annotation, electronic dictionaries, semantic field.*

Аннотация: Данная статья посвящена семантической разметке словосочетаний на узбекском языке в корпусе, а также освещены вопросы, которые следует учитывать при семантической разметке многозначных словосочетаний.

Ключевые слова: *семантическая метка, семантическая аннотация, электронные словари, семантическое поле.*

Hozirda shiddat bilan rivojlanib borayotgan axborot-texnologiya barcha sohalarga kirib bormoqda va bu sohalarning ham rivojlanishini jadallashtirmoqda. Bugungi kunda qaysi soha bo‘lmasin uni kompyuter texnologiyalarisiz tasavvur qilish qiyin. Shunday sohalardan biri tilshunoslikdir. Tilning funksional imkoniyatlaridan foydalanish borasida zamonaviy axborot texnologiyalari bizga juda ko‘p imkoniyat va qulayliklarni yaratdi. Jumladan, kompyuter tarjimasini, tahrir va tahlil dasturlari, elektron lug‘atlar, tezauruslar fikrimiz tasdig‘idir. Zamonaviy tilshunoslik yo‘nalishlaridan kompyuter lingvistikasi tobora kompyuter texnologiyalari bilan chambarchas bog‘lanib bormoqda. Bu esa o‘z-o‘zidan tilshunosligimizda har tomonlama qulay, zamonaviy elektron lug‘atlar, kompyuter dasturlari, onlayn darsliklar, maxsus platformalar yaratish vazifalari dolzarb masalalardan bir ekanligini ko‘rsatib turibdi. Prezidentimiz Sh.Mirziyoyev ta’kidlaganidek, “Ayni paytda aholi, ayniqsa, yoshlarimiz turli ma’lumotlarni asosan Internet va ijtimoiy tarmoqlardan olmoqda. Binobarin, ona tilimizni jozibali va qiziqarli tarzda taqdim etish, buning uchun o‘zbek tilining **kompyuter dasturlarini, onlayn darsliklar, elektron lug‘atlarini** yaratish kerak. Internet imkoniyatlaridan to‘liq foydalanib, o‘zbek tilini keng targ‘ib etish bo‘yicha ishlarimizni kuchaytirishimiz lozim” [1]. Bundan tashqari, Yangi O‘zbekistonda hozir har bir

soha raqamlashtirilmoqda. Tilshunoslik ham bundan mustasno emas. Davlatimizda bo'layotgan o'zgarishlar tilshunoslar oldiga o'zbek tilining lug'at tizimini, so'zlardagi ma'no torayishi, kengayishi, omonimiya, sinonimiya, antonimiya, paronimiya, graduonimiya, ko'p ma'nolilik hodisalarini, shuningdek, yangi paydo bo'lgan tushuncha va terminlarni chuqurroq tadqiq etish vazifalarini qo'yimoqda. Texnika taraqqiyoti tufayli qator davlatlarning o'z milliy korpuslari yaratildi. Bu borada o'zbek tili milliy korpusini yaratish ishlari ham boshlangan. Hozird o'zbek tilining ta'limiy korpusi yaratilgan. Va hozirda u tilshunos olimlarimiz, tadqiqotchilarimiz tomonidan boyitib borilmoqda.

O'zbek tilining milliy korpusi yaratilmoqda ekan, bu o'zbek tili iboralar lug'atini ham chuqurroq tadqiq etilishini talab qilmoqda. Ayniqsa, iboralarning o'ziga xosligini, uning rang-barangligini ko'rsatib beruvchi semantikasini korpusda teglash muhim masalalardan biri hisoblanadi.

Korpus lingvistikasida semantik razmetkalash, semantik teglarni yaratish muammolari to'g'risida A.M.Yelizarov, T.I.Reznikova, Y.I.Yakovchuk, O.Y.Shemanayeva, V.V.Ivanov, I.M.Boguslavskiy, V.P.Zaxarov kabi olimlar tomonidan bir qancha ishlar amalga oshirilgan.

O'zbek tilshunosligida o'zbek tili korpusi, o'zbek korpus lingvistikasi uchun semantik razmetka mavzusining chuqur o'rganilishi dastlab D.Ahmedovanning tadqiqot ishida [D.Ahmedova, 2020] uchraydi. Shuningdek, Sh.Hamroyevaning tadqiqotlarida semantik maydon, to'da va guruh tasniflarining semantik razmetka tizimini yaratishdagi ahamiyati haqida ba'zi fikrlar mavjud [Sh.Hamroyeva, 2018: 73].

Semantik teg (rus., razmetka) – til korpuslariga kiritiladigan birliklarning muayyan semantik kategoriya yoki kichikroq semantik guruh (leksik-semantik guruh, semantik maydon)ga tegishliligini bildiradigan, ma'noni xususiylashtiruvchi belgi, izohlar majmuyi [Захаров В.П., Богданова С.Ю., 2011: 161]. Korpusning semantik teglari so'z ma'no(lar)i spetsifikasi, so'zning omonimlik, sinonimlik bilan bog'liq izohlar majmuyini tuzish, so'zni kategoriyalash, uning mavzuiy guruh, LSG, semantik maydonga mansubligini belgilash, derivatsion xarakteristikasi, atash ma'nosi kabi belgilarni qamrab oladi [D.Ahmedova, 2020: 58]. Biror bir tilning korpusi uchun semantik teglar majmuyini yaratish uning tamoyillarini ishlab chiqish mazkur tilga xos xususiyatlardan kelib chiqadi. Shuning uchun o'zbek tili semasiologiyasi borasida qilingan tadqiqotlar, erishilgan natijalar o'zbek tili korpuslarida semantik teglar majmuyini yaratish, so'zlarning semantik teglash tamoyillarini ishlab chiqish uchun nazariy asos rolini bajaradi.

Boshqa razmetkalarda bo'lganidek, semantik teglashda ham yagona standart bo'lmasa-da, ammo raqam, harf yoki faqat raqamdan, shuningdek, aralash kodlardan foydalaniladi. Birinchi harf yoki raqam umumiy semantik ma'noni, keyingi belgi so'zning ma'nosini maxsuslashtiruvchi kichik semantik guruhni ifodalaydi. Semantik teg faqat so'z emas, ko'plab birikmalarni ham semantik guruhlariga birlashtirishga xizmat qiladi. Bunda har xil birikuvdagi bir ma'noni ifodalovchi

birikmalar bitta belgi bilan kodlanadi. Ibora tarkibidagi soʻzlar miqdorini bildiruvchi maʼlumotlar ham razmetka tarkibiga kiritiladi. Semantik teg korpusdagi soʻz maʼnosining ixtisoslashuvi, omonimlik, sinonimlik, maʼnoviy guruhga ajratish kabi muammolarni hal qiladi [D.Ahmedova, 2020: 58].

Til birliklarini semantik teglashda asosiy manba oʻzbek tilining izohli lugʻatidir [OʻTIL, 2000-2006]. Chunki lugʻat oʻzbek adabiy tilida keng isteʼmolda boʻlgan 80 000dan ortiq soʻz va soʻz birikmalarini, fan va texnika, sanʼat va madaniyatga oid boʻlgan termin(atama)larni, shuningdek, baʼzi dialektizmlar, tarixiy va eskirgan soʻzlarni ham oʻz ichiga olganligi sababli korpus tili uchun bebaho manba hisoblanadi. Til birliklarni semantik teglashda lugʻat maqolasining tuzilishi tajribasini qoʻllash yaxshi samara berishi shubhasiz. Chunki izohli lugʻatdagi bosh soʻz sifatida berilgan umumisteʼmol va cheklangan leksika korpusda ham izohli lugʻatdagi kabi farqli belgilanadi. Bunda farqli jihatlar ham boʻladi. Quyida izohli lugʻat maqolasining tuzilishi hamda iboralarni semantik teglashda undan foydalanish imkoniyatini koʻrib chiqamiz:

I. Turgʻun birikmalar ham lugʻatga kiritiladigan soʻzlarga qoʻyiladigan talablar asosida kiritilgan [OʻTIL, 2000-2006: 11]. Yaʼni umuisteʼmolda boʻlgan, shuningdek, qoʻllanish doirasi chegaralangan turgʻun birikmalardan koʻp ishlatiladigan, yozma nutqda ham koʻp uchraydiganlari kiritilgan:

– Ogʻzaki va yozma nutqda koʻp qoʻllaniladigan, turli sohaga oid birikma terminlar: *milliy daromad, daromad soligʻi, atrof muhit, maʼmuriy jazo, mehnat daftarchasi, qon bosimi, jarima toʻpi* va sh.k.

– Maʼnosi tarkibidagi soʻzlar maʼnosiga teng boʻlmaydigan (odatda “qoʻshma soʻz” deb qaraluvchi) *nazar qilmoq, javob bermoq, loyiq koʻrmoq, ozor bermoq* va sh.k.

– Tarkibidagi biror soʻz koʻchma maʼnoli boʻlgan birikmalar: *oq koʻngil, koʻngli oq, qora koʻngil, koʻngli qora, qoʻli gul* va sh.k.

– Koʻchma maʼnoli iboralar (frazologizmlar): *ammaning buzogʻi, uchchiga chiqqan, oyogʻini qoʻlga olmoq, toshingni ter* va sh.k.

– Izoh talab etuvchi maqol va matallar: *Temirni qizigʻida bosmoq. Sinamagan otning sirtidan oʻtma. Soʻqir koʻzga surma. Koʻza kunda emas, kunida sinadi* va sh.k.

Demak, turgʻun birikmalar va yasama soʻzlar korpusda alohida mustaqil lemma mavqeyida boʻladi va shunday teglanadi. Masalan, <ammamning buzogʻi>, <ilonning yogʻini yalagan>, <ichi qora>, <daromad soligʻi>, <uchiga chiqqan>, <koʻngli oq>.

Izohli lugʻatda bosh soʻz mustaqil qoʻllanmay, biror soʻz bilan yoki turgʻun ibora tarkibida kelgan boʻlsa, uning ketidan ostin-ustun ikki nuqta (:) qoʻyilib, soʻngra shu soʻz bilan berilgan:

LOM-MIM: lom-mim demaslik...

BELMA-BEL: belma-bel olishmoq...

Bu kabi birliklarni korpus birligi sifatida olish uchun ibora shakli lemma sifatida belgilandi: <belma-bel olishmoq>, <lom-mim demaslik>. Iboralar otga teng bo'lsa bosh kelishikda, fe'lga teng bo'lsa -moq shaklida beriladi ammo ba'zi iboralar nutqda ma'lum bir shakldagina qo'llaniladi. Shuning uchun mazkur tipdagi iboralar nutqda qanday shaklda bo'lsa, lug'atga ham xuddi shu shaklda beriladi. Masalan, *tomdan tarasha tushganday, yog' tushsa yalaguday, qovog'i osilgan, tepa sochi tikka bo'ldi* va sh.k.

Tilning lug'at boyligiga semantik maydon nazariyasidan kelib chiqib, korpusda tushunchadan so'zga boriladi, so'zlararo leksik-semantik munosabat orqali esa so'zdan tushunchaga borish mumkin [D.Ahmedova, 2020: 35]. Iboralar anglatadigan ma'no so'zlarning o'zaro munosabatidan hosil qilinadi faqat bu so'zlarning o'z ma'nosidan emas, balki ustama ma'nosi orqali ifodalanadi. Shuning uchun iboralarni korpusda teglashda so'zdan tushunchaga borish maqsadga muvofiq. Bunda yuqorida ta'kidlanganidek, izohli lug'at maqolasining tuzilishi tajribasidan foydalanish yaxshi natija beradi.

Mutaxassislar semantik annotatsiyalashning ikki xil: faset va shajara usulini farqlashadi [2]. Ko'pincha, korpus semantik razmetkasida har ikki usuldan ham foydalanishadi. Chunki so'z yoki turg'un birikmalar birdan ortiq ma'noni ifodalashi mumkin, shuning uchun har ikkalasidan ham foydalaniladi. So'z birdan ortiq ma'noni ifodalaganda shajara usuli to'g'ri natija bermaydi, ko'p ma'noli so'zlarni tasniflashda faset usuli qulay. Masalan, *ko'ngli aynidi, boshi aylandi* iboralari nutq vaziyatiga qarab quyidagi ma'nolarni bildiradi [Sh.Rahmatullayev, 1978: 154].

Ko'ngli aynidi kimning nimadan. 1. (“ovqat”dan). Organizmi qabul qilmagani tufayli behuzur bo'lmoq. *Bektemir “kabob”ni ko'p pishirdi, ko'ngli ayniguncha yedi.* Oybek “Quyosh qoraymas”.

2. (“ovqat”dan boshqa narsadan). Tabiatiga yoqmagani tufayli ijirg'anmoq. *Karim bilan Oygiz shuncha maqtadiki, eshitsangiz, ko'nglingiz aynib ketadi.* Sh.Rashidov “Bo'rondan kuchli”.

Boshi aylandi kimning. 1. Behud bo'lmoq, muvozanatni yo'qotmoq (organizmning kasallangan holati, bunda go'yo atrofdagi hamma narsa aylanib ketayotgandek tuyiladi). *U yerda ... allaqanday ma'nosiz, boshni aylantiradigan doimiy bir g'ovur-g'uvur...* Oybek “Qutlug' qon”.

2. nimadan. Esankiramoq. ... *muloyim qarab qo'yishlarini, ... jilmayib gapirishlarini sezardim, buning hammasidan boshim aylanardi.* O.Yoqubov “Muqaddas”

Ko'rib turganimizdek, ko'p ma'noli ibora ma'nolarini farqlash filtr yordamida amalga oshiriladi lekin birikma holidagi birliklarni kompyuterda bir qancha muammolar tug'diradi. Bunday birliklarni semantik teglashning bir necha yechimini taklif qilish mumkin. Ko'p ma'noli iboralar uchun alohida lingvistik ta'minot ishlab chiqilishi lozim. Bu jarayonning algoritmi taxminan quyidagicha bo'ladi [D.Ahmedova, 2020: 74]:

1) ta'kidlanganidek, izohli lug'atlarda mavjud ma'nolari yig'iladi va ulardan lingvistik ta'minot tuziladi;

2) iboralar lemmalanadi, ya'ni bitta lemma(so'z)ga tenglashtiriladi. Masalan, <boshi aylandi>, <ko'ngli aynidi> kabi.

Bizga ma'lumki, ko'p ma'noli so'zlar nutqda o'zi birikib kelayotgan “qo'shni” so'zga qarab belgilanadi. Korpus razmetkasi ham polisemantik xususiyatga ega bo'lgan so'zlarni shunday aniqlaydi. Bunday algoritmi yaratish uchun qoidalar majmuyi mutaxassis tomonidan lingvistik ta'minot bazasiga kiritilishi kerak.

D.Ahmedova o'z tadqiqoti [D.Ahmedova, 2020]da ko'p ma'nolilikni filtrlash algoritmi bo'yicha to'xtalib o'tgan. Bunda u “Razmetka tizimida birlik razryadini belgilash so'zning narsa/shaxs oti bilan birika olish qobiliyatiga qarab farqlanadi: *yengil* sifati narsa oti bilan birikma hosil qilganda “*og'ir emas*” ma'nosini anglatadi. Bu birlikning keying razryadlari esa nopredmet otlar taksonomlari bilan chegaralanadi”. Shunga amal qilgan holda ko'p ma'noli iboraning filtri quyidagi ko'rinishda bo'lishi mumkin:

Ibora	Kontekst	Yakuniy ma'no
<i>ko'ngli aynidi</i>	+ narsa oti (ovqat)	Sem=razryad “harakat-holat” taksonomik sinf: fizik holat
<i>ko'ngli aynidi</i>	+ shaxs + hid + ...	Sem=razryad “harakat-holat” taksonomik sinf: jismoniy holat

Shuni ham aytib o'tishimiz kerakki, semantik razmetka qanchalik mukammal tuzilgan bo'lmasin, so'zning ma'no qirralarini aniq ko'rsata olmaydi.

Demak, iboralarni semantik teglashda ular ikki yoki undan ortiq so'zdan tashkil topganligi uchun ularni bitta lemmaga biriktiriladi hamda ular anglatgan ma'no so'zlarning o'zaro aloqasidan hosil bo'layotganligini inobatga olib, so'zdan tushunchaga borish tamoyili asosida ma'nosi izohlanadi.

Foydalanilgan adabiyotlar:

1. Mirziyoyev Sh. O'zbek tili xalqimiz uchun milliy o'zligimizni va mustaqil davlatchilik timsoli, bebaho ma'naviy boylik, buyuk qadriyatdir. Prezident Sh.Mirziyoyevning o'zbek tiliga davlat tili maqomi berilganining o'ttiz yilligiga bag'ishlangan tantanali marosimidagi nutqi. 2019-yil, 21-oktyabr. <https://www.gazeta.uz/oz/politics/>

2. Axmedova D.B. Atov birliklarini o'zbek tili korpuslari uchun leksik-semantik teglashning lingvistik asos va modellari. Filol. fan. bo'yicha falsafa dokt. (PhD). diss. - Buxoro, 2020.



3. Hamroyeva Sh. O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Filol. Fan. bo'yicha falsafa dokt. (PhD) diss. - Buxoro, 2018. B-73.
4. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с. – С. 48-49.
5. Elov B., Axmedova X.I. Determining homonymy using statistical methods. // COMPUTATIONAL MODELS AND TECHNOLOGIES, Uzbekistan-Malaysia international conference, September 16-17th, 2022, Tashkent. 105-106.
6. Elov B., Axmedova X. Polisemantik so'zlarni semantik tahlil qilishning matematik modellari. O'zbekiston: til va madaniyat (Amaliy filologiya), 2021, 1(4). 77-85.
7. O'zbek tilining izohli lug'ati (A.Madvaliev tahriri ostida). 5 jildli. – Toshkent: “O'zbekiston milliy ensiklopediyasi” Davlat ilmiy nashriyoti, 2000-2006.
8. Куканова В.В. Принципы семантической разметки национального корпуса калмыцкого языка // http://kalmcorpora.ru/sites/default/files/kukanova_25.pdf.
9. Rahmatullayev Sh. O'zbek tilining izohli frazeologik lug'ati. – Toshkent: “O'qituvchi” nashriyoti, 1978. B-154.
10. Elov B., Axmedova X. Business Process Modeling That Distinguishes Homonymy Within Three Parts of Speeches in Uzbek Language. // 7th International Conference on Computer Science and Engineering, 14-16 September 2022, Turkey. 289-294.

MATNGA QAYTA ISHLOV BERISH TIZIMLARI MORFOLOGIK ANALIZATOR XUSUSIDA

Abduraxmanova Muqaddas Tursunaliyevna

ToshDO‘TAU doktoranti
f.f.n., dotsent

MParfi2005@yandex.ru

Olimova Muxlisa Parmonovna

O‘zMU 2-kurs talabasi

olimovamuxlisa556@gmail.com

Annotatsiya. Maqolada jahon tilshunosligida ATTdagi lingvistik dasturiy ta'minotlar, xususan, morfologik analizatorlar yaratish masalasi xususida fikr yuritilgan. Dunyo va o'zbek tilshunosligida morfoanalizatorlar ishlab chiqish borasidagi olib borilayotgan tadqiqot ishlariga to'xtalib o'tilgan. Mavjud morfologik tahlil algoritmlarining kelgusida qilinajak ishlarda katta asos vazifasini o'tashi misollar yordamida tahlil etilgan. Shuningdek, tizimdagi ayrim kamchiliklar sabablari ochib berilgan.

Kalit so'zlar : *morfologik analizator, korpus, trop, algoritm, lemma, token, morfotaktik imkoniyat.*

Annotation. The article discusses the fact that the issue of creating linguistics software, in particular, morphological analyzers, is one of today's urgent tasks in automatic editing and analysis in the world linguistics. Research work on the development of morfoanalyzers in world and Uzbek linguistics is discussed. The fact that existing morphological analysis algorithms will play a major role in future work is revealed with the help of examples. Also, the reasons for the deficiencies in the system that need to be corrected are explained.

Key words: *morphological analyzer, corpus, trop, algorithm, lemma, token, morphotactic opportunity.*

Аннотация. В статье рассматривается задача в создании лингвистического программного обеспечения, в частности, морфологических анализаторов, которая является одной из актуальных задач автоматического редактирования и анализа на сегодняшний день в мировой лингвистике. Обсуждается научно-исследовательская работа по развитию морфоанализаторов в узбекском языкознании. На примерах показано, что существующие алгоритмы морфологического анализа играют большую роль в дальнейшей работе. Также объясняются причины и недостатки в системе, которые необходимо исправить.

Ключевые слова: *морфологический анализатор, корпус, троп, лемма, токен, морфотактическая возможность.*

Dunyo tilshunosligida avtomatik tahlil va tahrir yo'nalishida imlo tekshiruvchi, morfologik va sintaktik tahlil qiluvchi tizimlarning lingvistik

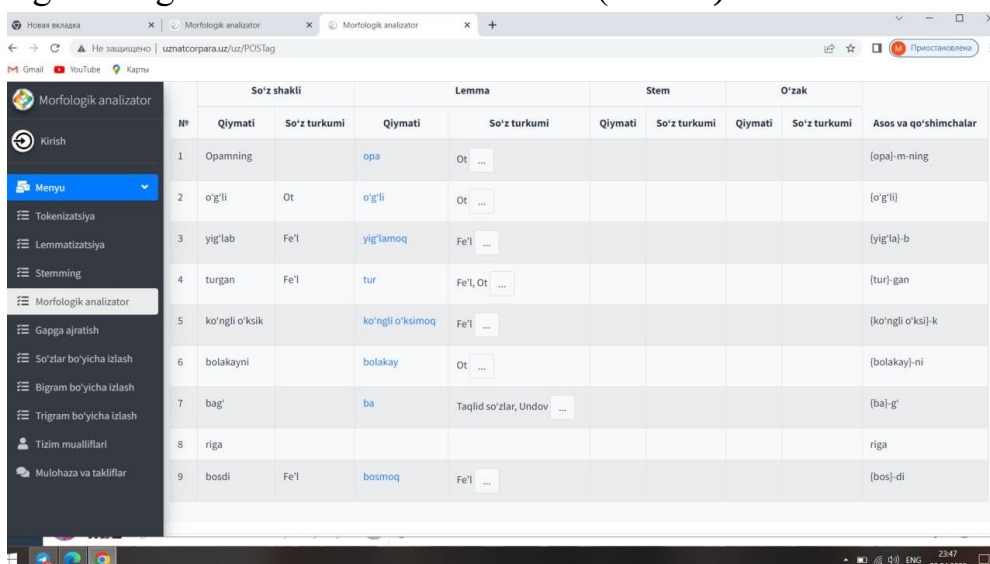


ta'minotlari ishlab chiqilgan. Ular tabiiy tilni qayta ishlashda til materialini tahlil va tahrir qiluvchi tezkor va har tomonlama tejamkor tizimlar yaratilishiga asos bo'lishi bilan birga mashina tarjimasi, elektron lug'atlar, parserlar (matndagi muayyan sintaktik birlik va strukturalarni avtomatik analiz algoritmi) ishlash sifatini oshirishda ham muhim ahamiyat kasb etmoqda. Zero, bugungi globallashuv jarayonida jahon kompyuter tilshunosligidan ortda qolmagan holda katta hajmli axborot matnlari ustida tezkor va samarali ishlash zarur. Shu ma'noda, avtomatik tahrir dasturlari ishlab chiqilmoqdaki, ular rus, ingliz, italyan, nemis, ozarbayjon, arab, turk, qirg'iz kabi tillardagi matnlarni tahrir qilishga qaratilgan.

Morfologik analizator so'zlar va so'z shakllarini lug'atdagi shakli bilan taqqoslash, ularning asosi va so'zning grammatik shakllarini aniqlash hamda ularga tavsif berish bilan shug'illanadigan dasturiy ta'minot hisoblanadi. Bugunga qadar jahon tilshunosligida bir qancha olimlar tomonidan morfoanalizatorlar yaratish masalasi o'rganilgan va salmoqli amaliy ishlar qilingan. Jumladan, 2004-yilda professorlar Konstans O'Sullivan va Charlotta Ebellar tomonidan yaratilgan "Ingliz va fransuz darslarida lug'atni o'rgatishda morfologik analizdan foydalanish" dasturi [Sullivan, 2004], "Ko'kborak" lug'atini shakllantirish [Sarkar, 2015], Malay va Tamil tillarining mashina tarjimasini yaratish morfologik tahlilning rivojiga turtki bo'lganligi [Rajeev, 2011], arab tilshunoslari Y. Jaafar va K. Bouzoubaalar yaratgan arab morfologik analizatori fikrimizning isboti bo'la oladi. Shuningdek, hozirgi kunda ham morfologik analizatorlar yaratish ustida jiddiy ishlar amalga oshirilmoqda. Misol tariqasida turkiy tillar doirasida vujudga kelayotgan algoritmlarni aytishimiz joiz. Chunonchi, Adnan Öztürel, Tolga Kayadelen va Isın Demirsahin turk tili morfologiyasining keng qamrovli modelini va uni amalga oshiruvchi ochiq manbali morfoanalizatorini taqdim etishgan [4]. Ushbu morfologik tahlil dasturidan morfologiya va sintaksisni to'liq qamrab olganligi tufayli turk tili modelini ishlab chiqishda foydalanish mumkin. Nafaqat turk tili, balki turkiy tillar agglyutinatив xarakterga ega bo'lganligi sababli bu tillarda morfoanalizator yaratilishida morfotaktik holat asosiy sanaladi, chunki asos (o'zak)dan keyin qaysi turdagi qo'shimchalar ketma-ketlikda qo'shilishi muayyan grammatik qoidaga asoslanadi. Xususan, o'zbek tilining morfotaktik imkoniyati quyidagicha bo'ladi: *Prefiks (old qo'shimcha) asos + yasovchi qo'shimcha + lug'aviy shakl yasovchi qo'shimcha + sintaktik shakl yasovchi qo'shimcha*. Ta'kidlash joizki, sof turkiy tillarda old qo'shimcha (prefiks) mavjud emas, biroq bu o'rinda fors-tojik tillaridan o'zlashgan *be-*, *no-*, *bo-*, *ba-*, *bar-*, *kam-*, *ser-*... kabi qo'shimchalar nazarda tutilgan. Tilshunoslik ham aniq fan bo'lib, istisnolarni qabul qilmasligi kerak. Bizningcha, o'zbek tilidagi *muzlatgich*, *o'chirgich* kabi so'zlarni ajratishda *muzlat + gich*, *o'chir + gich* shaklida bo'lishi yuqoridagi fikrimizni asoslaydi.

Bu qoidalarga asoslangan holda o'zbek tilida ham o'zbek tili korpusi tarkibida [5], <http://uznatcorpara.uz> da morfologik analizatorlar ishlab chiqilgan. Dasturiy ta'minot berilgan matndagi har bir tokenni tahlil qiladi va turkum, turli grammatik ma'nolar kabi ma'lumotlarini shakllantiradi. Qidiruvga berilgan so'zning turkumini

aniqlash uchun esa morfologik, semantik, sintaktik qoidalarga asoslanadi. Bu kabi lingvistik analizatorlar tabiiy tilga ishlov berish, axborotni qayta ishlash sifatini yaxshilash, Google translate tarjimonida avtomatik ishlov berish ko'nikmasini oshirish, o'zbek tilida turli korpuslar yaratish, korpus birliklariga avtomatik ishlov berish, korpusda qidiruv tizimini yo'lga qo'yishda muhim ahamiyat kasb etadi. Biroq hozirda mavjud morfonalizatorlar har tomonlama mukammal deya olmaymiz. Morfotahlil ishlash samaradorligi yuqori bo'lishi uchun tabiiy tillarning ayrim jihatlari, ya'ni omonimlik, paronimlik, polisemiya, frazeologizmlar, troplar to'sqinlik qiladi. Bu to'siqlarning yechim yo'llari izlanmoqda. Masalan, omonimlikni bartaraf etish texnologiyalari ishlab chiqilyapti. Buning uchun shakldosh so'zlarning har birini tasniflash lozim, ya'ni ularni lemma – gap bo'lagi va morfologik xususiyatlar majmuyi bilan taqqoslash mumkin. Kollokatsiya (semantik qurshov) usulidan ham o'zbek tilidagi omonim so'zlarni ajratishda foydalanish mumkin [Abdurahmanova, 2021]. Bu usul *oshiq* omonimik qatori misolida tahlil qilingan. Bunda tahlilga tortilgan *oshiq* so'zining yuzaga chiqqan barcha ma'nolari to'g'ri talqin etilganligi semantik qurshovning morfoanalizatorida unumdorlikni oshirishga xizmat qilishini ifodalaydi. Aytish joizki, semantik qurshov polisemiya, troplar masalasida ham ishonchli yechim bo'la oladi. Shuningdek, morfoanalizator to'g'ri ishlashiga putur yetkizadigan yana bir omil bu tahlil jarayonida so'zlarning morfonologik nuqtayi nazardan to'g'ri bo'linmasligi hisoblanadi. Bunga, asosan, so'zga qo'shimchalar qo'shilganda yuzaga chiqadigan fonetik o'zgarishlar sabab bo'ladi. Natijada so'zlar ma'nosi buzib talqin qilinadi, ma'lum so'z bo'g'inlarga bo'linib ikkita so'z sifatida namoyon bo'lib qoladi, so'z turkumlariga noto'g'ri tasniflanishlar kuzatiladi (1-rasm):



№	So'z shakli		Lemma		Stem		O'zak		Asos va qo'shimchalar
	Qiymati	So'z turkumi	Qiymati	So'z turkumi	Qiymati	So'z turkumi	Qiymati	So'z turkumi	
1	Opamning		opa	Ot ...					[opa]-m-ning
2	o'g'li	Ot	o'g'li	Ot ...					[o'g'li]
3	yig'lab	Fe'l	yig'lamoq	Fe'l ...					[yig'la]-b
4	turgan	Fe'l	tur	Fe'l, Ot ...					[tur]-gan
5	ko'ngli o'ksik		ko'ngli o'ksimoq	Fe'l ...					[ko'ngli o'ksi]-k
6	bolakayni		bolakay	Ot ...					[bolakay]-ni
7	bag'		ba	Taqlid so'zlar, Undov ...					[ba]-g'
8	riga								riga
9	bosdi	Fe'l	bosmoq	Fe'l ...					[bos]-di

1-rasm. Morfoanalizatorning ishlashidan namuna

Mazkur interfeysdan ko'rinib turibdiki, *Opamning o'g'li yig'lab turgan ko'ngli o'ksik bolakayni bag'riga bosdi* gapi shakl qiymati, so'z turkumlariga ajratish, asos va qo'shimchalarga bo'linish jihatlaridan yangilish tahlil qilingan.

Xatoliklar omonimlik (*turgan – tur*), morfonologik (*o‘g‘li, yig‘lab, bag‘riga*), asos va qo‘shimchaga ajratish bo‘yicha yuzaga kelgan.

Xulosa o‘rnida ayta olamizki, jahon tilshunosligi tajribasida ko‘plab morfoanalizatorlar ishlab chiqilgan, ammo bir tilning qonun-qoidalari boshqa bir tilga to‘g‘ri kelmasligi sababli o‘zbek tilining ichki imkoniyatlaridan kelib chiqib barcha qoidalar va istisnoli(dasturdagi) holatlarga asoslangan lingvistik tahlil metodlari yaratilishi galdagi vazifalardan biridir. Morfologik analizatorga til leksikasining chastotali so‘zlari, tildagi mavjud so‘zlarni qayta ishlashga oid qoidalar to‘plami kiritilishi kerak. Bundan tashqari, morfoanalizator so‘z tahlili haqidagi tahminlarni yaratishga imkon beradigan moslashuvchan sozlama sifatida shakllantirilishi kerak va bunday algoritm omonimlar muammosiga o‘xshash turli yanglishmovchiliklarni oldini oladi. Shu bilan bir qatorda, aniq tahlilni amalga oshirish uchun asos (o‘zak)ga qo‘shimchalar qo‘shilishi natijasida hosil bo‘ladigan fonetik (morfonologik) o‘zgarishlarlar bo‘yicha aniq qoliplar yaratilishi lozim.

Foydalanilgan adabiyotlar:

1. Constance O'Sullivan, Charlotte Ebel. Using morphological analysis to teach vocabulary in English and French classes. Teachers as Scholars Institute Princeton University. July, 2004. Marguerite Browning, Professor.
2. Partha Sarkar, Dr. Bipul Syam Purkayastha. Morphological analyzer in the development of bilingual dictionary (Kokborok - English) - An analysis for appropriate method and approach. // International journal of engineering and technology (IJET). Vol.4, Issue 10, April 2015 - 98-103pp.
3. Rajeev R. R., Jisha P. Jayan, Dr. Rajendran S. Morphological analyser and morphological generator for Malayalam – Tamil machine translation. // International journal of computer applications. Vol.13 – No.8, January 2011.
4. <http://aclanthology.org/W19-3110.pdf>
5. <https://uzbekcorpus.uz>
6. Abdurahmanova M. T., Rahmanova A. A. O‘zbek tili omonimlari uchun milliy teglar lug‘ati. – T.: Lesson press, 2021. – 9-11 b.
7. Elov B., Hamroyeva Sh., Axmedova X. Methods for creating a morphological analyzer. // 14th International Conference on Intellegent Human Computer Interaction, IHCI 2022, 19-23 October 2022, Tashkent.
8. Elov B., Hamraeva Sh., Elova D. Morfologik analizatorni yaratish usullari. O‘zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(1). 67-87.
9. Abdullayeva O. Morphological annotation system in the corpus of internet information texts in Uzbek language. // 7th International Conference on Computer Science and Engieering, 14-16 September 2022, Turkey. 160-164.

O‘ZBEK TILI MATNLARIDA KOREFERENSIYANI HAL QILISH BOSQICHLARI

Elova Dilrabo Qudratullayevna

Filologiya fanlari falsafa doktori PhD,
ToshDO‘TAU katta o‘qituvchisi

elova@navoiy-uni.uz

Annotatsiya. Tabiiy tilni qayta ishlash (Natural language processing, NLP) metodlari orqali odamlar va mashinalar o‘rtasida aloqa o‘rnatiladi. NLP – sun‘iy intellektning murakkab sohalaridan biri bo‘lib, inson tabiiy tili kompyuterlar o‘rganishi qiyin bo‘lgan semantik noaniqliklarga boydir. Bu muammoni bartaraf qilish usullaridan biri kontent mazmunini tushunishga xalal beradigan noaniq va umumiy mazmunni ifodalaydigan ifodalarni aniqlashtirishdan iborat. Kompyuterlar matndagi noaniq ifodalar qatnashgan jummalarni tushunishi osonroq bo‘lgan aniq ifodalarga almashtirish NLPda koreferensiya muammosini hal qilish orqali amalga oshiriladi. Ushbu maqolada o‘zbek tili matnlarida koreferensiya muammosi hal etish bosqichlari keltiriladi.

Kalit so‘zlar: *koreferensiya, anaphora, anaforik munosabat, anaforik so‘z, koreferensiyani hal qilish, katafora, antecedent, postsedent, eslatma.*

Abstract. Communication between people and machines is established through the methods of natural language processing (NLP). NLP is one of the complex areas of artificial intelligence, and human natural language is rich in semantic ambiguities that are difficult for computers to learn. One way to overcome this problem is to clarify expressions that express vague and general content that hinders the understanding of content content. In order to replace a sentence with ambiguous expressions in the text into an easier-to-understand expression, it is necessary to solve the problem of coreference in NLP. This article presents the stages of solving coreference problems in Uzbek language texts.

Key words: *coreference, anaphora, anaphoric relationship, anaphoric word, resolution of coreference, cataphora, antecedent, postsedent, note.*

Абстракт. Связь между людьми и машинами устанавливается с помощью методов обработки естественного языка (Natural language processing, NLP). NLP — одна из сложных областей искусственного интеллекта, а естественный человеческий язык богат семантическими неоднозначностями, которые трудно изучить компьютерам. Одним из способов преодоления этой проблемы является уточнение выражений, выражающих расплывчатое и общее содержание, затрудняющее понимание содержания. Для того чтобы заменить предложение с неоднозначными выражениями в тексте на более простое для понимания выражение, необходимо решить проблему кореферентности в NLP.

В данной статье представлены этапы решения проблем кореференции в текстах на узбекском языке.

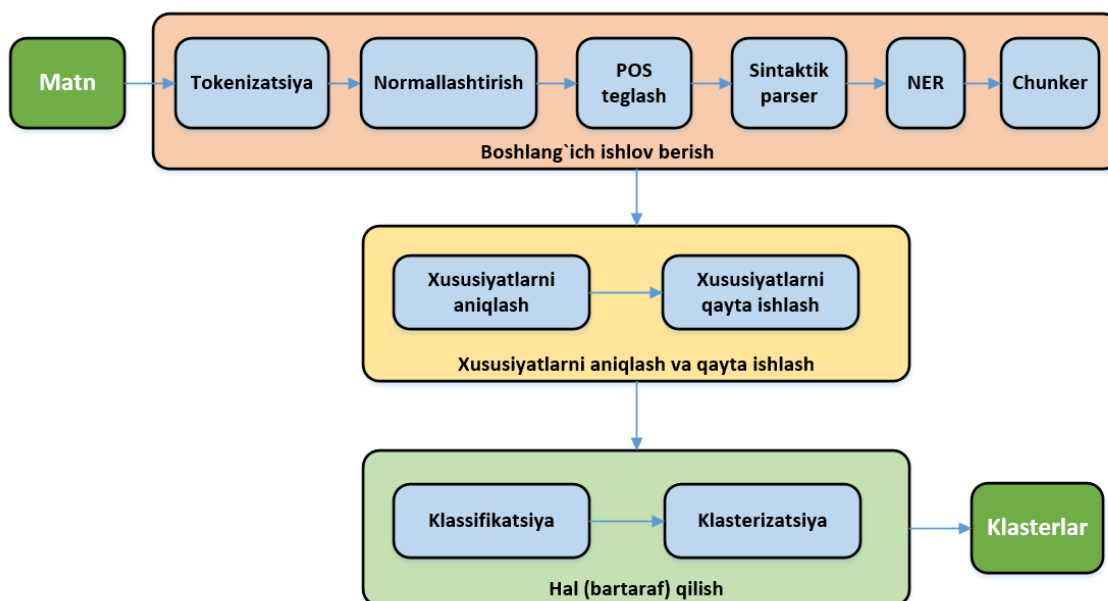
Ключевые слова: *кореференция, анафора, анафорическое отношение, анафорическое слово, разрешение кореферентности, катафора, антецедент, апостериор, нота.*

Koreferensiya muammosi hal qilish (Coreference resolution, CR) – bu matndagi bir obyektga tegishli bo‘lgan barcha eslatmalar (lingvistik ifodalar)ni aniqlashga asoslangan NLP vazifasi (1-rasm). Ushbu *eslatmalarni aniqlash* va *guruhlash (klasterlash)* orqali **noaniq ifodalarni obyektning aniq nomlari bilan almashtirib** matndagi koreferensiya muammosini bartaraf qilish mumkin. Obyekt nomlari **bir so‘zdan iborat token** bilan ham, yaxlit sintaktik butunlik - **otli birikmalar** bilan ham ifodalanishi mumkin.

“Men Tohir Malikni tanlayman, chunki u mening qiziqishlarimga mos asarlar yozgan” – dedi Akbar

1-rasm. Gapdagi bir obyektga tegishli bo‘lgan barcha eslatmalar

Koreferensiya muammosi hal etish – bu matnni *tushunish, ma’lumot olish, mashinali tarjima qilish, kayfiyatni tahlil qilish* yoki *hujjatlarni umumlashtirish* kabi ko‘plab NLP ilovalarida qo‘llanilishi mumkin bo‘lgan muhim NLP vazifasi [1]. Koreferensiya muammosini hal etish kompyuterlar matndagi noaniq ifodalar qatnashgan jummalarni tushunishi osonroq bo‘lgan aniq ifodalarga almashtirishning zamonaviy usuli hisoblanadi. Quyidagi 2-rasmda o‘zbek tili matnlarida koreferensiya muammosi hal etish bosqichlari keltirilgan.



2-rasm. Koreferensiya muammosi hal etish bosqichlari

Koreferensiya va anaforik munosabat

Shuni ta'kidlash kerakki, koreferensiya muammosini hal qilishda *matndagi havola (eslatma)lar* va *ular orasidagi koreferent bog'liqlikni aniqlash* lozim. Biroq, matnda bir nechta turdagi havolalar mavjud bo'lib, ularning ta'riflari munozarali masaladir.

Koreferensiya muammosini hal qilishda ajralib turadigan bir holat **anafora muammosini hal qilish (anaphora resolution, AR)** hisoblanadi. Matnda bir atama ikkinchisiga ishora qilganda, ikkinchisining yagona talqinini belgilaganda anafora koreferentlikni (anaforada har doim ham koreferentlik xususiyati namoyon bo'lmaydi) yuzaga keltiradi. [2].

Quyida keltirilgan misolda (1) va (2) ifodalar to'g'ridan-to'g'ri turli xil real obyektlarni nomlaganini ko'rish mumkin. Ammo bu ifodalar bir jumlada qo'llanganligi sababli, (2)ning talqini (1)ga tayanadi. Ushbu eslatmalar o'zaro koreferent bo'lmasa-da, anaforik munosabatni hosil qiladi.

Aziza ¹ tug'ilgan kuniga yaqin qolganda, **uyni** ² chinniday tozaladi.

Anaforani hal etish masalasi koreferensiyani bartaraf qilishdan farq qilsa-da, aksariyat hollarda biri boshqasiga teng keladi.

NLPda koreferensiyani hal qilishdagi asosiy terminlar ta'rifini keltiramiz:

– **Anafora** – o'zidan avval keltirilgan shaxs, narsa, holat kabilarni ko'rsatuvchi, ishora qiluvchi, antesedentga bog'liq so'z[3]. Anaforik so'z real borliqdagi bir obyektga yoki turli obyektlarga ishora qilishi mumkin. O'zbek tilshunoslida Azim Hojiyev anaforik so'z sifatida olmoshlarning ba'zi turlarini ko'rsatib o'tgan [4]. Lekin anaforik munosabatlar nafaqat olmoshlar vositasida, balki qarindoshlikni ifodalovchi nomlar, kasb-hunar nomlari, lavozim nomlari kabi so'zlar vositasida ham yuzaga keladi.

– **Katafora** – bu anaforaning bir ko'rinishi bo'lib, olmosh u nazarda tutgan otdan oldinroq ifodalanishi bilan farqlanadi [3].

– **Koreferent** – real borliqdagi bir obyektga ishora qiluvchi nomlar [3].

Anafora va Katafora

Anafora va kataforaning asosiy farqi shundaki, *anafora gapda u nazarda tutgan so'zdan keyin, katafora esa undan oldin joylashadi*. Anaforadan oldin joylashgan nom **antesedent**, kataforadan keyingi so'z esa **postsedent** deb ataladi.

O'zbekiston terma komandasi JChda g'alabani qo'lga kiritdi. Bu ularning ilk musobaqasi edi.
antecedent *anafora*

Maktabda uning baholari juda yomon bo'lishiga qaramay, Tomas Edison buyuk ixtirochi bo'ldi.
katafora *postcedent*

Antesedentlarni ajratish

Bunda (2) olmoshi bir nechta oldingi (1) ni anglatadigan anaforik ifoda.

Hamid Ne'matov va Asqar G'ulomov ¹ - buyuk tilshunos olimlar edi.

Ular ² ta'limga sezilarli hissa qo'shishda hamkorlik ham qilishgan.

Noaniq olmoshli havolalar

Ba'zi matnlarda ifodalangan olmosh va boshqa nomlar o'rtasida hech qanday bog'liqlik bo'lmasa-da, koreferensiya muammosini hal etish algoritmini yaratishda ushbu turdagi havolalarni e'tibordan chetda qoldirmaslik kerak. Bunday havolalarga gapdan chiqarilganda ham mazmunga putur yetkazmaydigan olmoshlarni, badiiy matnda, so'zlashuv uslubida ta'sirchanlikni oshirish, ta'kidni kuchaytirish maqsadida o'zi ko'rsatayotgan obyekt nomi bilan yonma-yon qo'llangan olmoshlarni, pleonastik qo'llangan olmoshlarni, rasmiy ish uslubida buyruq va talabnomalarda qat'iylikni oshirish uchun ishlatilgan olmoshlarni, tarjimai hol, tushuntirish xati yozilganda atoqli ot bilan yonma-yon kelgan olmoshlarni kiritishimiz mumkin.

Koreferensiyani hal qilish bosqichlari

Yuqorida keltirilgan fikr mulohazalardan kelib chiqqan holda, o'zbek tili matnlarida koreferensiyani hal qilish uchun quyidagi 3-bosqichdan iborat amallarni bajarish lozim:

1-bosqich. *Potentsial so'z, birikma(nomzod)larni aniqlash.*

2-bosqich. *Nomzodlarni guruhga ajratish.*

3-bosqich. *Olmoshlarni haqiqiy obyekt nomlari bilan almashtirish.*

1-bosqich. *Potensial nomzodlarni aniqlash.*

Koreferensiya muammosini hal qilishning dastlabki bosqichida alohida **so'zlar/tokenlar** yoki **spanlarni** qayta ishlash lozim [5].

Span nima? Ko'p hollarda, matndagi olmosh bog'langan obyekt bir nechta qo'shni tokenlardan iborat bo'ladi. Shu sababli koreferensiyani hal qilish NLP vazifasida span sifatida gapdagi o'zaro bog'liq tokenlar ketma-ketligi (otli birikmalar) tushuniladi. NLP vazifalarida aksariyat hollarda matndagi **NER** obyektlar **span** sifatida talqin qilinadi [6]. Shunigdek ko'plab **eslatma/havola**

termini ham span o'rnida qo'llaniladi. Ko'p hollarda bir necha tokenlardan iborat eslatmalardagi element (so'z)lar yaxlit ma'noni anglatadi.

1- bosqich amalga oshirilishi natijasida **olmoshlar**, ular ishora qilayotgan **nomzod spanlar ro'yxati** aniqlanadi.

2-bosqich. *Nomzodlarni guruhga ajratish*

Ushbu bosqichda nomzod spanlar muayyan shartlar va qoidalar asosida guruhlarga birlashtiriladi [7,8]. Nomzod spanlarni birlashtirish **klasterlash** yoki **guruhlash** jarayoni deb ataladi. Umumiy guruhga mansub spanlar umumiy ma'noni ifodalashi lozim. NLPdagi so'zlarni klasterlash *kutubxonadagi kitoblarni toifalash* vazifasiga o'xshaydi. Bugungi kunda guruhlash (toifalash)ning ko'plab usullari mavjud bo'lib, guruhdagi element (span)lar o'xshash xususiyatlarga ega bo'lishi va boshqa klasterlardan farq qilishi kerak.

3-bosqich. *Olmoshlarni haqiqiy obyekt nomlari bilan almashtirish.*

Matndagi **[men, mening]** va **[u]** olmoshlari mos ravishda “**Akbar**” va “**Tohir Malik**” real referent nomlari bilan almashtirilishi natijasida, olmoshsiz semantik jihatdan o'zgarmagan matn hosil qilinadi.

“**Men** **Tohir Malikni** tanlayman, chunki **u** **mening** qiziqishlarimga mos asarlar yozgan” – dedi **Akbar**

Xulosa

O'zbek tili matnlarida koreferensiyani hal qilish NLP vazifasining maqsadi har qanday noaniq ifodalarni ular ishora qilayotgan obyektlarning real nomlarini *aniqlash, guruhlash* va *almashtirishdan* iborat. Ushbu maqolada koreferentlik va anaforik munosabatlarning o'xshash jihatlari hamda farqli tomonlari muhokama qilindi. Shuningdek, ularning o'zbek tili matnlaridagi koreferensiya va anafora muammolarini hal qilish usullari keltirildi.

O'zbek tili matnlarida koreferensiyani hal qilish uchun *potensial so'z, birikma(nomzod)larni aniqlash, nomzodlarni guruhga ajratish* va *olmoshlarni haqiqiy obyektlar bilan almashtirish* bosqichlari amaliy misollar asosida tushuntirildi. Keyingi ilmiy tadqiqot ishlarida o'zbek tili matnlarida koreferensiyani hal qilish modellari haqida batafsil ma'lumot keltiriladi.

Foydalanilgan adabiyotlar:

1. Marta Maslankowska, Paweł Mielniczuk. Intro to coreference resolution in NLP.
2. Christopher Manning. Stanford lecture (CS224n). 2019
<https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1162/handouts/cs224n-lecture10-coreference.pdf>
3. R. Mitkov. Anaphora Resolution. Pearson Education, 2002.
4. Хожиев А. Тилшунослик терминларининг изоҳли луғати. – Тошкент: Ўзбекистон миллий энциклопедияси, 2002. – 168 б.



5. B.B.Elov, Sh.M.Khamroeva, Z.Y.Xusainova. NLP (tabiiy tilga ishlov berish)ning pipeline konveyeri (2022).
6. B.Elov. Tabiiy tilni qayta ishlash (nlp)da spacy modulidan foydalanish. Илим-фан ва инновацион ривожланиш, 2022 (4). 41-55
7. Urbizu, G., Soraluze, A., & Arregi, O. (2020). Sequence to Sequence Coreference Resolution. Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference, Crac.
8. Stylianou, N., & Vlahavas, I. (2021). A neural Entity Coreference Resolution review. In *Expert Systems with Applications* (Vol. 168). <https://doi.org/10.1016/j.eswa.2020.114466>

O'ZBEK TILIDAGI MATNLARDAN FE'L SO'Z TURKUMIGA OID SO'ZLARNI QOIDAGA ASOSLANGAN HOLDA QO'SHIMCHALAR ASOSIDA ANIQLASH

Sharipov Maqsud Siddiqovich

t.f.n, dots. Urganch Davlat Universiteti

maqsbek72@gmail.com

Yakubova Dilafruz Kadamboyevna

yaqubovadilafruz@gmail.com

Urganch Davlat Universiteti magistranti

Yuldashov Ollabergan Ergash o'g'li

ollaberganyuldashov@gmail.com

Urganch Davlat Universiteti o'qituvchisi

Annotatsiya. Bu ishda biz o'zbek tili matnlaridan fe'llarni aniqlashning qoidaga asoslangan algoritm yaratish jarayonini ko'ramiz. Fe'llarni matnlardan aniqlash va morfologik tahlil qilish o'zbek tili uchun tabiiy tillarni qayta ishlashning bir qator masalalarini (POS tagging, mashinali tarjimon va h.k.) yechishda foydali bo'ladi. Barcha turkiy tillar kabi o'zbek tili ham agglutinatив til bo'lib so'z o'zagi yoki negiziga qo'shimchalar qo'shilishi bilan yasaladi. Bu masalani yechish uchun tub fe'llar bazasi yaratildi, fe'lning so'z yasovchi va shakl yasovchi qo'shimchalari bazasi shakllantirildi. Bular yordamida o'zbek tilidagi matnlardan fe'l so'z turkumini aniqlovchi algoritm ishlab chiqildi.

Kalit so'zlar: *FSM, NFA, O'zbek tili, fe'l, FST*

Abstract. In this work, we will see the creation of a rule-based algorithm for identifying verbs from Uzbek language texts. Identifying verbs from texts and morphological analysis will be useful for solving a number of natural language processing issues (POS tagging, machine translation, etc.) for the Uzbek language. Like all Turkic languages, the Uzbek language is an agglutinative language and is formed by adding suffixes to the root or base of the word. To solve this problem, a base of basic verbs was created, and a base of word-forming and form-forming suffixes of the verb was formed. With the help of these, an algorithm was developed for identifying verbs from Uzbek texts.

Key words: *FSM, NFA, Uzbek language, verbs, FST.*

Абстракт. В этой работе мы увидим создание основанного на правилах алгоритма идентификации глаголов из текстов на узбекском языке. Выявление глаголов из текстов и морфологический анализ будут полезны для решения ряда задач обработки естественного языка (POS-теги, машинный перевод и т. д.) для узбекского языка. Как и все тюркские языки, узбекский язык является агглютинативным языком и образуется путем добавления суффиксов к корню

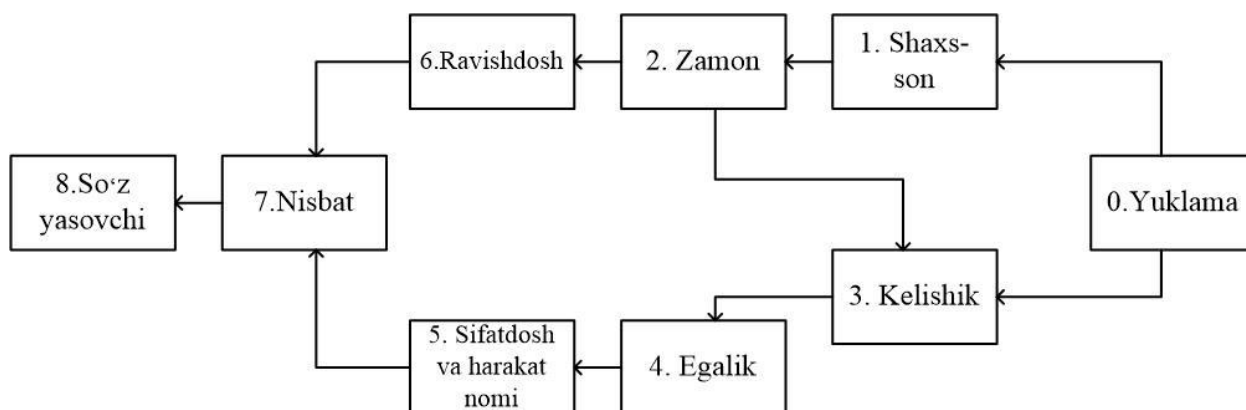
или основе слова. Для решения этой задачи была создана база основных глаголов, а также сформирована база словообразовательных и формообразующих суффиксов глагола. С их помощью был разработан алгоритм определения глаголов из узбекских текстов.

Ключевые слова: *FSM, NFA, Узбекский язык, глагол, FST.*

Tabiiy tillarni aniqlash (NLP: Natural Language Processing) sohasi boshqa sohalarga qaraganda o'zining dolzarbligi jihatidan ancha oldinlab ketdi va bu sohada barcha rivojlangan tillar qatori, o'zbek tili uchun ham bir qancha til resurslari va dasturiy vositalari ishlab chiqila boshlandi [1]–[3]. NLP sohasida lingvistik ma'lumotlar muhim rol o'ynaydi, shu jumladan, gap bo'laklari ham [4], [5]. O'zbek tili va boshqa ko'plab tillarda fe'l so'z turkumi juda katta rol o'ynab, ko'p hollarda gapning ildizi sifatida keladi. Shuning uchun bu so'z turkumi ustida izlanish olib borish muhim ahamiyat kasb etadi.

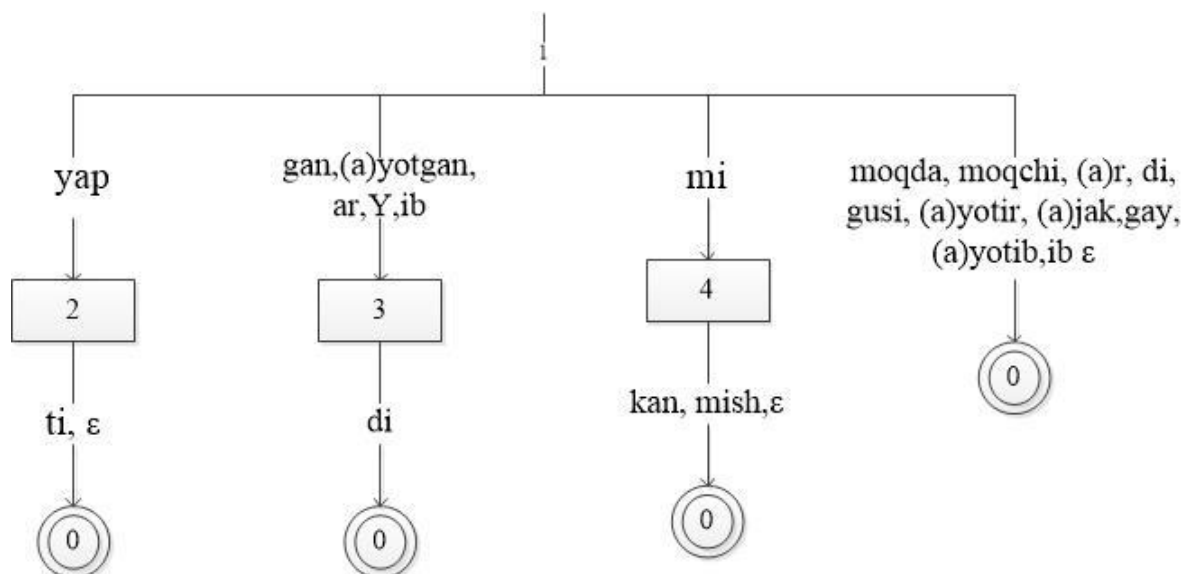
Ushbu maqolada o'zbek tilidagi matnlardan fe'llarni aniqlash algoritmi yaratish masalasi ko'riladi. O'zbek tilidagi fe'llar murakkab bo'lganligi uchun biz o'zbek tilidagi fe'llar uchun maxsus ma'lumotlar bazasini yaratdik. Fe'llar o'zbek tilida eng muhim va murakkab qismlardan biridir. Chunki fe'l butun jumlagga ta'sir qiladi va boshqaradi. Ushbu ish o'zbek tili uchun yaratilayotgan NLP modellar va dasturiy vositalar qatorida yangi ilmiy izlanishlar uchun imkon yaratadi. Shu paytgacha NLP sohasida o'zbek tili uchun bir qancha ilmiy ishlar qilingan, shular qatoriga o'zbek tili uchun yaratilgan nomuhim so'zlar dataseti [6], [7], lemmatizatsiya modellari [8], hamda sentiment tahlili datasetlari [9] kiradi.

O'zbek tili matnlardan fe'l so'z turkumiga oid so'zlarni aniqlashda biz tub fe'llar bazasini tuzdik, chunki tub fe'l qo'shimchalar bilan yasalmaganligi uchun uni FSMlar yordamida aniqlab bo'lmaydi. Yasama fe'llarni aniqlash uchun biz yuklamalar, shaxs-son, zamon, ravishdosh, sifat-dosh va harakat nomi, egalik, kelishik, nisbat, so'z yasovchi qo'shimchalar uchun alohida-alohida chekli avtomat (FSM)larni qurib chiqdik.



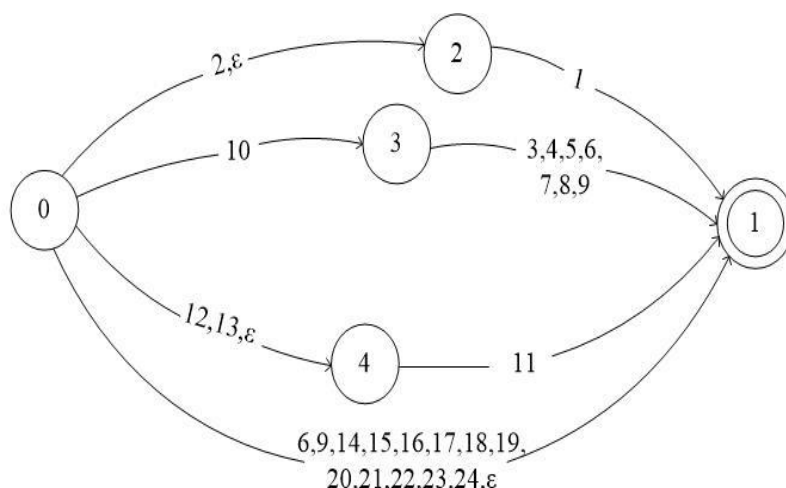
1-chizma. Fe'lni qo'shimchalar orqali aniqlash uchun FSM

Fe'lni qo'shimchalar orqali aniqlash 1-chizmada keltirilgan bo'lib bu yerda 9 ta Chekli avtomat (FSM)lar mavjud. Shulardan zamon (2)ning FSMni ko'rib chiqamiz.

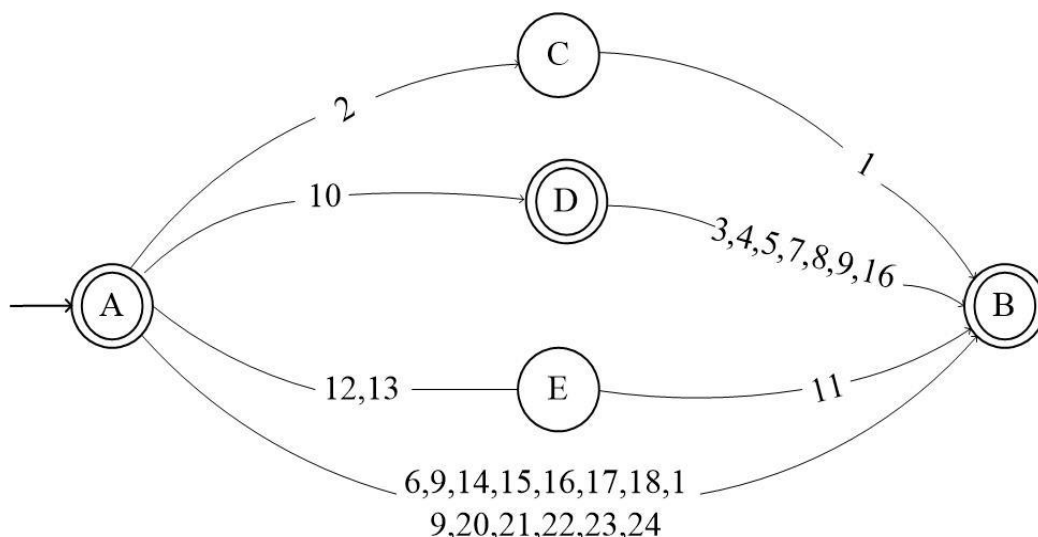


2-chizma: Zamon qo'shimchalarini qidirish FSM (chapdan o'ngga)

1. -yap	13. -mish
2. -ti	14. -moqda
3. -gan	15. -moqchi
4. -ayotgan	16. -gusi
5. -yotgan	17. -r
6. -ar	18. -ayotir
7. -a	19. -yotir
8. -y	20. -ajak
9. -ib	21. -jak
10. -di	22. -gay
11. -mi	23. -ayotib
12. -kan	24. -yotib



3-chizma: Zamon qo'shimchalarini qidirish deterministik bo'lmagan chekli avtomat (NFA)



4-chizma: Zamon qo'shimchalarini qidirish FSM (o'ngdan chapga)

Shunday istisnolarni aniqladikki, fe'l so'z turkumiga oid bo'lmagan so'zlar ham bu FSMda qo'shimchalarini topishi mumkin. Chunki fe'lning vazifadosh shakllaridan sifatdosh va harakat nomi ot so'z tukumiga hoslangan shakli bo'lganligi sababli otning sintaktik va lug'aviy qo'shimchalarini qabul qiladi. Misol uchun ot so'z turkumiga oid quyidagi *kitob-imiz(4)-ni(3)-mi(0)*, *mi*-qo'shimchasi 0-FSMda, *ni*- qo'shimchasi 3-FSMda, *imiz*- qo'shimchasi 4-FSMda topilganligi sababli bizning FSM bu so'zni fe'l deb chiqaradi. Quyida yana shunday istisnolar mavjud:

1. *qalam-imiz-ni-mi* (0,3,4)
2. *Xiva-dan-mi-siz-lar* (1,1,3,4)
3. *Xiva-dan-mi-siz-lar-a* (0,1,1,3,4)
4. *qizil-mi* (0)

5. *maktab-ga* (3)
6. *bir-ga-miz* (1,3)
7. *bir-ga-miz-a* (0,1,3)
8. *o'yin-i-ni* (3,4)

Soʻzlar faqat bitta *-miz*, *-ing*, *-imiz*, *-siz*, *-san*, *-man* qoʻshimchalari bilan tugagan boʻlsa u holda qoʻshimchani qirqib, bazaga murojat qilamiz.

Xulosa sifatida shuni aytish mumkinki, ushbu chekli avtomatlar bizga feʼl soʻz turkumiga oid soʻzlarni aniqlashda dastur tezligini oshirish uchun xizmat qiladi. Ushbu chizmalardan foydalangan holda Python dasturlash tilida dasturini tuzish maqsad qilindi.

Foydalanilgan adabiyotlar.

1. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.
2. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com
3. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.
4. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.
5. X. Madatov, M. Sharipov, and S. Bekchanov, “OʻZBEK TILI MATNLARIDAGI NOMUHIM SOʻZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021.
6. K. Madatov, S. Bekchanov, and J. Vičič, “Accuracy of the Uzbek stop words detection: a case study on "School corpus",” *arXiv preprint arXiv:2209.07053*, 2022.
7. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.
8. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” *arXiv preprint arXiv:2210.16006*, 2022.
9. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The



case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243

ЎЗБЕК ТИЛИ МАТНЛАРИНИ РАЗМЕТКАЛАШ

Турсунов Муҳаммадсолиҳ Саъдин ўғли

muhammadsolih927@gmail.com

СамДУ мустақил тадқиқотчиси

Аннотация. Ушбу мақолада ўзбек тили миллий корпуси (uzbekcorpora.uz)га матн юклаш ва сўзларни разметкалаш келтирилган.

Калит сўзлар: *MS Word, разметка, МББТ, лугат, автоматик разметка, грамматик ҳарактеристика.*

Abstract. This article describes how to upload text to the National Corpus of the Uzbek language (uzbekcorpora.uz) and organize words.

Key words: *MS Word, markup, MBBT, dictionary, automatic markup, grammatical characteristics.*

Абстрактный. В этой статье описывается, как загрузить текст в Национальный корпус узбекского языка (uzbekcorpora.uz) и упорядочить слова.

Ключевые слова: *MS Word, разметка, МББТ, словарь, автоматическая разметка, грамматические характеристики.*

uzbekcorpora.uz платформасида грамматик разметкалаш дастури *Nuxt JS, Python* ва *PostGreSql* маълумотлар базасини бошқариш тизимида (МББТ) тузилган. Матнларни корпусга киритиш жараёнини соддалаштириш учун грамматик разметкалаш икки босқичда бажарилади: матнни дастлабки форматлаш босқичи ва разметкалаш босқичи. Дастлабки форматлаш босқичида матннинг тузилмавий компонентлари аниқланади, яъни матнда сўзлар, гаплар, абзацлар ажратилади. Шунингдек, сўзларнинг матндаги ўрни аниқланади. Иккинчи босқичда грамматик разметкалаш амалга оширилади. Разметкалаш натижасида матндаги ҳар бир сўзшаклга ўзининг морфологик параметрлари ва сатрли атрибутлари тўплами бириктириб чиқилади. Қуйида бу иккала босқичнинг грамматик разметкалаш дастури орқали бажарилиши батафсил баён қилинади.

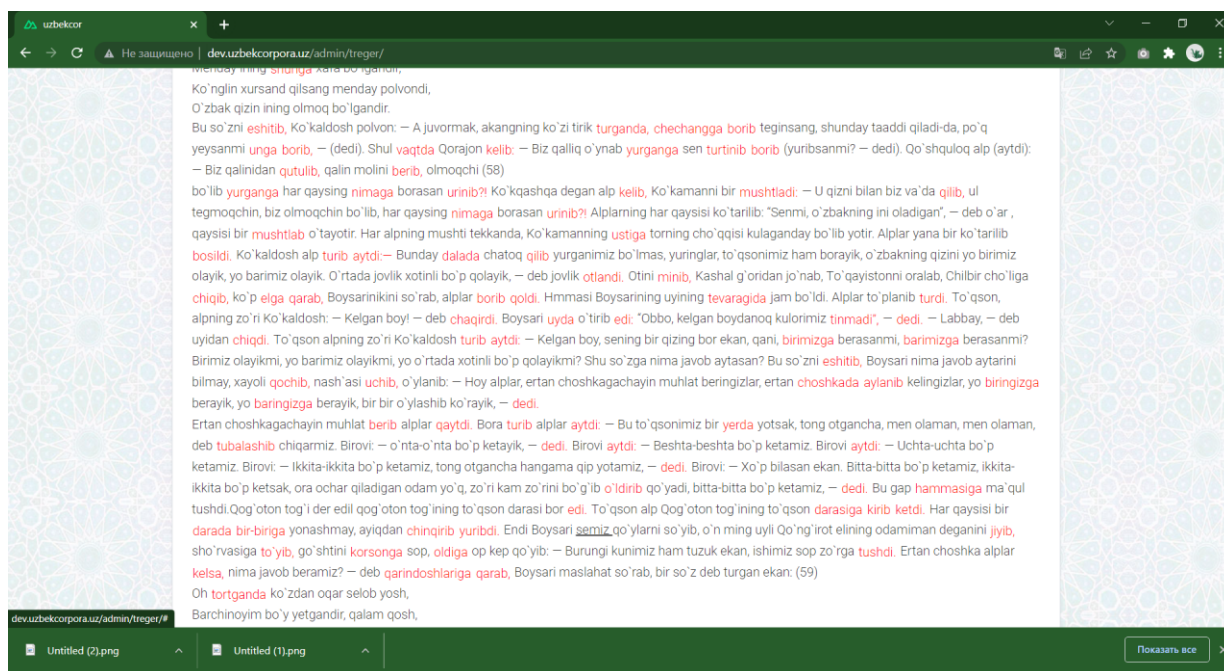
Дастлабки форматлаш босқичи. Бу босқичда корпусга киритиладиган файл дастлаб қандай шаклда бўлишидан қатъий назар, у *MS Word* дастури форматида ўтказилиши керак. *MS Word*нинг 2007 ва ундан юқори версияларини ишлатиш зарур, чунки уларда матн **.docx* форматда бўлиб, стандарт тизим асосида автоматик равишда тегланган ҳолда бўлади. Матнни тузилмавий компонентларга (сўзлар, гаплар, абзацлар) бўлиш *MS Word* дастури томонидан автоматик тарзда бажарилади. Матнда сўзларнинг ўрнини билдирувчи маълумотлар алоҳида файлга ёзилади.

Разметкалаш босқичи. Разметкалаш деб матнга ва унинг компонентларига махсус тегларни бириктириб чиқишни тушуниш керак. Махсус теглар икки хил бўлади: лингвистик теглар ва экстралингвистик (ташки) теглар. Лингвистик теглар матн элементларининг лексик, грамматик ва шунга ўхшаш бошқа хусусиятларини тавсифлайдиган маълумотлардан иборат. Экстралингвистик теглар эса муаллиф ҳақидаги ва матн ҳақидаги маълумотларни (муаллиф, номланиши, нашр йили ва жойи, жанр, тематика ва ҳақозо) тавсифлайди [Николаев И.С., 2016:320].

Корпуснинг улкан ҳажми билан мутаносиб равишда разметкалаш кўп вақт ва меҳнат талаб қиладиган жараёндир. Шунинг учун у автоматик тарзда бажарилиши зарур. Агарда разметкаланишнинг бирор бир тури ҳали автоматлаштирилмаган бўлса, у ҳолда бу ишни қўлда бажаришга тўғри келади [M.S.Tursunov, 2022].

Бу босқичда, дастлаб, автоматик разметкалаш дастури ишга туширилади.

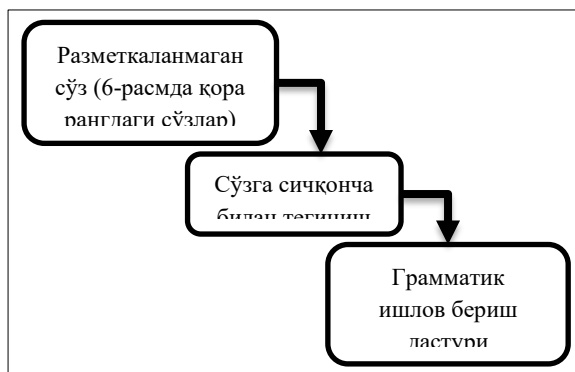
Натижасида матн *.docx форматдаги файл (1-расм) кўринишида тақдим этилиб, унда разметкаланган сўзлар қизил ва разметкаланмагани қора ранг ажратилган бўлади. (Грамматик луғат бўш бўлса, ҳамма сўзлар қора рангда чиқарилади.)



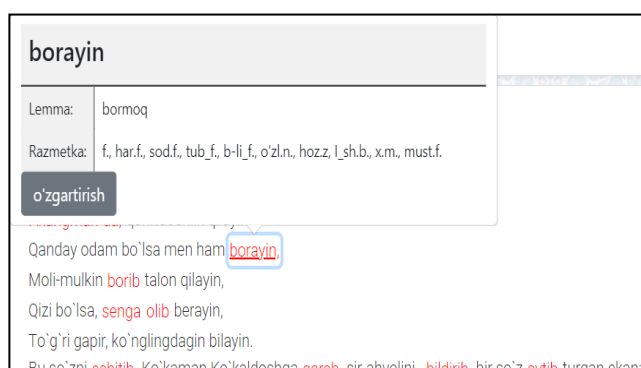
1-расм. Файлнинг дастлабки ҳолати

Бунинг учун, сўз устига курсор келтириб босилганда, сўзнинг морфологик характеристикалари ёзилган ойна экранда пайдо бўлади (2-расм). Бу ойнадаги маълумотлардан

а) сўз разметкаланмаган ҳолат

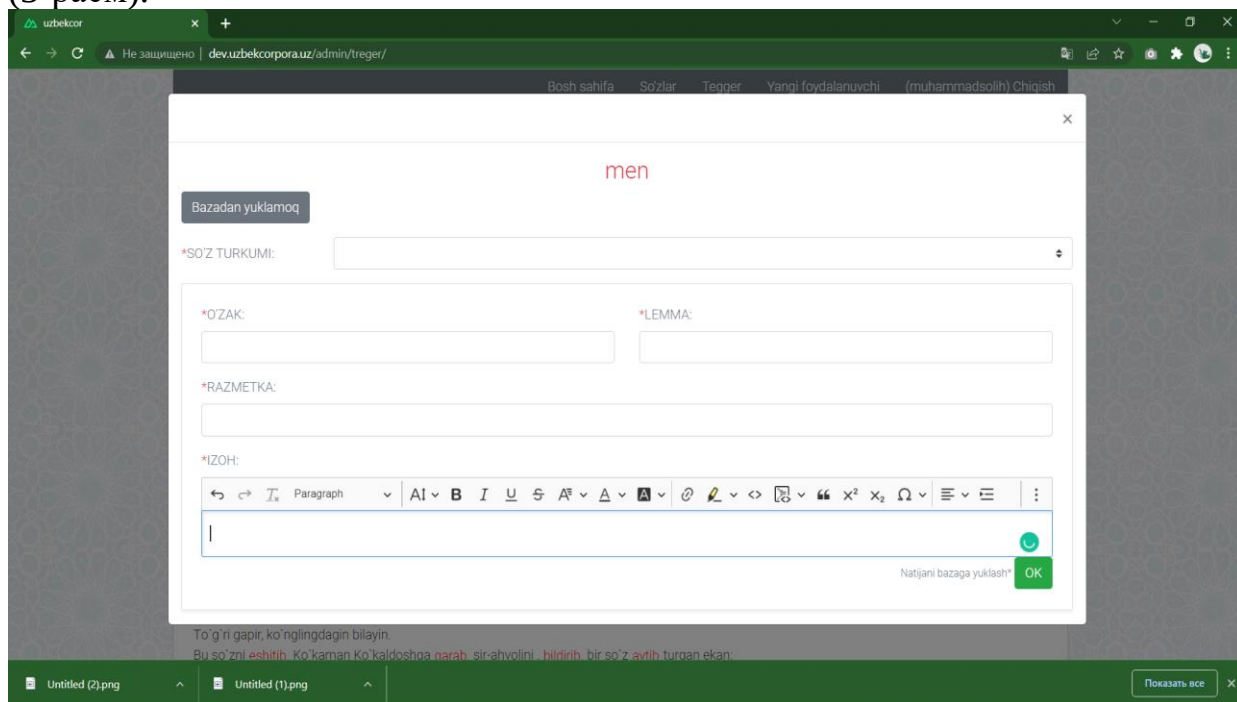


б) сўз разметкаланган ҳолат



2-расм. Сўзнинг грамматик теглари

Матндаги сўзларнинг ранглиридан фойдаланиб, сўз тегланган ёки тегланмагани аниқланади. Сўз ҳали разметкаланмаган бўлса, тегишли тугма босилади ва сўзнинг морфологик параметрларини белгилаш учун ойна очилади (3-расм).



3-расм. Сўзга морфологик характеристикалар бериш ойнаси

3-расмда сўзни морфологик разметкалаш мумкин бўлади. Бунда сўзга унинг туркуми, туркумга мос равишда грамматик белгилари, ўзаги, леммаси ва

изоҳ киритилади. Ойнадан туркумни танлаганимизда туркумга мос равишда грамматик белгилар рўйхати очилади. Рўйхатдан керакли белгилар танлаб олинади.

Сўзларни грамматик луғатга киритиш. Бу вазифани бажарувчи дастурий модул *Lug_Kirit* деб номланган. Луғатга янги бирлик киритиш фойдаланувчи томонидан интерактив режимда амалга оширилади. Бу жараёни *Lug_Kirit* дастури таъминлайди. Компьютер экранида *Suz_Rhati* дастури томонидан чиқарилган рўйхат мавжуд ва унда битта сўз бошқа ранг билан ажратиб кўрсатилган. *Lug_Kirit* дастури ёрдамида фойдаланувчи рўйхат бўйича ҳаракатланиб, жорий сўзни танлайди ва “Enter” тугмасини босади. Натижада экранда киритиш ойнаси пайдо бўлади. Фойдаланувчи бу ойнада сўралган маълумотларни клавиатурадан териб киритади ва “Базага киритиш” тугмасини босади. Шунда танланган сўз ва унга мос киритилган грамматик характеристикалар жамланиб, электрон луғат файли *LUGAT*га ва бошқа ёрдамчи файлларга ёзиб қўйилади. Шундан сўнг фойдаланувчи рўйхатдан навбатдаги сўзни танлайди ва базага киритади. Шу тариқа, берилган рўйхатдаги барча сўзлар *LUGAT* файлга киритилиб, грамматик луғат қадамма-қадам бойитиб борилади.

Теггер дастури матнларни корпусга киритиш учун фойдаланилади. Матнларни корпусга киритиш жараёнида, матнга грамматик ишлов берилади. Матндаги сўзларга грамматик маълумотлар яримавтомат тарзида тўлдирилади ва грамматик маълумотга эга бўлган сўзларнинг ранги қизил рангга ўзгаради ва фойдаланувчига разметкаланмаган сўзлар ҳамда разметкаланган сўзлар якқол ажралиб кўринади. Сўзни разметкалашнинг лингвистик таъминоти иловада (а илова) келтирилган. Бунини 1-расмда кўриш мумкин. Разметкаланмаган сўзларга грамматик маълумотлар бериш учун сўз устидан бир марта босилади ва сўзга грамматик ишлов бериш ойнаси ҳосил бўлади (8-расм). Сўз разметкалангандан сўнг сўзнинг ранги кўк рангга ўзгаради. Бу эса матн билан ишлаш вақтида разметкаси қўлда берилганлигини билдиради. Қизил рангдаги сўз устидан босилган вақтда унинг грамматик маълумотлари кўринади. Агар грамматик маълумотлар нотўғри берилган бўлса, 2-расмнинг б бандидаги ойнадан “O‘zgartirish” тугмаси босилади ва сўзга грамматик ишлов бериш учун 8-расмдаги ойна ҳосил бўлади. Матнга ишлов берилгандан сўнг матн қайта сақланади. Матнлар корпус таркибида *JSON* (б иловага қаранг) форматида сақланади.

Грамматик разметкалашни автоматлаштириш. Сўзларни грамматик таҳлил дастури учун махсус автоматик разметкалаш модули яратилган. Бу модул қўлда теглаш дастуридан олдин ишга туширилади. Унда ҳар бир сўз мавжуд грамматик луғат (МБнинг *Words* жадвалидан) таркибидан изланади ва агар у топилса, луғатда бор морфологик параметрлардан фойдаланиб, тег автоматик тарзда шакллантирилади ва сўзга бириктирилади. Грамматик луғатдан топилмаган сўзларнинг разметкаси бўшлиғича қолади. Матнда

такрорланувчи сўзлар учун луғатдан излаш фақат бир марта бажарилади ва натижа уларнинг ҳаммасига тадбиқ этилади.

Автоматик разметкаланининг ўзига хос камчиликлари ҳам мавжуд:

1) Матннинг грамматик луғатдан топилмаган сўзлари разметкасиз қолдирилади. Лекин, корпус яратиш жараёнида грамматик луғат кенгайиб бориши натижасида матннинг ундан топилмайдиган сўзлари улуши тобора камайиб боради. Шунинг учун, грамматик луғатни мумкин бўлган барча усулларни қўллаган ҳолда, мумкин қадар бойитиш зарур;

2) Грамматик луғатни кенгайтириб бориш натижасида омоним сўзлар муаммоси ҳам кўпроқ намоён бўла бошлайди. Бу муаммо шундан иборатки, луғатда бир сўз бир неча хил грамматик характеристикаларга эга бўлиб, улардан қайси бирини олиш керак деган савол пайдо бўлади. Муаммо сўзнинг контекстини таҳлил қилиш асосида ҳал қилинади.

Фойдаланилган адабиётлар:

1. Николаев И.С., Митренина О.В., Ландо Т.М., Прикладная и компьютерная лингвистика. Книга – Санкт Петербург: СПбГУ. 2016. – С. 320. - ISBN 978-5-9710-3472-8.
2. Kennedy G., An introduction to corpus linguistics. London: Longman, 1998.
3. M.S.Tursunov, Description of the management system programs of the national corpus of the Uzbek language // International journal of engineering mathematics, <https://iejemta.com/index.php/em/article/view/9/9>, volume 4 ISSUE 1, 2022, 32-42 p.
4. Elov B., Hamroyeva Sh., Abdullayeva O., Uzoqova M. O'zbek tilida pos tagging masalasi: muammo va takliflar. O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 45-63.

TEXT GENERATOR BASED ON LEXIS-SYNTACTIC TEMPLATES

Yodgorov Umidjon,

Senior Lecturer,

Tashkent State University of Uzbek Language and Literature,

yodgorov@navoiy-uni.uz

Abstract. In this paper, we consider a text generator developed using lexis-syntactic templates and compare it with a standard generator based on Markov chains.

Key words: *text generation, natural language processing, Markov chains, LSPL language.*

Абстракт. В данной работе мы рассматриваем генератор текста, разработанный с использованием лексико-синтаксических шаблонов, и сравниваем его со стандартным генератором на основе цепей Маркова.

Ключевые слова: *генерация текста, обработка естественного языка, цепи Маркова, язык LSPL.*

Annotatsiya. Ushbu maqolada leksik-sintaksik shablonlardan foydalangan holda ishlab chiqilgan matn generatori tadqiq etilgan hamda Markov zanjirlariga asoslangan standart matn generatori bilan solishtiriladi.

Kalit soʻzlar: *matn generatori, tabiiy tilni qayta ishlash, Markov zanjirlari, LSPL tili.*

Automatic text generation is one of the most common natural language processing tasks. Text generators are usually used in cases where there are no strict requirements for the quality of the text, and writing it manually takes a lot of time or is an uninteresting task - for example, compiling text summaries of data or writing texts for search engine optimization of web pages.

Among the algorithms for creating texts, two main approaches can be distinguished: methods based on rules allow achieving high quality texts, but require knowledge of the rules of the language and are time-consuming to develop, while methods based on machine learning depend only on training data, but more often they make grammatical and semantic errors in the created texts.

The most common approach to the task of automatic text generation is the use of Markov chains - a probabilistic model that does not depend on the rules of the language and does not require linguistic knowledge. The Markov model refers to machine learning methods.

Rule-based approaches to automatic text generation use lexis-syntactic patterns. This method has been studied in relation to texts in English, the most frequently analyzed in computational linguistics publications. In the case of Uzbek-language texts, lexis-syntactic templates are already used for tasks such as extracting

named entities, but not for creating texts: no works describing the use of templates in Russian text generators have been published at the moment.

In this paper, we consider a text generator developed using lexis-syntactic templates and compare it with a standard generator based on Markov chains.

The texts available on the websites online.navoiy-uni.uz and tsuull.uz were chosen as the material for the text generator.

2. Review of the work on the research topic

Since automatic text generation algorithms in English are widely used in various software systems, there are a significant number of English-language publications describing such algorithms. Here is a brief overview of the articles that is most relevant to the topic.

The most common language-independent approach to automatic text generation is Markov chains. An algorithm based on their application is presented in Sam Zhang [1]. The work carried out by the author includes the implementation of the English text generator in the Python programming language. Various collections of texts in English, including the WordNet corpus, were used as a training sample for the model. To improve the accuracy of the algorithm, the author uses a semantic similarity model in addition to the probabilistic approach.

Another popular machine learning algorithm used to create texts in natural language is the use of conditional random fields. An example of using this approach in the development of an English text generator is presented in Wei Lu et al. [3]. According to the authors, the use of conditional random fields makes it possible to efficiently generate natural language sentences based on their semantic representations. The algorithm proposed in the article consists of two stages. The first stage includes semantic analysis, which allows each sentence to automatically match its semantic representation. At the second stage, the text itself is created based on the obtained semantic representations.

The concept of a lexis-syntactic template as applied to texts in Russian is presented in the work of E.I. Bolshakova and others [2]. The authors give the definition of a lexis-syntactic template as a structural sample of a language construct that reflects its properties, and present their development - the LSPL language, which allows describing text using templates.

3. The purpose and objectives of the work

The purpose of the work is to implement a text generator based on lexis-syntactic patterns and compare it with a standard generator based on Markov chains.

The following tasks can be distinguished in the work on the project:

1. Collect a corpus of texts from online.navoiy-uni.uz and tsuull.uz
2. Develop templates for the main algorithm
3. Implement a template-based feedback generator
4. Implement a feedback generator using Markov chains and train it on the

collected data

5. Compare the results of the two generators

In this project, a rule-based method of automatic creation of texts is used. Lexis-syntactic patterns of texts are set manually and are used by the algorithm to determine the structure of generated sentences. To substitute words into sentences, the algorithm uses a dictionary formed on the basis of a corpus of real texts.

In the modern world, the amount of information is constantly growing, which leads to an increase in demand for automatic means of processing it. Automated text generation software is designed to automate the repetitive and uncreative process of writing supporting text such as summaries, web page indexing by search engines, and so on. However, natural language is difficult to formalize, and algorithms may experience inaccuracies, as automatically generated texts are inferior in quality compared to those created by a person. Research in the field of automatic text generation, aimed at developing new algorithms and improving existing ones, makes it possible to minimize the number of such inaccuracies and to generate texts that fully comply with the rules of the language and do not contain semantic errors.

4. Markov chains

A Markov chain with discrete time is a stochastic system that can be represented as a finite automaton - a set of states and transitions between them. Each of the transitions are assigned a probability value in such a way that for each of the states the sum of the transition probabilities is equal to one. An important property of a Markov chain is the fact that the distribution of transition probabilities from each state depends only on the state itself, but not on those preceding it. Thus, the Markov chain models a random process "without memory".

Markov chains can be used as a probabilistic natural language model. [1] The states of a finite automaton can be either individual words or N-grams - sequences of N words in a row in the text. The application of this model to the tasks of automatic text creation belongs to machine learning methods.

The process of creating text using Markov chains consists of two main stages. At the first stage, the model is trained using a corpus, a collection of texts. The learning process includes the calculation of transition probabilities: for this, the frequency of occurrence of consecutive pairs (or (N+1)-grams) of words in corpus texts is calculated. At the second stage of text creation, the initial state of the constructed finite automaton is selected and then, using a pseudo-random number generator, transitions are made to the next states, each of which adds the next N-gram to the generated text. The transition process is repeated until the desired text length is reached. The accuracy of the algorithm depends on the chosen value of N and on the number of texts in the corpus.

To illustrate that, let's give an example of a Markov chain scheme built on the basis of the input text. A good example is presented in the article [5], where a fragment of the text of song by The Smiths “How Soon Is Now” was used to build

the chain:

"I am the son
and the heir
of a shyness that is criminally vulgar
I am the son and heir
of nothing in particular".

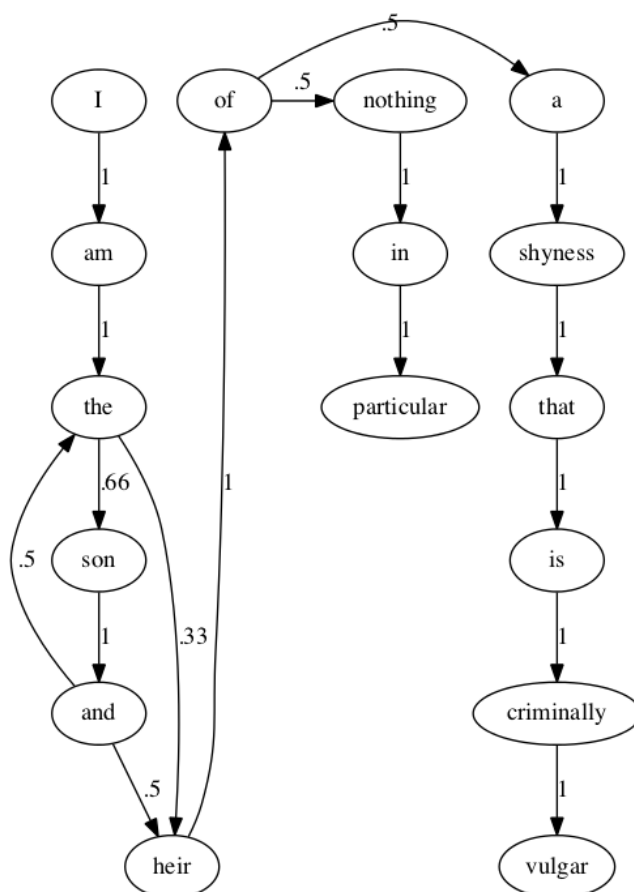


Рис. 1. Пример цепи Маркова [5]

For other languages, the chain is formed in a similar way: it is not the meaning of the words in the text that are important, but their order and frequency.

5. Lexis-syntactic patterns

Lexis-syntactic templates allow you to specify language constructs. A template consists of a logical structure and a semantic description. The use of lexis-syntactic patterns in the task of creating texts in natural language refers to rule-based methods and requires linguistic knowledge. [4, 2] In this paper, the lexis-syntactic template is a sentence scheme consisting of word types. An example of a template that is used in the algorithm presented in this work: a scheme of the form "property

person verb", which corresponds to the fragment of the review "I liked the iron".

The algorithm developed as part of the project is based on a rule-based approach to text processing, combined with the elements of machine learning: sentence templates are built based on manually written rules, and data is used to match them with words in the text creation process. extracted from real texts.

Steps of the algorithm:

0. Lexis-syntactic sentence templates are created manually
1. A list of model characteristics is compiled
2. Verbs and adjectives are extracted from real texts with characteristics
3. Text is formed: words are selected according to templates and using the data obtained in steps 1-2
4. The resulting text is processed in such a way that all neighboring word forms become grammatically consistent with each other

6.Comparison of generators

Generator properties	Template based generator	Generator based on Markov chain
Breaking grammar rules	In a sentence, it is possible only if there are errors in the implementation. With a sufficiently complete analysis of cases of agreement, individual sentences in the output are grammatically correct. But the coordination of proposals among themselves in this version is not provided	Perhaps due to the ambiguity of interpretations of many language constructions
Dependence on the size and variety of input data	Expressed for vocabulary, but not for grammar. The structure of sentences is defined by templates, and vocabulary of the generator is replenished when training on the input data	Expressed for both vocabulary and grammar: the only source for constructing a Markov chain is the input corpus of texts

mismatch semantic compatibility	of	Found in early versions of the generator (examples: "long iron", "ceramic instruction"). In the current version, agreement check protects against such situations: phrases encountered in the training corpus entered into the set of valid ones, and when generating text, it is checked whether the current phrase is in this set. A side effect of validation is the fact that it increases dependence on input data and reduces the variety of generated texts.	May occur within a sentence. Example: “The old iron suddenly broke down and I am grateful to him for a sensible offer.”
Variety language constructs	of	Low in the current version of the generator. By adding new sentence schemes to the template system, you can increase the variety by bringing texts closer in quality to real reviews. This is one of the possible tasks for further work.	High with a sufficient size of the training sample

7. Conclusion

Based on the results of the work carried out, the following conclusions can be illustrated:

1. Text generator based on lexical and syntactic patterns allows you to get grammatically correct sentences similar to fragments of real texts
2. There is no agreement between individual sentences in the generated texts, therefore, before applying to the creation of voluminous connected texts, the algorithm needs to be improved
3. In the tasks of creating texts for subsequent automatic processing (such as search engine optimization), the proposed algorithm has the same capabilities as the standard generator based on Markov chains
4. The number of grammatical and semantic errors in texts generated using



lexical and syntactic patterns is less than in texts generated using Markov chains

References

1. Sam Zhang. 2009. Natural Language Generation with Markov Chains and Grammar. Computer Systems Lab, 2009-2010
2. Большакова Е. И. и др. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов //Труды межд. конф. Диалог. – 2007. – С. 70-75.
3. Wei Lu, Hwee Tou Ng, Wee Sun Lee. 2009. Natural Language Generation with Tree Conditional Random Fields. Singapore-MIT Alliance. Department of Computer Science, National University of Singapore.
4. Kees Van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Real versus Template-Based Natural Language Generation: A False Opposition?. Comput. Linguist. 31, 1 (March 2005), 15-24.
5. How to fake a sophisticated knowledge of wine with Markov Chains. // URL: <http://www.onthelambda.com/2014/02/20/how-to-fake-a-sophisticated-knowledge-of-wine-with-markov-chains/>

RESTORAN SOHASIDAGI O'ZBEK TILIDAGI MATNLARNING SENTIMENT TAHLILI

Niyazmetova Kumushoy Ergash qizi

UrDU magistranti,

kumushoy.n@urdu.uz

Quriyozov Elmurod Rajabboy o'g'li

UrDU o'qituvchisi,

elmurod1202@urdu.uz

Annotatsiya. Foydalanuvchilar tomonidan yozilgan ko'plab sharhlarni sentimental tahlil qilish va tasniflash muammolarini yechish uchun foydali ma'lumotlarni olish tabiiy tilni qayta ishlashning muhim vazifasidir. Bu nafaqat mijozlar ehtiyojini qondirish uchun, balki kompaniyaning keyingi rivojlanishiga ham ta'sir qilishi mumkin. Ushbu maqolada biz Google xaritasidagi Toshkent shahrida joylashgan restoranlarga bildirilgan fikr-mulohazalaridan foydalanib Dataset tayyorlandi va logistic regressiyaga asoslangan modellaridan foydalanilib Sentiment tahlil qilindi. Umumiy baholash natijalari shuni ko'rsatadiki, agglyutinativ tillar uchun stemming kabi dastlabki ishlov berish bosqichlarini bajarish orqali tizim yaxshi natijalar beradi va natijada eng yaxshi ishlaydigan modelda 91% aniqlikka erishiladi.

Kalit so'zlar: *NLP, Sentiment tahlil, Dataset, Bag of word modeli, TF-IDF algoritmi.*

Abstract. Sentiment analysis of many comments written by users and extracting useful information for solving classification problems is an important task of natural language processing (NLP). This can affect not only customer satisfaction, but also the further development of the company. In this article, we have prepared a Dataset using feedback given to restaurants located in the city of Tashkent on the Google map and analyzed Sentiment using Logistic Regression models. Overall evaluation results show that the system performs well by performing pre-processing steps such as stemming for agglutinative languages, resulting in 91% accuracy in the best performing model.

Keywords: *NLP, Sentiment Analysis, Dataset, Bag of word model, TF-IDF algorithm.*

Аннотация. Анализ тональности многих комментариев, написанных пользователями, и извлечение полезной информации для решения задач классификации является важной задачей обработки естественного языка (NLP). Это может повлиять не только на удовлетворенность клиентов, но и на дальнейшее развитие компании. В этой статье мы подготовили набор данных, используя отзывы, полученные от ресторанов, расположенных в городе Ташкент, на карте Google, и проанализировали настроения с использованием

моделей логистической регрессии. Общие результаты оценки показывают, что система хорошо работает, выполняя этапы предварительной обработки, такие как выделение корней для агглютинативных языков, что обеспечивает точность 91% в самой эффективной модели.

Ключевые слова: *NLP, анализ тональности, набор данных, модель «мешок слов», алгоритм TF-IDF.*

Kirish.

Tabiiy tilni qayta ishlash (NLP) usullarining samaradorligi ko'plab ilovalarda belgilangan ma'lumotlar miqdoriga bog'liq. Sentiment tahlil qilish - bu iste'molchilar tomonidan bildirilgan fikrlarni tahlil qilish. Iste'molchilar odatda joylar/ovqatlar haqidagi fikr-mulohazalarini Google Map, Yelp va boshqalar kabi mashhur ilovalarga joylashtiradilar. Ular ko'pincha iste'molchilarni sharhlarda faol ishtirok etishga undaydi va foydalanuvchi tomonidan yaratilgan fikr-mulohazalar asosida iste'molchilarga o'z ehtiyojlarini to'liq qondirish imkonini beradi [1]. Tadbirkorlarga esa real vaqt rejimida shaxsiylashtirilgan xizmatlarni taqdim etishda yordam beradi.

Bundan tashqari, restoran sharhlari mijozlarning hissiy ehtiyojlari tarkibini ifodalaydi va iste'molchilar tanlovi haqida muhim ma'lumot manbayi hisoblanadi. Hozirgi vaqtda fikr-mulohaza yuritish, ayniqsa, yuqori manba tillari uchun chuqur o'rganish usullarini qo'llaganidan keyin juda yuqori aniqlik ko'rsatkichlariga erishdi [2]. O'zbek tili agglyutinativ til bo'lib, unda bir so'z ma'noli gap bo'la oladi. Bizning ma'lumotlarimizga ko'ra, restoran domenidagi fikr-mulohazalarga asoslangan hissiyotlarni tasniflash muammolari bo'yicha oldingi ishlar yetarli emas [3]. Shunday qilib, ushbu maqola uchun quyidagilarni hisobga olinadi:

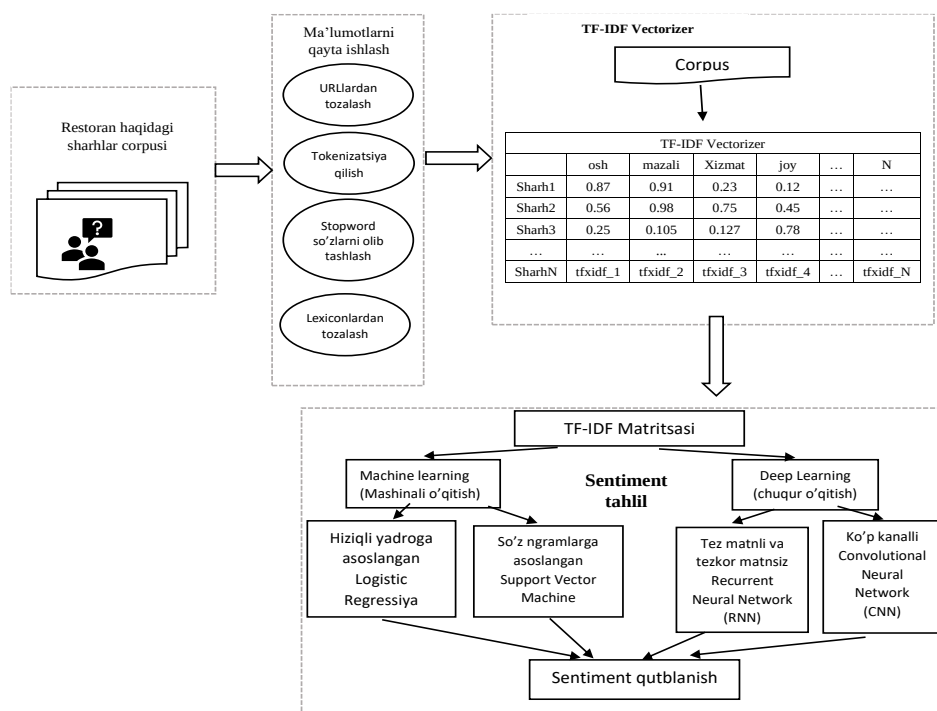
- Restoran domenining izohli korpusi his-tuyg'ularni tahlil qilish uchun yaratilgan bo'lib, u Google Xaritalardan o'zbek oshxonasi joylashuvi asosida to'planadi, bu yerda mahalliy milliy taomlar sharhlari asosiy maqsad hisoblanadi. Dataset katta xatolarni qo'lda olib tashlash va tozalashdan keyin 4500 ijobiy va 3710 salbiy sharhlarni o'z ichiga oladi. Annotatsiya jarayoni Google Xaritalar tomonidan ajratilgan 5 yulduzli fikr-mulohaza usuliga asoslanadi, bunda 1 dan 3 gacha ma'lumotlar to'plamini salbiy va 4 dan 5 gacha ijobiy deb hisoblaymiz. Biz ba'zi sharhlar ingliz, qirg'iz va rus kabi boshqa tillarga asoslanganligini aniqladik. Biz ularni e'tiborsiz qoldirishni istamadik, shuning uchun rasmiy Google Translate API yordamida ularni o'zbek tiliga tarjima qilishga qaror qildik.

- Datasetni oldindan qayta ishlash ikki bosqichda qo'llaniladi. Birinchi qadamlar URL manzillarini, tinish belgilarini va kichik harflarni olib tashlashdir. Ikkinchi qadam TF-IDF algoritmidan foydalangan holda to'xtash so'zlar ro'yxatini yaratgandan so'ng aniqlikni baholashga asoslangan ma'lumotlar to'plamidagi to'xtatuvchi so'zlarni e'tiborsiz qoldirish [4]; Keyin, biz stemming algoritmini qo'lladik.

O'zbek tilidagi so'zlarni elektron lug'ati asosida o'zbek tili nutqining bo'lagi: ot, sifat, son, fe'l, kesim, mayl, undoshlar uchun xulosa chiqarish kombinatsion yondashuvdan foydalaniladi. Algoritmni qo'llashning afzalliklari leksikadan xoli va uning murakkabligi bitta operatsiyani (tilning oxirlari lug'atiga ishora) bajarishga imkon beradi:

- so'zni qo'shimchalarga bo'lish;
- so'zlarni morfologik tahlil qilish.

So'nggi yillarda o'zbek tili uchun NLP sohasida bir qancha ishlar amalga oshirildi, jumladan Google Play ilovasi sharhlarini to'plash va tahlil qilish natijasida yaratilgan hissiyotlarni tahlil qilish ma'lumotlar to'plami, ikki turdagi ma'lumotlar: o'rta o'lchamli qo'lda izohlangan ma'lumotlar to'plami va katta hajmdagi ma'lumotlar to'plami ingliz tilidan avtomatik ravishda tarjima qilinadi [5]. Yana bir shunga o'xshash maqola o'zbek matnlarini fikr tasniflashda yangilik matnining kategoriyasi asosidagi xususiyatlarning ta'sirini o'rganilgan [6]. Semantik baholash bo'yicha ma'lumotlar to'plami so'z juftliklarida semantik o'xshashlik va bog'liqlik ballari, shuningdek, uning o'zbek tili uchun tahlili so'nggi ishda taqdim etilgan. NLP-da sun'iy intellektga asoslangan usullardan foydalanadigan o'sib borayotgan tendensiya mavjud, buni o'zbek tilidagi neyron transformatorlari - arxitekturaga asoslangan til modeli bilan o'zbek tilidagi ishda ko'rish mumkin [7].



1-rasm.
Tadqiqot
doirasi

Tuyg'ularni tahlil qilish sohasiga global nuqtayi nazardan qarashlar va fikrlardagi farqlarni hisobga olish g'oyasi bilan o'z ishlarida hissiyotlarni tahlil qilishning turli usullaridan, masalan, mashinani o'rganish va chuqur o'rganish usullaridan foydalangan ishlar mavjud. Twitter, Reddit, Tumblr va Facebook kabi mashhur ijtimoiy platformalarda mavjud bo'lgan fikrlardan iborat.

Metodologiya.

Ushbu maqolada biz restoran domeni ma'lumotlar to'plami uchun mashinali o'rganish va chuqur o'rganishga asoslangan hissiyotlarni tahlil qilish tizimini taklif qildik (1-rasm) [8]. Ramka veb-brauzer yordamida ma'lumotlarni yig'ish, oldindan ishlov berish (tozalash, to'xtatuvchi so'zlar, leksikonsiz stemming) [3], TF-IDF vazn matritsasi yaratish [9], [10], hissiyotlarni tahlil qilish uchun ML va DLni amalga oshirishni o'z ichiga oladi [11].



2-rasm: Google platformasida restoran uchun fikr qoldirish interfeysi namunasi

Ma'lumotlarni yig'ish.

Biz o'zbek tilida skanerlash uchun mavjud bo'lgan ko'p sonli ma'lumotlar to'plamini ko'rib chiqishdan boshlaymiz. Biroq Twitter yoki kino sharhlari kabi odatiy yondashuvlar o'zbek tiliga to'g'ri kelmaydi. Shu sababli, biz restoran sharhlarini to'plashga qaror qildik, chunki mahalliy aholi asosan restoranlar haqida fikr-mulohaza bildirishni yaxshi ko'rishadi. Bizning fikrimizcha, bu mantiqan to'g'ri, chunki o'zbek taomlari Mustaqil Davlatlar Hamdo'stligi (MDH, CA mamlakatlari) bo'ylab eng mashhur taomlardan biri hisoblanadi. Masalan, Kaliforniyaning aksariyat shaharlarida o'zbek oshxonasiga ixtisoslashgan band bo'lgan restoranlarni topish oson. Biz Toshkentdagi barcha mahalliy restoranlarni Google Xaritalardan ko'rib chiqdik. Birinchidan, biz kamida 3 ta ko'rib chiqishga ega bo'lgan 140 dan ortiq URL manzillar ro'yxatini tanladik va biz 2-rasmda ko'rsatilgan barcha ma'lumotlarni oldik. Ko'rish chog'ida Googlening spanga qarshi va DDOSga qarshi siyosatlarini ko'rib chiqdik, chunki ma'lumotlarni yig'ishda ma'lum cheklovlar mavjud.

Ma'lumotlarni oldindan qayta ishlash.

Ma'lumotlar to'plamini tekshirishda yulduzcha bilan baholangan matnlar to'plami qo'lda tuzatish talab qilingan. Faqat kulgichlar, ismlar yoki foydalanuvchi nomi eslatmalari, URL manzillari yoki maxsus ilova nomlari kabi boshqa ahamiyatsiz kontentdan iborat sharhlar olib tashlandi. O'zbek tilidan boshqa tillarda (asosan rus va ba'zilar ingliz tilida) yozilganlar rasmiy Google translate API

yordamida tarjima qilingan. O'zbekistonda odamlar rasmiy lotin alifbosidan foydalanishda-da, eski kirill alifbosidan foydalanish, ayniqsa, kattalar orasida bir xil darajada mashhur [12]. O'sha izohlar kirill alifbosida yozilganlar o'zbek mashinasi transliteratsiyasi vositasi yordamida lotinga o'tkazildi. Keyin, muhim ma'lumotlarga e'tibor qaratish uchun sharhlarimizdan past darajadagi ma'lumot so'zlarini olib tashlash uchun to'xtash so'zlarini qo'lladik. Modelimiz TF-IDF (Term chastotasi - teskari hujjat chastotasi) yordamida bitta so'zni to'xtatuvchi so'zlar to'plamini avtomatik aniqlashning tavsiya etilgan algoritmidir. Shundan so'ng, har bir so'z prefiks va qo'shimchalar tufayli so'z hajmini kamaytirish algoritmi leksikadan xoli bo'lgan asos yaratish vositasiga qayta ishlanadi. Asosiy g'oya - mos keladigan so'zlarning kombinatsion yondashuvidan foydalanish.

Baholash. To'plangan yangi ma'lumotlar to'plami mos ravishda 8×2 nisbatda baholash uchun *train-set* va *test-set* to'plamlariga bo'linadi. Trainset modelni qurish va o'qitish uchun, testset esa qurilgan modelni testlab ko'rish uchun kerak. Ma'lumotni tozalash jarayonidan so'ng bizda asl ma'lumotlar to'plami quyidagicha bo'ladi:

bu yerda $\vec{x}_i$ xususiyat vektorlarini va $\vec{y}_i$ izohli teglarni ifodalaydi:

$$(\vec{x}_i, y_i), \quad i = 1, 2, 3, \dots, N \quad (1)$$

$$\vec{x}_i = (x_{i1}, x_{i2}, \dots, x_{im}) \quad i = 1, 2, 3, \dots, N \quad (2)$$

N va m mos ravishda ko'rib chiqishlar soni va xususiyat vektorining uzunligiga teng. Keyin biz har bir xususiyat vektori $\vec{x}_i$ uchun TF-IDF ballarini hisoblaymiz, bu esa so'zlarni olish orqali vektorlashtiradi. Berilgan sharhdagi so'zning chastotasini va sharhlar orasidagi chastotani hisobga olish.

Barcha $\vec{z}_i$ larning yakuniy natijasi siyrak matritsa sifatida aniqlanadi.

$$\vec{z}_i = \text{TF}(x_i) \times \text{IDF}(x_i) \quad i = 1, 2, 3, \dots, N \quad (3)$$

Mashinani o'rganish algoritmlari. Logistik regressiya modeli

$$h(\vec{z}) = 1/(1 + \exp(-z))$$

$$P(y | \vec{z}) = \begin{cases} h(\vec{z}), & \text{if } y = +1(\text{ijobiy}) \\ 1 - h(\vec{z}), & \text{if } y = -1(\text{salbiy}) \end{cases} \quad (4)$$

Logistik regressiya modeli – bu tasniflash algoritmi bo'lib, o'zining eksponentsial va log chiziqli funksiyalari bilan mashhur. U diskret qiymatlar bilan ishlaydi va har qanday real qiymat funksiyasi 0 va 1 ni ko'rsatadi. Hissiyotni tahlil qilish shuni ko'rsatadiki, sharhlar (4) formula yordamida ijobiy yoki salbiy bo'ladi.

Natijalar va muhokamalar.

Ushbu bo'limda to'plangan hissiyotlarni tahlil qilish ma'lumotlar to'plamiga qo'llaniladigan mashinani o'rganish va chuqur o'rganish usullaridan foydalangan holda baholash jarayonida olingan natijalarning batafsil tavsifi keltirilgan.

Tajriba natijalari

Madel nomlari	Sentiment	Accuracy %	Recall %	F1 %	Precision %
Logistik regressiya (so'zlarga asoslangan n-gramm)	Ijobiy	88	98	93	89
	Salbiy	88	67	74	
Logistik regressiya (belgilarga asoslangan n-gramm)	Ijobiy	87	51	92	87
	Salbiy	83	97	64	
Logistik regressiya (so'z va belgilarga asoslangan n-gramm)	Ijobiy	95	95	92	91
	Salbiy	90	89	90	

1-jadval

Yuqorida qayd etilgan baholashning umumiy eksperiment natijalari amalga oshirildi va natijalarni 1-jadvalda ko'rish mumkin.

Xulosa. Ushbu maqolada biz o'zbek tili uchun restoran domenidagi yangi ma'lumotlar to'plamini ko'rsatdik, 8210 ta sharh, ijobiy yoki salbiy yorliqlar bilan izohlangan, u Google Xaritalardan poytaxt Toshkentdagi barcha restoran profillarining URL manzillaridan foydalangan holda ko'rib chiqilgan va ularga mos yulduz reytingi sifatida belgilangan. Keyin biz asosiy modellarimiz aniqligini oshirish uchun ma'lumotlar to'plamiga oldindan ishlov berishning to'liq bosqichlarini qo'lladik. Ma'lumotlar to'plamidagi eng yaxshi aniqlik natijasi (91%) so'z va belgilar n-grammlari bilan logistik regressiya modeli yordamida olingan. Yaqin kelajakda biz restoran sharhlarini amaliy darajada samarali tahlil qila oladigan ko'proq ma'lumotlarni to'plash orqali ishni kengaytirishni rejalashtirmoqdamiz. Shuningdek, ma'lumotlarni bo'lishda o'zaro tekshirish usullarini qo'llash orqali o'quv eksperimentlarini baholashning noto'g'riligini bartaraf etish bo'yicha ishlar olib borilmoqda.

Foydalanilgan adabiyotlar:

1. B. Pang, L. Lee, and others, "Opinion mining and sentiment analysis," *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
2. L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 8, no. 4, p. e1253, 2018.
3. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, "SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language," in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on*

Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 2022, pp. 199–206. [Online]. Available: www.scopus.com

4. X. Madatov, M. Sharipov, and S. Bekchanov, “O`ZBEK TILI MATNLARIDAGI NOMUHIM SO`ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021.

5. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers, 2022*, pp. 232–243.

6. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.

7. J. Mattiev, U. Salaev, and B. Kavsek, “Word Game Modeling Using Character-Level N-Gram and Statistics,” *Mathematics*, vol. 11, no. 6, p. 1380, 2023.

8. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings, 2022*, pp. 126–136. [Online]. Available: www.scopus.com

9. K. Madatov, S. Bekchanov, and J. Vičič, “Uzbek text summarization based on TF-IDF,” *arXiv preprint arXiv:2303.00461*, 2023.

10. K. Madatov, S. Matlatipov, and M. Aripov, “Uzbek text’s correspondence with the educational potential of pupils: a case study of the School corpus,” *arXiv preprint arXiv:2303.00465*, 2023.

11. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” *arXiv preprint arXiv:2205.10078*, 2022.

12. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.

MATNLARNING O'XSHASHLIGINI TAHLIL QILISH

Bozorov Obid

Mirzo Ulug'bek nomidagi O'zbekiston

Milliy universiteti

bozorov.obid@gmail.com

Annotatsiya. Ushbu ishda harflar chastotalarini matnlar tahlilida qo'llanilishi tadqiq qilingan. Chastotali tahlil usuli yordamida o'zbek tilidagi matnlarni moslik indeksi aniqlangan. Shuningdek, ochiq matnlarning moslik indeksini aniqlashning algoritmi taklif etilgan.

Kalit so'zlar: *chastotali tahlil, harflar chastotasi, matematik model, moslik indeksi, o'zaro moslik, tabiiy til.*

Abstract. In this work, the use of letter frequencies in text analysis is studied. Using the method of frequency analysis, the compatibility index of Uzbek texts was determined. Also, an algorithm for determining the compatibility index of open texts is proposed compatibility index of open texts has also been developed and implemented programmatically.

Key words: *frequency analysis, letter frequency, mathematical model, compatibility index, interchangeability, natural language.*

Аннотация. В данной работе изучается использование частот букв в анализе текста. Методом частотного анализа определен индекс сочетаемости узбекских текстов. Также предложен алгоритм определения индекса совместимости открытых текстов. Индекс совместимости открытых текстов также разработан и реализован программно.

Ключевые слова: *частотный анализ, частотность букв, математическая модель, индекс совместимости, взаимозаменяемость, естественный язык.*

Hozirda axborot himoyasi tizimlarini ishlab chiqishga va ularni takomillashtirishga alohida e'tibor qaratilmoqda. Tashkilotlarning axborot tizimlarida davlat sirlari yoki qonun bilan qo'riqlanadigan boshqa sirlarni tashkil etuvchi ma'lumotlar, shuningdek shaxsga doir ma'lumotlar muhofaza qilinishi, ulardan ruxsatsiz foydalanishning oldini olinishini ta'minlash hamda konfidensial ma'lumotlarning noqonuniy chiqib ketmasligi yuzasidan O'zbekiston Respublikasi qonun va hujjatlarida zarur dasturiy va tashkiliy-texnik choralar ko'rilishi shart qilib belgilangan. Shularni hisobga olib, davlat tashkilotlarida konfidensial axborotlarni ruxsatsiz chiqib ketishini bashoratlash va bartaraf qilishni hozirgi kunning dolzarb muammolaridan biri sifatida e'tirof etish mumkin [1-2].

Harflar chastotalari va ularning matnlar tahlilida qo'llanilishi. Matndagi harflar va so'zlarning chastotasi matn muallifi va matnning tegishli sohasi bilan bir-

biridan farqlanadi. Shuningdek, mualliflik ishlarini aniqlashda muallifning yozish usullari tahlili yordamida harf, bigram, trigram, soʻz chastotalari, soʻz uzunligi va gap uzunliklarini hisoblay olishlari mumkin va bu usul bilan matnlarning mualliflik huquqini himoyalash uchun ham ishlatishlari mumkin boʻladi. Oʻrtacha harflar chastotasini koʻplab belgili matnlar tahlili orqali hisoblash mumkin. Zamonaviy hisoblash tizimlaridan foydalanish orqali katta hajmli matnlar toʻplamlarini tahlil qilgan holda kerakli aniqlikdagi natijani olish qiyin hisoblanmaydi.

Faraz qilaylik, $A=(a_1, a_2, \dots, a_m)$ alifbo belgilari Z_m halqadan olingan $0, 1, 2, \dots, m-1$ raqamlar bilan raqamlangan boʻlsin. Shuningdek, N ta A alifbo belgilardan iborat $\bar{X}=(y_1, y_2, \dots, y_n)$ matn berilgan boʻlsin.

Taʼrif. $\bar{X}$ – alifbo belgilaridan tashkil topgan matn boʻlsin. $\bar{X}$ – matnning moslik indeksi deb:

$$I(\bar{X}) = \sum_{i=1}^m \frac{f_i(f_i - 1)}{n(n - 1)} \quad (1)$$

tenglik bilan aniqlangan qiymatga aytiladi [9].

Bunda, $\bar{X}$ – tahlil qilingan ochiq matnlar toʻplami;

f_i – alifboning i – belgisining matndagi uchrashlari soni;

n – matndagi barcha belgilar soni;

m – alifbodagi belgilar soni.

C# dasturlash tilida yaratilgan dastur yordamida 121 millionga yaqin belgilardan iborat oʻzbek tilidagi ochiq matnlarni (1) hisoblash formulasi yordamida tahlil qilib chiqildi va quyidagi natijalar olindi. Oʻzbek tilining moslik indeksi $I_{o'zbek} = 0,071698945$ qiymatiga teng boʻldi. Ushbu natijani ochiq manbaalardagi boshqa tillarning moslik indeksi bilan taqqoslanganda quyidagicha koʻrinish oldi (1-jadval).

1-jadval

Til	Rus tili	Ingliz tili	Oʻzbek tili	Italiyan tili	nemis tili	Fransuz tili
Moslik indeksi	0.0553	0.0644	0.071698945	0.0738	0.0762	0.0778

Hisoblash natijalaridan shuni aniqlash mumkinki oʻrganilayotgan matni qaysi tilga oidligini avtomatik ravishda aniqlash uchun oʻrganilayotgan matnning belgilari chastotalari aniqlash talab etiladi.

Berilgan matnning qaysi tilga yaqinligini aniqlash uchun quyidagi qiymat hisoblanadi:

$$X = \sum_{i=0}^m n_i k_i \quad (2)$$

bunda: η_i – tabiiy tilning i – harfining chastotasi, k_i – o'rganilayotgan matnning i – harfining chastotasi, m – alifbodagi belgilar soni.

Ushbu usul orqali matnning qaysi tildaligini aniqlash uchun matn uzunligi yetarlicha katta bo'lishi kerak. Agar, o'rganilayotgan matn uzunligi qisqa bo'lsa matnning chastota xarakteristikalarini tabiiy tilning xususiyatlaridan farq qilishi mumkin.

Amaliyotda konfidensial axborotlarni ruxsatsiz chiqib ketishini oldini olish tizimini aylanib o'tish uchun matndagi so'zlar va belgilarning o'rnini o'zgartirgan holda olib chiqishga urinishlar bo'ladi [3].

Faraz qilaylik, berilgan $A=(a_1, a_2, \dots, a_m)$ alifbo belgilaridan tashkil topgan $\bar{X}=(y_1, y_2, \dots, y_n)$ va $\bar{Y}=(y'_1, y'_2, \dots, y'_n)$ matnlarning qaysi tilga yoki sohaga yaqinligini aniqlash uchun matnlarning o'zor moslik indeksi hisoblanadi:

$$MI(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^m f_i f'_i}{nn'} \quad (3)$$

bunda, $\bar{X}, \bar{Y}$ – solishtirilayotgan matnlarning belgilari to'plamlari;
 f_i, f'_i – solishtirilayotgan matnlarning i – belgisining matnlardagi uchrashlari soni;

n, n' – solishtirilayotgan matnlardagi barcha belgilar soni;

m – alifbodagi belgilar soni.

Chastotali tahlil usuli yordamida o'zbek tilidagi matnlarni moslik indeksi hisoblab chiqilishi axborot tizimlarida mavjud matnlarning o'zbek tiliga oidligini avtomatik aniqlashning imkonini beradi. Shuningdek, konfidensial axborotlarni kibexujumchi tomonidan matndagi so'zlar yoki belgilarning o'rnini almashtirish orqali ruxsatsiz olib chiqib ketilishini oldini olishga xizmat qiladi. Shuningdek, ochiq matnlarning o'xshashligini aniqlashning algoritmi yordamida ishlab chiqilgan dasturiy ta'minot matnlarni solishtirish mumkin bo'ladi. Matnlarning o'zaro o'xshashligini aniqlash tizimini yaratilishi va uning amaliyotga joriy qilinishi milliy axborot resurslarini ishonchli himoyasini ta'minlaydi.

Foydalanilgan adabiyotlar:

1. Jo'rayev G.U., Bozorov O.N. «Elektron hukumat» sharoitida O'zbekiston Respublikasida konfidensial axborotlarni himoyalash muammolari va yechimlari// Toshkent shahri, 2019 yil 12-aprel. 108-111 b.
2. «Shaxsga doir ma'lumotlar to'g'risida»gi O'zbekiston Respublikasining Qonuni. -T., 2019 yil 2-iyul, O'RQ–547-son.
3. Radwan R. Tahboub, Yousef Saleh. Data Leakage/Loss Prevention Systems (DLP) NNGT Journal: International Journal of Information Systems. Volume 1, 2014. –P. 13-19.

O'ZBEKCHA MATNLARNI HISSIY TAHLIL QILISH UCHUN LEKSIKONGA ASOSLANGAN YONDASHUV

Allanazarova Sabohat Yusupboyevna

allanazarova.sabosh@gmail.com

ToshDO'TAU tayanch doktoranti

Annotatsiya. Axborot asrida ijtimoiy tarmoqlardan foydalanish orqali mijozlarning xizmat sifati, brend yoki mahsulotlar va hodisalarga bo'lgan munosabatlarini o'rganishga asoslangan tadqiqotlar ustunlik qilmoqda. Bunday onlayn sharhlar yordami bilan turli xil sohalarda mijozlarning fikrlarini o'rganish maqsadida qo'llanilishi mumkin. Hissiy tahlil qilishda qo'llaniladigan bir necha yondashuvlar mavud. Ushbu maqolada emotsional-ekspressivlik jihatidan hissiy baho ifodalovchi so'zlar lug'atidan foydalangan holda leksikonga asoslangan hissiy tahlil usuli ko'rib chiqiladi.

Kalit so'zlar: hissiy tahlil, ijtimoiy tarmoq, sharhlar, mijozlar fikri.

Abstract. In the information age, research based on the study of customer attitudes toward service quality, brand or products, and events through the use of social networks dominates. With the help of such online reviews, it can be used for the aim of studying the opinions of costumers in various fields. There are several approaches used in sentiment analysis. In this article, we will considered the method of lexicon based approach, using a dictionary of words expressing emotional-expressiveness.

Key words: sentiment analysis, social network, reviews, customer's opinion.

Аннотация. В век информации преобладают исследования, основанные на изучении отношения клиентов к качеству услуг, бренду или товарам и событиям посредством использования социальных сетей. С помощью таких онлайн-обзоров можно узнать мнение клиентов в различных отраслях. В анализе настроений используется несколько подходов. В данной статье рассматривается метод эмоционального анализа на основе лексики с использованием словаря слов, выражающих эмоциональную оценку с точки зрения эмоционально-экспрессивности.

Ключевые слова: анализ тональность, социальная сеть, отзывы, мнение клиентов.

Mijozlarning mahsulot yoki xizmatlar haqida qanday fikrda ekanligini tushunish biznesda har doim eng muhimlaridan biri bo'lib kelgan. Ijtimoiy tarmoq foydaluvchilarining xizmatlar va mahsulotlar haqidagi o'zgalarning fikri bilan qiziqishga o'chligi va bu fikrlarga asoslangan holda xizmatlar sifatini yaxshilashga qaratilgan biznes talabi NLP sohasi bo'lgan hissiy tahlil qilishni dolzarb tadqiqotlar

qatoriga qo'shdi. Katta ma'lumotlar davrida millionlab iste'molchilarning sharhlari va munozaralari internetda kuniga behisob aylanadi [1]. Bunday katta hajmli ma'lumotlarni qo'lda saralash va mavuviy tasniflash esa tasavvur qilganimizdek, murakkab vazifa. Ijtimoiy tarmoqdagi tizimga solinmagan matnlardagi hissiyotlarni avtomatik ravishda tahlil qilish uchun sentiment tahliliga ehtiyoj bor. Sentiment tahlili (ST) - yozma matnda ifodalangan fikrlarni miqdoriy jihatdan baholash va turkumlashtirish jarayonidir.

ST uchun qo'llanilishi mumkin bo'lgan ilk yondashuv – leksikaga asoslangan yondashuv bo'lib, u ikki xil ko'rinishda namoyon bo'ladi. Birinchisi **lug'atga asoslangan**, ikkinchisi esa, **korpusga asoslangan yondashuv**dir.

Lug'at asoslangan yondashuv. Lug'atga asoslangan yondashuvning asosiy g'oyasi matndan his-tuyg'u ifodalovchi so'zlarni ajratish uchun leksik ma'lumotlar bazasidan foydalanishdir. Har qanday tilda ST tizimini ishlash chiqish uchun asosiy vazifa bu hissiyot leksikonidir. Boshlang'ich kayfiyatdagi so'zlar to'plamiga (masalan: **yaxshi, yomon, ajoyib, qimmat**) asoslanib ularning qutblari qo'lda yig'iladi [2]. Dastlab, bu boshlang'ich to'plam katta bo'lishi shart emas, 30 ta fikr so'zlari yetarli [3]. Keyingi qadam leksik ma'lumotlar bazasidan mos sinonim va antonimlarni izlash orqali to'plamni boyitish uchun qutbli so'zlardan foydalanishdir. Resurslarga boy ingliz tilida o'nlab hissiyot leksikonlari (masalan, **SentiWordnet** [4], **Senticnet** [5], **NRC Emotion Lexicon** [6], **General Inquirer** [7], **AFIN** [8], **Wordnet-Affect** [9], **Opinion Lexicon** [10], **Subjectivity Lexicon** [11], **So-Cal** [12], **Micro-WNOp** [13], **SentiSense** [14]) mavjud. Boshqa tomondan, o'zbek tili resurslari kam til bo'lib, hissiyot leksikonlari hozircha ishlab chiqilmagan. Bunday lug'atlar qo'lda va yarim avtomat yoki avtomatik ravishda yaratilishi mumkin.

Lug'atlarni qo'lda ishlab chiqishda vazifa mahalliy tilda so'zlashuvchi tilshunos mutaxassislardan tarkib topgan guruh a'zolari tomonidan taqsimlanadi. Har bir nomzod so'z (bunda ot, fe'l, sifat, ravish, iboralar tahlilga tortilishi mumkin) a'zo etiketidan o'tgach, barcha tomonidan bir xil baholangan so'z o'zgarishsiz qoladi. Ba'zi birliklar esa o'zgarishi mumkin, sababi chunki har kim dunyoni turlicha qabul qiladi va tasavvur doirasi turlicha bo'ladi.

Mavjud leksikonlar orasida SentiWordNet (SWN) eng keng tarqalgan bo'lib, chunki u har bir sinsetga tayinlangan obyektiv, subyektiv va salbiy uchta belgining birlashishi hisoblanadi. Bundan tashqari, har bir sinset raqamli ballar yordamida belgilanadi, bu endi his-tuyg'u to'liq so'zning o'zi emas, balki so'zning ma'nosi bilan bog'liqligini ko'rsatadi. Bundan tashqari, bir vaqtning o'zida ijobiy va salbiy sinset ham hisobga olinadi. Bu muhim xususiyatlar SWN ni boshqa leksik manbalardan farqi hisoblanadi.

Har bir soha (xizmat ko'rsatish, media, fond bozorlari, siyosat va boshqalar) ixtisoslashgan lug'atlarga muhtoj bo'lsa-da, umumiy lug'atlar ham foydalidir, chunki ular ish jarayonida aniqlanishi va to'ldirilishi mumkin bo'lgan boshlang'ich material bo'lib xizmat qilishi mumkin. Buning sababi shundaki, har bir hissiyot atamalariga

tegishli his-tuyg'ular toifasi (**IJOBIY**: sevinch, mamnuniyat, ishonch; **SALBIY**: norozilik, nafrat, qo'rquv) va ball beriladi, bu so'z, jumla va hujjat darajasi kabi turli darajalarda ballni hisoblashda yordam beradi.

Korpusga asoslangan yondashuv sintaktik va semantik belgilarni ko'rib chiqishni o'z ichiga oladi. O'z-o'zidan korpusga asoslangan usul korpus tarkibidagi so'zlarning cheklanganligi sababli lug'atga asoslangan usuldan samaradorligi kamroq. Biroq ushbu yondashuvdan foydalanish turli sohalar va kontekst mavzusiga xos lug'atlarni shakllantirishga yordam beradi [15]. Ushbu usul uch qismdan iborat:

- Pos tegging usuli (har bir so'z ajratib olinadi);
- Har bir tokenga hissiy ohangiga qarab alohida ball beriladi;
- Gapning umumiy qutbliligi hisoblanadi.

Birinchidan, sharhlar oldindan qayta ishlanadi, ular tuzatiladigan jummalarga bo'linadi va jumladagi har bir so'zni belgilash va saqlash uchun “Matnni ajratish” (POS) usuli qo'llaniladi. Ikkinchidan, kontekstual ma'lumotlarga va jumla tuzilishiga asoslanib, gapning qutbliligini aniqlashga imkon beruvchi so'zni ajratib olish bosqichiga o'tiladi. Bu gap tarkibidagi iboralarni tanib olish imkonini beruvchi aspekt hisoblanadi. Oxirgi qism qoidaga asoslangan modul yordamida jumalarni obyektiv yoki subyektiv tasniflashdan iborat.

Leksikaga asoslangan yondashuv asosida subyektivlik tahlili ham amalga oshiriladi. Bu tahlilda har bir matn tarkibidagi gap, har bir gap tarkibidagi so'zlar subyektiv yoki obyektiv birlik ekanligi aniqlanadi. Subyektiv gapda hissiyotni idodalovchi so'z qatnashgan, obyektiv gapda esa hech qanday fikr so'zlari yo'q bo'ladi.

ID	Gap	Hissiyot so'zlari	Subyektiv/Obyektiv
S1	Yer quyosh atrofida aylanadi	-	obyektiv
S2	Men kecha telefon sotib oldim	-	obyektiv
S3	Juda shinam joy. Avvalo, uning atmosferasi yoqimli.	shinam, yoqimli	subyektiv

1-rasm. Subyektivlik tahlili na'munasi

Yuqoridagi 3-gapda **shinam** va **yoqimli** so'zlari bu gapni subyektiv deb topilishiga asos bo'lgan.

Ta'sirchan ma'no bo'yicha dastlabki ishlar [Osgood va boshq.] 1957-yilda ta'sirchan so'zlarning ma'nosini uchta muhim o'lchov (salbiy, ijobiy va betaraf) bo'yicha o'zgarishini aniqlashdan boshlangan.

Sentiment (hissiy) tahlilda tadqiqot obyekti sifatida restoran sharhlari, mahsulot yoki brend sifati, xizmat ko'rsatish, siyosat yoki mavjud vaziyat haqidagi sharhlar olinishi mumkin. Ijtimoiy tarmoqdagi sharhlar kulgich (emoji) va jargonlar bilan to'la bo'lganligi sababli, ularning barchasini boshqarish uchun ixtisoslashgan leksika kerak bo'ladi. Sentiment tahlil qilishda hissiyot so'zlari turli xil sohalarga sezgir va tasniflashda turlicha talqin qilinadi. Shu sababli, tasniflashda ixtisoslashgan lug'atlardan foydalanish maqsadga muvofiq. Restoranlar sharhlarida mayonez va yog'li sochiq kabi so'zlar salbiy ma'noga ega. Agar bu so'zlar sharhda paydo bo'lsa,

u odatda bu joyda salbiy baho sifatida baholanadi. Shubhasiz, matnlardagi bunday yashirin jihatni tahlil qilish alohida ahamiyatga ega.

Hissiyotlarni tahlil qilishda turli darajadagi vazifalar mavjud. Asosiy vazifa – berilgan matnda bildirilgan fikr ijobiy/salbiy yoki betaraf qutbliligini aniqlash. Qutblilik – berilgan gapning ijobiy yoki salbiy ekanligini ko'rsatadigan o'lchovdir. Biroz murakkabroq vazifa esa 1 dan 5 gacha bo'lgan shkala yordamida matndan hissiyot darajasini hisoblashdir, bu yerda 1 qiymati eng salbiy, aksincha, 5 qiymati eng ijobiyni anglatadi. Yana bir muhim masala - ma'lum bir jumlada qutblanish modifikatorining doirasini aniqlash (masalan, inkor ma'noni). Masalan, **bu yangi telefonning dizayni menga yoqadigan emas, lekin u ba'zi qiziqarli funksiyalarga ega** jumlasida **emas** faqat yoqadigan so'ziga tegishli bo'lib, **qiziqarli** so'zining qutblanishini o'zgartirmaydi.

Google Map ilovasida o'zbek restoran va oshxonalariga sharhlar berilgan va tegishli yulduzchalar bilan belgilangan. Bunda biz 5 yulduzli sharhlarda qanday ijobiy so'zlar, 1 yulduzchali sharhda qanday salbiy so'zlar uchrashini bilishimiz mumkin. Ilovaning kamchilik tomoni, Twitterda bo'lgani kabi avtomatik ijobiy/salbiy sharhlarga ajratilmagan.

Modifikator	Qutblilik
cheksiz	+ ijobiy
juda	+ ijobiy
bag'oyatda	+ ijobiy
mukammal	+ ijobiy
to'liq	+ ijobiy
ko'p	+ ijobiy
benihoya	+ ijobiy
behad	+ ijobiy
zo'rg'a	- salbiy
bazo'r	- salbiy
arang	- salbiy
deyarli	- salbiy
biroz	- salbiy
picha	- salbiy

2-rasm. O'zbek tilida uchrashi mumkin bo'lgan modifikatorlarning qisqacha ro'yxati

Norasmiy til birliklari, shu jumladan kulgichlar va jargonlar, ko'pincha ijtimoiy tarmoq foydalanuvchilari tomonidan o'z xabarlarini yetkazish uchun qo'llaniladi. Ingliz tili kulgichlar va jargonlar va hattoki haqorat so'zlari uchun leksik manbalar bilan yaxshi ta'minlangan bo'lsa-da, o'zbek tilidagi ST uchun bunday manbalarga muhtoj va o'zbekcha sharhlardagi norasmiy tilni tahlil qilishga to'sqinlik qiladi. Va aksincha, ijtimoiy tarmoqdagi sharhlarda bunday lingvistik manbalar ko'p uchraydi.

Xulosa. Internetning paydo bo'lishi mahsulotlar, xizmatlar, voqealar va siyosiy fikrlar kabi mavzular bo'yicha ma'lumotlarni oson va keng almashish uchun eshiklarni ochdi. O'rgatilgan resurslari kam o'zbek tilida hissiyot tahlili uchun lingvistik ta'minot manbalariga muhtoj.

Foydalanilgan adabiyotlar:

1. Dang, Yan, Yulei Zhang, and Hsinchun Chen. 2010. A lexicon enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53.
2. H. Han, Y. Zhang, J. Zhang, J. Yang, and X. Zou, “Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias”, pp. 1–11, 2018
3. B. Liu, “Sentiment analysis: Mining opinions, sentiments, and emotions,” *Cambridge University Press*, 2015.
4. Baccianella S, Esuli A, Sebastiani F. Sentiwordnet3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th International Conference on Language Resources and Evaluation. 2010, 2200–2204
5. Cambria E, Olsher D, Rajagopal D. SenticNet3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2014, 1515–1521
6. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436–465
7. Stone P J, Bales R F, Namenwirth J Z, Ogilvie D M. The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 1962, 7(4): 484–498
8. Nielsen F. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of Workshop on Making Sense of Microposts: Big Things Come in Small Packages. 2011, 93–98
9. Strapparava C, Strapparava C, Valitutti A. WordNet-affect: an affective extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation. 2004, 1083
10. Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 2015, 89: 14–46
11. Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 2005, 39(2): 165–210
12. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 2011, 37(2): 267–307
13. Cerini S, Compagnoni V, Demontis A, Formentelli M, Gandini G. Micro-WNOp: a gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In: Proceedings of Language Resources and Linguistic



Theory: Typology, Second Language Acquisition, English Linguistics. 2007, 200–210

14. De Albornoz J C, Plaza L, Gervais P. SentiSense: an easily scalable concept-based affective lexicon for sentiment analysis. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. 2012, 23–25

15. Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In LREC, 2010.

16. Elov B., Hamroyeva Sh., Xusainova Z. NLP (tabiiy tilga ishlov berish)ning vazifalari va zamonaviy yondashuvlar. TerDU, Filologik tadqiqotlar: til, adabiyot, ta'lim. 2022, 5-6.

MATNNI AVTOMATIK TAHLIL QILISHDAGI MUAMMOLAR

Suyunova Malika Odil qizi

Alisher Navoiy nomidagi Toshkent davlat
o'zbek tili va adabiyoti universiteti
Kompyuter lingvistikasi mutaxassisligi
2-bosqich magistranti,
malikasuyunova0@gmail.com

Annotatsiya. Bugungi kunda korpus lingvistikasi sohasida ko'plab ishlar amalga oshirilmoqda. Ushbu maqolada morfologik analizator yuzasidan qilingan ishlar, matnni avtomatik tahlil qilish, tahlil jarayonidagi muammolar, ko'p ma'noli va omonim so'zlarni korpusda farqlash masalasi, grammatik omonimlar bilan bog'liq muammolar va ularning yechimlari haqida so'z yuritiladi.

Kalit so'zlar: *korpus, kompyuterli grammatika, kompyuterli semantika, morfologik analizator, omonimlar, grammatik omonimlar.*

Abstract. Today, there are many work is being done in the field of corpus linguistics. This article discusses the work done on the morphological analyzer, automatic text analysis, problems in the analysis process, the issue of distinguishing multi-meaning and homonymous words in the corpus, problems related to grammatical homonyms and their solutions.

Keywords: *corpus, computer grammar, computer semantics, morphological analyzer, homonyms, grammatical homonyms.*

Абстрактный. На сегодняшний день ведется большая работа в области корпусной лингвистики. В данной статье рассматриваются проделанная работа по морфологическому анализатору, автоматический анализ текста, проблемы в процессе анализа, проблема различения многозначных и омонимичных слов в корпусе, проблемы, связанные с грамматическими омонимами и пути их решения.

Ключевые слова: *корпус, компьютерная грамматика, компьютерная семантика, морфологический анализатор, омонимы, грамматические омонимы.*

Kompyuter texnologiyalarida matn asosiy o'rinni egallaydi. Amaliy tilshunoslik sohasida tadqiqotchilar matnni bir tildan boshqa bir yoki bir necha tillarga avtomatik ravishda tarjima qiluvchi, matnni morfologik, sintaktik, semantik jihatdan avtomatik tahlil qiluvchi, matnda ishtirok etgan so'zlarning chastotasini ko'rsatuvchi va boshqa shu kabi dasturlarni yaratish ustida amaliy ishlarni olib borishmoqda. Xorijda bu kabi ishlar keng miqyosda amalga oshirilib, katta-katta hujjatlar matematik metodlar asosida tahlil qilinmoqda. 1950-yillardan boshlab matnni avtomatik tahlil qilish bo'yicha ko'plab tajribalar o'tkazildi. Ularning ko'pchiligi ko'p ishlatiladigan atamalar yoki iboralardan iborat matnni tahlil qilish

ustida olib borildi. Bu vaqtning tejalishiga, inson tomonidan qo'yiladigan kamchiliklarning kamayishiga olib keladi. Avallari ko'p vaqt, aqliy mehnat va kuch sarflab qilingan ishlar, endi tez fursatlarda inson aqli yordamida yuzaga kelgan dasturlar yordamida samarali natijalarga erishilmoqda. Shu bilan bir qatorda, axborot-qidiruv tizimida kiritmalarni to'g'ridan-to'g'ri tahlil va tahrir qilish imkoniyati kengaymoqda. Bugungi kunda matnni texnik tahlil qilish borasida yaratilayotgan ishlar asosan quyidagi yo'nalishda olib borilmoqda [Abjalova M., 2020:56]:

1. Kompyuterli grammatika
 - a) kompyuterli morfologiya
 - b) kompyuterli sintaksis
2. Kompyuterli semantika

Matnni avtomatik tahlil qilishda, eng avvalo, o'sha tildagi barcha so'zlarni o'zida jamlagan lug'at fondi muhim ahamiyatga ega. Chunki lug'at xalqning madaniyati, qadriyatlari va fikrlarini bilish uchun eshikdir. Matndagi so'zlar tahlil qilinayotganda bazaga kiritilgan lug'atga murojaat qilinadi. Bu usul bilan tahlil qilish eng keng tarqalgan va yaxshi samara bergan usullardan biridir. Lug'at matni tahliliga misol sifatida avtomatlashtirilgan kontent tahlilini keltirish mumkin. Bu tahlilda tez-tez ishlatiladigan so'zlar chiqarib tashlanadi. So'zlar asos shaklga ya'ni lug'atdagi ko'rinishga qadar kesib tashlanadi va matndagi so'zning chastotasi ko'rsatiladi.

Morfologik tahlilning vazifasi matndagi har bir so'zning qaysi turkumga mansubligini avtomatik tarzda muayyanlashtirish; so'zlarning qaysi leksik-grammatik sinfga tegishlilikini aniqlashdan iborat. Rus tilida formal morfologiya rivojlanganligi sababli buni amalga oshirishning keng imkoniyatlari mavjud. Ingliz tilida ham algoritm nisbatan oddiy: leksik ko'p ma'nolilik mavjud bo'lsa-da, morfoanaliz dasturlari (tegger) matndagi so'zning turkumini 90 foizda to'g'ri aniqlay oladi. Rus tili matnlarining morfologik analizini amalga oshirish uchun A. Zaliznyak grammatik lug'atining kompyuter versiyasidan foydalanilgan, ingliz tilining morfoanalizi Myuller grammatik lug'ati asosida amalga oshirilgan. Rus va ingliz morfoanalizini yaratish tajribasi shuni ko'rsatadiki, morfologik analizni amalga oshirish uchun formal grammatika yaratilgan bo'lishi kerak [Xampoeva III., 2020:37].

Matnni avtomatik tahlil qilishda ko'chma ma'noda qo'llangan so'zlar ya'ni ko'p ma'noli so'zlar, omonim so'zlarning qaysi ma'noda ishlatilayotganini kompyuterda aniqlash eng dolzarb masalalardan biridir. Chunki inson aqligina so'zning qaysi ma'noda kelayotganini matn mazmunidan kelib chiqib aniqlay oladi. Buni kompyuter dasturlari orqali amalga oshirish ko'pchilikni o'ylantirgan va har xil yondashuvlar kelib chiqqan. O'zbek tilshunosligida yangi soha hisoblangan korpus tilshunosligida matnni avtomatik tahlil qilishda ko'p ma'noli so'zlarni aniqlash, omonimiya muammosini yechish, ularni teglash kabi masalalar kun tartibiga qo'yilgan. Bu borada qator tadqiqotlar amalga oshirilgan. Jumladan, Sh.Gulyamova “O'zbek tili semantik analizatorining lingvistik asoslari” monografiyasida

polifunksional va omonim soʻzlarni korpusda farqlash uchun semantik filtr ishlab chiqqan. Bu nafaqat morfoanalizatorida, balki semantik analizatorida eng qiyin muammoni bartaraf etishga xizmat qiladi. Omonimiya hodisasi qadimdanoq dolzarb, keng muhokamalarga sabab boʻluvchi muammolardan biridir. Maʼlumki, soʻz har qanday tilning leksik boyligini tashkil etadi, bir-biri bilan oʻzaro muloqotda boʻladi, turli qadriyat hamda vazifalarni bajaradi. Soʻzlashuv nutqida soʻzlar maʼno jihatdan bir-biriga yaqin yoki aksincha, zid maʼnoga ega boʻladi. Baʼzi soʻzlar maʼno jihatdan yaqin boʻlmasa ham, koʻrinishi, shakli, imlo, talaffuzi bilan bir-biriga oʻxshaydi. Bunday soʻzlar guruhi tilning umumiy sistemasida omonimlar deyiladi. Omonim – grek tilidan olingan murakkab soʻz, *homos* “bir xil” va *onoma* “ism” degan maʼnolarni bildiradi. Shu bilan birga, omonimlar *bir xil nomlar* deb tarjima qilinadi [Гулямова III., 2021:53]. Sh. Rahmatullayevning “Oʻzbek tili omonimlarining izohli lugʻati” da omonimlar yozilishi bir xil fonemalardan iborat shakllar boʻlib, omonimiyada ifoda plani yetakchi turadi deb taʼrif berilgan[Rahmatullayev Sh., 1984:5].

Omonim hodisada yasama leksema va tub leksema, shuningdek, ularga grammatik maʼno ifodalovchi affiks qoʻshilganda voqe boʻladigan shakliy teng kelish ham qamrab olinadi. Masalan, *terim* soʻzshakli *ter* – feʼl leksemasiga *–im* leksema yasovchisining birikishi natijasida hosil qilingan leksemaning birlik, bosh kelishik formasiga teng.

Aniqroq boʻlishi uchun, bir necha turdagi omonimlarni ajratib koʻrsatish mumkin:

1. Nutqning bir qismi yoki lugʻatdagi shakli (lemma) farqlanadi;
2. Baʼzi morfologik xususiyatlar, masalan, kelishik, egalik va sonda (boy morfologiyali tillarda) farqlanadi;
3. Bir-biridan faqat mazmun jihatdan farqlanadi (bu holat semantik farqlanishda uchraydi) [Abjalova M., 2020:71].

Bugungi kunda lingvistik jihatdan tadqiqotlar olib borishda til korpuslarining ahamiyati katta. Chunki til korpuslari turli lingvistik topshiriqlarni bajarishga xizmat qiladi. Tilshunoslikning ilk va oʻrta bosqichlarida omonimiya hodisasining paydo boʻlishi, taraqqiyoti va yondosh hodisalarga munosabatlari oʻrganilgan boʻlsa, zamonaviy lingvistikada uning korpusda berilishi masalasi kun tartibiga qoʻyilgan. Sh. Hamroyeva taʼkidlaganidek, til korpusi elektron kutubxonadan farqli ravishda muayyan tilni tadqiq qilish, oʻrganish va oʻrgatish uchun zaruriy, foydali va qiziqarli matnlar toʻplanishini nazarda tutadi [Hamroyeva Sh., 2018:53].

Matnlarni avtomatik qayta ishlash bir necha bosqichga boʻlinadi va ulardan biri morfologik tahlil bosqichi hisoblanadi. Mazkur bosqichda har bir soʻzga morfologik tavsif beriladi: lemmasi (asosi), kelishigi, soni, darajasi, nisbati, shaxsi va hk. Morfoanalizning soʻzlarni morfologik teglash vazifasi omonim shakllar bilan murakkablashadi. Omonimiyani aniqlashning bir qator usullari aniqlangan. Mazkur usullarga tayanuvchi tizimlar oʻz navbatida quyidagi guruhni tashkil etadi:

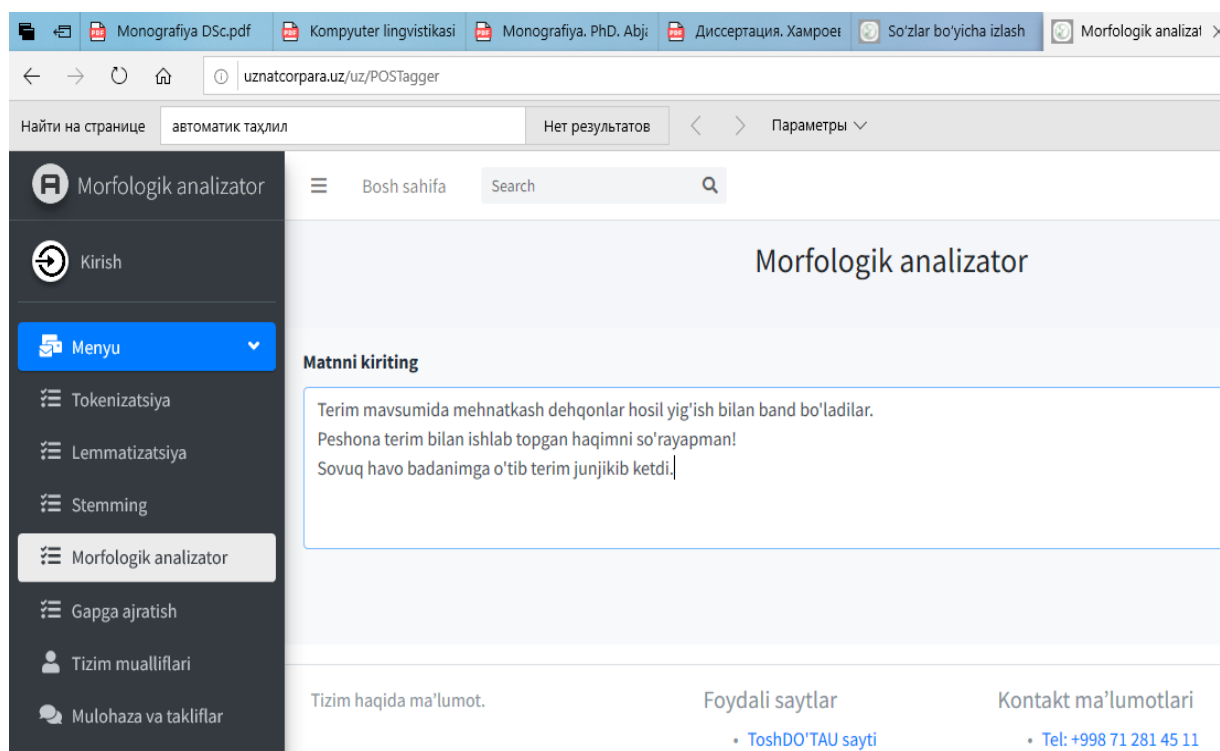
1. Qoʻlda yaratilgan qoidalarga asoslangan tizimlar.

2. Ehtimoliy modellar asosida yaratilgan va tavsiflangan korpuslarga tayanadigan tizimlar.

3. Ehtimollik modellari va qoidalarga asoslangan gibrid tizimlar[Abjalova M.,2021].

Biz tilimiz tabiati va xususiyatini inobatga olib, ulardan foydalanamiz. Omonimlarning bir qancha turlari mavjud bo'lib, ulardan leksik omonimlar lug'ati Sh.Rahmatullayev tomonidan tuzilgan. Leksik omonimlarni Sh.Gulyamova semantik analizatorida aniqlash uchun filtr ishlab chiqib muammoning yechimini topgan. Omonimlarning yana bir turi grammatik omonimlar bo'lib, ular so'zlarga grammatik shakllar qo'shilganda yuzaga keluvchi nutqiy jarayondir. Ayrim manbalarda leksik omonimlardan farqli ravishda grammatik omonimlar ham borligi haqida to'xtalib o'tilgan.

Morfologik analizator jarayonida so'zlar turkumga ajratilganda grammatik omonimlar shaklan bir xil bo'lganligi, lekin leksik omonimlar safiga kiritilmaganligi sababli qaysi turkumda va qaysi ma'noda kelayotganligini aniqlash juda qiyin. Quyida grammatik omonimlarning morfoanalizator jarayonida tahlil qilinishini ko'ramiz [<http://uznatcorpara.uz/uz/POSTag>]:



1-rasm. Morfologik analizatorga grammatik omonimlar ishtirokidagi gaplarning kiritilishi

Morfologik analizatorga grammatik omonimlar ishtirok etgan uchta gapni kiritdik(1-rasm) va tahlil qildik:

Morfologik analizator				
Kirish				
Menyu				
Tokenizatsiya				
Lemmatizatsiya				
Stemming				
Morfologik analizator				
Gapga ajratish				
Tizim mualliflari				
Mulohaza va takliflar				

№	So'z	Lemma	So'z turkumi	O'zak va qo'shimchalar
1	Terim	Terim	Ot ...	{Terim}
2	mavsumida	mavsum	Ot ...	{mavsum}-i-da
3	mehnatkash	mehnat	Ot ...	{mehnat}-kash
4	dehqonlar	dehqon	Ot ...	{dehqon}-lar
5	hosil	hosil	Ot ...	{hosil}
6	yig'ish	yig'ishmoq	Fe'l ...	{yig'ish}
7	bilan	bilan	Bog'lovchi ...	{bilan}
8	band bo'ladilar	band bo'lmoq	Fe'l ...	{band bo'li}-a-dilar

2-rasm. Morfoanalizatorga kiritilgan birinchi gapning tahlil natijasi

Birinchi gapda *terim* so'zi ot so'z turkumi sifatida to'g'ri tahlil qilingan(2-rasm). Ot so'z turkumining chuqurroq tahlilini tekshirib ko'ramiz(3-rasm)

Morfologik analizator													
terim													
Ot													
Tuzilishiga ko'ra (Ot)				Yasalishiga ko'ra (Ot)		Yasalish usuli (Ot)		Predmetning yakka/to'dasini bildirishiga ko'ra		Bevosita kuzatish mumkinligiga ko'ra		Turdosh ot	
Sodda ot	Qo'shma ot	Juft ot	Takror ot	Tub ot	Yasama ot	Kompozitsiya (ot)	Affiksatsiya (ot)	Yakka ot	Jamlovchi ot	Mavhum ot	Aniq ot	Narsa/buyum otlari	Shaxs otlari nomla
+					+		+	+			+		

3-rasm. Grammatik omonim so'zning tahlil qilinishi

terim													
Bosita kuzatish mumkinligiga ko'ra		Turdosh ot											
Mavhum	Aniq ot	Narsa/buyum otlari	Shaxs otlari nomlari	Inson tana a'zolari nomlari	Hayvon va uning a'zolari otlari	Oziq-ovqat nomlari	O'simlik va uning qismlari otlari	Makon/maskan otlari	Zamon otlari	Diniy-ilohiy otlar	Miqdor otlari	Faoliyat/jarayon otlari	
	+												+

4-rasm. Grammatik omonim so'zning tahlil qilinishi (davomi)

Terim so'zi ot:

1. Sodda ot
2. Yasama ot
3. Affiksatsiya
4. Yakka ot
5. Aniq ot
6. Faoliyat/jarayon otlari

Bunda morfoanalizator *terim* so'zi barcha jihatdan to'g'ri tahlil qilingan.
Keyingi gapning tahlil qilinishini ko'ramiz(5-rasm):

Kirish				
Menyu				
Tokenizatsiya				
Lemmatizatsiya				
Stemming				
Morfologik analizator				
Gapga ajratish				
Tizim mualliflari				
Mulohaza va takliflar				
9	Peshona	Peshona	Ot ...	{Peshona}
10	terim	terim	Ot ...	{terim}
11	bilan	bilan	Bog'lovchi ...	{bilan}
12	ishlab	ishlamoq	Fe'l ...	{ishla}-b
13	topgan	top	Fe'l, Ot, Sifat ...	{top}-gan
14	haqimni	haq		{haq}-im-ni
15	so'rayapman	so'ramoq	Fe'l ...	{so'ra}-yap-man

5-rasm. Grammatik omonim ishtirokidagi ikkinchi gapning tahlil qilinishi

terim														
Ot														
Tuzilishiga ko'ra (Ot)				Yasalishiga ko'ra (Ot)		Yasalish usuli (Ot)		Predmetning yakka/to'dasini bildirishiga ko'ra		Bevosita kuzatish mumkinligiga ko'ra		Turdosh ot		
Sodda ot	Qo'shma ot	Juft ot	Takror ot	Tub ot	Yasama ot	Kompozitsiya (ot)	Affiksatsiya (ot)	Yakka ot	Jamlovchi ot	Mavhum ot	Aniq ot	Narsa/buyum otlari	Shaxs otlari	nomla
+					+		+	+			+			

6-rasm. Grammatik omonim so'zning tahlil qilinishi

terim												
vosita tashish imkinligiga ra		Turdosh ot										
vhum	Aniq ot	Narsa/buyum otlari	Shaxs otlari nomlari	Inson tana a'zolari nomlari	Hayvon va uning a'zolari otlari	Oziq- ovqat nomlari	O'simlik va uning qismlari otlari	Makon/maskan otlari	Zamon otlari	Diniy- ilohiy otlar	Miqdor otlari	Faoliyat/jarayon otlari
	+											+

7-rasm. Grammatik omonim soʻzning tahlil qilinishi (davomi)

Ikkinchi gapda ham yuqoridagidek tahlil qilindi(6-7-rasm). Aslida *terim* yasama emas tub ot, faoliyat-jarayon oti emas narsa oti deb tahlil qilinishi kerak.

Morfoloqik analizator				
Kirish				
Menyu				
Tokenizatsiya				
Lemmatizatsiya				
Stemming				
Morfoloqik analizator				
Gapga ajratish				
Tizim mualliflari				
Mulohaza va takliflar				

16	Sovuq	Sovuq	Ot, Sifat ...	{Sovuq}
17	havo	havo	Ot ...	{havo}
18	badanimga	badan	Ot ...	{badan}-im-ga
19	o'tib	o't	Fe'l, Ot ...	{o't}-ib
20	terim	terim	Ot ...	{terim}
21	junjikib	junjikmoq	Fe'l ...	{junjik}-ib
22	ketdi			ketdi

8-rasm. Grammatik omonim ishtirokidagi uchinchi gapning tahlil qilinishi

Uchinchi gapda ham *terim* soʻzi yasama, faoliyat-jarayon oti sifatida tahlil qilindi. Aslida tub, inson ta'na a'zosi sifatida tahlil qilinishi kerak edi(8-rasm).

Ko'rinib turibdiki, avtomatik tahlil jarayonida grammatik omonimlarni aniqlab, ularni to'g'ri tahlil qilish nihoyatda kerakli masalalardan biridir. Buning uchun esa grammatik omonimlarni umumiy va xususiy birikuvchilarini aniqlab olish va shu asosda ularning semantik maydonini shakllantirib olish, barcha grammatik omonimlar uchun model ishlab chiqish kabi vazifalar kun tartibida turibdi.

Foydalanilgan adabiyotlar:

1. Хамроева Ш. Ўзбек тили морфологик анализаторининг лингвистик таъминоти. – GlobeEdit, 2020: 268.
2. Гулямова Ш. Ўзбек тили семантик анализаторининг лингвистик асослари. Монография – Тошкент: 2021: 277.



3. Rahmatullayev Sh. O'zbek tili omonimlarining izohli lug'ati. – Toshkent: O'qituvchi, 1984.
4. Hamroyeva Sh. O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Fil.fan. bo'yicha falsafa dokt. (PhD) diss. avtoref. – Qarshi, 2018.
5. Elov B., Hamroyeva Sh., Xusainova Z. NLP (tabiiy tilga ishlov berish)ning vazifalari va zamonaviy yondashuvlar. TerDU, Filologik tadqiqotlar: til, adabiyot, ta'lim. 2022, 5-6.
6. Elov B., Hamroyeva Sh., Elova D. Morfologik analizatorni yaratish usullari. O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(1). 67-87.
7. <http://uznatcorpara.uz/>

NOMUHIM SO‘ZLAR TUSHUNCHASI VA UNING AHAMIYATI

Mahmudjonova Gulshahnoz Ulug‘bek qizi

ToshDO‘TAU, kompyuter lingvistikasi

yo‘nalishi 1-kurs magistranti,

gulshahnozmahmudjonova@gmail.com

Annotatsiya: Ushbu maqolada nomuhim so‘zlar tushunchasi va uning ahamiyati haqida so‘z yuritiladi. Grammatik qoidalarga asoslangan nomuhim so‘zlar ro‘yxati mustaqil so‘zlardan, yordamchi so‘zlardan ishlab chiqilib, amaliyotga taklif etiladi.

Kalit so‘zlar: *nomuhim so‘zlar, semantik ma‘no, tabiiy tilni qayta ishlash, mashinali o‘qitish, grammatik usul, chastota, grammatik ma‘no, hujjat, yordamchi so‘zlar.*

Annotation: This article discusses the concept of stop words and its importance. A list of stop words based on grammatical rules is developed from independent words and auxiliary words and offered for practice.

Key words: *stop words, semantic meaning, natural language processing, machine learning, grammatical method, frequency, grammatical meaning, document, auxiliary words.*

Kirish. Nomuhim so‘zlar barcha tillarda mavjud bo‘lib, semantik ma‘noga ega bo‘lmagan so‘zlarga aytiladi. Kam ma‘noli ma‘lumotlarga ega bo‘lgan yoki mustaqil ma‘noga ega bo‘lmagan yoki barcha matnlarga xos keng tarqalgan so‘zlar nomuhim so‘zlari deb ataladi [Madatov X.A, Sharipov M.S, Bekchanov Sh.K, O‘zbek tili matnlaridagi nomuhim so‘zlar]. Dastlab hujjatlarni qayta ishlash va qidiruv tizimi algoritmlarini tezlashtirish uchun foydalanilgan nomuhim so‘zlar endi modellarni o‘qitish kabi mashinali o‘rganishni o‘z ichiga olgan tabiiy tilni qayta ishlash (NLP) vazifalarida: ma‘lumotlarini qidirish (IR), matnni sarhisob qilish, indekslash, ijtimoiy media platformalari, klasterlash va qidirish tizimlarida qo‘llanilib kelinmoqda. Nomuhim so‘zlar matnga semantik ma‘no deyarli yuklamasligi sabab, hujjatdan ularni olib tashlash ma‘lumotlar to‘plami hajmini kamaytirishi mumkin, bu esa ma‘lumotlarni tez va oson qayta ishlashda yordam beradi. Nomuhim so‘zlarni olib tashlash orqali hisoblash tez va ma‘lumotlar xotiradan ko‘p joy olmasligiga erishish mumkin. Ular **hech qanday mazmunli ma‘lumot bermaydi**, ayniqsa, matn tasniflash modellarida foydalanish yaxshi natija beradi. TF-IDF asosida o‘zbek tili matnlarini tasniflashda matnga oldindan ishlov berish jarayonida, ya‘ni lemmatizatsiya qilishdan oldin nomuhim so‘zlar matndan olib tashlangan [Mahmudjonova G.U. TF-IDF asosida o‘zbek tili matnlarini tasniflash, O‘zbek tilining milliy korpusi: muammolar va vazifalar mavzusidagi xalqaro ilmiy-amaliy anjumani, 2023.25.03]

Asosiy qism. Nomuhim soʻzlar hujjatning katta qismini tashkil etuvchi matn tarkibidagi soʻzlar boʻlib, ularning umumiy xususiyati shundaki, hujjatda muhim maʼlumotga ega emas; faqat grammatika tufayli ishlatiladi. Hujjatdagi matndan nomuhim soʻzlar tushirib qoldirilsa, matn mazmuni oʻzgarmaydi.

Matnning asl varianti: Farzandingizga ertaklardan birini oʻqing. Matn haqida bir nechta savol bering. Agar bola oʻqiy olsa, uni oʻzi qisqa hikoya oʻqishga taklif qiling va keyin uni qayta aytib bering.

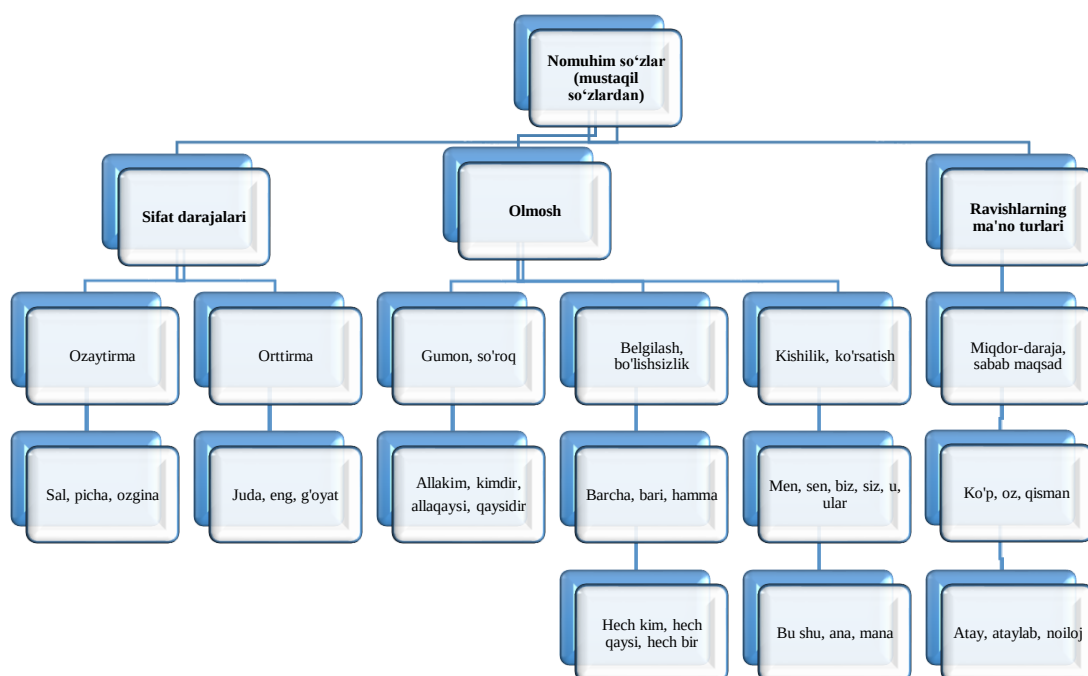
Matndan nomuhim soʻzlar olib tashlangandan keyingi holati: Farzandingizga ertaklardan oʻqing. Matn haqida savol bering. Bola oʻqiy olsa, qisqa hikoya oʻqishga taklif qiling, qayta aytib bering.

Nomuhim soʻzlarini aniqlash usullarini ikkita asosiy toifaga boʻlinadi:

1. Grammatika qoidalariga asoslanib
2. Statistik usullar.

Ushbu maqolada grammatik qoidalarga asoslanib aniqlash mumkin boʻlgan nomuhim soʻzlar taklif etiladi

Nomuhim soʻzlar roʻyxati, asosan, mustaqil soʻz turkumlaridan sifat darajalari, miqdor-daraja, sabab, maqsad ravishi, olmoshlar, yordamchi soʻzlar, modal, taqlid, undov soʻzlardan tuzilishi mumkin. Sababi, bunday soʻzlarning matn mazmuniga taʼsiri kam boʻladi, yaʼni gapda grammatik maʼno ifodalash uchun foydalaniladi. Quyida barchasi misolar orqali tushuntiriladi.



1-jadval. Nomuhim soʻzlar roʻyxati

Predmet, narsa-buyumning belgisini bildirib, qanday, qanaqa soʻroqlariga javob boʻluvchi soʻz turkumi sifat deyiladi. Sifatlar darajalanadi. Belgining meʼyoridan, ortiq yoki kam darajada ekanligini koʻrsatish sifat darajalari deyiladi.

Sifat darajalari 3 xil: oddiy, qiyosiy, orttirma daraja. Leksik usul bilan hosil qilingan sifatning qiyosiy va orttirma darajalari nomuhim soʻz boʻla oladi.

1	<i>Qiyosiy daraja sifat ishtirok etgan gap</i>	Bemorning sogʻligʻi kechagidan sal yaxshi
2	<i>Qiyosiy daraja sifat ishtirok etmagan gap</i>	Bemorning sogʻligʻi kechagidan yaxshi

Yuqorida berilgan gapni orttirma daraja shakli bilan kelganda:

1	<i>Orttirma daraja sifat ishtirok etgan gap</i>	Bemorning sogʻligʻi kechagidan gʻoyatda yaxshi
2	<i>Orttirma daraja sifat ishtirok etmagan gap</i>	Bemorning sogʻligʻi kechagidan yaxshi

Ot, sifat, son, ravish oʻrnida qoʻllanib, ularni koʻrsatishga xizmat qiluvchi va ularning vazifasini bajaruvchi soʻz turkumi olmosh deyiladi. Olmoshlar narsani, uning belgisi va miqdorini anglatmaydi, balki faqat ularni koʻrsatish, ularga ishora qilish uchun xizmat qiladi. Olmosh gapda ega, toʻldiruvchi, aniqlovchi boʻlib keladi. Gapda olmoshlar aniqlovchi vazifasida kelganda, gapda ularni tushurib qoldirish mumkin.

1	<i>Koʻrsatish olmoshi ishtirok etgan gap</i>	Bu qalam juda sifatli.
2	<i>Koʻrsatish olmoshi ishtirok etmagan gap</i>	Qalam juda sifatli

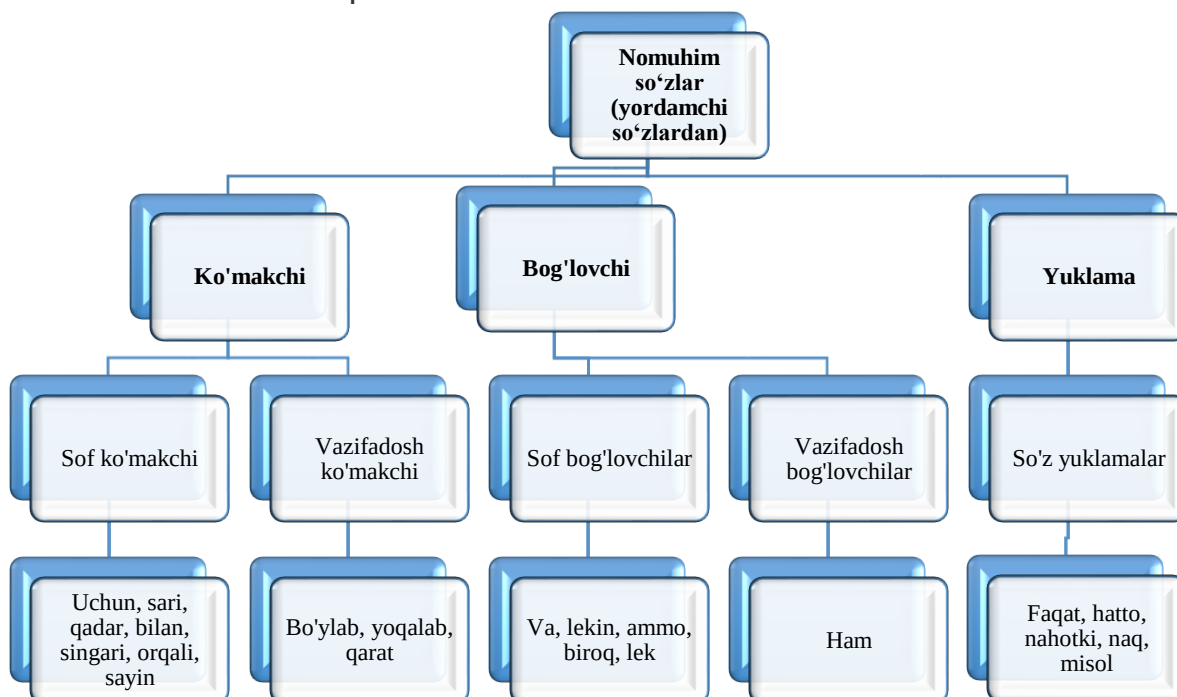
Gumon, belgilash, boʻlishsizlik, kishilik olmoshlari ham gapda aniqlovchi boʻlib kelganda nomuhim soʻzlar roʻyxatiga kiritish mumkin [Sayfullayeva.R, Mengliyev.B va boshqalar, Hozigi oʻzbek tili, Toshkent 2009, 345-bet].

1	<i>Belgilash olmoshi ishtirok etgan gap</i>	Gaplaringiz barchasi yolgʻon
2	<i>Belgilash olmoshi ishtirok etmagan gap</i>	Gaplaringiz yolgʻon

Ish-harakatning, holatning belgisini, belgining belgisini bildiruvchi soʻz turkumi ravish deyiladi. Ravish morfologik jihatdan oʻzgarmas boʻlib, asosan feʼllarga bogʻlanib keladi, hol vazifasini bajaradi. Ravishlarning maʼno turlaridarn miqdor-daraja, sabab, maqsad ravishlarini ham nomuhim soʻzlar roʻyxatiga kiritish mumkin.

1	<i>Miqdor-daraja ravish ishtirok etgan gap</i>	Otabek koʻp ishlardi
2	<i>Miqdor-daraja ravish ishtirok etmagan gap</i>	Otabek ishlardi.

Yordamchi so'zlar turkumiga ko'makchi, bog'lovchi, yuklama kiradi. Yordamchi so'zlar mustaqil (leksik) ma'no bildirmaydi, faqat grammatik ma'no iodalash uchun xizmat qiladi.



2-jadval. Nomuhim so'zlar (yordamchi so'zlar)

Gapda vosita, maqsad, sabab, vaqt, makon kabi sintaktik munosabatlarni ifodalash uchun qo'llanadigan so'zlar ko'makchilar deb ataladi. Ko'makchi o'zi bog'lanib kelgan so'zdan keyin kelib, so'zlar orasidagi birgalik, payt, sabab, maqsad kabi munosabatni ifodalashga xizmat qiladi. Ko'makchi ikki xil bo'ladi: 1. Sof ko'makchi 2. Ko'makchi vazifasidagi so'z (funksional ko'makchi) [6].

1	<i>Sof ko'makchi ishtirok etgan gap</i>	Chapga buriling, keyin to'g'riga yuring
2	<i>Sof ko'makchi ishtirok etmagan gap</i>	Chapga buriling, to'g'riga yuring

Qo'shma gap tarkibidagi sodda gaplar va gap bo'laklarini bog'lash uchun xizmat qiladigan yordamchi so'zlar bog'lovchi deyiladi. Bog'lovchilar qo'llanishiga qarab yakka bog'lovchilar va takroriy bog'lovchilarga ajratiladi. Bog'lovchilar vazifasiga ko'ra esa ikki xil bo'ladi: teng bog'lovchilar; ergashtiruvchi bog'lovchilar [6].

1	<i>Yakka bog'lovchi ishtirok etgan gap</i>	Onam va singlimga gul oldim.
2	<i>Yakka bog'lovchi ishtirok etmagan gap</i>	Onam, singlimga gul oldim.

Xuddi shu gap takroriy bog'lovchi bilan kelganda:

1	<i>Takroriy bog'lovchi ishtirok etgan gap</i>	Ham onamga, ham singlimga gul oldim
2	<i>Takroriy bog'lovchi ishtirok etmagan gap</i>	Onam, singlimga gul oldim.

Matndan teng va ergashtiruvchi bog'lovchilar olib tashlansa, gapning semantik ma'nosi deyarli o'zgarmaydi.

Ayrim so'z yoki gapga qo'shimcha ma'no ifodalash uchun qo'llanadigan yordamchi so'z yuklama deyiladi. Yuklama tuzilishi jihatidan ikki xil: a) affiks holidagi yuklamalar:

-chi, -a (-ya), -da, -u (-yu), -oq (-yoq), -gina (-kina,-qina): b) so'z holidagi yuklamalar: axir, faqat, xuddi, nahotki, hatto, hattoki[6].

1	<i>So'z holidagi yuklama ishtirok etgan gap</i>	Axir , kelaman degandingiz
2	<i>So'z holidagi yuklama ishtirok etmagan gap</i>	Kelaman degandingiz.

Yuqoridagi misollardan ko'rinib turibdiki, yordamchi so'zlar nomuhim so'zlar bo'la oladi.

Fikrning voqelikka munosabatini bildirgan so'zlar modal so'zlar deyiladi. Modal so'zlar ma'no jihatdan quyidagi turlarga bo'linadi:

1. Fikrning aniqligini, realligini tasdiqlovchi modal so'zlar: albatta, shubhasiz, so'zsiz

2. Fikrning taxminligini, gumonligini, noaniqligini bildiruvchi modal so'zlar: ehtimol, chamasi, shekilli, aftidan, haytovur, chog'i.

3. Fikrning o'zaro munosabatini va tartibini bildiradi(yoki fikrning oldingi fikrga bog'liqligini bildiradi): demak, xususan, aksincha, masalan, jumladan, chunonchi,

4. Modallikni ifodalaydi: lozim, mumkin, shart, zarur.

5. Tasdiqni bildiradi: bor, mayli, xo'p, ha.

6. Inkorni bildiradi: yo'q, aksincha.

7. Afsuslanish, ajablanishni bildiradi: afsuski, attang, ajabo.

1	<i>Modal so'z ishtirok etgan gap</i>	Demak , ertaga natijalar aniq bo'ladi.
2	<i>Modal so'z ishtirok etmagan gap</i>	E, Ertaga natijalar aniq bo'ladi.

Demak, nomuhim so'zlar ro'yxatiga modal so'zlar ham kiradi. Shu jumladan, undov so'zlarni ham nomuhim so'zlar qatoriga kiritish mumkin.

His-hayajonni, xitobni bildirishga xizmat qiladigan soʻzlar undov soʻzlar deyiladi: undovlar quyidagi xususiyatlarga ega:

- 1) leksik (lugʻaviy) maʼno anglatmaydi;
 - 2) gap boʻlagi vazifasida kelmaydi va gap boʻlaklari bilan grammatik jihatdan bogʻlanmaydi;
 - 3) undovda ohang muhim oʻrin tutadi;
 - 4) kelishik, egalik, shaxs, zamon kabi grammatik maʼnolarni bildirmaydi.
- Masalan, undov soʻz gapda ishtirok etganda:

1	<i>Undov soʻz ishtirok etgan gap</i>	Oh , buncha shirin olma.
2	<i>Undov soʻz ishtirok etmagan gap</i>	Buncha shirin olma

Taqlid soʻzlar ikki xil:

1. Holatga (obrazga) taqlidni bildiruvchi soʻzlar: yalt, yarq, milt-milt, bilq, dir-dir.
 2. Tovushga taqlidni bildiruvchi soʻzlar: tars, turs, chars, gumbur-gumbur.
- Taqlid soʻzlarni ham nomuhim soʻzlar roʻyxatiga kiritish mumkin.

1	<i>Taqlid soʻz ishtirok etgan gap</i>	U dir-dir qaltirardi
2	<i>Taqlid soʻz ishtirok etmagan gap</i>	U qaltirardi

Yuqoridagi misollardan koʻrinib turibdiki, nomuhim soʻzlar gap tarkibini grammatik shakllantirishda ishlatilinish, semantik maʼno ahamiyati kam.

Nomuhim soʻzlarni hujjatdan olib tashlashning ijobiy va salbiy tomonlari bor. Ijobiy tomoni nomuhim soʻzlar matndan chuqur oʻrganish (deep learning) va mashinali oʻrganish (machine learning) modellarini qoʻllashdan oldin olib tashlanadi. Chunki nomuhim soʻzlar hujjatda tez-tez uchraydi va bu soʻzlarni tasniflash yoki klasterlash uchun ishlatilishi mumkin boʻlgan muhim maʼlumotlarni saqlamaydi va bu soʻzlar olib tashlanganda maʼlumotlar toʻplamining hajmi kamayadi va modelni qoʻllashga ketgan vaqt qisqaradi. Salbiy tomoni notoʻgʻri tanlangan va olib tashlangan nomuhim soʻzlar matn maʼnosini oʻzgartirishi mumkin. Har doim ham ular matnda olib tashlanmaydi. Nomuhim soʻzlarni olib tashlash matn ustida bajarayotgan vazifaga va erishmoqchi boʻlgan maqsadga bogʻliq. Misol uchun, agar his-tuygʻularni tahlil qilish (sentiment analiz) vazifasini bajara oladigan modellarda bunday soʻzlarini olib tashlash notoʻgʻri xulosaga olib keladi. Shuning uchun nomuhim soʻzlarini tanlashda ehtiyot boʻlish kerak. Quyida misol bilan tushintiriladi.

1	Siz bu ishni bugun qilishingiz shart emas .

2

Siz bu ishni bugun qilishingiz shart.

Agar bu gapda nomuhim soʻz sifatida “emas” soʻzi olib tashlansa, gapga teskari maʼno beradi. Yaʼni ish bugun qilinishi shart boʻlib qoladi. Gap mazmuni butunlay oʻzgarib ketadi.

Nomuhim soʻzlarning taʼsiri aniq boʻlsa-da, nomuhim soʻzlar roʻyxatini ishlab chiqish uchun aniq koʻrsatma yoʻq. Shuning uchun turli tizimlar turli tillar uchun turli oʻlchamdagi nomuhim soʻz roʻyxatidan foydalanadi. Ingliz tili uchun SMARTsystem 571 soʻzdan iborat nomuhim soʻz roʻyxatini taklif qildi, Fox 421 soʻz hajmini taklif qildi va DIALOG axborot xizmati kabi tijoriy tizim faqat 9 soʻzdan iborat kichikroq nomuhim soʻzlar roʻyxatidan foydalandi [Tijani O.D. Akinwale A.T. Onashoga S.A. Adeleke E.O, An auto generated approach of stop words using Aggregated analysis, Conference: 2017 Nigeria Computer Society International Conference]. Lekin ingliz tilida doimiy ravishda “a, the, and” [<https://gist.github.com/sebleier>] kabi soʻzlar foydalaniladi. Bundan tashqari rus tilida “с, а, аля” [<https://github.com/stopwords-iso/stopwords-ru>], turk tilida “belki, chunki, dahi” [<https://github.com/tkorkunkaya/Turkish-Stopwords>] kabi soʻzlar nomuhim soʻzlar sifatida githab platformasiga yuklangan. Oʻzbek tili uchun nomuhim soʻzlar roʻyxati Elov B., Hamroyeva Sh., Abdullayeva O., Xusainova Z. tomonidan ishlab chiqilgan boʻlib githab platformasiga [<https://github.com/Gulshahnoz/UzbekStopwords.git>] yuklandi. Nomuhim soʻzlar hujjatlarning konteksti yoki maʼlumotlariga hissa qoʻshmaydi va ular indekslash paytida, shuningdek, maʼlumot olish tizimi tomonidan soʻrov qilishdan oldin olib tashlanishi kerak. Biroq, turli hujjatlar toʻplamida yagona sobit nomuhim soʻzlar roʻyxatidan foydalanish qidiruv samaradorligiga zarar yetkazishi mumkin [Dolamic L. and Savoy J., When stopword lists make the difference. Journal of the American Society for Information Science and Technology, 61(1):200–203, 2010]. Chunki bir hujjatda nomuhim hisoblangan soʻz boshqa bir hujjatda ahamiyatli boʻlishi mumkin [Sankar S.S., Pal S., Effect of stopwords in Indian language IR, Indian Academy of Science, 2021. 25 August]

Nomuhim soʻzlarning xususiyatlari

1. Matnda juda kam qiymatiga ega boʻladi.
2. Matnda paydo boʻlish chastotasi yuqori boʻladi
3. Kamdan-kam hollarda soʻrov soʻzlari/qidiruv soʻzlari sifatida ishlatiladi, yaʼni maʼlumotlarni qidirishda bunday soʻzlardan foydalanilmaydi.
4. Kirish soʻzlar, olmosh, modal, taqlid, undov soʻzlar yordamchi soʻzlar: bogʻlovchi, koʻmakchi yuklama kabi soʻzlar boʻlishi mumkin.
5. Matn uchun umumiy soʻzlar boʻlib, muayyan sohada maxsus ishlatililmaydi.
6. Tilning qurilishi uchun zarurdir.

Xulosa. Nomuhim soʻzlar – bu hech qanday maʼlumot qiymatiga ega boʻlmagan va hujjatning katta qismini tashkil etuvchi matn tarkibidagi soʻzlar. Bunday soʻzlardan mashinali oʻqitish modellarida maʼlumotlar hajmini qisqartirish va modellar tez ishlashi uchun foydalanish mumkin. Matnlarni tasniflovchi

modellarda nomuhim soʻzlarni olib tashlash yaxshi natija beradi. Biroq his-tuygʻularga asoslangan modellarda bunday soʻzlarni olib tashlash notogʻri xulosaga olib kelishi mumkin.

Foydalanilgan adabiyotlar:

1. Dolamic L. and Savoy J., When stopword lists make the difference. Journal of the American Society for Information Science and Technology, 61(1):200–203, 2010
2. Madatov X.A., Sharipov M.S., Bekchanov Sh.K., Oʻzbek tili matnlaridagi nomuhim soʻzlar, <https://www.researchgate.net/publication/351951567>.
3. Elov B., Xusainova Z., Xudayberganov N. Oʻzbek tili korpusi matnlari uchun TF-IDF statistik koʻrsatkichni hisoblash. SCIENCE AND INNOVATION INTERNATIONAL SCIENTIFIC JOURNAL VOLUME 1 ISSUE 8 UIF-2022: 8.2 | ISSN: 2181-3337. 1774-1785.
1. Mahmudjonova G.U. TF-IDF asosida oʻzbek tili matnlarini tasniflash, Oʻzbek tilining milliy korpusi: muammolar va vazifalar mavzusidagi xalqaro ilmiy-amaliy anjumani, 2023.25.03
2. Sankar S.S., Pal S. Effect of stopwords in Indian language IR, Indian Academy of Science, 2021. 25 August
3. Sayfullayeva R., Mengliyev B. va boshqalar. Hozigi oʻzbek tili, Toshkent 2009, 345-bet
4. Tijani O.D. Akinwale A.T. Onashoga S.A. Adeleke E.O, An auto generated approach of stop words using Aggregated analysis, Conference: 2017 Nigeria Computer Society International Conference
5. <https://github.com/tkorkuncakaya/Turkish-Stopwords>
6. <https://gist.github.com/sebleier>
7. <https://github.com/stopwords-iso/stopwords-ru>
8. <https://github.com/Gulshahnoz/UzbekStopwords.git>

SODDA GAPLARDAGI MAZMUNY MOSLIK, O'XSHASHLIK VA SHAKLIY FARQLILIKNING LEKSIK VOSITALARI HAMDA ULARNING LINGVISTIK MODELLARI

Muratbekova Shoira Dilshodbek qizi

Alisher Navoiy nomidagi Toshkent davlat
o'zbek tili va adabiyoti universiteti
Kompyuter lingvistikasi mutaxassisligi
2-kurs magistranti

Annotatsiya. O'zbekistonda ilmiy-tadqiqotning rivojlanishi, original ilmiy matnlarning paydo bo'lishi, muayyan lingvistik nazorat (ko'chirmakashliklarning oldini olish uchun) tartibini ishlab chiqishni taqozo qiladi. Buning uchun elektron shaklda mavjud ilmiy matnlarning mazmuniy o'xshashlik darajasini aniqlaydigan dasturiy ta'minot yaratilishi zarur.

Kalit so'zlar: *model, lingvistik model, Levenshteyn masofasi, Shingle algoritmi, Jakkard algoritmi.*

Abstract. The development of research in Uzbekistan, the emergence of original scientific texts requires the development of a certain linguistic control (to prevent duplication). For this, it is necessary to create software that determines the level of similarity of the content of scientific texts available in electronic form.

Key words: *model, linguistic model, Levenstein distance, Shingle's algorithm, Jaccard's algorithm.*

Аннотация. Развитие научных исследований в Узбекистане, появление оригинальных научных текстов требует разработки специальной процедуры лингвистического контроля (для предотвращения плагиата). Для этого необходимо создать программное обеспечение, определяющее степень сходства содержания научных текстов, имеющих в электронном виде.

Ключевые слова: *модель, лингвистическая модель, расстояние Левенштейна, алгоритм Шингла, алгоритм Жаккара.*

Tilshunoslikda ham model tushuncha sifatida lingvist tomonidan tilning sun'iy ko'rishdagi yangidan tartiblashtirilishi, ya'ni odatdagi ko'rnishidan ko'ra anchagina soddalashtirilgan shaklga keltirilishi. Bu o'rinda lingvistik maqsadlarda tilning asl nusxasiga tegishli jihatlarni, so'zlar, gap qoliplari (xatti-harakatlarini takrorlaydi)ga taqlid qiladi.

Tilshunoslikda modellarning ko'plab ta'riflari mavjud:

- model – har qanday matn birliklarining (so'z, gaplar) turi, qolipi (til qolipi);
- model – til obyektlarini tavsiflash uchun belgilar, sxemalardir;

- model– qat’iy metatilga ega bo‘lgan tuzilishning rasmiylashtirilgan nazariyasi.

Tilshunoslikda modellashtirishning asosiy maqsadi shaxsning integral lingvistik qobiliyatini modellashtirish va bu orqali uning ishlarini tez, oson va qulay amalga oshirishga sharoit yaratish. Lingvistik model tushunchasi strukturaviy tilshunoslikda vujudga kelgan, ammo 60–70-yillarda ilmiy foydalanishga kirgan. Zamonaviy tilshunoslikda “model” atamasining mazmuni asosan “nazariya” atamasi nomi ostida Yelmslev[1] tomonidan ilgari surilgan. Model faqat yetarlicha aniq ifodalangan va yetarlicha rasmiylashtirilgan nazariya nomiga loyiq deb hisoblanadi[1]. Har bir model ideal holda, kompyuterda amalga oshirilishi kerak.

Tilshunoslikda “lingvistik model” tushunchasi struktur lingvistikaning ta’siri natijasida paydo bo‘lgan va keyinchalik kompyuter lingvistikasini paydo bo‘lishiga olib kelgan[2]. Bundan ma’lum bo‘ladiki, lingvistik model hodisa sifatida struktur tilshunoslik va kompyuter lingvistikalarini bog‘lab turadi. Modelni qurish nafaqat lingvistik hodisalarni aks ettirish vositalaridan biri, balki til haqidagi bilimlarning haqiqatini tekshirishning obyektiv amaliy mezon sifatida ham alohida ahamiyatli ishlardandir. Til o‘rganishning boshqa usullari bilan birlikda modellashtirish nutq faoliyatining yashirin mexanizmlari, uning nisbatan ibtidoiy modellardan tilning mohiyatini to‘liqroq ochib beradigan yanada mazmunli modellarga o‘tishi haqidagi bilimlarni chuqurlashtirish vositasi sifatida ishlaydi, desak mubolag‘a bo‘lmasa kerak.

Tilshunoslikda modellashtirish tizim sifatida o‘z prinsipiga ega va uning ba’zi quyi tizimlari boshqalarni modellashtiradi, masalan, yozma nutq tizimi og‘zaki nutq modelidir; yozma tilda biz bir nechta modellar (bosma, qo‘lda yozilgan) bilan ishlaymiz; ifoda rejasi kontent rejasi modeli.

Modellashtirish usuli odatda imo-ishora tizimlariga asoslanadi, lekin tilning o‘zi ham ishoralar tizimi, ya’ni so‘zlarni so‘zlar bilan modellashtirishdir. Har qanday model, shu jumladan lingvistik model ham rasmiy bo‘lishi kerak. Model, agar boshlang‘ich ma’lumot hamda matnlarni va ular bilan ishlash qoidalarini (yangi obyektlar va fikrlarni shakllantirish yoki joylashtirish qoidalari) aniq va ixcham ko‘rsata olsa, albatta ko‘zlangan maqsad hosil bo‘ladi hamda bu ko‘rinishida rasmiylik yuzaga keladi. Ideal holda, har qanday rasmiy model matematik tizim hisoblanadi. Shuning uchun, ma’lum ma’noda, rasmiyatchilik tushunchasi matematika, aniqlik yoki noaniqlik tushunchasiga teng bo‘lishi talab qilinadi. Rasmiylik, aniqlik, noaniqlik – bular nazariya taqdim etilgan tilning xususiyatlaridir. O‘z-o‘zidan bu xususiyat rasmiy nazariya bashoratlarining obyektiv eksperimental ma’lumotlar bilan mos kelishini ta’minlamaydi. Nazariyaning to‘g‘riligi uni tasdiqlash yoki rad etishga qodir bo‘lgan aniq tajribalarni o‘rnatishga imkon beradi, ammo nazariyaning aniqligi va haqiqati o‘rtasida zarur mantiqiy bog‘liqlik bor yoki yo‘q bo‘lishi amaliy tajribalarsiz xulosa berish mantiqan xato bo‘lgan hodisa. Rasmiy model eksperimental ma’lumotlar bilan u yoki bu talqin orqali bog‘lanadi. Modelni talqin qilish – bu model obyektlari (ramzlari) o‘rniga ma’lum bir mavzu sohasidagi



obyektlarni, masalan, tilni almashtirishning ehtimollik yoki qat'iy qoidalarini ko'rsatishni anglatishi bilan xarakterlanadi.

Ishimizning dastlabki qismlarida ko'chirmakashlik va uning oldini olish bilan bog'liq mavjud ahvolni, ya'ni mualliflik huquqining himoya qilinishining huquqiy asoslari yetarli bo'lsa-da, bu borada ilmiy yondashuvlarning ancha ortda qolayotganligini ta'kidlagan edik. Falsafada miqdor o'zgarishi sifat o'zgarishiga ta'sir qilishi qonuniyat sifatida qaralgani holda, O'zbekiston Respublikasi Mustaqilligiga o'ttiz yildan oshganiga qaramay, tadqiqotlarni amalga oshirish masalasida, ayniqsa tilshunoslikda jahon tilshunosligi bilan bellasha oladigan darajada dolzarb muammolar yechimi ko'zga tashlanmayapti. Albatta, qonun kuchli, biroq inson ehtiyojlari undanda kuchli, degan qarashga ko'ra, o'zganing ishini noqonuniy o'zlashtirish, ilmiy odobni chetlab o'tish jiddiy muammo bo'lib qolmoqda. Yuqorida ta'kidlaganimizdek qonunchilik nimani qilish mumkin, nimani qilish mumkin emas, taqiqlangan ishni qilib qo'lga tushganda jazolanishini amalga oshira oladi. Biroq mualliflik huquqini buzilishi ilmiy asoslab berilmasa, chuqur tahlil metod va vositalariga ega bo'lmasa, samarali natija kutish o'rinsiz va imkonsiz bo'lib qolaveradi. Bugungi kunda rivojlangan mamlakatlarda ham amalga oshirilayotgan ilmiy-tadqiqot ishlarining original matnligini aniqlash jiddiy muammo bo'lib turibdi. Oliy ta'lim muassasalari tomonidan ilmiy ishlarni internetga joylashtirib borish internetdan arzon va oson foydalangan holda bunday ishlarni topib olish bu boradagi muammolarni avj olishiga sabab bo'lmoqda. Albatta xorijda ilmiy matnlarning originalligini tekshiruvchi ko'plab tijoriy dasturlar borligini yuqorida ham ta'kidlagan edik. Biroq bunga qarshi undanda ko'proq tijoriy takliflar ham mavjudligi, xususan, “aqlli sinonimizatorlar” hamda dasturni aldash bilan bog'liq usullarning tinimsiz ishlab chiqarilishi ham jiddiy va dolzarb muammo. Internetdagi onlayn antiplagiat dasturlardan foydalangan holda o'zbek tilidagi ilmiy matnlarda sodda darak gaplarning mazmuniy o'xshashlik darajasini aniqlovchi plagiatga qarshi dastur uchun lingvistik baza tayyorlash hamda uning ishlash metodlarini belgilab olish bo'yicha o'zimizning taklif va tavsiyalarimizni ishlab chiqdik.

Kompyuter lingvistikasida lingvistik modul termini bugungi kunda muhim ahamiyat kasb etmoqda. Boisi tabiiy tilni kompyuter tiliga o'tkazilishi, ya'ni kompyuter tizimi vositasida matnga ishlov berish yo'llarini kashf etish kuzatilmoqda. Bu borada chet tillarining lingvistik dasturlari ishlab chiqilgan va bugungi kunda ular takomillashtirilmoqda. Lingvistik modul ana shunday lingvistik dasturlarning mustaqil tarkibiy qismlari, ya'ni dasturiy ta'minotning muayyan lingvistik jarayonini qamragan qismi hisoblanadi[2]. Aslida til nazariyasi ham mavjud tilning o'z xususiyatidan kelib chiqqan, uning o'ziga xos tomonlarini muayyan tartibga ko'ra tizimli ravishda foydalanish uchun tayyorlab ishlov berishdir. Ya'ni amaliyotdan nazariya kelib chiqqan, bugungi kunda nazariyadan amaliyotga mukammallashib qaytish hodisasi ancha faollashgan.

Matnlarda o'xshashlikni aniqlash bo'yicha *Jakkard o'lchovi* (floristik umumiylik ko'effitsiyenti, fransuz ko'effitsiyenti kommunikatsiya, nemis

olimi)[3]dan foydalaniladi (1901-yilda Pol Jakkard tomonidan taklif qilingan ikkilik o'xshashlik o'lchovidir).[1] :
$$K_J = \frac{c}{a + bc}$$
 bu yerda a – birinchi sinov maydonidagi turlarning soni, b – ikkinchi sinov maydonidagi turlarning soni, c – 1 va 2-saytlar uchun keng tarqalgan turlarning soni. Bu ma'lum bo'lgan birinchi o'xshashlik darajasi. Adabiyotda ko'effitsiyent muallifining familiyasi, shuningdek Jakkard deb berilgan. Jakkard ko'effitsiyenti turli xil modifikatsiyalar va yozuvlardagi ekologiya, geobotanika, molekulyar biologiya, bioinformatika, genomika, proteinomika, informatika va boshqa sohalarida faol qo'llaniladi. Jakkard o'lchovi Sorensen o'lchoviga va cheklangan to'plamlar uchun Sokal-Snit o'lchoviga teng.

O'zbek tilida bitta jumlani turli xil usulda, masalan so'zlarning joyini o'zgartirib yoki so'zlarni sinonimlar bilan almashtirib ifodalash mumkin. Ikki gapning o'xshashligini aniqlash zarurati kichik amaliy masalani hal qilishda paydo bo'lgan. Ko'effitsiyentlarni aniqlash uchun oddiy o'lchovlar va yig'ilgan statistikadan foydalaniladi. Qisqacha aytganda, vazifani quyidagicha ifodalash mumkin: “Yangi jumlar turli manbalarda ma'lum chastotada keladi. Bir xil fakt bo'yicha ikkita jumla bo'lmasligi uchun chiqish ma'lumotlarini filtrlash kerak.”

Ikkita jumlani taqqoslash

Ikkita qatorning o'xshashlik darajasini aniqlash muammosini hal qilishning bir necha usullari mavjud.

Levenshteyn masofasi[4]

Bir satrni boshqasiga aylantirish uchun qancha operatsiyalarni (qo'shish, o'chirish yoki almashtirish) bajarish kerakligini ko'rsatuvchi raqamni qaytaradi.

Xususiyatlari:

- oddiy dastur;
- so'z tartibiga bog'liq;
- chiquvchi natija - bu raqam;
- chiquvchi natijani biror narsa bilan solishtirish kerak.

Masalan, “*Men maktabda dars beraman.*” va “*Men universitetda dars beraman.*” Bu gaplarni bir-biriga moslashishi uchun qancha ko'p o'zgartirish qilsa foizi past, qancha kam o'zgarish bo'lsa foizi baland chiqadi. Ya'ni “*men*” o'zgarmadi, “*universitet*” o'zgardi, “*dars*” o'zgarmadi, “*beraman*” o'zgarmadi. Bunda 75 foiz ko'chirilgan deb beradi. Vaholanki, misol qilingan ikki jumla ma'no-mazmuni jihatidan bir-biridan keskin (universitetdagi o'qituvchining va universitetdagi ta'limning maktabdan, maktab o'qituvchisidan ko'plab farq qiladigan jihatlari mavjud) farq qiladi. “*Men maktabda dars beraman.*” va “*Men maktabda dars beraman*” desa umuman ikki jumlaning bir-biriga moslash operatsiyasi amalga oshirilmaydi va o'z-o'zidan 100 foiz ko'chirilgan deb qo'ya qoladi.

Shingle algoritmi[4]

U matnlarni shingillalarga (inglizcha - tarozilarga), ya'ni 10 so'zdan iborat zanjirlarga (kesishmalar bilan) ajratadi, shingillalarga xesh funksiyalarni qo'llaydi, matritsalarini oladi, ularni bir-biri bilan taqqoslaydi.

Xususiyatlari:

- algoritmnin amalga oshirish uchun matematik qismini batafsil o'rganish kerak;
- katta matnlar ustida ishlaydi;
- jumlar tartibiga bog'liq emas.

Bu algoritmda, satrlar, so'zlar solishtirib chiqiladi, so'z tartibining o'zgarishi ahamiyatga ega emas, ortiqcha so'zlarni soniga qarab foizni chiqarib beradi. “*Men maktabda ustozman.*” va “*Men maktabda o'qituvchiman.*” Bu o'rinda “ustoz” va “o'qituvchi”ning sinonim ekanligini bilmaydi va 70 foiz o'xshash deb chiqaradi.

Jakkard algoritmi

Bu algoritmda esa so'z emas, harflar taqqoslanadi. “*Men maktabda o'qituvchi.*” va “*Men maktab o'qituvchisi.*” Bu o'rinda –da va –si qo'shimchalari ortiqcha, 76 foiz o'xshash deb beradi. Bunga yana misol sifatida asr va asir so'zlarini 90 foiz o'xshash deb chiqarishini ta'kidlash orqali jiddiy kamchiliklar bor ekanligini e'tirof etamiz.

Mantiqiy vektorlarning tarkibiy qismlari, ya'ni faqat ikkita 0 va 1 qiymatlarini oladigan komponentlardan foydalanilganda, o'lchov Tanimoto koeffitsiyenti yoki kengaytirilgan Jakkard koeffitsiyenti sifatida tanilgan. Agar obyektlar turlarning paydo bo'lishi bilan taqqoslansa (ehtimollik talqini), ya'ni uchrashish ehtimoli hisobga olinadigan bo'lsa, u holda Jakkard o'lchovining analogi *Iversen ehtimollik* o'lchovi bo'ladi[4].

Bu usulda mosliknigina tekshirish imkoniyati mavjud, ya'ni so'zlar sinonimga o'zgartirilganda, gap bilan sintaktik aloqaga kirishmaydigan bo'laklar qo'shilganda mazmuniy o'xshashlikni aniqlashda murakkabliklar yuzaga keladi. Ya'ni, bu usulda faqatgina aynan o'xshash so'zlarga taqqoslanadi. Bizga esa, aynan o'xshash bo'lmagan so'zlarning semantik o'xshashligini aniqlash muhim hisoblanadi.

Xulosa o'rnida shuni aytish mumkinki, ilmiy matnlarning o'xshashlik darajasini aniqlovchi plagiatga qarshi dasturlar uchun gap bilan sintaktik aloqaga kirishmaydigan bo'laklarning maxsus ro'yxatini shakllantirish lozim. Ilmiy matnlardagi sodda darak gaplarning mazmuniy o'xshashlik darajasini aniqlashda so'z birikmalarida so'z va so'z birikmasi sinonimligi holatini modellashtirishda o'zbek tilining izohli va sinonimlar lug'atlaridan foydalanish maqsadga muvofiq[5].

Foydalanilgan adabiyotlar:

1. Hamroyeva Sh. O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Fil.fan. bo'yicha falsafa dokt. (PhD) diss. avtoref. – Qarshi, 2018.



2. Elov B., Hamroyeva Sh., Xusainova Z. NLP (tabiiy tilga ishlov berish)ning vazifalari vazamonaviy yondashuvlar. TerDU, FILOLOGIK TADQIQOTLAR: TIL, ADABIYOT, TA'LIM. 2022, 5-6.

1. <https://studopedia.info/2-73325.html>

2. <https://habr.com/ru/post/341148/>

3. Murtazayev A. Ilmiy matndagi sodda darak gaplarning mazmuniy o'xshashligi va lingvistik modellari. – Qo‘qon – 2021. – 133b.

UZBEK AUTOMATIC SPEECH RECOGNITION MODELS USING DEEP LEARNING TECHNIQUES

Salaeva Makhliyo

Master student at Urgench State University,
makhliyo.salaeva@urdu.uz

Kuriyozov Elmurod

Teacher at Urgench State University,
elmurod1202@urdu.uz

Salaev Ulugbek

Ph.D. Student at Urgench State University,
ulugbek.salaev@urdu.uz

Annotatsiya. O'zbek tili kabi resurslari kam bo'lgan tillar uchun nutqni avtomatik aniqlash (ASR: Automatic Speech Recognition) tizimlarini yaratish yuqori sifatli nutq ma'lumotlari yo'qligi sababli qiyin vazifa hisoblanadi. Ushbu maqolada biz audio kitoblardan ularning transkriptlari bilan ma'lumotlar to'plamini va o'zbek tili uchun ASR tizimini yaratish uchun ma'lumotlar manbasini yaratish uchun ochiq manba loyihasini oldik. Ma'lumotlar to'plamiga asoslanib, u nutqni aniqlash uchun ikkita Deep Learning modeli bo'yicha treyning o'tkazildi. Bizning tajribalarimiz shuni ko'rsatadiki, model ASR ko'rsatkichlarida 18,2% WERga erishgan.

Kalit so'zlar: *Nutqni avtomatik aniqlash, o'zbek tili, takrorlanuvchi neyron to'rlar, konvolyutsion neyron to'rlar.*

Abstract: Building Automatic Speech Recognition (ASR) systems for less-resourced languages such as Uzbek is a challenging task due to the lack of high-quality speech data. In this article, we obtained a dataset using from audiobooks with their transcripts and the open-source project to create source of data for building ASR systems for Uzbek language. Based on the dataset we performed training by two Deep Learning models for speech recognition. Our experiments show that the model achieved 18.2% WERs in ASR performance.

Keywords: *Automatic Speech Recognition, Uzbek language, Recurrent Neural Networks, Convolutional Neural Networks*

Аннотация: Создание систем автоматического распознавания речи (ASR) для менее ресурсоемких языков, таких как узбекский, является сложной задачей из-за отсутствия высококачественных речевых данных. В этой статье мы получили набор данных, используя аудиокниги с их транскрипциями и проект с открытым исходным кодом для создания источника данных для построения систем ASR для узбекского языка. На основе набора данных мы провели обучение по двум моделям глубокого обучения для распознавания речи. Наши эксперименты показывают, что модель достигла 18,2% частота ошибок слов (WER) в производительности ASR.

Ключевые слова: *Автоматическое распознавание речи, узбекский язык, рекуррентные нейронные сети, сверточные нейронные сети*

Introduction.

Linguistic resources in the field of natural language processing mainly correspond to widely spoken languages spoken by more than half of the world's population. The availability of language resources has enhanced the progress of Text-to-Speech (TTS), Automatic Speech Recognition (ASR), and AI-based systems. However, the development and research of technologies for low-resource natural languages requires the development of sufficient resources in these types of languages.

Automatic speech recognition refers to the task of recognizing human speech and presenting it in text form. Advanced technologies have been developed in this field of research in recent decades. ASR is typically seen in user-facing applications such as virtual agents, live captioning, and clinical notes, where accurate speech transcription is critical. ASR is an important component of speech AI, which is a set of technologies that help people talk to computers by voice. Recently, great progress has been made in the field of ASR using various deep learning approaches. Although technology giants have created advanced speech recognition engines for English, European and Asian languages, there is little research on the development of ASR systems such as Uzbek which is in the preliminary stage for creating language resources. This is due to the absence of a standard speech corpus of the Uzbek language, as well as dialectal differences [1]. A major challenge in developing an

ASR system is collecting sufficient speech corpus and data to train and test of the system. Other obstacle of building speech recognition models for Uzbek language is that it is an agglutinative language where the addition of numerous suffixes to the root of a word can create entirely new words[2], leading to a substantial increase in vocabulary size. With this problem in mind, a dataset consisting of approximately 11 hours of transcribed audio recordings of 16 speakers of different genders and ages was first collected. This data set is intended for the ASR task.

Related work.

There are several preliminary studies aimed at identifying speech in the context of the Uzbek language, of which [1] developed an initial Uzbek speech corpus consisting of 3500 sentences. A proposed ASR system based on HMM and DNN for the Uzbek language [3] allowed to obtain about 105 hours of transcribed speech data through the process of training the model on the basis of data consisting of 108,000 words. Similarly, the authors of [4] developed a system for recognizing spoken speech using 10 hours of transcribed audio. The works of [5] focus on voice command recognition systems based on a finite number of dictionaries.

It should be noted that the datasets used in these works are very limited and specialized for narrow application domains. Other existing Uzbek datasets are either

too expensive or not publicly available [6]. As a result, it limits the possibility of creating and testing an ASR system using a sufficiently large open-source Uzbek speech corpus. That is why it is of the first importance to develop an open-source Uzbek speech corpus of sufficient size.

Regarding the NLP research on Uzbek language, there has been a recent sharp increase in the development of NLP resources and tools specifically for Uzbek language. These include a tool for parts-of-speech tagging [7], removal of stopwords [8], Latin-Cyrillic transliteration [9], datasets for performing sentiment analysis [10], [11], stopwords [12], and text classification [13], as well as methodology for word game modeling [14].

Data collection.

Collecting high-quality speech data is crucial for building accurate speech recognition models. One approach to collecting speech data is to use platforms such as Mozilla Common Voice [15], which is a crowdsourcing project for collecting and validating speech data in multiple languages. Another approach is to collect data from audiobooks and their transcripts. Audiobooks are a good source of speech data as they provide a diverse range of speakers, speaking styles, and accents. Additionally, their transcripts provide a reliable source of text for aligning with the speech data, which is essential for training speech recognition models. By using a combination of these approaches, obtained diverse dataset that can improve the accuracy of their speech recognition models.

Methodology.

The ASR system aims to convert an input audio signal $x = (x_1, x_2, \dots, x_T)$ of certain length T into a sequence of words or characters (i.e., labels) $y = (y_1, y_2, \dots, y_N)$, $y_N \in V$ from a specific vocabulary V . These labels can be either words or characters. The output sequence is determined by finding the most probable sequence of labels $\hat{y} = \arg \max_{y \in V} p(y|x)$. The ASR system typically

involves several steps such as pre-processing, feature extraction, classification, and language modeling (Figure 1).

The initial step of ASR is pre-processing which focuses on enhancing the quality of the audio signal by eliminating noise and filtering the signal. Additionally, in ASR, feature extraction is a critical step, where different techniques are employed to generate a set of coefficients or values. It is necessary for this step to be resistant to quality-related factors like noise or the echo effect. The objective of the classification model is to identify the spoken text from the input signal. It takes the pre-processed features as input and produces the corresponding output text. The objective of the classification model is to identify and interpret the spoken language in the input signal. After extracting the relevant features in the pre-processing stage, the model produces the corresponding text as output. On the other hand, the language

model (LM) plays a crucial role in capturing the semantic information and grammatical rules of a language. LM is vital for recognizing the output tokens generated by the classification model and for correcting the output text.

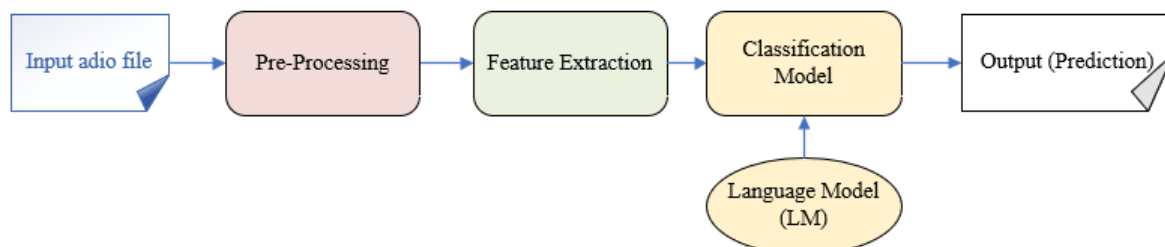


Figure 1. General overview of ASR system process.

Recurrent Neural Networks (RNNs) have been widely used in Automatic Speech Recognition (ASR) tasks due to their ability to model sequential data. Unlike feedforward neural networks, RNNs can capture the temporal dependencies between inputs by passing information from one time step to the next through a hidden state. This makes RNNs effective in modeling the dynamic nature of speech signals, which often exhibit complex temporal patterns.

In ASR, RNNs are typically used as acoustic models, taking in a sequence of acoustic features and outputting a sequence of probabilities over phoneme, grapheme, or word labels. One popular variant of RNNs is the Long Short-Term Memory (LSTM) network, which has been shown to effectively capture long-term dependencies in speech signals. However, RNNs are also known to have some limitations, such as difficulty in capturing long-term dependencies in sequences and the vanishing gradient problem. This has led to the development of other neural network architectures such as Convolutional Neural Networks (CNNs) and more recently, Transformer Networks, which have shown significant improvement in ASR tasks.

The series of experiments conducted on the ASR dataset using RNNs and CNNs models. After training the models, they evaluated their performance based on the word error rate (WER) metric. The results showed that the RNNs model achieved a WER of 21.6%, while the CNNs model achieved a WER of 23.5%. The lower WER score of the RNNs model suggests that it was more accurate in transcribing speech to text (Table 1). The experiments demonstrate the potential of both RNNs and CNNs models in ASR applications and highlight the importance of selecting the appropriate model for the specific task.

Table 1. Training results on the test dataset

Model Name	WER	Number of Layers	Number of Neurons	Batch size	Activation function
RNNs	21.6%	4	512	64	ReLU
CNNs	23.5%	5	256	32	Tanh

Conclusion.

In conclusion, we have developed a high-quality but smaller Uzbek speech corpus containing transcribed audio recordings spoken by numerous speakers. The corpus has been thoroughly checked by native speakers to ensure accuracy and reliability. Our ASR experiments demonstrated the effectiveness of both the RNNs and CNNs in transcribing Uzbek speech and model trained on the Uzbek speech corpus achieved 19.4% and 18.2% WERs on the validation and test sets respectively, using both RNNs and CNNs. We plan to continue increasing the size and quality of our dataset and conduct further ASR experiments by using Deep Learning approaches in the future. Our findings suggest that our Uzbek speech corpus and ASR models can serve as a valuable resource for future research in the field of Uzbek speech processing.

References

1. M. Musaev, I. Khujayorov, and M. Ochilov, “Automatic Recognition of Uzbek Speech Based on Integrated Neural Networks,” in *Advances in Intelligent Systems and Computing*, 2021. doi: 10.1007/978-3-030-68004-6_28.
2. A. Mukhamadiyev, I. Khujayorov, O. Djuraev, and J. Cho, “Automatic Speech Recognition Method Based on Deep Learning Approaches for Uzbek Language,” *Sensors*, vol. 22, no. 10, 2022, doi: 10.3390/s22103683.
3. M. Musaev, S. Mussakhojayeva, I. Khujayorov, Y. Khassanov, M. Ochilov, and H. Atakan Varol, “USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021. doi: 10.1007/978-3-030-87802-3_40.
4. M. Musaev, I. Khujayorov, and M. Ochilov, “Development of integral model of speech recognition system for Uzbek language,” in *14th IEEE International Conference on Application of Information and Communication Technologies, AICT 2020 - Proceedings*, 2020. doi: 10.1109/AICT50176.2020.9368719.
5. M. Musaev, I. Khujayorov, and M. Ochilov, “The Use of Neural Networks to Improve the Recognition Accuracy of Explosive and Unvoiced Phonemes in Uzbek Language,” in *2020 Information Communication Technologies Conference, ICTC 2020*, 2020. doi: 10.1109/ICTC49638.2020.9123309.
6. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” *arXiv preprint arXiv:2205.10078*, 2022.
7. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.



8. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.
9. U. Salaev, E. Kuriyozov, and C. Gómez-Rodriguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com
10. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com
11. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodriguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243.
12. X. Madatov, M. Sharipov, and S. Bekchanov, “O‘ZBEK TILI MATNLARIDAGI NOMUHIM SO‘ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021.
13. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.
14. J. Mattiev, U. Salaev, and B. Kavsek, “Word Game Modeling Using Character-Level N-Gram and Statistics,” *Mathematics*, vol. 11, no. 6, p. 1380, 2023.
15. R. Ardila *et al.*, “Common voice: A massively-multilingual speech corpus,” in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020.

O'ZBEK TILINING FORMAL GRAMMATIKASINI YARATISH

UDK: 811.512.133`42

O'ZBEK TILIDA SODDA GAP STRUKTURALARI

Jurayeva Nilufar Vayitovna

Geologiya fanlari Universiteti dotsenti

mina-uzb@mail.ru

Atayeva Dildora Abdullayevna

UrDU magistr

datayeva19910202@gmail.com

Annotatsiya. Maqolada o'zbek tilidagi sodda gaplar strukturalari va gap bo'laklarining formal modellari keltirilgan. Keltirilgan strukturalar asosida sodda gaplar strukturalarining elektron bazasi ishlab chiqildi.

Kalit so'zlar: kompyuter lingvistikasi, sodda gap strukturalari, ega, kesim, ikkinchi darajali bo'laklar.

Annotation. The article presents structures of simple sentences and formal models of sentence fragments in the Uzbek language. Based on the mentioned structures, an electronic database of simple sentence structures was developed.

Аннотация. В статье представлены конструкции простых предложений и формальные модели фрагментов предложений в узбекском языке. На основе указанных структур была разработана электронная база данных структур простых предложений.

Bugungi kunda o'zbek tilida sodda gaplarni sintaktik tahlil qilish uchun dasturlar hali to'liq ishlab chiqilmagan, shuning uchun bu masalani yechish muhimdir.

Hozirgi kunda ilm fandagi noqulayliklarni bartaraf etish uchun juda katta ishlar olib borilmoqda, ayniqsa, axborot almashinuvi daqiqalar ichida millionlab kuzatilayotgan paytda uni qayta ishlash va tahrir qilish asosiy ishlardan biridir.

Har bir til asrlar davomida rivojlanib, boyib boradi. Jahon miqyosida o'zbek tilini dunyoviy tillar darajasiga olib chiqish hozirda eng muhim vazifalardan biridir. Buni amalga oshirishda, o'zbek kompyuter lingvistikasini yanada rivojlanishiga e'tibor qaratilgan.

“Kompyuter lingvistikasi” fanining maqsadi, tabiiy tillarning matematik modellari qurish, lingvistik muammolarni hal qiluvchi kompyuter dasturlarini ishlab chiqishga yo'naltirilgan. Modellar asosida matnlarni tahrirlash, kompyuter tarjimai, kompyuter lug'atlari, tillarni o'rgatish hamda bilimlarni baholash bo'yicha kompyuter dasturlarini ishlab chiqish mumkin bo'ladi. Tabiiy tilni qayta ishlash (NLP)dan maqsad kompyuterlarga strukturlangan matnlarni tahlil qilish va

tushuntirish. Tabiiy tilni qayta ishlash(NLP) sun'iy intellektning bir bo'lagi bo'lib, uning maqsadi kompyuterlar va insonlar o'rtasidagi o'zaro aloqalarni o'rnatish.

Ushbu maqolada gap bo'laklari, gap strukturasi va turlarini o'rganadigan grammatika qismi sintaksislarini ko'rib chiqamiz.

O'zbek tilida sodda gap strukturasi ko'rinishi.

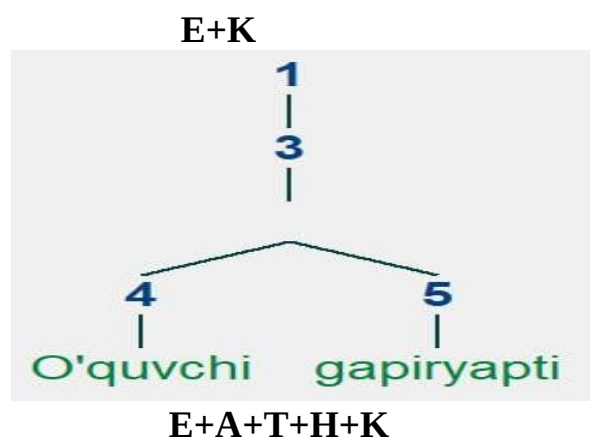
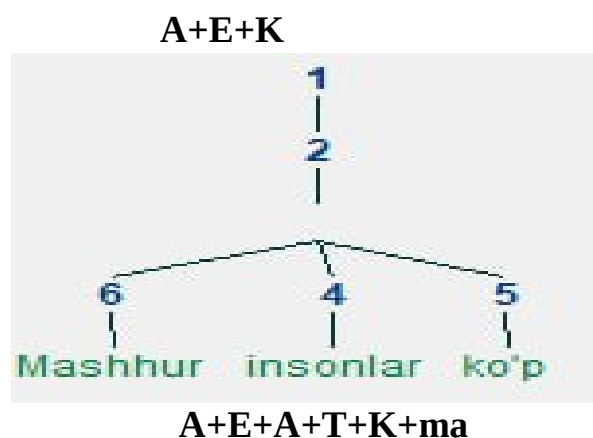
EGA + ATH + KESIM

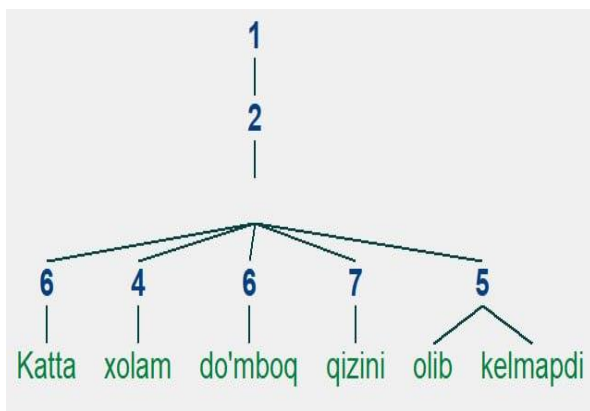
Bu yerda **ATH** gapning ikkinchi darajali bo'laklari, ya'ni **A**-aniqlovchi, **T**-to'ldiruvchi, **H**-hol

A+E+K	Qiziqarli kitoblar keldi.
E+A+K +mi?	Aziza a'lochi o'quvchimi?
E+T+H+K	Men she'rni ifodali o'qidim.
A+E+A+T+K+ma	Katta xolam do'mboq qizini olib kelmabdi.
E+T+K+ma+mi	Dadam sovg'a bermadimi?
H+A+T+E+K	Darsda o'sha domlaga men savol berdim.
E+H+K	Oyim darrov keldi.
H+A+E+K	Qishda kuchli qor yog'adi.
A+E+T+K+ma+mi	Mening fikrim sizni qiziqtirmadimi?
E+A+T+K+ma+mi	Domla yangi mavzuni o'tmadimi?
A+E+A+T+K+ma+mi	Katta opam bu chelakni ko'rmadimi?
H+A+E+A+T+K	Bugun kichik singlim bechora itni uchratib qolibdi.

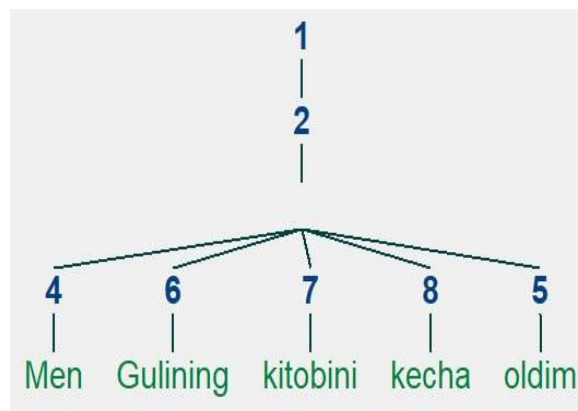
Yuqoridagi sturukturalar asosidagi gaplarni daraxt ko'rinishida tasvirlash mumkin.

Bu yerda 1-Sodda gap, 2-Sodda yoyiq gap, 3-Sodda yig'iq gap, 4-Ega, 5-Kesim, 6-Aniqlovchi, 7-to'ldiruvchi, 8-Hol. Bu belgilashlar orqali daraxtlar quyidagicha ifodalandi.

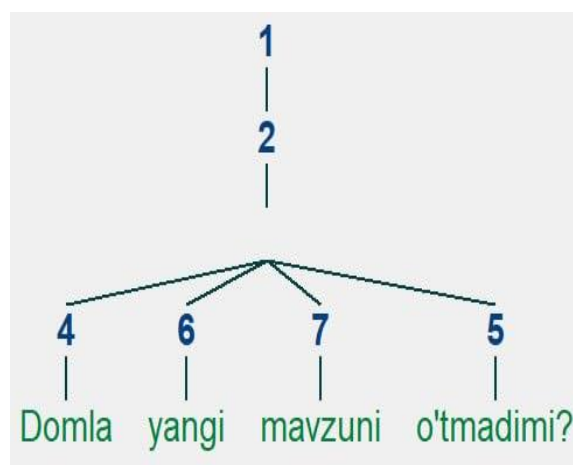
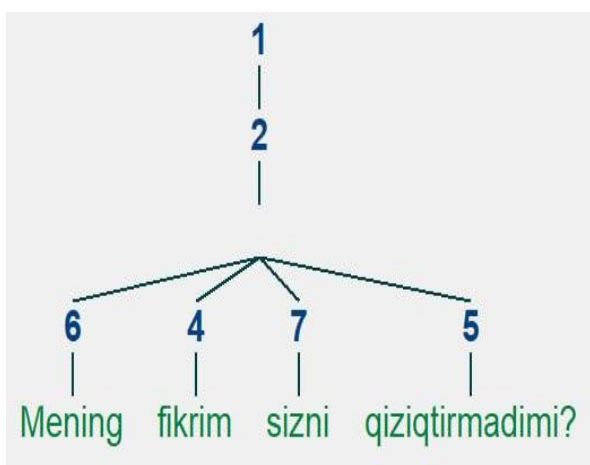




A+E+T+K+ma+mi?



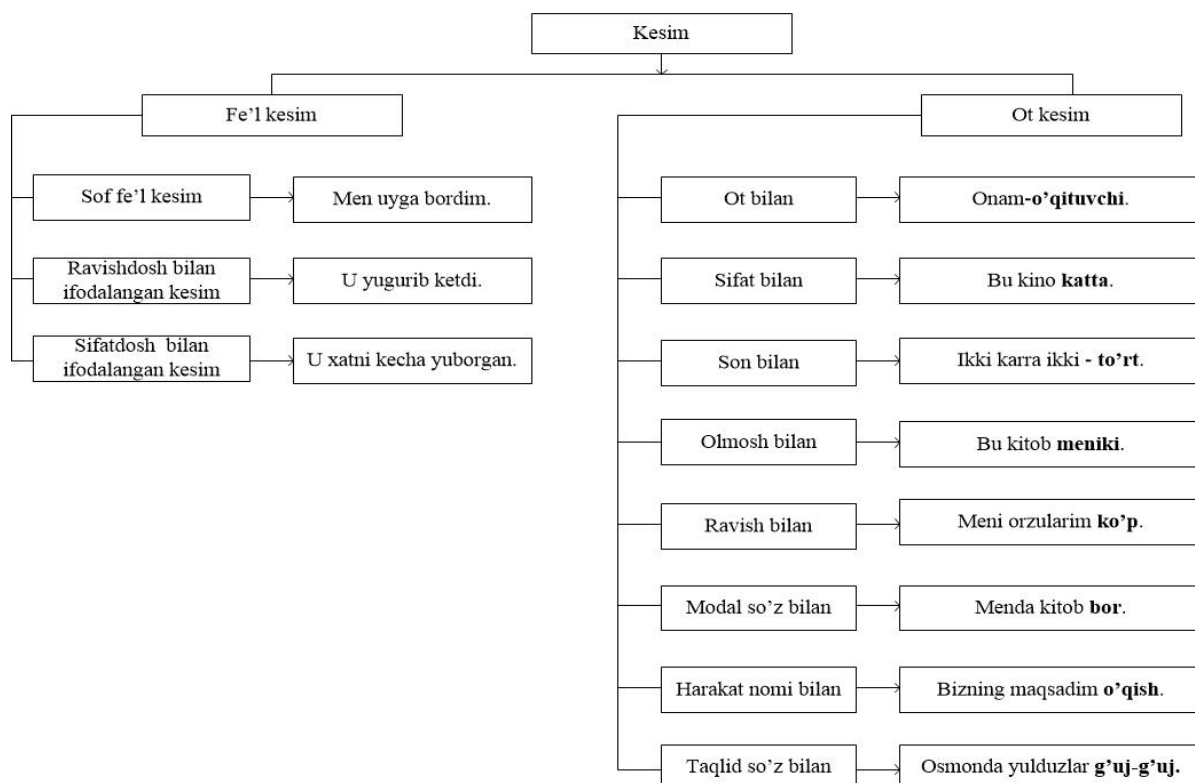
E+A+T+K+mi?



Gap fikr, tuyg'u va istakni bildiruvchi so'zlar uyushmasi ma'lum darajada shakllangan vositadir. O'zbek tilida gap bo'laklari 5 ta: ega, kesim, aniqlovchi, to'ldiruvchi va hol.

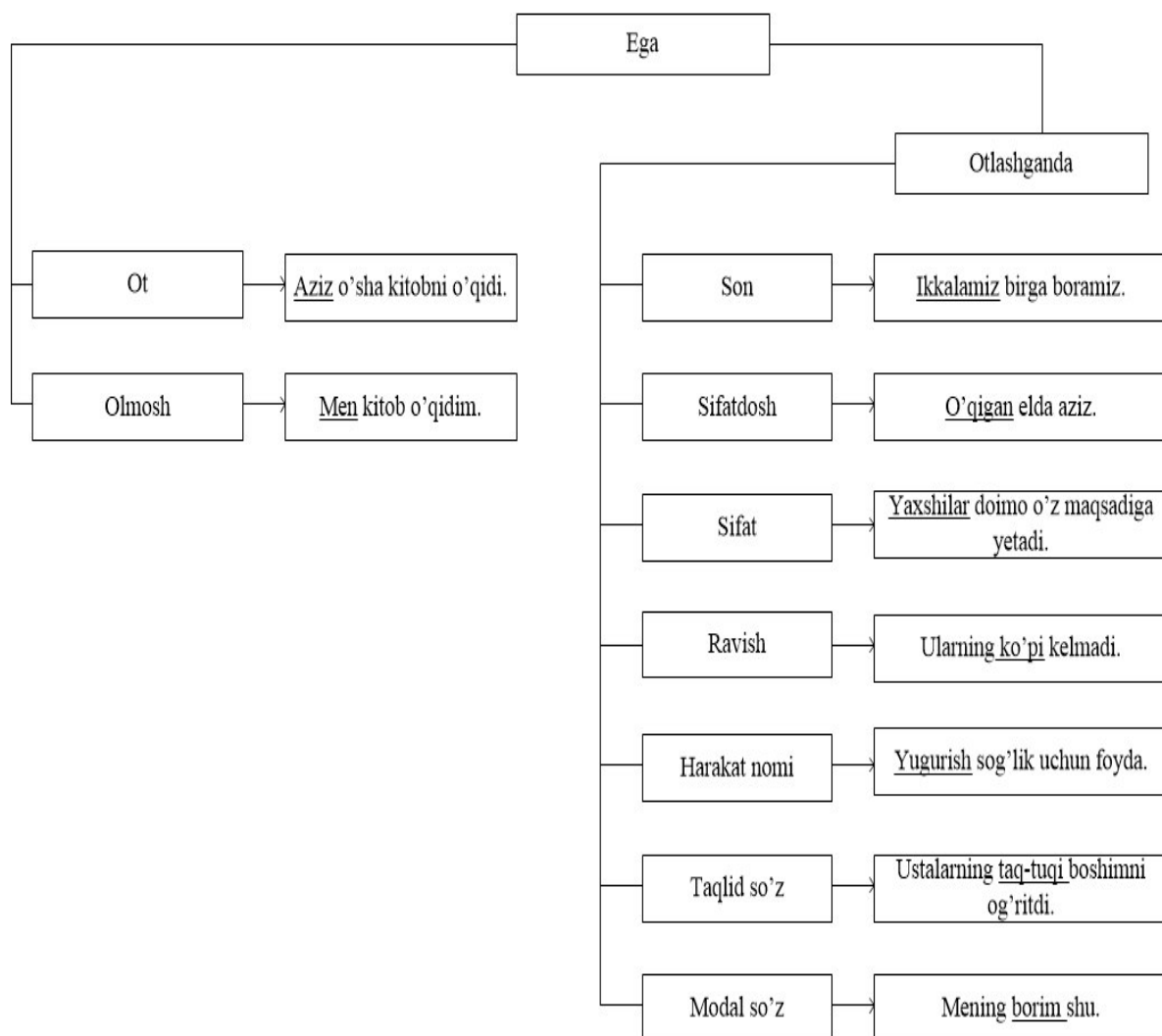
Gapning gap sifatida shakllanishida muhim bo'lgan, uning mazmuniy va grammatik markazi hisoblanuvchi bo'lak-kesim. Shu sababli u gapning eng muhim bo'lagi, bosh bo'lak hisoblanadi. Kesim gap markazi bo'lib, u tasdiq-inkor, zamon, mayl, shaxs-son ma'nolarini ifodalab keladi, hukmni ifodalaydi va gapning boshqa bo'laklarini o'z atrofida birlashtiradi. Kesim qaysi turkum so'zlari bilan ifodalanishiga ko'ra ikki xil: fe'l kesim, ot kesim. Fe'lning turli shakllari orqali ifodalanган kesim **fe'l kesim** deyiladi.

Fe'l kesimlar quyidagi so'z turkumlari bilan ifodalanadi ega:



Ega kesimga ergashib kesimda ifodalangan belgi, harakat-holatning bajaruvchisini ko'rsatib, kesimdan keyingi bo'lak hisoblanadi. Gap kesimining qo'shimchalaridan anglashilib turgan shaxs-son ma'nosini aniqlashtiruvchi bo'lak egadir.

Ega quyidagi so'z turkumlari bilan ifodalanadi.



Gapning kesimi yoki egasiga tobelanib, ularga ergashib keladigan bo'laklar ikkinchi darajali bo'laklar deyiladi. Ikkinchi darajali bo'laklarga aniqllovchi, to'ldiruvchi, hol kiradi. Ular fe'l, ot, sifat, son, ravish, taqlid so'zlar, olmoshlar bilan ifodalanishi mumkin.

Hol kesim ifodalagan ish-harakat, holat, voqea-hodisaning o'rnini, payti, miqdor-darajasi, sababi, maqsadi, bajarilish sharti, qurol vositasi, qanday vaziyatda bajarilgani kabi belgi-xususiyatlarini ifodalab keladi. Hol hamisha kesimga tobelanadi.

Hollar ravish so'zlar, kelishik shaklidagi va ko'makchili otlar bilan, sifat, son, taqlid so'z va undovlar bilan, fe'ning xoslangan shakllari bilan ifodalanishi mumkin. Hollar o'zi bog'lanib kelgan hokim bo'lakning qanday belgisini bildirishiga ko'ra quyidagi turlarga bo'linadi: 1. Ravish holi (vaziyat holi); 2. Daraja-miqdor holi; 3. O'rin holi; 4. Payt holi; 5. Sabab holi; 6. Maqsad holi.

1. Eshik birdan ochildi.	E+H+K
2. Bu gapni ikkinchi aytman.	A+T+H+K
3. Bolalar maktab tomon otlanishdi.	E+H+K

4. **Ertaga** ta'til boshlanadi. **H+E+K**

5. O'yinni **noiloj** tugatdik. **E+H+K**

6. U o'qigani ketdi. **E+H+K**

To'ldiruvchi fe'lga bog'lanib kelib predmetlik manosini ifodalaydi va u ikkiga bo'linadi:

1. Vositasiz to'ldiruvchi. Men **ukamni** uxlatdim. **E+T+K.**

2. Vositali to'ldiruvchi. Akam **mendan** katta. **E+T+K.**

Aniqlovchi hokim bo'lakdan anglashilgan narsaning qandayligi, kimga yoki nimaga qarashli ekanini bildiradi. Aniqlovchi fe'l-kesimga bog'lanmaydi. Narsaning qanday belgisini aniqlashiga ko'ra aniqlovchi uchga bo'linadi. 1. Sifatlovchi aniqlovchi. 2. Qaratqich aniqlovchi 3. Izohlovchi aniqlovchi.

Sodda gaplar strukturalari hozirda 64 tani tashkil qildi va ularga mos misollar keltirildi. Python dasturlash tilida bu strukturalarni egasi, kesimi va ikkinchi darajali bo'laklarini topish dasturi tuziladi.

Foydalanilgan adabiyotlar

1. Erkaboyeva N. O'zbek tilidan ma'ruzalar to'plami. - Toshkent "Akademnashr", 2016

2. Hamroyeva O. Ona tili. Maruzalar to'plami. – Toshkent, "Akademnashr", 2021.

3. Ismoilova S.A., Botirboyev F.P., Jurayeva N.V.. Eganing formal modelarning kompyuter bazasi.

4. Sharipov M., Kuriyozov E., Yuldashev O., Sobirov O. "UzbekTagger: The rule-based POS tagger for Uzbek language," arxiv preprint arxiv: 2301.12711, 2023.

5. Jurayeva N.V. Formal nazariyada o'zbek tilidagi ot so'z turkumining shakllari. Образование и наука в XXI Веке.

6. Botirboyev F.P., Jurayeva N.V. O'zbek tilidagi qaratqich aniqlovchining formal modellari. "Экономика и социум" №4 (83), 2021.

7. Jurayeva N.V. Qarshiboyeva N.A. NG harf birikmasini bo'g'inlarga ajratuvchi va izoh beruvchi kompyuter dastur. "SCIENTIFIC PROGRESS" Scientific Journal ISSN: 2181-1601 Volume: 1, ISSUE: 3

8. Jurayeva N., Xasanova C.X. Formal gramatika nazaryasi "Kompyuter lingvistikasi: muammolar, yechim, istiqbollar" Respublika ilmiy-texnikaviy konferensiya Vol. 1 №. 01 (2021)

9. Sharipov M., Mattiev J., Sobirov J., Baltayev R. "Creating a morphological and syntactic tagged corpus for the Uzbek language," arxiv preprint arxiv:2210.15234, 2022.

10. Rahimov A. Kompyuter lingvistikasi asoslari. "Akademnashr", 2011-yil.

ӨЗГЕ ТІЛДІ МЕКТЕПТЕ ҚАЗАҚ ТІЛІНЕН ОРЫНДАЛАТЫН ЖАТТЫҒУЛАРДЫҢ ЛИНГВИСТИКАЛЫҚ АСПЕКТІЛЕРІ

Есенова Гулшат Сериковна

Қ. Жұбанов атындағы Ақтөбе өңірлік
университетінің аға оқытушысы,
гуманитарлық ғылымдар магистрі

Резюме. В данной статье рассматриваются лингвистические аспекты упражнения при изучении казахского языка в школах с русским языком обучения.

Resume. This article devotes to linguistic aspects of exercises in learning Kazakh in Russian language schools

Қазіргі таңда тіл үйренуші қазақ тілін қоғамдық-әлеуметтік ортаға қатысты қарым-қатынас құралы ретінде игереді. Қазақ тілін коммуникативтік тұрғыдан қатынас құралы ретінде игеру – оның қоғамдық-әлеуметтік қызметін білуді және бағалауды көздейді. Осы жағынан алғанда қазақ тілді оқытуда тұлғаның сөйлеу қабілетін арттыратын, танымдық көзқарасын қалыптастыратын, өз ойын еркін әрі түсінікті, жүйелі түрде жеткізуге көмектесетін білім алғаны жөн. Сондықтан оқушының қазақ тілді жетік меңгерген, жан-жақты дамыған жеке тұлға болып тәрбиеленуі – қазақ тілі пәнінің мақсатын айқындауы қажет. Осы тұрғыда тілді, оның ішінде қазақ тілін оқытудың сапасын арттыру жайының лингвистикалық негізі қазақ тіл біліміне байланысты екені баршамызға белгілі.

Қазақ тілін оқыту әдістемесінің теориялық негіздерінің бірі – тіл білімдік немесе лингвистикалық негіз. Олай болса, оның негізгі нысаны да қазақ тілінің жалпы заңдылықтарымен таныстыру. Қазақ тілі – бай сөздік қоры бар, өзіндік дыбыс құрамы, сөзжасам мен сөздерді түрлендіру жүйесі бар, сөз тіркестері мен сөйлем жүйесі бар, солар арқылы сөйлетуді қамтамасыз ететін құбылыс - деп қарасақ, оны екінші тіл ретінде басқа ұлт өкілдеріне оқытудың өзіндік ерекшеліктері мен әдіс-тәсілдері де бары анық. Бұндай әдіс-тәсілдерді қарастыратын, ретке келтіретін қазақ тілін оқыту әдістемесі ғылымы. Қазақ тілін оқыту әдістемесі қазақ тіл білімі оқытатын тіл салаларын, тілдік материалдарды негізге ала отырып, өзара тығыз байланыста оқытуды белгілейді.

Тілдік материалдың байланысын сабақта жақсы, дұрыс ұйымдастыру тілді сапалы да тиімді үйретуге әкеледі.

Лингвистика тіл туралы ғылым, ал қазақ тілін оқытудың сапасын арттыру әдістемесі қазақ тілінің заңдылықтарын оқыту мен үйретудің жаңа тиімді жолдарын қарастыратыны белгілі [1.6].

Профессор Н.Оралбаева жоғары оқу орнына арнап жазған оқулығында «Қазақ тілін орыс мектептерінде оқыту әдістемесінің мақсаты- мұғалімдерді қазақ тілін орыс мектебінде оқытудың ғылыми негізімен қаруландыру, қазақ тілін екінші тіл ретінде өзіндік ерекшелігін меңгерту, қазақ тілін орыс мектебінде оқытудың түрлі әдіс-тәсілдері, тиімді жолдарымен таныстыру, оқушыларды қазақша сөйлеуге дағдыландыруда оларды пайдалана білуге бейімдеу, бағыттау» деген болатын.

Сөйтіп, қазақ тілін оқыту әдістемесі қазақ тіл білімі оқытатын тіл салаларын, тілдік материалдарды негізге ала отырып, өзара тығыз байланыста оқытуды белгілейді. Тілдік материалдардың байланысын сабақта жақсы ұйымдастыру – тілді сапалы үйретуге негіз болады.

Қазақ тілінің өзіне тән заңдылықтары мен грамматикалық ережелерін үйренбей, тіл үйренуші өз ойын басқаларға жеткізе алмайды. Сондай-ақ қазақ тілінде сөйлесе де алмайды. Осы орайда Қ.Жақсылықова қазақ тілін оқытудың сапасын арттыру әдістемесінің лингвистикалық негізін анықтауда алдымен мынадай екі мәселеге көңіл бөлген жөн дейді:

«1. Қазақ тілінің жүйесіне, қазақ тіл білімінің негізгі заңдылықтарына сүйену;

2. Қазақ және орыс тілдері жүйесін салыстыра талдау мәліметін ескеру» [2.33].

Бұл жерде яғни қазақ тілінің негізгі заңдылықтарын үйретуде жаттығу орындалады. Айта кететін жайт:

Профессор Н.Оралбаева мен К.Жақсылықовалардың оқулықтарында тілдік жаттығулар мен сөйлеу жаттығулары деп топтастырылуын көптеген зерттеушілер құптайды.

Ж.Ноғайбаеваның «Дыбыс үндестігі арқылы орыс бөлім студенттерін қазақша сөйлеуге үйрету әдістемесі» атты диссертациялық жұмысында «Дыбыс үндестігін сөйлеуде дұрыс қолдану дағдысын қалыптастыратын жаттығулар жүйесі» атты тақырып берілген. Осы бөлімде зерттеуші әдіскерлердің еңбектеріндегі (В.П.Успенский, Н.И.Гез, И.В.Рахманов, И.А.Грузинская, В.С.Цетлин, А.Күзембаева, Б.Мешінбаевалардың) жаттығу жұмыстарын жүйелеп талдаған. Олардың әрқайсысының жаттығулар топтастыруда ерекшеліктері де бар екендігі айтылады.

Зерттеуші өз тақырыбы бойынша ғылыми деректерге, тәжірибелі оқытушылар пікірлеріне сүйене отырып, тілді үйрету жаттығуларына қойылатын талаптарды былайша анықтайды:

1. Жаттығуларда сөйлеу мақсатының көзделуі;
2. Жаттығуда сөйлеу түрлерінің толық қамтылуы;
3. Жаттығуларды орындатудағы қиындықтың ерекшеліктері;
4. Жаттығуда дыбыс үндестігінің заңдылығын қайталау толық қамтылуы, онымен байланысты болуы;

5. Жаттығу мазмұнының қазақ халқының мәдениеті, өнері дәстүрімен байланысты болуы;
6. Жаттығулардың мазмұны дұрыс, қызықты, түсінікті болуы;
7. Жаттығулардың мазмұны тіл үйренушілердің өміріне, мамандығына байланысты болуы көзделеді [3. 59-60].

Қазақ тілі материалдарына дағдыландыруда тіл материалдарының өзіне тән белгілерін қамтитын дағды түрлері жүзеге асырылады. Бұндай жұмыс процесінде оқушыларда толық түсінік қалыптасады.

Ә.Исабаев тіл үйренушілердің дағдысын қалыптастыруда жұмыстың мынадай түрлері орындалады дейді:

- 1) мұғалімнің дайындаған карточкалары іске асырылады;
- 2) оқулықтағы жаттығулардың шарттары орындалады;
- 3) схемалар мен кестелер жасалады;
- 4) ауызша мысалдар келтіріледі;
- 5) ауызша сөйлемдер құрастырылады;
- 6) мәтінге лайықты шағын шығармалар жазылады;
- 7) картина мен сурет бойынша сөйлемдер құрастырылады;
- 8) орфографиялық сөздіктер бойынша жұмыс жүргізіледі;
- 9) жергілікті материалдар пайдаланылады.

Зерттеуші жаттығу жұмыстарын орындау барысында мына төмендегідей дағдылар жетіледі дейді:

- 1) алған білі еске түсіріледі;
- 2) теориядан алған білімі бекітіледі;
- 3) көру қабілеті нығаяды;
- 4) есте сақтау қабілеті артады;
- 5) байымдау қабілеті артады;
- 7) сенімділік қабілеті күшейеді;
- 8) байқау қабілеті дамиды [4.32].

Яғни, білім беру әдістері жаттығу орындату кезінде студенттердің қазақ тілінің теориялық материалдарын психикалық функциялары негізінде меңгеруіне, оны тіл үйренушілердің ойында қалыптастыруды және практикалық жағынан шебер қолдана алуына мүмкіндік туғызады.

А.Күзембаева жаттығулар жалпы мынадай талаптарға сай жүргізілуі тиіс дейді:

1. Жаттығуларды орындайтындай білім болу керек;
2. Жаттығуды не үшін орындайтынына көзін жеткізу арқылы сапалы орындату;
3. Жаттығу оқушылардың қызығуын, ынтасын арттыруға тиіс;
4. Жаттығулар кездейсоқ болмай, белгілі бір жүйемен орындалып отыруы шарт;
5. Таяуда өтілген материалдарға ғана негізделмей, бұрын өтілген кейбір материалдарды да қамтуы қажет;

6. Бір жаттығу ұзақ уақытқа созылмауға тиіс;
7. Шығармашылық жаттығулар көбірек болуы шарт;
8. Әр жаттығу оқушы еңбегінің нәтижесін көрсететіндей болу керек;
9. Жаттығулар оқушы ойына әсер етуі жағынан аналитикалық-синтетикалық болғаны жөн;
10. Жаттығу жұмысының барысында білімі бекіп, өз бетімен жұмыс жасау дағдылары қалыптасады.

Қ.Қадашева өзінің «Қазақ тілін оқыту әдістемесі» атты кітабында (еңбегінде) «Қазақ тілін екінші тіл ретінде оқыту (ЕТО)» деген тарауында жаңаша жаңғыртып оқытудың әдістемелік түрлерін, ұстанымдарын жетілдіруге, тәсілдерін барынша өзгертуге тырысады, оның мақсаты мен мағынасы арасындағы сәйкессіздікті жоюды басшылыққа алады. Сондай-ақ лингвистикалық ұстаным туралы былайша түйін жасайды: «Лингвистикалық ұстанымдар: мәдениет пен тілді байланыстырып оқыту. Тілді оқытқанда сол үйренетін тілді тұтынушы халықтың мәдениетін, тарихын, әдебиетін, дәстүр-салтын және тілін байланыстыра үйрету табысты болады» [4. 42].

Ал енді «ЖОО-дағы қарым-қатынасқа түсу ұстанымы» деген тақырыпта жаңаша жаңғыртып оқыту үйренушінің дербес белсенділігін арттыруға негізделсе, ондай белсенділік, жаңаша жаңғырту жаттығуларымен орындалады дей келе, мұндай жаттығулар үйренушінің дербес белсенділігіне бағытталып, үнемі есінде сақталуы және активтендірілуге арналуы, сонымен қатар жетілдірілген түрде берілуі қажет,- дейді. Жаттығуларды орындау кезінде нақты талдау тәртібі мен жолы, оны орныдау жылдамдығы ескерілетінін, сонымен бірге сөйлеудің жалпы мазмұны мен тілдік байлығын да назар аудару керектігін қадап айтады.

Ғалым- әдіскер мұндай жаттығуларды орындау үшін үйренушілерге өз бетімен іздеу өрнектері арқылы жұмыс істеуге дайындық жұмыстары ұсынылуы керек деген. Оған мынадай тапсырма түрлерін көрсетеді (нұсқайды):

- көзбен байқау арқылы қабылданған лингвистикалық мазмұнын өз сөзімен айтуға арналған жаттығулар;
- жадында қалған модельдер арқылы лингвистикалық мазмұнын айтуға арналған жаттығулар;

жадында суреттеу арқылы айтуға арналған жаттығулар [4.84].

Ғалым аталған еңбегінде жаттығуларды былайша топтаған:

1. Жаттығу-бақылау;
2. Кешенді жаттығу;
3. Өнімді-рецептивті жаттығулар

Сонымен қатар, жаттығуларды материалдың сипатына, орындалу түріне қарай да сипаттайды [4.22-23].

Мемлекеттік тілді мемлекеттік қызметшілерге (ересектерге, тіл үйренуші жеке тұлғаларға) жеделдете деңгейлік оқыту стандартына (2007 ж.)

жазылым әрекеті барысында орындалатын жұмыс түрлерінің бірі ретінде арнаулы коммуникативтік жаттығулар мен тапсырмалар жүйесі берілген [5.23].

Бұдан шығатын қорытынды: кез-келген зерттеуші тіл үйретуде жаттығу жұмыстарынсыз өз жұмысын ілгері жүргізе алмайды. Бұл жөнінде Қарлығаш Сарбасованың сөзін келтіре кетелік: «Тақырып түсіндірілгеннен кейін, жаттығу жұмыстары орындалады. Дұрыс ұйымдастырылған бұл жұмыстың нәтижесі біріншіден, оқушының не тіл үйренушінің білімін нақтылайды, әрі нығайтады. Екіншіден, өздігінен жұмыс істеуге жаттықтырады, ой қабілетін арттырады. Үшіншіден, қажетті кезінде алған білімін тәжірибеде қолдана білу дағдыларын қалыптастырады. Сонымен, сабақтағы жаттығу жұмыстарының басты мақсаты – білімнің берік те, тиянақты болуын қамтамасыз ету» [Қазақ тілі мен әдебиеті, 1994,6. №5-6. 19 б.].

Тіл үйретуде, меңгертуде жаттығу жұмыстарының орны ерекше екенін бәріміз білеміз, түсінеміз. Тілдің лингвистикалық аспектілері туралы сөз еткенде тіл салаларын меңгерту мәселелері туралы айтылмай тұрмайды. Сондықтан жаттығулардың мынадай түрлері де орындалады:

Жаттығу жұмыстарының бір түрі фонетикалық жаттығулар. Өзге ұлт балалары қазақ тіліне тән дыбыстарды айтуда қиналады, осы сияқты кемшіліктерді жою мақсатында, қазақ тіліне тән дыбыстардың артикуляциясы, олардың қолданысы, орны, емлесі жөнінде мәліметтер беріледі, жаттығу жұмыстары орындалады. Грамматикалық жаттығулардың ішінде морфологиялық жаттығулар ерекше маңызды орын алады. Өйткені ана тіліміздің морфологиялық құрылысын саналы түрде меңгертуге бірден-бір ыңғайлы жол – морфологиялық жаттығулар сөз тұлғасы мен сөз таптарын талдауға байланысты жүргізіледі. Оларды жеке-жеке көрсетсек, шамамен мына талаптар деуге болады.

Сөз тұлғасына байланысты:

1. әуелі негізгі түбірлерді, одан әрі туынды түбірлерді, соңында түбірлес сөздерді талдайтындай жағдай жасау;

2. белгілі бір сөздің ішінен негізгі түбірді тапқызу үшін сол негізгі түбірлерден жасалатын туынды сөздермен салыстыру жұмысын жүргізу;

3. туынды сөздер мен негізгі түбірді ажырату барысында мағына жағына назар аударту талаптарын ескеру.

Сөз таптарына байланысты:

1. әр сөз табының өзіне тән ерекшелігін таныту;

2. жалғау түрлерін меңгерту, әрі олардың белгілі бір сөз табына телініп жалғанатынын саналы түрде ұғындыру;

3. жалғауларды ауызша және жазбаша қолдана білу талаптары.

Жалпы алғанда, морфологиялық жаттығуларды, оның ішінде сөз таптары бйынша жүргізілетін жаттығулар мынадай жұмыс түрлерін қамтиды:

1) Мәтіннен керекті сөздерді тапқызу;

- 2) Керекті сөздерді қосып, өздерінше сөйлемдер немесе әңгіме құрастырту;
- 3) Сөздің мағынасын түсіндіртіп, сөйлемдегі қызыметін көрсету;
- 4) Сөздердің лексикалық немесе грамматикалық мағыналарының қалай не арқылы өзгергенін айтқызу;
- 5) Мағына мен тұлға жағынан қандай ұқсастықтары мен айырмашылықтары барын салыстыру арқылы анықтау;
- 6) Ұқсас сөздерді топтау;
- 7) Байқаған құбылыстары жөнінде қорытынды шығарту т.с.с. жұмыс түрлері.

Грамматикалық жаттығулардың бір түрі – синтаксистік жаттығулар. Оған сөйлем мүшелері, сөйлем түрлері, сөз тіркесі, сөздердің байланысу жолдары мен тәсілдерін таптыру сияқты жұмыс түрлері және оған қоса пунктуацияға қатысты жаттығу түрлерін жатқызуға болады.

Қазақ тілін оқытуда жаттығу жұмысының маңызы ерекше. Жаттығу үнемі тұрақты жүргізілетін болса, алған білім де тиянақты, сапалы болмақ. Яғни жаттығу нәтижесінде тіл үйренушілер алған тілдік (теориялық) білімдерін практикада дұрыс қолдана алуға дағдыланады. Сөйтіп, жаттығулар орындау арқылы тілді меңгерушілердің алған білімдері, берік әрі нақты болады. Сонымен бірге өз бетімен жұмыс істеуге үйренеді, ойлау қабілетін ұштайды, дамыта алады. Себебі тіл үйренуші жаттығу орындау кезінде үнемі талдау-жинақтау жұмысын жасай алады. Демек, жаттығу кешенді, жүйелі жұмыс дағдысын қалыптастырады.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР:

1. Оралбаева Н., Жақсылықова Қ. Мектептерде қазақ тілін оқыту әдістемесі. Алматы. Ана тілі, 1996.- 6б.
2. Жақсылықова К. Қазақ тілін сапалы оқытудың кейбір лингвистикалық аспектілері // жинақта. Алматы, 1997.
3. Ноғайбаева Ж. Дыбыс үндестігі арқылы орыс бөлім студенттерін қазақша сөйлеуге үйрету әдістемесі. ҚД. Алматы, 2001.
4. Қадашева Қ. Қазақ тілін оқыту әдістемесі. Алматы: Мұрагер, 2005.- 176 б.
5. Мемлекеттік тілді мемлекеттік қызметшілерге (ересектерге, тіл үйренуші жеке тұлғаларға) жеделдете оқыту стандарты. Астана, 2007.
6. Қазақ тілі мен әдебиеті // 1994, №5-6.
7. Исабаев Ә. Қазақ тілін оқытудың дидактикалық негіздері (Көмекші оқу құралы).- Алматы: Қазақ университеті, 1993.- 60 б.
8. Қасабек Қ., Күзембаева А., Орыс мектептерінде қазақ тілін оқыту әдістемесінің практикалық және зертханалық жұмыстары (Оқу құралы).-



Алматы: Ы.Алтынсарин атындағы қазақтың білім академиясының
Республикалық баспа кабинеті, 1999.- 190 б.

СТИЛИ РЕЧИ В ДИСКУРСЕ ПЕРСОНАЖА

Айтбенбетова А.К

Муратбек Б.К

Молдабаева З.М

APY им К.Жубанова., Казахстан, г.Актобе

Аннотация: В статье рассматриваются лингвостилистические и коммуникативно-прагматические особенности речи в дискурсе персонажа в художественном диалоге. А также лингвостилистический и эстетический потенциал диалоговой формы речи литературных персонажей очень высок и требует дальнейшего изучения в самых разных аспектах и ракурсах

Ключевые слова: *Художественная литература, лингвостилистика, художественный диалог, особенности речи персонажа.*

Abstract: the article deals with the linguistic-stylistic and communicative-pragmatic features of the character's speech in the artistic dialogue. As well as the linguistic and aesthetic potential of the dialogue form of speech of literary characters is very high and requires further study in a variety of aspects and angles

Keywords: *Fiction, linguistic stylistics, artistic dialogue, features of the character's speech.*

В любом художественном произведении авторская речь (речевая сфера автора) занимает гораздо больший объем, чем речь персонажа и опирается на лингвостилистические средства литературного языка (нормативной речи). Литературный язык – это «образцовый нормализованный язык, нормы которого противопоставляются диалектам и просторечию [1]. В художественном тексте и авторская речь и речь персонажа имеют свои отличительные лингвистические особенности и конструктивные признаки, имеющие выражение в интонации, соотношении порядка слов, характера лексики и синтаксических конструкций». Авторская речь в его разновидностях и речь персонажа может быть охарактеризована с точки зрения коммуникаций, т.е. в аспекте человеческой деятельности, так как язык является не только системой знаков, но и специфической человеческой деятельностью.

В прозаическом художественном произведении в наиболее общем виде можно выделить два типа текстов: повествование и диалог, которые представляют собой неоднородные в структурном отношении формы реализации художественной речи: речевую сферу автора и речевую сферу персонажа. В художественной прозе типы текстов в зависимости от коммуникативных установок подразделяется на определенные композиционно-речевые структуры, к которым О.Г.Ревзина относит: авторское монологическое слово, прямую речь (художественный диалог), внутренний

монолог, несобственно-прямую речь[2]. Лингвистическими особенностями авторского монологического слова могут выступать обращение (повествование) от первого и третьего лица, когда отправитель выражен в форме «я» или когда в третьеличной форме он (автор) не имеет эксплицитного выражения и авторское монологическое слово функционально оформлено в виде описания, повествования и рассуждения.

Под прагматикой понимается «один из планов или аспектов исследования языка в их отношении к тому лицу или лицам. В какой мере предложения в лексическом, морфологическом и синтаксическом плане являются корректным, а то, при каких обстоятельствах, в каких коммуникативных ситуациях, с какими предпосылками и намерениями произносятся данные высказывания и достигают ли они при этом коммуникативных целей. Прагмалингвистика описывает факты языка в плане его употребления, в аспекте человеческой деятельности и стремится раскрыть, в каких условиях и с какой целью в данном случае говорит человек» [3]. Например, Повесть «Материнское поле» — это еще одна из повестей, которая описывает жизнь аила в годы войны. Сюжет прост: колхозница Толгонай, старая женщина, потерявшая в войне мужа и троих сыновей, растит сына своей невестки, которая умерла во время родов, и рассказывает о своей жизни. Герои произведения — Толгонай, ее семья, аилчане и земля, которая связана с народом, перенесшим большие физические и моральные трудности в военные годы.

Повесть начинается с диалога Толгонай с землей и продолжается ее воспоминаниями, которые прерываются диалогами с землей. Впервые в творчестве Ч. Айтматова в этих диалогах земля раскрывается как среда жизни, мышления героини, и в ней война показана как историческое событие, которое нарушает жизнь людей, приносит горе.

В необычном построении композиции повести выразилось новаторство Ч. Айтматова. Он использует прием обрамления, то есть основное повествование прерывается диалогами и, как пишет исследователь М. Плисецкий, «это напоминает по своему построению большую инструментальную музыкальную форму, где чередуются две основные темы» [10. С. 289], под которыми исследователь имеет ввиду судьбу Толгонай и судьбу Алиман. Критик Л. Лебедева отмечает: «История Алиман могла бы стать основой самостоятельного произведения — она как бы искусственно подчиненная целому часть основного сюжета — оттого и нарушается цельность восприятия повести» [9. С. 304].

В повести углубляется психологизм взаимоотношений героев, ставятся философские вопросы, которые в предыдущих произведениях звучали не так выражено. Изображается единая судьба людей, потеря близких, сила женского начала. Лирично описываются сцены переживаний Толгонай о том, что вдруг Алиман решится уйти, как ей сказать, как не оплошать, не обидеть. И ее

нежелание, чтобы она уходила. А потом прощение. Сцена после проводов Касыма на войну и слова Суван-кула, полные истины: «Кто мы были с тобой, Толгонай? Вот с этим народом мы стали людьми. Так давай поровну будем делить с ним все — добро и беды. Когда хорошо было, все были довольны, а теперь, выходит, каждый будет думать только о себе да на судьбу свою плакаться? Нет, так будет нечестно. Если Алиман убивается — так это дело другое, она не видела в жизни того, что мы видели. А ты — мать. Запомни это. Если потребуется, все уйдем» [8. С. 141].

В сферу интересов лингвистической прагматики входят проблемы и вопросы отношений и связей субъекта речи и адресата (слушателя, читателя), речевая стратегия и тактики говорящего, тактика самой речи и ее виды, интонация говорящего, прагматические пресуппозиции и целый ряд коммуникативных характеристик, входящих в сферу адресата: его общеобразовательный и культурный уровень, психология, характер, уровень понимания и оценки субъекта, отношение к высказываемому факту и т.д.

В казахском языкознании прагматика художественного текста исследована в монографии Г.Азимжановой, в которой триада «язык – речь – субъект» рассматривается в прагматической целостности и связанность, целостность, эмотивность художественного текста показаны в аспекте изоморфности и адекватности с семантическими, синтаксическими и прагматическими структурами и характеристиками текста [4].

Прагматический аспект языкового функционирования с наибольшей выраженностью проявляется в устной коммуникации, для которой характерна так называемая «разговорная речь». Определение понятия «разговорная речь» считается дискуссионными. Есть мнения, согласно которым разговорная речь является основной существования языка и самой всеобщей формой языка и в разговорной речи сходится вся проблематика языкознания и ее касается почти любой теоретический вопрос. Термином «разговорная речь» предлагается назвать неподготовленную непринужденную речь носителей литературного языка, обнаруживающуюся в условиях непосредственного общения при неофициальных отношениях между говорящими и отсутствии установки на сообщения, имеющие официальный характер [5]. Особенности разговорной речи могут выступать стилиобразующими факторами и характеристиками речи персонажа в художественном тексте. Так, Г.Азимжанова на примере речи персонажей (диалогического характера) рассказов С.Муратбекова показывает стилиобразующие средства, выступающие на фонетическом и синтаксическом уровнях. К таким средствам Г.Азимжанова относит: 1) протяжное (долгое) произношение некоторых гласных и согласных в речи персонажа; 2) эмпатическое ударение в реплике персонажа; 3) деформация слова; 4) типовые синтаксические конструкции разговорной речи; 5) релятивы; 6) порядок слов, присущий разговорным конструкциям; 7) незаконченные предложения; 8) иноязычные вкрапления; 9) использование осложненных синтаксических

конструкций; 10) использование слов – экспрессивов; 11) использование в речи пословиц и поговорок; 12) выпадение отдельных слов в предложении; 13) использование словесных повторов; 14) различные типы прономинации, продуцирующие характер экспрессии и настроение персонажа [6].

Характеристика речи персонажа не исчерпывается названными выше фонетико-синтаксическими средствами литературно-разговорной речи, так как план персонажа в его композиционно-речевых проявлениях и приемах довольно обширен. Так, Н.А.Кожевникова выделяет речь персонажа в следующих проявлениях: прямая речь в форме внутренней речи или внутреннего монолога; необозначенная прямая речь; смешанные формы чужой речи; несобственно-авторское повествование, стилизация [7].

Кроме плана персонажа, Н.А.Кожевникова выделяет и другие прозаические словесноэстетические структуры (повествование от первого лица и его разновидности; субъективное авторское повествование), характерное для типов повествования в русской литературе 19 – 20 вв. Известно, что в казахском языкознании прозаические композиционно-речевые структуры (тексты в тексте, субтексты) в таком широкомасштабном плане еще не исследованы.

К одной из относительно последних работ в казахском языкознании, имеющих отношение к разработке некоторых проблем речи или языка персонажей, можно отнести кандидатскую диссертацию Енсебаевой К.К. «Стилистические функции языка персонажей современной художественной литературы». В названном диссертационном исследовании выясняются экспрессивные значения и стилистическая нагрузка языка персонажей, выявляются языковые элементы, участвующие в раскрытии образов, а также стилистические особенности употребления лексических элементов и синтаксических конструкций в языке персонажей.

В казахском языкознании разговорная речь рассматривается в ее двух разновидностях: литературно-разговорная речь и устная разговорная речь. Устная разговорная речь отличается и от письменной, и от устной формы литературной речи рядом особенностей, к которым могут быть отнесены: ее функционирование (употребление) в бытовом общении, в играх, на работе, на улице и др. ситуациях, предполагающих устное общение.

Устная разговорная речь в ее употреблении предполагает произвольность речевого акта, неофициальность ситуации общения, преимущественно диалоговая форма и т.д. Некая спонтанность диалоговой формы устной разговорной речи персонажа является ее важнейшей характерной особенностью. «...Собственно диалог по существу не имеет «программы». Именно «неподготовленность», выступающая в данном случае как прием, стала отличительным качеством диалога. «Разговорность» сказывается прежде всего в том, что в диалоге нет ведущего». Каждый из

участников качинает не с вопросов к собеседнику, а с реализации своих собственных позиций».

Разговорная форма диалоговой речи в художественном произведении диктует особенности синтаксического и лексического (лингвостилистического) построения и функционирования данной формы речи. Устно-разговорная диалоговая речь строится с учетом того, чтобы она была понятно (общедоступно) и адекватно воспринимаема. В диалоге соблюдается жесткая экономия лексических и синтаксических средств – в рамках диалогической речи функционируют эллиптические, незаконченные, однословные и т.д. предложения и очень ограниченные по объему словосочетания. Несмотря на то, что диалогу присущи многие (если не все) характерные особенности устной разговорной речи, тем не менее диалог в художественном произведении имеет свою специфику, отличающую его от собственно устной формы разговорного диалога. Диалог в художественном тексте прямо и опосредованно зависит от характера и особенностей художественного произведения, выступающего в качестве эстетически организованного сложного словесного целого. Диалог в художественном тексте подчинен правилам времени, ритма и темпа, вне которых не существует художественный текст.

Если диалог реальном общении выполняет чисто коммуникативно-прагматическую функцию, то художественном тексте, он, характеризую речь персонажей, выполняет стилеобразующую, т.е. эстетическую функцию. Говоря другими словами, любая лексико – стилистическая единица художественного текста, в том числе и диалог, подчинены задачам, целям, идее, теме художественного произведения как эстетически целого и подчинены авторской воле и таланту языковой личности писателя. Исходя из этой позиции, можно говорит о диалоге, как об одной из важных категорий художественного текста, выполняющей стилистические функции языка персонажей и рисующей образ персонажей художественного произведения.

Диалогическая речь как образ – и стилеобразующее средство может характеризовать персонаж с разных сторон как одного из героев и действующих лиц художественного произведения. Так, исследователь А.И.Смайл раскрывает связь особенностей диалогической речи с образами персонажей, считая, что стилистические, грамматические, синтаксические и лексические средства диалога помогают построить и раскрыть характер и образ персонажей – героев литературного произведения.

Таким образом, речь персонажа, проявленная в форме диалога, имеет свои отличительные лингвостилистические и коммуникативно – прагматические особенности, о части которых говорилось в данном сообщении. Лингвостилистический и эстетический потенциал диалоговой формы речи литературных персонажей очень высок и требует дальнейшего изучения в самых разных аспектах и ракурсах. Так, например, по свидетельству

К.К.Кенжебаевой диалог в художественном тексте может создавать так называемый микрообразили ситуационный образ, а также художественный образ персонажей. Речь (язык персонажей), продуцированная в форме диалога, может характеризовать тот или иной образ персонажа – действующего лица литературного произведения с точки зрения употребления в ней элементов «деревенской речи» и «городской речи». Приведенные выше факты свидетельствуют о лингвостилистическом и эстетическом потенциале и особенностях речи персонажа в лингвостилистических, лингвопрагматических и др.аспектах.

Литература:

1. Гак В.Г.Прагматика, узус и норма речи. Иностранные языки в школе. 1982. №5, стр.30
2. Ревзина О.Г. Методы анализа художественного текста//Структура и семантика художественного текста? Доклады 7 Международной конференции. М., 1999, стр.259
3. Земская Е.А.Русская разговорная речь. Тексты. М., 1978, стр.65
4. Ёзимжанова Г. Көркем проза прагматикасы. Алматы, 2005, 13 – 14 бб.
5. Қазақ әдеби тілінің ауызша түрі. «Ғылым» баспасы. Алматы, 1987, 120 б.
6. Ахманова О.С. Словарь лингвистических терминов. М., 1966
7. Кожевникова Н.А. Типы повествования в русской литературе 19-20 вв. М., 1994
8. Айтматов Ч. Т. В Германии: исследования, статьи, рецензии о творчестве писателя / Составитель Е. Шагалова. — Бишкек, 1991.
9. Асаналиев К.А. Открытие человека современности. — Фрунзе, 1968.
10. Асаналиев К.А. Новая книга Ч. Айтматова // Ч. Айтматов: очерки, статьи и рецензии о творчестве писателя. — Фрунзе, 1975.

UDK: 811.512.133`42

O‘ZBEK TILIDAGI MATNLARDAN SIFAT, RAVISH, YORDAMCHI SO‘Z VA ORALIQ SO‘ZLAR TURKUMIGA OID SO‘ZLARNI ANIQLASH

Sharipov Maqsud Siddiqovich

UrDU t.f.n, AT kafedrası dotsenti

maqsbek72@gmail.com,

Jumaniyozova Zebiniso Xushnud qizi

UrDU magistranti

zebinisojumaniyozova0922@gmail.com

“Ma’mun” universiteti NTM o‘qituvchisi

Annotatsiya. O‘zbek tilshunosligida boshqa so‘z turkumlari singari sifat, ravish va yordamchi so‘z turkumi ham katta ahamiyat kasb etadi. Barcha turkiy tillar kabi o‘zbek tili ham agglutinatív til bo‘lib so‘z o‘zagi yoki o‘zakga affiks (qo‘shimcha)lar qo‘shilishi bilan yasaladi. O‘zbek tilidagi matnlardan sifat, ravish, yordamchi so‘z va oraliq so‘zlar turkumiga oid so‘zlarni aniqlash uchun tub sifat va tub ravish so‘zlar bazasi yaratildi. Ushbu maqolada o‘zbek tilidagi matnlaridan sifat, ravish va yordamchi so‘z turkumiga tegishli so‘zlarni aniqlashning qoidaga asoslangan algoritm yaratish jarayoni ko‘rib chiqiladi.

Kalit so‘zlar: *ma’lumotlar bazasi, algoritm, affikslar, lemma, chekli avtomat (FSM), tokenlash, deterministik bo‘lmagan chekli avtomat (NFA).*

Abstract. In Uzbek linguistics, like other word groups, adjectives, adverbs and auxiliary words are of great importance. Like all Turkic languages, the Uzbek language is an agglutinative language and is formed by adding affixes to the stem or stem. A database of basic adjectives and basic adverbs was created to identify words related to adjectives, adverbs, auxiliary words, and intermediate words from Uzbek texts. In this article, the creation of a rule-based algorithm for identifying words belonging to adjectives, adverbs and auxiliary words from Uzbek language texts is considered.

Key words: *database, algorithm, affixes, lemma, finite state machine (FSM), tokenization, Non-Deterministic finite automata (NFA).*

Абстракт. В узбекском языкознании, как и в других группах слов, большое значение имеют прилагательные, наречия и вспомогательные слова. Как и все тюркские языки, узбекский язык является агглютинативным языком и образуется путем добавления аффиксов к основе или основе. База данных основных прилагательных и основных наречий была создана для выявления слов, связанных с прилагательными, наречиями, вспомогательными словами и промежуточными словами из узбекских текстов. В данной статье рассматривается создание основанного на правилах алгоритма выявления слов,

входящих в состав прилагательных, наречий и вспомогательных слов из текстов на узбекском языке.

Ключевые слова: база данных, алгоритм, аффиксы, лемма, конечный автомат (FSM), токенизация, недетерминированные конечные автоматы (NFA).

Tildagi barcha soʻzlar oʻziga xos belgi-xususiyatlari, soʻrogʻi va qanday umumlashgan maʼno ifodalashiga va gapda biror gap boʻlagi vazifasida kelishiga koʻra soʻz turkumlariga ajratiladi. Soʻz turkumlari maʼno va vazifasiga koʻra quyidagi guruhlariga boʻlinadi:

1. Mustaqil soʻzlar: Feʼl, ot, sifat, son, olmosh va ravish;
2. Yordamchi soʻzlar: Bogʻlovchi, koʻmakchi va yuklama;
3. Oraliq soʻzlar: Undov, taqlid va modal soʻzlar;

Hozirgi kunda zamonaviy texnologiyalar imkoniyatlari samaradorligini yaxshilash maqsadida matndan sifat, ravish, yordamchi va oraliq soʻz turkumiga tegishli soʻzlarni aniqlash zarurati tugʻiladi. Bu masalani tadqiqot maqsadi va vazifalari, oʻrganilganlik darajasini aniqlash maqsadida maʼlumotlar toʻplash ishiga kirishildi. Maʼlumotlar bazasini toʻplashda Sayfullayeva R., Mengliyev B. va boshqalarning “Hozirgi oʻzbek adabiy tili” va Erkaboyeva N. “Oʻzbek tilidan maʼruzalar toʻplami” darsliklari asos sifatida olindi. Bu maʼlumotlardan foydalangan holda oʻzbek tilining izohli lugʻati 1-2-3-4-5-jildlaridan sifat va ravish soʻz turkumiga doir soʻzlar ajratib olindi. Oʻzbek tili izohli lugʻatining 1-2-3-4-5-jildlaridan hammasi boʻlib, 1601 ta sifat va 877 ta ravish soʻz turkumiga doir soʻzlar aniqlandi va jadval koʻrinishida shakllantirildi. Ajratib olingan maʼlumotlar bazasi txt. fayl sifatida baza qilib kiritildi. Bazadan foydalangan holda sifat va ravish soʻz turkumiga oid soʻzlar tub va yasama soʻzlar guruhlariga ajratib olindi. Sifat va ravishning shakl yasovchi va soʻz yasovchi qoʻshimchalari bazasi shakllantirildi.

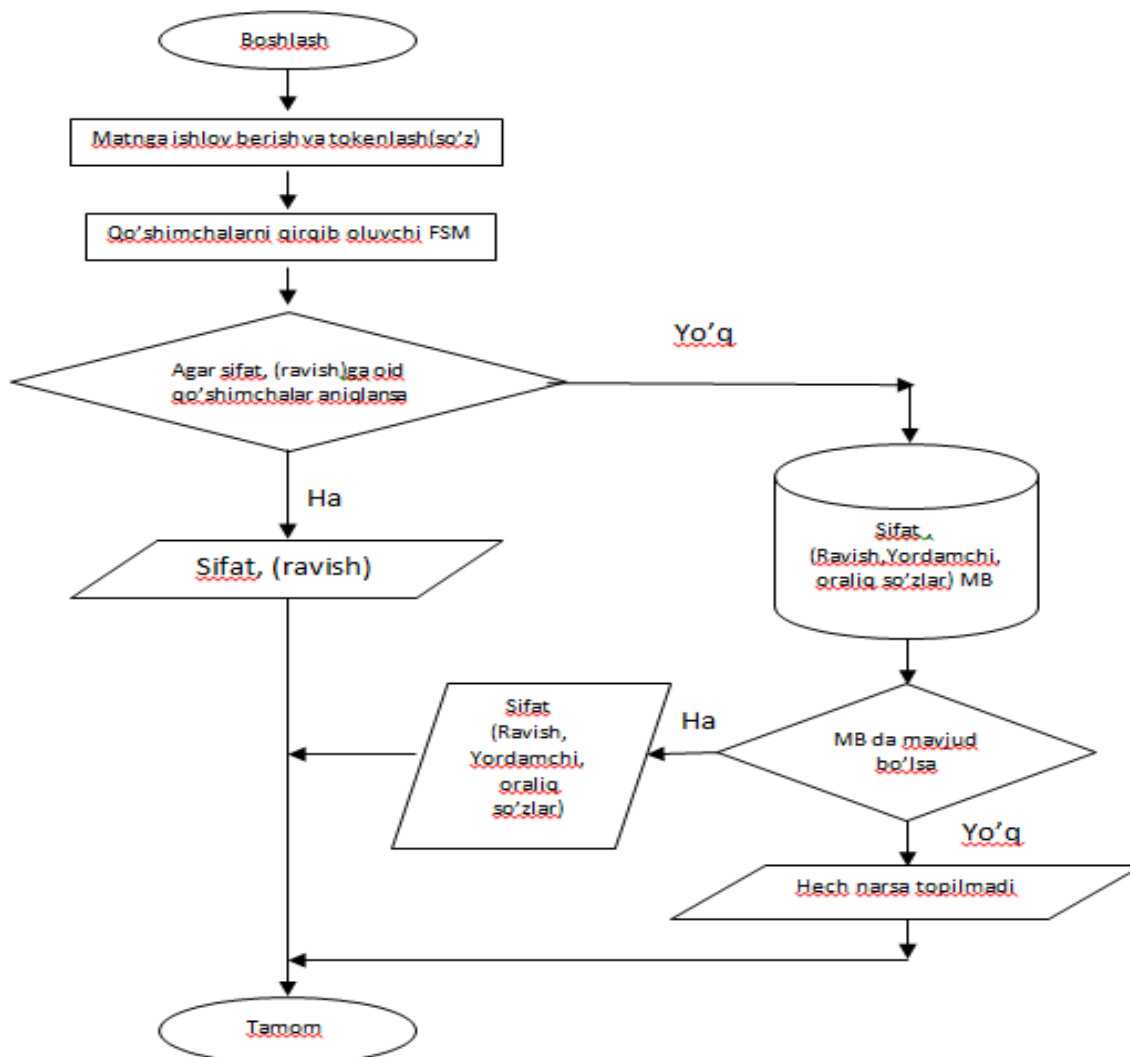
Soʻz turkumlari	Sifat lemma yasovchi affikslar	Ravish lemma yasovchi affikslar
Otdan	56	18
Feʼldan	40	16
Sifatdan	11	5
Ravishdan	3	4
Sondan	0	0
Olmoshdan	2	3
Taqlid soʻzlardan	3	0
Undov lemmadan	0	0
Modal lemmadan	2	2
Jami yasovchi affikslar	117	48

2-jadval: Boshqa soʻz turkumlaridan sifat va ravish lemma yasovchi affikslar miqdori

Oʻzbek tilida sifat va ravish soʻz turkumiga tegishli soʻzlarni aniqlashda old qoʻshimchalardan ham foydalanish mumkin.

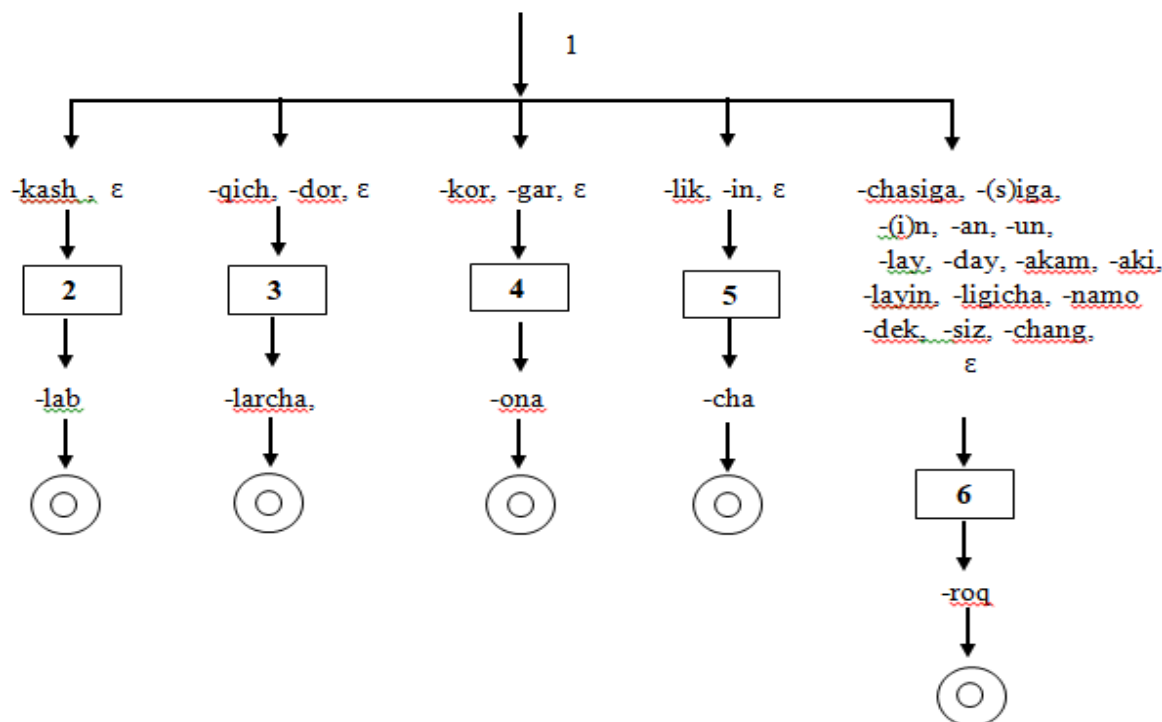
Gʻayri-	+ Sifat(Adj) = Sifat(Adj)	gʻayri+tabiiy(Adj)=gʻayritabiiy(Adj)
Ser-		ser+shox(N)=sershox(Adj)
Be-		be+baxt(N)=bebaxt(Adj)
Xush-		xush+bichim(N)=xushbichim(Adj)
Bad-		bad+axloq(N)=badaxloq(Adj)
Ba-	+ Ot(N) = Sifat(Adj)	ba+havo(N)=bahavo(Adj)
No-		no+umid(N)=noumid(Adj)
Bo-		bo+adab(N)=boadab(Adj)
Bar-		bar+kamol(N)=barkamol(Adj)
Ham-		ham+fikr(N)=hamfikr(Adj)
Bar-		bar+vaqt(N)=barvaqt(Adv)
	+ Ot(N) = Ravish(Adv)	
Be-		be+ixtiyor(N)=beixtiyor(Adv)

Bular yordamida oʻzbek tilidagi matnlardan sifat va ravish soʻz turkimini aniqlovchi algoritm ishlab chiqildi.



3-jadval: Matndan sifat va ravish so‘z turkumiga tegishli so‘zlarni aniqlovchi
algoritm

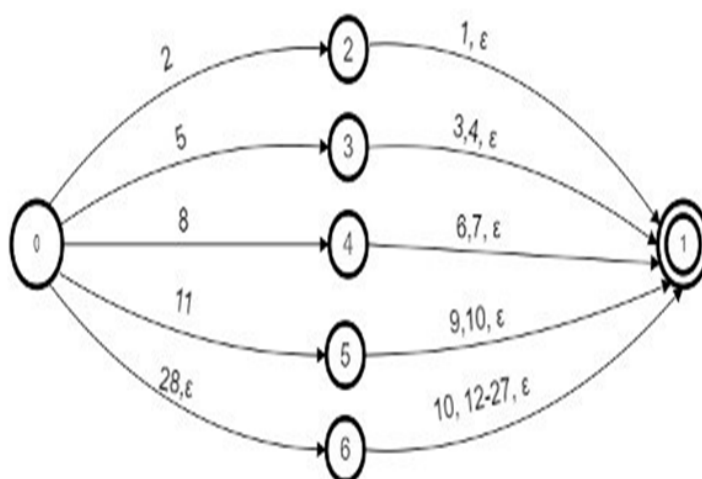
Algoritm asosan eng birinchi berilgan matnni tokenlab oladi, ya’ni matnni so‘zlarga ajratib olinadi. Tub sifatlarni va so‘z o‘zaklarini ma’lumotlar baza (MB)dan qidirib topadi. Yasama sifat va ravishlarni aniqlash uchun sifat va ravishning shakl yasovchi va so‘z yasovchi qo‘shimchalari uchun alohida-alohida chekli avtomat (FSM)larni qurib olindi.



4-jadval: Ravish soʻz turkumi yasovchi qoʻshimchalarni qidirish chekli avtomat (FSM)

Chekli Avtomat (FSM)larga asoslangan holda deterministik boʻlmagan chekli avtomat (NFA) tuzib olindi:

1. -kash
2. -lab
3. -qich
4. -dor
5. -larcha
6. -kor
7. -gar
8. -ona
9. -lik
10. -in
11. -cha
12. -chasiga
13. -siga
14. -iga
15. -n
16. -an
17. -un
18. -lav
19. -day
20. -akam
21. -aki
22. -lavin
23. -ligicha
24. -namo
25. -dek
26. -siz
27. -chang
28. -roq



5-jadval: Ravish soʻz turkumi yasovchi qoʻshimchalarni qidirish deterministik boʻlmagan chekli avtomat (NFA)

O'zbek tilida yordamchi va oraliq so'z turkumiga tegishli so'zlar qo'shimcha olmaganligi sababli bu so'z turkumiga tegishli so'zlarni ma'lumotlar bazasidan topiladi. O'rganishlar natijasida yordamchi so'zlar: bog'lovchilardan 33 tasi, ko'makchi 82 tasi, yuklama 30tasi va oraliq so'zlar: taqlid so'zlar 163tasi, undov so'zlar 82tasi va modal so'zlar 67 tasi aniqlandi va ma'lumotlar bazasiga kiritildi.

Xulosa qilib aytish mumkinki, ushbu ma'lumotlar bazasi bizga sifat, ravish, yordamchi va oraliq so'z turkumiga oid so'zlarni aniqlashda asos bo'lib xizmat qiladi. Ushbu ma'lumotlar bazasidan foydalangan holda Python dasturlash tilida algoritm tuzish maqsad qilindi.

Foydalanilgan adabiyotlar.

1. “O'zbekiston milliy ensiklopediyasi” Davlat ilmiy nashriyoti. -T.: 2020.
2. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.
3. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com
4. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.
5. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.
6. X. Madatov, M. Sharipov, and S. Bekchanov, “O'ZBEK TILI MATNLARIDAGI NOMUHIM SO'ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021.
7. K. Madatov, S. Bekchanov, and J. Vičič, “Accuracy of the Uzbek stop words detection: a case study on "School corpus",” *arXiv preprint arXiv:2209.07053*, 2022.
8. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.
9. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” *arXiv preprint arXiv:2210.16006*, 2022.

O‘ZBEK TILIDA RAVISH SO‘Z TURKUMINI MODELLASHTIRISHDA SO‘Z YASALISHI MASALALARI VA ULARNING MORFOANALIZATORDA TAHLILI

Abdullayeva Oqila Xolmo‘minovna

ToshDO‘TAU Kompyuter lingvistikasi va raqamli
texnologiyalar kafedrasida katta o‘qituvchisi, PhD

abdullayevaoqila@gmail.com

Nizomova Fotima Bo‘rixo‘ja qizi

ToshDO‘TAU 2-kurs talabasi

nizomovafotima@navoiy-uni.uz

Annotatsiya. O‘zbek tilida so‘z turkumlarini, jumladan, ravish so‘z turkumini modellashtirish, uning lisoniy qolipini ishlab chiqish masalasi zaruriy vazifa hisoblanadi. Maqolada ravish so‘z turkumiga oid nazariy qarashlar tahlilga tortildi va qiyoslandi. Ravish so‘z turkumi va uning morfoanalizatoridagi ahamiyati jihatidan fikrlar bildirildi. Ravish so‘z turkumi va uning yasash qonuniyatlari borasidagi muammolarga kompyuter lingvistikasi jihatidan yechim berish, algoritmini ishlab chiqish borasida fikr yuritildi. Shuningdek, ravishning yasama yoki tub ekanligi to‘g‘risida nazariya taqdim etildi va tilshunos olimlarning bir nechta asarlarida isbotlangan fikrlari qayd etildi. Nazariy fikrlar misollar orqali isbotlandi.

Kalit so‘zlar: *ravish, so‘z yasalishi, lingvistik modellashtirish, algoritm.*

Annotation. It is a necessary task to model word groups in the Uzbek language, including the adverb word group, and to develop its linguistic form. In the article, the theoretical views on the lexical phraseology were analyzed and compared. Opinions were expressed in terms of the adverb vocabulary and its importance in the morphoanalyzer. It was thought about giving a solution to the problems of the adverb word group and the laws of its formation in terms of computer linguistics, and developing an algorithm. In addition, a theory was presented about whether adverb is artificial or fundamental, and the opinions of linguists, proven in several works, were noted. Theoretical ideas are proved by examples.

Key words: *adverb, word formation, linguistic modeling, algorithm.*

Аннотация. Необходимой задачей является моделирование групп слов в узбекском языке, в том числе наречие групп слов, и разработка их языковой формы. В статье проанализированы и сопоставлены теоретические взгляды на лексическую фразеологию. Были высказаны мнения относительно лексики наречие и ее важности в морфоанализаторе. Задумывалось дать решение проблем словосочетания наречие и закономерностей его образования с точки зрения компьютерной лингвистики, и разработать алгоритм. Кроме того, была представлена теория о том, является ли наречие искусственным или

фундаментальным, а также отмечены мнения лингвистов, подтвержденные в нескольких работах. Теоретические идеи подтверждаются примерами.

Ключевые слова: наречие, словообразование, лингвистическое моделирование, алгоритм.

Kirish. So‘z yasash leksikaning o‘shida muhim rol o‘ynaydigan faktorlardan biri bo‘lib (uning mahsuli tilda so‘zlarni ko‘paytiradi, leksikaga xizmat qiladi), yo‘llari ham, vositalari ham har xil (affiks yordami bilan yasash: affiksatsiya, so‘zlarni qo‘shib yasash: kompozitsiya va boshqalar), ular butun bir sistemani tashkil qiladi.

So‘z yasalishi tilshunoslik fanining ayrim bir bo‘limi, mustaqil lingvistik soha, tarmoq bo‘lib, u so‘zlarning yasalishini, yangi so‘z hosil qilishning qonun-qoidalarini, modellarini, vositalarini, shu bilan bog‘liq holda, so‘zlarning strukturasini tekshiradi. Demak, bu sohaning obyekti so‘zdir. So‘zni leksikologiya ham, grammatika ham tekshiradi, lekin bu ikki xil tekshirish bir-biridan ma’lum belgilar bilan ajralib turadi. [G‘ulomov, 1973:7].

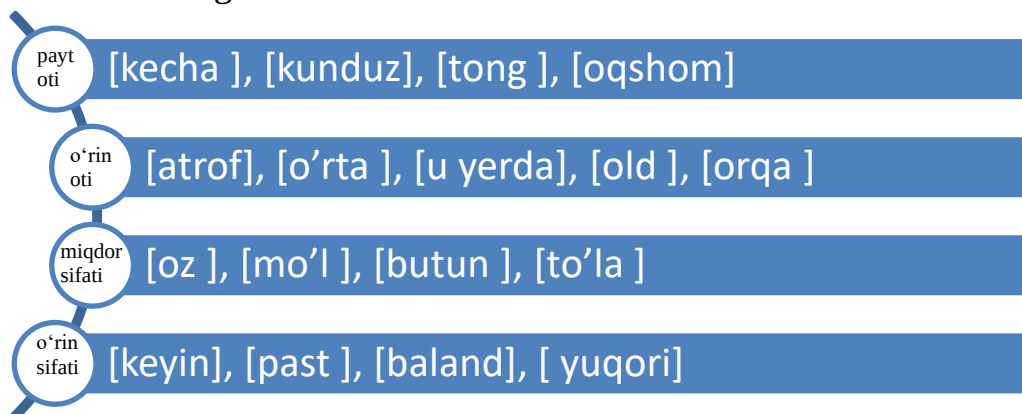
Leksika doimiy harakatda bo‘lib, uning rivojlanishi, boyish yo‘llari har xil: boshqa tillardan so‘z olish, so‘z yasash (betonchi, atomchi kabi), ayrim grammatik kategoriyalarning leksikalizatsiyasi (birdan, to‘satdan, tezda kabi so‘zlarning to‘la paradigma ko‘rinishlariga ega bo‘lmay, butunligicha bir formada “qotishi”), so‘z komplekslarining ajralmas, semantik bir butun holga kelib, bir so‘zning xususiyatlarini eslatuvchi so‘zga teng birlik hosil qilishi: frazeologik xarakterdagi birliklar (neftni qayta ishlash sanoati, mehnat rezervlari boshqarmasi kabi) va boshqalar. Bularning eng asosiylari boshqa tildan so‘z olish va so‘z yasashdir. [G‘ulomov, 1973:21].

Ravish

Ravish va uning umumiy grammatik ma’nosi.

Ravish hamisha tilshunoslikning muammoli masalasidan bo‘lib kelgan. Uni o‘rgangan olim V.V.Vinogradov “mustaqil so‘zning hech bir guruhiga sig‘may qolgan so‘z ravish turkumiga yig‘ilgan” deb haqqoniy baho bergan edi. Uning ma’no va vazifasi ot, sifat, son, olmoshga yaqin turishi arab va Yevropa tilshunosligida qayd etilgan. Sharq va g‘arb tilshunosligi ma’no va vazifasi jihatidan emas, balki boshqa birortasiga xos bo‘lmagan belgi-o‘zgarmaslik belgisi bilan ismdan farqlaydi. Biroq o‘zbek tili grammatik qurilishining tavsifiga bag‘ishlab yozilgan ko‘plab adabiyot va darslikda ravishning o‘zgarmaslik tabiati inobatga

olinmagan.



Tilshunoslar o'zbekcha ravishni rus tilidagi “напечение”ning muqobili sifatida talqin qilishgan, natijada, ravish turkumi doirasida o'rganilgan. [Sayfullayeva, 2009, 254].

Asosan harakatning, qisman predmetning belgi, miqdor yoki holatini bildiruvchi o'zgarma so'z- ravishning so'z turkumi sifatidagi umumiy grammatik ma'nosi. [Sayfullayeva, 2009:255]

Ravishda derivatsiya. Tilshunosligimizda ravish yasalihi munozarali. Chunki ravishning asosiy qismi-asli boshqa turkumga mansub soddalashgan, yaxlitlashgan leksema. Shu boisdan aksariyat tilshunoslar ravish yasalihini inkor etishib, ravishlashish (boshqa turkum so'zining ravishga o'tishi) mavjudligini tan olishadi.

Ravishlar tarkibiga e'tibor berilsa, ularda turli so'z yasash qolipi mavjudligi ma'lum bo'ladi:

[ot+an]	[ot+chasiga]	[ot+larcha]	[ot+ona]	[ot+lab]	Qo'shma ravish [olmosh+ot], (har gal, har zamon)	[ravish+ot] (hali zamon, hali beri)	[son+ot] (bir yo'la, bir dam)	Boshqa turkumdagi takroriy so'z ravish bo'lishi mumkin: yuzma-yuz, quruqdan- quruq

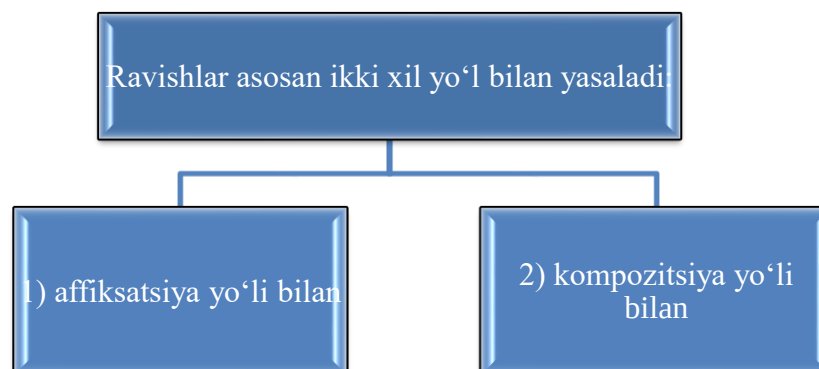
Tilshunoslar ravish turkumining yasalihi sistemasiga ega emasligini [-an], [-ona] qo'shimchasining arabcha va tojikcha so'z tarkibida uchrashini, [-larcha], [-chasiga], [-lab] kabining grammatik ma'no ifodalashi bilan va [-iga-, siga], [-chasiga] kabi so'zning sanoqli darajada ekanligi bilan izohlaydi. Ravish gapda ko'pincha fe'lga bog'lanib hol, otga bog'lanib sifatlovchi vazifasida keladi. Bu ravish - umumiy grammatik ma'nosining tarkibiy qismi. [Sayfullayeva, 2009:256-257]. Ravishning o'ziga xos yasovchi affikslarga ega bo'lishi ham uning morfologik belgisi sanaladi. -chasiga (mardchasiga), -larcha (vahshiylarcha), -iga, -siga (ko'ndadalangiga,

uzunasiga), lab (haftalab), -cha (askarcha, atroflicha, shoshilgancha), -dek, -day (o'qdek, toshday) kabi.

Tuzilish jihatdan ayrim ravishlar kelishik formasini olganga o'xshab ketadi.

otlarga	olmoshlarga	sifatlarga	ravishdoshlarga
• quyiga, tubdan	• shundan, unda-bunda	• sekinroq, bab-baravar	• ertalab, bora-bora, turib turib

Bu o'xshashlik tasodifiy hodisa bo'lmay, balki ravish so'z turkumi sifatida shakllanayotgan payt o'z tarkibi o'sha turkumlardan ajralib chiqqan so'zlar hisobiga to'ldirgan. Hatto shunday ravishlar borki, ular hozirgi kunda boshlang'ich formalaridan tamomila uzoqlashgan: olg'a, onda, sonda, jo'rttaga, tashqari, yana kabi. [Fuzailov, 1973:528]



1. Affiksatsiya yo'li bilan ravish yasalihi tilimizda qadimdan mavjud bo'lib, ayniqsa, keyingi vaqtlarda eng unumli usulga aylangan. Ravish yasovchilar ham o'z navbatida unumli va unumsiz affikslarga ajralib turadi. [Fuzailov, 1973, 529].

Unumli yasovchilar	• -cha • -larcha • -chasiga • -iga • -siga • -lab • -deb • -day	Unumsiz yasovchilar	• -ona • -an • -n (-in) • -lay • -simon

Ravish yasalihiining ikkinchi asosiy usuli kompozitsiyadir.

Kompozitsiya yo'li bilan so'zlar o'zaro qo'shiladi, tuziladi yoki takrorlanadi, natijada ravish hosil bo'ladi. Bu yo'l bilan ravish yasalihini boshqa turkumdagi shu xil yasalihi bilan tamoman tenglashtirib bo'lmaydi.

So‘z qo‘shish bilan yasashda birmuncha ravishlar tarkibida morfologik
formalarning ham mavjudligini kuzatamiz:

-bir yerda, bu
yoqqa, ketma-ket

shu yerdan,
kundan-kunga,
nom-banom

bekordan-bekor,
qo‘yarda-
qo‘ymay,

Bu formalar qo‘shma ravish yasashda muhim ahamiyatga ega.
[Fuzailov, 1973, 532].

Ma'lumki, sinonimik qatorni tashkil etgan ba'zi qo'shma ravishlar aynan
so'zlarning ishtirokidan vujudga kelishiga qaramay, ularning morfologik
shakllanishidan farq bo'ladi: **bir yo'l, bir yo'la, bir yo'li, bir yo'lasi; yo'l-yo'la,**
yo'lma-yo'l, yo'l-yo'lakay, har dam, har damda, damma-dam; kun-kundan,
kundan-kunga, kundan kun; nari-beri, naridan-beri kabi. [Fuzailov, 1973, 533].

Ravishlar tuzilishiga ko'ra sodda, qo'shma, murakkab va juft bo'ladi.

Sodda ravishlar birgina o'zakdan tashkil topib, tub yoki yasama bo'ladi.
Yasovchi affiksarsiz qo'llanadigan ravishlar tub-ravishlar deyiladi: tez, sekin, keyin,
oz, ko'p, picha, kecha.

Maxsus affikslar bilan yasalgan ravishlar yasama ravishlar deyiladi:
to'xtovsiz, oyday, mardlarcha, g'olibona, mardona, qaxramonlarcha

- Ravishlar quyidagi affikslar yordamida yasaladi:

1) -an: taxminan, qisman, tasodifan;

2) -siz: to'xtovsiz, so'zsiz;

3) -lab: grammlab, kilolab, aravalab, ko'plab;

4) -cha; o'zbekcha, menincha;

5) -an: taxminan, qisman, tasodifan;

6) -n {-in): ertan, yashirin, yozin-qishin, ochin-to'qin;

7) -incha: yashirincha, ko'pincha;

8) -ligicha: tirikligicha, butunligicha;

9) -ona: mardona, do'stona, g'olibona, g'oyibona;

10) -lay, (-layin): butunlay, tiriklay, tiriklayin;

11) -day (-dek): rutday, metinday, oyday, paxtadek, mushtdek.

[Asqarova M. Abduraxmonov X. 1981, 108].

Yuqoridagi manbalardagi qoidalarga e'tibor beradigan bo'lsak ravish so'z
turkumining yasalishi munozarali masalaga aylanib turibdi. Bu masalaga Azim
Hojiyevning “O'zbek tilida so'z yasalishi” qo'llanmasidagi quyidagi fikrlarini
keltirib o'tish lozimdir.

Hozirgi o'zbek adabiy tilida so'z yasashining asosiy usuli affiksatsiya
usulidir. Hozirgi o'zbek adabiy tilida so'z yasovchi affikslar, aniqrog'i, so'zning

morfem strukturasida soʻz yasovchi sifatida ajraladigan affikslar juda koʻp. Lekin ularning hammasi ham yangi-yangi soʻzlar yasashda qatnashavermaydi. Ulardan ayrimlari hozirda ham yangi soʻzlar yasashda qoʻllanayotgan boʻlsa, baʼzi affikslar esa soʻz yasash funksiyasini deyarli, yoki butunlay toʻxtatgan. Hozirgi oʻzbek adabiy tili soʻz yasash sistemasining tahlilida birinchi tipdagi, yaʼni soʻz yasash funksiyasi davom etayotgan affikslar yordamida soʻz yasashini yoritish muhimdir. Maʼlumki, ayrim soʻz turkumlarida soʻz yasashda affiksoidlar ham qatnashadi. Ular yordamida ham sodda yasama soʻzlar hosil qilinadi. Shu tufayli bunday soʻz turkumlarida, affikslar yordamida soʻz yasashidan soʻng, affiksoidlar bilan soʻz yasashi ham berildi. [Hojiyev, 1989:35].

Hozirgi oʻzbek adabiy tilida ravish yasashining asosiy usuli affiksatsiya usulidir. Lekin ravish yasovchi affikslar u qadar koʻp emas. Shuningdek, ravish yasovchi aktiv affikslar yordamida hosil boʻlgan ravishlar ham koʻp miqdorda emas. Hozirgi oʻzbek tilida ravish yasovchi mahsulli affikslar, yaʼni soʻz yasash funksiyasi toʻxtamagan affikslar deb, -cha, -chasiga, -larcha, -lab affikslarini koʻrsatish mumkin. [Hojiyev, 1989, 93]. -cha affiksi feʼldan boshqa deyarli barcha mustaqil soʻz turkumlarga oid leksemalardan ravish yasaydi. Lekin, birinchidan, uning yordamida yasalgan ravishlar koʻp emas. Ikkinchidan, bu affiks yordamida ravish yasashi maʼlum sistemani tashkil etmaydi. Masalan, u ot va sifatdan ravish yasaydi: zarracha, ayyorcha kabi. *Lekin -cha affiksi ot yoki sifatning maʼlum bir leksik-semantik turidan umumiy soʻz yasash maʼnosiga ega boʻlgan, soʻz yasashining maʼlum tipini hosil qiluvchi ravishlar yasamaydi.* [Hojiyev, 1989:93].

-cha affiksi orqali yasalgan ayyorcha, yashirincha soʻzlari tarz ravishidir. Lekin bular bir xil soʻz yasashi maʼnosiga ega emas: ayyorcha «ayyorga xos tusda»; yashirincha «yashirin holda». Quyidagi misolda yana boshqacha maʼno: *Sobirovning qilgʻilikni qilib, kolxozni tashlab ketishi ustida uzoq toʻxtaydi va bunga ham boleshevichka baho berib oʻtadi* (A. Sodiriy). [Hojiyev, 1989:93].

-chasiga. Bu affiks asosan shaxs bildiradigan otlarga qoʻshilib, harakat-holatning asosdan anglashilgan shaxs yoki narsaga xos tarzda, shunga xos usul bilan bajarilishini bildiruvchi ravish yasaydi: yigitchasiga, mardchasiga, dehqonchasiga, chapanichasiga, tentakchasiga kabi. [Hojiyev, 1989:94].

-larcha. Hozirgi oʻzbek adabiy tilida ravish yasovchilar ichida eng faoli deb -larcha affiksini keltirish mumkin. Bu affiks koʻplik koʻrsatkichi -lar bilan -cha affiksining birikuvidan hosil boʻlgan. Shu sababli, -larcha affiksi yordamida yasalgan ravishlarda -lar affiksiga xos maʼno sezilib turadi.

-larcha affiksi asosan baʼzi otlardan, aniq shaxs bildiruvchi otlardan va belgi bildiruvchi soʻzlardan ravish yasaydi. Bu tipdagi ravishlar boshqa harakatning yoki belgining belgisini asosdan anglashilgan narsa-predmetning harakatiga, belgisiga qiyoslab oʻxshashligini, shu kabi ekanini bildiradi: otalarcha gʻamxoʻrlik qilmoq, qardoshlarcha yordam kabi. [Hojiyev, 1989:94].

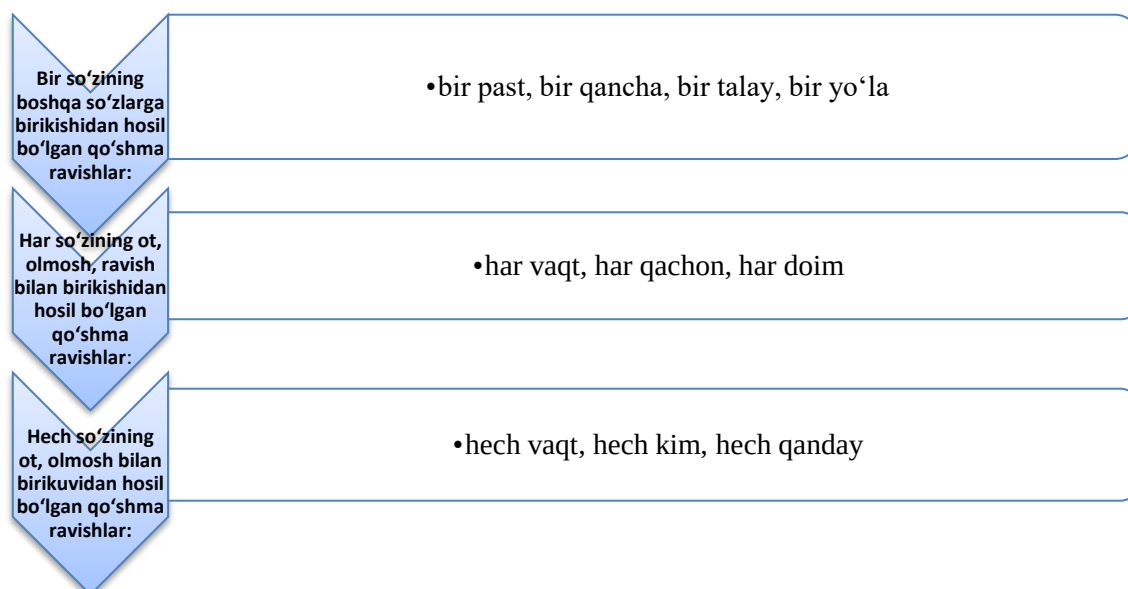
-lab. Bu affiks miqdor, hajm-oʻlchov, vaqtni bildiruvchi soʻzlardan ravish yasaydi: oylab, metr lab, grammlab kabi.

Vaqt birliklarini bildiruvchi soʻzlarga qoʻshilganda, harakatning asosdan anglashilgan vaqtda, shu vaqt davomida bajarilishini bildiruvchi ravish yasaydi: azonlab (uygʻotdi), yillab (koʻrishmadilar) kabi. Miqdor, oʻlchov birliklarini anglatuvchi soʻzlarga qoʻshilganda, harakatning asos bildirgan oʻlchov-miqdor bilan bajarilishini (bildiruvchi ravishlar yasaladi: kilolab, tonnalab, metr lab, qarichlab, paqirlab kabi. [Hojiyev, 1989:95].

-chang. Bu affiks kiyim-bosh bildiruvchi ayrim soʻzlarga qoʻshilib, shaxsning faqat asosdan anglashilgan kiyimda ekanlik holatini, shunday holatda biror harakatni bajarishni bildiruvchi ravish yasaydi: kavushchang, mahsichang, koʻylakchang, maykachang kabi.

Yuqorida keltirilgan ravish yasovchilardan tashqari yana bir qator affikslar borki, ular yordamida yasalgan ravishlar oz yoki koʻp miqdorda boʻlishidan qatʼiy nazar, hozirgi tilda yangi ravish (yoki ravishlar) yasamaydi yaʼni ular soʻz yasash funksiyasini toʻxtatgan. [Hojiyev, 1989:96].

Kompozitsiya usuli bilan yasalgan qoʻshma soʻzlar hozirgi oʻzbek tilida anchagina miqdorni tashkil etadi. Shu bilan birga kompozitsiya usulida yangi-yangi soʻzlar yasalishi hozir ham davom etyapti. Ana shu ikki faktor asosida boʻlsa kerak, hozirgi oʻzbek tilidagi soʻz yasash usullari haqida gap borganida kompozitsiya usuli soʻz yasashining faol, mahsuldor usuli sifatida taʼriflanadi. Lekin kompozitsiya usulining aktivlik, mahsuldorlik darajasini belgilashda bu usul bilan yasalgan qoʻshma soʻzlarning miqdori emas, balki mazkur usulning hozirgi oʻzbek adabiy tili soʻz yasash sistemasida tutgan oʻrni, hozirgi soʻz yasashida ishtiroki, ish koʻrish darajasi asosga olinishi kerak boʻladi. [Hojiyev, 1989:98].



Keltirilgan qoʻshma ravishlar maʼlum model (andoza, sxema) asosida yasalsa ham, lekin ular soʻz yasash asosi va yasovchi formantdan iborat strukturaga ega emas. Chunki, ular tarkibida regulyar qatnashayotgan komponent (bir, sar, hech) soʻz

yasovchi (formant) ga xos funksiya va ma'noga ega emas - qo'shma ravishlar ma'nosini komponentlari ma'nosi bilan asoslab (motivlab) bo'lmaydi. Shu tufayli tarkibida regulyar qatnashuvchi komponenti bo'lgan bunday ravish turlari so'z (ravish) yasallishining ma'lum bir tipini hosil qilmaydi. ***Umuman, hozirgi o'zbek adabiy tilida kompozitsiya usuli bilan yangi ravishlar yasallishi kuzatilmaydi.*** [Hojiyev, 1989:109].

Yuqorida keltirib o'tgan fikrlarimizning yana bir dalili sifatida Azim Hojiyevning O'zbek tili morfologiyasi, morfemikasi va so'z yasallishining nazariy masalari-Toshkent, 2010-yilda nashrdan chiqqan asaridagi quyidagi fikrlarni yana bir bor keltirib o'tish lozimdir.

So'z yasallishi sistemasi. Til birligi-yasama so'z. So'z turkumlarida so'z yasallishi masalasi.

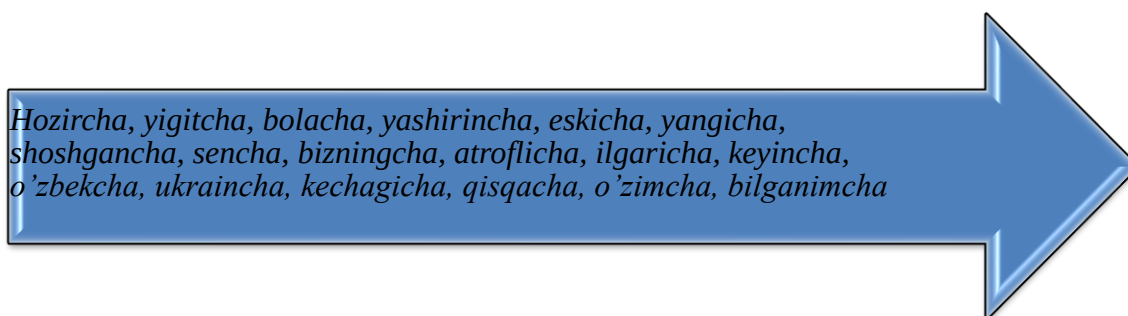
So'z yasallishi sistemasining faqat mustaqil so'z turkumlari uchun xos bo'lishi tabiiy holdir. Xuddi shu talabga ko'ra mustaqil so'z turkumlaridan son va olmosh turkumlari so'z yasallish hodisasiga, so'z yasallishi sistemasiga ega emas. (Shu boisdan ham o'quv yurtlari uchun chiqarilgan o'zbek tilidan qo'llanma va darsliklarda son va olmosh turkumlariga bag'ishlangan bo'limlarda so'z yasallishi hodisasi keltirilmaydi). Faqatgina ravish turkumida so'z yasallishi masalasida turli fikr mulohazalar davom etyapti.

Ma'lumki, oliy o'quv yurtlari uchun nashr etilgan o'zbek tili darsliklarining barchasida, boshqa mustaqil so'z turkumlari qatori, ravish yasallishi alohida mavzu sifatida beriladi. Demak, o'zbek tilida ravish yasallishi hodisasi borligi deyarli barcha mutaxassislar tomonidan tan olinadi. Biroq masalaga yuqorida aytilgan talablar asosida yondashilsa, o'zbek tilida ravish so'z turkumining so'z yasallish sistemasiga ega emasligi inkor etib bo'lmaydigan fakt ekanligi ma'lum bo'ladi. [Hojiyev, 2010:173].

Biror so'z turkumida so'z yasallishi sistemasi bor-yo'qligi haqida fikr yuritilar ekan, avvalo, bu so'z turkumida yasama so'zga xos mohiyatga ega bo'lgan til birligi, aniqrog'i, so'z yasallish asosi va so'z yasovchidan iborat tarkibga ega bo'lgan til birliklari-yasama so'zlar bor-yo'qligi belgilanishi kerak bo'ladi.

Masalaga shu tarzda yondashishning o'ziyoq o'zbek tilida ravish so'z turkumi so'z yasallishi sistemasiga ega emasligini ma'lum qilib qo'yadi. Buni o'zbek tilida eng faol, sermahsul ravish yasovchi deb ta'riflanayotgan birgina -cha affiksi misolida aniq ko'rish mumkin. O'zbek tili ravish sistemasi haqida fikr yuritilgan ishlarning barchasida -cha affiksining sermahsul ravish yasovchi ekanligi aytiladi va bu fikrni tasdiqlash uchun misollar keltiradi. Biroq keltirilgan misollar ularning tarkibidagi -cha affiksining ravish yasovchi umuman so'z yasovchi emasligini aniq tasdiqlaydi. [Hojiyev, 2010:174].

-cha. Bu affiks ot, sifat, olmosh va fe'llarga qo'shib, harakatning qay tarzda bajarilishi, harakat protsessi so'zning negizida ifodalangan predmet, holatga o'xshatilganligi ma'nolarini anglatgan ravishlar yasaydi:



*Hozircha, yigitcha, bolacha, yashirincha, eskicha, yangicha,
shoshgancha, sench, bizningcha, atroflicha, ilgarcha, keyincha,
o'zbekcha, ukraincha, kechagicha, qisqacha, o'zimcha, bilganimcha*

Diqqat bilan qaralsa, berilgan ta'rif va keltirilgan misollarning o'ziyoq -cha affiks so'z yasovchiga xos xususiyatlarga ega emasligini aniq ko'rsatadi. Masalan, -cha affiksli so'z "harakat prosessi so'zning negizida ifodalangan predmet, holatga o'xshatilganligi yoki bir belgining boshqa belgi bilan ma'nolarini anglatar" ekan, bu so'zlar tarkibidagi "cha" ni so'z yasovchi emasligini tasdiqlaydi. [Hojiyev, 2010:174].

Chunki o'xshatish, o'xshashligini ifodalash so'z yasovchilarga xos xususiyat emas, so'z yasovchi yangi lug'aviy ma'noli til birligi-yasama so'z hosil qiladi. Qolaversa, birinchidan, hozircha, yigitcha, bolacha, atroflicha so'zlari so'z yasash asosi+soz yasovchidan iborat tarkibli birlik-yasama so'z emas, ikkinchidan, shoshgancha, bizningcha, o'zimcha, bilganimcha, so'zlaridagi -cha so'z shakliga qo'shilyapti. So'z yasovchi birlik esa bunday xususiyatga ega emas.

Uchinchidan, ukraincha, bolacha, ilgargicha yasama so'z -lug'aviy birlik emas, shuning uchun izohli lug'atlarda bosh so'z (izohlovchi so'z) sifatida berilmaydi va hokazo. Xullas, o'zbek tilidagi -cha affiksi so'z yasovchiga xos xususiyatlarga ega emas. U o'z mohiyatiga ko'ra shakl yasovchi morfemalarga, ko'makchilarga o'xshaydi. Aniqroq qilib aytganda, o'zbek tilida birdan ortiq shakl yasovchi morfemaning o'zaro birikuvidan so'z yasovchi, jumladan, ravish yasovchi qo'shma affiksning yuzaga kelishi hodisasi yo'q va bo'lishi mumkin ham emas... [Hojiyev, 2010, 175].

Ravish yasalishi haqidagi gaplarni yakunlar ekanmiz, bir faktni alohida ta'kidlashni lozim topamiz. Qizig'i shundaki, o'zbek tili ravish yasalishi haqida fikr yuritilgan ishlarda yasama ravish deb keltirilgan fakt-misollarning barchasi o'zbek tilida ravish yasalishi hodisasi, ravish yasash sistemasi borligini emas, aksincha, yo'qligini isbotlashga xizmat qiladi. Bu holat qanday natijalar berishini o'zingiz tasavvur qilavering. Bunday holatning mavjudligi esa so'z yasalishi sistemasi til birligining mohiyati haqida aniq tasavvurga ega emaslik, ravish turkumida shunday til birligi bor-yo'qligini belgilashda yasama so'zga xos yuqorida ko'rilgan xususiyatlar asosida ish ko'rmaslik oqibatidir. [Hojiyev, 2010:176].

So'z yasalishi hodisasi mustaqil so'z turkumlarida mavjud. Mustaqil so'z turkumlarining ham ot, sifat, va fe'l turkumlari so'z yasalishi sistemasiga ega. Yasama so'zning va so'z yasalishi tiplarining mohiyati asosida ish ko'rilsa, buning

sabablari jumladan, o'zbek tilida ravish yasashi hodisasi yo'qligi, ravish turkumi so'z yasashi sistemasiga ega emasligi ham ma'lum bo'ladi. Shu asosda e'tibor bilan qaralsa, yasama ravish deb keltirilayotgan fakt-misollarning barchasi o'zbek tilida ravish yasashi hodisasi ham, yasama ravishlar ham yo'qligini tasdiqlaydi. [Hojiyev, 2010, 179].

Xulosa qilib shuni aytish mumkinki, ravish so'z turkumining yasama emasligining isboti bizga bir qancha analizatorlarda qulayliklar yaratadi. Ravish yasovchi qo'shimchalarning ba'zilar murakkab qo'shimchalar sirasiga kiradi. Va ularni tub va murakkablikdan aniq ajratib olishni mashinaga ayrim algoritmlar bilan ishlab chiqish muammoli masalalardan biri hisoblanadi. Do'stlarcha - do'stona, mardlarcha - mardona, vahshiylarcha - vahshiyona, yigitchasiga, mardchasiga, dehqonchasiga, chapanichasiga, tentakchasiga kabi ravishlarni mashinaga yaxlit holda yasama emas deb teglaganimiz bizga juda ko'p qulayliklarni keltirib chiqaradi va mashinaning aniq ishlashiga olib keladi desak xato bo'lmaydi. Chunki, do'stlarcha, vahshiylarcha, yigitchasiga, mardchasiga, dehqonchasiga kabi ravishlarni asos morfema va qo'shimcha morfemaga ajratishda bir qancha qiyinchiliklar kelib chiqadi. Shuning uchun mashinaga yuqoridagi nazariy fikrlarni hisobga olgan holda, ushbu ravishlarni yaxlit holda, ma'lum algoritmlar yordamida teglagan ma'qul desak xato bo'lmaydi.

Buni quyidagi misollar yordamida ham ko'rishimiz mumkin.



Morfologik analizator

Matnni kiriting

Hozircha, yigitcha, bolacha, yashirinch, eskicha, yangicha, shoshgancha, sencha, bizningcha, atroflicha, ilgaricha, keyincha, o'zbekcha, ukraincha, kechagicha, qisqacha, o'zimcha, bilganimcha

Analiz



№	Soʻz shakli		Lemma		Stem		Oʻzak		Asos va qoʻshimchalar
	Qiymati	Soʻz turkumi	Qiymati	Soʻz turkumi	Qiymati	Soʻz turkumi	Qiymati	Soʻz turkumi	
1	Hozircha	Ravish	Hozircha	Ravish ...	hozir	Ravish	hozir	Ravish	{Hozircha}
2	yigitcha	Ot	yigitcha	Ot, Ravish ...	yigitch		yigitch		{yigitcha}
3	bolacha		bolacha	Ot ...	bola	Ot	bola	Ot	{bolacha}
4	yashirincha		yashirincha	Ravish ...	yashir	Fe'l	yashir	Fe'l	{yashirincha}
5	eskicha		eskicha	Ravish ...	eski	Fe'l, Sifat	eski	Fe'l, Sifat	{eskicha}
6	yangicha	Ravish	yangicha	Ravish ...	yangi	Ravish, Sifat	yangi	Ravish, Sifat	{yangicha}
7	shoshgancha		shoshmoq	Fe'l ...	shosh	Fe'l	shosh	Fe'l	{shosh}-gancha
8	sencha		sencha	Ravish ...	sen	Olmosh	sen	Olmosh	{sencha}
9	bizningcha		bizon	Ot ...	bizon		bizn		{bizon}-ing-cha
10	atroflicha	Ravish	atroflicha	Ravish ...	atrof	Ot	atrof	Ot	{atroflicha}
11	ilgaricha		ilgaricha	Ravish ...	ilgari	Ravish	ilgari	Ravish	{ilgaricha}
12	keyincha		keyincha	Ravish ...	keyin	Ravish	keyin	Ravish	{keyincha}
13	o'zbekcha		o'zbekcha	Sifat ...	o'zbek	Ot	o'zbek	Ot	{o'zbekcha}
14	ukraincha		ukraincha	Ravish ...	ukrain	Ot	ukrain	Ot	{ukraincha}

Foydalanilgan adabiyotlar:

1. A.F.G'ulomov, Sh.U.Rahmatullayev, M.S.Sodiqova, M.A.Asqarova, A.P.Xojiyev, C.F.Fuzailov, Sh.Sh.Shoabdurahmonov, R.Q.Qo'ng'urov O'zbek tili grammatikasi-Toshkent-1975.



2. M. Asqarova, X. Abduraxmonov O'zbek tili grammatikasining praktikumi -Toshkent, 1981.
3. R.Sayfullayeva. Hozirgi adabiy o'zbek tili – Toshkent, 2009.
4. A.Hojiyev. O'zbek tili so'z yasalishi-Toshkent, 1989.
5. A.Hojiyev.O'zbek tili morfologiyasi, morfemikasi va so'z yasalishining nazariy masalari-Toshkent, 2010.
6. Elov B., Hamroyeva Sh., Axmedova X. Methods for creating a morphological analyzer. 14th International Conference on Intellegent Human Computer Interaction, IHCI 2022, 19-23 October 2022, Tashkent.
7. Abdullayeva O., Bozorqulova O'. Korpus lingvistikasida lingvistik tahlillar: muammo va yechim. // O'zbek tili tadqiqi va ta'limi muammolari mavzusidagi xalqaro ilmiy-amaliy anjumani materiallari. – Toshkent, 27-oktabr 2021. – B. 296-302.
8. Abdullayeva O., Nizomova F. O'zbek tilida sifat so'z turkumini modellashtirishda so'z yasalishi va omonimiya masalasi. // PROBLEMS OF RESEARCH AND EDUCATION OF THE UZBEK LANGUAGE mavzusidagi xalqaro ilmiy-amaliy anjumani materiallari, 23-dekabr 2022-yil. – B. 314-321.
9. <http://uznatcorpara.uz/>

AUTOMATIC DETECTION TECHNOLOGIES FOR STOPWORDS IN UZBEK LANGUAGE

Aziz Iskandarov

iskandarovazizbek7@gmail.com

Shermatov Boburjon

boburshermetov912@gmail.com

Elmurod Kuriyozov

elmurod1202@urdu.uz

Urgench State University

Annotatsiya. Nomuhim soʻzlar (stopwrods) tabiiy tilni qayta ishlashda (NLP) keng tarqalgan masala boʻlib, matnni tahlil qilish, maʼlumotlarni qidirish va boshqa NLP ilovalarida muammolarni keltirib chiqarishi mumkin. Ushbu tadqiqot ishida oʻzbek tilida nomuhim soʻzlarni avtomatik aniqlash texnologiyalarini ishlab chiqishga eʼtibor qaratilgan. Biz oʻzbek matnlarida nomuhim soʻzlarni aniqlash uchun mashinali oʻrganish (machine learning) algoritmlariga, xususan, Support Vector Machines (SVM) va Random Forest (RF) tasniflagichlariga asoslangan metodologiyani taqdim etamiz. Bizning metodologiya oʻzbek tilida oʻtkazilgan oldingi tadqiqotdan olingan “Maktab korpusi” deb nomlangan oʻzbek tilidagi matnlarning qoʻlda annotatsiya qilingan maʼlumotlar toʻplamidan foydalangan holda baholanadi va baholash natijalari yuqori aniqlik koʻrsatkichlariga erishdi. Natijalarimiz shuni koʻrsatadiki, taklif etilayotgan yondashuv oʻzbek matnlarida nomuhim soʻzlarni aniqlashning samarali usuli boʻlib, oʻzbek tilidagi matnlarni qayta ishlash uchun NLP ilovalari ish faoliyatini yaxshilashda ishlatilishi mumkin.

Kalit soʻzlar: *Oʻzbek tili, nomuhim soʻzlarni aniqlash, mashinali oʻqitish.*

Abstract. Stopwords are a common problem in natural language processing (NLP), which can cause issues in text analysis, information retrieval, and other NLP applications. This research paper focuses on the development of automatic detection technologies for stopwords in the low resource Uzbek language. We present an approach based on machine learning algorithms, specifically Support Vector Machines (SVM) and Random Forest (RF) classifiers, to identify stopwords in Uzbek texts. The approach is evaluated using a manually annotated dataset of Uzbek language texts called “School Corpus”, from a previously conducted research, and the evaluation results achieve high accuracy rates. Our results demonstrate that the proposed approach is an effective method for identifying stopwords in Uzbek texts and can be used to improve the performance of NLP applications for Uzbek language processing.

Keywords: *Uzbek language, Stopwords detection, Machine learning.*

Аннотация. Стоп-слова — распространенная проблема при обработке естественного языка (NLP), которая может вызывать проблемы при анализе

текста, поиске информации и других приложениях NLP. Это исследовательская работа посвящена разработке технологий автоматического обнаружения стоп-слов в узбекском языке. Мы представляем подход, основанный на алгоритмах машинного обучения, в частности на машинах опорных векторов (SVM) и классификаторах случайного леса (RF), для выявления стоп-слов в узбекских текстах. Подход оценивается с использованием вручную аннотированного набора данных текстов на узбекском языке под названием «Школьный корпус» из ранее проведенного исследования, и результаты оценки достигают высоких показателей точности. Наши результаты показывают, что предложенный подход является эффективным методом определения стоп-слов в узбекских текстах и может использоваться для повышения производительности приложений NLP для обработки узбекского языка.

Ключевые слова: *Узбекский язык, обнаружение стоп-слов, машинное обучение.*

1. Introduction.

Stopwords are commonly used words that are removed from text during NLP preprocessing, as they carry little meaning and can lead to noise in the analysis. However, the set of stopwords can vary depending on the language and domain of the text. In Uzbek language, there is a lack of comprehensive stopword lists available for NLP applications, which can affect the performance of text analysis and information retrieval tasks. Advancements in the field of Machine Learning [1] and Neural Networks [2], [3] make it possible to successfully identify stopwords from given texts.

In this paper, we propose a machine learning-based approach for automatic detection of stopwords in Uzbek language texts. Our approach utilizes Support Vector Machines (SVM) and Random Forest (RF) classifiers to identify stopwords. The proposed approach is evaluated using the already available annotated dataset of Uzbek texts called “School Corpus” [4], and we demonstrate its effectiveness in identifying stopwords in Uzbek texts.

Uzbek language. Uzbek is a Turkic language spoken by more than 30 million people mainly in Uzbekistan, but also in other Central Asian countries and by diaspora communities. It has a rich morphology and syntax, with complex inflectional and derivational systems, which can make language processing tasks challenging. The Uzbek language has its unique features and characteristics, such as vowel harmony and postpositions instead of prepositions. As there is a growing need for NLP applications to process Uzbek language data, developing effective techniques for tasks such as automatic detection of stopwords is crucial for improving the accuracy and effectiveness of these applications [5].

2. Related Work.

Several studies have been conducted on automatic stopword detection in different languages, including English, French, Chinese, and other languages [6], [7]. The most commonly used methods for stopword detection are rule-based and frequency-based approaches. However, these approaches require manual effort and do not always perform well, especially for languages with complex morphology and syntax.

Machine learning-based approaches have been shown to be effective in stopword detection tasks. In recent years, various machine learning algorithms, such as SVM, RF, and Naive Bayes (NB), have been applied to stopword detection tasks in different languages [8], [9].

Regarding the stopwords detection for Uzbek language, there has been a great deal of work, mostly focusing on the creating of the stopwords list for the language by Madatov et. al. [10], [11], which they introduce a methodology to extract Uzbek stopwords based on TF-IDF method. Apart from that, there has been a recent upward trend in creating NLP resources and tools for the Uzbek language, such as Latin-Cyrillic transliteration tool [12], datasets for sentiment analysis [13] and text classification [14], as well as tools for stemming [15], [16] and lemmatizers [17].

3. Dataset.

To perform our analysis on Uzbek stopwords, we used previously established ground rules and methodology for the Uzbek language stopwords by Madatov et. al. [18], who created the list of stopwords from their own dataset called “Uzbek Corpus”, which was constructed using online study materials comprising of 731,156 words, of which 47,165 are unique words.

We performed several preprocessing steps on the dataset, including the re-creation of annotated dataset by annotating the “School Corpus” they created using the above-mentioned unigrams. Further preprocessing steps took place, such as tokenization, stemming, and removing punctuation and numbers. We also converted all the text to lowercase to reduce the dimensionality of the feature space. The dataset was randomly split into training and testing sets with a ratio of 80:20.

4. Methodology.

4.1. Feature Extraction.

We used two feature extraction techniques to represent the text for classification: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). The BoW representation counts the occurrence of each word in the text, while TF-IDF measures the importance of a word in a document relative to the entire corpus.

4.2. Classification.

We used two classifiers, SVM and RF, to classify the texts into stopword and non-stopword categories, as the machine learning methods for classification prove themselves to perform good [8]. The SVM classifier was implemented with a linear kernel and a regularization parameter of 0.1. The RF classifier was implemented with 100 decision trees and a maximum depth of 10.



5. Evaluation and Results.

We evaluated the performance of our approach using several metrics, including accuracy, precision, recall, and F1-score. The results are presented in the following section.

Our experiments show that the SVM classifier achieved an accuracy of 98.2% and an F1-score of 0.98, while the RF classifier achieved an accuracy of 97.8% and an F1-score of 0.98. The precision and recall scores were also high, indicating that the classifiers performed well in identifying stopwords in Uzbek texts.

6. Conclusion and Future Work.

In this paper, we presented a machine learning-based approach for automatic detection of stopwords in Uzbek language texts. Our approach utilizes Support Vector Machines (SVM) and Random Forest (RF) classifiers to identify stopwords in Uzbek texts, and it was evaluated using a manually annotated dataset of 500 Uzbek texts. The results demonstrate that the proposed approach achieves high accuracy rates in identifying stopwords in Uzbek texts.

Our approach can be useful in various NLP applications for Uzbek language processing, such as text analysis, information retrieval, and text classification. By identifying and removing stopwords, the performance of these applications can be improved significantly.

Future work includes expanding the dataset and exploring other feature extraction techniques and machine learning algorithms to improve the performance of the approach. Additionally, the approach can be extended to other languages with similar characteristics to Uzbek, such as other Turkic languages. Overall, the proposed approach provides a useful tool for improving the accuracy and effectiveness of NLP applications for Uzbek language processing.

References

1. J. Mattiev and B. Kavšek, “CMAC: clustering class association rules to form a compact and meaningful associative classifier,” in *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part I* 6, 2020, pp. 372–384.
2. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.
3. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243.



4. K. Madatov, S. Bekchanov, and J. Vičič, “Accuracy of the Uzbek stop words detection: a case study on" School corpus",” *arXiv preprint arXiv:2209.07053*, 2022.
5. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings*, 2022, pp. 199–206. [Online]. Available: www.scopus.com
6. J. Savoy, “A stemming procedure and stopword list for general French corpora,” *Journal of the American Society for Information Science*, vol. 50, no. 10, 1999, doi: 10.1002/(SICI)1097-4571(1999)50:10<944::AID-ASI9>3.0.CO;2-Q.
7. L. Dolamic and J. Savoy, “Brief communication: When stopword lists make the difference,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, 2010, doi: 10.1002/asi.21186.
8. J. Mattiev, U. Salaev, and B. Kavsek, “Word Game Modeling Using Character-Level N-Gram and Statistics,” *Mathematics*, vol. 11, no. 6, p. 1380, 2023.
9. J. Mattiev and B. Kavšek, “ACHC: Associative Classifier Based on Hierarchical Clustering,” in *Intelligent Data Engineering and Automated Learning—IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings 22*, 2021, pp. 560–571.
10. K. Madatov, S. Bekchanov, and J. Vičič, “Lists of uzbek stopwords,” Univerza na Primorskem, Inštitut Andrej Marušič, 2021.
11. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.
12. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com
13. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com
14. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.
15. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a morphological and syntactic tagged corpus for the Uzbek language,” *arXiv preprint arXiv:2210.15234*, 2022.
16. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.



17. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” *arXiv preprint arXiv:2210.16006*, 2022.

18. X. Madatov, M. Sharipov, and S. Bekchanov, “O`ZBEK TILI MATNLARIDAGI NOMUHIM SO`ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021

O'ZBEK TILIDAGI MATNLARDA OT VA OLMOSH SO'Z TURKUMLARIGA TEGISHLI SO'ZLARNI ANIQLOVCHI ALGORITM YARATISH

Sharipov Maqsud Siddiqovich

Texnika fanlari nomzodi, PhD

maqsbek72@gmail.com

UrDU dotsenti

Mavlonova Mahliyo Mavlon qizi

mmm399197@gmail.com

UrDU magistranti

Sobirov Og'abek Ollayor o'g'li

sobirovogabek0409@gmail.com

UrDU magistranti

Annotatsiya. Ushbu maqolada o'zbek tili matnlarida ot va olmosh so'z turkumlarini aniqlashning qoidaga asoslangan algoritmini yaratish ko'rib chiqilgan. Ot va olmoshlarni matnlarda aniqlash va morfologik tahlil qilish o'zbek tili uchun tabiiy til jarayonlarining (TTJ) bir qator masalalarini (*POS tagging, mashinali tarjimon va h.k.*) yechishda foydali bo'ladi. Barcha turkiy tillar kabi o'zbek tili ham agglyutinativ til bo'lib so'z o'zak yoki negiziga qo'shimchalar qo'shilishi orqali yasaladi. Bu masalani yechish uchun tub bo'lgan ot va olmosh so'zlar bazasi yaratildi, bundan tashqari ot va olmoshning so'z yasovchi va shakl yasovchi qo'shimchalar bazasi shakllantirildi.

Kalit so'zlar. *Tabiiy til jarayonlari (TTJ), teglash, so'z turkumlari, ot, olmosh, asos, qo'shimchalar, lemma, lemmatizatsiya, Chekli holatlar mashinasi (ChHM).*

Annotation. This article deals with the creation of a rule-based algorithm for identifying noun and pronoun word classes in Uzbek language texts. Identification and morphological analysis of nouns and pronouns in texts will be useful in solving a number of issues (*POS tagging, machine translation, etc.*) of natural language processes (NLP) for the Uzbek language. Like all Turkic languages, the Uzbek language is an agglutinative language, formed by adding suffixes to the root or base of the word. To solve this problem, a database of basic nouns and pronouns was created, in addition, the database of word-forming (derivational affixes) and form-forming (inflectional affixes) of nouns and pronouns was created.

Key words. *Natural Language Processing (NLP), POS-tagging, word classes, noun, pronoun, stem, affixes, lemma, lemmatization, Finite State Machine (FSM).*

Аннотация. В данной статье речь идет о создании основанного на правилах алгоритма определения классов имен существительных и местоимений в текстах на узбекском языке. Выявление и морфологический анализ существительных и местоимений в текстах будет полезен при решении

ряда вопросов (POS-тегирование, машинный перевод и др.) обработка естественного языка (ОЕЯ) для узбекского языка. Как и все тюркские языки, узбекский язык является агглютинативным языком, образованным путем добавления суффиксов к корню или основе слова. Для решения этой задачи была создана база данных основных существительных и местоимений, кроме того, создана база данных словообразовательных (словообразовательных аффиксов) и формообразующих (флективных аффиксов) существительных и местоимений.

Ключевые слова. Обработка Естественного Языка, тегирование части речи, классы слов, существительное, местоимение, основа, аффиксы, лемма, лемматизация, *Конечный Автомат (КА)*.

Kirish. Matnlarda soʻzlarni qaysi soʻz turkumiga tegishli ekanini aniqlovchi algoritmi ishlab chiqish tabiiy til jarayonlarining (TTJ) asosiy jarayonlaridan biridir. Ushbu maqolada oʻzbek tili matnlarida ot va olmosh soʻz turkumlarini aniqlashning qoidaga asoslangan algoritmini yaratish koʻrib chiqilgan. Bunda ushbu masalani yechish uchun oʻtning *soʻz yasovchi* va *shakl yasovchi* qoʻshimchalar bazasi yaratildi, bundan tashqari tub boʻlgan *olmosh* soʻzlar bazasi shakllantirildi va shular asosida qoidaga asoslangan Chekli Holatlar Mashinasi (ChHM) yordamida algoritm ishlab chiqildi.

Tabiiy til jarayonlarining bir qancha sohalarida koʻplab ishlar olib borilgan, biz bu boradagi bir nechta ishlarni koʻrib chiqdik. Misol uchun: matnlarni teglash qoidalarga asoslangan tizimlardan ishlab chiqiladi (Kupiec, 1992), sintaktik tahlil qilingan datasetning yaratilishi (Sharipov et al., 2022), gap boʻlaklarini ajratib beruvchi dataset va dasturiy vositasi (Sharipov et al., 2023), soʻzlar oʻxshashlik va bogʻliqlik darajalarini oʻz ichiga olgan semantik tahlil dataseti (Salayev et al., 2022), soʻz lemmasini topish algoritmi (Sharipov & Sobirov, 2022). Ushbu ishlar qatorida matnlarda soʻzlarni qaysi soʻz turkimiga tegishli ekanini aniqlash tabiiy til jarayonlarining yana bir muhim masalasi ekanini koʻrishimiz mumkin.

Metodologiya. Tadqiqot maqsadi va vazifalari, oʻrganilganlik darajasini aniqlash maqsadida ilmiy yoʻnalishlarda erishilgan yutuqlar, shu bilan birga ushbu mavzu yuzasidan maʼlumotlar toʻplashdan ishiga kirishildi. Maʼlumotlar bazasi “Hozirgi oʻzbek adabiy tili”(SAYFULLAYEVA et al., 2009) va “Oʻzbek tilidan maʼruzalar toʻplami”(Erkaboeva, 2016) darsliklariga murojaat qilindi. Otlar haqida maʼlumotlar toʻplash ishlari tugagach, Oʻzbek tilining izohli lugʻati 1-2-3-4-5-jildlaridan ot va olmosh turkumiga doir soʻzlar ajratib olindi. Oʻzbek tili izohli lugʻatining 1-2-3-4-5-jildlaridan umumiy hammasi boʻlib, 14154 ta oʻtning izohi bilan jadvashtirildi. Ajratilgan maʼlumotlar bazasi txt. fayl sifatida baza qilib kiritildi. Tadqiqotning keyingi bosqichida ot va olmosh soʻz turkumlariga doir oʻziga xos xususiyatlar kiritib chiqildi. Yaʼni ot va olmosh soʻz turkumlari asos va soʻz yasovchi, lugʻaviy shakl yasovchi va sintaktik shakl yasovchi qoʻshimchalarga ajratib chiqildi. Oʻzbek tilida jami soʻz yasovchi qoʻshimchalar, taxminan, 150 dan



	qo'shi m- chalari							Birlilik	Ko'plik		
Ot	-kor -chi -zor -dosh	- n ik i	- da gi	- gac ha	- da y - de k	- cha(k) - (a)lo q - chOq (- chiq) - (b)oy -bek -bibi - Xon(-jon) -gina - posh sha	- l a r	- (i) m - (i) ng - (s) i	- (i)miz - (i)ngi z -lari	-ning -ni -ga -da -dan	Paxtakor niki Gulchidek Gulzorlri mizning Sinf doshd ekmiz

1-jadval. Ot so'z turkumining qo'shimchalari.

Kishilik	Ko'rsatish	Belgilash (har bilan bog'liq barcha so'zlar)	Bo'lishsizlik (hech bilan bog'liq barcha so'zlar)	So'roq	O'zlik	Gumon
men	u	hamma	hech nima	kim	o'z	kimdir
sen	bu	barcha	hech bir	nima	o'zimga	nimadir
u	shu	bari	hech kim	qayer	o'zimda	qandaydir
biz	o'sha	jami	hech qanday	qanday	o'zimdan	qayerdir
siz	ana	jamiyki	hech qayer	qanaqa	o'zimniki	qaysidir
ular	mana			qaysi	o'zimnikiga	nechtadir

2-jadval. Olmosh so'z turkumining tub so'zlari.

Xulosa qilib aytish mumkinki, ushbu ma'lumotlar bazasi bizga ot va olmosh so'z turkumiga oid so'zlarni aniqlashda asos bo'lib xizmat qiladi. Ushbu ma'lumotlar bazasidan foydalangan holda Python dasturlash tilida algoritm tuzish maqsad.

Natija. Ushbu ish orqali ishlab chiqilgan algoritmning ChHM chizmasi maqola oxirgi betida keltirilgan.

Xulosa. Tabiiy til jarayonlari masalalarida so'zlarni qaysi so'z turkumiga tegishli ekanini aniqlash masalasi asosida o'zbek tili matnlarida ot va olmosh so'z turkumlarini aniqlash masalasi ko'rib chiqildi, kelajakda ushbu masala dasturini tuzish reja qilindi.

Foydalanilgan adabiyotlar:

1. Erkaboeva, N. (2016). O'zbek tilidan ma'ruzalar to'plami. Akademnashr.
2. Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, 6(3), 225–242.
3. Salaev, U., Kuriyozov, E., & Gómez-Rodríguez, C. (2022). SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language. 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - Held in Conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 199–206. www.scopus.com
4. Sayfullayeva R., Mengliyev B., Boqiyeva G., Qurbonova M., Yunusova Z., Abuzalova M. Hozirgi o'zbek adabiy tili. -Toshkent, 2009.
5. Sharipov, M., Kuriyozov, E., Yuldashev, O., Sobirov, O. (2023). UzbekTagger: The rule-based POS tagger for Uzbek language. *ArXiv Preprint ArXiv:2301.12711*.
6. Sharipov, M., Mattiev, J., Sobirov, J., & Baltayev, R. (2022). Creating a morphological and syntactic tagged corpus for the Uzbek language. *ArXiv Preprint ArXiv:2210.15234*.
7. Sharipov, M., & Sobirov, O. (2022). Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language. *ArXiv Preprint ArXiv:2210.16006*

O'ZBEK TILIDAGI FE'LLI BIRIKMALARNING FORMAL MODELLARI

Xajiyeva Durdona Shonazarovna

Urganch davlat universiteti magistri, durdonaxajiyeva@gmail.com

Jo'rayeva Nilufar Vayitovna

Geologiya fanlari universiteti dotsenti, minauzb@mail.ru

Annotatsiya. Bu maqolada fe'lli birikmaning formal modellari ifodalangan bo'lib, kompyuter lingvistikasi fanining avtomatik tarjima, avtomatik sintaktik tahlil va morfologik tahlil hamda NLP – tabiiy tilning qayta ishlanishi yo'nalishlarida lingvistik baza sifatida ishlatilishi mumkin.

Kalit so'zlar: *Kompyuter lingvistikasi, formal grammatika, sintaksis, so'z birikmasi (otli birikma, fe'lli birikma).*

Abstrakt. The article describes formal subject models that can be used as a linguistic base in the field of automatic translation of computational linguistics, automatic parsing and NLP - natural language processing.

Key words: *Computational linguistics, formal grammar, syntax, word combinations.*

Аннотация. В данной статье представлены формальные модели сочетания глаголов, которые могут быть использованы в качестве лингвистической базы в областях автоматического перевода, автоматического анализа синтаксиса и НЛП - Нейролингвистическое программирование.

Ключевые слова: *Компьютерная лингвистика, формальная грамматика, синтаксис, словосочетания.*

Kirish. Kompyuter texnologiyalari tilshunoslikka ham kirib keldi. Mashina tarjimasi deyilganda, bir tildagi matnni ikkinchi bir tilga kompyuter yordamida, tez tarjima qilish nazarda tutiladi. Tarjima muammolarni hal qilishda bir nechta metodlar ishlab chiqilgan. Mazkur ishda shu metodlardan biri formal model tuzish metodidan foydalanilgan. Albatta, olingan natijalar kompyuter lingvistika fani uchun aniqrog'i o'zbek kompyuter lingvistikasi uchun juda dolzarb, uning rivojlanish uchun muhim va ayni damda kerak bo'lgan ma'lumotlar hisoblanadi.

O'zbek tilining formal grammatika bo'yicha ma'lumotlar ko'p emas. Shulardan A.Po'latovning "Kompyuter lingvistikasi" kitobini va N.V.Jo'rayevaning tadqiqot ishlarini aytib o'tish mumkin va mazkur ishda bu yerda keltirilgan ma'lumotlardan va berish uslubidan foydalanilgan. Avval ayrim tushunchalarga formal nazariya nuqtayi nazarida qisqacha ta'rif berib o'taylik.

Tabiiy til grammatikasida: Soʻz - tilning narsa hodisalar, jarayon va xususiyatlarini nomlash uchun xizmat qiladigan eng muhim struktura maʼnoviy birligi.

Sintaksis - yunoncha “tuzish”, “biriktirish”, “yopishtirish” degan maʼnoni anglatadi. Biror tildagi soʻzlarning soʻz birikmalariga va gaplarga birikish usullari aynan shu boʻlimda oʻrganiladi.

Asosiy qism. Feʼlli birikma. Feʼlli soʻz birikmasida hokim soʻz feʼl bilan ifodalanadi: daftarni ochmoq, kitob varaqlamoq, tez gapirmoq, zavq bilan kulmoq.

Soʻz birikmasi tobe va hokim soʻzning qaysi vosita orqali bogʻlanishiga koʻra 3 xil boʻladi: 1) boshqaruv; 2) moslashuv; 3) bitishuv.

Boshqaruv

Tobe soʻzning hokim soʻzga tushum, joʻnalish, oʻrin-payt va chiqish kelishigi qoʻshimchalari yoki koʻmakchilar yordamida bogʻlanishi boshqaruv deyiladi.

Masalan: maktabda oʻqimoq, uyga ketmoq.

Feʼl boshqaruvi		
Uzum yemoq	Uzumni yemoq	Uzumdan yemoq

Tobe boʻlak hokim boʻlakka koʻmakchilar yordamida ham bogʻlanadi.

Masalan: kelajak haqida suhbatlashmoq, telefon orqali gaplashmoq.

Kelishik va qoʻshimchalar sinonimiyasi			
Kelishiklar orqali ifodalangan grammatik maʼno aniq	Koʻmakchilar orqali ifodalangan grammatik maʼno mavhum		
Uyga ketmoq	Uy tomon ketmoq	Uyga tomon ketmoq	Uy tomonga ketmoq
Maktabdan chiqmoq	Maktab tomon ketmoq	Maktabga tomon ketmoq	Maktab tomondan chiqmoq

Baʼzan kelishik va koʻmakchilar birgalikda qoʻllaniladi. Masalan, dala tomon ketmoq - dalaga tomon ketmoq - dala tomonga ketmoq.

Bitishuv

Soʻz birikmalari tarkibidagi tobe va hokim soʻzning hech qanday qoʻshimchalarsiz, faqat ohang va maʼnoga koʻra bogʻlanishi bitishuv deyiladi.

Masalan: tez termoq, chiroyli gapirmoq.

Moslashuv.

Tobe soʻzning hokim soʻzga qaratqich kelishigi qoʻshimchasi yordamida, hokim soʻzning esa tobe soʻzga egalik qoʻshimchalari yordamida bogʻlanishi moslashuv deyiladi.

Masalan: kitobning varagʻi, bizning gulimiz.

Moslashuv asosida otli birikma hosil boʻladi.

Fe'lli birikmada so'zlar o'zaro bitishuv va boshqaruv asosida birikadi.

Fe'lli birikmaga formal nazariya sifatida ta'rifini uning formal modellarini tushunish kerak. Tobe so'zning sintaktik tahlilini va uning turli so'z turkumlari bilan ifodalanishini ko'rib chiqamiz. Bunda biz oddiy tahlil uslubidan foydalanib, uni kompyuter tiliga moslashtiramiz.

Fe'lli birikma				
	So'z turkumi	Turi	Shakli	Misol
	Ot	Atoqli ot	$So'z + fe'l$	Atirgul olmoq
			$So'z + k.q + fe'l$	Alisherni kutmoq
		Turdosh ot: shaxs oti	$So'z + fe'l$	Chol qimirlamoq
			$So'z + k.q + fe'l$	Qizga gapirmoq
		narsa oti	$So'z + fe'l$	Kitob o'qimoq
			$So'z + k.q + fe'l$	Gulni hidlamoq
		o'rin-joy oti	$So'z + k.q + fe'l$	Uyga kirmoq
			$So'z + fe'l$	Uy olmoq
		faoliyat – jarayon oti	$So'z + fe'l$	Terimchilik qilmoq
			$So'z + k.q + fe'l$	Terimchilikni o'rganmoq
		Aniq ot	$So'z + k.q + fe'l$	Uyni tozalamoq
			$So'z + fe'l$	Kema qurmoq
		Mavhum ot	$So'z + k.q + fe'l$	Borliqni anglamoq
			$So'z + fe'l$	Hayol surmoq
		Sodda ot	$So'z + k.q + fe'l$	Bog'bonga bermoq
			$So'z + fe'l$	Bog'bon qilmoq
		Qo'shma ot	$So'z + k.q + fe'l$	Taklifnomani olmoq
			$So'z + fe'l$	Guldasta yasamoq
		Juft ot	$So'z + fe'l$	Ota-ona bo'lmoq
			$So'z + k.q + fe'l$	Ota-onani sevmok
		Takror ot	$So'z + fe'l$	Parcha-parcha qilmoq
			$So'z + k.q + fe'l$	Choy –poyni damlamoq
			-----	----- -----Θ Gul-gul qilmoq (ravish)
	Sifat	Asliy sifat	$So'z + fe'l$	chiroyli kiyinmoq
		Nisbiy sifat	-----	-----
	Sifat (otlashganda)		$So'z + k.q + fe'l$	Yaxshilikni ulashmoq
			$So'z + fe'l$	Yomonlik qilmoq



Son	Tartib son	<i>So 'z +fe' l</i>	Beshinchi bo' lmoq
		<i>So 'z +k.q +fe' l</i>	Beshinchini tanlamoq
	(Miqdor son): sanoq son	<i>So 'z +k.q +fe' l</i>	Uchga ko' paytirmoq
		<i>So 'z +fe' l</i>	Yetti o' lchamoq
	dona son	<i>So 'z +fe' l</i>	Yettita bermoq
		<i>So 'z +k.q +fe' l</i>	Yettitani bermoq
	taqsim son	<i>So 'z +fe' l</i>	Uchtadan bermoq
	chama son	<i>So 'z +fe' l</i>	Beshtacha qolmoq
		<i>So 'z +k.q +fe' l</i>	Beshtachani solmoq
	jamlovchi son	<i>So 'z +fe' l</i>	Ikkovi kelmoq
		<i>So 'z +k.q +fe' l</i>	Ikkoviga olmoq
	Kasr son	<i>So 'z +k.q +fe' l</i>	Yarmini topmoq
		<i>So 'z +fe' l</i>	Yarmi qolmoq
	Sodda	<i>So 'z +k.q +fe' l</i>	Beshni ayirmoq
		<i>So 'z +fe' l</i>	Bir qilmoq
	Qo' shma	<i>So 'z +k.q +fe' l</i>	O' n ikkiga ko' paytirmoq
		<i>So 'z +fe' l</i>	O' n ikki qilmoq
Olmosh	O' zlik olmoshi	<i>So 'z +fe' l</i>	O' zi bajarmoq
		<i>So 'z +k.q +fe' l</i>	O' zini o' ldirmoq
	Kishilik olmoshi	<i>So 'z +fe' l</i>	Seni kutmoq
		<i>So 'z +k.q +fe' l</i>	unga bermoq
	Ko' rsatish olmoshi	<i>So 'z +k.q +fe' l</i>	Buni bilmoq
	So' roq olmoshi	<i>So 'z +k.q +fe' l</i>	qayerga bormoq
	Belgilash olmoshi	<i>So 'z +k.q +fe' l</i>	Hammaga aytmoq
	Bo' lishsizlik	<i>So 'z +k.q +fe' l</i>	Hech qayerni tozalamasdi
	Gumon olmoshi	<i>So 'z +fe' l</i>	Qachondir kelmoq
		<i>So 'z +k.q +fe' l</i>	Allakimdan qo' rqmoq
	sodda	<i>So 'z +k.q +fe' l</i>	Hammani belgilamoq
	Qo' shma	<i>So 'z +k.q +fe' l</i>	Har nimani demoq
	juft	<i>So 'z +k.q +fe' l</i>	Siz-u bizga ishonmoq

		takror	<i>So 'z + k.q + fe' l</i>	Hamma-hammasini tugatmoq
	Ravish	Sabab maqsad ravishi	<i>So 'z + fe' l</i>	Ataylab sotmoq
		Holat ravishi	<i>So 'z + fe' l</i>	Qo'qqisdan kirmoq
		Payt ravishi	<i>So 'z + fe' l</i>	Hamisha kutmoq
		Miqdor daraja ravishi	<i>So 'z + fe' l</i>	Qittay ichmoq
		O'rin ravishi	<i>So 'z + fe' l</i>	shu topda kulmoq
	Fe'l		<i>So 'z + fe' l</i>	Kulib gapirmoq
	Modal		<i>So 'z + fe' l</i>	Aftidan demoq
	Modal (vazifadosh)		<i>So 'z + k.q + fe' l</i>	Aftidan bilinmoq
	Taqlid		<i>So 'z + fe' l</i>	Pitir-pitir qilmoq
	Undov		<i>So 'z + fe' l</i>	Uf demoq
	Birikma li ifodalar		<i>So 'z + k.q + so 'z</i>	Hujjatni imzolamoq
				Hujjatga imzo chekmoq
	Turg'un bog'lanma	Fraza (Ibora)	<i>So 'z + k.q + fe' l</i>	Qulog'iga quymoq (dorini qulog'iga quymoq)
		-----		Qulog'iga quymoq (gapini qulog'iga quymoq)

k.q - kelishik qo'shimchasi (-ning, -ni, -ga(-ka,-qa), -da,-dan) .

Kishilik olmoshi: *men, sen, u, biz, siz, ular.*

O'zlik olmoshi: *o'zi.*

So'roq olmoshlari: *kim, nima, qanday, qanaqa, qaysi, nechta, nechanchi, qancha, nima uchun, nega, nima sababdan, nima maqsadda, qayerda, qani, qayerga, qachon, qachondan.*

Belgilash olmoshlari: *bari, barcha, hamma, butun, ba'zi, jami, hamma-hammasi, har bir, har qanday, har narsa, har kishi.*

Ko'rsatish olmoshi: *bu, shu, u, o'sha, ana, mana, mana bu, mana shu, ana bu, ana shu, u-bu, ul-bul, uncha-buncha, shu-shu, o'sha-o'sha, uncha-muncha.*

Gumon olmoshlari: *kimdir, qachondir, qandaydir, qaysidir, nimadandir, allaqaysi, allanarsa, allakim, allaqanday, allaqachon, allaqancha, bir kishi, bir narsa, bir nima.*

Bo'lishsizlik olmoshi: *hech kim, hech nima, hech qanday, hech qayerda.*

Holat ravishi: *jim, yayov, chindan, qo'qqisdan, piyoda, majburan, tasodifan, astoydil, eskicha, qatorsiga, birma-bir, yakkama-yakka, yonma-yon, bafurja, arang, zo'rg'a, mardlarcha, yaqqol, yangicha, o'rtacha, bazo'r, bazur, bemalol,*

beomon, beto'xtov, to'xtovsiz, bexosdan, tez, asta, sekin, ohista, birdan, birdaniga, bir-bir, omon-eson.

Payt ravishi: *endi, hozircha, so'ngra, hanuz, hamisha, doimo, hamon, dastlab, erta-indin, qishin-yozin, kecha-kunduz, bugun, ertaga, ertalab, ilgari, oldin, qadim, so'ng, keyin, avval, burun, boya, yana, kecha, indin, allaqachon, har doim, har kuni, har gal, har lahza, har yili, har doim, har onda, har fursatda, shu onda, shu topda, shu lahzada, shu fursatda, hozir.*

O'rin ravishi: *olg'a, ichkari, tashqari, nari, beri, uzoqdan, yaqindan, olisda, yuqorida, pastda, u yoqda, bu yoqqa, u yerda, bu yerda, har yerda, har joydan, oldindan.*

Miqdor - daraja ravishi: *sal, oz, picha, ko'p, kam, ancha, xiyol, xiyla, bioz, bir talay, jinday, eng, nihoyatda, juda sal-pal, bag'oyat, qittay.*

Maqsad-sabab ravishi: *ataylab, atayin, atay, jo'rttaga, qasddan, azza-bazza, noiloj, noilojlikdan, chor-nochor, ilojsizlikdan.*

Modal so'zlar: *albatta, balki, shaksiz, so'zsiz, shubhasiz, ehtimol, chamasi, shekilli, aftidan, darhaqiqat, haqiqatdan, xullas, demak, umuman, avvalo, avvalambor, masalan, jumladan, attang, afsus, xayriyat.*

Taqlid so'zlar: *g'a-g'a, di-dit, qiy-chuv, qars-qurs, piq-piq, piqir-piqir, duk-duk, ship-ship, miyov, maa, mo'o', taraq-turuq, lip-lip, jimir-jimir, yarq, yilt, yal-yal, dir-dir, qalt-qalt, xippa, g'ippa, gup-gup, duv-duv, shart, cho'rt, paqqa, shig', jiz, jaz, shuv, chirt, lik-lik, gir-gir, yalt-yalt, yilt-yilt, lapang-lapang, apil-tapil, dikang-dikang, mo'lt-mo'lt, lo'q, lang, dang, ding, pag'a-pag'a, ong-tong, hang-mang, vij-vij, vijir-vijir, bij-bij, bijir-bijir.*

Undov so'zlar: *uf, oh, iye, o'h-ho', bay-bay, o, ur-ree, e, voy, hey, tiss, ey, hoy, xo'sh, chuh, beh-beh, pish-pish, drr, chip-chip, salom, xayr, alvido, rahmat, qulluq, allo, marhamat, balli, barakalla, tasanno, ofarin.*

Muhokama. Jadvalda keltirilgan ma'lumot fe'lli birikmaning formal ko'rinishi sifatida olsa bo'ladi. Formal grammatikada, til birliklarini jadval yoki formula bilan ifodalaydi. Lekin formal nazariya yozilishda o'zimiz uchun to'g'ri ketma ketlikni aniqlashtirib olishimiz kerak bo'ladi. To'g'ri so'z birikmasi strukturasi sifatida T.S (tobe so'z) +k.q (yoki ko'makchi)+H.S (hokim so'z) ketma - ketlikni olamiz. Va shu struktura asosida so'z birikmasini tuzamiz. Bu yerda har bir tobe so'z qaysi so'z turkumi bo'lishidan qat'iy nazar hokim so'z faqat bir so'zdan iborat bo'lishi kerak. Mazkur ishda faqat bir so'z bilan ifodalangan holat qaraldi.

Xulosa. So'z birikmasi sifatida turli so'z turkumlari bo'la oladi. Turli so'zlarni tobe so'z sifatida kelishligi uchun u otlashgan shaklda bo'lishi kerak va ko'plik qo'shimchasi, egalik qo'shimchasi va qo'shimchasiz shaklda kela oladi. Yoki shu qo'shimchalar kombinasiyasi shaklida ham kelishi mumkin. Formal nazariya jihatdan, tobe so'zning belgisiz holatida kelganda uning joylashgan o'rni va qaysi so'z turkumi bilan ifodalanganligini aniqlash mumkin.

REFERENCES



1. Botirboev F.P., Juraeva N.V. O'zbek tilidagi qaratqich aniqlovchining formal modellari. Экономика и Социум. № 4(83)
2. Жўраева, Н. В., Султанов, Р. О., Абдуллаева, С. А., Рахимжонов, В. А. (2020). Systematization of word combinations in the uzbek language. Наука и Мир, 2(6), 65- 68.
3. Po'latov A. 2011. Kompyuter lingvistikasi. Toshkent: Akademnashr.
4. Rahimov A. 2011. Kompyuter lingvistika asoslari. Toshkent: Akademnashr.
5. Juraeva N.V. «The formal model of the grammar of the uzbek language» //European Journal of Research and Reflection in Educational Sciences// Vol.8 No. 4, 2020 Part II ISSN 2056-5852
6. Juraeva N.V., Dustmurodova G.X., Abdullaeva Sayora //Rasmiy-idoraviy uslub grammatikasini kompyuterga kiritishdagi kriteriyalar // "Ilm fan va ta'limning rivojlanish istiqbollari" mavzusidagi ko'p tarmoqli, ilmiy, masofaviy, onlayn konferensiyasi. 27.04.2020
7. R.B., Madaminova N.A.// O'zbek tilidagi "NG" harf birikmasi elektron bazasini tuzishda klassifikatsiyalash va sistemalash muammolari// "Til va adabiyot ta'limi" O'zbekiston respublikasi xalq ta'limi vazirligining ilmiy-metodik jurnali. 2020 yil. 7- son/ ISSN 2010-5584
8. Jurayeva N.V., Ismoilova S.A., Botirboyev F.P. // Eganing formal modellarining kompyuter bazasi // "ACADEMIC RESEARCH IN EDUCATIONAL SCIENCES" nomli konferensiyasi// VOLUME 2 | ISSUE 5 | 2021 ISSN: 2181-1385//2021
9. Erkaboyeva N. O'zbek tilidan ma'ruzalar to'plami. -Toshkent, 2021. Yosh kuch.

UDK: 811.512.133`42

DESIGNING A TREE-BASED MODEL FOR ANALYZING THE SYNTAX OF UZBEK LANGUAGE SENTENCES

Gayrat Matlatipov

Urgench State University

gayrat@urdu.uz

Mardona Karimova

Master student at Urgench State University

karimovam95@mail.ru

Ulugbek Salaev

Urgench State University

ulugbek.salaev@urdu.uz

Abstract: Syntax analysis is a fundamental component of natural language processing that enables machines to understand the structure of sentences. In this paper, we propose a tree-based syntax analysis method using context-free grammar for the Uzbek language, which is a low-resource language with complex morphology and syntax. We developed a context-free grammar for the Uzbek language and created a parser that generates a syntax tree for a given sentence. This work contributes to the development of natural language processing tools for Uzbek language, which can facilitate the analysis of Uzbek text data.

Keywords: *context free grammar, syntax analysis, parser tree, Uzbek language*

Abstrakt: Sintaksis tahlili tabiiy tilni qayta ishlashning asosiy komponenti bo'lib, u mashinalarga jumlarar tuzilishini tushunishga imkon beradi. Ushbu maqolada biz murakkab morfologiya va sintaksisga ega kam resursli til bo'lgan o'zbek tili uchun kontekstsiz grammatikadan foydalangan holda daraxtga asoslangan sintaksik tahlil usulini taklif qilamiz. Biz o'zbek tili uchun kontekstsiz grammatikani ishlab chiqdik va berilgan gap uchun sintaksis daraxtini iyerarxiyalash jarayonlari qaralgan. Ushbu ishda o'zbek tilidagi matn ma'lumotlarini tahlil qilishni osonlashtiradigan o'zbek tili uchun tabiiy tillarni qayta ishlash vositalarini ishlab chiqishga yordam beradi.

Kalit so'zlar: *kontekstsiz grammatika, sintaksis tahlili, iyerarxiyalash, o'zbek tili.*

Аннотация: Синтаксический анализ является фундаментальным компонентом обработки естественного языка, который позволяет машинам понимать структуру предложений. В этой статье мы предлагаем метод анализа синтаксиса на основе дерева с использованием контекстно-свободной грамматики для узбекского языка, который является мало ресурсным языком со сложной морфологией и синтаксисом. Мы разработали контекстно-

свободную грамматику для узбекского языка и создали парсер, который генерирует синтаксическое дерево для заданного предложения. Наша работа способствует развитию инструментов обработки естественного языка для узбекского языка, которые могут облегчить анализ текстовых данных на узбекском языке.

Ключевые слова: контекстно-свободная грамматика, синтаксический анализ, дерево синтаксического анализа, узбекский язык.

One of the important tasks in NLP is syntax analysis, which involves parsing sentences to understand their structure and relationships between words. However, there is a lack of resources and research in this area for the Uzbek language. In this paper, we proposed a tree-based syntax analysis model for the Uzbek language using context-free grammar (CFG). We also provide a detailed process of implementing the CFG for a sample Uzbek sentence.

The process of performing the task of creating a tree-based syntax analysis model using a context-free grammar for the Uzbek language involves the following steps:

- Developing a context-free grammar (CFG) for the Uzbek language: This involves identifying the various parts of speech, such as nouns, verbs, adjectives, etc., and their corresponding rules and structures.
- Implementing the CFG in a parser: A parser is a tool that can read a sentence and determine its syntactic structure according to the CFG. There are several parser options available.
- Testing the parser: The parser is tested using sample sentences in the Uzbek language to ensure that it can accurately determine their syntactic structure.
- Constructing the syntax tree: Once the parser has determined the syntactic structure of a sentence, a syntax tree can be constructed to represent the hierarchical relationships between the various parts of speech in the sentence.

The (Almamatova, 2022) provides a comprehensive overview of the study of phrasemes in Uzbek linguistics, including their syntactic and semantic properties. The analysis of phrasemes is essential for understanding the structure of sentences in Uzbek language and can help improve natural language processing tasks. The (Sharipov et al., 2022) works have also attempted to create tagged corpora for low-resource languages using different methodologies and tools. Some of these works include the creation of tag sets and annotation tools specific to the target language.

Parsing algorithms provide a way to recognize the strings of a language and assign each string a syntactic analysis. However, before we can analyze sentences in Uzbek, we must first separate them into word groups. To accomplish this, we define word groups based on the structure of the Uzbek language. Additionally, we define part-of-speech tags for each word in the sentence to further aid in the analysis process.

Table 2. Part-of-speech tags for Uzbek language words

Tag	POS (Uzbek)	POS (English)	Sample
N	Ot	Noun	kitob
V	Fe'l	Verb	aytmoq
ADJ	Sifat	Adjective	ko'p
ADV	Ravish	Adverb	sekin
PRN	Olmosh	Pronoun	bu
CONJ	Bog'lovchi	Conjunction	bilan, va
NUM	Son	Numeric	yetti
PART	Yuklama	Particle	nafaqat
IMIT	taqlid	imitation	tuq-tuq
INTJ	undov	interjection	urra
MOD	modal	modal	darvoqe
AUX	ko'makchi	auxiliary	tufayli

These are three important concepts related to syntax analysis in natural language processing:

- 1) The hierarchical structure of phrases allows them to be combined to form larger structures, creating a tree-like structure that can represent the sentence's syntax.
- 2) Context-free grammars provide a way to describe the phrase structure of a sentence, allowing us to analyze and understand its syntax.
- 3) The rules in a context-free grammar describe how the phrases are structured, providing a framework for analyzing the sentence's syntax.

CFG is a formal way of describing the syntactic structure of a language. It consists of a set of rules, also known as productions, which specify how the elements of the language can be combined to form valid sentences (Bazrafkan, 2022). Each rule has a left-hand side, which specifies a symbol that can be replaced, and a right-hand side, which specifies the replacement.

<i>N</i>	A set of non-terminals (phrase labels)
<i>T</i>	a set of terminals (words)

P	a finite set of rules or productions
S	a distinguished nonterminal symbol called the start symbol

Here is an example of a CFG for Uzbek language:

Kamola kechqurun kichik mushukni ko'rdi

$S = NP + VP$

$NP = ADj + N$

$NP = N$

$VP = V + N$

$VP = V$

$N = Kamola \mid Mushukni$

$V = ko'rdi \mid aytadi$

$ADJ = kichik \mid katta$

$NP = Prn \quad u, men$

$NP = Noun \quad kitob$

$NP = Prn^* + Num^* + ADJ^* + Noun \quad Mening + qizil + kuylakim$

$NP = NP + CNJ + NP \quad men + va + Saida$

$VP = Prep + VG \quad u + tez ketdi$

$VP = NP + VG \quad bu kitoblarni + oldi$

$VP = NP + Prep + VG \quad bu kitoblarni + kechqurun + oldi$

$VG = ADV + V \quad tez + yugirdim$

Note that the part marked with * indicates that it can be omitted.

Table 1. Using of CFG on the Uzbek Sentence

$S \rightarrow NP + VP$	<i>U va Kamola eralab poyezda ketishdi</i>
$NP \rightarrow PRN$	<i>U</i>
$NP \rightarrow CONJ$	<i>va</i>
$NP \rightarrow PP$	<i>Ertalab</i>
$NP \rightarrow N$	<i>Kamola \mid poyezda</i>
$VP \rightarrow V$	<i>ketdi</i>
$VP \rightarrow N + V$	<i>Poyezda + ketishdi</i>
$VP \rightarrow PP + N + V$	<i>ertalab + poyezda + ketdishdi</i>

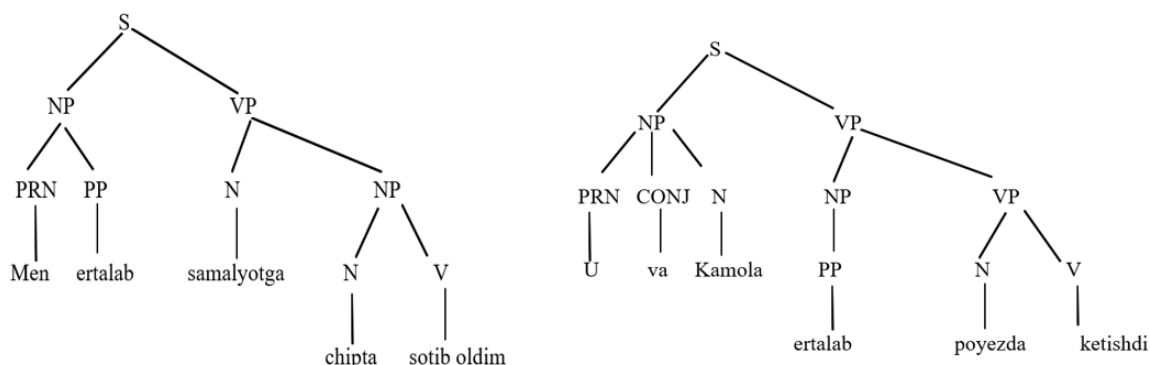


Figure 2. Tree structure generated by CFG-based syntax analysis of Uzbek language sentence

In conclusion, we have successfully created a tree-based syntax analysis model for the Uzbek language sentences using context-free grammar. The implementation of this model involved defining word groups and assigning part-of-speech tags to them. We also presented an example of the context-free grammar rules for Uzbek language. This approach can be used for further development of natural language processing tools for Uzbek language, such as text-to-speech synthesis, machine translation, and sentiment analysis. Additionally, this work can serve as a foundation for future research on the syntax of Uzbek language and its comparative analysis with other languages.

References

1. Almamatova Sh. Uzbek Phraseology as A Science. Frontline Social Sciences and History Journal (2022).
2. Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language, CEUR Workshop Proceedings (2022)
3. Kurbanova M. Analysis and Explication of Syntax of Uzbek Language. Language, Society, Personality and Creativity Of Nizami Ganjevi. 2021. – p 468-472.
4. Bazrafkan M. A New Top-Down Context-Free Parsing for Syntactic Pattern Recognition. International Journal of Mathematical Models and Methods in Applied Sciences (2022) 16.

MORFEM LUG‘ATLARDA TOVUSH O‘ZGARISHLARINING MODELLASHTIRILISHI MASALALARI

Berdiyeva Habiba Abdumannon qizi

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

Kompyuter lingvistikasi mutaxassisligi 2- bosqich magistranti

habibahoshimovaa@gmail.com

Ilmiy rahbar: f.f.d., prof. B.Mengliyev

Annotatsiya. Ushbu maqolada morfemik lug‘atlarda tovush o‘zgarishlarini lingvistik modellashtirish haqida va bunga bugungi kundagi zarurat, qanday manbalar lingvistik modellashtirish uchun ilmiy manba bo‘la oladi va shu kabi bir qator savollarga javob beriladi. Morfemik lug‘atlarda tovush o‘zgarishlarini lingvistik modellashtirish tilshunoslik sohasidagi ushbu yo‘nalishning osonlashuviga yordam beradi.

Kalit so‘zlar: *tovush o‘zgarishi, lingvistik model, modellashtirish, morfem lug‘at, qo‘shimcha.*

Abstract. In this article, about the linguistic modeling of sound changes in morphemic dictionaries and the need for it today, what sources can be scientific sources for linguistic modeling, and a number of similar questions will be answered. Linguistic modeling of sound changes in morphemic dictionaries helps facilitate this direction in linguistics.

Keywords: *sound change, linguistic model, modeling, morpheme dictionary, suffix.*

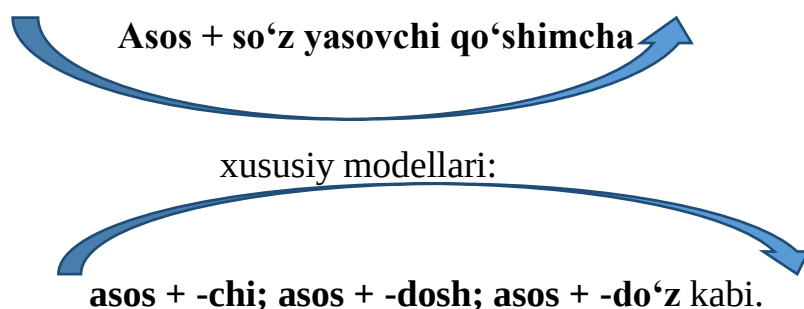
Аннотация. В данной статье о лингвистическом моделировании звуковых изменений в морфемных словарях и необходимости его сегодня, какие источники могут быть научными источниками для лингвистического моделирования и на ряд подобных вопросов будут даны ответы. Лингвистическое моделирование звуковых изменений в морфемных словарях помогает развитию этого направления в области языкознания.

Ключевые слова: *изменение звука, языковая модель, моделирование, словарь морфем, сложение.*

Model (lotincha «modelus» so‘zidan olingan bo‘lib, «nusxa», «andaza», «o‘lchov», «me‘yor» ma‘nolarini anglatadi) tabiiy fanlar yoki umuman fanda shunday moddiy qurilma, grafik, sxema, umuman, bilish vositasi sifatida tushuniladiki, u muayyan original-obyekt haqidagi ma‘lum otlar majmuyi tarzida yuzaga kelgan hosila-obyekt demakdir. Boshqacha tushuntirganda, model tabiiy obyektlarning imitatsiyasidir (o‘xshashi, taqlidiy ko‘rinishi), u o‘zbek tilidagi qolip, andaza so‘zlariga mos keladi [6].

Modellashtirish har bir fan obyektini soddalashtiruvchi metoddir. Lingvistik birliklarni modellashtirish bu belgilar tarkibidagi elementlarning barqaror munosabatlariga asoslanadi. Shuning uchun ham butunlik tarkibidagi elementlar o'rtasida munosabatlarning barqaror va beqaror turlarga ajratilishi lingvistik modellashtirish uchun katta ahamiyatga ega. Modellashtirish barcha fanlar uchun xos bo'lgan umumilmiy metod hisoblanadi.

Lingvistik tadqiqot modellari. Bunda muayyan til hodisalari asosida olib borilgan tadqiqot jarayonini aks ettiradi. Masalan, o'zbek tilida morfologik usul asosida so'z yasallishining umumiy modeli:



Modellashtirish tilshunoslikda strukturalizm yoʻnalishi taʼsirida faol tatbiq qilina boshlandi.

Lingvistik modellashtirish - bu belgilarni bajarayotgan vazifasiga koʻra beqaror va barqaror kabi turlarga ajratish hisoblanib, tilshunoslikka E.Sepir, L.Blumfild, R.Yakobson, N.Chomskiy, Harris, Hokket kabi olimlarning ilmiy izlanishlari orqali kirib kelgan [1;73]. 20-asrning 60-70-yillarida oʻzining taraqqiyotiga erishgan. Lingvistik modellashtirish tadqiq qilnayotgan belgining soddalashishiga, uning oʻz hosilasi mavjud ekanligiga, model belgisining oson qayd qilinishiga yordam beradi. Lingvistik yoʻnalishda aynan tovush oʻzgarishlarining modellashtirilishini nazarda tutadigan boʻlsak, ular inson nutqiy faoliyatiga daxldor modellashtirish hisoblanadi va bu modellar nutq oʻzgarishlarining aniq faktlariga tayangan holda yuzaga keltiriladi.

Inson shaxsida ruh, tafakkur va til chambarchas bogʻliqdir. Demak, ijodiy va mustaqil taffakkurli shaxs tarbiyasida ona tili va uning taʼlimini tashkil etuvchi vositalarning roli beqiyos.

Maʼlumki, oʻrta maktab ona tili taʼlimining konseptual asoslarida uning zaruriy vositasi ikkita ekanligi taʼkidlanadi: darslik va lugʻatlar. Oʻquvchilarning ijodiy tafakkurini va nutqini oʻstirishning zaruriy omillaridan biri lugʻatlar ustida ishlashdir. Lugʻatlar oʻquvchilarning ijodiy tafakkuri, mustaqil fikrlash darajasi, ijodiy fikr mahsulini ogʻzaki va yozma shakllarda ravon bayon etishga olib keladigan behad qimmatli va boy soʻz zahirasi vujudga keltiradi.

Morfemik lugʻatda modellashtirish muammolari bir necha omillar tufayli yuzaga kelishi mumkin. Ushbu omillardan baʼzilari:

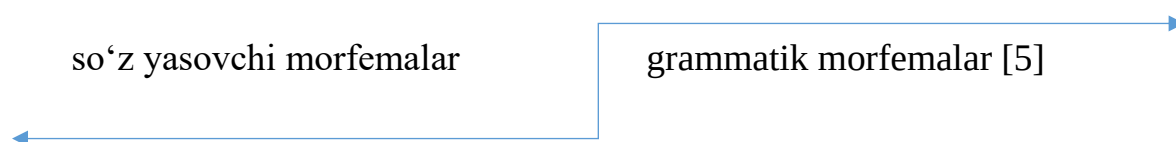
1. Aniq ma'nolilik: Morfemik lug'at tarkibidagi so'zlar ba'zan ko'p ma'noli bo'lishi mumkin, ya'ni ular bir necha talqinga ega bo'ladi. Ushbu noaniqlik kontekstdagi so'zlarning ma'nosini modellashtirish va bashorat qilishni qiyinlashtirishi mumkin.

2. Kontekstga bog'liqlik: Morfemik lug'at tarkibidagi so'zlarning ma'nosi ko'p jihatdan ular qo'llanilgan kontekstga bog'liq bo'lishi mumkin. Model so'zning ma'nosini aniq bashorat qilish uchun kontekstni tushunishi kerak.

3. Neyron tarmoq cheklovlari: Tabiiy tilni qayta ishlash uchun keng qo'llaniladigan neyron tarmoqlar morfemik lug'atlarning murakkabligi va o'zgaruvchanligini boshqarishga qodir bo'lmasligi mumkin.

4. Ko'p ma'nolilik: morfemik lug'atlarda bir nechta ma'noli so'zlar bo'lishi, bu esa chalkashliklarni keltirib chiqarishi va modellashtirish jarayonini yanada murakkablashtirishi mumkin.

Ma'noli qismlarga ajralmaydigan, so'zning nomustaqil tarkibiy qismiga morfema deyiladi. Morfemalar funksional-semantik jihatdan ikki guruhga bo'linadi:



So'zning morfem tarkibini ajratish. Har bir so'zning morfemalarga ajratilishi undagi qismlarning hozirgi o'zbek adabiy tilida o'zaro munosabatga kirisha olish xususiyatiga asoslanadi. So'zning tarkibini ajratishda, avvalo, o'zak ajratilib, uning lug'aviy ma'nosi aniqlanadi. O'zakka qaysi qo'shimcha turi qo'shilmasin, u har doim o'zakdagi ma'no bilan aloqador bo'ladi va yangi so'z yasalganda ham o'zakning ma'nosi saqlanib qoladi. Masalan, *suvoqchilik* so'zining tarkibi *suva+q+chi+lik*. Ushbu kasb-hunar bilan bog'liq soha otining o'zakdagi suvash harakati bilan aloqasi uzilmagan.

So'z tarkibidagi o'zgarishlar. Davrlar o'tishi bilan so'zlarda ma'no taraqqiyoti yuz berib, yasalma o'z atash ma'nosidan uzoqlashishi, kuchli tovush o'zgarishiga uchrashi yoki ayrim morfologik shakllar o'z funksiyasini yo'qotib, «qotib qolishi» mumkin. Agar o'zakdagi ma'no bilan yasalma ma'nosi o'rtasida aloqa uzilsa, morfemalar o'zak tarkibida soddalanishga uchragan bo'lsa, bunday so'zlar ma'noli qismlarga ajralmaydi, unga yaxlit so'z sifatida qaraladi. So'z yasovchi morfemalarda ham, lug'aviy va sintaktik shakl yasovchi morfemalarda ham bu holatni kuzatish mumkin: *qishloq* (aslida qishlaydigan joy, bugungi kunda doimiy yashash joyi, ya'ni qish bilan aloqa uzilgan), *ovloq* (aslida ovlaydigan joy, bugungi kunda pana, xoli joy), *qovurg'a* (qopurg'a – yurak, o'pka kabi a'zolari qoplab turadigan organ), *qopqoq*, *yaproq*, *bo'yin*, *yurak*, *ichak*, *uyg'on*, *uyg'ot*, *uyg'oq*, *yumshamoq*, *o'xshamoq*, *qurshamoq*, *olcha* (aslida olu-cha), *qizilcha* (1. kichik qizil toshmali kasallik. 2. qizartirib qovurilgan go'shtli taom), *ko'kcha* (qovun navi), *Onaxon*, *Olmaxon*, *Qizlarxon*, *Beklar* (mahallasi), *qizg'ish*, *sarg'ish* (aslida qizil-ish,

sariq-ish), *ko'rish* (salomlash), *kelish* (bir qarorga kelmoq), *tortinmoq*, *cho'qinmoq*, *botinmoq*, *kechirmoq*, *surishtir*, *cho'milmoq*, *kelin*, *kechasi*, *kunduzi*, *tunlari*, *chamasi*, *mazmuni*, *qisqasi*, *yaxshisi*, *rosti*, *birdan*, *birga*, *birvarakayiga*, *kunda* va h.

Ko'p hollarda so'zning tarkibini aniqlashda so'zdagi fonetik o'zgarishlar qiyinchilik tug'diradi. Tillarni kuzatishdan ma'lumki, morfologik ko'rsatkichlar yo mustaqil ma'noli leksemalardan, yoki har xil qo'shimchalarning o'zaro birikib yangi qo'shimcha hosil qilish usulidan shakllanadi. Ko'pgina so'zlar qo'shilib ketishi natijasida ikkinchi qism morfemaga aylanib ketishi mumkin: *gapir* (*gap ur*), *tupir* (*tuf ur*), *supur* (*suv ur*), *berkit* (*berk et*), *to'lat* (*to'la et*), *bo'shat* (*bo'sh et*), *yo'qot* (*yo'q et*), *yo'qol* (*yo'q o'l*), *aytoladi* (*ayta oladi*), *ketaver* (*keta ber*), *ketvormoq* (*ketib yubormoq*), *boryapti* (*borib yotibdi*).

Hozirgi o'zbek adabiy tilida so'zga qo'shimcha qo'shilganda o'zakda fonetik o'zgarishlarning quyidagi ko'rinishlari kuzatiladi.

Tovush tushishi: *sariq-ay / sarg'ay, past-ay / pasay, og'iz-aki / og'zaki, o'yin-a / o'yna, o'rin-ash / o'rnash, ayir-im / ayrim, ayir-il / ayril, buyur-uq / buyruq; ikki-ov / ikkov, olti-ala / oltala, yetti-ovlon / yettovlon; men-ning / mening, sen-ni / sening; burun-i / burni, shahar-im / shahrim, o'g'il-ing / o'g'ling.*

Tovush orttirilishi: *isi-q / issiq, qot-iq / qattiq, tib-iy / tibbiy, his-iy / hissiy, shu-cha / shuncha, bu-day / bunday; u-ga / unga, shu-da / shunda, bu-dan / bundan, haq-i / haqqi, parvo-i / parvoyi, obro'-ing / obro'ying.*

Tovush almashishi:

a ~ o: tara-q / taroq, so'ra-q, / so'roq, qiyna-v / qiynov, so'ra-v / so'rov;

o ~ a: son-a / sana, ong-la / angla, uosh-a / uasha, ot-a / ata;

i ~ u: tinti-v / tintuv, o'qi-v / o'quv, sovi-q / sovuq, quri-q / quruq;

q ~ g': qishloq-i / qishlog'i, yutuq-im / yutug'im, buyruq-i / buyrug'i;

k ~ g: yurak-i / yuragi, eshik-im / eshigim, tilak-ing / tilaging;

So'z tarkibi lug'ati o'rni bilan imlo lug'ati vazifasini ham o'tay oladi. Chunki so'zlarning tarkibiy tuzilishi haqida qat'iy bilimga ega bo'lish, o'zakka turli vazifadagi morfemalar qo'shilishi natijasida yuz beradigan fonetik o'zgarishlarni anglab olish, shubhasiz, yozuvdagi har xilliklarni bir shaklga keltirishga yordam beradi.

Peterburg lingvistik maktabining yirik namoyandasi I.A.Boduen de Kurtene A.I.Vachinyayevning «Imlo lug'ati»ni tahlil qila turib, so'zlarni morfemalarga ajratishning imlo uchun ahamiyati juda katta ekanligini alohida ta'kidlagan edi.

O'zakni ajrata bilish, so'z tarkibidagi tovush o'zgarishlarini fahmlash so'zni to'g'ri yozishga imkon beradi. Shu sababli lug'atda ayrim so'zlarning etimologiyasi, ya'ni kelib chiqishiga ishora qilib ketishgan: anglashilmovchilik (*ong-la-sh-il-ma-vchilik*), sanoqsiz (*son-a-q-siz*), atoqli (*ot-a-q-li*) kabi [5].

Umuman olganda, morfemik lug'atlarda modellashdirish muammolari tabiiy tilning murakkabligi va o'zgaruvchanligi tufayli mavjud. Ushbu muammolarga



qaramay, tabiiy tilni qayta ishlash sohasidagi yutuqlar ushbu sohalardagi modellarning ish faoliyatini yaxshilashda davom etmoqda.

Foydalanilgan adabiyotlar:

1. Baxter et al. 2006 — G. Baxter, R. Blythe, W. Croft, A. McKane. Utterance selection model of language change. *Physical Review E* 73, pp. 046118–1 — 046118–20.
2. Brinton 1891 — D. Brinton. The American race. A linguistic classification and ethnographic description of the native tribes of North and South America. N. D. C. Hodges, New York. Brown et al. 2013.
3. C. Brown, E. Holman, S. Wichmann. Sound correspondences in the world's languages. *Language* 89.1, pp. 4–29. Brugmann 1967.
4. Hamroyeva Sh. O'zbek tili morfologik analizatorining lingvistik ta'minoti. Monografiya. “Globe edit”. – Toshkent-2020.
5. Mengliyev B., Bahriddinova B. O'zbek tilining so'z tarkibi lug'ati. – Toshkent. “Yangi asr avlodi”. 2007.
6. Rahimov A. Kompyuter lingvistikasi asoslari. – Toshkent. “Akademnashr” nashriyoti, 2011.
7. www.fayllar.org
8. www.ziyouz.com

O‘ZBEK TILIDAGI SO‘Z BIRIKMALARINI MODELASHTIRISH MUAMMOLARI MASALASI

Xudayarova Sabura

ToshDO‘TAU Kompyuter lingvistikasi mutaxassisligi
1-kurs magistranti

Annotatsiya. Matnlarni avtomatik tahlil va tahrirlash, shu bilan birga ularni avtomatik tarjima qilish jarayonida sintaktik birliklar, xususan, so‘z birikmasi alohida o‘rin tutadi. Bu esa so‘zlarning o‘zaro bog‘lanish usullari, so‘z birikmalarining hosil bo‘lishida ishtirok etuvchi turli elementlar hamda tobe va hokim qismlarning ma‘lum bir so‘z turkumlari bilan ifodalanishiga ko‘ra keltirilgan turli tasniflar asosida so‘z birikmasining muayyan bir yagona tasnifini belgilash, ushbu tasnifga tayangan holda so‘z birikmalarining aniq va qat’iy lisoniy sintaktik qoliplar [LSQ]ni shakllantirish mazkur maqolada tahlil etildi.

Kalit so‘zlar: *So‘z birikmasi, modelashtirish, lisoniy sintaktik qolip, avtomatik tahlil.*

Abstract. In the process of automatic analysis and editing of texts, as well as their automatic translation, syntactic units, in particular, word combinations, have a special place. This is a word on the basis of different classifications given according to the ways of connecting words, various elements involved in the formation of word combinations, and the representation of subordinate and dominant parts by certain word groups. Determination of a certain single classification of the combination, the formation of clear and strict linguistic syntactic patterns of word combinations based on this classification was analyzed in this article.

Key words: *Syntax, modeling, linguistic syntactic pattern, automatic analysis.*

Абстракт. В процессе автоматического анализа и редактирования текстов, а также их автоматического перевода особое место занимают синтаксические единицы, в частности словосочетания. Это слово на основе различных классификаций, данных по способам соединения слов, различным элементам, участвующим в образовании словосочетаний, представлению придаточных и господствующих частей определенными словосочетаниями. В данной статье проанализировано формирование четких и строгих лингвистических синтаксических моделей словосочетаний на основе этой классификации.

Ключевые слова: *Синтаксис, моделирование, языковая синтаксическая модель, автоматический анализ.*

Sintaktik tahlilni amalga oshiruvchi dasturiy ta‘minot uchun so‘z birikmalarining LSQlari shakllantirilar ekan, bunda barcha bir-biriga mos keluvchi,

shu bilan birga bir-biridan farq qiluvchi soʻz birikmasiga doir turli tasniflarni oʻrganib, istisnoli holatlar bartaraf etilgandagina, bizningcha, masalaga toʻgʻri yondashilgan boʻladi. Soʻz birikmasining LSQlarini modellashtirish esa sintaktik analizatorning bir qismi boʻlgan soʻz birikmasini tahlil qiluvchi dasturiy taminot uchun asos boʻlib xizmat qiladi. Til birliklarini modellashtirish quyidagi umumiylik, xususiylik hamda oʻzgarmaslik kabi jihatlari bilan xarakterlanadi. Ushbu jihatlarni soʻz birikmasi doirasida koʻrib oʻtamiz: aqli bola: 1. umumiylik tobe soʻz + hokim soʻz = T+H 2. xususiylik sifat +ot=S+OT 3. oʻzgarmaslik ushbu birikma oʻrnida boshqa shu kabi sifat+ot koʻrinishidagi koʻplab birikmalarni hosil qilish mumkun, biroq LSQ ham va unga asoslangan model ham oʻzgarmaydi. Soʻz birikmalari turli soʻz turkumlari hamda biriktiruvchi elementlarning turli grammatik koʻrsatkichlar bilan ifodalanishi LSQlar hamda modellarning oʻzgarishiga sabab boʻladi. Soʻz birikmasining LSQlari hamda uning modellari haqida soʻz borar ekan, ushbu fikrlarga eʼtiborimizni qaratmoqchimiz. “Soʻz birikmalarining modeli [qolipi] tarixan tayyor holda mavjud va odamlar shu qolipga suyangan holda, soʻzlarni bir-biriga bogʻlaydilar. Shu jihatdangina soʻz birikmalari doimiy xarakterga ega boʻladi.” Yuqorida soʻz birikmasi qolipi va modeli haqida fikr yuritilar ekan, qolip va model tushunchasi bir maʼnoda qoʻllanilganligini koʻramiz. Bundan tashqari A.Rahimov “Kompyuter lingvistikasi asoslari” oʻquv qoʻllanmasida ham model tushunchasi qolip va andoza soʻzlari bilan mos kelishini aytib oʻtadi. Ushbu tushunchalar bir maʼnoni anglatgandek boʻladi nazarimizda, biroq ular tilshunoslik nuqtayi nazaridan bir-biridan farq qiladi. Masalan, maktabga bormoq kabi soʻz birikmasining qolipi quydagicha: ot+ga+feʼl ushbu qolipni ifodalash uchun turli modellardan foydalanish mumkin. Soʻz birikmasini modellashtirishda, asosan anʼanaviy tilshunoslikda keltirilgan soʻz birikmalarining qoliplari oʻrganildi. Bu oʻrinda quyidagi tasnifni koʻrib oʻtamiz:

Soz birikmalari quyidagi xususiyatlari asosida tasnif qilinadi:

1. Soʻz birikmalari qismlarining grammatik tabiatiga koʻra.
2. Soʻz birikmalarining tuzilishiga koʻra.

Soʻz birikmalari qismlarining grammatik tabiatiga koʻra tasnif qilinganda, hokim aʼzoning qaysi soʻz turkumidan ekanligi, tobe aʼzoning sintaktik vazifasi asosga olinadi.

Soʻz birikmalari hokim qismning qaysi soz turkumidan ekanligiga koʻra quyidagi turlarga ajratiladi: 1) otli birikma; 2) feʼlli birikma; 3) sifatli birikma; 4) ravishli birikma; 5) modal soʻzli birikma.

1. Otli birikma. Bunday birikmaning hokim soʻzi vazifasida ot, otlashgan soʻzlar keladi va u quyidagi koʻrinishlarga ega boʻladi:

- hokim qism vazifasida ot keladi: *gʻayratli odam, samimiy inson, ezgu niyatli kishi.*
- hokim qism sifatida olmosh keladi: *bolaning oʻzi, Karimaning oʻzi.*
- hokim soʻz sifatida otlashgan sifat keladi: *kishilardan eng mard, qizlarning sarasi.*
- hokim qism vazifasida otlashgan son keladi: *talabalarning beshtasi, oʻquvchilardan biri.*

- hokim qism vazifasida otlashgan olmosh keladi: *tilaklarning barchasi, kutib oluvchilarning hammasi*.
- hokim soʻz vazifasida otlashgan sifatdosh keladi: *onamlarning aytganlari, bolalarning oʻqiyotgani va h.k.*
- hokim qism vazifasida otlashgan ravish keladi: *koʻp soʻzning oʻzi, ishlarning koʻpi*.
- hokim soʻz sifatida otlashgan undov soʻz keladi: *ularning dod-voyi, bolalarning urasi*.

- hokim soʻz vazifasida otlashgan taqlid soʻz keladi: *otlarning dupur-dupuri, toʻplarning gumbur-gumburi*. Koʻrinadiki, otli soʻz birikmasi qismlari bitishuv, muvofiqlashuv va boshqaruv yoʻli bilan grammatik aloqaga kirishadi..

2. Feʼlli birikma. Bunday birikmaning hokim qismi feʼl bilan ifodalanadi: *tez ishlamoq, oldinda yurmoq, diqqat bilan tinglamoq, kulib gapirmoq*.

Feʼlli birikmaning qismlari oʻzaro bitishuv va boshqaruv yoʻli bilan boglanadi.

3. Sifatli birikma. Mazkur birikmada hokim soʻz otlashmagan sifat bilan ifodalanadi, birikmaning qismlari boshqaruv yoʻli bilan sintaktik munosabatga kirishadi: *akasidan kuchli, singlisidan ziyrakroq*.

4. Ravishli birikma. Bunday birikmaning hokim soʻzi otlashmagan ravish bilan ifodalanadi. Ravishli soʻz birikmalarida tobe va hokim qismlar oʻzaro boshqaruv yoʻli bilan bogʻlanadi: *raketadan tez, toshbaqadan sekin*.

5. Modal soʻzli birikma. Bunda modal soʻz birikmaning hokim soʻzi sifatida keladi. Birikma boshqaruv yoʻli bilan tuziladi: *bizga kerak, oʻquvchilarda bor*.

Soʻz birikmalari, tobe qismning sintaktik vazifasiga koʻra, quyidagi turlarga boʻlinadi: 1) aniqllovchili birikma; 2) toʻldiruvchili birikma; 3) holli birikma.

1. Aniqllovchili birikma. Bunday birikmalarning tobe qismi aniqllovchining barcha turlari vazifasini bajaradi. Shunga koʻra ularni: 1) sifatlovchili birikma, 2) qaratuvchili birikma, 3) izohlovchili birikmalarga ajratish mumkin. Masalan: *yozma ish, ilmiy maqola, beshta kitob, allaqanday ovozlari* – sifatlovchili birikma; *iqtisodiyotning taraqqiyoti, maktab hovlisi* - qaratuvchili birikma; *oʻqituvchi Salimova, professor Qodirov* - izohlovchili birikma.

Aniqllovchili birikma bitishuv va boshqaruv yoʻli bilan tuziladi.

2. Toʻldiruvchili birikma. Tobe qismi toʻldiruvchi vazifasini bajaradigan soʻz birikmasi toʻldiruvchili birikma deb yuritiladi. Bunda qismlar oʻzaro boshqaruv yoʻli bilan birikadi. Masalan: *yoshlarga nasihat qilmoq, odamlarni qadrlamoq, talabalar uchun darslik, ilm-fan haqida soʻzlamok*.

Soʻz birikmalarining barcha LSQlarini toʻplash bilan birga oʻzbek tilining xususiyatlarini aks ettiruvchi ramzlar orqali ushbu qoliplarga modellar taklif etildi. Lisoniy sintaktik qoliplarni shakillantirish hamda modellashtirish jarayonida baʼzi muammoli masalalar koʻzga tashlanadi. Modellashtirishdagi ushbu muammoli jihatlarni bartaraf etish sintaktik tahlilni amalga oshiruvchi dasturni har tamonlama aniq ishlashini taʼminlaydi.

1-muammoli vaziyat:

Tobe soʻz va hokim soʻz shaxs va sonda moslashib, qaratqich kelishigi va egalik qoʻshimchalari yordamida birikish orqali moslashuvli birikmalarni hosil qiladi. Ushbu birikmalarning qolipi quyidagicha ifodalanadi: tobe soʻz+ning +hokim soʻz +egalik qoʻshimchalari bu qolipning umumiy koʻrinishi boʻlib, tobe va hokim qismlar turli soʻz turkumlari orqali ifodalanadi. Masalan, ot +ning +ot +egalik qoʻshimchasi *oʻquvchining savoli, maktabning bogʻi* qaratqich kelishigi koʻp hollarda belgisiz qoʻllanilganda ham birikma mazmuniga taʼsir koʻrsatmaydi. *Baxt manzili, koʻcha eshigi* kabi birikmalarda qaratqich kelishigi ifodalanmagan boʻlsa-da, qaratqich va qaralmish munosabati mavjud. Tilimizda shunday birikmalar borki, yuqoridagi *baxt manzili, koʻcha eshigi* kabi birikmalar qolipiga mos keladi, biroq mazmun jihatdan farqlanadi. *Navoiy kutubxonasi, bobur bogʻi* kabi birikmalar mazmunan atalganlik, umundan ajralganlikni ifodalashi bilan birga biron bir obyektни nomlab kelmoqda. Shunday ekan, ushbu birikmalar soʻz birikmasi doirasida emas, balki NERlar sifatida oʻrganilishi maqsadga muvofiq deb oʻylaymiz.

2-muammoli vaziyat:

Oʻzbek tili grammatikasi darsligida soʻz birikmalarining turli konstruksiyalari alohida tahlil qilingan boʻlib, bu oʻrinda *zavq bilan oʻynamoq, qunt bilan oʻqimoq* kabi birikmalarning qoliplariga eʼtiborimizni qaratmoqchimiz. Ushbu birikmalar tahlil qilinar ekan bitishuv yoʻli bilan birikkanligi hamda *zavq bilan, qunt bilan* soʻzlari *bilan* koʻmakchisi bilan birgalikda harakat holatini bildirib, ravishni ifodalalashi taʼkidlanadi. Bilamizki, *qalam bilan chizmoq, pichoq bilan kesmoq* kabi birikmalar koʻmakchi orqali bogʻlangan boʻlib, boshqaruvli birikmani hosil qiladi. Bu oʻrinda, *zavq bilan oʻynamoq* hamda *qalam bilan yozmoq* kabi soʻz birikmalarining lisoniy sintaktik qoliplariga eʼtibor qaratilsa, ushbu birikmalar qolipi shaklan bir-biriga oʻxshash ekanligi koʻzga tashlanadi. *Zavq bilan oʻynamoq* ot +koʻmakchi +feʼl, ushbu qolipning modeli quyidagicha: [OT_k +F], *pichoq bilan kesmoq*: ot +koʻmakchi +feʼl ushbu qolip modeli ham [OT_k +F] koʻrinishida shakllantiriladi. Matnlar sintaktik jihatdan avtomatik tahlil qilinar ekan, birikma turlarini aniqlash jarayonida keltirilgan qolipga mos ravishda shakllanuvchi shu kabi barcha birikmalar, boshqaruvli birikmalar sifatida ajratiladi. Shunday ekan, mavhum otlar oldidan bilan koʻmakchisi kelganda ravishni ifodalovchi soʻzlar guruhini shakllantirish va ularning alohida oʻziga xos modellarini ishlab chiqish orqali koʻzlangan maqsadga erishish mumkin boʻladi.

3-muammoli vaziyat:

1. oʻqib chiqmoq [koʻmakchi feʼlli soʻz qoʻshilmasi]; 2. kulib gapirmoq [feʼl+ib +feʼl] soʻz birikmasi bir qaraganda ushbu birliklarni bir-biridan farqlab olish qiyin. Koʻmakchi feʼlli soʻz birikishidan hosil boʻlsa, KFSQ esa yetakchi hamda koʻmakchi feʼldan tashkil topadi. Koʻmakchi feʼl oʻzining maʼno mustaqilligini butkul yoʻqotadi. Bir-biridan farq qiluvchi birliklarni shakllantiruvchi qoliplarning oʻxshash jihatlarini koʻrib oʻtmoqchimiz. 1. yetakchi feʼl +ib +koʻmakchi feʼl, feʼl +ib +feʼl koʻrinadiki qoliplar bir-biridan shaklan farq qilmaydi. Berilgan qoliplar asosida modellar ishlab chiqilib, sintaktik tahlilni amalga oshiruvchi dastur bazasiga

kiritilar ekan, sintaktik tahlil jarayonida fe'l +ib +fe'l ko'rinishidagi so'z birikmalarini ko'makchi fe'lli so'z qo'shilmasidan farqlashi qiyinlashadi va so'z birikmalarini aniqlash qobiliyati pasayadi. O'zbek tilida aytib bermoq, tutib olmoq, ko'rib chiqmoq kabi bir qancha fe'llar mavjud bo'lib, ularning barchasi yig'ilib ko'makchi fe'lli so'z qo'shilmasi guruhini shakllantirish bilan bir qatorda ularning o'ziga xos alohida modellarini ishlab chiqilishi masalaning hal etilishidagi yechimlardan biri bo'ladi, deb o'ylaymiz.

4-muammoli vaziyat:

Ochiqdan ochiq, tomdan tomga, o'zidan o'zi kabi qo'shma so'zlar qismlari ajratib yoziladi. Ushbu qo'shma so'zlar ham bir qancha so'z birikmalarini hosil qiladi. Masalan: tomdam tomga sakramoq, ochiqdan ochiq gapirmoq shu kabi bir qancha so'z birikmalarini modellashtirish biroz munozarali fikrlarni yuzaga keltiradi. Mana shunday murakkablikga ega bo'lgan shu bilan birga istisnoli holatlarni yuzaga keltiruvchi birikmalarni aniqlash muhim ahamiyat kasb etadi. Bundan tashqari so'z birikmalari sodda hamda murakkab birikmalarga ajratilar ekan, murakkab birikmalarni modellashtirishda aniq va qat'iy to'xtamga kelinish lozim, shundagina o'zbek tilidagi barcha so'z birikmalarining mukammal modeli yaratiladi.

5-muammoli vaziyat:

So'zlar o'zaro nafaqat grammatik jihatdan, balki leksik-semantik jihatdan mos ravishda bog'langandagina so'z birikmasini hosil qiladi. Ko'rinadiki, so'z birikmalarining bog'lanishida leksik jihat ham muhim rol o'ynaydi. stolning qo'li, otning barmog'i kabi birikmalar shaklan grammatik jihatdan mos ravishda bog'langan bo'lib, leksik jihatdan bog'lanish imkoniyatiga ega emas. Bunday birikmalarni sintaktik tahlilni amalga oshiruvchi dastur orqali aniqlash anchayin murakkab masala. Ushbu masalaning yechimi leksik tahlilni amalga oshiruvchi dasturning yaratilishi bilan bog'liq. So'z birikmalarini modellashtirishda barcha xususiyatlar e'tiborga olinishi, mukammal dasturning yaratilishi uchun asos bo'lib xizmat qiladi. So'z birikmalari va uning lisoniy sintaktik qoliplarini o'rganib, tahlil qilar ekanmiz bajarilishi lozim bo'lgan vazifalar talaygina. Ushbu vazifalarni bajarilishi esa ba'zi muammolarning yechimi bo'lib xizmat qiladi. So'z birikmalarining yagona va optimal modellarini ishlab chiqish hali ham yechimini kutayotgan masalalardan biridir.

Xulosa qilib aytganda, o'zbek tilshunosligida so'z birikmasi tuzilish strukturasiga ko'ra, grammatik munosabatga kirishishiga ko'ra, ba'zi shakllarda so'zga, ba'zi o'rinlarda gapga tenglanishiga ko'ra murakkab tuzilma hisoblanadi. Ular uchun ayni bir qolipni belgilab olish qiyin hisoblanadi.

Foydalanilgan adabiyotlar:

1. Anorbekova A., Mirzayeva Sh. Hozirgi o'zbek adabiy tili. – Toshkent, 2011.
2. G'ulomov A., Asqarova M. Hozirgi o'zbek adabiy tili. Sintaksis. - Toshkent, 1965.



3. Mahmudov N., Nurmonov A. O'zbek tilining nazariy grammatikasi (sintaksis). – Toshkent, 1995.
4. Muhiddinova H., Xudoyberganova D. va b. Hozirgi o'zbek adabiy tili. – Toshkent, 2006.
5. O'zbek tili grammatikasi. II tom. (mualliflar jamoasi). – Toshkent, 1978.
6. Sayfullayeva R., Mengliyev B. va b. Hozirgi o'zbek adabiy tili. – Toshkent, 2009.
7. Sharipov M. Hozirgi o'zbek tilida so'z birikmalari sintaksisi. – Toshkent, 1918.
8. Toshnazarova D., Ibragimova A. So'z birikmalari ta'rifiga yangicha yondashuv. Til va adabiyot ta'limi jurnali, 4-son. – Toshkent, 2020.
9. Elov B., Hamroyeva Sh., Xusainova Z. NLP (tabiiy tilga ishlov berish)ning vazifalari va zamonaviy yondashuvlar. TerDU, FILOLOGIK TADQIQOTLAR: TIL, ADABIYOT, TA'LIM. 2022, 5-6.

KOMPYUTER LINGVISTIKASIDA SUN'IIY INTELLEKT VA MASHINA TARJIMASI

MASHINA TARJIMASIDA MATNNI MOSLASHTIRISH USULLARI

Xamroyeva Shahlo,

f.f.d., dotsent shaxlo.xamroyeva@navoiy-uni.uz

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

Matyakubova Noila,

doktorant nailya89mm@mail.ru

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

Annotatsiya. Matnni moslashtirish turli xil mashina tarjimai tizimlarining muhim jarayonidir. Bu vazifa dastlabki matnning so'zlari, jumalari yoki paragraflari va ularning tarjimai (parallel korpus) o'rtasidagi yozishmalarni aniqlashdan iborat. Parallel korpusni moslashtirishning ikkita asosiy yondashuvi mavjud: statistik usullar va leksikaga asoslangan usullar.

Kalit so'zlar: *Matnni moslashtirish, mashina tarjimai, statistik usullar, leksik usullar, tabiiy tilni qayta ishlash.*

Abstract. The text alignment is an important process of different Machine Translation systems. This task consists in identifying correspondences between words, sentences or paragraphs of a source text and their translation (parallel corpus). There are two main approaches to perform parallel corpus alignment: the statistical-based methods and lexical-based methods.

Keywords: *Text alignment, machine translation, statistical methods, lexical methods, natural language processing.*

Аннотация. Выравнивание текста — важный процесс различных систем машинного перевода. Эта задача заключается в выявлении соответствий между словами, предложениями или абзацами исходного текста и их переводом (параллельный корпус). Существует два основных подхода к параллельному выравниванию корпусов: статистические методы и методы на основе лексики.

Ключевые слова: *Выравнивание текста, машинный перевод, статистические методы, лексические методы, обработка естественного языка.*

Kirish. Tabiiy tilni qayta ishlash (NLP) - bu odamlar muloqot qilish uchun foydalanadigan tilni yaratish va tushunishga qaratilgan kompyuter fanining sohasi hisoblanib u juda ko'p muhim vazifalar va ilovalarga ega; ulardan biri matnni avtomatik ravishda bir tildan ikkinchi tilga tarjima qilishni maqsad qo'llaniladigan mashina tarjimai(MT)dir. MTni amalga oshirish uchun lug'atlar, statistika yoki misollarga asoslangan bir nechta usullar mavjud. Ushbu usullarning o'ziga xos

afzalliklari va kamchiliklari bo'lsa ham, ular matnni moslashtirish jarayoni kabi ba'zi usullarni keng foydalanadi. Matnni moslashtirish (TA) jarayoni paragraflar, jumlar, manba matnlarning so'zlari va ularning tarjimalari o'rtasida muvofiqlikni o'rnatish uchun parallel korpusni tashkil qilishdan iborat. Parallel korpus turli tillardagi ikkita matn to'plami sifatida belgilanishi mumkin, bu to'plamlardan biri manba matni, ikkinchisi esa ularning tarjimasi hisoblanadi.

Asosiy qism. Parallel korpusni moslashtirishning ikkita asosiy yondashuvi mavjud: statistik usullar va leksikaga asoslangan usullar. Leksikaga asoslangan yondashuvlar mavjud leksik bilimlarga tayanadi, masalan, antonim va sinonimlar, so'zning tarjimalari va boshqalar; Statistik yondashuvlar esa leksik bo'lmagan ma'lumotlarga tayanadi, masalan, jumla uzunligi, gapning o'rnini, takrorlanish chastotasi, ikki tilda gap uzunligi nisbati kabi holatlarga.

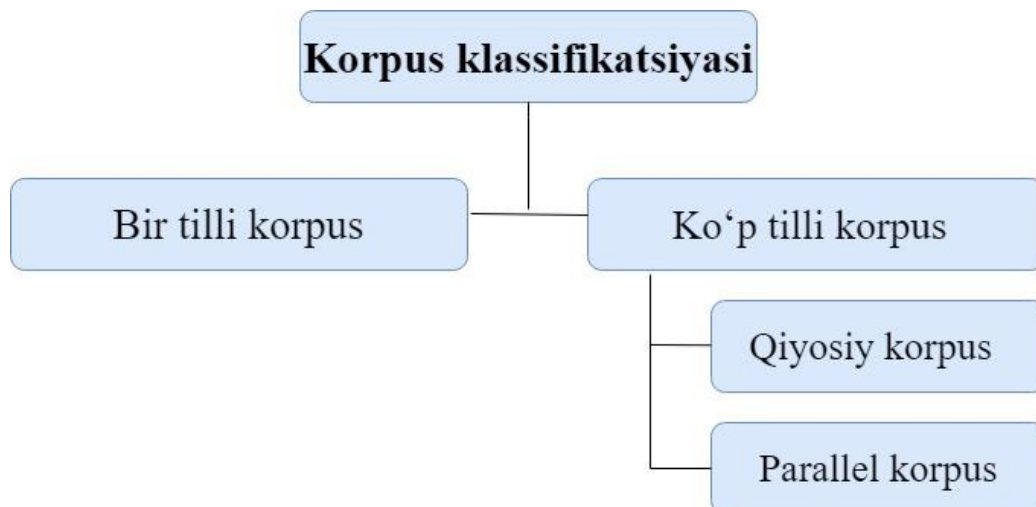
TA jarayonini amalga oshirish uchun biz yozma korpusdan foydalanishimiz kerak. Buni tadqiqot uchun, ayniqsa tarjima dasturlarini ishlab chiqish va tabiiy tilni qayta ishlash uchun foydalaniladigan matnlar to'plami yoki katta hajmli matn sifatida aniqlash mumkin. Korpuslar nomlangan yoki nomlanmagan bo'lishi mumkin. Nomlangan korpuslar turli atributlarni yoki lingvistik ma'lumotlarni aniqlash uchun izohlanadi, masalan, korpus tarkibidagi hujjatlarning mavzulari yoki so'zlarning nutq qismi va boshqalarni. Masalan, "atirgullar" so'zi uchun korpusda belgilangan atributlar ot, ko'plik va boshqalar bo'lishi mumkin. Nomlanishi mumkin bo'lgan lingvistik ma'lumotlar uning lemmasi, ya'ni ma'lum bir lug'at bo'yicha so'zning to'g'ri ma'nosi va boshqalar bo'lishi mumkin. Rus va ingliz kabi tillarda, otning jinsini ifodalovchi belgilar qo'shilishi mumkin. Nomlangan korpus **/sentence/word/lemma/pos/id** ni aniqlaydigan maxsus belgi bilan belgilanadi [Moore, R. C. 2005,. 1-8b]. Ammo, nomlanmagan korpuslar lingvistik ma'lumotlarga ega emas va aniq tuzilishga ham ega emas. Bu ko'pincha elektron pochta yoki tezkor xabarlar, hujjatlar yoki ijtimoiy media xabarlar kabi foydalanuvchilar tomonidan yaratilgan ma'lumotlardir.

Korpusning har xil turlari mavjud bo'lib mashina tarjimasi sohasida bir tilli va ko'p tilli korpusning tasnifi muhim ahamiyatga egahisoblanadi. Ko'p tilli korpuslar bir nechta tillardagi matnlar bo'lib, ularni quyidagi kichik toifalarga bo'lish mumkin:

1. Parallel korpusni turli tillardagi ikkita matn to'plami sifatida aniqlash mumkin, bu to'plamlardan biri manba matnlari, ikkinchisi esa ularning tarjimalari. Ushbu matnlarning har biri bitextlar deb nomlanishi mumkin [Harris,B. 8-10]. Parallel korpus bir yo'nalishli, ikki tomonlama yoki ko'p yo'nalishli bo'lishi mumkin. Masalan, Navoiyning "Xamsa" dostonlari va ularning turli tillardagi nusxalarini parallel korpus deb hisoblash mumkin.

2. Qiyosiy korpus - bu bir xil asosiy mavzuni yorituvchi, lekin uni ko'rib chiqish uslubida farq qiluvchi turli tillardagi matnlar to'plamidir. Bu shuni anglatadiki, qiyosiy korpus manba matni va ularning tarjimasi hisoblanmaydi. Misol tariqasida jurnallar yoki yangiliklarni uzatish tizimlaridan olingan yangi maqolalar

to'plamlari olish mumkin, chunki ular turli tillarda bir xil voqeaga ishora qiladi va shuning uchun qiyosiy korpus deb hisoblanishi mumkin [Simões, A. 2004. 5-8b].



1-jadval Korpus klassifikatsiyasi

Korpusni moslashtirish - bu manba matnlarning paragraflari, jumalari yoki so'zlari va ularning tarjimalari o'rtasida muvofiqlikni o'rnatish uchun parallel korpusni tashkil qilishdan iborat. Shunga qaramay, parallel korpuslarni avtomatik ravishda moslashtirish ba'zi til juftliklari uchun katta ahamiyatga ega vazifa hisoblanadi.

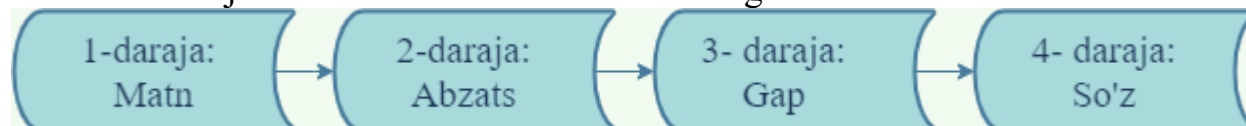
Kanadalik tilshunos Elliot Maklovich MTda qo'llaniladigan moslashtirish usullarini 4 darajaga bo'lib, ularni quyidagicha tasniflagan [Macklovitch, E., & Hannan, M.L.1998. 41-57b.]:

1-darajali moslashtirish: matn yetarlicha uzun bo'lmaganda butun matnni moslashtiradi.

2-darajali moslashtirish: bu daraja abzatlarni moslashtirishni takrorlaydi.

3-darajali moslashtirish: jumalarni moslashtirishni tavsiflaydi.

4-darajali moslashtirish: bitextlar orasidagi so'zlarni moslashtiradi.



2-rasm. MTda qo'llanadigan moslashtirish darajalari

Korpusni moslashtirish yondashuvlari

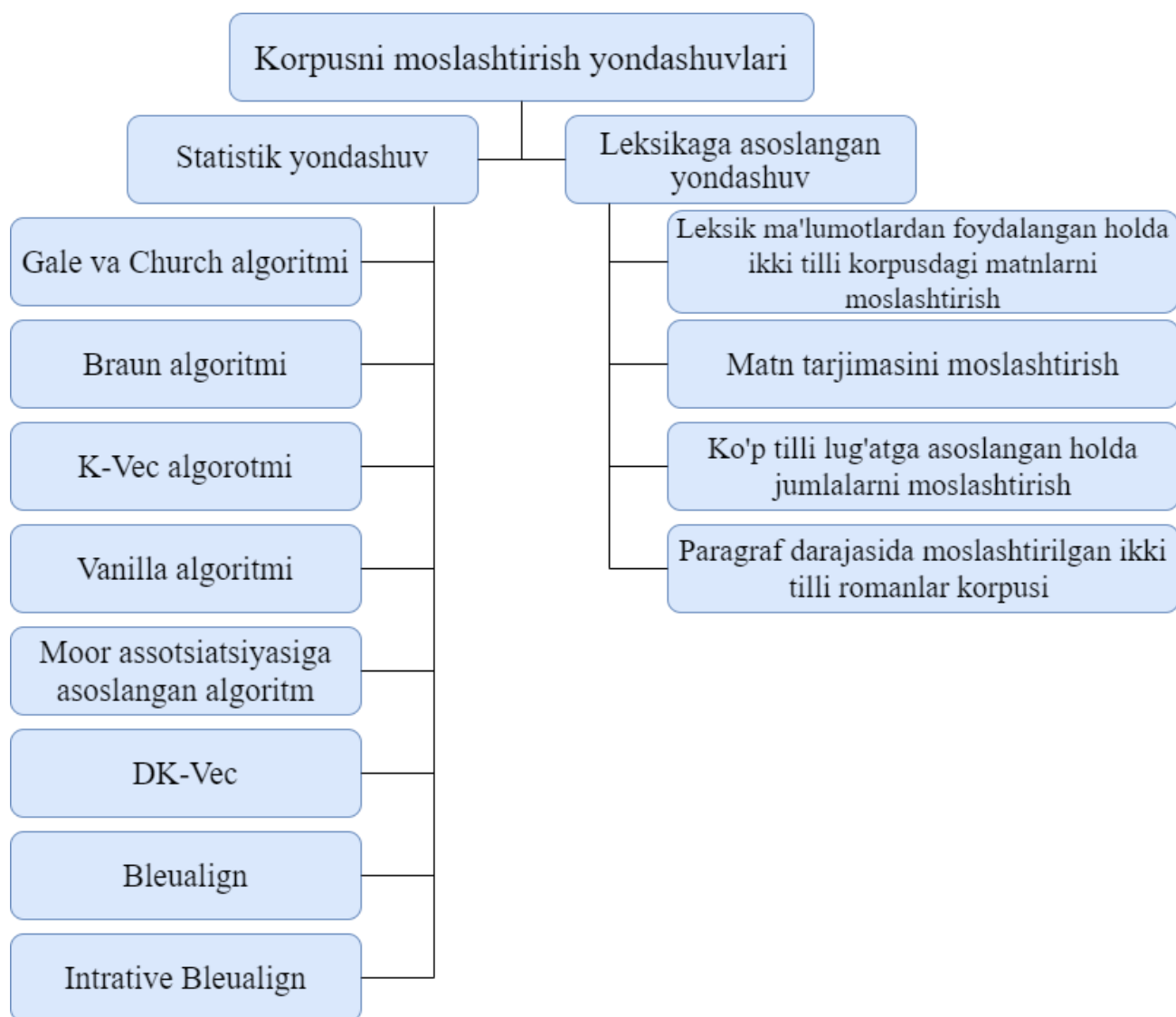
Parallel korpusni moslashtirishning ikkita asosiy yondashuvi mavjud: birinchi yondashuv statistik ma'lumotlarga asoslanadi, ikkinchisi esa qo'shimcha lingvistik bilimlarni qo'llaydi. Ushbu farqning asosi qayta ishlash usullaridan mustaqil ravishda qayta ishlanadigan ma'lumotlar turi bilan bog'liq [Gelbukh, A., Sidorov, G., & Vera-Félix, J. Á.2006. 16-23b).]. Ushbu yondashuvlar asosida bir



nechta texnikalar ishlab chiqilgan bo'lib, ularning har biri o'zining afzalliklari va kamchiliklariga ega.

Leksikaga asoslangan yondashuvlar mavjud leksik resurslarga, masalan, katta hajmdagi ikki tilli lug'atlar, tillar haqida ma'lumot olish uchun maxsus atamalar ro'yxatiga, antonim va sinonimlar, so'zning tarjimalari va boshqalarga tayanadi. Ushbu usullar statistik ma'lumotlarga asoslangan usullardan ko'ra sekinroq va tilga bog'liq bo'ladi. Ushbu usullarning asosiy kamchiligi shundaki, ularning ishlashi ko'p jihatdan moslashtirish jarayonida qo'llaniladigan leksik ma'lumotlarga bog'liq. Biroq, bu usullarning ko'pchiligi statistik usullardan ko'ra yaxshiroq natijalarga erishish mumkinligi uchun ishlab chiqilmoqda.

Statistik yondashuvlar leksik bo'lmagan ma'lumotlarga tayanadi, masalan, jumla uzunligi, jumlaning o'rni, birgalikda sodir bo'lish chastotasi, ikki tildagi jumla uzunligi nisbati. Bu usullar moslashtirish jarayonini tezlashtiradi, biroq, ushbu usullarning asosiy kamchiligi shundaki, ularning ishlashi maqsadli matn va bitextlarning manba matni o'rtasidagi tarkibiy o'xshashlikka bog'liq [Gelbukh, A., Sidorov, G., & Vera-Félix, J. Á.2006. 16-23b]. 3- rasmda bugungi kunda korpus moslashuvida keng qo'llanilayotgan usullar va algoritmlar keltirilgan.



3- rasm. Korpusni moslashtirish yondashuvlari

Gale va Church algoritmi - ushbu algoritm "bir tildagi uzunroq jumlar boshqa tildagi uzunroq jumalarga tarjima qilinadi va qisqaroq jumlar qisqaroq jumalarga tarjima qilinadi" g'oyasi asosida ishlaydi va moslashtirish tamoyila aynan shu g'oyaga asoslangan.[Gale, W. A., & Church, K. W. 1993. 75-102b] Ushbu algoritm faoliyati uchun paragraflarda allaqachon moslashtirilgan parallel korpus talab qilinadi. Ushbu algoritm matnlarni moslashtirish uchun jumlarning uzunligini belgilarda hisobga oladi. Bular har bir jumlar juftligi (manba matnlardan biri va maqsadli matnlardan biri) uchun masofa o'lchovi deb ataladigan qiymatni hisoblash uchun ishlatiladi. Masofa o'lchovi qanchalik past bo'lsa, jumlarning o'zaro mos kelishi ehtimoli shunchalik yuqori bo'ladi.

Braun algoritmi - "Braun klastering" deb ham nomlanadi, chunki u dastlab 1990-yillarda Braun universitetida ishlab chiqilgan. Gale va Church algoritmiga o'xshab, bu algoritm matnlarni moslashtirish uchun jumlar uzunligini hisobga oladi, ammo farqi shundaki, o'lchovda jumla uzunligi so'zlarda hisobga olinadi.

Unda Kanada Hansard korpusining TEX formatiga kiritilgan teglardan moslashish jarayonida foydalaniladi. Algoritm asosiy va kichik bog'lanish nuqtalarini ko'rib chiqadi va moslashtirishni ayni ushbu ikki bosqichda amalga oshiradi.[Brown, P. F., Lai, J. C., & Mercer, R. L. 1991. 169-176b]

K-Vec algoritmi- bu algoritm Paskal Fung va Kennet Uord Cherxi [7] tomonidan ishlab chiqilgan sentence aligner vositasi hisoblanadi. K-vec juda muhim xususiyatga ega; u jumla chegaralariga bog'liq emas. Ushbu xususiyatdan foydalanib, algoritm o'xshash bo'lmagan tillarni moslashtirishni maqsad qiladi; masalan, ingliz va yapon, yevropa tillari kabi. Ushbu algoritm "agar ikkita so'z bir-birining tarjimasini bo'lsa, ular ikkita bo'lmagan so'zlarga qaraganda bir xil segmentlarda paydo bo'lish ehtimoli ko'proq" tamoyili asosida yaratilgan[Fung, P., & Church, K. W. 1994.1096-1102b]. Bittekstdagi so'z boshqa so'zning tarjimasini ekanligini aniqlash uchun ushbu algoritm ularning tegishli matnda taqsimlanishining o'xshashligiga e'tibor qaratadi.

Vanilla Aligner – 1997-yilda Pernilla Danielson va Daniel Ridings tomonidan taqdim etilgan va Gale va Church algoritmining takomillashtirilganidir. Xuddi o'zidan oldingisi kabi, bu ham sentence aligner vositasi hisoblanib, jumlar chegaralariga bog'liq. Ushbu alignerning asosiy afzalligi SGML formatidagi bitextlar bilan ishlashga mosligidir [Danielsson, P., & Ridings, D.1997. 75-80b]. Bitextlarni SGML formatida ishlatishning afzalliklaridan biri shundaki, standart shakl yoki tuzilma o'rnatilishi mumkin va u jumla chegaralarini osonroq aniqlashga yordam beradi.

Moor assotsiatsiyasiga asoslangan algoritm – ushbu algoritm asosiy ikkita tamoyilga asoslanib yaratilgan: birinchidan, ular sentence alignerlar phrase (iboraga asoslangan) alignerlarni yaratish uchun yaxshi debocha bo'lishi mumkinligiga qattiq ishonishgan; ikkinchidan, o'sha vaqtga qadar taqdim etilgan algoritmlar, masalan, Braun yuqori hisoblashda birmuncha murakkabliklar va kamchiliklarga ega, ammo past hisoblashda proporsional yaxshi aniqlik bilan ishlaydi. Moor so'zlarni moslashtirishning uchta turli strategiyasini taqdim etadi [Moore, R. C. 2005. 1-8b).]:

- 1) bir so'zni bitta ma'nosi bilan moslashtirish
- 2) bir so'zni bir nechta ma'nolari bilan moslashtirish
- 3) Tokenlarni mosligini tanlash.

Har bir strategiyada bir nechta muammolarni yechish uchun ikki yoki undan ortiq usullar mavjud va ularning barchasi leksikon tarjimasini yaratishda qo'llanilgan Log-Likelihood-Ratio (LLR) assotsiatsiyasi o'lchoviga asoslangan.

DK-vec algoritmi – (Dinamik K-vec algoritmi) hozircha DK-vec o'zining ajdodi K-vec algoritmiga asoslangan bo'lib, u ikkita so'z bir-birining tarjimasini bo'lsa, bir segmentda paydo bo'lish ehtimoli ko'proq degan faraz ostida ishlaydi. Biroq, bu odatda o'xshash bo'lmagan tillarda sodir bo'lmaydi. Bundan tashqari, k-vec algoritmi tilning eski ma'lumotlarini yoki uning ishlashini pasaytiradigan korpus xususiyatlarini hisobga olmaydi.

Bleualign - ushbu sentence aligner Riko Sennrich va Martin Volk tomonidan yaratilgan. Uning asosiy g'oyasi moslashtirish jarayonida yordam berish uchun MT tizimi va tarjima baholovchi, BLEUdan foydalanishdir [Sennrich, R., & Volk, M. 2010. 10-13b]. Alignerni yaxshiroq tushunish uchun BLEU haqida bilish kerak. BLEU - bu MT tarjima natijasi sifatini baholovchi algoritm. Ushbu sifatni o'lchash uchun BLEU MT tarjima natijasini bir yoki bir nechta inson tarjimalari bilan mosligini taqqoslash orqali baxolanadi.

Intrative Bleualign - Ushbu sentence aligner algoritmi 2011 yilda Riko Sennrich va Martin Volk tomonidan moslashtirish jarayonida mashina tarjima qilish tizimidan foydalanishdagi kamchiliklarni chuqurroq tahlil qilish natijasida yaratilgan. Sennrich va Volk MT asosidagi alignerlar manba matnining to'g'ri tarjimasiga kuchli bog'liqligini aniqladilar va MT tizimlari odatda moslangan matnlar bilan oziqlanganligini hisobga olsak, aylana bog'liqligi mavjudligi aniq bo'ladi. Ushbu qaramlikni bartaraf etish uchun ushbu algoritm moslashtirishni amalga oshirish uchun bootstrapping (yuklash) usulini taqdim etadi.

Ikki tilli korpusda jummalarni moslashtirishning leksik yondashuvlari turli tillardagi tegishli jummalarni aniqlash uchun leksik ma'lumotlardan foydalanishni o'z ichiga oladi. Ushbu yondashuv turli tillardagi tegishli jumlar o'xshash leksik tarkibga ega bo'ladi degan taxminga asoslanadi. Leksik asosda moslashtirishning keng tarqalgan usullaridan biri ikki tilli lug'atdan foydalanishga asoslangan. Ushbu yondashuv boshqa tilda to'g'ridan-to'g'ri tarjimalari bo'lgan bir tildagi so'zlarni aniqlashni va keyin tegishli jummalarni moslashtirish uchun ushbu tarjimalardan foydalanishni o'z ichiga oladi. Misol uchun, agar ingliz tilidagi "cat" so'zining o'zbek tilidagi "mushuk" ga to'g'ridan-to'g'ri tarjimasi bo'lsa, unda bu so'zlarni o'z ichiga olgan jummalarni moslashtirish mumkin.

Leksik yondashuvda moslashtirishning yana bir usuli statistik modellardan foydalanishni o'z ichiga oladi. Bu modellar jumlar orasidagi yozishmalarni birgalikdagi so'zlarning chastotasiga qarab aniqlash uchun ehtimollik algoritmlaridan foydalanadi. Misol uchun, agar turli tillardagi ikkita jumlada ko'plab umumiy so'zlar mavjud bo'lsa, ular mos kelishi ehtimoli ko'proq. Umuman olganda, leksikaga asoslangan yondashuvlar ikki tilli korpusdagi jummalarni moslashtirish uchun samarali bo'lishi mumkin, ayniqsa, birlashtirilayotgan tillar ko'p qarindosh bo'lsa (ya'ni, o'xshash shakl va ma'nolarga ega so'zlar). Biroq, bu usullar leksik tuzilmalari juda xilma-xil bo'lgan tillar uchun yoki idiomatik iboralar bilan ishlashda unchalik samarali bo'lmasligi mumkin.

Paragraf darajasida moslashtirilgan ikki tilli romanlar korpusi - bu ikki tildagi matnlar to'plami bo'lib, ular paraqraflari asosida tekislangan. Ushbu turdagi korpusda bir tildagi har bir paraqraf boshqa tildagi tegishli paraqraf bilan birlashtirilib, matnlarni oson taqqoslash va tahlil qilish imkonini beradi. Paraqraf darajasida moslashtirilgan ikki tilli romanlar korpusini yaratish odatda bir necha bosqichlarni o'z ichiga oladi. Birinchidan, har ikki tildagi romanlarni tanlab, raqamlashtirish kerak. Keyinchalik, har bir romandagi paraqraflarni aniqlash va

ajratish kerak. Keyin bir tildagi paragraflar boshqa tildagi tegishli paragraflar bilan mos kelishi kerak. Bu moslashtirish qo'lda yoki aligner dasturi yordamida amalga oshirilishi mumkin.

Moslash tugallangandan so'ng, ikki tilli korpus turli maqsadlarda ishlatilishi mumkin, masalan:

1. Qiyosiy adabiy tahlil: Tadqiqotchilar turli tillardagi romanlarning uslublari, mavzulari va hikoya tuzilmalarini solishtirishlari mumkin.

2. Tilni o'rgatish va o'rganish: O'qituvchilar korpusdan til o'rganuvchilarning o'qishni tushunish va tarjima qilish ko'nikmalarini yaxshilash uchun materiallar ishlab chiqishda foydalanishlari mumkin.

3. Mashina tarjimasi: Korpusdan mashina tarjimasi modellarini ularning aniqligi va ravonligini oshirish uchun o'rgatish uchun foydalanish mumkin.

Umuman olganda, paragraf darajasida moslashtirilgan ikki tilli romanlar korpusi tadqiqotchilar, til o'rganuvchilar va mashina tarjimasi tizimlarini ishlab chiquvchilar uchun qimmatli manba bo'lishi mumkin.

Xulosa. Korpus tilshunosligida statistik va leksik moslashtirish so'zlar yoki iboralarni parallel yoki taqqoslanadigan korpuslar o'rtasida moslashtirish uchun ishlatiladigan ikkita usuldir. Ikkala texnikaning ham afzalliklari va kamchiliklari bor va ulardan foydalanish aniq tadqiqot savollari va tadqiqot maqsadlariga bog'liq. Statistik moslashtirish katta hajmdagi ma'lumotlar bilan ishlashda va asosiy e'tiborni avtomatik qayta ishlashga qaratilayotganda foydali bo'lsa, leksik moslashtirish muayyan so'z yoki iboralarni semantik va sintaktik tahlil qilishga qaratilgan bo'lsa ko'proq mos keladi.

Foydalanilgan adabiyotlar:

1. Brown, P. F., Lai, J. C., & Mercer, R. L. (1991, June). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics.(169-176b)

2. Danielsson, P., & Ridings, D. (1997, February). Practical presentation of a “vanilla” aligner. In *TELRI Workshop in alignment and exploitation of texts*, February.

3. Fung, P., & Church, K. W. (1994, August). K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics. (1096-1102b)

4. Fung, P., & McKeown, K. (1994). Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. *arXiv preprint cmp-lg/9409011*.

5. Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1), (75-102b).



6. Gelbukh, A., Sidorov, G., & Vera-Félix, J. Á. (2006). A bilingual corpus of novels aligned at paragraph level. In *Advances in Natural Language Processing*. Springer Berlin Heidelberg. (16-23 b)
7. Harris, B. (1988). Bi-text, a new concept in translation theory. *Language Monthly*, 54,(8-10b).
8. Kit, C., Webster, J. J., Sin, K. K., Pan, H., & Li, H. (2004). Clause alignment for Hong Kong legal texts: A lexical-based approach. *International Journal of Corpus Linguistics*, (29-51b).
9. McEnery, A., & Xiao, R. Z. (2008). Paralell and comparable corpora: what are they up to?. *Incorporating Corpora: Translation and the Linguist. Translating Europe. Clevedon: Multilingual Matters*.
10. Macklovitch, E., & Hannan, M. L. (1998). Line ‘em up: advances in alignment technology and their impact on translation support tools. *Machine Translation*, 13(1),(41-57b).
11. Meyers, A., Kosaka, M., & Grishman, R. (1998, October). A multilingual procedure for dictionary-based sentence alignment. In *Conference of the Association for Machine Translation in the Americas*. Springer Berlin Heidelberg. (187-198b)
12. Moore, R. C. (2005, June). Association-based bilingual word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Association for Computational Linguistics. (1-8b)
13. Sennrich, R., & Volk, M. (2010, November). MT-based sentence alignment for OCRgenerated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado.
14. Sennrich, R., & Volk, M. (2011, May). Iterative, MT-based sentence alignment of parallel texts. In *18th Nordic Conference of Computational Linguistics, NODALIDA*
15. Simões, A. (2004). *Parallel corpora word alignment and applications* (master thesis). Universidade do Minho, Braga.
16. Natural Language Toolkit (NLTK): <https://www.nltk.org/>
17. British National Corpus (BNC): <https://www.natcorp.ox.ac.uk/>
18. Text Encoding Initiative (TEI): <https://tei-c.org/>

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ОБУЧАЮЩАЯ УЗБЕКСКОМУ ЯЗЫКУ КАК ИНОСТРАННОМУ НА БАЗЕ АНАЛИЗАТОРА РЕЧИ

Ибрагимова Сайёра Номазовна,

докторант, snmiraeva@mail.ru

Научно-исследовательский институт

Развития цифровых технологий и искусственного интеллекта

Аннотация. Ушбу мақолада келтирилган мавзу хозирги замонда жуда долзарб мавзулардан бири ҳисобланади. Таниб олиш технологияларининг кенг жорий этилиши одамларнинг турли билим соҳалари бўйича малакаларини оширишга имкон беради, улар орасида чет тилини ўргатиш алоҳида уринда туради. Нутқ сигналини фонемалар орқали матнга айлантиришнинг таклиф қилинаётган модели ҳақида қисқача тасниф келтириб ўтилган, шунингдек ўзбек тилини ўрганувчисининг ўзига хос хусусиятларини ҳисобга олган ҳолда лингвистик симулятор яратишнинг долзарблиги кўрсатиб ўтилган.

Аннотация. Тема, представленная в данной статье, является довольно актуальной в современном мире. Повсеместное внедрение технологий распознавания позволяет улучшать навыки людей в различных областях знаний, среди которых особое место занимает обучение иностранным языкам. В данной статье представлен краткий обзор работы технологии преобразования речи в текст, определение важности распознавания речи в современных условиях, а также актуальности создания лингвистического тренажера для обучения узбекскому языку с учетом особенностей говорящего.

Annotation. The topic presented in this article is quite relevant in the modern world. The widespread introduction of recognition technologies makes it possible to improve people's skills in various fields of knowledge, among which teaching foreign languages occupies a special place. This article provides a brief overview of the work of speech-to-text technology, determining the importance of speech recognition in modern conditions, as well as the relevance of creating a linguistic simulator for teaching the Uzbek language, taking into account the characteristics of the speaker.

Ключевые слова: *Лингвистический тренажер, распознавание речи, нейронная сеть, классификация, речевой сигнал.*

Обучение иностранных граждан фонетике узбекского языка – одна из важнейших задач в силу ряда причин: методологических, коммуникативных, этических и др. От освоения норм произношения зависит, в первую очередь, успешность овладения всеми видами речевой деятельности [1,2]. Кроме этого, многие документы часто оформляются в узбекском языке, то ошибки в произношении, написании и восприятии звуков могут стать причиной грамматических нарушений иностранных граждан.

Использование индивидуализации в обучении, в свою очередь, позволяет решить ряд вопросов, связанных с особенностями субъекта, осваивающего язык, с особенностями влияния родного языка, особенностями ряда внешних факторов, в первую очередь, обучающей среды. Именно влиянием этих факторов обусловлено стремление создать гибкую обучающую систему, представляющую собой речевой тренажер, где идет диалог между компьютером и человеком.

Лингвистический тренажер должен поддерживать как набор классических упражнений, развивающих навыки чтения и письма, так и упражнения на правильное звучание речи и аудирование. Виды учебного материала классического типа, предусмотренные в тренажере: работа с текстами, упражнения, выполнение тестов, работа со словарями. Дополнительно вводятся:

- работа с текстом, который набирает учащийся – его воспроизведение, анализ текста на грамматику;
- работа с речью, воспроизводимой учащимся;
- аудиальные тесты и упражнения.

Для вывода результата правильного произношения слов и допустимых ошибок, анализируется входной речевой сигнал для дальнейшего его распознавания. В большинстве систем распознавания речи присутствует этап сравнения входного сигнала с имеющимися эталонами.

Очень часто при распознавании речевых сигналов различные шумы, возникающие от аппаратуры или же различные окружающие шумы мешают. К тому же некоторые неточности в распознавании возникают из-за особенностей диктора, т.е. его диалект, дефекты речи при разговоре или же временные изменения характеристик голоса, например, хрип при болезни, что приводит к сложностям при распознавании речевых сигналов [3]. Кроме этого, немаловажную роль играет акцент обучающегося. Так как узбекский язык имеет специфические буквы такие как ғ, қ, ў, х, нг [4]. Их распознавание является немного сложной при присутствии акцента у обучающегося.

Процесс распознавания речи включает две основные составляющие: лингвистическая и акустическая.

Акустическая составляющая выполняет:

1. Преобразование речевого сигнала в цифровую форму, выделение фонем из слов;
2. Предварительная обработка речевого сигнала (нормализация, шумоподавление);
3. Вычисление признаков речевых единиц.
4. Сравнение принимаемого речевого сигнала с заранее подготовленным эталоном, описывающего определенный звук (буква, слово) в узбекской речи.

Лингвистическая составляющая системы преобразования речи в текст включает этапы, на которых привлекается априорная информация, характерная для используемого языка:

1. Создание базы знаний вероятностей сочетаний фонем, накладывающих ограничение на структуру слова, подчиняясь особенностям распознаваемого языка;
2. Построение корректной фонетической транскрипции из последовательности фонем;
3. Разработка базы словоформ узбекского языка;
4. Формирование грамматической формы слова;
5. Сравнение текстового написания входного речевого сигнала с эталонным текстом для вывода грамматических ошибок.

Предлагается нейросетевая модель пофонемного преобразования речи в текст для дальнейшей пофонемной проверки правильности произношения узбекских слов (рис. 1).

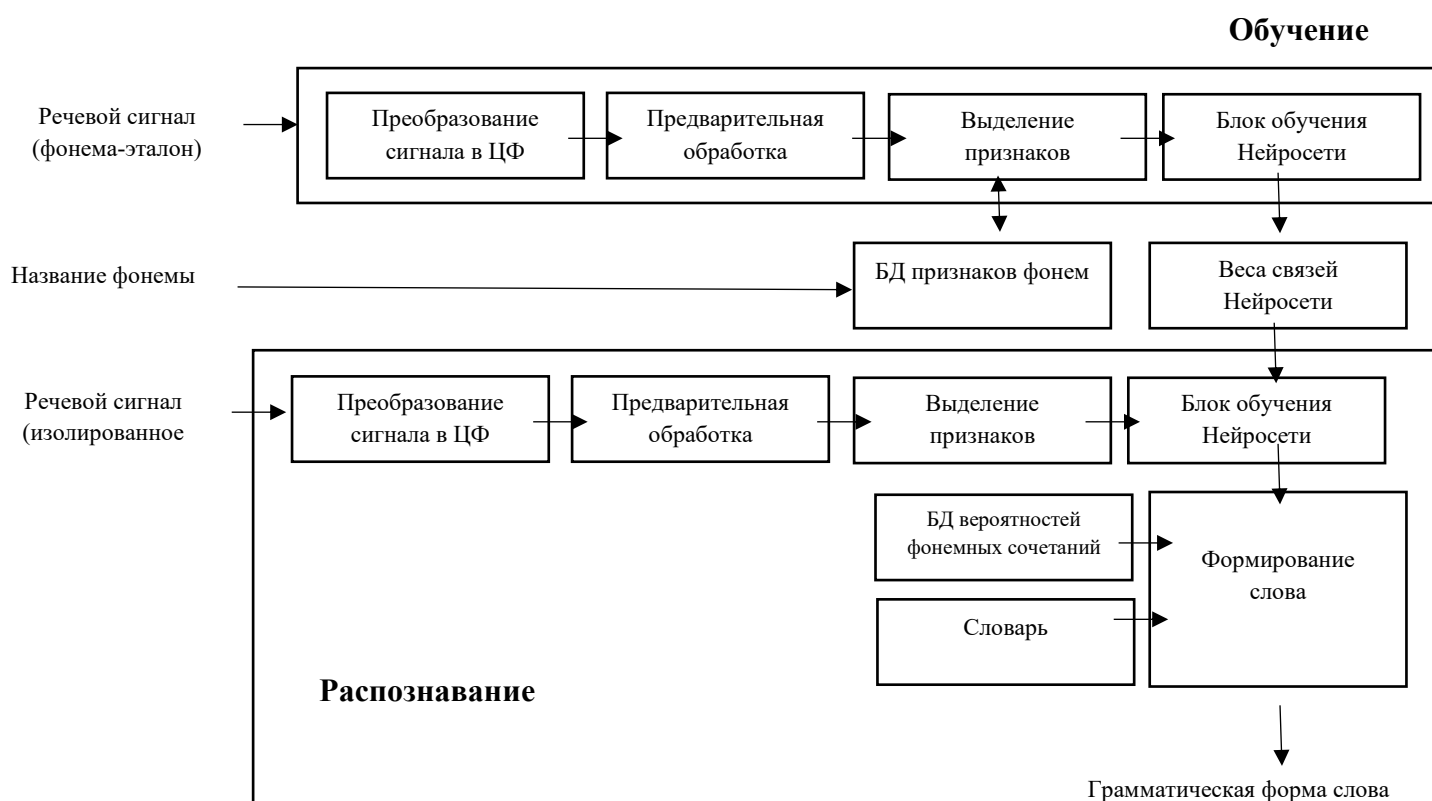


Рисунок 1. Структурная схема систему «Речь-текст»

- Выделяются следующие особенности данной модели:
- распознавание фонем осуществляется с помощью нейронной сети;
 - для повышения качества построения фонетической транскрипции применяются вероятности сочетаний фонем узбекского языка;

- в состав системы входит словарь, позволяющий формировать корректные грамматические формы слов;

- расширение словаря не требует дополнительного обучения системы, так как в качестве речевых единиц для распознавания использовались фонемы.

В режиме обучения на вход системы подается последовательность фонем, представляющая собой обучающее множество. Запись речевого сигнала осуществляется через микрофон, подключенный к входу звуковой карты компьютера, блок оцифровки сигнала выполняет преобразование сигнала в цифровую форму с заданными параметрами частоты дискретизации. Далее оцифрованный сигнал поступает на блок предварительной обработки. На данном этапе осуществляется очистка сигнала от шума, удаление участков сигнала, содержащих речевые паузы, проводится нормализация сигнала и его разбиение на фиксированные интервалы во временной области, на которых будут определяться характеристики.

В блоке выделения признаков рассчитываются параметры, характеризующие данную фонему-эталон. Вычисленные характеристики сохраняются в базе данных признаков. Следующим этапом является обучение нейронной сети, которое проводится на обучающем множестве, представленном в виде наборов признаков фонем. Результатом обучения нейросети является такая подстройка ее весов, при которой приложение некоторого множества входов связей нейросетей, обученной под конкретного диктора сохраняются и в дальнейшем используются при распознавании.

В режиме распознавания на вход системы подается речевой сигнал, в виде изолированного слова. К нему применяются этапы преобразования в цифровую форму, предварительной обработки и выделения признаков. В блоке распознавания рассчитанные векторы параметров последовательно подаются на входы обученной нейросети, которая их классифицирует. Таким образом, на выходе блока распознавания формируется последовательность распознанных сегментов фонем, которая затем передается в блок формирования слова, где с помощью БД вероятностей фонемных сочетаний и словаря происходит построение грамматической формы распознаваемого слова.

Описанная выше модель является частью лингвистического тренажера, которая позволяет преобразовать входной речевой сигнал в текст для дальнейшего показа ошибок на уровне фонем.

необходимо разработать интеллектуальную систему, учитывающую технологические особенности распознавания речевых сигналов, и которая будет применяться в качестве обучающего тренажера узбекскому языку, а также создать обширную базу данных эталонных моделей произношения звуков, букв и слов на узбекском языке.

В процессе обучения навыкам разговорной речи в интеллектуальной системе будет включать в себя 4 стадии:

1. Первичный тест обучаемого для дальнейшего выявления уникальных характеристик голоса, дефектов речи.

2. По ответам тестовой проверки обучаемому будет предложен прохождение курса упражнений для постановки речи. На данной стадии система выдает информацию о совпадении или несовпадении у обучающегося произнесенного с услышанным. В случае несовпадения система предлагает пройти упражнение еще раз для достижения правильной постановки звучания иностранных слов. Если система выдала положительный результат о совпадении, считается, что обучающийся успешно выполнил задание.

3. Промежуточное тестирование для сопоставления первичного тестирования с полученным результатом после прохождения определенного количества упражнений. Если система выявила повышение процента совпадения с эталонной моделью произношения, следовательно, курс упражнений остается неизменным, в противном случае система предлагает другие упражнения для постановки речи.

4. Итоговое тестирование обучающегося.

Таким образом, по итогу обучения на данном тренажере обучающийся улучшит навыки правильного произношения, написания и разговорной речи на узбекском языке, и присутствие учителя уже не будет обязательным.

Заключение

Представлен двухкомпонентная модель системы пофонемного преобразования речи в текст в лингвистических тренажерах для вывода грамматических ошибок обучаемого. Рассмотрены этапы акустической и лингвистической составляющих. Указаны основные стадии интеллектуального интерфейса лингвистического тренажера. Дальнейшим шагом научной работы будет являться сравнение двух речевых сигналов, т.е. сравнение входного сигнала с эталонным для проверки правильности произношения узбекских слов.

Список литературы:

1. Бернштейн С.И. Вопросы обучения произношению (применительно к преподаванию русского языка иностранцам). Вопросы фонетики и обучения произношению. М: Изд-во Моск. ун-та, 1975.

2. Бузанова Т.В. Роль и место фонетики в процессе обучения иностранцев русскому языку на начальном этапе // Русская и сопоставительная филология: Лингвокультурологический аспект / Казан. гос. ун-т. филол. Фак-т. Казань: Казан. гос. ун-т., 2004. С. 53-56.

3. Верещагина Е. А., Фомина Е. Р., Восприятие иностранного языка при помощи аудиального канала восприятия для обучения иностранному языку, 39 международная научно-методическая конференция «Актуальные вопросы качества образования», 8 декабря 2015 г.



4. Рахимова К.Н. (2018). Сфера применения информационных компьютерных технологий при изучении узбекского языка. Наука и образование сегодня, № 4(27), 64-65.

5. Н. В. Бордовской (ред.) (2011). Современные образовательные технологии (2-е изд., стр. 432). М.: КНОРУС

NUMPY KUTUBXONASI MUHITINI SOZLASH

Rizvanov Qodir Qaxramon o'g'li

Alisher Navoiy nomidagi
Toshkent davlat o'zbek tili va adabiyoti universiteti
Kompyuter lingvistikasi
mutaxassisligi 2-kurs magistranti

Annotatsiya: Ushbu maqolada Data science va sun'iy intellekt sohalarini o'rganish uchun kerakli bo'lgan Python dasturlash tilining ma'lumotlar bilan ishlovchi Numpy kutubxonasida ishlash muhitini sozlash keltirib o'tiladi.

Kalit so'zlar: *Data science, ma'lumotlar, tahlil, Numpy, muhandis, massiv, python, kutubxona*

Numpy Python dasturlash tilining kutubxonasi bo'lib, fan va sohalarning deyarli barchasida qo'llaniladi. Numpy kutubxonasi Pythonda raqamli ma'lumotlar bilan ishlash uchun universal standart hisoblanadi. Numpy foydalanuvchilari boshlang'ich kodlovchilardan tortib, eng zamonaviy ilmiy sanoat tadqiqotlari va ishlanmalari bilan shug'ullanuvchi tajribali tadqiqotchilargacha bo'lgan barchani o'z ichiga oladi.

Numpy kutubxonasi ko'p o'lchovli massiv va matrisali ma'lumotlar bilan ishlashni o'z ichiga oladi. U bir xil n o'lchovli massiv obyekti bo'lgan **ndarray** ni samarali ishlash usullari bilan ta'minlaydi. NumPydan massivlarda turli xil matematik amallarni yuqori tezlikda bajarish mumkin. Bu Python kutubxonasi bo'lib, u ko'p o'lchovli massiv obyekti, turli hosila obyektlari hamda massivlar ustida tezkor operatsiyalar, jumladan, matematik, mantiqiy, shakllarni manipulyatsiya qilish, saralash, tanlash, kiritish/chiqarish uchun tartiblar assortimentini taqdim etadi.

NumPy paketining *o'zagida ndarray* obyekti joylashgan. Bu bir xil ma'lumotlar turlarining n o'lchovli massivlarini qamrab oladi, ko'plab operatsiyalar ishlash uchun kompilyatsiya qilingan kodda bajariladi. NumPy massivlari va standart Python ketma-ketliklari o'rtasida bir nechta muhim farqlar mavjud:

- NumPy massivlari Python ro'yxatlaridan farqli o'laroq (dinamik ravishda o'sishi mumkin) yaratilayotganda qat'iy belgilangan hajmga ega. *Ndarray* o'lchamini o'zgartirish yangi massivni yaratadi va asl nusxasini o'chiradi.

- NumPy massividagi elementlarning barchasi bir xil turdagi ma'lumotlarga ega bo'lishi kerak va shuning uchun xotirada bir xil o'lchamda bo'ladi. Istisno: (Python, shu jumladan NumPy) obyektlar massivlariga ega bo'lishi mumkin, bu esa turli o'lchamdagi elementlarning massivlarini yaratishga imkon beradi.

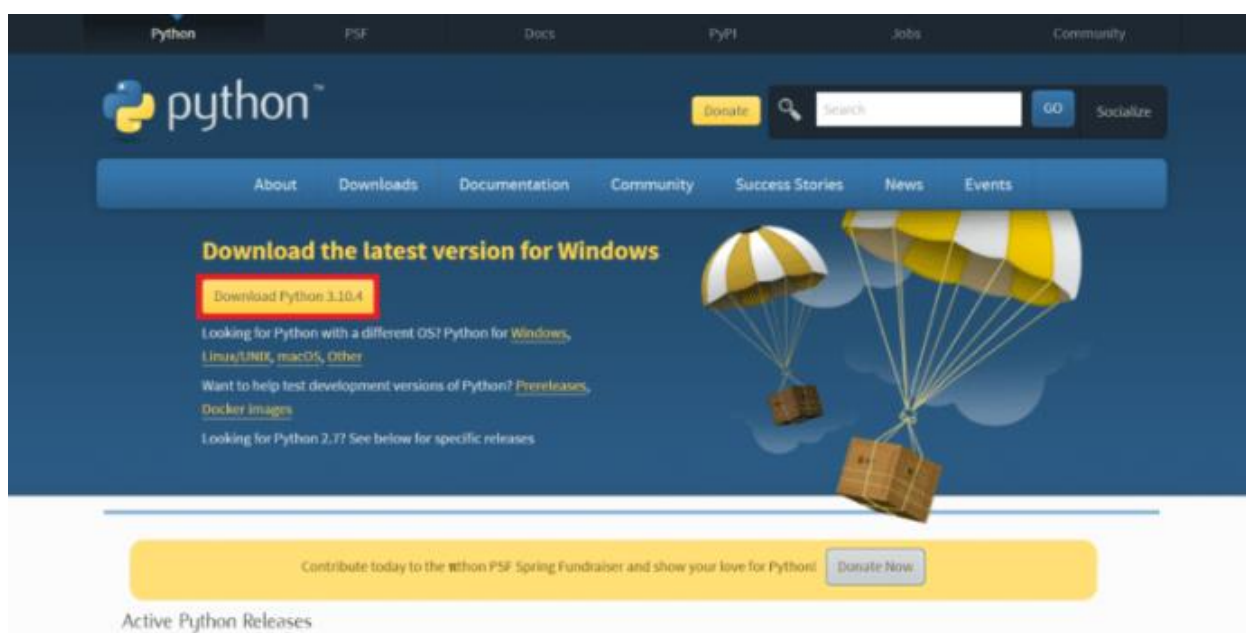
- NumPy massivlari katta miqdordagi ma'lumotlar bilan ilg'or matematik va boshqa turdagi operatsiyalarni osonlashtiradi. Odatda, bunday operatsiyalar Pythonning o'rnatilgan ketma-ketliklaridan foydalanish mumkin bo'lganidan ko'ra samaraliroq va kamroq kod bilan amalga oshiriladi.

• Pythonga asoslangan ilmiy va matematik paketlarning ko'pligi NumPy massivlaridan foydalanmoqda; Ular odatda Python-ketma-ket kiritishni qo'llab-quvvatlasa ham, ular qayta ishlashdan oldin bunday kirishni NumPy massivlariga aylantiradi va ko'pincha NumPy massivlarini chiqaradi. Boshqacha qilib aytganda, Python-ga asoslangan bugungi ilmiy/matematik dasturiy ta'minotning ko'p qismini (ehtimol ko'pini) samarali ishlatish uchun Pythonning o'rnatilgan ketma-ketlik turlaridan qanday foydalanishni bilish yetarli emas - NumPy massivlaridan qanday foydalanishni ham bilish kerak.[1]

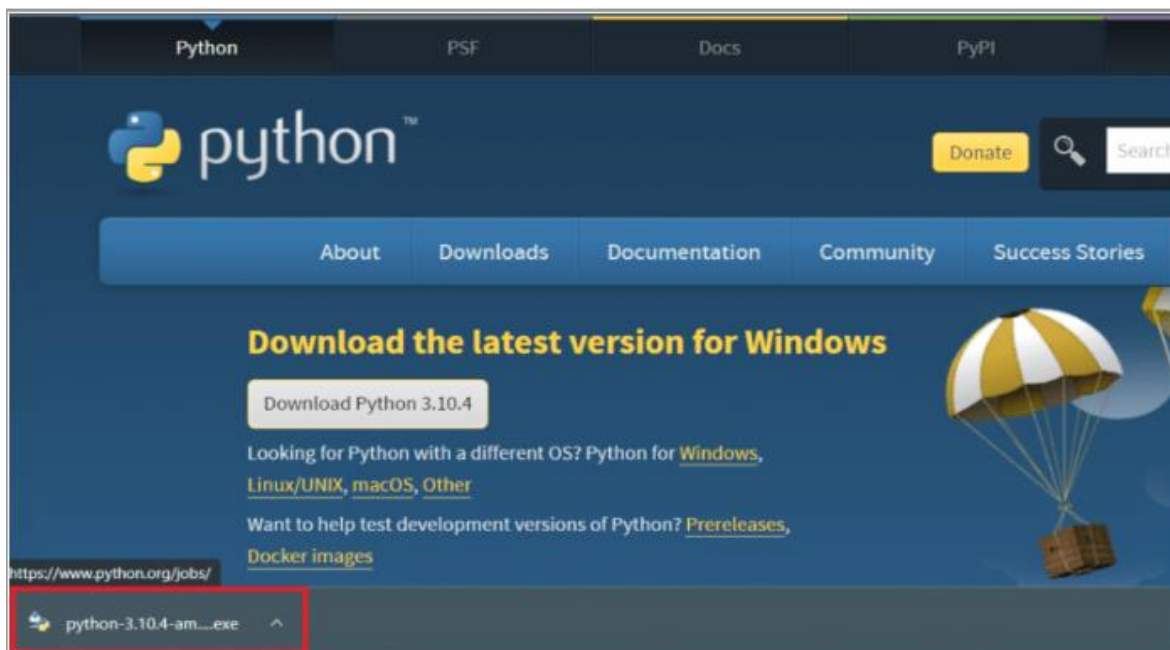
Python va Numpy kutubxonasini o'rnatish

1-qadam. Python dasturini o'rnatish

NumPyni o'rnatish jarayonidagi birinchi qadam Pythonni o'rnatishdir. NumPyni olishdan oldin tizimingizda eng so'nggi Python versiyasi o'rnatilgan bo'lishi kerak. Pythonni Windowsga qanday o'rnatishni bilish uchun quyidagi bosqichlarni o'qing va bajaring.[2]

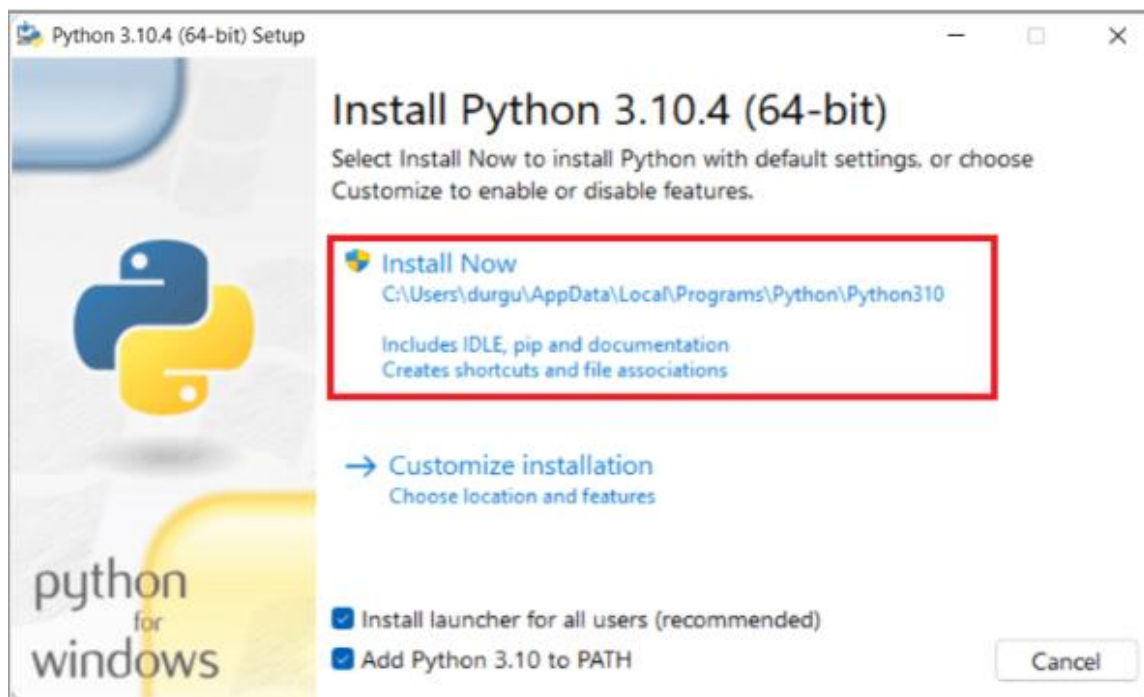


1-rasm. Python rasmiy veb-saytidan Python dasturini yuklab olish



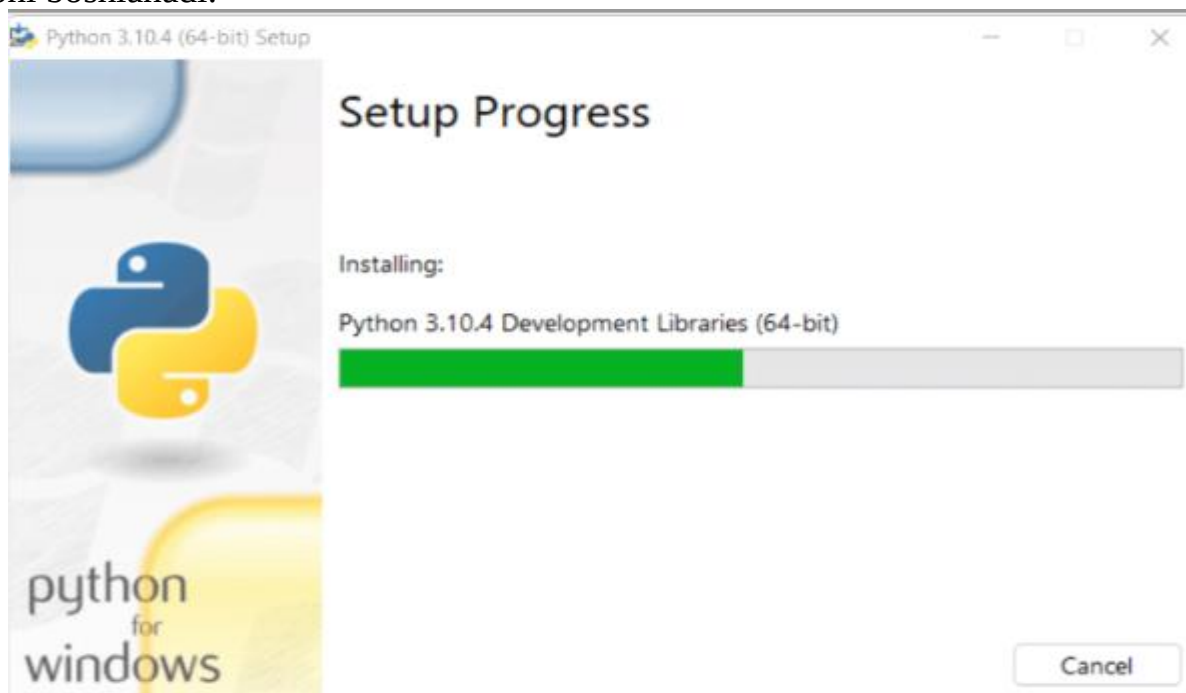
2-rasm. Python.exe dasturini o'rnatish

Demak, python.org rasmiy veb-saytiga kirganimizda bizni operations sistemamiz va qurilmamizga to'g'ri keladigan dastur variant asosiy ekranga chiqadi. Dasturning oxirgi versiyasida eng oxirgi o'zgarishlar mavjud bo'ladi. Shunda oxirgi versiyani tanlagan ma'qul bo'ladi. Hozir birni ekranimizda python 3.10.4 versiyani taklif qilyapti, demak ushbu versiyani qurilmaga o'rnatamiz.



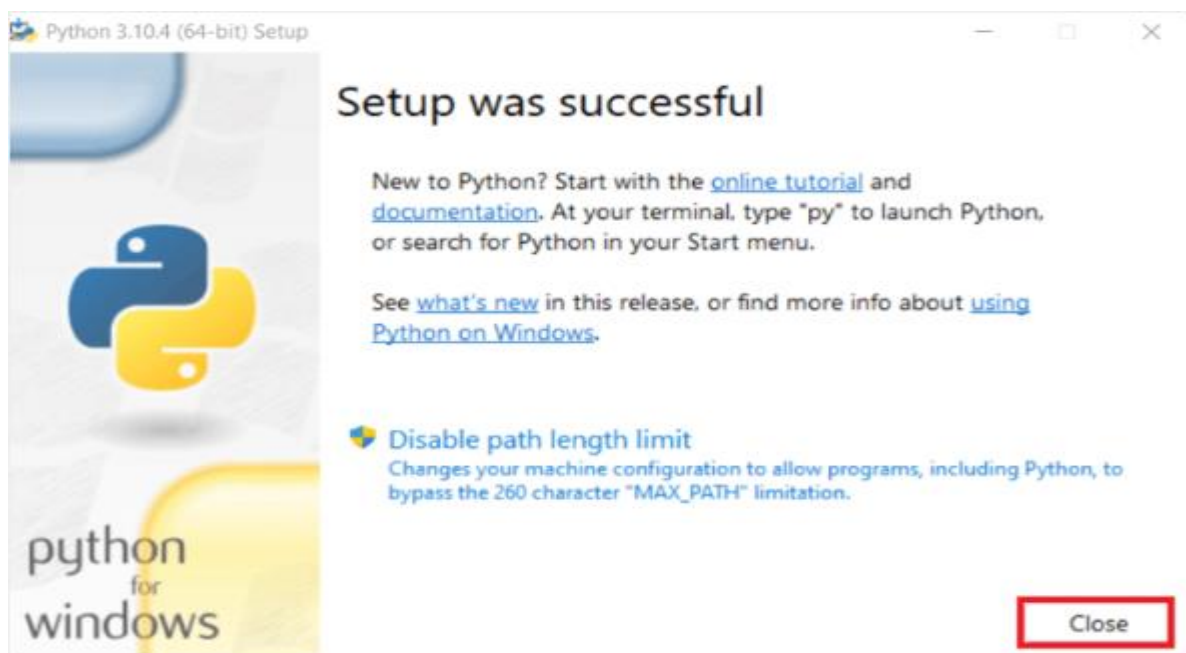
3-rasm. Endi o'rnatish variantini bosing.

O'rnatish jarayonida bizda yuqoridagidek oyna chiqadi va ushbu oynadan kerakli tanlanmalarni belgilab, “Install Now” ni tanlaymiz va quyidagicha o'rnatish jarayoni boshlanadi.



4-rasm. O'rnatish jarayoni avtomatik ravishda **O'rnatish jarayoni** oynasida boshlanadi.

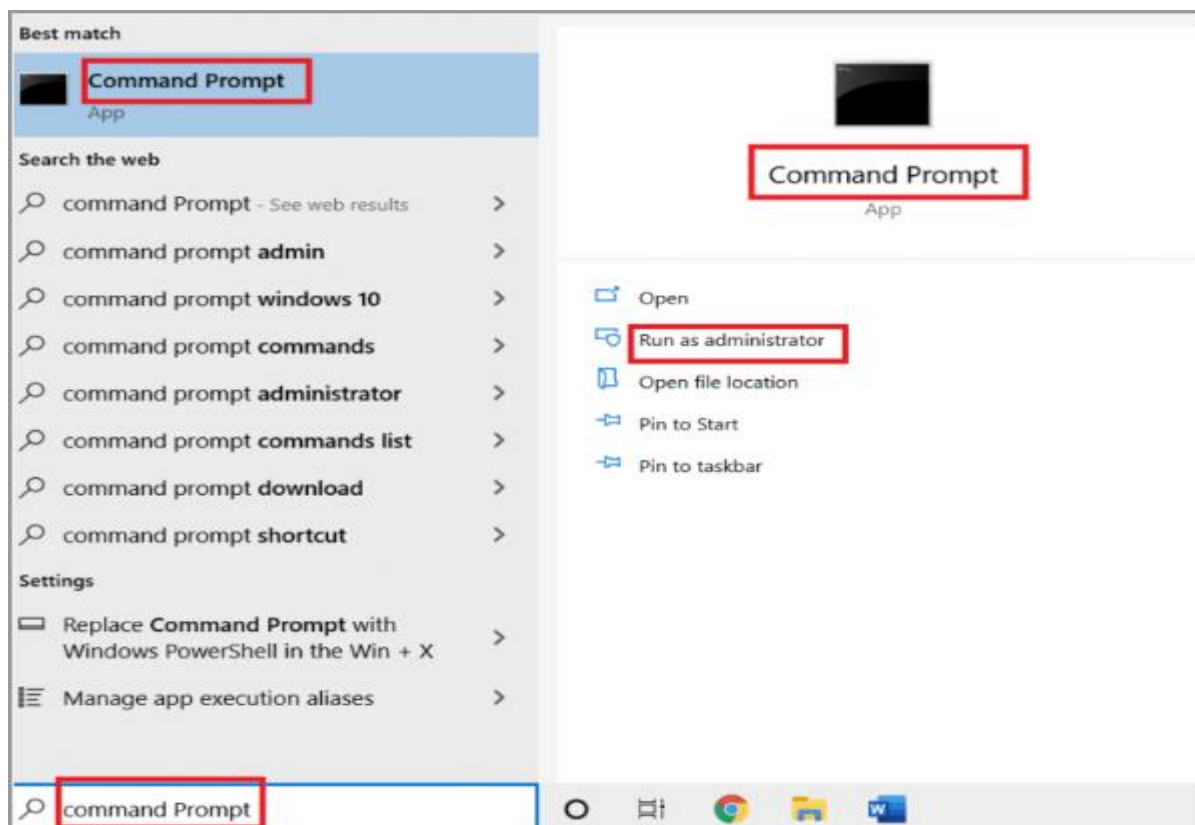
Ushbu jarayondan keyin bizda O'rnatish muvaffaqiyatni yakunlanganligini ko'rsatsa demak bizni qurilmamizda dastur o'rnatilib bo'lindi. “Close” tugmasi orqali oynani yopishimiz mumkin.



5-rasm. Nihoyat, ko'rsatilganidek, o'rnatish muvaffaqiyatli bo'lganligi haqida xabar olasiz . Yopish variantini bosib

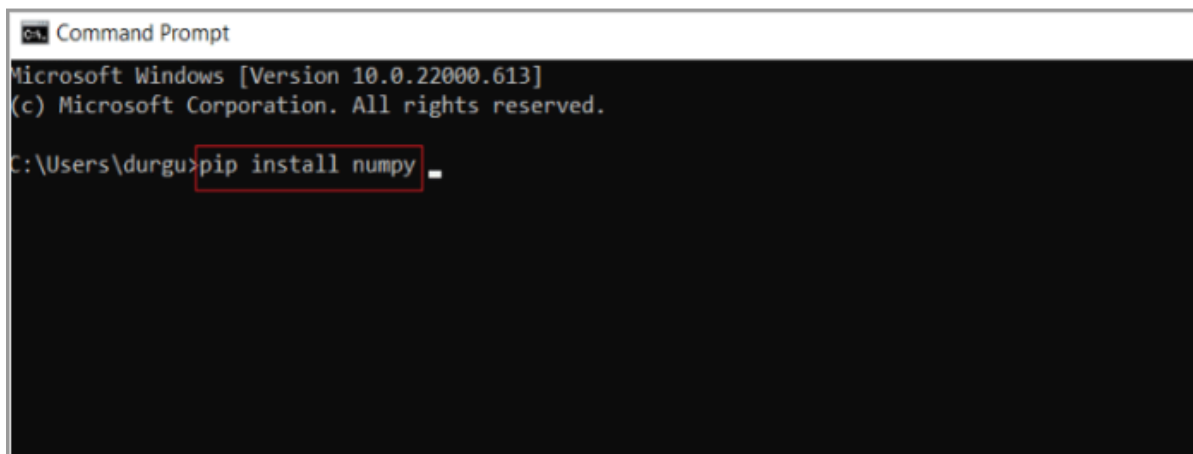
2-qadam. Numpy o'rnatish

Bizda keyingi qadam, qurilmaga Numpy kutubxonasini o'rnatish bo'ladi. Bunda kompyuter izlash joyiga “Command Prompt” deb izlaymiz va uni administrator sifatida ishga tushiramiz. Ketma-ketlik quyidagicha davom etadi.



6-rasm. Windows tugmachasini bosib, buyruq satrini kiriting va Administrator sifatida ishga tushirishni bosib

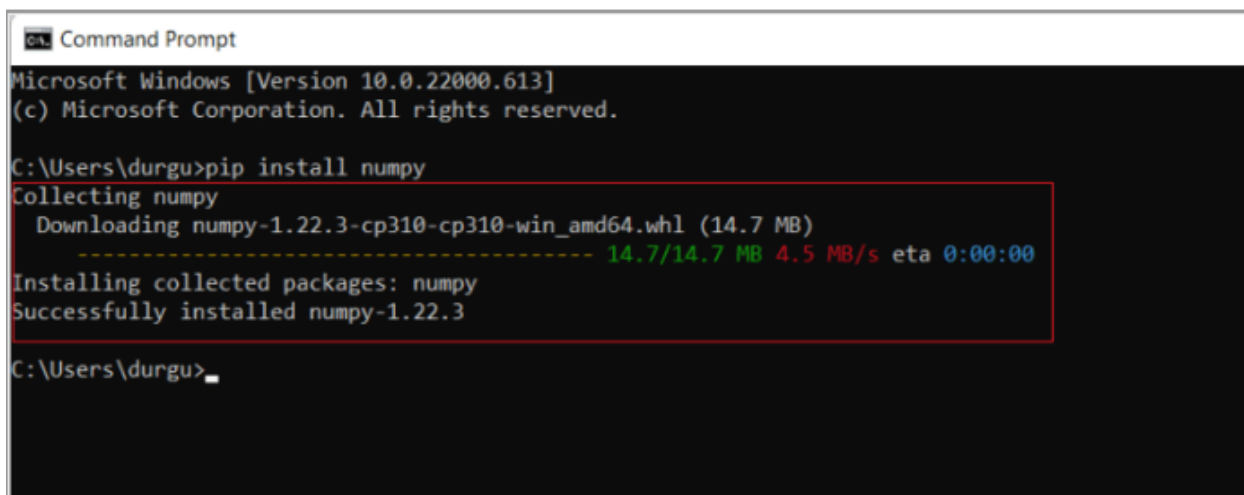
Ishga tushirilganda biz Numpy kutubxonasini o'rnatish uchun `pip install numpy` buyrug'ini beramiz. Birozdan keyin avtomatik kutubxonani o'zi yuklay boshlaydi. O'rnatish jarayonini quyidagicha kuzatishimiz mumkin. Numpy o'rnatilganligini tekshirish uchun shu “qora oyna”ga `numpy -version` buyrug'ini beramiz. Bizga shunda o'rnatilgan versiyani ekranga chiqaradi. Bizni ekranda bu **numpy 1.22.3** kabi ko'rsatilib turibdi.



```
Command Prompt
Microsoft Windows [Version 10.0.22000.613]
(c) Microsoft Corporation. All rights reserved.

C:\Users\durgu>pip install numpy
```

7-rasm. NumPy o'rnatilishini boshlash uchun **pip install numpy** buyrug'ini kiriting va Enter tugmasini bosong



```
Command Prompt
Microsoft Windows [Version 10.0.22000.613]
(c) Microsoft Corporation. All rights reserved.

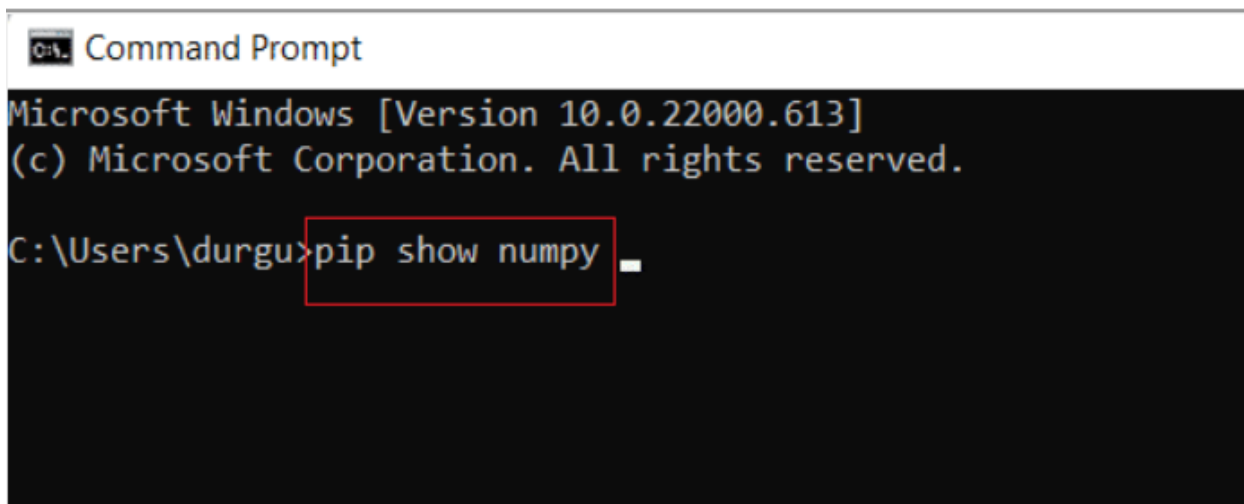
C:\Users\durgu>pip install numpy
Collecting numpy
  Downloading numpy-1.22.3-cp310-cp310-win_amd64.whl (14.7 MB)
----- 14.7/14.7 MB 4.5 MB/s eta 0:00:00
Installing collected packages: numpy
Successfully installed numpy-1.22.3

C:\Users\durgu>
```

8-rasm. NumPy paketini yuklab olish va o'rnatish avtomatik ravishda boshlanadi va tugaydi. Siz xabarni ko'rasiz: **numpy (-versiya)** muvaffaqiyatli o'rnatildi

3-qadam: NumPy o'rnatilishini tekshiring

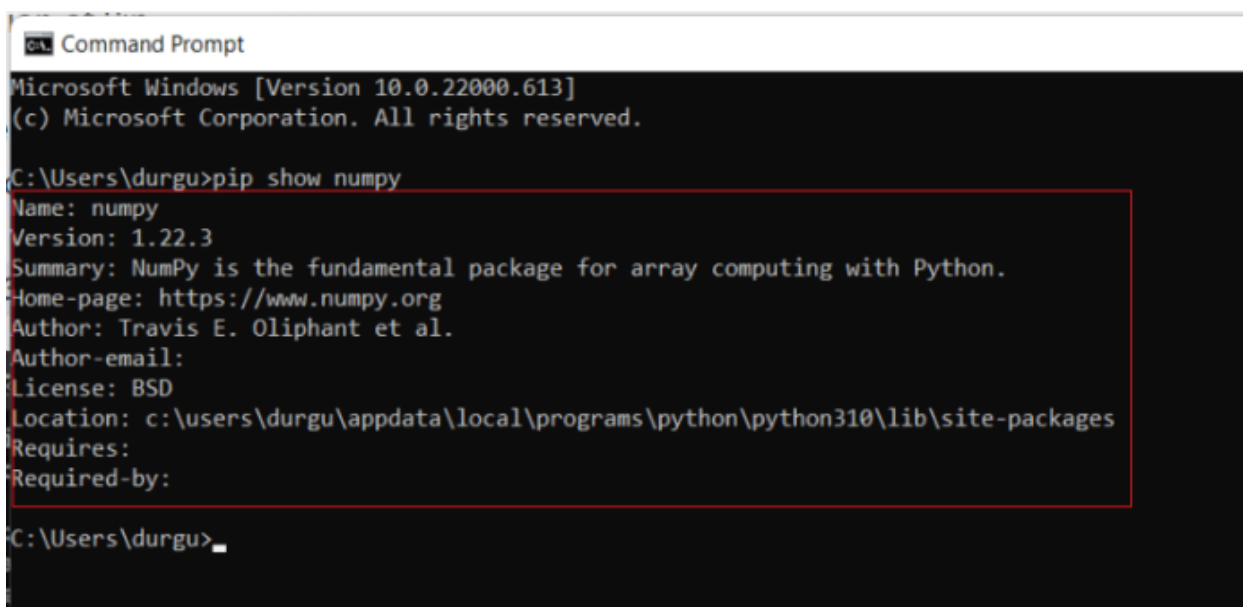
Keyingi qadamda esa biz Numpy o'rnatilgan joyni va o'rnatilish bo'yicha batafsil ma'lumot berilganini ko'rishimiz mumkin. Buni uchun biz "qora oynaga" **pip show numpy** buyrug'ini kiritganimizda ko'rishimiz mumkin.



```
Microsoft Windows [Version 10.0.22000.613]
(c) Microsoft Corporation. All rights reserved.

C:\Users\durgu>pip show numpy
```

9-rasm. Buyruq satrini ishga tushiring va **pip show numpy** buyrug‘ini kiriting va NumPy Python paketlarining bir qismi ekanligini tekshirish uchun Enter tugmasini bosing



```
Microsoft Windows [Version 10.0.22000.613]
(c) Microsoft Corporation. All rights reserved.

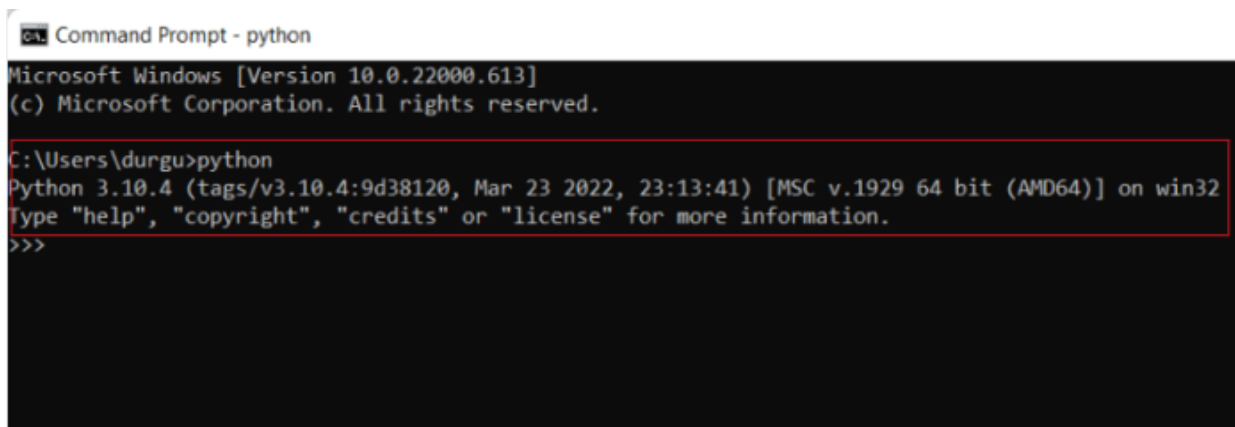
C:\Users\durgu>pip show numpy
Name: numpy
Version: 1.22.3
Summary: NumPy is the fundamental package for array computing with Python.
Home-page: https://www.numpy.org
Author: Travis E. Oliphant et al.
Author-email:
License: BSD
Location: c:\users\durgu\appdata\local\programs\python\python310\lib\site-packages
Requires:
Required-by:

C:\Users\durgu>
```

10-rasm. Chiqish sizga **NumPy versiyasini** tizimda saqlanadigan joyini ko‘rsatadi

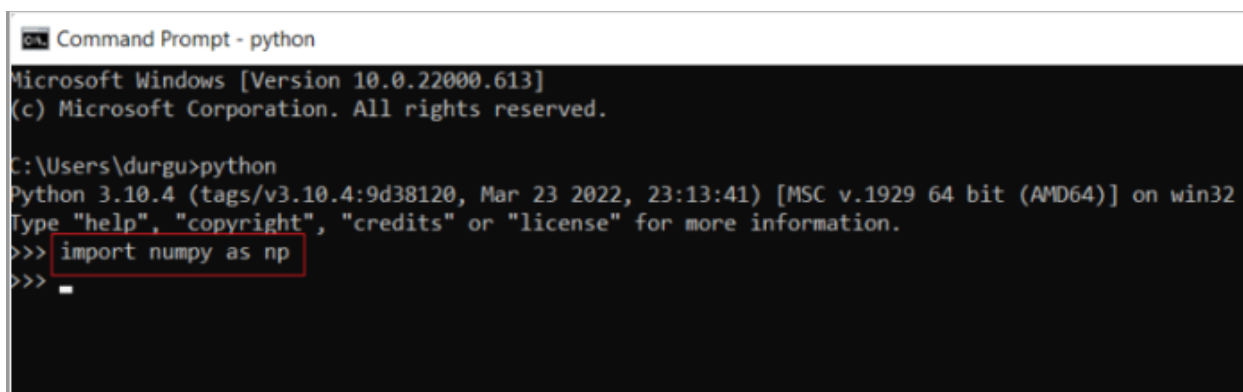
4-qadam: NumPy paketini import qiling

Muvaffaqiyatli o‘rnatilgan yoki yo‘qligini tasdiqlash uchun NumPy kutubxonasini import qilishingiz mumkin. Xuddi shunday qilish uchun keyingi bosqichlarni o‘qing va bajaring.[4]



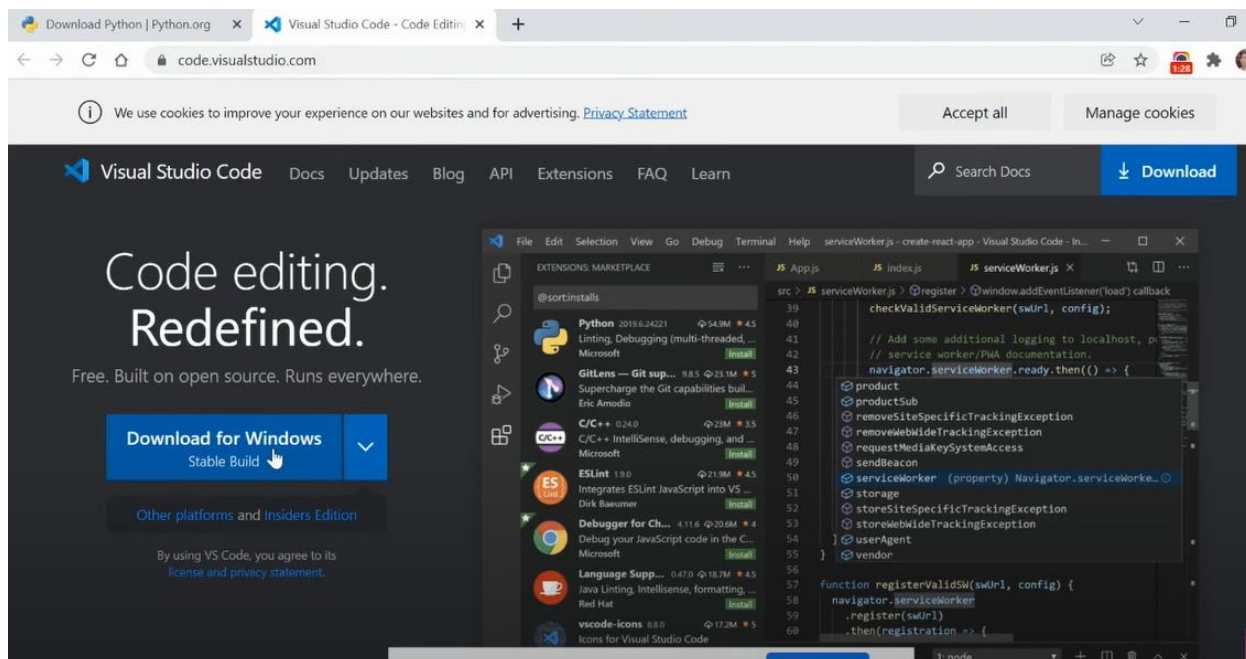
11-rasm. **Buyruq satrida Python** ni kiriting va Python muhitini yaratish uchun **Enter tugmasini bosing**.

Buyerda biz Numpy kutubxonasi o'rnatilgan yoki o'rnatilmaganligini tekshirish uchun uni import qilib ko'ramiz. Import qilish uchun esa import numpy buyrug'ini berishimiz kerak bo'ladi. Va biz numpy kutubxonasini muvaffaqiyatli import qildik.



12-rasm. Endi import numpy ni np buyrug'i sifatida kiriting va Enter tugmasini **bosing**.

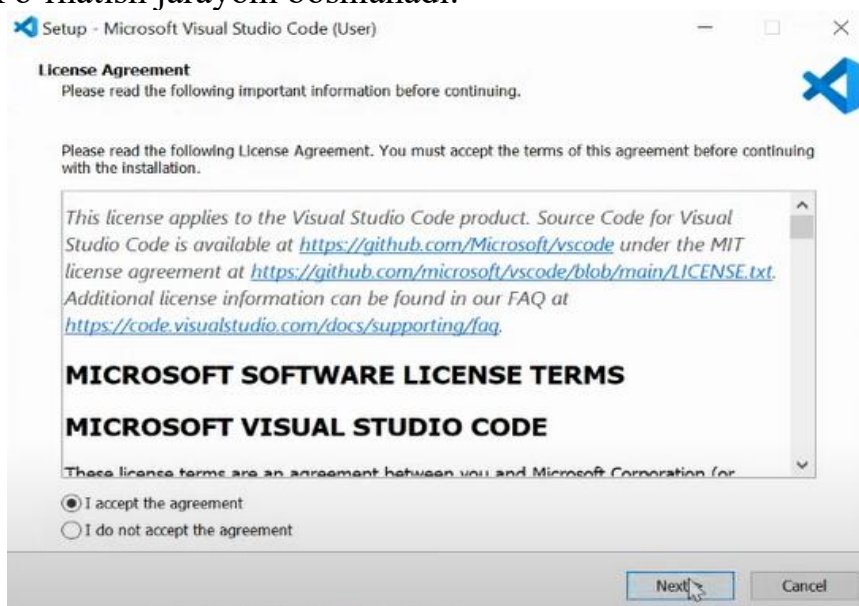
Endi siz NumPy kutubxona paketini muvaffaqiyatli import qildingiz va uning funksiyalari va sinflaridan turli NumPy obyektlari uchun foydalanishingiz mumkin.[3]



13-rasm. Visual Studio code dasturini rasmiy veb-saytidan yuklab olish va o'rnatish.

Endi bizga dasturdan va kutubxonadan foydalanish uchun kerakli muhitni yuklab, qurilmamizga o'rnatib olishimiz kerak bo'ladi. Bu yerda biz Visual Studio code Dasturini o'zimizga qulay muhit deb tanlab oldik.

Birinchi navbatda ushbu dasturni rasmiy veb-saytiga kirib yuklab olamiz. **code.visualstudio.com** havolasi orqali kirganimizda bu yerda ham bizga qurilmamiz uchun eng mos keladigan versiyasini asosiy ekranga yuklash sifatida chiqarib beradi. Yuklashni tanlab yuklab olamiz va quyidagicha oynada biz rozilik berganimizdan keyin avtomatik o'rnatish jarayoni boshlanadi.



14-rasm. Visual Studio code o'rnatish

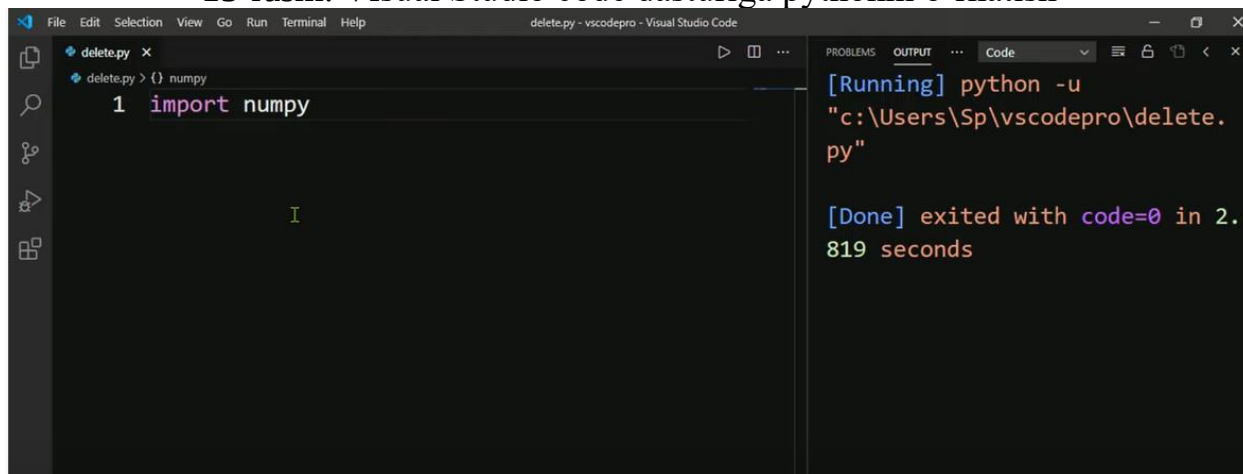
Dasturni o'rnatib bo'lgach biz dasturga endi python va numpyni o'rnatib olishimiz kerak bo'ladi. Chunki biz avvalgi qadamlarda qurilmamizga o'rnatdik, lekin hali endi o'rnatilgan bu dastur Python va Numpyni tanimaydi.

Shunda dasturni bo'limidan kerakli bo'lgan python dasturini tanlab, install tugmasini bosamiz. Shunda bizga python dasturini o'rnatib beradi.

Shundan keyin asosiy oynaga o'tib Numpy kutubxonasini import qilib olishimiz kerak. Buni uchun asosiy oynada **import numpy** buyrug'i bizga yordam beradi.[5]



15-rasm. Visual Studio code dasturiga pythonni o'rnatish



16-rasm. Numpy paketini import qilish

Demak, yuqorida Python dasturini o'rnatish, Numpy kutubxonasini o'rnatish, Visual Studio code dasturini o'rnatish va unga Python dasturini o'rnatish, Visual Studio code dasturiga Numpy kutubxonasini import qilishlarni ko'rib chiqdik. Kelgusida uning imkoniyatlari haqida alohida maqola qilishni o'yladik.

Foydalanilgan adabiyotlar:



1. Python for Data Analysis – Wes McKinney
2. NumPy 1.5 – Ivan Irdis
3. SciPy and NumPy
4. https://numpy.org/doc/stable/user/absolute_beginners.html
5. <https://havola.uz/e/numpy>
6. <https://community.uzbekcoders.uz/post/ma-lumotlar-tahlili-data-analysis-uchun-python-dasturlash-tili-distributorl--5ee9f044d73750310952c31a>
7. <https://techcult.com/how-to-install-numpy/>
8. <https://www.youtube.com/watch?v=I-E09N5-lzk>

NER (NAMED ENTITY RECOGNITION – NOMLANGAN OBYEKTNI TANIB OLISH) OBYEKT LARI TARKIBIDA TOPONIMLARNI AVTOMATIK ANIQLASH

Samatboyeva Madina To'liqinjon qizi

ToshDO'TAU,

Kompyuter lingvistikasi mutaxassisligi magistranti

msamatboyeva@gmail.com

Annotatsiya: Ushbu maqolada NER (Named Entity Recognition) obyektlari tarkibiga kiruvchi toponim (joy nomi)lar haqida fikr yuritildi va ularning turlari tadqiq qilindi. Maqolada atoqli va turdosh ot tushunchasiga ham ta'rif berildi. Toponim turlari jadvallar asosida misollar bilan birgalikda keltirib o'tildi.

Kalit so'zlar: *NER, obyekt, nom, nomlangan obyekt, turdosh ot, atqli ot, joy nomi, toponim, leksika.*

Abstract: In this article, toponyms included in NER (Named Entity Recognition) objects were considered and their types were studied. The article also defined the concept of common nouns and common nouns. Types of toponyms are given together with examples on the basis of tables.

Key words: *NER, object, name, named object, cognate noun, proper noun, place name, toponym, lexicon.*

Аннотация: В этой статье рассматриваются топонимы (географические названия), которые являются частью объектов NER (Named Entity Recognition), и исследуются их типы. В статье также дается определение понятия имя собственное и имя существительное нарицательное. Типы топонимов приведены вместе с примерами на основе таблиц.

Ключевые слова: *NER, объект, имя, названный объект, имя существительное нарицательное, имя собственное, топоним, лексика.*

Bizni o'rab turgan har qanday geografik obyektning nomi bor va bu “nom” obyektlarni bir-biridan farq qilish uchun xizmat qiladi. Shu sababdan, tildagi ko'pgina so'zlar nomlarga (atoqli otga) aylangan [1]. Kundalik hayotimizdagi “kutubxona”, “maktab”, “bog'cha” so'zlari ham obyektни bildiruvchi so'zlar, ammo ular turdosh ot (appellyativ – lotincha “atoqli ot aksi”)lardir. Turdosh va atoqli ot tushunchalari o'zi anglatib kelgan obyektning “nomlanish” xususiyatiga ko'ra bir-biridan farq qiladi. Ya'ni, nomlangan hamda nomlanmagan obyekt mazmunning kengayishi va qayta nomlanish sifati bilan ajralib turadi.

NE (Named Entity) – ot so'z turkumi doirasida nomlangan obyekt – “atoqli ot”ni ifodalaydi [2]. Nomlangan obyektlarni avtomatik aniqlovchi jarayon – **NER (Named Entity Recognition)** – nomlangan obyektни tanib olish bo'lib, matndagi

barcha nomli obyektlar aniqlanadi. NER **tabiiy tilni qayta ishlash (NLP – Natural Language Processing)** ning ko‘plab sohalarida qo‘llaniladi [3]. NER obyektlaridan biri – *joy nomlari (toponim)*dir.

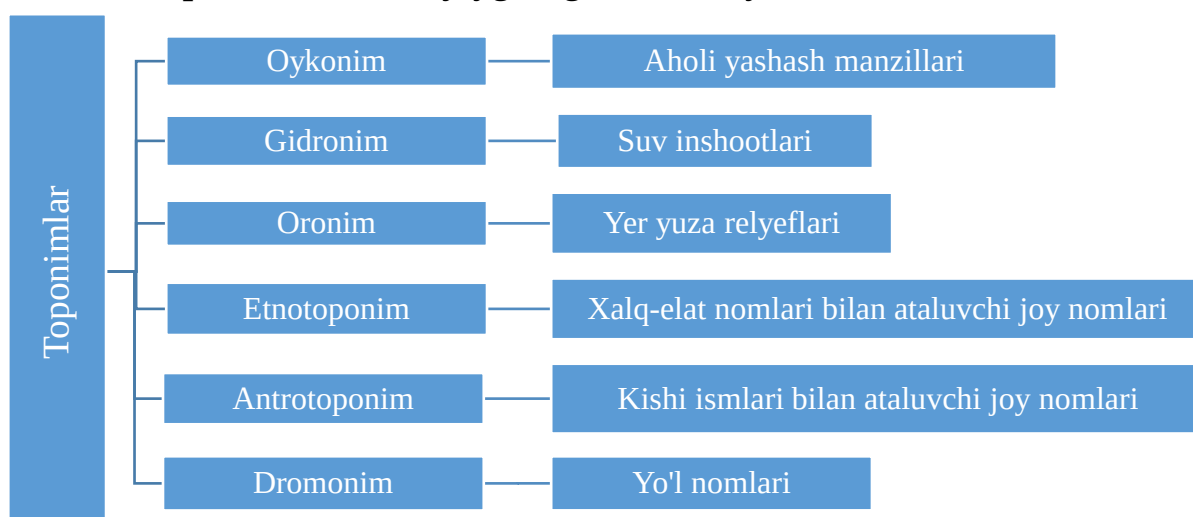
Toponimlar

Yuqoridagi turdosh ot sifatida keltirilgan misollar (“kutubxona”, “maktab”, “bog‘cha”) “joy”ni bildiruvchi obyektlarni ifodalaydi. Ushbu so‘zlarni eshitganimizda, ko‘z oldimizda umumiy “bino” ko‘rinishiga ega obyektlar tasvirlanadi. Agar ushbu joyni bildiruvchi obyektlar qayta nom bilan nomlansa (masalan, “Alisher Navoiy kutubxonasi”, “Avloniy maktabi”, “Lolazor bog‘chasi”), ular atoqli otga aylanadi va “toponimlar” deb yuritiladi. Joy nomlari, **geografik nomlar** yoki **toponimlar** deb ataladi. Toponimlarni **toponimika** fani o‘rganadi. **Toponimika** yunoncha **topos** - joy va **onoma (yoki onima)** - nom so‘zlaridan tarkib topgan.

Toponimlar til lug‘at tarkibining bir qismi bo‘lib, til qonuniyatlariga bo‘ysunadi [4]. Toponimlar tilshunoslik (lingvistika)ning onomastika (yunoncha “onomastike” - nomlash, nom qo‘yish san’ati [5] bo‘limida o‘rganiladi.

Toponimlar joy xususiyatini, uning tarixini, ushbu obyekt bilan bog‘liq voqea-hodisani anglatgan oddiy so‘z va nomning birikuvidan hosil bo‘lgan atoqli otdir. Toponimlar doirasida uchta tushunchani farqlab olish lozim: **“Toponimika”**, **“Toponimiya”**, **“Toponim”**.

- **Toponimika** – joy nomlari, ularning tuzilishi, geografik nomlarning hosil bo‘lishi va rivojlanishi bilan shug‘ullanuvchi *fan*;
- **Toponimiya** – *joy nomlari yig‘indisi* (masalan, Buxoro toponimiyasi – Buxoro joy nomlari yig‘indisi);
- **Toponim** – bu bitta joyga tegishli *xususiy nom*.



1-rasm. Toponim turlari

Toponimlar tilshunoslik nuqtayi nazari bilan o‘rganilganda nomning paydo bo‘lishi, uning tarixi, yaratilishi, leksik-semantik tasnifi, lingvistik strukturasi, so‘z

turkumi, uning atoqli otligi, qaysi tilga mansub leksika ekanligi, toponim bilan bog'liq til qonuniyatlari o'rganiladi. Ammo, rivojlanib borayotgan asrimizda, lingvistik qarashlar kompyuter tiliga moslashtirilib, zamonaviy elektron dasturlar yaratilayotgan bir davrda, qo'lda joy nomlarini aniqlash va ular bilan bog'liq statistik ilmiy tadqiqot ishlarini olib borish mushkul.

Ma'lum bir hududdagi geografik nomlarning yig'indisi shu hududda yashovchi xalqlarning asrlar davomida, nomlar yaratishdagi ijod mahsuli hisoblanadi. Dunyoda qancha geografik nom borligini hech kim aniq bilmaydi. Taxminiy hisoblarga ko'ra, butun yer sharida yarim milliarddan ortiq geografik nom bor. Holbuki, ular qatoriga soy, jilg'a, buloq, quduq, jar, qir, mahalla, guzar, ko'cha kabi mikrotoponimlar kirmaydi. Agar, yer yuzidagi barcha katta-kichik joy nomlarini hisobga olish imkoni bo'lganda, son-sanoqsiz raqam hosil bo'lar edi [1]. Ushbu zaruratlar matnni qayta ishlash jarayonida joy nomlarini avtomatik aniqlash ishini amaliyotga joriy etish kerakligini taqozo etmoqda.

1. **Oykonim** – aholi yashash manzillari bo'lib, “polinimiya”, “urbanomiya” atamallari bilan ham yuritiladi. Aholi yashash manzillari tuzilish tarkibi turli leksikalardan iborat bo'ladi.

Toponimlarning *oykonim (aholi yashash manzillari)*lar turi hosil bo'lishida ishtirok etadigan leksikalar:

- Suvga oid leksikalar;
- Yer yuza relyeflariga oid leksikalar;
- O'simlik, daraxtlarga oid leksikalar;
- Hayvonlarga oid leksikalar;
- Kasb-hunarga oid leksikalar.

1-jadval. Oykonimlar hosil bo'lishida ishtirok etadigan leksikalarga misollar

№	Leksika turi	Oykonimni hosil qiluvchi leksika	Hosil bo'lgan toponim(lar)
1	Suvga oid leksikalar	Daryo	Daryobo'yi, Daryoliq, Sirdaryo, Qoradaryo
		O'zak – “suvi ba'zan qurib qoladigan kichik ariq”	Ko'ko'zak, Injenero'zak, Sario'zak
		Ro'd, ro'dbor – “daryo”, “kanal”	Ro'dak, Labiro't, Ro'dasoy
		Kanal	Kanalboyi, Kanalobod, Kanalpasti
2		Adir	Bo'lakadir, Adirmoq

	Yer yuza relyeflariga oid leksikalar	Bel	Beltoʻv, Beltepa
		Dara	Omondara, Palandara
		Doʻng	Doʻng, Doʻngariq
3	Oʻsimlik, daraxtlarga oid leksikalar	Paxta	Paxta, Paxtakor, Paxtakon, Paxtakesh
		Ajriq	Ajriqti
		Arpa	Arpa, Arpakoʻl, Arpalikoʻl, Arpapoya
		Archa	Archazor, Yakkaarcha
4	Hayvonlarga oid leksikalar	Aygʻir	Aygʻir, Aygʻirtepa
		Baliq	Baliqli, Baliqlikoʻl
		Burga	Burgali, Burgalik
		Boʻri	Boʻrisoy, Boʻritepa
5	Kasb-hunarga oid leksikalar	Aravachi	Aravakashlar, Aravachilar
		Baxshi	Baxshilar, Baxshiyon, Baxshitepa
		Baliqchi	Baliqchi, Baliqchilar
		Boʻzchi	Boʻzachi, Boʻzchi

2. **Gidronimlar** – (yunoncha “hidro” — suv, “nom” — nom) suv inshootlari nomi boʻlib, okean, dengiz, koʻl, soy, jilgʻa, quduqning har biri suv obyektidir [4]. Hidronimlarning ham tuzilish tarkibi suv inshootiga tegishli boʻlgan leksikalardan iborat boʻladi.

2-jadval. Hidronimlar hosil boʻlishida ishtirok etadigan leksikalarga misollar

№	Gidronimlarga oid leksikalar	Hosil boʻlgan toponim(lar)
1	Dengiz	Dengizkoʻl
2	Daryo	Sirdaryo, Qashqadaryo, Qoradaryo
3	Ariq	Yangiariq, Toshariq, Moyariq, Ariqboʻyi, Beshariq
4	Koʻl	Koʻliyon, Koʻlob, Labikoʻl, Luqlikoʻl, Oqkoʻl
5	Rud	Shahrud (shahr i rud), Bolorud, Labirud (lab i rud)

6	Nahr (nar, nor)	Nahr ash-Shosh, Ulug'nor, Norin, Minor,
7	Quduq	Sho'rquduq, Uchquduq, Jalolquduq, Oqquduq

Bundan tashqari yana bir qancha topotermin – gidronimlarni bildiruvchi (yoki ularga ishora qiluvchi) leksikalar mavjud. Masalan, *seldara*, *akvatoriya*, *ob*, *qayirma*, *obshor*, *chashma*, *havza*, *quvur*, *hovuz*...

3. **Oronimlar** – tog'-u toshlarga emas, balki re'lyefning salbiy shakllari - vodiylar, daralar, jarliklar, soyliklar, o'yiq-o'ypotlar, shuningdek, tekislik, past tekisliklar va qumliklar nomlaridir. [4]

5-jadval. Oronimlar hosil bo'lishida ishtirok etadigan leksikalarga misollar

№	Oronimlarga oid leksikalar	Hosil bo'lgan toponim(lar)
1	Ranglar (Oq, qora, sari(q), qo'ng'ir...)	Qo'ng'iro't, Qo'ng'irobod, Qoraqo'rg'on, Saribozor
2	Bel – Baxmal, Zomin tumanlarida “dovon”	O'rtabel, Ayribel, Beltov, Oqbel, Qumbel
3	Dovon – “tog'dan oshaverishga yaqin dam olinadigan tekisroq maydoncha”	Kandirdovon
4	Egar (Navoiyda va qurama shevasida – “iyar”)	Egarchi, Iyartosh, Egarbozor
5	Zov (zovlin)	Zovurak, Avezov, Surxzov
6	Qir (qirang, qirpang)	Qaqir, Qirguli, Qoraqir, Uzunqir

4. **Etnotoponim** – xalq-elat nomlari bilan ataluvchi toponim (joy nomlari)dir. Toponim – joy nomiga nisbat berilganda, nomning kelib chiqishi ko'p hollarda tarixiy aspektlarga asoslanadi. Buni O'zbekiston toponimiyasi jihatidan tadqiq etsak, mamlakatimiz xududida qadimdan turli qabila-urug'lar birgalikda yashab kelishgan. Tarixiy manbalarda ham O'zbek xalqining o'zi 99 urug'dan iborat bo'lganligi haqida ma'lumotlar keltiriladi. Qaysi bir urug' yangi bir joyda o'troq o'rinishsa, shu joy nomini o'z urug'i nomi bilan atashgan. Hozirgi kunda joy nomlari etnotoponim bo'lsa, ushbu hududda qachonlardir ushbu urug'ga mansub xalq yashagan deb taxmin qilinadi.

6-jadval. Etnotoponimlar hosil bo'lishida ishtirok etadigan

leksikalarga misollar

№	Etnotoponim oid leksikalar	Hosil bo'lgan toponim(lar)
1	Qozoq	Qozoq, Qozoqon
2	O'zbek	O'zbekon, Yuqori O'zbekon
3	Dovon – “tog‘dan oshaverishga yaqin dam olinadigan tekisroq maydoncha”	Kandirdovon
4	Egar (Navoiyda va qurama shevasida – “iyar”)	Egarchi, Iyartosh, Egarbozor
5	Zov (zovlin)	Zovurak, Avezov, Surxzov
6	Qir (qirang, qirpang)	Qaqir, Qirguli, Qoraqir

6-jadval. Dramonim (yo‘l, ko‘cha nomlari)larga misollar

№	Etnotoponim oid leksikalar	Hosil bo'lgan toponim(lar)
1	“Yo‘l” so‘zi qo‘shib yoziladigan dramonimlar	Yangiyo‘l, Toshyo‘l, Eskiyo‘l
2	“Yo‘l” so‘zi ajratib yoziladigan dramonimlar	Samarqand xalqa yo‘li, Buyuk ipak yo‘li, Yangi yo‘l
3	“Yo‘l” so‘zi chiziqcha(defis) bilan yoziladigan dramonimlar	Toshkent-Jizzax yo‘li, Toshkent-Guliston yo‘li

Xulosa qilib shuni aytish mumkinki, toponimlarni o‘rganish va ularni avtomatik aniqlovchi dastur modellarini yaratish matnlar ustidagi kompyuter yordamida bajariladigan ishlarni osonlashtiradi. Matnlar qayta ishlanadi va qidiruvga asoslangan dasturlarda ham ushbu modellar keng qo‘llaniladi. Toponimlar aniqlanganda matn mazmuniga asoslangan asosiy fikrni avtomatik olishga ham yordam beradi.

Foydalanilgan adabiyotlar:

1. Q. Hakimov. Toponimika. “Mumtoz so‘z” nashriyoti. Toshkent. 2016. 5 b, 47 b.
2. Rachna Jain, Abhishek Sharma, Gouri Sankar Mishra, Parma Nand, and Sudeshna Chakraborty. Named Entity Recognition in English Text. Journal of Physics: Conference Series. 2020
3. Susan Li. Named Entity Recognition with NLTK and SpaCy. 2018.1p
<https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>



4. S.Qorayev. Toponimika. O'zbekiston faylasuflari milliy jamiyati nashriyoti Toshkent. 2006. 5-6 b, 47 b, 66 b
5. <https://qomus.info/oz/encyclopedia/o/onomastika>

LEKSIK SINONIMLARNI ANIQLASH UCHUN WORD2VEC VA ROBERTAFORMASKEDLM MODELLARIDAN FOYDALANISH. SYN- ROBERTA MODELI HAQIDA

Uzoqova Mohiyaxon Tuyg'un qizi
ToshDO'TAU

Kompyuter lingvistikasi mutaxassisligi 1-kurs magistranti
mohiyaxonuzokova@gmail.com

Annotatsiya. Leksik sinonimlarni lug'atga asoslangan usul, ya'ni avvaldan shakllantirilgan sinonimlar bazasi bilangina avtomatik aniqlash to'laqonli o'zini oqlamasligi mumkin. Sababi sinonim sifatida taqdim etilgan so'zlar kontekstga mos tushmasligi mumkin. Bu esa o'zbekcha sinonimayzer dasturidan foydalanishda noqulaylik tug'diradi. Ushbu maqolada sinonimlarni kontekstni inobatga olgan holda aniqlashda foydali deb topilayotgan Word2Vec va RoBERTa modellariga nazar tashlandi. Qolaversa, Word2Vec, RoBERTa modeli va barcha uslub va ohang doirasida teglangan sinsetlar bazasi integratsiyasi asosida ishlaydigan Uz-Synonymizer loyihasiga to'xtalib o'tildi.

Kalit so'zlar: Word2Vec, RoBERTa, vektorayzer, tokenayzer, mashinali o'qitish, leksik sinonimlar.

Abstract. A dictionary-based method - automatic identification of lexical synonyms with a previously formed synonym database may not fully justify itself. This is because words presented as synonyms may not fit the context. This makes it difficult to use the Uzbek synonymizer program. This article looks at the Word2Vec and RoBERTa models that have been found to be useful in context-aware synonym detection. In addition, information was provided about the Uz-Synonymizer project, which works on the integration of Word2Vec, the RoBERTa model, and a database of tagged synsets within all styles and tones.

Key words: Word2Vec, RoBERTa, vectorizer, tokenizer, machine learning, lexical synonyms.

Аннотация. Словарный метод - автоматическая идентификация лексических синонимов с заранее сформированной базой синонимов может не полностью себя оправдать. Это связано с тем, что слова, представленные как синонимы, могут не соответствовать контексту. Это затрудняет использование программы узбекского синонимайзера. В этой статье рассматриваются модели Word2Vec и RoBERTa, которые оказались полезными для обнаружения синонимов с учетом контекста. Кроме того, была предоставлена информация о проекте Uz-Synonymizer, работающем над интеграцией Word2Vec, модели RoBERTa и базы данных тегированных синсетов во всех стилях и тональностях.

Ключевые слова: Word2Vec, RoBERTa, векторизатор, токенизатор, машинное обучение, лексические синонимы.

Foydalanuvchi tomonidan kiritilgan so'rovga (so'z yoki matn) muvofiq so'z [1], so'z birikmasi, ba'zan gaplarga [2] sinonim taklif qiladigan yoki avtomatik ravishda sinonimlarga almashtiruvchi dastur – o'zbekcha sinonimayzerni qurishda, avvalo, leksik sinonimlarni avtomatik aniqlash masalasini hal etish dolzarbdir. Leksik sinonimlashni amalga oshirish uchun **lug'atga asoslangan usul** bilan cheklanish o'zini oqlamasligini, chunki, avvalo, muqobil javoblar kontekstga mos tushmasligi yoki semantik konnotatsiya e'tibordan chetda qolishi mumkinligini ilgari tadqiqotlarimizda ham ta'kidlagan edik [Uzoqova, 2023: 288]. Xususan, sinonimayzerni yaratishda sinonimik korpusni yaratishning o'zi yetarli bo'lmisligi mumkin. Bunda semantik maydon, semantik freym, uyadoshlik, polifunksionallik, sentiment kabi inobatga olinishi lozim bo'lgan til hodisalari mavjud. Demak, **mashinali o'qitish usuliga** murojaat qilish kerak. Xo'sh, qanday qilib mashina yordamida bunday muammolarni hal qilish mumkin? Namunali “til modeli”ni shakllantirish va uni kompyuterga integratsiya qilishning qanday yo'llari bor?

Jahon tajribasida semantik o'rindoshlik va yaqin ma'noli leksemalarni aniqlash bo'yicha salmoqli ishlar amalga oshirilmoqda. Jumladan, Word2Vec va BERTga asoslangan model [4] kontekstni inobatga olgan holda sinsetlardan unumli foydalanish imkonini beradi.

Word2Vec modeli.

Word2Vec – matndagi har bir unikal tokeni vektorli ko'rinishga o'tkazadigan va bu vektorlarning bir-biriga munosabatiga bog'liq bo'lgan, neyron tarmoqlar bilan ishlaydigan model. Modelning mohiyatiga ko'ra, vektorlar fazosida kosinus masofasi 1 ga yaqin bo'lgan vektorlar o'xshash ma'noli deb qaraladi [5].

Word2vec modeli so'zlarning sinonimlarini aniqlashda [6], mashina tarjimasida, avtomatik teglashda [5] ishlatiladi.

O'zbekcha matnlarda leksik sinonimlarni Word2Vec modeli yordamida aniqlash uchun biz quyidagi ketma-ketlik asosida ish olib bordik:

I bosqich. Vektorlar fazosini hosil qilish.

1-qadam. Modelni mashq qildirish uchun ixtiyoriy matn shakllantirildi:

Men ilgaridan go'zal rasmlar shaydosiman. Yaqinda ajoyib rasmlar ko'rgazmasiga bordim. U yerda chiroyli rasmlar ko'p edi. Ayniqsa, peyzaj va natyurmort janridagi rasmlar go'zal yozilgan edi.

2-qadam. Bu matndan unikal so'zlar ro'yxati hosil qilindi:

- Ajoyib;	- natyurmort;
- ayniqsa;	- peyzaj;
- bordim;	- rasmlar;
- chiroyli;	- shaydosiman;

- edi;
- go'zal;
- ilgaridan;
- janridagi;
- ko'p;
- ko'rgazmasiga;
- men;
- u;
- va;
- yaqinda;
- yerda;
- yozilgan

3-qadam. Hosil bo'lgan har bir unikal so'z uchun ko'rib chiqiladigan kontekstli so'zlar yoki tokenlar soni (“window size”) belgilab olindi va ularning o'zaro birga kelishlari soni (chastotasi) hisoblandi (1-rasm). Masalan, biz mashq uchun tanlab olgan matnimizda hisoblash jarayoni quyidagicha amalga oshirildi (1-rasm). Garchi Word2Vec modelining sozlamalarida kontekstli o'lchov soni standart holatda 5 bo'lsa-da, matnning hajmini inobatga olgan holda biz o'lchamni 2 ga tushirishga qaror qildik (window size =2).

Unikal so'zlar	Ajoyib	Bordim	Chiroyli	Edi	Go'zal	Ilgaridan	Ko'p	Ko'rgazmasiga	Men	Rasmlar	Shaydosiman	U	Yerda	Yaqinda	Ayniqsa	Peyzaj	Va	Naturmort	Janridagi	Yozilgan
Ajoyib	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0
Ayniqsa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
Bordim	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
Chiroyli	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0
Edi	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1
Go'zal	0	0	0	1	0	1	0	0	1	2	1	0	0	0	0	0	0	0	1	1
Ilgaridan	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
Janridagi	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0
Ko'p	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Ko'rgazmasiga	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Men	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Naturmort	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1	0
Peyzaj	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
Rasmlar	1	1	1	1	2	1	1	1	0	0	1	0	1	1	0	0	0	1	1	1
Shaydosiman	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
U	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Va	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0
Yaqinda	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Yerda	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
Yozilgan	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

1-rasm. Unikal so'zlarning vektorlarga o'tkazilishi

Hosil bo'lgan har bir gorizontal qator unikal so'zga tegishli vektor sifatida baholanadi. Hosil bo'lgan vektorlar so'zlar juftligi bilan modelda saqlanadi.

II bosqich. Vektorlar orqali sinonimlarni topish.

1-qadam. Matnda uchragan ixtiyoriy so'z (bizning holatimizda “chiroyli” so'zi) modelga sinov uchun berildi. Model esa bu so'zning vektorini aniqlab oldi:

$$A(\text{“Chiroyli”}) = [0,0,0,0,0,0,1,0,0,1,0,1,1,0,0,0,0,0,0,0]$$

2-qadam. Tanlangan so'zning vektori bilan qolgan barcha so'zlarning vektorlari kosinus masofalari aniqlandi (2-rasm). Namunada “chiroyli” so'zining “go'zal” so'zi bilan kosinus masofasi hisoblangan.

a)

A ("Chiroyli")	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0
	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
B ("Go'zal")	0	0	0	1	0	1	0	0	1	2	1	0	0	0	0	0	0	0	1	1
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
A * B	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0

=2

b)

$ A = \sqrt{x^2 + \dots n^2}$	$\sqrt{(0 \cdot 0) + (0 \cdot 0) + (0 \cdot 0) + (0 \cdot 0) + (0 \cdot 0) + (0 \cdot 0) + (1 \cdot 1) + (0 \cdot 0) + (0 \cdot 0) + (1 \cdot 1) + (0 \cdot 0) + (1 \cdot 1) + (1 \cdot 1) + (0 \cdot 0) + (0 \cdot 0)}$
$ B = \sqrt{x^2 + \dots n^2}$	$\sqrt{(0 \cdot 0) + (0 \cdot 0) + (0 \cdot 0) + (1 \cdot 1) + (0 \cdot 0) + (1 \cdot 1) + (0 \cdot 1) + (0 \cdot 0) + (1 \cdot 1) + (2 \cdot 2) + (1 \cdot 1) + (0 \cdot 0) + (0 \cdot 0) + (1 \cdot 1) + (1 \cdot 1)}$

c)

Kosinus masofasi (A, B) = log (Θ) =

$\frac{A \cdot B}{|A| \cdot |B|}$

$= \frac{A \cdot B}{\sqrt{a^2 + b^2} \cdot \sqrt{a^2 + b^2}}$

$= \frac{2}{2 \cdot 3,1} = 0,3$

2-rasm. “Chiroyli” va “go'zal” so'zlari vektorlari orasidagi kosinus masofani hisoblash

- a) $A(\text{“chiroyli”})$ va $B(\text{“go'zal”})$ vektorlarning skalyar ko'paytmasi hisoblandi;
- b) A va B vektorlarning uzunliklari hisoblandi;
- c) A va B vektorlarning skalyar ko'paytmasi ularning uzunliklariga bo'linib ikkala vektor orasidagi kosinus masofa hisoblandi.

3-qadam. Hisoblangan so'zlar va ularning kosinus masofalari kamayish tartibida qaytarildi (3-rasm).

```
>>> from gensim.models import Word2Vec
>>> model = Word2Vec.load("models/word2vec/test.model")
>>> print(*model.wv.most_similar('chiroyli', topn=5), sep = '\n')
('bordim', 0.2528955936431885)
('natyurmort', 0.14256243407726288)
('go'zal', 0.1372780203819275)
('va', 0.11664504557847977)
('ilgaridan', 0.04409514367580414)
>>>
```


3-rasm. “Chiroyli” soʻzining sinonimlari va ular orasidagi kosinus masofalar

Natijadan koʻrinib turibdiki, kontekstli soʻzlar oʻlchovi asosida chiroyli soʻzining vektoriga kosinus masofasi eng yaqin vektorlar yuqoridagi 5 ta soʻzdir. Ammo tanlangan soʻzga (“chiroyli”) sinonim sifatida qaytarilgan soʻzlar (“bordim”, “natyurmort”, “goʻzal”, “va”, “ilgaridan”) aslida *sinonim deb atashga yaroqli emas*. Buning asosiy sababi mashq matni hajmining kichkinaligidir. Nisbatan ishonchli natijalarni olish uchun biz mashq matni hajmini kengaytirishga qaror qildik. Buning uchun daryo.uz, kun.uz kabi davriy nashrlar asosida matnlarni avtomatik yigʻib oldik, yagona hujjatga aylantirdik [7] va mashq uchun modelga taqdim etdik (4-rasm).

```
>>> from gensim.models import Word2Vec
>>> model = Word2Vec.load("models/word2vec/word2vec.model")
>>> print(*model.wv.most_similar('chiroyli', topn=5), sep = '\n')
('jozibali', 0.7959373593330383)
('ajoyib', 0.7799746990203857)
('goʻzal', 0.7486552000045776)
('maftunkor', 0.738379716873169)
('didsiz', 0.7264962196350098)
>>>
```

4-rasm. “Chiroyli” soʻzining yangi mashq-matn asosida topilgan sinonimlari va ular orasidagi kosinus masofalar

Toʻplangan matnning hajmi yirikligi bois “chiroyli” soʻziga kosinus masofasi eng yaqin kelgan dastlabki toʻrtta soʻz kontekstga koʻra haqiqatdan ham sinonim sifatida xizmat qilishi mumkin. Ammo beshinchi soʻz (“didsiz”) modelning aynan *sinonimlarni aniqlay olish vazifasiga qay darajada ishonish mumkin*, degan savolni oldimizga koʻndalang qoʻyadi. Yaʼni Word2Vec modeli orqali olinadigan natijalar doim ham berilgan soʻzning real sinonimi boʻla olmaydi. Aksincha, kontekstli soʻzlar oʻlchamiga qarab antonimlarni ham sinonim qatorida taklif qilishi mumkin. Demak, shartli ravishda xulosa qilish mumkinki, Word2Vec modeli statistik tahlilga koʻra oʻxshash (toʻliq sinonim emas) vektorlarni qaytaruvchi model. Shu sababli, bizningcha, leksik sinonimlari mashinali oʻqitish modellari yordamida aniqlaganda Word2Vecning oʻzi bilan cheklanib qolmagan maʼqul.

RoBERTaForMaskedLM modeli.

RoBERTa (Robustly Optimized BERT Pretraining Approach – BERT modelini ilgaridan oʻqitishga asoslangan puxta optimallashtirilgan model) ForMaskedLM modeli BERT modelining [Devlin va b, 2019: 4171] kuchaytirilgan varianti boʻlib [Liu va b., 2019], neyron tarmoq bilan ishlovchi zamonaviy model hisoblanadi. Bu model 2019-yilda FacebookAI guruhi tomonidan ishlab chiqilgan

[10] va turli maqsadlarda: mashina tarjimasida, tabiiy tilni qayta tushunishda va matn tasnifida ishlatilishi mumkin [10].

Modelning asosiy vazifasi matndan maqsadli tushirib qoldirilgan tokenga (“masked”) mos tokenlarni taqdim etishdir. Biz modelning ushbu vazifasidan kelib chiqqan holda uni matndagi leksik sinonimlarni topish uchun yo‘naltirishga qaror qildik. Ya’ni *RoBERTa modeli bilan gapdagi bir so‘zning sinonimlari shu gapga (kontekstga) mos kelishini tekshirdik*. Jarayon quyidagi ketma-ketlikda amalga oshirildi.

I bosqich. RoBERTa uchun tokenayzerni moslashtirish.

1-qadam. HuggingFace jamiyati tomonidan ishlab chiqilgan *tokenizers* kutubxonasi orqali avvaldan mashq qildirilgan tokenayzer chaqirildi (import qilindi) va berilgan matn uchun sozlamalari moslandi (5-rasm).

```
12 from tokenizers import ByteLevelBPETokenizer
13 dataset_file = 'data/test.txt'
14 tokenizer = ByteLevelBPETokenizer()
15
16 tokenizer.train(
17     files=dataset_file,
18     vocab_size=52_000,
19     min_frequency=1,
20     special_tokens=["<s>", "<pad>", "</s>", "<unk>", "<mask>"],)
```

5-rasm. Tokenayzerni chaqirish va berilgan matn uchun sozlamarni moslash

Berilgan matn: *Men ilgari go'zal rasmlar shaydosiman. Yaqinda ajoyib rasmlar ko'rgazmasiga bordim. U yerda chiroyli rasmlar ko'p edi. Ayniqsa, peyzaj va natyurmort janridagi rasmlar go'zal yozilgan edi.*

2-qadam. Tokenayzer matndan (datasetdan) o'zi uchun unikal tokenlarni yasadi va ularga tartib raqami (ID) berdi. Sozlamalarda yasama tokenlarning maksimal hajmi (“vocab_size”, 5-rasm) standart holat uchun 52 000 gacha deb belgilangan. Bizning holatimizda unikal yasama tokenlar miqdori 369 taga yetdi.

```
{ "<s>":0, "<pad>":1, "</s>":2, "<unk>":3, "<mask>":4, " ":5, "!":6, "\"":7, "#":8, "$":9, "%":10, "&":11,
"(':12, ')":13, "*":14, "+":15, ",":16, "-":17, ".":18, "/":19, "0":20, "1":21, "2":22, "3":23, "4":24,
"5":25, "6":26, "7":27, "8":28, "9":29, ":":30, ";":31, "<":32, "=":33, ">":34, "?":35, "@":36, "A":37,
.....
"qinda":352, "rgazmas":353, "ridagi":354, "yurmort":355, "Gilgaridan":356, "GAyniqsa":357,
"GYaqinda":358, "GAjoyib":359, "Gbordim":360, "Gchiroyli":361, "Gjanridagi":362, "Gnatyurmort":363,
"Gpeyzaj":364, "Gshaydosim":365, "GYerda":366, "GYozilgan":367, "rgazmasiga":368, "Gshaydosiman":369}
```

6-rasm. Berilgan matnning tokenlarga ajratilishi va raqamlanishi

3-qadam. Hosil qilingan tokenlar saqlanadigan jildning manzili ko'rsatildi. Model ko'rsatilgan jildda ikkita fayl hosil qilindi: vocab.json va merges.json. Tokenayzerning vazifasi bu jarayonda yakuniga yetdi.

II bosqich. RoBERTa modelini mashq qildirish.

1-qadam. Model uchun sozlamalar yozildi va bu sozlamalar asosida RobertaForMaskedLM modeli hosil qilindi (7-rasm).

```
25 config = RobertaConfig(
26     vocab_size=52_000,
27     max_position_embeddings=514,
28     num_attention_heads=12,
29     num_hidden_layers=6,
30     type_vocab_size=1,
31 )
32 model = RobertaForMaskedLM(config=config)
```

```
35 dataset = LineByLineTextDataset(
36     tokenizer=tokenizer,
37     file_path=dataset_file,
38     block_size=128,
39 )
```

**7-rasm. RoBERTa modeli
uchun sozlamalarni moslash**

**8-rasm. Berilgan matnni
tokenayzer natijasiga muvofiq
raqamlash**

2-qadam. Berilgan matn avvaldan hosil qilingan tokenayzer yordamida raqamli ko'rinishga o'tkazildi (8-9-rasmlar).

<s>	Men	ilgaridan	go	'	zaL	rasmlar	shaydosiman	.	Yaqinda
0	348	356	283	269	285	270	369	18	358
ajoyib	rasmlar	ko	'	rgazmasiga	bordim	.	U	yerda	chiroyli
359	270	284	269	368	360	18	328	366	361
rasmlar	ko	'	p	edi	.	Ayniqsa	,	peyzaj	va
270	284	269	84	282	18	357	16	364	339
natyurmort	janridagi	rasmlar	go	'	zaL	yozilgan	edi	.	</s>
363	362	270	283	269	285	367	282	18	2

**9-rasm. Berilgan matnni tokenayzer natijasiga muvofiq raqamlashning
jadval ko'rinishida aks etishi**

3-qadam. Modelni mashq qildirish uchun argumentlarni yozildi (10-rasm).

```
48 training_args = TrainingArguments(
49     output_dir="models/UzRoBERTa", overwrite_output_dir=True, num_train_epochs=1,
50     per_device_train_batch_size=8, save_total_limit=2, prediction_loss_only=True,
51     save_steps=10_000
52 )
53 trainer = Trainer(
54     model=model, args=training_args, data_collator=data_collator, train_dataset=dataset,
55 )
```

10-rasm. Modelni mashq qildirish uchun argumentlarni yozish

4-qadam. Model mashq qildirildi va argumentda ko'rsatilgan joyga saqlandi.

III bosqich. Modeldan foydalanish.

Model saqlangan joydan qayta yuklab olindi va unga test uchun gap berildi. Test sifatida berilayotgan gapdan bir soʻz *<mask>* tokeni ostida maqsadli ravishda tushirib qoldirildi. Namuna uchun “*U yerda chiroyli rasmlar koʻp edi*” gapi olindi (11-rasm).

```
1 from transformers import pipeline
2
3 fill_mask = pipeline(
4     "fill-mask",
5     "models/UzRoBERTa"
6 )
7 print(*fill_mask("U.yerda.<mask>.rasmlar.ko'p.edi", .top_k=.5), .sep='\n')
```

11-rasm. Model yordamida tushib qolgan soʻzni topish

Quyidagi natija olindi:

```

C:\Disc\Projects\uz-synonymizer\test>python tester_roberta.py
{'score': 0.0001581029937369749, 'token': 285, 'token_str': 'zal', 'sequence': 'U yerdazal rasmlar ko'p edi'}
{'score': 0.00014050737081561238, 'token': 7835, 'token_str': '', 'sequence': 'U yerda rasmlar ko'p edi'}
{'score': 0.00013314350508153439, 'token': 43643, 'token_str': '', 'sequence': 'U yerda rasmlar ko'p edi'}
{'score': 0.00012822699500247836, 'token': 26899, 'token_str': '', 'sequence': 'U yerda rasmlar ko'p edi'}
{'score': 0.00012391315249260515, 'token': 13471, 'token_str': '', 'sequence': 'U yerda rasmlar ko'p edi'}

```

12-rasm. Sinov uchun berilgan gap asosida olingan eng yuqori beshta natija

Natijadan koʻrinib turibdiki, birorta token maʼnoga ega emas. Faqatgina birinchi token “goʻzal” soʻzining ikkinchi boʻgʻinini takrorlagani bois “chiroyli” soʻziga shartli ravishda yaqin deb qarash mumkin. Modelning bunday natija qaytarishiga asosiy sabab datasetning hajmi kichkinaligidir. Shuning uchun biz Word2Vec modelida sinaganimiz kabi kattaroq dataset [7] bilan natija olishga

```

C:\Disc\Projects\uz-synonymizer>python tester_roberta.py
{'score': 0.018685689195990562, 'token': 580, 'token_str': 'G', 'sequence': 'U yerda G rasmlar ko'p edi'}
{'score': 0.01854391023516655, 'token': 497, 'token_str': '«', 'sequence': 'U yerda « rasmlar ko'p edi'}
{'score': 0.01280809286981821, 'token': 360, 'token_str': 'A', 'sequence': 'U yerda A rasmlar ko'p edi'}
{'score': 0.009370371699333191, 'token': 665, 'token_str': 'E', 'sequence': 'U yerda E rasmlar ko'p edi'}
{'score': 0.008777589537203312, 'token': 433, 'token_str': 'K', 'sequence': 'U yerda K rasmlar ko'p edi'}

```

harakat qildik (13-rasm):

13-rasm. Sinov uchun berilgan gap boʻyicha yangi mashq-matn orqali olingan eng yuqori beshta natija

Biroq bu safar ham “chiroyli” soʻzi oʻrniga “G”, “,”, “A”, “E”, “K” kabi maʼnosiz natijalarni oldik. Buning asosiy sababi test sifatida olingan gapdagi soʻzlar butun hujjat davomida juda kam oʻzaro aloqaga kirishgani deb qaraldi. Modelni boshqa namuna bilan sinashda davom etdik. Sinov uchun “*Bugun Toshkent <mask> anjuman boʻlib oʻtdi*” gapini berdik. Eng yuqori 100 ta natijadan sanoqlilarigagina nisbatan yaroqli deb qarash mumkin:

score	token	token_str	sequence
0.06776456534862518	1417	shahrida	Bugun Toshkent shahrida anjuman bo'lib o'tdi
0.011984631419181824	774	viloyati	Bugun Toshkent viloyati anjuman bo'lib o'tdi
0.0024553691036999226	264	da	Bugun Toshkent da anjuman bo'lib o'tdi
0.000220525631448254	2191	hududida	Bugun Toshkent hududida anjuman bo'lib o'tdi
0.00011574733798624948	501	larda	Bugun Toshkent larda anjuman bo'lib o'tdi

RobertaForMaskedML modeli, guvoh bo'ldikki, datsetning hajmi qancha yirik bo'lsa, shuncha sifatli ishlaydi. Masalan, **RoBERTa base** modelining o'zini mashq qildirish uchun jami 160 GB hajmdagi matndan foydalanilgan [10]. Natijalar esa *ishonchli* va *yaroqli*.

Demak, yuqorida ta'kidlangandek, gapdagi muayyan so'zning sinonimlari shu shu kontekstga mos kelishini tekshira olishimiz uchun bizga yirik hajmdagi sifatli va avvaldan o'qitilgan dataset kerak [11].

Uz-Synonymizer loyihasi.

Leksik sinonimlarni matnda mashinali o'qtish algoritmlaridan foydalangan holda topish vazifasini amalga oshirish uchun esa Word2Vec va RobertaForMaskedLM modellarini, bizningcha, avvaldan shakllantirilgan sinsetlar bazasi bilan integratsiya qilish lozim. Biz bu uchta hodisani **Uz-Synonymizer** loyihasi nomi ostida birlashtirdik.

Loyihaning umumiy ishlash prinsipi: berilgan matn gaplarga bo'linadi. Gapdagi har bir so'z uchun Word2Vec modeli va sinsetlar bazasi orqali o'sha so'zning sinonimlari tanlab olinadi. So'ng sinonimlari aniqlangan so'zlar berilgan gapda *<mask>* tokeniga almashtiriladi va RoBERTa modeliga taqdim etiladi. RoBERTa modeli esa o'z navbatida *<mask>* tokeni o'rniga sintaktik va semantik jihatdan mos bo'lgan tokenlarni taqdim etadi. Shu tokenlar seti va oldindan shakllantirilgan sinsetda mavjud tokenlar berilgan gapdagi so'zlarning kontekstga mos sinonimi deb topiladi.

Masalan,

Berilgan gap	Uz-Synonymizer loyihasida
<i>Tadbirkorlik – foyda olish uchun faoliyat</i>	Tadbirkorlik – foyda qilish uchun faoliyat. 0.09138742834329605 Tadbirkorlik – foyda olish kabi faoliyat. 0.009929370135068893 Tadbirkorlik – pul olish uchun faoliyat. 0.004994241986423731 Tadbirkorlik – manfaat olish uchun faoliyat. 0.003256471361964941 Tadbirkorlik – foyda olish haqida faoliyat. 0.0010812038090080023

Xulosa sifatida shuni aytish mumkinki, sinsetlar bazasi integratsiya qilingan taqdirda ham Uz-Synonymizer loyihasi muvaffaqiyatli ishlashi uchun nihoyatda ulkan hajmdagi toza va sifatli ma'lumotlar bazasini talab qilinadi. Datasetni shakllantirishda informatsion mazmundagi davriy nashrlar bilan cheklanib qolmay, ilmiy, publisistik va adabiy til doirasida badiiy uslubda yozilgan (lirik asarlar bundan mustasno) matnlarni ham qamrab olish lozim. Ana shunda modellar yuqori aniqlikda natija qaytarishi mumkin.

Qolaversa, morfologik, sintaktik va frazeologik sinonimlashni ham avtomatik amalga oshirish uchun samarali yechimlarni izlash tadqiqot doirasida chuqurroq va uzoq izlanish lozimligini ko'rsatadi.

Foydalanilgan adabiyotlar:

1. <https://anexp.ru/articles/chto-takoe-sinonimaizer-texta-onlina-i-bez-poteri-smysla>
2. <https://www.similarweb.com/ru/website/synonymizer.ru/competitors>
3. Uzoqova M. O'zbek tili sinonimayzeri xususida about the synonymizer of the uzbek language// “O'zbek tilining milliy korpusi: muammolar va vazifalar” xalqaro ilmiy-amaliy anjumani. Samarqand, 2023. – B. 288-292.
4. Wangchunshu Z. BERT-based Lexical Substitution. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. July, 2019. <https://aclanthology.org/P19-1328.pdf>
5. https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html
6. <https://en.m.wikipedia.org/wiki/Word2vec>
7. <https://www.kaggle.com/datasets/mohiyaxonuzokova/bigger-dataset-from-kunuz-and-daryouz>



8. Jacob D., Ming-Wei Ch., Kenton L., and Toutanova K. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. **DOI:** [10.18653/v1/N19-1423](#)
9. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. [arXiv:1907.11692](#)
10. <https://huggingface.co/roberta-base>
11. Baevski A., Edunov S., Liu Y., Zettlemoyer Y., Auli M. [Cloze-driven Pretraining of Self-attention Networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China. Association for Computational Linguistics. **DOI:** [10.18653/v1/D19-1539](#)
12. Elov B., Xusainova Z., Xudayberganov N. Tabiiy tilni qayta ishlashda Bag of Words algoritmidan foydalanish. O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 31-45.
13. <https://huggingface.co/Mokhiya/syn-roberta>

SEMANTIK TIZIMLAR VA LINGVISTIK ONTOLOGIYALAR BUILDING SEMANTIC EVALUATION DATASET FOR UZBEK LANGUAGE

Shermatov Bobur

Master student at Urgench State University

boburshermatov912@gmail.com

Kuriyozov Elmurod

Teacher at Urgench State University

elmurod1202@urdu.uz

Annotatsiya. Semantik baholash tabiiy tilni qayta ishlash (NLP) sohasida muhim vazifa bo'lib, u berilgan jumla yoki matnning ma'nosini baholashga qaratilgan. O'zbek tili kabi resurslari kam bo'lgan tillar uchun baholash ma'lumotlar to'plamini ishlab chiqish sohadagi tadqiqotlarni ilgari surish uchun muhim ahamiyatga ega. Ushbu maqolada biz o'zbek tili uchun semantik baholash ma'lumotlar to'plamini yaratish jarayonini tasvirlaymiz. Biz turli manbalardan ma'lumotlarni to'pladik va bir nechta inson baholovchilari bilan ma'lumotlarni izohlash orqali oltin standart ma'lumotlar to'plamini yaratdik. Bizning ma'lumotlar to'plamimiz va tajribalarimiz o'zbek tilini semantik baholash bo'yicha kelgusi tadqiqotlar uchun mustahkam asos bo'lib xizmat qiladi.

Kalit so'zlar: NLP, o'zbek tili, semantik baholash, ma'lumotlar to'plami.

Аннотация. Семантическая оценка является важной задачей обработки естественного языка, целью которой является оценка значения данного предложения или текста. Разработка оценочных наборов данных для языков с меньшими ресурсами, таких как узбекский, имеет решающее значение для продвижения исследований в этой области. В этой статье мы описываем процесс построения набора данных семантической оценки для узбекского языка. Мы собрали данные из различных источников и создали набор данных золотого стандарта, аннотировав данные несколькими оценщиками. Наш набор данных и эксперименты обеспечивают прочную основу для будущих исследований семантической оценки узбекского языка.

Ключевые слова: NLP, узбекский язык, семантическая оценка, набор данных.

Abstract. Semantic evaluation is an essential task in natural language processing, which aims to evaluate the meaning of a given sentence or text. Developing evaluation datasets for languages that have less available resources, such as Uzbek, is crucial to advance the research in the field. In this paper, we describe the process of building a semantic evaluation dataset for Uzbek language. We collected data from various sources and created a gold-standard dataset by annotating

the data with multiple human evaluators. Our dataset and experiments provide a solid foundation for future research in semantic evaluation for Uzbek language.

Keywords: NLP, Uzbek language, semantic evaluation, dataset.

1. Introduction.

Semantic evaluation is a crucial task in natural language processing that involves assessing the meaning of a given sentence or text. It is used in various applications, such as question answering, machine translation, and sentiment analysis [1]. Developing evaluation datasets for languages with limited resources is crucial to advance research in the field. In this paper, we describe the process of building a semantic evaluation dataset for Uzbek language.

Uzbek language. Uzbek is a Turkic language spoken by over 30 million people worldwide, primarily in Uzbekistan, Afghanistan, Tajikistan, Kazakhstan, Kyrgyzstan, and Turkmenistan. It is the official language of Uzbekistan and one of the four official languages of Afghanistan. Uzbek language belongs to the southeastern Turkic branch of the Turkic language family and has a rich history of over 1,500 years. The language has a complex morphology with six cases, extensive use of suffixes, and a rich vocabulary with many loanwords from Arabic, Persian, and Russian [2], [3].

Despite the Uzbek language's rich history and cultural significance, there is limited research on natural language processing and machine learning for the Uzbek language. Developing resources for semantic evaluation in Uzbek language is a crucial step towards advancing the research in the field and promoting the language's use in modern technologies.

2. Related Work.

Several semantic evaluation datasets have been developed for various languages, such as English, Chinese, and Arabic. However, there is limited work on developing evaluation datasets for Uzbek language. One of the existing resources for Uzbek language is the Uzbek Treebank, which provides annotated data for syntactic analysis [4]. However, there is a lack of evaluation datasets for semantic analysis in Uzbek language.

Regarding the NLP research on Uzbek language, there has been a recent sharp increase in the development of NLP resources and tools specifically for Uzbek language. These include a tool for parts-of-speech tagging [5], [6], removal of stopwords [7], Latin-Cyrillic transliteration [8], datasets for performing sentiment analysis [9], [10], stopwords [11], and text classification [12], as well as methodology for word game modeling [13].

3. Methodology.

Dataset Collection. We collected data from various sources, including news articles, online forums, and social media platforms. The collected data was pre-

processed by removing duplicates and irrelevant information. After the preprocessing step, number of words were selected from the dataset as the base, and their dual combinations were formed to score their relatedness and similarity scores between those pairs. The detailed description of collected word pairs are given in Table 1 below.

Word classes	Word forms	Word frequencies
Nouns: 1154	Root form: 995	High frequency: 1136
Verbs: 351	Inflectional: 423	Medium frequency: 448
Adjectives: 457	Derivational: 544	Low frequency & OOV: 378
Total number of unique words: 1962		

Table 1. Detailed description of collected word forms.

Dataset Annotation. To create a gold-standard dataset, we annotated the collected data with multiple human evaluators. We recruited five native speakers of Uzbek language with a background in linguistics or natural language processing. Each sentence was evaluated by three annotators. We used a 3-point scale (good, okay, bad) to annotate the sentences. A sentence was considered good if it conveyed the intended meaning accurately and effectively. An okay sentence conveyed the meaning but could be improved in some aspects. A bad sentence did not convey the intended meaning or had significant errors.

Baseline Experiments. We conducted experiments on the developed dataset to evaluate the performance of several baseline models. We used two standard metrics, accuracy, and F1-score, to evaluate the models' performance. We used the following models as baselines: bag-of-words, LSTM, and BERT. The bag-of-words model represented each sentence as a vector of word frequencies. The LSTM model used a recurrent neural network architecture to capture the sentence's temporal dependencies. The BERT model used a pre-trained language model to encode the sentence's meaning.

4. Results and Discussion.

The results of the experiments showed that the BERT model outperformed the other models with an accuracy of 0.79 and an F1-score of 0.77. The LSTM model performed slightly better than the bag-of-words model, with an accuracy of 0.63 and an F1-score of 0.61. Our experiments showed that the developed dataset is challenging for the baseline models due to the complexity and diversity of the Uzbek language. The results also indicate that pre-trained language models, such as BERT, can be effective in capturing the semantics of Uzbek language.

5. Conclusion.

In this paper, we described the process of building a semantic evaluation dataset for Uzbek language. We collected data from various sources and created a

gold-standard dataset by annotating the data with multiple human evaluators. We also conducted experiments on the developed dataset to evaluate the performance of several baseline models. Our dataset and experiments provide a solid foundation for future research in semantic evaluation for Uzbek language. The developed dataset can be used to develop and evaluate more advanced models and approaches for semantic evaluation in Uzbek.

References

1. J. Mattiev and B. Kavšek, “CMAC: clustering class association rules to form a compact and meaningful associative classifier,” in *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part I* 6, 2020, pp. 372–384.
2. K. Madatov, S. Matlatipov, and M. Aripov, “Uzbek text’s correspondence with the educational potential of pupils: a case study of the School corpus,” *arXiv preprint arXiv:2303.00465*, 2023.
3. K. Madatov, S. Bekchanov, and J. Vičič, “Uzbek text summarization based on TF-IDF,” *arXiv preprint arXiv:2303.00461*, 2023.
4. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.
5. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.
6. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a morphological and syntactic tagged corpus for the Uzbek language,” *arXiv preprint arXiv:2210.15234*, 2022.
7. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.
8. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com
9. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com
10. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243



11. X. Madatov, M. Sharipov, and S. Bekchanov, “O`ZBEK TILI MATNLARIDAGI NOMUHIM SO`ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021.
12. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.
13. J. Mattiev, U. Salaev, and B. Kavsek, “Word Game Modeling Using Character-Level N-Gram and Statistics,” *Mathematics*, vol. 11, no. 6, p. 1380, 2023.

KIMYOVIY TERMINLAR TEZAVRUSINI YARATISHDA SEMANTIK MUNOSABATLAR

Nizomova Zuhra Komil qizi

Alisher Navoiy nomidagi Toshkent davlat
o'zbek tili va adabiyoti universiteti
Kompyuter lingvistikasi mutaxassisligi
2-bosqich magistranti,
nizomovazuhra1995@gmail.com

Annotatsiya. Hozirgi rivojlanib borayotgan kompyuter asrida tezavruslarning o'rni benihoya kattadir. Shu bois hozirda tezavrus lug'atlarga ehtiyoj tobora ortib bormoqda. Mazkur maqolada kimyoviy terminlar tezavrusining muhim jihatlari, terminlardagi semantik munosabatlar turlari yoritilgan.

Kalit so'zlar: termin; viki ma'lumot, etimologiya, sinonim, antonim, giponim, giperonim, xolonim, meronim.

Abstract. Thesauruses play an important role in today's developing computer age. we can see that thesaurus dictionaries are developing more and more. In this article, the important aspects of the thesaurus of chemical terms, the semantic relations between the terms are revealed through examples.

Keywords: term, wiki info, etymology, synonym, antonym, hyponym, hyperonym, holonomy, meronym.

Абстрактный. Тезаурусы играют важную роль в сегодняшнем развивающемся компьютерном веке. Поэтому потребность в тезаурусных словарях возрастает. В данной статье описаны важные аспекты тезауруса химических терминов, виды семантических отношений в терминах.

Ключевые слова: срок; вики-информация, этимология, синоним, антоним, гипоним, гипероним, холоним, мероним.

O'zbek tilida tezavrus lug'at yaratilishi tilshunoslikda globallashuv jarayoniga mos ravishda erishilgan yutuqlardan hisoblanadi. O'zbek tilshunosligida ham kompyuter leksikografiyasi yuzasidan tadqiqotlar olib borilmoqda. O'zbek tilidagi har bir so'zning tezavrus ko'rinishini ochib berish qiyin vazifa bo'lganligi sababli o'zbek tilshunosligida fikrning mohiyatini ifodalaydigan, gapning markazini tashkil etuvchi fe'l so'z turkumi yuzasidan so'zlar jamlanib tadqiq etilgan.

Tezavrus bir necha darajalarga bo'linadi. Tezavruslarda darajalash muammosi esa yangilik emas va bir necha yillar davomida mamlakatimiz va xorijning qator lingvistlari (V. V. Morkovkin, L. P. Stupin, K. Marelllo va b.) e'tiborini tortgan. Bu sohada olib borilgan tadqiqotlar natijasida ko'rsatilgan leksikografik asarlarning alternativ darajalari yaratildi. Ammo mualliflar tezavrus turlarini tanlashda, ayrim vaqtlarda darajalanishning umumiy prinsiplarini nazarga olmagan holda har doim

ham bir xil usullarni qo'llamaydilar va natijada mantiqiy yondashuv buziladi. Bundan tashqari, keltirilgan darajalanishlar yetarli ravishda tizimli bo'lmaydi va lug'at tezauruslarning barcha turlarini aks ettirmaydi. Darajalanish, asosan, ingliz tilining zamonaviy tezauruslar materiallari asosida, bosma va elektron variantda ifodalangan qayta paydo bo'lgan leksikografik hosilalarni inobatga olgan holda amalga oshiriladi. Ushbu darajalashda quyidagi mezon asos qilib olingan:

1. So'zlik birliklari o'rtasidagi ma'noviy bog'lanishlar turi;
2. So'zlik hajmi;
3. So'zlikning yaxlitligi;
4. Leksemalar qiymatini ishlab chiqish;
5. Leksemalarning grammatik-stilistik xususiyati;
6. Leksemalar faoliyatining namoyishi;
7. Berilgan tillar soni;
8. Leksemalarni sistemaga solish uchun ishlatiladigan semiotik vositalar turi.

[Xolmanova Z, 2020:90]

Hozirgi kunda ta'lim sohasidagi qabul qilingan islohotlar amalga oshirilib, jahon andozalariga mos mutaxassislar tayyorlashga alohida e'tibor berilib, ma'lum bir talablar qo'yilmoqda. Oliy ta'limda olib borilayotgan islohotlardan eng muhim talablaridan biri bu rivojlangan chet el mamlakatlari adabiyotlarini tarjima qilish va o'quv jarayoniga tatbiq etish malakali mutaxassislar tayyorlashning muhim omili hisoblanmoqda. Lekin bugungi kunda kimyo sohasidagi chet el adabiyotlarini tarjima qilish uchun tegishli izohli lug'atlar yetarli darajada mavjud emas. Shuning uchun kimyoga oid terminlarning tezaurus lug'atini tayyorlash va nashr etish muhim amaliy ahamiyat kasb etadi. [B.F.Muxiddinov, T.I.Nurmurodov, X.M.Vapoyev 2020:4]

Kimyoviy atamalarning tezaurusini yaratish har doim qiyin vazifa bo'lib kelgan, chunki kimyoviy terminologiya murakkab, kontekstga juda bog'liq va doimo rivojlanib boradi. Kimyoviy atamalarning tezaurusini yaratishda semantik muammolar ayniqsa qiyin, chunki ular quyidagi munosabatlarni o'z ichiga oladi.

1. Sinonimik (ma'nodoshlik) munosabati.
2. Giper-giponimik (jins-tur) munosabati.
3. Xolo-meronimik (butun-bo'lak) munosabati.
4. Graduonimik (leksik darajalanish) munosabati.
5. Funksionimik (vazifadoshlik) munosabati.
6. Zidlik munosabati [Abjalova M, 2022:92].

Kimyoviy terminlarning tezaurus lug'atini yaratishda dastavval baza shakllantiriladi. Baza kimyoviy terminlarni jamlash bilan boshlanadi. Bazada quyidagi xususiyatlar ochib beriladi va semantik munosabatlar o'rnatiladi;

- termin nomi;
- terminning ta'rifi;
- viki ma'lumot linki;
- etimologiyasi;
- sinonimi;

- antonimi;
- giponimi;
- giperonimi;
- xolonimi;
- meronimi.

Kimyoviy atama bir nechta ma'noga ega bo'lsa, noaniqlik paydo bo'ladi, bu chalkashlik va noto'g'ri identifikatsiyaga olib kelishi mumkin. Masalan, “kislota” atamasi xlorid kislotasi kabi protonlarni beruvchi moddani yoki sirka kislotasi kabi COOH funksional guruhini o'z ichiga olgan moddani nazarda tutishi mumkin. Tegishli kontekstsiz “kislota”ning qaysi ta'rifi qo'llanilishini aniqlash qiyin bo'lishi mumkin. Shunga o'xshash so'zlarda noaniqlik paydo bo'lmasligi uchun kimyo fanining tezaurus lug'atini yaratish dastlab anorganik kimyo terminlari asosida yaratish maqsad qilingan.

Omonimiya deganda yozilishi va talaffuzi bir xil, ammo ma'nosi har xil bo'lgan atamalar tushuniladi. Kimyoda bu ko'pincha gidroksil guruhini yoki Ogayo shtatining qisqartmasini ifodalashi mumkin bo'lgan “OH” kabi qisqartmalar bilan sodir bo'ladi. Bu aniq kimyoviy atamalarni qidirishda xatolarga olib kelishi mumkin.

Ko'p ma'nolilik - bu bitta atama bir-biriga bog'liq bo'lgan bir nechta ma'noga ega bo'lgan hodisa. Masalan, “qo'rg'oshin” metall elementga yoki qalamlarda ishlatiladigan materialga ishora qilishi mumkin. Kimyoviy atamalarning tezaurusi bilan ishlashda chalkashliklarni keltirib chiqarmaslik uchun ushbu ko'p ma'nolardan xabardor bo'lish kerak.

Tilning lug'at boyligini tizim sifatida o'rganishda leksemalarning gipo-giperonimik munosabatlariga tayanish muhim ahamiyatga ega: u tabiat va jamiyatdagi narsa-predmetlar, voqea-hodisalarning tildagi nomlari bo'lgan leksemalarning ma'nolarini va shu ma'nolar orqali borliqdagi narsa-hodisalarning o'zlari haqidagi tushuncha-tasavvurlarni umumlashtirish va farqlash imkonini beradi. Dastlabki tadqiqotlarda, shuningdek, misollarda namoyon bo'lishi jihatidan jins-tur va butun-bo'lak munosabatlari bir leksik-semantik guruhga birlashtirilgan. Ammo jins-tur munosabati o'zaro teng bog'lanishga asoslanadi va so'zlar bir dominanta (uya) so'z atrofida birlashadi, butun-bo'lak munosabati esa butunni tashkil etuvchi yoki butunning qismlarga bo'linuvchi okkozional xarakterga ega munosabat turi hisoblanadi. Masalan, tana – odam tanasi = butun tushuncha, tana, bosh, bo'yin, elka, qo'l, bel, qorin, oyoq qatoridagi tana so'zi va odam gavdasining qismlarini ifodalovchi lug'aviy birliklar orasida o'zaro tobelanish va bo'ysunish munosabatlari mavjud emas. Tana so'zi 1) jon, ruh so'zlari bilan (tana → jon, ruh), 2) shox, ildiz, barg so'zlari (tana → shox, ildiz, barg), 3) bosh, bo'yin, elka, qo'l, bel, qorin, oyoq so'zlari bilan (tana → bosh, bo'yin, elka, qo'l, bel, qorin, oyoq), 4) po'stloq, mag'iz so'zlari bilan (tana → po'stloq, mag'iz) bog'lanadi. Ushbu to'rt bog'lanish butun ma'nosini ifodalovchi tana so'zi nutq sharoiti bilan bog'liq holda bir matnda bir so'z bilan bog'lansa, ikkinchi bir matnda, yana boshqa bir so'z bilan bog'lanib keladi. Ko'rinadiki, butun-bo'lak munosabati nisbiy paradigmatic bo'lib,

unda assotsiativ aloqa mavjud. Bir leksik paradigmatic qatorni hosil qiluvchi daraxt, archa, qayin, terak, tol, chinor tushunchalarini ifodalovchi lug'aviy birliklar o'zaro teng bog'lanish asosida aloqaga kirishadi. Bu qatordagi daraxt so'zi umumiylik, yaxlitlik ma'nosini ifodalovchi so'z sifatida archa, qayin, terak, tol, chinor tushunchalarini shu daraxtning alohida-alohida turlarini ifodalovchi tushunchalar sifatida biriktiradi, qamrab oladi. Bu o'rinda daraxt, archa, terak, qayin, tol, chinor turidagi so'zlar orasida butun-bo'lak munosabati yuzaga kelmaydi. Chunki daraxt, archa, qayin, terak, tol, chinor so'zlari o'zaro jins-tur munosabatini ifodalash asosida bog'langan so'zlar qatori orasidagi farq daraxt, archa, qayin, terak, tol, hinor so'zlari orasidagi mantiqiy semantik aloqa bilan daraxt, shox, ildiz, barg so'zlari orasidagi o'zaro aloqa chog'ishtirilganda yaqqol namoyon bo'ladi. Daraxt so'zi jins tushunchasini ifodalovchi so'z sifatida archa, qayin, terak, tol so'zlari bilan bog'langan. Bu qatordagi daraxt so'zi markaziy, bosh so'z sifatida shu jinsga nisbatan tur tushunchalarining nomlarini ifodalab keluvchi so'zlarning barchasini o'ziga biriktirib keladi. Ayni paytda daraxt so'zi butun tushunchasining nomini ifodalovchi lug'aviy birlik sifatida shu butunlik tarkibiy qismlarining nomini ifodalovchi ildiz, tana, shox, barg so'zlari bilan bog'lanib, bir leksik-semantik qatorni hosil qiladi. So'zlarning giponimik qatori giperonim va giponimdan iborat. Tarkibi kengroq so'zni giperonim, tor mazmundagi tegishli so'zni giponim deb atashadi: idishlar (giperonim), tarelka, chashka (giponimlar) [Abjalova M. 2022:111-113].

Kimyoviy terminlarda gipo-giperonimik munosabatlarni o'rnatishda bir muncha muammolarga duch kelishimiz mumkin. Chunki bu yerda xolonim-meronim munosabatlarini farqlash ham muhim ahamiyat kasb etadi. Ular orasida farqni ajratib olish uchun esa yuqorida gipo-giperonimga oid tushuncha va quyidagi xolonim-meronim haqida tushunchalarni to'g'ri anglab yetishimiz oldimizda turgan kimyoviy terminlar orasida semantik munosabatlar o'rnatilishining asosi bo'lib xizmat qiladi.

termin nomi	giponim	giperonim	xolonim	meronim
Beton	asfalt, plastobeton	beton	beton	trepel, diatomit, pemzalar, vulqon kuli, turn, domna shlaklar, ohak, gips
Boksitlar	diasporbyo mitli boksit, gidrargillitli (gibbsitli)	boksit	boksit	kremnezem, temir oksidi
Vosklar	ozokerit, lanolin	vosklar	vosklar	yuqori yog'lar,



				yuqorimaleku lyarbir atomli spirtlarning murakkab efirlari
Vud qolishmasi	vud qotishmasi	qotishma	vud qotishmasi	Bi, Pb, Sn, Cd
Autooksidlanish	o'z-o'zidan oksidlanish	qaytarilish	Autooksidla nish	Oksidlanish

Ma'lumki, o'tgan asrning o'rtalaridan leksik-semantik guruhlardagi leksemalarning tizimli munosabatlari sinonimik va antonimik jihatdan o'rganila boshlandi. O'sha vaqtda meronimik yoki xolo-partonim munosabatlar tilshunoslar e'tiboriga tushadi. Xolonim (qad. yunoncha ὅλος = «butun» + ὄνομα = «nom») muayyan tushuncha(lar)ning butun holati (ko'inishi), butun haqida tushuncha va uning nomi boshqa tushunchaga nisbatan uning «butun»i tushuniladi. Masalan, kompyuter – qattiq disk va protsessorga nisbatan xolonim, uy – poydevor va tom tushunchalariga nisbatan xolonim. Meronim (qad. yunon. μέρος = «qism» va ὄνομα = «nom») boshqa tushunchaning tarkibiy qismi, meronim ayrim adabiyotlarda partonim (lot. pars, chiq.k. – partis = «qism») deyiladi. Xolonim tarkibiy qismlarini birlashtiruvchi butunlik, meronym esa o'sha butunning tarkibiy qismlari, ya'ni boshqa tushuncha(lar) butunning ajralmas qismi sifatida namoyon bo'ladi. Shu bois butun – qism munosabati qadimgi yunoncha nom bilan xolo-meronim yoxud xolonimmeronim tarzida qo'llaniladi. UzNet lingvistik ma'lumotlar bazasida xolonimlar butun va meronimlar esa qism / bo'lak tarzida nomlangan.

Ushbu semantik muammolarni hal qilish uchun kimyoviy atamalarining tezaurusi ularning ma'nolari va ishlatilishining batafsil tavsifini berishi kerak. Har qanday noaniqlikni aniqlashtirish uchun sinonimlar va ta'riflar kiritilishi kerak va har bir atama uchun kontekstni hisobga olish kerak. O'zaro havolalar tegishli atamalarni ko'rsatish va kimyoviy terminologiyani bog'lash uchun ishlatilishi kerak.

Xulosa. Kimyoviy terminlarning tezaurusini yaratish kimyo sohasidagi mutaxassis bilimlarini mustahkamlash, yuzaga kelishi mumkin bo'lgan semantik muammolarni chuqur tushunishni talab qiladi. Kimyo sohasidagi aniq muloqot va kimyoviy tadqiqotlar muvaffaqiyati uchun bu sohada ehtiyotkorlik bilan ishlash juda muhimdir.

Foydalanilgan adabiyotlar:



1. Kimyoga oid terminlarning o'zbekcha-ruscha-inglizcha izohli lug'ati (I qism) [Matn] ilmiy nashr/ B.F.Muxiddinov, T.I.Nurmurodov, X.M.Vapoyev va boshqalar. - Toshkent: Muharrir nashriyoti, 2020. - 336 b.
2. Abjalova, Manzura. O'zbek tili ontologiyasi: yaratish texnologiyasi va konsepsiyasi [Matn] : M.A. Abjalova. – Toshkent: Nodirabegim, 2022. – 212 b.
3. Xolmanova Z.T. Kompyuter lingvistikasi. – T.: “Asian Book House”, 2020. - 90 b.
4. Abjalova M., Uzakova M. Software and Linguistic Features of Uzbek Synonymizer. // IEEE – UBMK – 2022: 7th International Conference on Computer Science and Engineering. 24-26 September 2022. Istanbul – Turkey. – pp. 177-181.
5. Abjalova M.A. O'zbek tilidagi matnlarni tahrir qilishda uslub masalasi // Ta'lim tizimida ijtimoiy-gumanitar fanlar, – Toshkent, 2013. – №2. – B. 120-125
6. Abjalova M.A. Tahrir va tahlil dasturlarining lingvistik modullari. [Matn]: monografiya / M.A. Abjalova. – Toshkent: Nodirabegim, 2020. – 176 b. ISBN 978-9943-6939-0-6
7. Rahimov A. Kompyuter lingvistikasi asoslari. Toshkent: Akademnashr. 2011:160.
8. Gulyamova Sh., Ahmedova X. O'zbek tili semantik analizatori uchun omonim so'zlar ma'lumotlar bazasini shakllantirish masalasi xususida. So'z san'ati xalqaro jurnali. Toshkent 2021.

O'ZBEK↔INGLIZ SINONIMLARI BAZASINING TUZILISHI VA TARKIBI

Salomov Abbos Alixon o'g'li
ToshDO'TAU

Kompyuter lingvistikasi mutaxassisligi 2-kurs magistranti
Ilmiy rahbar: **Abjalova Manzura Abdurashetovna**

Annotatsiya. Ushbu maqolada lingvistik ta'minot nima ekanligi, o'zbek tilida sinonimlarning o'rni, ingliz tilidagi sinonimlarning ahamiyati ularning farqlari, o'zbek↔ingliz sinonimlari bazasini yaratish va uning tuzilishi hamda tarkibi haqida so'z yuritiladi.

Kalit so'zlar: *sinonimlar, lingvistik ta'minot, lingvistik baza, antonimlar, mashina tarjimas, avtomataik tarjimon dasturlar.*

Annotation. In the article what a linguistic database is, the role of synonyms in Uzbek language, the significance of synonyms in English language and the differences between them, creating the database of Uzbek↔English synonyms and its construction and structure are discussed.

Keywords: *synonyms, linguistic support, linguistic database, antonyms, Machine translation, automatic translation programs.*

Аннотация. В статье обсуждаются, что такое лингвистическая база данных, роль синонимов в узбекском языке, значение синонимов в английском языке и различия между ними, создание базы данных узбекско-английских синонимов, ее построение и структура.

Ключевые слова: *синонимы, лингвистическое обеспечение, лингвистическая база данных, антонимы, Машинный перевод, программы автоматического перевода.*

Kirish. Lingvistik baza yaratish har bir tilda bugungi kunda olib borilayotgan ishlardan biri o'laroq, zamonaviy soha hisoblangan Kompyuter lingvistikasining qaysidir ma'noda bir qismidir. Chunki lingvistik baza websaytlar va turli xil dasturlarning ishlashiga xizmat qiladi. Qolaversa, bir tilning rivojlanishi uchun ham muhim omil hisoblanadi. Kompyuter lingvistikasi ham aynan tabiiy tilni rivojlantirishga qaratilgan soha hisoblanadi.

Bugunga qadar kompyuter lingvistikasida butun dunyo bo'ylab ko'plab yutuqlarga erishildi. Buning natijasida turli til korpuslari, online ma'lumotlar bazalari, til o'rganishga xizmat qiluvchi dasturlar, web sahifalar va avtomatik tarjimon dasturlar yaratildi. Albatta, bu ishlarning hammasining zahirida tilning lingvistik bazalari yotadi.

Lingvistik dasturiy ta'minot (inglizchadan *lingware* = *linguistic* + *software*) - tabiiy tilda audio ma'lumotlar, chizmalar (OCR) va matnlarni tahlil qilish, qayta ishlash, saqlash va qidirishni ta'minlaydigan kompyuter dasturlari va ma'lumotlari.

Asosiy qism. O'zbek–ingliz sinonimlari bazasini yaratishda qilinadigan ishlarning eng avvalida o'zbek tilidagi mavjud sinonimlarni jamlash va ularni boyitish lozim. Bu borada Azim Hojiyevning 1996-yilda ishlab chiqilgan “O'zbek tilining sinonimlari” kitobidan foydalanib ularni jadval asosida shakllantiriladi. Keyingi o'rinda tildagi har bir sinonimik qatorning eng ko'p qo'llaniladigan va uslubiy betaraf bo'lgan dominantasi aniqlanadi. Qolgan sinonim so'zlarning barchasi qaysi uslubga aloqadorligi tekshirilib jadval shakllantiriladi. Keying o'rinda ushbu jadvaldan joy olgan sinonim so'zlarning hammasi tarixiy yoki tarixiy so'z emasligiga ko'ra farqlab chiqiladi. Bunda ushbu so'zlarning bugungi kunda ham qo'llanilishi va qo'llanilmasligiga ko'ra guruhlash ma'qul deb topildi.

Sinonimik jadval (Автосохраненный).xlsx									
B	C	D	E	F	G	H	I	J	K
Sinonimik qator	Izohi	Dominantasi	Ilmiy	Badiiy	So'zlashuv	Rasmiy-idoraviy	Publitsistik	Eskirgan	Eskirmagan
a'mol, amal	Qilingan yoki qilinadigan ish(lar), amal.	amal		amal		a'mol		a'mol	amal
buyuk, ulug', a'zam, azim	Ko'rinish, tuzilish e'tibori bilan g'oyat katta, ulkan.	azim		ulug', buyuk, a'zam	azim				ulug', buyuk, azim, a'zam
a'zo, qism, uzv	Odam yoki hayvon organizmining muayyan bir vazifani bajaruvchi qismi.	a'zo	a'zo	uzv	qism			uzv	a'zo, qism
ilalabad, toabad, mangu, umrbod, abadiy	Doimiy, abadiy, mangu, o'lmas	abadiy		abadiy, mangu, umrbod, toabad, ilalabad				ilalabad, toabad	abadiy, mangu, umrbod
abadiylashtirmoq, manguilashtirmoq	Umrini cheksiz uzaytirmoq	abadiylashtirmoq		manguilashtirmoq, abadiylashtirmoq					abadiylashtirmoq, manguilashtirmoq
ardoqlamoq, papalamoq, pupalamoq, ehtiyotlamoq, ehtiyot qilmoq, avaylamoq, ayamoq	Yomon ta'sirdan, salbiy hodisadan himoya qilmoq.	avaylamoq	ehtiyot qilmoq	ardoqlamoq	ehtiyotlamoq, papalamoq, pupalamoq, ayamoq				ardoqlamoq, papalamoq, pupalamoq, ehtiyotlamoq, ehtiyot qilmoq, avaylamoq, ayamoq
abas, foydasiz, befoйда, behuda	Befoyda bo'lgan, hech nafi yo'q	befoyda		abas, behuda	befoyda, foydasiz			abas	foydasiiz, befoйда, behuda
abdol, xudojo'y, dindor	e'tiqodiga ega dindor	xudojo'y		abdol	xudojobefoyda, foydasiz	dindor		abdol	dindor, xudojo'y
abgor, nochor, xarob	yomon ahvolda	xarob		nochor, xarob	abgor				abgor, nochor, xarob
abjaq bo'lmoq, majaq				abjaq bo'lmoq					abjaq bo'lmoq, majaq

1-rasm. O'zbek tili sinonimlari jadvali.

O'zbek va ingliz tillarini bog'lashda rus tili asosiy vosita hisoblanadi. Negaki hozircha mukammal darajadagi o'zbekcha-inglizcha lug'atlar yaratilmagan. Shunga ko'ra dastlab tarjimada o'zbekcha-ruscha, so'ngra ruscha-inglizcha lug'atlardan foydalanildi.

Ko'rinadiki, har bir jadval kompyuter dasturi uchun ma'lumotlar bazasi hisoblangani holda bir-biri uchun ham axborot banki vazifasini o'taydi. Demak, kompyuter dasturi uchun lingvistik ta'minot yaratishda ham uzviylikka, sistemalilikka qat'iy rioya qilinadi.

Keyingi o'rinda ingliz tilida uchraydigan sinonimlar yig'ib olindi. Bir vatqning o'zida sinonimlar bilan birgalikda antonimlar ham yig'ib olinadi. Bunda ishning tushunarliroq bo'lishi ko'zda tutiladi.

SYNONYMS AND ANTONYMS

1. AWKWARD

Syn: clumsy, ungainly, ponderous, rough

Ant: clever, dexterous, apt, skilful

2. APARENT

Syn: evident, obvious, perceptible, distinct

Ant: masked, obscure, indistinct, doubtful

3. ALLURE

Syn: entice, fascinate, tempt, seduce

Ant: repulse, repel, deter, distaste

4. ABORTIVE

Syn: vain, useless, fruitless, unproductive

Ant: effectual, productive, successful, progressive

5. ABOMINABLE

Syn: odious, detestable, repugnant, aversion

Ant: likeable, pleasant, affectionate, enjoyable

2-rasm. Ingliz tili sinonimlar va antonimlar ro'yxati.

Ingliz o'zbek sinonimlarini muvofiqlashtirib chiqish jarayonida bir nechta muammolar uchrashi mumkin. Ulardan biri sinonim so'zlarning sonidir. Aytaylik bir tilda 5 ta, ikkinchi tilda 7 ta sinonim uchrasa, ular qanday muvofiqlashtiriladi? Bu savolga filologiya fanlari bo'yicha falsafa doktori (DSc) Manzura Abjalova Abdurashetovnaning fikrlariga ko'ra aynan tarjimasi chiqmagan so'zlar uchun dominanta so'zni muvofiqlashtirish kerak.

Masalan, “antikvar” so'zining o'zbek tilidagi sinonimlari quyidagilardir:

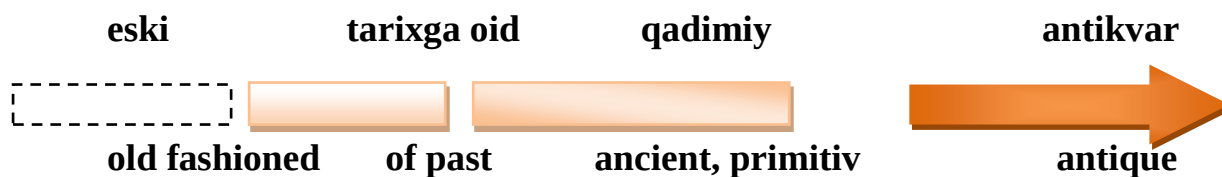
antikvar, eski, qadimiy, tarixga oid

Ingliz tilida esa bu so'zning quyidagi shakllari mavjud:

antique, ancient, old fashioned, primitive, of past

Ushbu sinonim so'zlarning soni bir xil emas. O'zbek tilida 4 ta, ingliz tilida 5 ta sinonim mavjud. “***Antique***” so'zi “***antikvar***” so'zi bilan bir ma'noga ega. “***Ancient***” va “***primitive***” so'zlari esa “***qadimiy***” so'ziga sinonim bo'la oladi. “***Old fashioned***” so'zi “***eski***”, “***of past***” so'zi esa “***tarixga oid***” so'zi bilan tarjimada semantik muvofiqlikka ega.

Bundan tashqari, sinonimlar o'z tarjimasi bilan muvofiqlashtirib chiqilgandan keyin ularning graduonimiyasi ham yaratilishi ko'zda tutilgan, ya'ni sinonimlar o'zlarining ma'no darajalariga ko'ra tartiblab chiqiladi. Bunda so'zlarning ijobiy yoki salbiy ma'no anglatishi yoki kuchsiz-kuchli ma'noga egaligiga ko'ra quyidagicha tartiblanadi:



3-rasm. Muvofiqlashtirilgan o'zbek↔ingliz sinonimlarining graduonimiyasi.

Xulosa. Ushbu sinonimlar bazasida faqat o'zbek tilidagi sinonimik qatorlar uslubiga ko'ra tarixiy eskirgan yoki eskirmaganiga ko'ra tahlilga ega. Shuningdek har bir sinonimik qator uchun umumiy qisqacha izoh ham berib o'tilgan. Umuman olganda, o'zbek va ingliz tillarida mavjud sinonimlarning muvofiqlashtirilib va darajalab chiqilishi katta ilmiy ahamiyatga ega. Ayniqsa, bu ish Mashina tarjimasini va avtomatik tarjimon dasturlar uchun muhim lingvistik ta'minot bo'lib xizmat qiladi. Tarjimadagi so'z tanlash bilan bog'liq muammolar ham o'z yechimini topadi.

Foydalanilgan adabiyotlar:

1. Abjalova M.A. Tahrir va tahlil dasturlarining lingvistik modullari. [Matn] : monografiya / M.A. Abjalova. – Toshkent: Nodirabegim, 2020. – 176 b. ISBN 978-9943-6939-0-6
2. Abjalova M.A. O'zbek tili ontologiyasi: yaratish texnologiyasi va konsepsiyasi. [Matn] : monografiya / M.A. Abjalova. –Toshkent: Nodirabegim, 2021. – 215 b. ISBN 978-9943-7804-5-3.
3. Abjalova M., Uzakova M. Software and Linguistic Features of Uzbek Synonymizer. // IEEE – UBMK – 2022: 7th International Conference on Computer Science and Engineering. 24-26 September 2022. Istanbul – Turkey. – pp. 177-181.
4. Abjalova M.A. O'zbek tilidagi matnlarni tahrir qilishda uslub masalasi // Ta'lim tizimida ijtimoiy-gumanitar fanlar, – Toshkent, 2013. – №2. – B. 140-145.
5. Abjalova M.A. Синонимайзер (синонимизатор) в образовательной корпусе узбекского языка. // Родные языки и культуры в современном изменяющемся мире. – Тува, 2022. – 79-84 с.
6. Abjalova M., Bozorov S. Kvazisinonimlar nutqiy imkoniyat sifatida. // Актуальные вызовы современной науки, LX Международная научная конференция 26-27 июня 2021 г. – С. 123-126.
7. Bergelson M.B. Boguslavskaya O. Yu. Gaft R.I. Intellektual tizimlarda til faoliyatini modellashtirish.
8. Hojiyev A. O'zbek tilining sinonimlari. 1996. Toshkent. – 374 b.
9. https://www.researchgate.net/publication/2614910_The_Creation_Distribution_And_Use_Of_Linguistic_Data_The_Case_Of_The_Linguistic_Data_Consortium
10. <https://www.thesaurus.com/browse/database>
<https://learn.microsoft.com/en-us/sql/relational-databases/synonyms/create-synonyms?view=sql-server-ver16>

KOMPYUTER LEKSIKOGRAFIYASI VA LINGVODIDAKTIKASI

AXBOROT TEXNOLOGIYALARI SOHASIGA DOIR XORIJIY TILLARDAN O'ZLASHGAN SO'ZLAR TADQIQI

Ilmiy rahbar: Atamuratov Rasuljon Kadirjonovich

ToshDO'TAU dots.v.b., PhD

Shakirova Sevara Toxirjon qizi

ToshDO'TAU Noshirlik ishi yo'nalishi 3-kurs talabasi

Annotatsiya: Mazkur maqolada o'zbek tiliga o'zga tillardan axborot texnologiyalari sohasiga doir kirib kelgan o'zlashma so'zlarning tahlili hamda ularning tilning leksik tizimiga moslashishi jarayoni to'g'risida so'z yuritilgan.

Kalit so'zlar: *axborot texnologiyalari, o'zlashma so'zlar, terminologiya, termin, atamalar tizimi.*

Аннотация: В данной статье рассматривается анализ заимствованных слов, пришедших в узбекский язык из других языков, относящихся к сфере информационных технологий, и процесс их адаптации к лексической системе языка.

Ключевые слова: *информационные технологии, идиомы, терминология, термин, система терминов.*

Annotation: This article discusses the analysis of borrowed words that came into the Uzbek language from other languages related to the field of information technology, and the process of their adaptation to the lexical system of the language.

Key words: *information technologies, idioms, terminology, term, system of terms.*

Zamonaviy texnologiyalarning jadal taraqqiyoti sari ona tilimizda yangi so'zlar ham ko'payib bormoqda. O'zbek tili leksikasi muntazam ravishda ichki va tashqi manbalar hisobiga boyib bormoqda. XX-XXI asrlarda jamiyat hayotida bo'lgan o'zgarishlar so'zlarning o'zlashish va iste'moldan chiqish jarayonini tezlashtirdi. O'zlashma so'zlarning 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, mavjud so'zning lug'aviy ma'nolaridan birini yangi ma'noda qo'llash hamda boshqa tildan so'z qabul qilish orqali hosil qilinadi. O'zbek tiliga xorijiy so'zlarning o'zlashish yo'llari turlichadir.

Agar tushuncha aynan o'zbekcha nomlanishga ega bo'lmasa, o'zining ijobiy va samarali ta'siriga ega bo'lsa, to'g'ridan-to'g'ri o'zlashtiriladi. O'zlashma so'zlar axborot texnologiyalari sohasida ham keng qo'llanmoqda. Yangi texnik qurilmalarning paydo bo'lishi bilan bir vaqtda ularni belgilovchi atamalar paydo bo'ladi. O'zbek tilini rivojlantirishining hozirgi bosqichi bir qator qo'shimcha va

intralingvistik sabablarga ko'ra xorijiy tillardan o'zlashishning faollashishi bilan tavsiflanadi. Axborot texnologiyalari terminologiyasida boshqa terminologik tizimlarga qaraganda ko'proq xorijiy o'zlashma so'zlar mavjud va ularning aksariyati ingliz tilidan olingan. Masalan, IT - ingliz tilidan olingan “Information Technology” so'zlarining qisqartmasi bo'lib, o'zbek tilida “Axborot texnologiyalari” deb yuritiladi. Shunday bo'lsa ham, bu so'z IT ko'rinishida hozirgi kunda og'zaki va yozma nutqda talaffuz qilinadi va ishlatiladi.

Tadqiqotimiz davomida axborot texnologiyalari sohasida o'zga tillardan o'zbek tiliga o'zlashgan, hozirda keng qamrovda rasmiy muloqot va nashrlarda ishlatib kelinayotgan o'zlashgan so'zlar tarkibi keltirilgan.

IT Park (IT-ingliz tilidan olingan “*Information Technology*” so'zlarining qisqartmasi) IT Park – istiqbolli startap loyihalarni ishga tushirish va bozorga chiqarishni ta'minlash uchun mo'ljallangan obyektlar, binolar va inshootlar majmuasi, IT-kompaniyalar uchun, shu jumladan, ilmiy va ta'lim tashkilotlari bilan integratsiyalashgan holda hududiy erkin iqtisodiy zona. IT sohasida faol va iqtidorli insonlar buxgalteriya hisobi, yuridik, marketing va ta'lim orqali o'z g'oyalarini haqiqiy biznes loyihalariga aylantirish uchun haqiqiy imkoniyatga ega bo'ladigan joy.

Messenjer (inglizcha *kuryer, messenjer*) — matnli xabarlar, audioyozuvlar, fotosuratlar va boshqa multimediyalarni bir zumda almashish uchun mo'ljallangan dastur. Ilovalar (telegram, whatsapp, Instagram) kompyuter, smartfon, planshetga o'rnatiladi va internet orqali ishlaydi.

Akkaunt (inglizcha *account – hisob*) – bu saytda ro'yxatdan o'tgandan so'ng yaratiladigan foydalanuvchining shaxsiy sahifasi yoki shaxsiy kabineti.

Bot (ingliz tili “robot” so'zidan olingan bo'lib) – bu foydalanuvchilar uchun mo'ljallangan interfeyslar orqali avtomatik ravishda berilgan jadvalga muvofiq har qanday harakatlarni bajaradigan maxsus dastur asosida ishlaydigan virtual robot (telegram ijtimoiy tarmog'ida keng qamrovda qo'llaniladi).

Ban (inglizcha *taqiqlash, qulflash*) – bu internetdagi foydalanuvchilarning xatti-harakatlarini nazorat qilish usullaridan biri. Telegram ijtimoiy tarmog'ida nojo'yi xatti-harakat qilgan foydalanuvchi kanal admini tomonidan uch marta ogohlantiriladi. O'zgarish yuz bermagan taqdirda, kanal admini tomonidan foydalanuvchi tarmoqda qulflanadi, unga tarmoqdan foydalanish taqiqlab qo'yiladi.

SMM (inglizcha *Social Media Marketing*) – ijtimoiy tarmoqlarda marketing bu mahsulotlar, tovarlar va xizmatlarni ijtimoiy tarmoqlarda tanitish, uni yurgizish.

Heshteg (inglizcha *hashtag – hash belgisi (#) va teg-label*) – kalit so'z, xabarning mavzusi yoki bir nechta so'zlari, teg (belgi) mikrobloglar va ijtimoiy tarmoqlarda qo'llaniladi.

Vidjet (inglizcha *widjet bezak*) – bu kompyuter ish stoli, smartfon, planshet ekrani yoki veb-sayt sahifasida ma'lumotlarni aks ettiruvchi kichik grafik dastur.

Big Data (inglizcha *katta ma'lumotlar*) – katta hajmdagi ma'lumotlarga nisbatan qo'llanadigan termin hisoblanadi (odatda terabayt, ekzabayt va petabaytlar darajasida).

Sketch (inglizcha *eskiz*) – kelajakdagi ishning muhim xususiyatlari va elementlarini tavsiflashga imkon beradigan vizual loyihaning eskizlari, sxemalari, qoralamalari. Sketch grafik dizayn sohasida keng qo'llanadi.

Trecking (inglizcha *kuzatish*) – kompyuter grafikasida harakatlanuvchi obyektlarning o'rnini kamera yordamida aniqlash.

O'zbekistonda **trekking** so'zi aholi va tadbirkorlik subyektlarining davlat organlari va tashkilotlariga murojaatlarini ko'rib chiqish holatini onlayn rejimida kuzatish samarali tizimlarini joriy etish ma'nosini anglatadi.

Internet-banking (inglizcha) – bu mijozlarning depozit hisobvaraqlarini, jumladan bank kartalariga ochilgan hisobvaraqlarini, internet orqali boshqarish huquqini beruvchi tizimdir.

Startup (inglizcha *startup, startup company*) — bu tadbirkor tomonidan kengaytiriladigan biznes modelini izlash, ishlab chiqish va tasdiqlash uchun amalga oshiriladigan kompaniya yoki loyiha.

Texnopark (inglizcha *technopark*) – innovatsion loyihalarni amalga oshirish uchun hududiy, ilmiy, texnologik va texnik bazadir.

ChatGPT (inglizcha *Generative Pre-trained Transformer*) – bu OpenAI tomonidan yaratilgan sun'iy intellektga asoslangan chatbot.

blokcheyn (inglizcha *blockchain* “*block*” - blok, “*chein*” – zanjir) — raqamli transformatsiyaning negizidagi texnologiyalardir. Ularning har biri sanoatning ayrim sohalarida inqilobiy o'zgarishlarni amalga oshirishga qodir va birlashgan holda o'z salohiyatini sezilarli darajada oshirishi mumkin.

Xab (inglizcha) – havola almashinuviga ilova bo'lgan SEO doiralarida taniqli havolalarni sotib olish va sotish tizimi.

Feedback (inglizcha *Feedback*) – shaxs yoki guruhning har qanday harakat, hodisa, ma'lumotlarga munosabati.

Promo kod (inglizcha *promo-kod*) – maxsus shartlarda mahsulot yoki xizmatni sotib olish huquqini beruvchi harflar yoki raqamlardan iborat belgilar to'plami.

Fleshmob (inglizcha *flesh-mob, flesh-lahza, mob-olomon*) – oldindan rejalashtirilgan ommaviy harakat. Odamlarning katta guruhi jamoat joyida yig'ilib, oldindan belgilangan ommaviy harakatlarni amalga oshiradi.

Brending (inglizcha *Branding*) – bu kompaniyaning brend identifikatorini yaratish jarayoni. Ushbu jarayon, logotip, shior, vizual dizayn kabi brendni qo'llab-quvvatlaydigan materiallarni taqdim etadi.

Tilimizda o'z muqobiliga ega bo'lgan so'zlar ham o'zlashgan holda qo'llanmoqda. Vaholanki, ayrim o'zlashgan so'zlarning tilimizda o'z muqobili bor. Misol uchun *trend* so'zi. Kundalik hayotimizda *trend* so'zini musiqa, kiyim yoki ranglarga nisbatan trendga chiqdi sifatida ishlatilganini ko'rish mumkin. Ammo



trend so'zining o'zbek tilidagi muqobili sifatida *ommabop* so'zi mavjud. Bunday o'zlashma so'zlar talaygina.

Xulosa o'rnida keltirish mumkinki:

O'zlashma so'zlar jamiyat va davr ehtiyoji bilan paydo bo'ladi va qo'llanilishi bo'yicha doimiylik tusini oladi.

a) O'zbek tili leksikasida inglizcha o'zlashma so'zlar soni ortib bormoqda.

b) O'zlashma so'zlarning muqobil variant yo'qligi sabab hech qanday o'zgarishsiz qo'llanmoqda.

c) Ayrim o'zlashma so'zlar o'zbek tili izohli lug'atiga kiritilmagan.

d) 2 xil qo'llanilayotgan o'zlashma so'zlar imlo bo'yicha ko'pgina xatoliklarga sabab bo'lishi mumkin. Shu sababli ularning faqatgina bitta ma'noda qo'llanilishini ta'minlash lozim.

Aytish joizki, bugungi kunda ingliz tili xalqaro til sifatida o'z mavqeyini tobora mustahkamlamoqda va ilmiy tadqiqot tiliga aylanmoqda. Shu bois global va ijtimoiy tarmoq hamda ilmiy manbalar orqali ingliz tiliga oid so'zlarning tarqalishi kengaymoqda. Ingliz tilida faol qo'llanilayotgan so'zlar termin mavqeyiga aylanmoqda. Shu bois turli manbalar orqali tilga kirib kelayotgan o'zlashma so'zlarni o'zlashish jarayonida ularni to'g'ri va o'rinli qo'llash, tarjima imkoniyati cheklangan yoxud asl o'zlashma semalarini namoyon etadigan milliy so'z ifodasi bo'lmagan holdagina o'zlashma so'zlarni matnlarda qo'llash o'rinli hisoblanadi.

O'ZBEKISTON MEDIA MAYDONIDA SPORT SAYTLARINING RIVOJLANISH TENDENSIYALARI

Boborajabov Ergash

O'zDJTU mustaqil tadqiqotchi-izlanuvchisi

Annotatsiya. Maqolada internet tarmog'idagi sport saytlari xususida so'z yuritiladi. Shuningdek, sport saytlari faoliyatini yanada kengaytirish, saytning uzluksiz reformasini amalga oshirish yutuq va kamchiliklari haqida tahlillar keltiriladi.

Kalit so'zlar: *Internet, sport saytlari, sport jurnalisti, reklama, milliy sport, internet jurnalistikasi.*

Abstract. The article talks about sports sites on the Internet. Also, the analysis of the achievements and shortcomings of the further expansion of the activity of sports sites, the implementation of continuous reform of the site is presented.

Key words: *Internet, sports sites, sports journalist, advertising, national sports, internet journalism.*

Аннотация. В статье рассказывается о спортивных сайтах в Интернете. Также представлен анализ достижений и недостатков дальнейшего расширения деятельности спортивных площадок, осуществления непрерывного реформирования площадки.

Ключевые слова: *Интернет, спортивные сайты, спортивный журналист, реклама, национальные виды спорта, интернет-журналистика.*

Jahon jurnalistikasining eng yosh va navqiron yo'nalishi bo'lgan Internet jurnalistikasi bugun paydo bo'lganiga chorak asr bo'lib ulgurmay bir necha yuz yillarga tatiydigan taassurotlar paydo qilmoqda. Internet nafaqat jurnalistika uchun, balki, jamiyatning barcha jabhalari uchun ham bitmas-tuganmas imkoniyatlar manbayiga aylanib ulgurdi. Mazkur platforma davlat uchun eng tezkor hujjatlar aylanmasidan tortib, byurokrtiyaga chek qo'ygan instrument vazifasini o'tayotgan bo'lsa, tadbirkorlar uchun eng katta va xaridorgir "rasta" vazifasini bajarmoqda; shifokorlar uchun masofadan turib eng murakkab operatsiyalarni bajarish imkoniyatini bergan "dastgoh" bo'lsa, jurnalistlar uchun eng katta axborot makoniga aylangan megapolis vazifasini o'tamoqda.

Dunyo globallashuv jarayonida ekan, har bir sohaning o'z vektorlari bo'yicha kengayishini, takomillashishini, ixtisoslashishini va yana yangi yo'nalishlarga adaptatsiya bo'lishini taqozo etadi. Jumladan, sport jurnalistikasining yangi formati bo'lgan veb-saytlarning axborot bozoriga kirib kelishi mazkur sohaning murakkab jihatlarida ixtisoslashuviga zamin yaratdi. Jamiyatda ro'y berayotgan jarayonlar dinamikasidan qat'i nazar, sport jurnalistikasi sohasidagi ixtisoslik mualliflaridan



nafaqat professional sport bilimlariga, jumladan, nozik nuyanslarga ega bo'lishni, balki harakatchanlikni, katta hajmdagi raqamli ma'lumotlar bilan ishlash qobiliyatini ham talab qiladi. Shu bilan birga, publitsistning lingvistik va stilistik mahorati ham katta rol o'ynaydi. Buning natijasida jurnalist nomga ega bo'ladi va sport ishqibozlarining xushmuomala va talabchan auditoriyasini ifodalovchi o'zining doimiy o'quvchilari doirasini yaratadi.

O'zbekiston Respublikasi Prezidenti Shavkat Mirziyoyev ta'biri bilan aytganda, “dunyoda hech qachon bir xil tong otgan emas”. Mazkur ibora paydo bo'lganidan buyon ming yillar o'tdi. Bu – butun koinot, tabiat va bashariyat doimiy o'zgarishda, yangilanishda, deganidir. Davrlar o'taveradi, zamonlar almashaveradi. Bu yorug' olamga kelgan har bir odamzod avlodi turli o'zgarish va yangilanishlar sari boraveradi [Мирзиёев, 2022:4]. So'nggi yillarda mamlakatimizning barcha sohalarida bo'lganidek, ommaviy axborot vositalarida ham tub o'zgarishlar, yangilanishlar, evrilishlar sodir bo'ldi. Ishlash tizimi mazmun va shakl nuqtayi nazaridan o'zgardi, uning yangi formatlari shakllandi. Ularda beriladigan materiallar mazmun-mohiyati va mundarijasi yangicha uslublarda namoyon bo'la boshladi.

Ayniqsa, global tarmoq bo'lgan internet jurnalistikasi zamon va makon tushunchalaridan o'zib ketdi. Bu borada millionlab muxlis va ishqibozlarga ega soha sport jurnalistikasi har qachongidan-da ko'proq muvaffaqiyat qozondi. Aynan mazkur sohada yurtimizda erishilayotgan yutuqlar, mavjud muammolar, istiqboldagi maqsadlarni o'rganuvchi “O'zbekistonda sport saytlari: paydo bo'lishi, faoliyat yuritishi va rivojlanish tendensiyalari” mavzusidagi tadqiqot ishi birinchi marta tizimli asosda tahlilga tortilishi hisoblanadi.

“Sport jurnalistikasining hozirgi holati biz bilgan yigirma-yigirma besh yil avval mualliflar asosan musobaqalar haqidagi reportajlar va sportchilar haqidagi xabarlar bilan cheklanib qolganidan ko'p jihatdan farq qiladi”, deb yozadi taniqli sport jurnalisti A.Galinskiy. O'quvchilar orasida onlayn sport ommaviy axborot vositalarining mashhurlikiga qaramay, onlayn sport jurnalistikasining ko'pgina nazariy jihatlarini hali ham to'liq tushunilmagan. Sport televideniyesi yoki sport matbuoti tarixiga bag'ishlangan asarlar mavjud, O'zbekiston Internet-mediasini rivojlantirish bo'yicha ishlar mavjud, ammo ilmiy hamjamiyat hozirgacha O'zbekistondagi sport saytlari asosida rivojlanib borayotgan internet jurnalistikasining hozirgi holatini yoritib beradigan to'liq tadqiqotni taqdim etmadi.

Mazkur maqolada sport saytlari rivojlanish istiqbollari va O'zbekiston media bozorining hozirgi holatini hisobga olgan holda soha saytlarini, internet portallarini tasniflashga harakat qilingan. Ish onlayn medianing muhim tarkibiy qismlarini har tomonlama o'rganishga qaratilgan: auditoriya, ijtimoiy tarmoqlar, davlat va nodavlat sport saytlari va ularning roli, uslubi, mualliflari, vizualizatsiyasi, interaktivligi va bloglari, reklama, kontent-tematik o'ziga xosliklari ochib berilgan.

Sport jurnalistikasi va sport saytlarining rivojlanish tendensiyalari bo'yicha bir guruh olimlar tomonidan tadqiqot ishlari olib borilgan [2] [3] [4]. Jahon jurnalistikasining muhim yo'nalishlaridan biri bo'lgan sport jurnalistikasini tahlil



etish, xususan uning rivojlanish tendensiyalari va muammolari bo'yicha nazariy va amaliy o'rganishga yo'naltirilgan ilmiy izlanishlar dunyoning yetakchi ilmiy markazlari va oliy ta'lim muassasalari, jumladan, Hartford universiteti (AQSH), Hartford universiteti (Buyuk Britaniya), Leeds Beckett University (Buyuk Britaniya), University of Galway (Irlandiya), National University of Singapore (Singapur), University of Hong Kong (Xitoy), Hebrew University of Jerusalem (Isroil), M.V.Lomonosov nomidagi Moskva davlat universiteti – MGU (Rossiya), O'zbekiston davlat jismoniy tarbiya va sport universiteti (O'zbekiston) O'zbekiston jurnalistika va ommaviy kommunikatsiyalar universiteti, (O'zbekiston), O'zbekiston davlat jahon tillari universiteti (O'zbekiston)da olib borilmoqda.

Shuningdek, Rossiya Federatsiyasida K.Alekseyev [Алексеев, 2013] [Алексеев, 2021], E.Voytik [Войтик, 2010], D.Desyuk, S.Ilchenko [Ильченко, 2018], O.Shmidt o'z asarlarini sport televideniyesi, radiosi va matbuoti tarixi va rivojlanishiga bag'ishlagan. Internet jurnalistika muammolari haqida A.Amzin, S.Balmayeva, E.Vartanova, I.Zasurskiy, M.Lukinalar yozgan. Biroq, onlayn sport jurnalistikasining media makonida hukmronligi va uning talabi va ommabopligiga qaramay, bu hodisa bo'yicha hali to'liq va ko'lamli tadqiqotlar mavjud emas. Ushbu muammo haqida tadqiqotchilar E.Vartanova, L.Kruglova, A.Zamkov [Амзин, 2020], [Болмаева, 2010] [Замков, 2017], shuningdek, ayrim jurnalistik amaliyotchilar qisman yozganlar, ammo bugungi kunda sport saytlarining tuzilishini o'zgartiruvchi tegishli texnologik va auditoriya siljishlarining to'liq tavsifi hali mavjud emas. Media-industriya, raqobatbardosh ustunliklarni, funksional salohiyatni va yetakchi sport onlayn media auditoriyasi bilan o'zaro munosabatlarning eng samarali shakllarini aniqlashda mazkur tadqiqot ishimiz dolzarb hisoblanadi.

“2022-2026 yillarga mo'ljallangan Yangi O'zbekiston taraqqiyot strategiyasi to'g'risida”gi O'zbekiston Respublikasi Prezidentining Farmoni [12] sohaga doir ishlarning rivojlanishini yanada tezlashtirdi. Jumladan, jismoniy tarbiya va sport bilan muntazam shug'ullanadigan aholi sonini 2026 yilda 33 foizga chiqarish, respublika yoshlari o'rtasida milliy sport turlarini ommalashtirish, O'zbekistonda sportning ommaviyligini oshirish maqsadida milliy sport turlari va xalq o'yinlarini rivojlantirish [13] bu kabi harakatlarni internetdagi sport saytlari va ijtimoiy tarmoqlar orqali ommalashtirish vazifalari dolzarb ahamiyat kasb etadi.

2017-2021 yillarda O'zbekiston Respublikasini rivojlantirishning beshta ustuvor yo'nalishi bo'yicha Harakatlar strategiyasida yoshlarga oid davlat siyosatini takomillashtirish, mamlakat yoshlarini jismonan sog'lom, ruhiy va intellektual rivojlangan, mustaqil fikrlaydigan, qat'iy hayotiy nuqtai nazarga ega, Vatanga sodiq yoshlarni tarbiyalash, demokratik islohotlarni chuqurlashtirish va fuqarolik jamiyatini rivojlantirish jarayonida ularning ijtimoiy faolligini oshirish, yosh avlodning ijodiy va intellektual salohiyatini qo'llab-quvvatlash hamda amalga oshirish, bolalar va yoshlar o'rtasida sog'lom turmush tarzini shakllantirish, ularni jismoniy tarbiya va sportga keng jalb etish [14] kabi dolzarb vazifalar davlat siyosati darajasida belgilab olindi. Bu kabi qamrovi keng va mas'uliyati yuksak vazifalar har

bir millat va xalqni ma'nan sog'lom, aqlan yetuk bo'lib rivojlanishida ulkan platforma vazifasini o'taydi.

Respublikaning taniqli olim va mutaxassislari X.Do'stmuhammad, P.Allamberganova, Yu.Artiqova, A.Karimov [Каримов, 2012] tadqiqotlarida televideniye sohasiga oid turli muammolar tahlilga tortilgan. “O'zbekiston matbuoti”, “Ijtimoiy fikr” jurnallarida, “O'zbekiston adabiyoti va san'ati”, “Hurriyat” va boshqa matbuot nashrlarida mavzu atrofida kichik maqolalar, taqrizlar uchraydi, xolos. N.A.Abduazizova, M.Xudoyqulov, F.A.Mo'minov, K.T.Irnazarov, F.I.Mo'minova, X.Do'stmuhammad, B.A.Do'stqorayev, Sh.Qudratxo'jayev, R.Muhammadiyev, S.Umirov, K.Ismoilova, Yu.Hamdamiyov, Ya.Mamatova, M.Mirsoatova, O.Toshboyev va boshqalarning tadqiqotlari, kitoblari, maqolalarida ham chiqishlarga oid ayrim fikr-mulohazalar bildirilgan. Bevosita sport jurnalistikasi bo'yicha rus olimlari S.Mixaylov, A.Gost [Михайлов, 2012], mamlakatimizdan A.Imomxo'jayev [Имомхўжаев, 2006] lar tomonidan sport jurnalistikasiga oid ayrim yo'nalishlar tadqiq etilgan bo'lsa-da, sport saytlari faoliyatining yoritilishi maxsus o'rganilmagan. Mazkur mavzu atrofida tibbiy va ijtimoiy nuqtai nazardan ilmiy anjumanlar doirasida chiqishlar, bakalavrlik va magistrlik dissertatsiyalari himoya qilingan. Biroq, bu sport jurnalistikasi sohasidagi, sport saytlari rivojlanishining zamonaviy tendensiyalaridagi ilmiy-nazariy bo'shliqni to'ldirish uchun yetarli emas.

Foydalanilgan adabiyotlar:

1. Мирзиёев Ш.М. Янги Ўзбекистон тараққиёт стратегияси. Тўлдирилган иккинчи нашри. – Тошкент: “Ўзбекистон” нашриёти, 2022. 4-бет.
2. <https://www.hartford.edu>
3. <https://www.hud.ac.uk>
4. <https://www.leedsbeckett.ac.uk>
5. Алексеев К.А. Спортивная журналистика: учебник для магистров / К.А. Алексеев, С.Н. Ильченко. – М.: Издательство Юрайт, 2013.
6. Алексеев К.А. Новейшая история отечественной журналистики Учебное пособие / — Москва; Берлин: Директ-Медиа, 2021. — 400 с.,
7. Войтик Е.А. Спортивная медиакоммуникация в России в начале XXI в. — Томск: Издательский Дом ТГУ, 2013., Десюк Д.В. Языковая специфика спортивного телетекста / Д. В. Десюк // Знание. Понимание. Умение. - 2010. - Вып. 4. - С. 187-190.
8. Ильченко С.Н. Основы журналистской деятельности Учебник и практикум для академического бакалавриата / С. Н. Ильченко. — М.: Издательство Юрайт, 2018. — 311 с. — (Серия: Бакалавр. Академический курс).
9. Амзин А.А. Интернет-журналистика. Как писать хорошие тексты, привлекать аудиторию и зарабатывать на этом / Москва: Издательства АСТ, 2020. – 400 с.



10. Болмаева С.М. Медиаконвергенция и мультимедийная журналистика [Текст] : [материалы к обучающим семинарам] / - Екатеринбург: 2010 (Екатеринбург: Типография "Аграф"). – 146.
11. Замков А.В., Крашенинникова М.А., Лукина М.М., Цынарёва Н.А. Роботизированная журналистика: от научного дискурса к журналистскому образованию // Медиаскоп. 2017. Вып. 2.
12. “2022-2026 йилларга мўлжалланган Янги Ўзбекистон тараққиёт стратегияси тўғрисида”ги Ўзбекистон Республикаси Президентининг Фармони 28.01.2022 йилдаги 60-сонли ПФ.
13. <https://lex.uz/docs/5841063#5844198> 29.11.2022й. “2022-2026 йилларга мўлжалланган Янги Ўзбекистон тараққиёт стратегияси тўғрисида”ги Ўзбекистон Республикаси Президентининг Фармони 28.01.2022 йилдаги 60-сонли ПФ.
14. 14. 2017-2021 йилларда Ўзбекистон Республикасини ривожлантиришнинг бешта устувор йўналиши бўйича Ҳаракатлар стратегияси.
15. Каримов А. И. Аудиовизуал журналистика (тележурналистика), - Т.: ЎзДЖТУ, 2012.
16. Михайлов С.А., Мост А.Г. “Спортивная журналистика”, М.: Аспект-пресс, 2002.
17. Имомхўжаев А. И. Футбол – кувончим, дардим, фахрим, Т., “Шарқ” НМАК, 1996.

O'ZBEK MILLIY INTERNET SEGMENTIDA SAYTLARNING YARATILISHI, FAOLIYATI VA OMMALASHISHI

Boborajabov Ergash,

O'zDJTU mustaqil tadqiqotchi-izlanuvchisi

Annotatsiya. Maqolada internet jurnalistika yoki onlayn nashr – bu ommaviy axborot vositasi sifatida ro'yxatdan o'tgan va jurnalistika tamoyillariga muvofiq tegishli funksiyalarni bajarishga qaratilgan veb-saytlar xususida so'z yuritiladi.

Kalit so'zlar: *jurnalistika, internet jurnalistika, veb sayt, ommaviy axborot vositalari, an'anaviy jurnalistika.*

Абстракт. В статье говорится об интернет-журналистике или сетевом издании, представляющем собой веб-сайт, зарегистрированный в качестве средства массовой информации и направленный на выполнение соответствующих функций в соответствии с принципами журналистики.

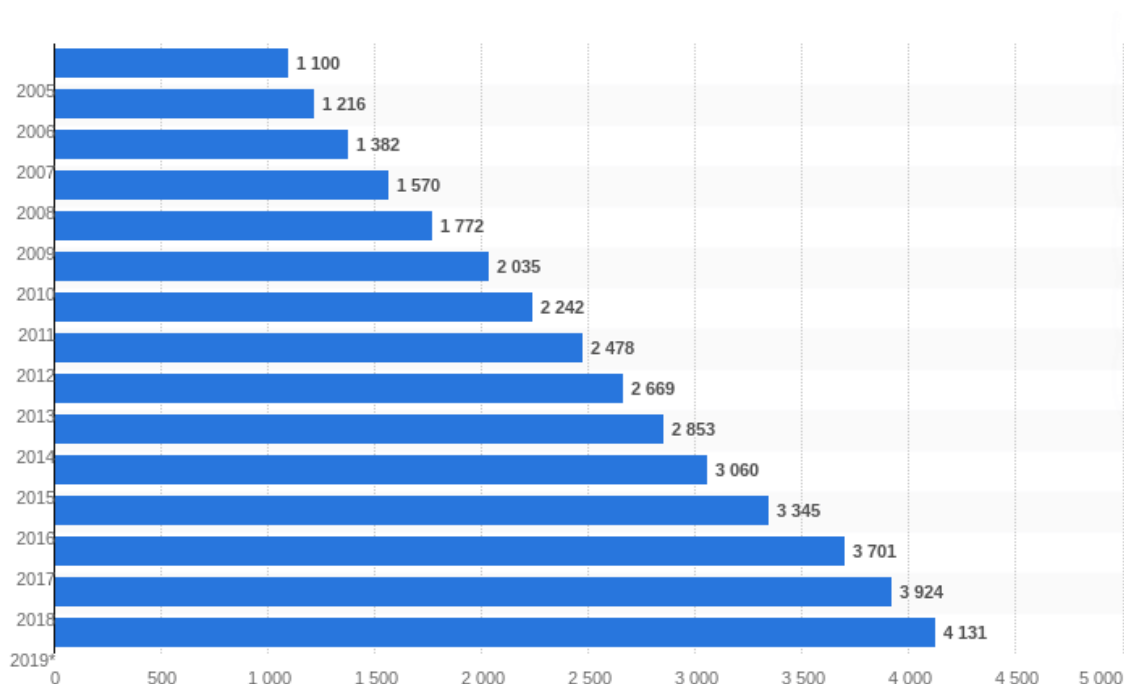
Ключевые слова: *журналистика, интернет-журналистика, сайт, СМИ, традиционная журналистика.*

Abstract. The article refers to internet journalism or online publication, which is a website registered as a mass media and aimed at performing relevant functions in accordance with the principles of journalism.

Keywords: *journalism, internet journalism, website, mass media, traditional journalism.*

Internet kundalik hayotimizning tobora muhim qismiga aylanib bormoqda. Ishonchimiz komilki, ko'plab insonlar kundalik hayotida jismoniydan ko'ra ekranlar orqali muloqot qilish uchun ko'proq vaqt sarflashadi. Bir narsa aniq: ish, dam olish, o'yin-kulgi va munosabatlarni o'z ichiga olgan WORLD WIDE WEBning roli kundan kunga oshsa oshadiki, lekin yo'qolmaydi. Eng muhim narsadan boshlaylik: Internetdan qancha odam foydalanadi? 2020-yilda 4,54 milliard faol foydalanuvchi, 4,18 milliard noyob mobil foydalanuvchi bo'lgan. Dunyoda ijtimoiy tarmoqlardan faol foydalanuvchilar ro'yxati 3,8 milliard kishiga yetdi va ularning deyarli har biri mobil qurilmalari orqali ijtimoiy tarmoqlardan foydalanadi – aniqrog'i esa – 3,75 milliard kishi. Sayyoramizning hozirgi aholisi 8,0 milliard kishini tashkil qiladi, shundan yer aholisining 56 foizi (shu jumladan bolalar) faol internet foydalanuvchilari hisoblanishadi. Bugungi kunda har 10 kishidan 6 nafari internetga kirish imkoniga ega [1]!

Internetning paydo bo'lishi va faoliyatga yo'naltirilishi ortidan dunyo global o'zgarishlar maydoniga aylandi, har bir yangilik va kashfiyotlarning tez suratlar bilan ommalashishi, teleradio va matbuotning mazkur tizimga integratsiya qilinishi ortidan katta hajmdagi axborotlar aylanmasi vujudga keldi.



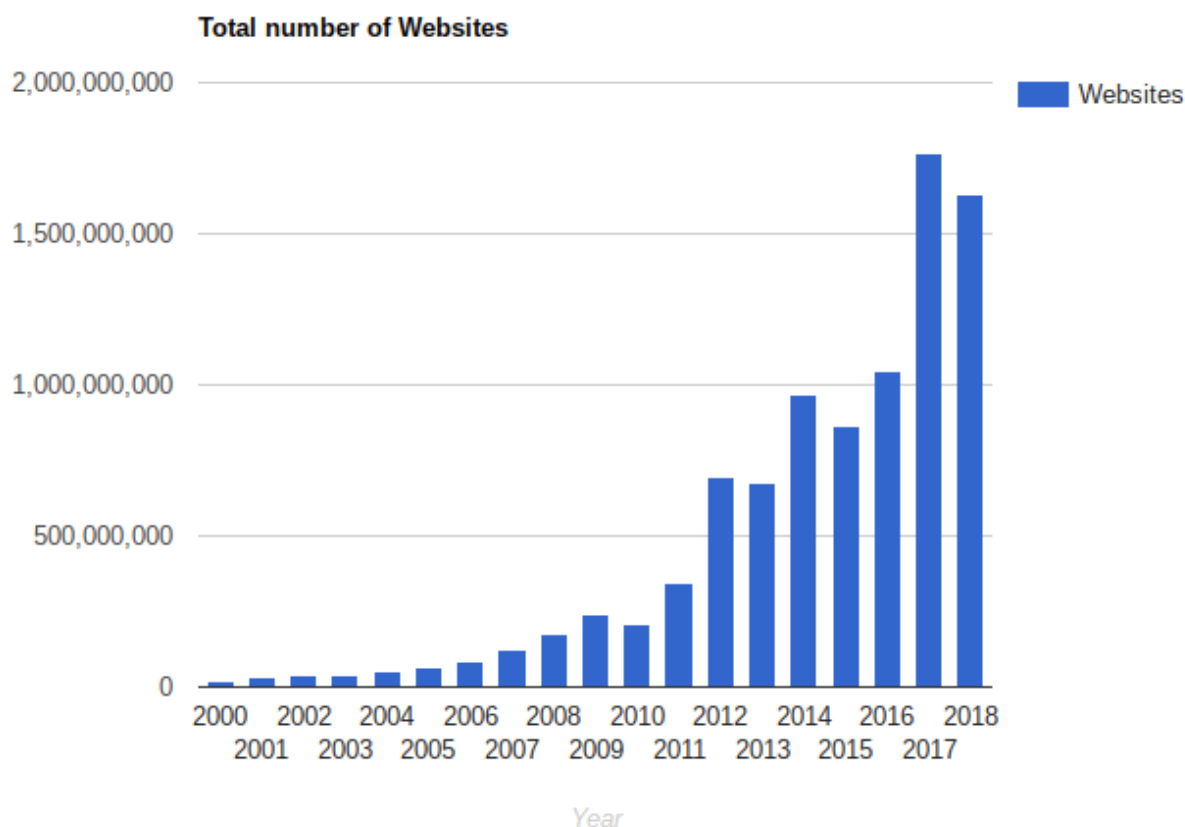
1-rasm. Dunyo miqyosida internetdan foydalanish

Keling, buni o'ylab ko'raylik: dunyo aholisining o'rtacha o'sishi hozirda 1% atrofida, Internetga kirish darajasi esa yiliga 8,2% ni tashkil qiladi. Joriy ko'rsatkichni 2017-yildagi 3,7 milliard foydalanuvchi bilan solishtiring. 2005-yilgacha sayyoramizda 1 milliarddan kam faol foydalanuvchi bo'lgan, ammo bu ko'rsatkichni ikki barobarga oshirib, 2 milliardga yetkazish uchun atigi 6 yil kerak bo'ldi. An'anaga ko'ra, Internetdan eng ko'p foydalanadigan mintaqasi Osiyo hisoblanadi (axir u dunyodagi eng ko'p aholi yashaydigan mintaqadir). 4,54 milliard internet foydalanuvchisidan 2,4 milliardga yaqini Xitoy va boshqa Osiyo mamlakatlarida yashaydi. Biroq, eng tez o'sib borayotgan foydalanuvchilar guruhi Afrikada bo'lib, o'tgan yillarda raqamli jihatdan ancha orqada qolgan ushbu qit'ada har 10 nafar yangi foydalanuvchidan 5 nafari ro'yxatdan o'tgan [1].

Bu qadar katta auditoriya ehtiyojini qondirish uchun zarur ma'lumotlar bazasini muomalaga chiqarish, bu uchun esa son-sanoqsiz saytlar, sahifalar va boshqa tarqatish vositalari zarur edi. Aynan ana shu talab ortidan vujudga kelgan internet saytlari bugun mamlakatimiz internet dunyosida ham ming-minglab tashkil etildi va faoliyati yo'lga qo'yildi. O'zbekistonda 2016-yil uchinchi choragiga kelib davlat rahbarining o'zgarishi va hukumatda olib borilgan oqilona islohotlar o'z natijasini ko'rsata boshladi. Raqamli jurnalistikaning unsurlarini o'zida mujassam etgan yangiliklar saytlari va onlayn nashrlar 2016-yil oxiriga kelib anchayin erkin so'z ayta boshladi. Buning misoli sifatida ularga qilinayotgan murojaatlar, tashriflar va kuzatishlar statistikasidagi dinamikaning ortishini ko'rishimiz mumkin.

Birinchisi 1991-yilda ingliz fizigi Tim Berners - Li tomonidan yaratilgan. 2020-yildagi eng so'nggi internet statistik ma'lumotlariga ko'ra, dunyoda 1,8

milliardga yaqin faol veb-saytlar mavjud va ularning soni har soniyada ortib bormoqda. Ajablanarlisi shundaki, hamma saytlarga ham odamlar tashrif buyurishmaydi. Hisobotlar shuni ko'rsatadiki, internet-trafikning yarmidan ko'pi robotlar tomonidan ishlab chiqariladi va bu ulush har yili ortib bormoqda [1]. Quydagi diagrammada aynan veb-saytlarning yillar kesimida o'sib borish holati tasvirlangan (2-rasm).

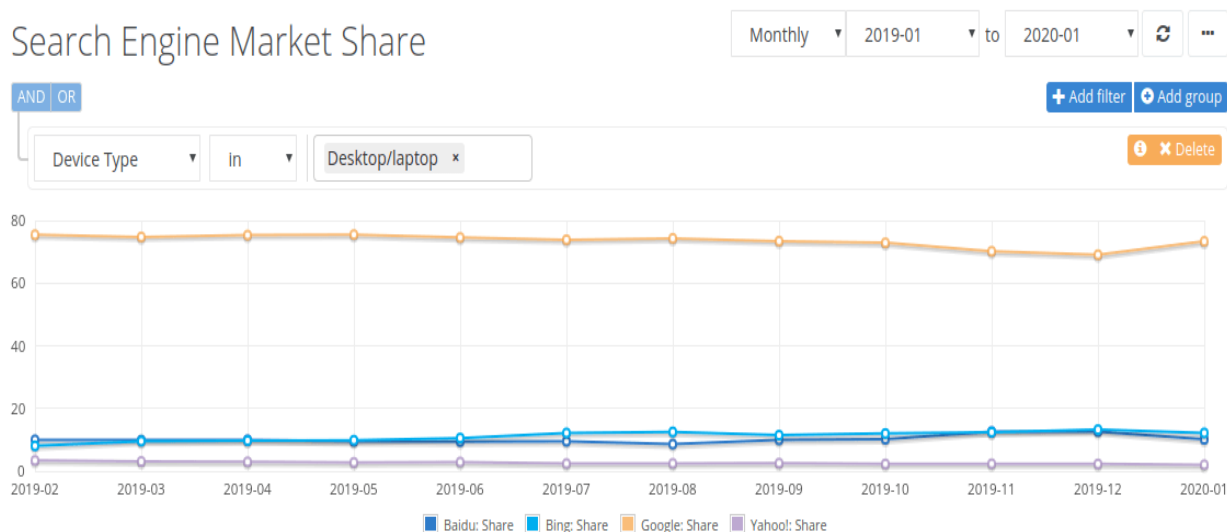


2-rasm. Veb-saytlarning yillar kesimida o'sib borish holati

Bugun oddiygina Internet saytlaridagi “Sport” sahifasi ham imkon darajada professional yondashilib yuritiladi. Sababi saytning har bir sahifasini va undagi ma'lumotlarni behuda narsa bilan to'ldirishdan qochish, berilayotgan ma'lumotlar va uning manbalaridagi aniqlik kuzatuvchilar e'tiboridan chetda qolmayapti. Hozir saytlarga tashrif buyuruvchi har bir inson, ayniqsa u qaysidir jamoa yoki sportchining ishqibozi bo'ladigan bo'lsa ularni aldab ham, yolg'on ma'lumotlar bilan ovuntirib ham bo'lmaydi. Negaki, hozirgi avlod vakillari o'ta sinchkov, har qanday ma'lumotni alternativ manbalardan tekshiruvchi, ochiq interenet manbalaridan foydalanishni yaxshi tushunadigan, o'zi uchun eng zarur ma'lumotlarni oladigan “Alfa” avlod vakillari, desak mubolag'a bo'lmaydi. Sababi Internet ommalashganiga o'ttiz yil bo'layotganini hisobga olsak, bugungi kunda oxirgi o'ttiz yillik avlod vakillari aynan Internet ichida tug'ildi va u bilan yashamoqda. Shu bois, beriladigan har qanday ma'lumotlarni aniq manba va vositalar asosida taklif etish, uni yengil

xazm bo'ladigan variantda taqdim etish, kuzatuvchilar reaksiyasini e'tibor markazida tutish, ular bildiradigan har qanday munosabatga tezkorlik va holislik bilan javob berib borish, veb-sayt obro'sini oshirish bilan bir qatorda uning kuzatuvchilari sonini ham ortishiga sabab bo'ladi. Shuningdek, foydalanuvchilarning mazkur saytlarni oson topishi, o'ziga kerakli vaqtda va makonda unga kirish uchun zarur qulayliklar yaratish, har bir saytni Internetda mavjud eng tezkor qidiruv tizimlariga integratsiya qilish va ulardan foydalanishda kuzatuvchilar madaniyatini shakllantirish ham eng muhim vosita hisoblanadi.

Biz har kuni ish, dam olish, o'yin-kulgi va o'zimizni rivojlantirish uchun ma'lumotlarga kirish uchun qidiruv tizimlaridan foydalanamiz. Agar siz taxmin qilmagan bo'lsangiz, Google qidiruv tizimlarining so'zsiz yetakchisi bo'lib, global qidiruv tizimlari bozorining 73 foizini egallaydi. Bozor ulushining qolgan qismi Bing, Baidu, Yahoo!, DuckDuckGo, Yandex va Ask o'rtasida taqsimlanadi. Google har kuni 3,5 milliarddan ortiq so'rovni qayta ishlaydi, bu soniyada taxminan 40 000 so'rovga teng. Bu kompaniya to'g'ridan-to'g'ri onlayn reklamadan 134 milliard dollardan ortiq daromad oladi [2]. Quyidagi jadvalda 2019-yil yanvar oyidan 2020-yil yanvar oyigacha dunyo bo'yicha qidiruv saytlariga qilingan murojaatlar va ularning statistikasi aks etgan (3-rasm).



3-rasm. Dunyo bo'yicha qidiruv saytlariga qilingan murojaatlar va ularning statistikasi

Tadqiqotimiz davomida mamlakatimizdagi top-10 talikka kiruvchi yangiliklar saytlariga bir kunda, bir haftada va bir oy davomida tashrif buyuruvchilar va ko'rishlar soni tahlilga tortildi. Tahlil jarayonida top-10 talikdagi saytlarning o'zni almashib turishi mumkinligini inobatga olib aynan mazkur kunlar davomida eng yuqori ko'rsatkichda turgan saytlar tahlil qilindi. Bu tahlillar 2022-yil dekabr oyi va 2023-yil yanvar oyi davomida amalga oshirilganligini alohida ta'kidlash lozim.

O'zbekistondagi yangiliklar veb-saytlari reytingida birinchi top-10 talikka kiruvchi veb-saytlarga bir kunda tashrif buyuruvchilar va ko'rishlar soni quyidagicha [3] (1-dadval):

O'rin	Resurs nomi	Tashrif buyuruvchilar	Ko'rishlar
1	Daryo — Новости Узбекистана и мира	299847	1708986
2	Qalampir – Правда бывает горькой	164517	429370
3	Aniq.uz – только самые точные новости!	55369	139880
4	Nuz.uz — Независимая газета «Новости Узбекистана»	40514	55764
5	Xabar.uz — Информационно-аналитический портал	33785	74144
6	«Даракчи» – Информационно-развлекательный портал	19087	31092
7	Sputnik Uzbekistan	16768	45962
8	Газета «Халқ сўзи»	12826	17600
9	Solaneews.uz	11296	18321
10	Azon.uz – новостной духовный портал	10251	16709
		664 260	2 537 828

1-jadval. O'zbekistondagi yangiliklar veb-saytlari reytingida birinchi top-10 talikka kiruvchi veb-saytlarga bir kunda tashrif buyuruvchilar va ko'rishlar soni

O'zbekistondagi yangiliklar veb-saytlari reytingida birinchi top-10 talikka kiruvchi veb-saytlarga bir haftada tashrif buyuruvchilar va ko'rishlar soni quyidagicha [4] (2-jadval):

O'rin	Resurs nomi	Tashrif buyuruvchilar	Ko'rishlar
1	Daryo — Ўзбекистон ва жаҳон янгиликлари	1 702 426	11 447 861
2	Qalampir - Haqiqat achchiq bo'ladi	1 126 857	3 317 342
3	Aniq.uz - qiziq, qisqa va aniq xabarlar!	215 318	708 248
4	Ўзбекистон Янгиликлари	203 378	350 605
5	«Xabar.uz» axborot-tahliliy portali	179 556	449 892
6	Axborot-ko'ngilochar "Darakchi" portali	99 222	183 133
7	Sputnik Ўзбекистон	96 895	281 386
8	«Халқ сўзи» газетаси	72 687	102 447
9	muallimlar.uz	70 840	106 109
10	ЎЗА - Ўзбекистон Миллий ахборот агентлиги	61 541	157 662
Umumiy		3 828 720	17 104 685

2-jadval. O'zbekistondagi yangiliklar veb-saytlari reytingida birinchi top-10 talikka kiruvchi veb-saytlarga bir haftada tashrif buyuruvchilar va ko'rishlar soni

O'zbekistondagi yangiliklar veb-saytlari reytingida birinchi top-10 talikka kiruvchi veb-saytlarga bir oy davomida tashrif buyuruvchilar va ko'rishlar soni quyidagicha [5] (3-jadval):

O'rin	Resurs nomi	Tashrif buyuruvchilar	Ko'rishlar
1	Daryo — Ўзбекистон ва жаҳон янгиликлари	3 957 076	26 600 577
2	Qalampir - Haqiqat achchiq bo'ladi	2 201 997	7 064 885
3	«Xabar.uz» axborot-tahliliy portali	397 816	1 001 331
4	Aniq.uz - qiziq, qisqa va aniq xabarlar!	367 772	1 330 950
5	Ўзбекистон Янгиликлари	299 976	465 699
6	Axborot-ko'ngilochar "Darakchi" portali	247 109	424 510
7	Sputnik Ўзбекистон	228 681	645 702
8	«Халқ сўзи» газетаси	211 487	312 340
9	Bugun yangiliklar: O'zbekiston va dunyodagi eng so'nggi yangiliklar — Solanews.uz	193 975	451 198
10	Azon.uz - axborot-tahliliy portal	156 987	276 337
Umumiy		8 262 876	38 573 353

3-jadval. O'zbekistondagi yangiliklar veb-saytlari reytingida birinchi top-10 talikka kiruvchi veb-saytlarga bir oyda tashrif buyuruvchilar va ko'rishlar soni

Tadqiqotimizda mazkur yo'nalishdagi tahlillar www.uz tizimida mavjud statistikaga asoslanib o'rganildi. O'zbekistonda www.uz tizimi ochiq holda barcha saytlar statistikasini yuritadi. Shu orqali resurslar ommabopligi baholanadi. Har bir saytda www.uz tizimining hisoblagichi o'rnatiladi va ularning reytingi shakllantiriladi. Ushbu tahlil natijalariga ko'ra, shuni aniqladikki, mamlakatimizda birgina yangiliklar saytlarining top-10 taligiga kiruvchi saytlarga bir oy davomida tashrif buyuruvchilari soni 8 262 876 nafarni, sahifalar bo'ylab ko'rishlar soni esa 38 573 353 marotabani tashkil etadi. Bugungi kunda mamlakatimizda aynan yangiliklar saytlarining soni 184 tani tashkil etishini, biz faqat reytingdagi birinchi top-10 talikdagi saytlarnigina tahlilga tortganimizni inobatga oladigan bo'lsak, bu ko'rsatkichlar yanada ortishi muqarrar.

Foydalanilgan adabiyotlar:

1. <https://sdvv.ru/articles/elektronnaya-kommertsiya/statistika-interneta-2020> 23.01.2023.
2. <https://sdvv.ru/articles/elektronnaya-kommertsiya/statistika-interneta-2020> 24.01.2023.
3. https://www.uz/uz?cat_id=359 21.01.2023.



4. https://www.uz/uz/stat/visitors/ratings?cat_id=359&cat_id=359&VisitorRating%5Bp 21.01.2023.
5. <https://www.uz/uz/stat/visitors/ratings?VisitorRating%5Bperiod%5D=monthly&VisitorRating> 21.01.2023.

KOMPYUTER LEKSIKOGRAFIYASI. MULTIMEDIA VA LUG‘AT

Qurbonova Roza

ToshDO‘TAU 1-bosqich talabasi

Annotatsiya. Maqolada kompyuter leksikografiyasi haqida, uni takomillashtirish lug‘atchilikni rivojlantirish haqida, shuningdek, multimedia bilan lug‘atni biriktirib, multi-lug‘at yaratish haqida fikr yuritiladi.

Kalit so‘zlar: *kompyuter leksikografiyasi, multimedia, ma‘lumotlar ombori, leksikografiya, kompyuter leksikografiyasi, kompyuter lug‘atlari, kitobiy lug‘atlar, online lug‘atlar, elektron lug‘atlar.*

Abstract. The article discusses computer lexicography, its improvement, the development of lexicography, as well as the creation of a multi-dictionary by combining a dictionary with multimedia.

Keywords: *Computer lexicography, multimedia, data warehouse, lexicography, computer lexicography, computer dictionaries, book dictionaries, online dictionaries, electronic dictionaries.*

Абстракт. В статье рассматривается компьютерная лексикография, ее совершенствование, развитие лексикографии, а также создание мультисловаря путем объединения словаря с мультимедиа.

Ключевые слова: *компьютерная лексикография, мультимедиа, база данных, лексикография, компьютерная лексикография, компьютерные словари, книжные словари, онлайн-словари, электронные словари.*

Kompyuter leksikografiyasi - tilshunoslikning muayyan til leksikonidagi so‘zlarning ma‘lumotlar bazasi va programmalashtirish tillari yordamida ishlash tizimiga ega bo‘lgan dasturlar yaratish bilan shug‘illanuvchi sohasi. [1]

Leksikografik asar odamlarning ma‘lum bir taraqqiyot darajasining guvohidir, u jamiyatdagi ishlab chiqarish munosabatlarining holati va darajasini belgilaydi, lug‘atda ta‘riflangan til tashuvchisi – etnosning falsafiy, siyosiy, diniy, ilmiy-texnik fikri rivojlanadi. Shu bilan birga, lug‘atning o‘zi ma‘lum bir tarixiy rivojlanish davrida ma‘lum jamiyatning ilm-fan yutug‘i bo‘lgan madaniyat asari kabi ko‘rinadi. Yaxshi lug‘at san‘at yoki musiqa asarlari singari inson ruhini yaratish, vahiy sifatida xalq tarixiga kiritilgan. Har qanday loyihasini amalga oshirishga kirishishdan oldin uzoq vaqt davomida tayyorgarlik ishlari va tajribalarni o‘tkazish kerak. P.N.Denisov lug‘at tayyorlashda quyidagilarga aniqlik kiritish kerakligini ta‘kidlagan:

- 1) uning hajmi;
- 2) manbalar;
- 3) lug‘atning lingvistik kredosi. [1]

Matnning formal ajraladigan eng kichik birligi soʻz hisoblanadi, lekin u maʼno anglatuvchi eng kichik birlik morfemalardan tashkil topadi. Soʻz tarkibi oʻzak morfemalar, prefikslar (old qoʻshimchalar) va suffikslar (oʻzakdan keyin keladigan qoʻshimchalar)dan iborat boʻladi.[1]

Leksikografiya tilshunoslik fanining amaliy sohalaridan biri hisoblanadi. U yunoncha «lexikos» - «soʻz», «lugʻat», «grapho» - “yozaman” degan maʼnolarni anglatadi. Leksikografiya boʻlimida lugʻatlar, ularni tuzish yoʻllari va tamoyillari oʻrganiladi.

“Lugʻat” arabcha “til”, “sheva”, “lahja”, “soʻz”, “ibora” kabi maʼnoni bildiradi. Hozirgi kunda ushbu soʻz ikki maʼnoda: 1) muayyan tilda, uning hududiy yoki ijtimoiy lahjasida mavjud boʻlgan, shuningdek, u yoki bu yozuvchi asarlarida uchraydigan soʻzlar yigʻindisi, yaʼni leksika; 2) soʻzlar (yoki morfema, soʻz birikmasi, ibora) muayyan tartibda (alifboli, uyali, mavzuli) joylashtirilgan, tavsiflanuvchi birliklar, ularning kelib chiqishi, maʼnolari, yozilishi (imlosi), talaffuzi, uslubiy mansubligi, boshqa tillarga tarjimai haqida maʼlumotlar jamlangan kitob maʼnolarida qoʻllanadi. [1]

Avtomatizatsiya ijtimoiy hayotning barcha jabhalariga kirib kelishi natijasida lugʻatlarni elektron holatda yaratish imkoniyati yuzaga keldi. Kompyuter yordamida lugʻatlar bilan ishlashning optimallasuvi natijasida kompyuter leksikografiyasi yoʻnalishi shakllandi. Kompyuter leksikografiyasi amaliy tilshunoslikning muhim tarkibiy qismi boʻlib, unda lingvistik va dasturiy taʼminot asosida ishlaydigan, kodlash va dekodlash prinsipi asosida yaratilgan kompyuter lugʻatlari, ularni tuzish dasturlari, algoritmlari oʻrganiladi. Mazkur sohaning rivojlanishi natijasida turli nomlarda elektron lugʻatlar yaratilmoqda. Leksikografiya sohasiga kompyuter texnologiyalarining tatbiq etilishi lugʻat tuzish ishlarini ancha yengillashtiradi.

Kitobiy lugʻatlar va elektron lugʻatlarning farqi quyidagilarda koʻrinadi:

1. Kitobiy lugʻatlar kartotekalar asosida tuziladi. Elektron lugʻatlar (elektron kartotekalar) asosida yaratiladi.

2. Kitobiy lugʻatlar qoʻlyozma holatda, kompyuter lugʻatlari elektron variantda tuziladi.

3. Kitobiy lugʻatlar qoʻlyozma holida tahrir qilinadi; elektron lugʻatlar nisbatan tayyor lingvistik baza asosida yaratiladi.

4. Kitobiy lugʻatlar bosma holida nashr etiladi; elektron lugʻatlarda lingvistik taʼminot dasturiy taʼminot bilan uygʻunlashtiriladi.

5. Kitobiy lugʻatlar tuzilishi sahifalar ketma-ketligiga tayanadigan chiziqlilik tamoyiliga boʻysunadi. Elektron lugʻatlar strukturasi gipertekst texnologiyasiga asoslangan boʻladi, bu esa foydalanuvchiga lugʻat maqolalarining ixtiyoriy qismiga tezkor murojaat qilish imkonini beradi.

Elektron lugʻatlarda har bir soʻzga mutanosib keluvchi kod ishlab chiqiladi va qoʻlanadi, kodni qayta ishlash jarayonida zaruriy boʻlgan maʼlumotlar, tarjimalar, sinonim, antonim va sharhlarga ega boʻlish mumkin.

Soʻzlarni kodlashtirish quyidagicha amalga oshiriladi. Maʼlumotlar fayllarga joylashtiriladi, uning har bir elementi 3 qismdan iborat boʻladi: 1) soʻzning tartib raqami; 2) soʻz; 3) kod. «Soʻzning tartib raqami» (yaʼni uning adresi). Soʻzlarning tarjimai, sinonimlari va antonimlarini koʻrsatish ularning birinchi harflarini kodlashdagi tartib raqamlari bilan birgalikda keltirish bilan amalga oshiriladi.

«Soʻz» - alfavit harflari bilan yozilgan oddiy soʻz.

«Kod» - raqam va harflar ketma-ketligi boʻlib, unda soʻzning barcha zaruriy morfologik, sintaktik, leksik xususiyatlari haqidagi maʼlumotlar jamlangan boʻladi: grammatik maʼlumot, adreslar, tarjima, sinonim, antonim, mutanosib sharhlar.

Chastotali lugʻatlar matnda soʻzlarning qoʻllanish chastotasi (qanday miqdorda ishlatilishi)dan kelib chiqqan holda tayyorlanadigan soʻzliklar hisoblanadi. Chastotali lugʻatlar statistik leksikografiyaning rivoji natijasida yuzaga kelgan. Bunday lugʻatlar, ayniqsa, til oʻqitish jarayonida leksik minimumlarni aniqlashda muhim ahamiyatga ega. Soʻzlarning chastotasini aniqlash lingvostatistik tahlillar yordamida amalga oshiriladi. Chastotali lugʻatlar kompyuterning lingvistik informatsion bazasi hisoblanadi.

Til – kishilar orasidagi eng muhim aloqa vositasi, tafakkur quroli. Har bir xalq tilida uning tarixi, xarakteri, badiiy ijodi va tafakkurining oʻziga xos xususiyatlari oʻz ifodasini topadi.

Kompyuter lingvistikasi fani 1960-yilda AQShda shakllandi. Mazkur fan XX asrning oʻrtalarida paydo boʻlgan matematik lingvistika fani asoslaridan kelib chiqqan. Tabiiy tillar grammatikasi formal modeli kompyuterda tarjima qilish, tillarga oʻrgatish, bilimlarni diagnostika qilish, matnlarni tahrir etish kabi amaliy dasturlarni ishlab chiqishga asos boʻlgani holda kompyuter lingvistikasi yoʻnalishiga zamin yaratdi.

Kompyuter lingvistikasi fanining asosiy maqsadi lingvistik masalalarni hal qilishning kompyuter dasturini yaratishdir. Uning asosiy vazifalariga tillarga oʻqitish, bilimlarni tekshirish, matnlarni turli jihatdan tahrirlash va mashina tarjimai uchun moʻljallangan dasturlarni ishlab chiqish kabilar kiradi. Kompyuter lingvistikasi bir-biri bilan chambarchas bogʻliq boʻlgan uch yoʻnalishni oʻz ichiga oladi: 1) lingvostatistika; 2) injener lingvistikasi; 3) algoritmik (dasturlash) lingvistika.

Mening fikrim lugʻatchilikda, yaʼni lugʻat tuzishda (ayniqsa elektron lugʻatlarda) rasmlardan foydalanishni ilgari surish toʻgʻri boʻladi.

1.Yosh bolalar uchun bu qiziqarli.

2.Yoshi kattalarda kundalik hayotda uchraydigam muammolardan biri ularning lotin alifbosiga asoslangan oʻzbek tilini birdan anglay olmasligi, tushunmasligi, ularga kirill alifbosidagi oʻzbek tili tushunarli va aniq boʻlganligi sababli, rasmiy lugʻat kattalar uchun ham tushunarli va foydali.

3.Tarjima lugʻatlarda faqat chet tilini oʻqib, yodlanadi, tarjimai rasmda ayon boʻladi.

4. Tafakkurni rivojlantiradi va mantiqni shakllantiradi.

Multimedia (inlizcha multimedia atamasi, lotincha multum – ko‘p, medium – vosita, muhit so‘zlaridan olingan) – texnik va dasturiy vositalar majmuasi bo‘lib, foydalanuvchilarga turli xil ma‘lumotlar (matn, grafika, ovoz, video) ning bitta yagona axborot muhitida birlashtirgan holda muloqot rejimida ishlash imkoniyatini beradi.

Multimedia – bu harakatsiz tasvir, harakatli video, kompyuter grafikasi va matn animatsiyalari, nutq va yuqori sifatli ovoz kabilar bilan ishlash imkonini beruvchi interfaol (muloqotli) tizim, boshqacha aytganda, matn, grafika, animatsiya, ovoz, video, nutq kabi ma‘lumotlarni kompyuter xotirasiga kiritish, qayta ishlash, saqlash, uzatish va namoyish etish imkoniyatini beruvchi texnologiyalar yig‘indisi.

Multimedianing rivojlanish tarixi.

Multimedia-texnologiyalari informatikaning eng dolzarb va mashxur yo‘nalishlaridan biri hisoblanadi. Bu yo‘nalishning asosiy maqsadi “tasvir, tekst va matn ovoz, video animatsiya va boshqa tasviriylar effektlar (Simulation) kolleksiyasiga ega bo‘lgan, hamda interaktiv va boshqa vizual boshqaruv mexanizmiga ega bo‘lgan” mahsulotlarni yaratishdan iborat. Bu tushuncha 1988-yil yangi texnologiyalarni tatbiq qilish va foydalanish muammolari bilan shug‘ullanuvchi yirik Yevropa komissiyasi tomonidan bayon qilingan.

80 – yillar oxirida multimedia – texnologiyalarini gumanitar sohalarda qo‘llashga bo‘lgan qiziqishlarni oshishi, bevosiat atoqli amerikalik kompyuter – biznesmeni Bill Geyts nomi bilan bog‘liq. U turli “muhitli”: tasvir, ovoz, animatsiya va gipermatnli sistema asosidagi multimedia mahsulotlarini yaratish va amaliyotda muvaffaqiyatli qo‘llash g‘oyasini ilgari surgan (“National Art Gallery. London”).

Aynan mana shu mahsulot o‘zida multimedianing quyida keltirilgan 3 ta asosiy prinsipi mujassamlashtirib olgan edi:

1. Ma‘lumotlarni inson sezuvchi turli ma‘lumot muhitlari kombinatsiyasi yordamida tasvirlash.

2. Mahsulotdagi ma‘lumotlarni tasvirlashni turli xil liniyalarining mavjudligi (xususan unda ma‘lumotni tasvirlash liniyasini mahsulotda mavjud bo‘lgan boshqaruv mexanizmi orqali insonning o‘zi belgilaydi).

3. Interfeysning badiiy dizayni va navigatsiya vositalari.

Lug‘at tuzish ilmini o‘rgandik, multimediani o‘rgandik, quyidagi misollardan ko‘raylik multi-lug‘at qanday shakllanadi va rivojlanadi.

1. Asosiy so‘z: Bus

2. audio (qanday o‘qilishi)

3. rasm (harakatli, signal ovozlari bilan ishlanadi.)

4. o‘zbek tilida qanday ayilishi audiosi ham (yosh bolalar uchun).



5. berilgan so‘zga aloqador joylar, hodisalar va so‘zlar rasmi beriladi.



Bus stop



bus station

6. gap yoziladi masalan:



I'm waiting for the bus at the bus stop.

7. audio (gap o'qiladi, bolalar uchun bo'lsa o'zbek).

8. bu lug'atda tasvirlar harakatga keladi, avtobus, yurib saylgohga boradi, odamlar uni kutadi va unga chiqadi, harakatli tasvirlar, audiolarga boy gaplardan tashkil topgan paydo bo'ladi natijada.

Foydalanilgan adabiyotlar:

1. B.Mengliyev, Sh.Xamroyeva, D.Axmedova “Kompyuter leksikografiyasi”, O'quv qo'llanma, Toshkent – 2022 (13-bet).
2. M.M.Nafasov, F.R.Murodova, Z.R.Murodova, I.I.Hayitova, “Ta'limda multimedia ilovalari” Buxoro, 2019, 7-bet.
3. S.Normamatov “Leksikografiya asoslari”, Toshkent – 2020.
4. https://uz.wikipedia.org/wiki/Kompyuter_leksikografiyasi
5. Xolmanova Z., Kompyuter lingvistikasi, Toshkent – 2020 “Asian Book House” (57-bet)
6. www.savol-javob.com

O'ZBEK TILI UCHUN STEMMERNI ISHLAB CHIQISH

Aloyev Narzillo Rahmatullayevich

ToshDO'TAU tayanch doktoranti,

vip.alayev@gmail.com

Xuydayberganov Nizomaddin Uktambay o'g'li

ToshDO'TAU

Kompyuter lingvistikasi va raqamli texnologiyalar

kafedrası stajor-o'qituvchisi

nizomaddin@navoiy-uni.uz

Annotatsiya. Ushbu maqolada o'zbek tili leksik birliklarini axborot qidirish tizimlaridan izlashga mo'ljallangan stemmerni ishlab chiqish usullari taqdim etiladi. U chiziqli hisoblash murakkabligiga ega bo'lib, uning sinov muvaffaqiyatining koeffitsienti 96,1%ni tashkil qiladi. Ushbu maqolada o'zak shakllanishidagi ehtimollik ko'rinishining batafsil tavsifi beriladi.

Kalit so'zlar: *stemmer, NLP, agglyutinativ til, stem, morfologik analizatorlar, IR, ma'lumot olish tizimlari, o'zbek stemmeri modeli.*

Abstract. This article presents the methods of developing a stemmer for searching Uzbek lexical units from information search systems. It has linear computational complexity, and its trial success rate is 96.1%. This article provides a detailed description of the probability view in core formation.

Key words: *stemmer, NLP, agglutinative language, stem, morphological analyzers, IR, information retrieval systems, Uzbek stemmer model.*

Аннотация. В данной статье представлены методы разработки стеммеров для поиска узбекских лексических единиц из информационно-поисковых систем. Он имеет линейную вычислительную сложность, а показатель успешных тестов составляет 96,1%. В данной статье дается подробное описание появления вероятности при формировании ядра.

Ключевые слова: *стеммер, НЛП, агглютинативный язык, основа, морфологические анализаторы, ИР, информационно-поисковые системы, узбекская стеммерная модель.*

Kirish

Ingliz tili kabi tillarda stemlash jarayoni nisbatan oddiyroq kechadi. Buning sababi esa so'z shakllarining morfologik o'zgarishlari cheklangan bo'ladi. Boshqa tomondan esa, o'zbek tili kabi agglyutinativ tillarda stemlash hamon jiddiy muammo bo'lib qolmoqda, chunki ular nazariy jihatdan cheksiz mavjud so'z shakllarini hosil qilish imkoniyatiga ega. Morfologik analizatorlar agglyutinativ tillarda ma'lumot qidirish tizimlari uchun stemmer sifatida yagona aniq vosita hisoblanadi [1,2]. Ammo bu holda, agglyutinativ tillarda IR tizimlari uchun stemlash, qarama-qarshi talablar juftligi bilan aniqlanishi mumkin bo'lgan paradoksni keltirib chiqaradi: IR

tizimlaridan agglyutinativ tillar uchun xotiraga saqlash murakkabligini bartaraf etish va ish faoliyatini yaxshilash uchun stemlashni talab qiladi, lekin, morfologik analizatorlardan stemmer sifatida foydalanishda yuzaga keladigan hisoblash murakkabligining yuqori darajasi tufayli IR tizimlarining umumiy ishlash koeffitsienti pasayadi. Biz taklif qilayotgan stemlash usuli, ushbu qiyinchilikni samarali yengish uchun ishlatilishi mumkin.

Ushbu maqolada o'zbek IR tizimlari uchun taqdim etilgan statistik muhitga asoslangan ehtimollik stemlash modelini taqdim etiladi. Shuningdek, taklif qilingan model tahlillar va tajriba natijalari bilan tasdiqlanadi. Shunday qilib, yangi stemmer **O(n)** hisoblash murakkabligiga ega bo'lgan holda yuqorida ko'rsatilgan to'siqlarni yengib o'tadi [3].

Stemlash va ma'lumot olish

IR (ma'lumot olish) tizimlari katta hajmdagi elektron hujjatlardan to'plangan ma'lumotlarni qayta ishlash uchun ishlatiladi. Hujjat haqidagi ma'lumotlar asosan so'zlarning semantikasidan iborat bo'ladi. Demak, IR tizimlari haqiqatdan semantika vakillari hisoblangan so'zlarni boshqaradi/saqlaydi. Hujjatdagi so'z matn ketma-ketligidagi grammatik qo'llanishiga qarab turli morfologik shakllarga ega bo'ladi, lekin o'zagi bilan ifodalangan semantika o'zgarishsiz qoladi. Shuning uchun, IR tizimlari odatda xotiraga saqlash muammolarini bartaraf etish va ish faoliyatini oshirish uchun so'z shakllari o'rniga stemlardan foydalanadi.

Lovins tomonidan o'zak ta'rifi “bir xil o'zakli barcha so'zlarni umumiy shaklga qisqartirish tartibi, odatda har bir so'zdan uning hosilaviy va flektiv qo'shimchalarini olib tashlash” orqali berilgan [4]. Ingliz tili kabi analitik tillarda IRni aniqlash agglyutinativ tillarga qaraganda ancha oson kechadi. Misol tariqasida, Porterning ingliz tili uchun algoritmi cheklangan miqdordagi kaskadli qayta yozish qoidalariga asoslanadi va deyarli barcha so'z shakllarini atigi **1200** ga yaqin morfologik o'zgarishlarni tugatish orqali qabul qilishi mumkin. Lekin o'zbek tilda agglyutinativ til sifatida, hatto taxminan **23000** ta stemlar bo'lsa ham; nazariy jihatdan cheksiz turli xil so'z shakllari yaratilishi mumkin [5].

Morfologik murakkablikning ushbu darajasida lug'atni o'zgartirish modellari va o'rnatilgan qoidalarni ifodalaydigan morfologik analizatorlari, morfologik tahlil qilish uchun qabul qilingan standart hisoblanadi. Ammo bu morfologik analizatorlar ikki darajali til modeliga asoslanadi. IR tizimlari uchun ikkita to'siq mavjud: Hisoblashning murakkabligi morfologik analizatoridan stemmer sifatida foydalanishdan kelib chiqadi va aksincha, stemmersiz, ya'ni, saqlash murakkabligi bir xil semantik, ammo morfologiyasi boshqacha bo'lgan so'zning har bir turli shaklini indekslashdan kelib chiqadi [3,4].

Shu sababli, xotiraga saqlash hajmini boshqara olish va qoniqarli ish darajasiga erishish uchun agglyutinativ tillar uchun IR tizimlaridan stemlash talab qilinadi. Biroq, o'z navbatida, IR da samarali foydalanish uchun chiziqli hisoblash murakkabligi, algoritmlarni stemlash uchun muhimdir. Agglyutinativ tillarda stemming zarurligini isbotlash uchun Porter stemmeri ko'plab inglizcha so'zlarni o'z

ichiga olgan **newswire** ma'lumotlar bazasiga qo'llanilib ma'lum bir siqish darajasiga erishildi va shuningdek, 400000 dan ortiq o'zbekcha so'zni o'z ichiga olgan (kun.uz satidan olingan) siyosiy yangiliklar ma'lumotlar bazasiga qo'llanilgan stemmer 1-jadvalda o'z aksini topgan.

1-jadval. O'zbek tilidagi matnlarni stemmlash ko'rsatkichlari

korpus	so'z tokenlari	unikal terminlar	siqish (%)
O'zbek	422,112	32,370	87,3
inglizcha	650,192	18,384	38,1

Yuqoridagi 1-jadvaldan ko'rinib turganidek, ingliz tilidagi ma'lumotlar bazasida ko'proq so'z tokenlari mavjud bo'lishiga qaramay; o'zbek tilidagi alohida atamalar ingliz tilidan ko'ra ko'proq. Bu o'zbek tilidagi so'zlarning xilma-xilligini ko'rsatadi hamda, bu siqish nisbati natijalari bilan ham tasdiqlangan. 38,1% siqish darajasi 87,3% dan ancha kichik, bu esa ingliz tili uchun morfologik o'zgarishlarning past darajasini aniq ko'rsatadi. Biroq, o'zbek tili ma'lumotlar bazasi uchun kuzatilgan **87,3%** siqish nisbati esa qanchalik stemlash zarurligini ta'kidlab turadi.

Stemlash jarayoni bog'liq masalalar

IR maqsadlari uchun foydalaniladigan stemlash algoritmlarini to'rtta guruhga bo'lish mumkin:

- *table lookup*;
- *successor variety*;
- *n-gram*;
- *affix removal*.

Table Lookup algoritmlarida, indekslangan terminlar sifatida ishlatilishi mumkin bo'lgan barcha atamalar va ularning asoslari jadvalda saqlanadi. Shunday qilib, so'rovlar va hujjatlardan atamalarni **juda tez stemlash mumkin**. Ammo amalda ingliz tili uchun bunday unikal terminlar jadvali mavjud emas va boshqa tabiiy tillar uchun esa ba'zi indeks atamalar domen yoki kontekstga bog'liq bo'ladi. Shuningdek, shu tartibdagi jadvaldagi saqlashga ketadigan resurslar muammoligicha qolaveradi.

Successor Variety algoritmlari asosida olingan natijalar katta qismidan kuzatilgan fonemalarning taqsimlanishi bilan birga so'z yoki morfema chegaralarini aniqlashga asoslangandir.

N-gramm asosidagi stemmer so'rov va hujjat o'rtaidagi bo'shliq shartlari o'xshashlik o'lchovidan foydalanadi. Uzunligi m bo'lgan atama $m-n+1$ n -grammga ega. Har bir shart juftligi uchun har ikkala shartning yagona n -grammasidan va o'xshashlik matritsasidan hisoblangan koeffitsientlar tuziladi. O'xshashlik matritsasi tuzilgandan so'ng, atamalar yagona bo'g'inli klasterlash usuli yordamida klasterlanadi.

Affikslarni ajratishga asoslangan stemmerlarda stendlarni aniqlash uchun atamalardan qo'shimchalar va/yoki prefikslarni olib tashlaydigan algoritm mavjud. Yuqorida eslatib o'tilganidek, Porterning stemmeri affikslarni olib tashlash algoritmlariga yaxshi misol bo'la oladi. To'rtta stemmerlar guruhining barchasi ingliz tili kabi analitik tillar yoki fransuz va nemis kabi ba'zi Hind-Yevropa tillari uchun ishlab chiqilgan hisoblanadi. Ular turkiy, turkman, ozarbayjon yoki azeri, qashqay va o'g'uz tillarini ham o'z ichiga olgan janubi-g'arbiy yoki o'g'uz guruhidagi turkiy agglyutinatativ tillar kabi inflektiv paradigmaga ega bo'lgan tillarda yuqori darajadagi morfologiyaga dosh bera olmaydi.

Shuningdek, o'zbek tili uchun ishlab chiqilgan bir qator stendlash algoritmlari mavjud bo'lib, ular morfemalarni belgilash va aniqlash kabi ba'zi usullarni kerakli aniqlik darajasiga tushirish orqali morfologik analizatorni IR tizimiga moslashtirishga qaratilgan harakat bo'lib hisoblanadi [6]. Ushbu turdagi algoritmlar oldindan tuzilgan elektron leksikondan ehtimoliy o'zakni izlaydilar, so'ngra esa, o'zbek tilining morfotaktikasiga ko'ra qo'shimchalarning birikmalarini qidirib, so'z shaklining qolgan qismini aniq qo'shimchalar qatoriga moslashtirish orqali uning to'g'riligini tekshiradi. Chunki ularning tayanch tushunchasi morfologik tahlillar bo'lib hisoblanadi. Tekshirishning keyingi bosqichining qat'iylik darajasi ham aniqlikni, ham, ayniqsa, hisoblash murakkabligini belgilaydi; yuqori darajadagi tekshirish qat'iyligi bilan, aniqlik yuqori bo'ladi, lekin hisoblash murakkabligi NP-murakkabligiga yaqin bo'ladi va aksincha. Shuning uchun, bu stemmerlar uchun hisoblash murakkabligi darajasi chiziqli nuqtadan NP-Hardga yaqin nuqttagacha, ular foydalanadigan morfologik bilim darajasi qanchalik oshsa shunga qadar yana oshadi [7,8].

O'zbek tilidagi stemmerni ishlab chiqish maqsadida turkiy tillar oilasiga mansub tillarda olib borilgan ilmiy tadqiqotlar o'rganib chiqildi. Jumladan, turk tilidagi stendlash muammolariga birinchilardan bo'lib Köksal tomonidan yechim topilishi harakati boshlangan [9]. Bu nolingvistik nuqtayi nazaridan olib qaraganda boshqalardan butunlay farq qiladi. U so'z shaklining o'zagi sifatida boshidanoq so'z shaklining o'zgarish uzunlikdagi qismini qabul qiladigan stemmerni ishlab chiqdi. Dastlab, Köksalning ta'kidlashicha, 5 ta stem uzunligi sinovlarida eng yaxshi natija beradi. Ammo amalda barcha hujjatlar to'plami uchun umumiy belgilangan uzunlik mavjud emas. Muayyan mavzu bo'yicha ma'lum bir hujjat to'plami bilan yaxshi ishlaydigan qiymat boshqa mavzu ostidagi hujjatlar to'plami bilan ishlaydigan IR tizimi ish unumdorligini kamayishiga olib kelishi mumkin.

Taklif etilayotgan modelimizda biz ilgari tahlil qilingan matn ma'lumotlar bazasidan to'plangan umumiy morfologik bilimlar statistikasidan modelni takomillashtirish uchun pragmatik sifatida foydalanamiz. Oldindan hisoblangan statistik ma'lumotlardan foydalangan holda, biz chiziqli hisoblash murakkabligining saqlanishini ham ta'minlashga erishdik.

O'zbek stemmeri modeli

Har qanday agglyutinativ tildagi soʻz shaklini $S_n = h_1 h_2 \dots h_n$ boʻlgan satr orqali bilan belgilash mumkin. Bunda har bir h_i ($i = 1, 2, \dots, n$) tegishli A alifbosining elementi boʻlib, n esa harflar sonini ifodalaydi. Ushbu tadqiqotda 26 harfdan iborat oʻzbek alifbosi va boʻsh belgi uchun pastki chiziq “_” dan quyidagi tarzda foydalanamiz:

$$A = \{a, b, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, x, y, z, o', g', sh, ch, ng, ', ' _'\}$$

Quyidagi yozuvni har qanday S_n satrning $1 \leq i \leq j \leq n$ qanoatlantiruvchi satr ostisini belgilashda moslashtiramiz,

$$S_n[i : j] = h_i h_{i+1} \dots h_j$$

$$S_n[: j] = h_i h_{i+1} \dots h_j \text{ va } S_n[i :] = h_{i+1} \dots h_n$$

Yuqoridagi belgilashga asoslangan holda,

$S_n[i : i + 1] = h_i h_{i+1}$ maxsus satr osti $(h_i, h_2)_i$ harflarining tartiblangan juftligi bilan belgilanadi.

Bu yerda i ($i = 1, 2, \dots, n - 1$) sub-indeks boʻlib, ushbu qatordagi tartiblangan juftlikning boshlangʻich pozitsiyasini va $h_1 = h_i, h_2 = h_{i+1} \in A$ ni bildiradi.

$i = n$ uchun juftlik oldindan tayyor qilib qoʻyilgan $(h_n, ' _')_{i=n}$ orqali aniqlanadi. Shunday qilib, tadqiqotda har qanday $S_n = h_1 h_2 \dots h_n$ satr n ta tartiblangan juftlikka ega boʻladi.

Har qanday oʻzbek tilidagi soʻz shaklda $1 \leq j \leq n_{max}$ pozitsiyasida paydo boʻlishi mumkin boʻlgan $(h_i, h_2)_j$ harflarining tartiblangan juftligi (n_{max} – oʻzbek soʻz shakli ega boʻlishi mumkin boʻlgan maksimal harflar soni) va $n \geq j$ uchun $S_n = h_1 h_2 \dots h_n$ bilan belgilangan maʼlum bir soʻz shakl, $(h_i, h_2) \in S_n$ yozuvi $i = j$ uchun $h_1, h_2 = (h_1, h_2)_j$ sharti bilan S_n da i ($1 \leq i \leq n$) pozitsiyasida tartiblangan $(h_i, h_2)_i$ juftlik mavjudligini anglatadi.

Shuningdek, har qanday soʻz shaklini barcha $1 \leq m \leq n$ uchun $S_n^m = (g_m, e_m)$ boʻyicha satr ostining tartiblangan juftligi sifatida ifodalash uchun yana ikkita $g_m = s_n[: m]$ va $e_m = s_n[m:]$ belgilashlarni kiritamiz.

Namuna maydoni va tartiblangan juftlik ehtimoli

Faraz qilaylik, L toʻplam $i = 1, 2, \dots, n_{max}$ pozitsiyalari uchun har qanday oʻzbek tilidagi soʻz shaklida paydo boʻlishi mumkin boʻlgan barcha mumkin boʻlgan tartiblangan $(h_1, h_2)_i$ harf juftliklarining yigʻindisi boʻlsin. Demak, L namuna maydonini quyidagicha aniqlash mumkin:

$$L = \{(h_1, h_2)_i | h_1, h_2 \in A \text{ va } 1 \leq i \leq n_{max}\}$$

$G_k, E_k, T_k \in L$ va $1 \leq k \leq n_{max}$ boʻlgan G_k, E_k va T_k toʻplamlari aniqlangan eventlarni quyidagicha ifodalash mumkin:

$$G_k = \{(h_1, h_2)_i | i = k \text{ va } (h_1, h_2)_i \in g_m \text{ va } 1 \leq m \leq n_{max}\}$$

$$E_k = \{(h_1, h_2)_i | i = k \text{ va } (h_1, h_2)_i \in e_m \text{ va } 1 \leq m \leq n_{max}\}$$

$$T_k = \{(h_1, h_2)_i | i = k, h_1 = S_n[k:k], h_2 = S_n[k+1, k+1], 1 \leq i \leq n_{max}\}$$

Shunday qilib, $S_n = h_1 h_2 \dots h_n$ bilan belgilangan har qanday berilgan so'z shaklining $i = 1, 2, \dots, n$ pozitsiyalarida tartiblangan har bir $(h_1 h_2)_i$ juftligi uchun yuqoridagi uchta to'plamda bo'lish ehtimolini quyidagicha aniqlash mumkin,

$$Pr(S_n[i: i + 1] \in G_i) = Pr((h_1, h_2)_i \in G_i) = P_G((h_1, h_2)_i) \quad (1)$$

$$Pr(S_n[i: i + 1] \in E_i) = Pr((h_1, h_2)_i \in E_i) = P_E((h_1, h_2)_i) \quad (2)$$

$$Pr(S_n[i: i + 1] \in T_i) = Pr((h_1, h_2)_i \in T_i) = P_T((h_1, h_2)_i) \quad (3)$$

Bu yerda (1) tenglama $(h_1, h_2)_i$ tartiblangan juftlik berilgan so'z shaklining o'zak qismida bo'lish ehtimolini bildirsa, (2) tenglama ham $(h_1, h_2)_i$ tartibli juftlik berilgan so'z shaklining qo'shimcha qismida bo'lish ehtimolini bildiradi va nihoyat, (3) tenglama $(h_1, h_2)_i$ tartiblangan juftlik berilgan so'z shaklining o'zak qismi va qo'shimcha qismi o'rtasida bo'lish ehtimolini bildiradi.

Ehtimoliy stemlash tasdig'i

Ushbu tadqiqotdagi ehtimollik modelimiz aniqligi quyida keltirilgan tajriba orqali tekshiriladi.

1-taklif. $[P_E((h_1, h_2)_m) > P_G((h_1, h_2)_m)]$ va $P_T((h_1, h_2)_{m-1}) \geq \alpha$ tenglamani qanoatlantiradigan $1 \leq m \leq n$ butun qiymat mavjud bo'lsa,

Bu hodla $S_n^m = (g_m, e_m)$ so'z shakli va $0 \leq \alpha \leq 1$ o'zgarmas uchun, g_{m-1} so'z shaklining ehtimoliy o'zagi deb aytiladi.

1-ta'rif. $S_n = h_1 h_2 \dots h_n$ ($1 \leq n \leq n_{max}$) bilan belgilangan har qanday so'z shaklning ehtimollik asosi e_{m-1} satr ostini ajratib olish va g_{m-1} satr ostini ma'lum $0 \leq \alpha \leq 1$ uchun 1-taklifni qanoatlantiradigan m pozitsiyasida ushbu so'z shaklining o'zagi sifatida qabul qilishdan iborat.

Tajriba va natijalar

Tajriba va test ma'lumotlar bazalari o'zbek tili korpusidagi matnli ma'lumotlar bazasidan tanlandi. Bu o'zbekcha xabar matnlari to'plami bo'lib, 1 milliondan ortiq so'z shakllariga ega. Tanlangan ma'lumotlar bazalarining xossalari va natijalari quyidagi 2-jadvalda keltirilgan.

2-Jadval. Test ma'lumotlar bazasi va tajribalar xususiyati

Ma'lumotlar bazasi	So'zlar soni	Unikal so'zlar	Unikal o'zaklar	Noma'lum o'zaklar	Bajarilgan (to'g'ri)	Juftlik soni	Samaradorlik (%)
Tajribaviy	210,129	32,122	9,568	0	0	5,648	0
TEST	211,489	32,563	9,253	3,212	8,918	-	95.8

Ikkala ma'lumotlar bazasidagi terminlar morfologik tahlil natijalariga ko'ra o'zak va qo'shimcha qismlarni aniqlash uchun oldindan qayta ishlangan. Bunda, ma'lumotlar bazalari boshqa qayta ishlanmaydi, masalan: *nomuhim so'zlarni olib tashlash, noto'g'ri yozilgan so'zlarni aniqlash va hokazolar.*

Shuni ta'kidlash lozimki, agar morfologik analizator ma'lumotlar bazasi uchun stemmer sifatida ishlatilsa, siqilish nisbatining mutlaq yuqori chegarasi **74%** ni tashkil qiladi (ya'ni, 32,563 ta alohida terminlarning barchasi ularning **9,253** haqiqiy stemiga to'liq qisqartirilganda) va taklif qilingan o'zak siqish nisbati **72%** ga yetishi mumkin.

(1), (2) va (3) tenglamada berilgan ehtimollar quyidagi formulalar bilan hisoblanadi:

$$P_G((h_1, h_2)_i) = f_{g,i} * w_{g,i} / N; P_E((h_1, h_2)_i) = f_{e,i} * w_{e,i} / N;$$

$$P_T((h_1, h_2)_i) = f_{t,i} * w_{t,i} / N$$

Bu yerda

– $f_{g,i}$, $f_{e,i}$ va $f_{t,i}$ - tartiblangan harf juftlarining mos ravishda o'zak qismida, qo'shimcha qismida va so'z shaklining o'zak va qo'shimcha qismi o'rtasidagi tranzitda paydo bo'lishi eventlari bo'yicha tajriba bazasidan kuzatilgan chastotalar soni;

– $w_{g,i}$, $w_{e,i}$ va $w_{t,i}$ tuzatish omillari ilgari mos keladigan $f_{g,i}$, $f_{e,i}$ va $f_{t,i}$ chastotalar uchun ijobiy haqiqiy qiymatlar hisoblanadi.

– **N** - tajriba ma'lumotlar bazasida buyurtma qilingan juftliklarning umumiy soni.

Tajriba ma'lumotlar bazasiga ishlov berilgandan so'ng, **L** namunaviy maydoni uchun umumiy soni **7148** ta unikal tartiblangan harflar juftligi kuzatildi. Ulardan **3142** juft G_k to'plamiga, **1381** juft E_k to'plamiga va **367** juft T_k ga tegishli.

G_k va E_k to'plamlar kesishmasida **1397** ta unikal tartiblangan juftliklar mavjud. Darhaqiqat, bu kuzatuvlar asosida barcha tartiblangan juft harflarning **50%** i faqat o'zak qismida, **19%** faqat qo'shimcha qismida va **24%** o'zak va qo'shimcha qismlarida bo'lishlarini ko'rsatadi.

Ushbu statistik ma'lumotlar biz taklif qilayotgan model nima uchun muvaffaqiyatli bo'lganini aniq tushuntirib beradi. Test ma'lumotlar bazasida **9253** ta alohida stem mavjud bo'lib, ulardan **3212** tasi eksperimental ma'lumotlar bazasida yo'q. Bu nisbat juda katta hisoblanadi va taxminan **40%** noma'lum stemlarni tashkil qiladi. Stemmer **32563** ta test shartlari uchun **68962** ta mumkin bo'lgan stemlarni ishlab chiqardi. Shunday qilib, har bir atama uchun taxminan **1,5** ta stem ishlab chiqarilgan.

Aslida, bu **68962** hosil bo'lishi mumkin bo'lgan o'zaklarning aksariyati tan olingan o'zbek so'z shakllaridir. Lekin ular kerakli o'zakdan farqli semantikaga ega bo'lgan root (o'zak)ga tushiriladi yoki kerakli o'zakdan farqli bo'lgan hosila yoki flektiv shaklga o'zlashtiriladi.

Hisob-kitob natijasida, kerakli stemming aniq mosligini qabul qildik. Natijada, biz taklif qilgan stemmer sinov shartlarining **94,2%** uchun to'g'ri stemlarni ishlab chiqarishga erishdi. Bizning stemmer **32563** tadan **672** tasi turlicha bo'lgan **1055** ta atamani stemlarga ajrata olmadi. Biroq, o'tkazib yuborilgan atamalarning aksariyati

oddiy IR tizimida ishlatiladigan har qanday matn protsessorlari tomonidan yo'q qilinishi mumkin bo'lgan **neologizmlar**, **qisqartmalar** yoki **olmoshlardir**.

Xulosa

Ushbu maqolada agglyutinativ tillar uchun ehtimollikka asoslangan yangi stemming modeli taqdim etildi. Taklif qilingan model chiziqli hisoblash murakkabligiga ega bo'lib, **94,2%** aniqlik darajasiga ega. Maqolada keltirilgan stemming modelini fleksiyon paradigmalariga ega bo'lgan boshqa agglyutinativ tillarga qo'llash mumkin.

Foydalanilgan adabiyotlar ro'yxati:

1. Hakkani-Tür, D. Z., Oflazer, K., & Tur, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Language Resources and Evaluation*, 36(4). <https://doi.org/10.3115/990820.990862>
2. Jurafsky, D., & Martin, J. (2014). Speech and Language Processing. In *Speech and Language Processing*. (Vol. 3).
3. B.Elov, Sh.Hamroyeva, X.Axmedova. (2022). Methods for creating a morphological analyzer, *14th International Conference on Intellegent Human Computer Interaction, IHCI 2022, 19-23 October 2022, Tashkent*.
4. Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(June).
5. Porter, M. F. (2006). An algorithm for suffix stripping. *Program*, 40(3). <https://doi.org/10.1108/00330330610681286>
6. B.B.Elov, Sh.M.Khamroeva, Z.Y.Xusainova. Yashirin markov modeli vositasida o'zbek tili matnlarini POS teglash (2022).
7. Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash. O'zbek amaliy filologiyasi istiqbollari / Respublika ilmiy-amaliy konferensiya to'plami. Elektron nashr / ebook. – Toshkent: ToshDO'TAU, 26.10.2022. 159-163 b.
8. Muljono, Afini, U., & Supriyanto, C. (2017). Morphology analysis for Hidden Markov Model based Indonesian part-of-speech tagger. *Proceedings - 2017 1st International Conference on Informatics and Computational Sciences, ICICoS 2017, 2018-January*. <https://doi.org/10.1109/ICICOS.2017.8276368>
9. Köksal, A.: Bilgi Eri im Sorunu ve Bir Belge Dizinleme ve Eri im Dizgesi Tasarım veGerçekle tirimi. Doçentlik Tezi. Fen Bilimleri Enstitüsü, Bilgisayar BilimleriMühendisli i Anabilim Dalı, Hacettepe Üniversitesi, Ankara (1979) (12) (PDF) *Stemming in Agglutinative Languages: A Probabilistic Stemmer for Turkish*.

COGNITIVE INTERPRETATION OF THE CATEGORY OF INTENSITY. INTENSITY AS A MEANS OF FOREGROUNDING

Gulyamova Ozodaxon Shavkatovna
Alisher Navoiy Tashkent State University of
Uzbek Language and Literature
English Philology Department
ozodahongulyamova@gmail.com

Abstract: *The article analyses the category of intensity as a means of foregrounding, as well as the intertextual functions of the strong position of the text, as a way of forming the reader's worldview through the prism of the author's individual picture of the world.*

Key words: *category of intensity, foregrounding, author's individual picture of the world, intensifiers, expressiveness, emotiveness*

Аннотация: *В статье анализируется категория интенсивности как средство актуализации, а также интертекстуальные функции сильной позиции текста, как способ формирования читательского мировоззрения через призму индивидуально-авторской картины мира.*

Ключевые слова: *категория интенсивности, выдвижение, индивидуально-авторская картина мира, интенсификаторы, экспрессивность, эмотивность.*

Annotatsiya: *Maqolada intensivlik kategoriyasi aktuallashtirish vositasi sifatida, shuningdek, matnning kuchli pozitsiyasining matnlararo funktsiyalari, muallifning dunyoning individual surati prizmasi orqali o'quvchi dunyoqarashini shakllantirish usuli sifatida tahlil qilinadi.*

Kalit so'zlar: *intensivlik kategoriyasi, ilg'orlik, dunyoning individual mualliflik surati, intensifikatorlar, ekspressivlik, emotivlik.*

In the process of perceiving a literary text as a complex structural and semantic unity, certain signals play a key role, helping to build an aesthetic effect. Such signals contribute to linking the stylistic significance of the text with the perception and interpretation of the text as a whole. One of these signals is the category of foregrounding, in the construction of which intensity plays the most active role. According to the general recognition of scientists, the work of representatives of the Prague Linguistic Circle is the starting point on the way to the introduction of the concept of “foregrounding” into the field of linguistics, as well as its further special consideration in modern linguistic and stylistic studies, in particular, in cognitive stylistics. In the works of I.V. Arnold, the phenomenon of extension is primarily considered within the framework of decoding style as “methods of formal



organization of the text, focusing the reader's attention on certain elements of the message and establishing semantically relevant relationships between the elements of one or more often at different levels" [Arnold I.V., 2002:99]. Arnold presents a wide stylistic differentiation of types of advancement (convergence, cohesion and deceived expectation) as well as highlighting them.

In modern linguistics, there is a tendency for a systematic approach, when studies are carried out not of individual elements, and not even their relationships, but the study of entire systems with numerous complex constituent units that form one single whole due to relationships. Each link in this unified chain performs certain functions, translating specific semantic attitudes, closely interacting with each other, thereby serving to organize a complex structure of content. The idea of a particular text emerges like a mosaic, where each individual element complements and sets off the neighboring one, sometimes changing the semantic load, so the meaning of the text is not a simple sum of the meanings of its parts. So, nomination is "a technique that determines the great importance of placing in the foreground in terms of its importance of one or another language norm, which acts as a search stimulus or "key" in the processes of language processing of information".

The nomination contributes to the interpretation of the text from the correct angle, activates not only background knowledge, but also judgments, attitudes and emotions, facilitates the search for relevant information, reduces the need for large amounts of information [Kubryakova E.S., 2004:560].

One of the tasks of the category of intensity from the standpoint of the principle of advancement is the study of verbal means of its implementation, namely the convergence of stylistic devices, cohesion, the effect of deceived expectations. The analysis of the linguistic material showed that category of intensity can equally act as a means of foregrounding. This is facilitated by the factor of stylistic saturation, a wide variety of stylistic means of emotional, evaluative, figurative and expressive nature. All this contributes to the "foregrounding" and the achievement of its effectiveness, category of intensity, characterized by stylistic marking, which manifests itself in the convergence of stylistic devices, is considered as one of the methods of foregrounding, marking the most conceptually significant information.

Let us turn to the analysis of category of intensity in the nomination position in the novel *The Land of Decoration*:

I KNOW ABOUT faith. The world in my room is made out of it. Out of faith I stitched the clouds. Out of faith I cut the moon and the stars. With faith I glued everything together and set it humming. This is because faith is like imagination. It sees something where there is nothing, it takes a leap, and suddenly you're flying.

This passage contains a large number of stylistic intensifiers involved with varying degrees of emphasis in the text of the work. The first graphic intensifier is the capitalization of the first three words. However, the fact that each chapter of this



book begins in this style - writing the first three words in capital letters, this deviation from graphic norms loses its “non-triviality” to some extent, being part of the individual author's style. That is, this graphic technique can be considered an example of advancement only from the point of view of generally accepted classical norms of literature. The next intensifier is inversion. *Out of faith I stitched the clouds. Out of faith I cut the moon and the stars.* This syntactic device serves to focus the reader's attention on the word faith, highlighting faith as the basis of the entire universe of the character. The author maximally focuses the reader's attention on the concept of faith, resorting to repeated repetition. A vivid example of a semantic intensifier is also a metaphorical epithet - when a character declares that he carved out the moon and stars from faith, and wove clouds from faith. This author's technique helps the reader to better understand the socio-psychological portrait of the girl, to realize the primary role of faith in the character's life, explaining that not only the artificial world created by the character, but also the girl's real world was based on faith. She believed that the Land of Decoration she created was real. Through faith, she saw houses, people, animals, and so on down to the smallest detail, in the world she created. Where others saw only a pile of garbage and useless trash.

The analysis of the following passage from Georges Orwell's novel "1984" demonstrates:

Every day newspapers, magazines, photographs, films, posters and books were all changed. The past was changed. The Party was always right. The Party had always been right. The Records Department, where they destroyed all the old copies of everything, was the largest department in the Ministry of Truth, but there was no truth. The new copies were not true and the old copies had not been true either.

The first thing that undoubtedly attracts the reader's attention is a long synonymous row: *newspapers, magazines, photographs, films, posters and books.* The syntactic intensifier in the form of parallel constructions is used by the author to create a certain rhythm of the narration. The same purpose is served by intensifiers used at the level of constructing sentences. The sentences are short, jerky - *The past was changed. The Party was always right. The Party had always been right.* The author tries to convey the emotional state of the character and all the unbearability of the reality surrounding him, using the most concise short sentences that reflect the scarcity and routine of everyday life. A world in which all "superfluous and interfering" emotions are prohibited. People receive short, jerky orders and just as silently, without further ado, carry them out. The author also resorts to an intensifying device-repetition. The sentence *The Party was always right* is immediately repeated, but there is an element of strengthening the statement at the grammatical level - in the second sentence, the correction of time, *The Party had always been right*, this statement becomes an indisputable axiom. The next author's neologism **Ministry of Truth**, described as the largest ministry, and especially the one that followed it, but



there was no truth, emphasizes the absurdity of what is happening in the world described in the novel. The reader is clearly aware of the deceit and hypocrisy of the system that exists in the world invented by the author, and wonders if this world is such an invented one, and if there are too many analogies in modern life with this world. In the sentence The new copies were not true and the old copies had not been true either, the author resorts to the method of nomination - the effect of deceived expectations. After all, having read the first part of the sentence The new copies were not true, the reader subconsciously expects a refutation. What followed instead and the old copies had not been true either draw the final touches, revealing the whole palette of falseness and deception of the system.

It is worth noting that, being a way of foregrounding, the category of intensity contributes to the creation of the author's subjective picture of the world within certain limits of artistic space and time. This picture of the world is the result of comprehending an impressive set of information about the world, along with the author's personal experience as a whole. It should be noted that the linguistic personality of the author manifests itself when using CI as a means of promotion when using the entire palette of stylistic means.

Analyzing the following passage from The Land of Decoration, we came to the following conclusions:

PEOPLE DON'T BELIEVE in very much. They don't believe politicians and they don't believe ads and they don't believe things written on packets of food in the grocery store. Lots of them don't believe in God either. Father says it's because science has explained so many things that people think they should be able to know how everything happens before they believe it, but I think there is another reason.

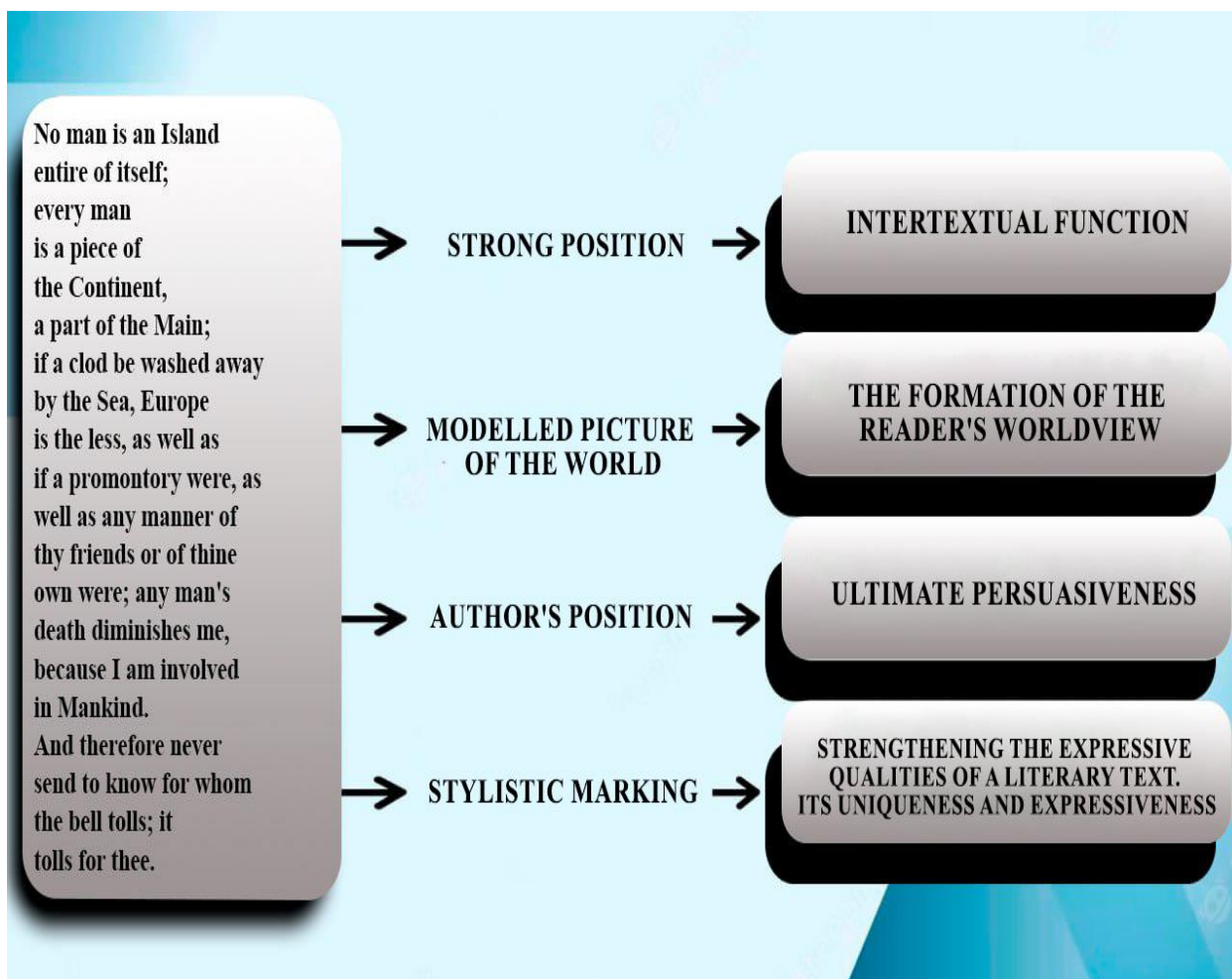
I think people don't believe things because they are afraid. Believing something means you could be wrong, and if you're wrong you can get hurt. For instance, I thought I could climb the whole way round my room without touching the floor, and it hurt when I fell down. All the important things, like whether someone loves you or something will turn out right, aren't certain, so we try to believe them, whereas all the things you don't have to wonder about, like gravity and magnetism and the fact that women are different from men, you can bet your life on but you don't have to.

This passage draws the attention of the reader because is in the position of foregrounding, which is achieved by: a) placing the text in a strong position (the beginning of the chapter) b) convergence of stylistic devices - syntactic intensification - repetition, epithet, antithesis, comparison, inversion; d) an expression of the author's position regarding the vital values of human life. Since this passage is the beginning of the chapter, it not only attracts the reader's attention, but also helps to place accents and mark landmarks both for the entire chapter in particular and for the entire subsequent narrative [Arnold I.V.,1981:298] Here the

personality of the author enters into reflections through the transmission of the thoughts of the character.

The main position in literary texts is occupied by the title, epigraph, first paragraphs. Here we will focus on the role of the epigraph as one of the means of foregrounding. The epigraph, altogether with the title, in a concise form expresses the concept of the literary text, simultaneously acting as a category of intensity. At the same time, there is some autonomy of the epigraph, its autosemantic and intertextual character. Indicative in this respect is the epigraph from E. Hemingway's novel "For Whom the Bell Tolls". The entire poem by John Donne is used here as an epigraph:

No man is a island, intire of it selfe; Every man is a piece of the Continent, a part of the maine; if Cloud be washed away by the Sea Europe is the lesse, as well as if a Promontorie were, as well as if a Manner of thy friends or of thine owne were; any mans death diminishes me, because I am involved in Mankinde; And therefore never send to know for Whom the bell toll; it tolls for thee



John Donne's poem about the responsibility of people, about the exceptional value of human life, corresponds to Hemingway's ideas about the need to "save the



world", about the role of the individual in the common struggle for freedom and justice. Category of intensity from the standpoint of foregrounding is a complex speech and cognitive construction capable of realizing a large number of explicit and implicit meanings, the decoding of which contributes to the interpretation of conceptual information. The connection between the epigraph and the entire text of the novel is evident both in the title and in separate fragments: *“But remember, this that as long as we can hold them here we keep the fascists tied up. They can't attack any other country until they finish with us and they can never finish with us.”* Stylistic marking, in this example expressed by the convergence of the stylistic device, contributes to the fact that the category of intensity, in this case the epigraph, appears in a strong position of the text, thereby marking the most significant information. Thus, the epigraph performs a pre-text function here, directing the reader's perception in the right direction.

Category of intensity performs the most important function, namely, it contributes to a deeper penetration into the ideological and philosophical essence of the work and creates a context that allows the reader to follow the logic of the author's thoughts, makes the ideas presented more convincing. Stylistically rich texts have the power of influence not only as a means of removing the uncertainty of the modeled picture of the world and disclosing conceptual information, but also as a means of forming the reader's worldview.

Thus, this passage attracts the attention of the reader, because. is in the position of promotion, which is achieved: a) by placing the text in a strong position (the beginning of the text); b) lack of direct connection with the content of the text; c) convergence of stylistic devices - metaphor, epithet, antithesis, comparison, inversion; d) an expression of the author's position regarding the vital values of human life. The personality of the author appears in an emotive context, which often plays a role in shaping the concept of the entire work, contributing to a deeper penetration into the ideological and philosophical essence of the work and creating a context that allows the reader to follow the logic of the author's thoughts, giving the ideas presented more persuasiveness.

The analysis of the linguistic material showed that category of intensity can act as one of the means of foregrounding the most significant information in a literary text, and this is created by its stylistic marking and various ways of emphasizing information: a) convergence of stylistic device; b) in a strong position i.e. in a position of nomination which manifests itself in violation of the storyline of the narrative.



References

1. Arnold I.V. Stylistics. Modern English: Textbook for Universities. 5th ed., rev. and additional – M.: Flinta: Nauka, P. 2002. – 384
2. Arnold I.V. Stylistics of Modern English. Decoding Style. - L.: Education, P.1981. - 298 .
3. Ernest Hemingway. For whom the Bell Tolls. United States. Charles Scribner's Sons. 1940
4. George Orwell. 1984. Great Britain, Secker & Warburg, 1949.
5. Grace McCleen. Land of Decoration. Henry Holt and Co. 2012
6. Karasik V.I. Cultural Dominants in Language // Linguistic Personality: Cultural Concepts. - Volgograd - Arkhangelsk: Change, 1996. - p. 3-16.
7. Kubryakova E.S. Language and Knowledge. On the Way to Gaining Knowledge About Language: Parts of Speech from a Cognitive Point of View. The Role of Language in the Knowledge of the World. - M.: Languages of Slavic Culture, P.2004. - 560
8. Normurodova N.Z. Cognitive-stylistic Features of Graphic Means as an Expression of the Author's Conceptual Picture of the World // Foreign Philology – 2021– No. 1, P. 49-55.
9. Harris Z. S. Discourse analysis. Language 28:1–30, L., 1952. – 189 p.
10. Werth P. Text Worlds: Representing Conceptual Space in Discourse. London: Longman, 1999

MUNDARIJA

SO'Z BOSHI

Sirojiddinov Sh.S. Kompyuter lingvistikasi – milliy tilshunosligimiz istiqboli ..	4
Isajon S. Kompyuter lingvistikasi tadriji xususida	5
Jamoliddinova O.R. Kompyuter lingvistikasi istiqboli yoshlar qo'lida	6

YALPI YIG'ILISH MA'RUZALARI

Eşref Adalı Alfabe (Abece) ve Bilişim	7
Mengliyev B.R. "So'zlar birikuvchanligi lug'ati" va uning ahamiyati.....	15
Qarshiyev A.B. O'zbek tili matnlarining razmetkalangan korpusini modellashtrish	18
Xolmanova Z.T. Sun'iy intellekt: BORIS	26

TABIIY TILNI QAYTA ISHLASH (NLP)

Галимов М.Р., Бурнашев Р.А., Гатиатуллин А.Р., Прокопьев Н.А. Проектирование геоинформационной системы "диалектолог" с использованием нечеткой логики	34
Sharipov M.S, Adinayev X.S, Raximov R.R. O'zbek tili uchun tabiiy tilni qayta ishlashda (NLP) mashinali o'rganishning (Machine Learning) o'rni.....	46
Elova D.Q., Abdusalomova SH.A Koreferensiya va anafora talqini	53
Qurbonova R.U., Ro'zimboyeva S.N. Data mining techniques in natural language processing: a review	58
Axmedova X.I., Mirzaxmedova P. So'z ma'nosini aniqlashda mashinali o'qitish yondashuvlari	64
Xusainova Z.Y. O'zbek tilida stemmingni amalga oshirishning gibrid statistik yondashuvi	70
Axmedova X.I., Xushmurodova M. So'zlarni semantik tahlil qilish algoritmnlari	77
Xusainova Z.Y., Qosimova S.F, Baxtiyorova N.D., Turayeva F.SH. BPE algoritmi asosida o'zbek tilidagi matnlarni tokenlash	83
Xusainova Z.Y., Xusanboyeva J.I, Bekqulova S.F., Bekchanova M.O. O'zbek tili matnlarini qayta ishlashda BOW algoritmidan foydalanish.....	92
Rakhmonova M.A. Syntactic parsing approaches for english, turkish and uzbek	100
Raxmonova M.A. O'zbek tilidagi sodda gaplar uchun sintaktik parsing modeli	108
Adilova M. NLP vazifalari va yo'nalishlari	118

KORPUS LINGVISTIKASI

Elov B.B., Amirqulov M.A. O'zbek-ingliz tillari parallel korpusiga qo'yiladigan lingvistik va ekstralingvistik talablar	126
Islomov I.X. O'zbek tili milliy korpusi uchun litosferaga oid terminlar mikrotizimi tasnifi	133

Babayev S.S., Ruzimboyev X.R. Data science in corpus linguistics	138
Xolmonova I.A. Turk-o'zbek parallel korpusiga matn tanlashda muhim xususiyatlar	143
Bo'riyev I.E. Korpusda o'zbek tilidagi iboralarini semantik teglash masalasi	147
MATNGA QAYTA ISHLOV BERISH TIZIMLARI	
Abduraxmanova M.T, Olimova M.P. Morfologik analizator xususida	153
Elova D.Q. O'zbek tili matnlarida koreferensiyani hal qilish bosqichlari	157
Sharipov M.S., Yakubova D.K., Yuldashev O.E. O'zbek tilidagi matnlardan fe'l so'z turkumiga oid so'zlarni qoidaga asoslangan holda qo'shimchalar asosida aniqlash	163
Tursunov M.S. O'zbek tili matnlarini razmetkalash	169
Yodgorov U.S. Text generator based on lexis-syntactic templates	174
Niyazmetova K.E., Quriyozov E.R. Restoran sohasidagi o'zbek tilidagi matnlarning sentiment tahlili	181
Bozorov O.N. Matnlarning o'xshashligini tahlil qilish	188
Allanazarova S.Y. O'zbekcha matnlarni hissiy tahlil qilish uchun leksikonga asoslangan yondashuv	191
Suyunova M.O. Matnni avtomatik tahlil qilishdagi muammolar	197
Mahmudjonova G.U. Nomuhim so'zlar tushunchasi va uning ahamiyati	205
Muratbekova Sh.D. Sodda gaplardagi mazmuniy moslik, o'xshashlik va shakliy farqlilikning leksik vositalari hamda ularning lingvistik modellari	213
Salaeva M.I., Kuriyozov E.R., Salaev U.I. Uzbek automatic speech recognition models using deep learning techniques	219
O'ZBEK TILINING FORMAL GRAMMATIKASINI YARATISH	
Jurayeva N., Atayeva D. O'zbek tilida sodda gap strukturalari	225
Есенова Г. Өзге тілді мектепте қазақ тілінен орындалатын жаттығулардың лингвистикалық аспектілері	231
Айтбенбетова А.К., Муратбек Б.К., Молдабаева З.М. Стили речи в дискурсе персонажа	238
Sharipov M.S, Jumaniyozova Z. O'zbek tilidagi matnlardan sifat, ravish, yordamchi so'z va oraliq so'zlar turkumiga oid so'zlarni aniqlash.....	244
Abdullayeva O.X., Nizomova F.B. O'zbek tilida ravish so'z turkumini modellashtirishda so'z yasashi masalalari va ularning morfoanalizatorida tahlili	250
Iskandarov A., Shermatov B., Kuriyozov E.R. Automatic detection technologies for stopwords in Uzbek language	261
Sharipov M.S., Mavlonova M.M., Sobirov O.O. O'zbek tilidagi matnlarda ot va olmosh so'z turkumlariga tegishli so'zlarni aniqlovchi algoritmi yaratish	268
Xajiyeva D.SH., Jo'rayeva N.V. O'zbek tilidagi fe'lli birikmalarning formal modellari	273
Matlatipov G., Karimova M., Salayev U.I. Designing a tree-based model for analyzing the syntax of Uzbek language sentences	280



Berdiyeva H.A. Morfem lug‘atlarda tovush o‘zgarishlarining modellashtirilishi masalalari	285
Xudayarova S. O‘zbek tilidagi so‘z birikmalarini modellashtirish muammolari masalasi	290
KOMPYUTER LINGVISTIKASIDA SUN’IY INTELLEKT VA MASHINA TARJIMASI	
Hamroyeva SH.M, Matyakubova N. Mashina tarjimasida matnni moslashtirish usullari	296
Ibragimova S.N. Интеллектуальная система обучающая узбекскому языку как иностранному на базе анализатора речи	304
Rizvanov Q.Q. NUMPY kutubxonasi muhitini sozlash	311
Samatboyeva M.T. NER (Named Entity Recognition – nomlangan obyektni tanib olish) obyektlari tarkibida toponimlarni avtomatik aniqlash	322
Uzoqova M.T. Leksik sinonimlarni aniqlash uchun Word2Vec va RoBERTaForMaskedLM modellaridan foydalanish. Syn-Roberta modeli haqida	329
SEMANTIK TIZIMLAR VA LINGVISTIK ONTOLOGIYALAR	
Shermatov B., Kuriyozov E.R. Building semantic evaluation dataset for uzbek language	340
Nizomova Z.K. Kimyoviy terminlar tezaurusini yaratishda semantik munosabatlar	345
Salomov A.A. O‘zbek-ingliz sinonimlari bazasining tuzilishi va tarkibi	351
KOMPYUTER LEKSIKOGRAFIYASI VA LINGVODIDAKTIKASI	
Atamuratov R.K., Shakirova S.T. Axborot texnologiyalari sohasiga doir xorijiy tillardan o‘zlashgan so‘zlar tadqiqi	355
Boborajabov E. O‘zbekiston media maydonida sport saytlarining rivojlanish tendensiyalari	359
Boborajabov E. O‘zbek milliy internet sigmentida saytlarning yaratilishi, faoliyati va ommalashishi	364
Qurbonova R. Kompyuter leksikografiyasi. Multimedia va lug‘at	371
Aloyev N.R., Xudayberganov N.U. O‘zbek tili uchun stemmerni ishlab chiqish.....	375
Gulyamova O.Sh., Cognitive interpretation of the category of intensity. Intensity as a means of foregrounding.....	383