ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 2 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 2 (6)

compling.tsuull.uz

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

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

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

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 43-53

JAHON TILSHUNOSLIGIDA TREEBANKLAR TASNIFI

Oqila Abdullayeva[1]
O‘g‘iloy Bozorqulova[2]

Annotatsiya. Treebanklar tilshunoslikda katta ahamiyatga ega bo‘lib, ular tabiiy tilni avtomatlashtirish, kompyuter lingvistikasi, mashinada tarjima va tilshunoslik tadqiqotlari uchun muhim resurs hisoblanadi. Ushbu maqolada jahon tilshunosligida treebanklarning turlari, tasnifi va ularning qo‘llanilish sohalari haqida so‘z yuritiladi.

Kalit so‘zlar*: til korpusi, lingvistik teglash, Prague Treebank, Universal Treebank, English Treebank, morfologik tahlil, POS teglash, sintaktik tahlil, semantik tahlil.*

Kirish

Tabiiy tilni kompyuter yordamida tadqiq qilish va qayta ishlash, odatda, to‘rt jihatni: formallashtirish, algoritmlash, dasturlash va amaliyotga qo‘llashni o‘z ichiga oladi [Hovy, Lavid, 2010; Arista, 2022].

Sintaktik parsing gapning grammatik tuzilishi bilan birga, gapning to‘g‘ri yoki noto‘g‘ri tuzilganligini, gap chegaralarini va so‘zlarning o‘zaro birikish munosabatlarini tahlil qiladi. Sintaktik tahlil bilan bog‘liq tadqiqotlar o‘rganilganda treebank atamasini uchratish mumkin. Treebanklar til korpuslarida matni sintaktik va semantik tahlil qilish usulidir. 80-yillarda Jeofri Lich tomonidan kiritilgan bu termin matnni ham sintaktik, ham semantik xususiyatlarini kompozitsion tahlil tasviri daraxt strukturasiga o‘xshatilgan [Sampson, 2003]. Treebank (yoki “daraxt banki”) bu lingvistik ma’lumotlar bazasi bo‘lib, tabiiy til matnlarining sintaktik

[1]*Abdullayeva Oqila Xolmo‘minovna - filologiya fanlari bo‘yicha falsafa doktori. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti doktoranti (DSc).*
E-pochta: *abdullayeva.oqila@navoiy-uni.uz*
ORCID: 0000-0002-2524-4832
[2]*SBozorqulova O‘g‘iloy Erkin qizi - Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi yo‘nalishi 1-kurs magistranti.*
E-pochta: *ogiloybozorqulova62@gmail.com*

tuzilmalari annotatsiyalangan korpuslarini o‘z ichiga oladi. Bu atama NLP va kompyuter lingvistikasi sohalarida ishlatiladi. Treebanklar tilshunoslarga va tilni qayta ishlash bo‘yicha mutaxassislarga tahliliy ma’lumotlarni taqdim etadi, xususan:

1. Sintaktik daraxtlar: Treebankda har bir jumla sintaktik daraxt shaklida ifodalanadi. Bu daraxt jumlaning tuzilishini (masalan, so‘z turkumlari va ularning o‘zaro bog‘lanishlari) ko‘rsatadi.

2. Qo‘lda yoki avtomatik annotatsiya: Treebanklar qo‘lda, avtomatik yoki ushbu usullarning kombinatsiyasi yordamida yaratilishi mumkin.

3. Til texnologiyalarida qo‘llanilishi: treebanklar, shuningdek, mashina tarjimasi, til modellarini qurish, jumlalarni tahlil qilish va tushunish tizimlarida qo‘llaniladi.

Asosiy qism

Dunyoda keng qo‘llaniladigan treebanklardan biri Penn Treebank bo‘lib, u ingliz tilining sintaktik va semantik ma’lumotlarini o‘z ichiga oladi. Shuningdek, boshqa ko‘plab dunyo tillari uchun treebanklar yaratilgan.

1. Penn Treebank – Ingliz tilidagi constituent (tarkibiy qismlar tahlili) treebanklarning eng mashhuri bo‘lib, NLP va mashinaviy o‘qitishda asosiy manbalardan biridir.

Penn Treebank – bu tabiiy tilni qayta ishlash (NLP) va lingvistik tadqiqotlar uchun keng qo‘llaniladigan, ingliz tilidagi matnlardan tashkil topgan sintaktik va semantik annotatsiyalangan treebankdir. U Pensilvaniya universitetida ishlab chiqilgan va NLP algoritmlarini rivojlantirishda asosiy resurslardan biri hisoblanadi.

Penn Treebankning xususiyatlari:

1. Ma’lumot manbalari:

Ingliz tilidagi korpuslar: gazeta maqolalari (masalan, Wall Street Journal), romanlar va boshqa turli matnlar.

Tegishli annotatsiyalar: sintaktik daraxtlar, so‘z turkumlari (POS) teglari va semantik belgilashlar.

2. Annotatsiya turi:

Jumlalar kontekstli grammatik tuzilmalari bilan daraxt shaklida ifodalanadi.

So‘z turkumlari (POS teglari) har bir so‘z uchun aniq ko‘rsatilgan.

Sintaktik tahlil daraxtlari faqatgina grammatik munosabatlarni emas, balki so‘zlar o‘rtasidagi bog‘lanishlarni ham aks ettiradi.

3. Maqsadi:

NLP modellarini o‘qitish va sinovdan o‘tkazish uchun standartlashtirilgan ma’lumotlar to‘plamini taqdim etish.

Sintaksis va semantikani tahlil qilishni avtomatlashtirishda asosiy vosita bo‘lish.

Ishlatilishi:

1. Tilni o‘qitish va tahlil qilish:

So‘z turkumlarini aniqlash (POS tagging).

Jumlalarni sintaktik tahlil qilish (Parsing).

2. Til texnologiyalari:

Mashina tarjimasi, suhbatlashuv tizimlari va chatbotlar; Til tushunish modellarini rivojlantirish.

3. Standart baholash:

Penn Treebankning qismi bo‘lgan Wall Street Journal korpusi ko‘pincha NLP modellarini sinash uchun standart benchmark sifatida ishlatiladi.

Muayyan model va vositalar:

Penn Treebank asosida mashhur NLP vositalari ishlab chiqilgan, masalan, Stanford Parser, Berkeley Parser va NLTK (Python kutubxonasi).

Bu treebank NLP sohasida asosiy resurs hisoblanadi va ko‘plab zamonaviy til modellarining rivojlanishiga hissa qo‘shgan.

2. Universal Dependencies (UD)

Bu bog‘liqlik treebanklari ko‘p tillarni qamrab oladi va bir xil formatda annotatsiya qilingan.

Universal Dependencies (UD) – bu lingvistik ma’lumotlarni birlashtirish uchun mo‘ljallangan ochiq loyihadir. UD turli tillar uchun sintaktik tuzilmalarni aniqlash va ularni bir xil formalizm asosida tavsiflashga qaratilgan.

UD bir qancha maqsadlarni ko‘zlaydi:

1. Turli tillar uchun umumiy tahlil usuli: UD yordamida barcha tillar uchun bir xil qoidalar va annotatsiya sxemalari qo‘llaniladi, bu esa tillararo taqqoslashni osonlashtiradi.

2. Universallik va soddalik: UDning qoidalari barcha tillar uchun mos kelishi, lekin shu bilan birga oddiy va tushunarli bo‘lishi lozim.

3. Kompyuterda foydalanish uchun yaroqlilik: UD ko‘pincha tabiiy tilni qayta ishlash (NLP) vazifalarida ishlatiladi, masalan, sintaktik tahlil yoki mashinani o‘qitish algoritmlari uchun.

UDning asosiy komponentlari

1. Annotatsiya darajalari:

Morofologik belgilar (features): So‘zlarning grammatik belgilarini (masalan, jins, son, kelishik) ko‘rsatadi.

Sintaktik bog‘lanishlar (dependency relations): So‘zlar o‘rtasidagi sintaktik aloqalarni belgilaydi (masalan, “subj” — ega, “obj” — to‘ldiruvchi, “mod” — aniqlovchi kabi).

So‘zlar (tokens): Matndagi alohida birliklar.

2. Teglar:

POS teglar – so‘z turkumlari (part-of-speech tags): Universal POS teglar, masalan, NOUN, VERB, ADJ va boshqalar.

Xususiyatlari: So‘zlar bilan bog‘liq qo‘shimcha ma’lumotlar, masalan, vaqt (tense), kelishik (case) va boshqalar.

3. Sintaktik aloqalar: So‘zlar o‘rtasidagi munosabatlarni belgilovchi struktura. Misol:

“Men kitobni o‘qidim” jumlasi uchun bog‘lanishlar:

Men → ega (subj)

kitobni → to‘ldiruvchi (obj)

o‘qidim → asosiy fe’l (root)

UD loyihasining ahamiyati shundaki, Universal Dependencies lingvistlar va tabiiy tilni qayta ishlash bo‘yicha tadqiqotchilar uchun juda foydali, chunki u til o‘rganish, taqqoslash va mashina o‘rganish jarayonlarini soddalashtiradi. UD ayni paytda yuzlab tillarni qamrab olgan va katta ma’lumotlar to‘plamlarini taklif etadi.

3. Prague Dependency Treebank

Chex tiliga asoslangan va chuqur bog‘liqlik sintaksisini ko‘rsatadi.

Prague Dependency Treebank (PDT) – bu chex tilining sintaktik va semantik tahlili uchun yaratilgan keng ko‘lamli korpus bo‘lib, u Praga universiteti (Charles University) tomonidan ishlab chiqilgan. PDTning asosiy maqsadi tabiiy tilning strukturalarini chuqurroq tushunish va lingvistik nazariyalarni tahlil qilishga yordam berishdir.

Asosiy xususiyatlari

1. Sintaktik daraja: PDT so‘zlar o‘rtasidagi sintaktik bog‘lanishlarni aniqlaydi. Ushbu bog‘lanishlar dependency grammar (bog‘lanish grammatikasi) tamoyillariga asoslangan bo‘lib, jumla tuzilishidagi asosiy so‘z (masalan, fe’l) bilan unga bog‘liq boshqa so‘zlar o‘rtasidagi aloqalarni tavsiflaydi.

2. Semantik daraja (tectogrammatical level): PDTning o‘ziga

xos jihati shundaki, u faqat sintaktik bog‘lanishlar bilan cheklanmaydi, balki so‘zlarning semantik rollarini ham aniqlaydi. Bu darajada faqat semantik jihatdan muhim elementlar (masalan, so‘zlarning ma’nosi) hisobga olinadi.

3. Ko‘p darajali tahlil:

Analitik daraja (analytical layer): Morfologik va sintaktik ma’lumotlar.

Tektogrammatik daraja: Semantik tuzilishlar va rollar.

Surface syntax: Yuzaki tuzilish tahlillari.

PDTning komponentlari

1. Korpus: PDT yuz minglab so‘zlardan tashkil topgan katta hajmli chex matnlari to‘plamidir. U matnlar bir necha qatlamda tahlil qilingan.

2. Annotatsiya qoidalari: har bir so‘zga morfologik va sintaktik teglar qo‘yilgan, so‘zlar o‘rtasidagi munosabatlar aniqlangan. Sintaktik bog‘lanishlar uchun (dependency grammar) asosida aniqlangan teglar ishlatiladi.

3. Semantik ma’lumotlar: har bir fe’l uchun argument-struktura (masalan, ega, to‘ldiruvchi) va ularning semantik rollari belgilangan.

PDTning qo‘llanilishi

1. Tilshunoslik tadqiqotlari: PDT orqali tilning sintaktik va semantik jihatlarini chuqur o‘rganish mumkin.

2. NLP (Tabiiy tilni qayta ishlash): PDT mashinali o‘qitish algoritmlarini o‘rgatish uchun ishlatiladi, ayniqsa, chex tilidagi NLP loyihalarida qo‘llaniladi.

3. Lingvistik qiyoslash: PDT boshqa tillar uchun yaratilgan dependency treebanklar bilan taqqoslash imkonini beradi.

Prague Dependency Treebank va Universal Dependencies bog‘liqligi: PDT Universal Dependenciesning oldingi ko‘rinishlaridan biri sifatida qaraladi. Chex tilining PDT orqali yaratilgan tahlili UDning rivojlanishiga katta hissa qo‘shgan. UD ko‘plab tamoyillarini PDTdan ilhomlangan holda ishlab chiqilgan.

4. TIGER Corpus. Nemis tilidagi konstituent daraxtlari asosida tuzilgan treebank. Quyidagi jadvalda Treebanklarning tasnifi bo‘yicha ma’lumotlar keltirilgan.

1-jadval Treebanklarning tasnifi

Tasnif me’zoni	***Turi***	***Izoh***	***Misollar***
Sintaktik nazariya asosda	*Konstituent daraxtlar*	*Gapning tarkibiy qismlarini (NP, VP, PP) tahlil qilishda foydalaniladi*	*Penn Treebank, Cambridge Treebank*
	Bog‘liqlik daraxtlari	*So‘zlarning o‘rtasidagi munosabatlarni tasvirlaydi*	*Universal dependencies,* *Prague dependency treebank*
Tilning o‘ziga xosligiga qarab	*Bir tilga oid Treebanklar (monolingual)*	*Faqat bir tilning ma’lumotlarini o‘z ichiga oladi*	*English, Penn Treebank, Tatar Treebank*
	Ko‘p tilli Treebanklar (multilingual)	*Bir nechta tillarni qamrab oladi*	*Universal dependencies (UD)*
Annotatsiya darajasiga qarab	*To‘liq annotatsiyalangan treebanklar*	*Sintaktik semantik tuzilmani to‘liq o‘z ichiga oladi*	*Prague dependency Treebank*
	Qisman annotatsiyalangan Treebanklar	*Faqatgina sintaktik yoki cheklangan darajadagi ma’lumotlarni o‘z ichiga oladi*	*Treebank-3*

Ham semantik, ham sintaktik ma’lumotlarni o‘z ichiga olgan daraxt banklari zamonaviy tillarda semantik-sintaktik bog‘lanishlarni modellashtirish uchun juda muhim hisoblanadi. Ular murakkab va qo‘shma gaplarda turli semantik munosabatlarni (masalan, faktlarni) ifodalovchi bo‘laklar orasidagi bog‘lanishlarni, gap uzunligini va turlarini aniqlashga xizmat qiladi. Bunday banklarga misol sifatida Groningen Meaning Bank va Treebank Semantics Parsed Corpusni keltirish mumkin. Shuningdek, semantik daraxt banklarini taqqoslash natijalari 2-jadvalda keltirilgan. Ushbu jadvallardan ko‘rinib turibdiki, barcha semantik daraxt banklari semantik ma’lumotlarni o‘z ichiga oladi. Ularning aksariyati leksik ma’lumotlarni ham o‘z ichiga oladi, ammo UCCA va AMR loyihalari bundan mustasno. Chunki ular tilga bog‘liq bo‘lmagan semantik tavsiflarni yaratishga

qaratilgan. Sintaktik ma'lumotlar asosan konstitutsiya daraxtlari shaklida taqdim etiladi, bunda FrameNet alohida e'tiborga loyiqdir, shu sababli u so'zlar orasidagi bog'lanishlarni o'rganishga qaratilganligi sababli ma'lumotlarni bog'liqlik daraxtlariga yaqin formatda saqlaydi. Hikoyaviy ma'lumotlar esa kamdan-kam hollarda taqdim etiladi, chunki ushbu banklardagi matnlarning faqat kichik bir qismi hikoyalarni ifodalaydi.

Quyidagi jadvalda Semantik Treebanklarning tasnifi bo'yicha ma'lumotlar keltirilgan.

2-jadval Semantik Treebanklarning tasnifi

Semantik Treebanklar ro'yxati	*So'zlar statistikasi*
PropBank	*Taxminan 3 500 ta murakkab fe'llar*
FrameNet	*1 224 ta grammatik shakl* *13 640 ta leksika* *202 229 ta annotatsiyalangan so'zlar to'plami*
AMR	*3 ta korpus* *47274 ta gaplar*
TSPC	*4 ta korpus* *299 ta matn* *17 796 treebank shakliga solingan gaplar*
Groningen Semantik Banki	*10 102 ta hujjatlar* *63 256 ta gap* *1 388 847 token*
UCCA	*Ingliz tili vikipediasidan (160K token)*

Xulosa

Jahon tilshunosligida treebanklar tabiiy tilni qayta ishlash (NLP) va tilshunoslik tadqiqotlarida muhim vosita sifatida o'rganilmoqda. Ular sintaktik tuzilishlarni aniq va tizimli ravishda ko'rsatish imkoniyatini yaratib, tilshunoslarga tillarning ichki strukturasi haqida chuqurroq tushuncha beradi.

Treebanklarning tasnifi ularning sintaktik nazariyalariga, annotatsiya darajasiga va tilga xos xususiyatlariga asoslanadi.

Tasniflar orqali treebanklar bir nechta toifaga bo'linadi: konstituent daraxtlar (tarkibiy qismlarni tahlil qilish), bog'liqlik daraxtlari, bir tilli va ko'ptilli treebanklar, to'liq va qisman annotatsiya qilingan treebanklar. Har bir tur o'zining alohida qo'llanish sohasiga ega bo'lib, tadqiqotlarda yoki NLP tizimlarida foydalanilishi mumkin. Treebank tizimini rivojlantirish uchun mavjud korpuslarni matnli ma'lumotlar bilan boyitish yoki konteksli ma'lumotlarga boy yangi

korpus yaratishga harakat qilish kerak. Matnli ma’lumotlar bilan kengaytirilgan semantik daraxt banklari sxemalaridan foydalanib, korpus teglarini kengaytirish va matn ma’lumotlarini tahlil qilish kabi muhim bog‘lanishlarni aniqlashga qaratilgan mashinaviy o‘qitish algoritmlarini qo‘llash mumkin.

Kelajakda ko‘plab tillar va ko‘ptilli treebanklarni yaratish, annotatsiya jarayonini avtomatlashtirish va yangi lingvistik yondashuvlarni ishlab chiqish, bu sohaning rivojiga hissa qo‘shadi. Shuningdek, treebanklar yordamida tilshunoslikni chuqurroq o‘rganish, mashinaviy tarjima, nutqni tanib olish va boshqa NLP ilovalarini rivojlantirishda muhim qadamlar qo‘yilishi kutilmoqda. Treebanklar nafaqat ilmiy tadqiqotlar uchun, balki amaliy dastur va tizimlarni yaratishda ham katta ahamiyatga ega.

Foydalanilgan adabiyotlar

Bamman D., Crane G. Treebanking Ancient Greek: The Perseus Project. Literary and Linguistic Computing, 26(1), 1-16. 2011.

Hajič J., et al. The Prague Dependency Treebank: A New Version. In Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014.

Handbook of Linguistic Annotation. Journal of Quantitative Linguistics. https://doi.org/10.1080/09296174.2018.1424495

Marcus M.P., Santorini B., Marcinkiewicz M.A. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2). 1993, p. 313-330.

Nivre J., et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016.

Sampson, G. ‘Reflections of a dendrographer.’ In A. Wilson, P. Rayson and T. McEnery (eds.) Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech, Frankfurt am Main: Peter Lang, pp. 157-184. 2003.

Silveira R., et al. A Detailed Study of the Stanford Dependency Treebank. In Proceedings of the ACL Workshop on Linguistic Annotation, 2014.

Tsvetkov Y., Dredze M. Evaluating Cross-Lingual Word Embeddings with Multilingual Treebank Data. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language

Processing, EMNLP 2015.

Universal Dependencies (UD). Universal Dependencies Documentation, 2024. Retrieved from https://universaldependencies.org/

Xia F., Palmer M. A Survey of Treebanking Tools and Techniques. Language Resources and Evaluation, 35(2), 201-226. 2001.

Xusainova Z.Y. BPE algoritmi asosida tokenizatsiya jarayonini amalga oshirish // O‘zbekiston milliy universiteti xabarlari jurnali. Toshkent, 2023. №1/3/1. – B.296-298.

Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash // “O‘zbek amaliy filologiyasi istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiyasi – Toshkent, 2022. №.1. – B. 154-163.

THE CLASSIFICATION OF TREEBANKS IN WORLD LINGUISTICS

Oqila Abdullayeva[1],
O‘g‘iloy Bozorqulova[2]

Abstract. Treebanks are of great importance in linguistics and are an important resource for natural language automation, computer linguistics, machine translation and linguistic research. This article will talk about the types, classification and areas of their application of treebanks in World linguistics.

Key words: *language corpus, linguistic tagging, Prague Treebank, Universal Treebank, English Treebank, morphological analysis, POS tagging, syntactic analysis, semantic analysis.*

References

Bamman D., Crane G. Treebanking Ancient Greek: The Perseus Project. Literary and Linguistic Computing, 26(1), 1-16. 2011.

Hajič J., et al. The Prague Dependency Treebank: A New Version. In Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014.

Handbook of Linguistic Annotation. Journal of Quantitative Linguistics. https://doi.org/10.1080/09296174.2018.1424495

Marcus M.P., Santorini B., Marcinkiewicz M.A. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2). 1993, p. 313-330.

Nivre J., et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016.

Sampson, G. ‘Reflections of a dendrographer.’ In A. Wilson, P. Rayson and T. McEnery (eds.) Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech, Frankfurt am Main: Peter

[1]*Abdullayeva Oqila Xolmo‘minovna – PhD, post-doctorate student at Tashkent State University of Uzbek Language and Literature named after Alisher Navo’i.*
E-pochta: *abdullayeva.oqila@navoiy-uni.uz*
ORCID: 0000-0002-2524-4832
[2]*Bozorqulova O‘g‘iloy Erkin qizi - Master of degree. Alisher Navo’i Tashkent State University of Uzbek Language and Literature.*
E-pochta: *ogiloybozorqulova62@gmail.com*

Lang, pp. 157-184. 2003.

Silveira R., et al. A Detailed Study of the Stanford Dependency Treebank. In Proceedings of the ACL Workshop on Linguistic Annotation, 2014.

Tsvetkov Y., Dredze M. Evaluating Cross-Lingual Word Embeddings with Multilingual Treebank Data. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015.

Universal Dependencies (UD). Universal Dependencies Documentation, 2024. Retrieved from https://universaldependencies.org/

Xia F., Palmer M. A Survey of Treebanking Tools and Techniques. Language Resources and Evaluation, 35(2), 201-226. 2001.

Xusainova Z.Y. BPE algoritmi asosida tokenizatsiya jarayonini amalga oshirish // O‘zbekiston milliy universiteti xabarlari jurnali. Toshkent, 2023. №1/3/1. – B. 296-298.

Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash // “O‘zbek amaliy filologiyasi istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiyasi – Toshkent, 2022. №.1. – B.154-163.

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

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

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

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

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

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