O‘ZBEK TILI MATNLARIDA PUNKTUATSION TAHLIL QILISH UCHUN KORPUS YARATISH

Sharipov Maqsud Siddiqovich
Urganch davlat universiteti, Kompyuter ilmlari kafedrasi dotsenti
Adinayev Xushnudbek Saylboyevich
Muhammad al-Xorazmiy nomidagi TATU Urganch filiali o‘quv-uslubiy boshqarma boshlig‘i
Yusupova Mukaddas Muratovna
Urganch Ranch texnologiya universiteti 1-kurs magistranti

Annotatsiya. Biz o‘zbek tili matnlarida tinish belgilarini tahlil qilish uchun yangi korpusni taqdim etamiz. Punktuatsiya korpusi – bu matnlarni to‘g‘ri tinish belgilar bilan belgilangan holda yig‘ilgan ma’lumotlar to‘plami bo‘lib, u tabiiy tilni qayta ishlash (NLP) va punktuatsiyani tiklash modellari uchun ishlatiladi. Punktuatsiya korpusini yaratish uchun matnlarni badiiy kitoblar, maktab darsliklari, maqolalar va internet saytlaridan olindi.

Kalit so‘zlar. Tinish belgilari, NLP, korpus, ASR.

Аннотация. Мы представляем новый корпус для анализа знаков препинания в узбекских текстах. Корпус пунктуации – это коллекция текстов, отмеченных правильными знаками препинания, которая используется для обработки естественного языка (NLP) и моделей восстановления пунктуации. Для создания пунктуационного корпуса были взяты тексты из художественной литературы, школьных учебников, статей и веб-сайтов.

Ключевые слова. Пунктуация, НЛП, корпус, ASR.

Abstract. We present a new corpus for analyzing punctuation in Uzbek texts. A punctuation corpus is a collection of texts marked with correct punctuation marks, which is used for natural language processing (NLP) and punctuation restoration models. To create the punctuation corpus, texts were taken from fiction books, school textbooks, articles, and websites.

Keywords: Punctuation, NLP, corpus, ASR.

Kirish

Punktuatsiya korpusi matnlardan tinish belgilarini avtomatik tiklash, nutqdan matnga (ASR) tizimlari uchun punktuatsiya qo‘shish, mashinaviy tarjima va NLP modellari sifatini oshirish, Avtomatik yozuvni yaxshilash va grammatik tahlil qilish sohalarida qo‘llanilishi mumkin. Ko‘plab raqamli audio va video ma’lumotlar har kuni efirga uzatiladigan yangiliklar va teleradiostansiyalar kabi ommaviy axborot vositalari tomonidan ishlab chiqariladi. Nutqni kirish sifatida qabul qiladigan va tegishli matnni ishlab chiqaradigan avtomatik nutqni aniqlash (ASR) tizimlari bunday manbalarga ma’lumot qidirish, ma’lumot olish, mashina tarjimasi,

ISSN 3030-3907

indekslash va sarlavha qo‘yish kabi turli vazifalarga tayyor bo‘lish uchun qo‘llanilishi mumkin.

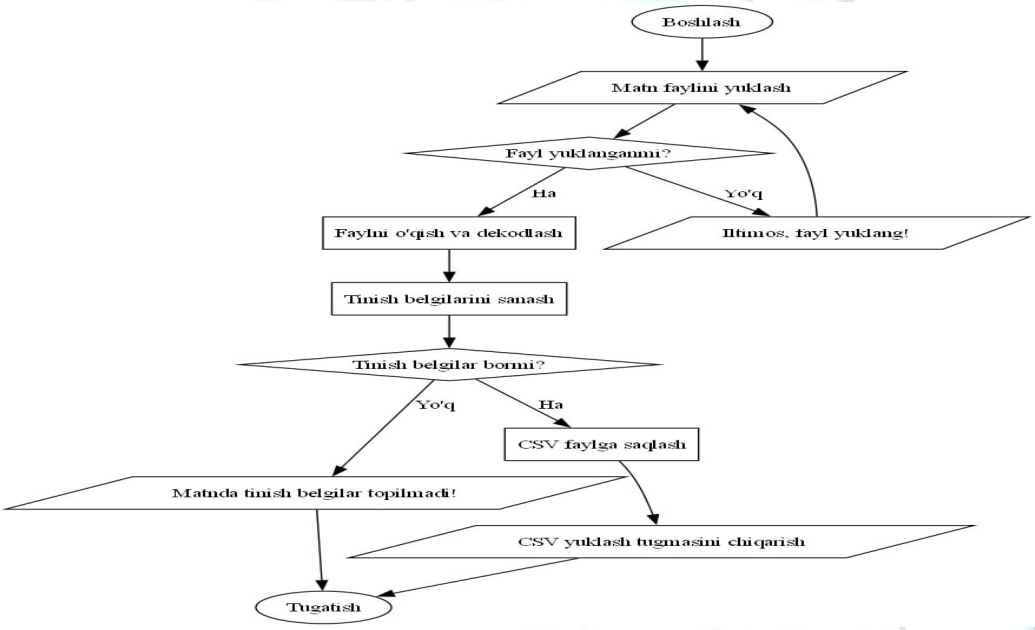
Deyarli barcha zamonaviy ASR tizimlari oddiygina tinish belgilarini o‘z ichiga olmaydi va ko‘pincha kichik harflar formatida so‘zlarning boshlang‘ich ketma-ketligini ishlab chiqaradi. Bu ma’lumotni qidirish va tasniflash kabi vazifalar uchun foydali bo‘lishi mumkin bo‘lsa-da, ASR chiqishi tinish belgilari va to‘g‘ri bosh harflar [1] kabi boshqa ma’lumotlardan foydalanishi mumkin. Ushbu maqolada biz o‘zbek tili matnlardagi tinish belgilarini tahlil qilish vazifasi uchun matnlarni yig‘ishni maqsad qilganmiz. Bizga ma’lum bo‘lishicha, o‘zbek tili uchun hozirgacha bunday mos korpus ishlab chiqilmagan.

METHODOLOGIYASI

Yaratilgan korpus quyidagicha ishlatilishi mumkin:

- Mashina o‘rganish – Punktuatsiyani tiklash modellari uchun.
- ASR tizimlari – Nutqni matnga o‘girishdan tinish belgilarini to‘g‘ri qo‘shish masalasida.
- Grammatik tekshirish – Matnlarni avtomatik tekshirish va tahrir qilishda.

Korpusni tinish belgilari bo‘yicha statistik tahlil qilish uchun quyidagicha algoritm va unga mos dastur kodi ishlab chiqildi.




```
import re
import pandas as pd
import ipywidgets as widgets
from IPython.display import display
from google.colab import files
def count_punctuation(text):
```


ISSN 3030-3907

```
    punctuation_marks = [".", ",", ":", "!", "?", "–", "(", ")", ";","..."]  # Tinish belgilar ro‘yxati
        punctuation_counts = {mark: 0 for mark in punctuation_marks}
        for mark in punctuation_marks:
        punctuation_counts[mark] = text.count(mark)
        return {k: v for k, v in punctuation_counts.items() if v > 0}
def upload_and_process(change):
        uploaded_file = file_upload.value
        if uploaded_file:
        file_content = list(uploaded_file.values())[0]['content'].decode('utf-8')  # UTF-8 dekodlash
        result = count_punctuation(file_content)
        if not result:
        print("Matnda tinish belgilar topilmadi!")
        else:
        df = pd.DataFrame(list(result.items()), columns=["Belgi", "Soni"])
        display(df)
        df.to_csv("punctuation_count.csv", index=False)
        print("Natijalar CSV faylga saqlandi. Faylni yuklab olish uchun quyidagi tugmani bosing:")
        download_button = widgets.Button(description="CSV faylni yuklab olish")
        display(download_button)
        download_button.on_click(lambda x: files.download("punctuation_count.csv"))
        else:
        print("Iltimos, fayl yuklang!")
file_upload = widgets.FileUpload(description="Matn faylini yuklash", accept='.txt', multiple=False)
process_button = widgets.Button(description="Tahlil qilish")
process_button.on_click(upload_and_process)
display(file_upload, process_button)
```

Bu dastur yordamida o‘zbek tili matnlarida tinish belgilari qatnashishi bo‘yicha tahli o‘tkazish mumkin.

NATIJA

Korpusda mavjud bo‘lgan 30 turdagi matn, 137 848 gap va 1 642 860 so‘zdan iborat matnlar dastur yordamida statistic tahli amalga oshirildi. Korpusda barcha tinish belgilari statitikasi 1-jadvalda keltirilgan.

ISSN 3030-3907

1-jadval

№	Matn turlari	Undov belgilari soni	So‘roq belgilari soni	Vergullar soni	Nuqtalar soni	Tirelar soni	Qo‘shtirnoqlar soni	Qavslar soni	Ko‘p nuqtalar soni	Ikki nuqtalar soni	Nuqtali vergullar soni
1.	Adabiyot	1143	969	8315	9948	2438	1777	478	771	485	91
2.	Anatomiya	1	0	7649	10565	3608	138	1253	0	837	2407
3.	Biologiya	0	222	3849	4522	793	496	605	0	245	307
4.	Botonika	2	229	2104	2480	5	128	351	0	30	12
5.	Diniy	20	209	4098	5533	297	940	476	30	185	108
6.	Fizika	6	1487	5904	10591	2	26	1124	3	609	91
7.	Geografiya	0	250	7134	5157	518	386	867	0	54	26
8.	Huquq	5	120	1422	1388	50	99	53	11	98	154
9.	Ijtimoiy soha	0	11	325	332	7	101	9	1	35	9
10.	Informatika	22	82	784	2268	40	183	112	1	183	140
11.	Iqtisodiyot	1	0	362	305	9	53	10	0	8	9
12.	Kimyo	18	334	2549	3298	451	103	571	6	444	158
13.	Kino	3	1	449	368	27	141	19	0	5	7
14.	Matematika	22	580	4656	9683	3	150	1938	278	1064	1736
15.	O‘qish kitobi	193	420	2528	2762	0	264	7	88	265	11
16.	O‘zbek tili kitob	56	297	11984	10992	1415	570	2488	37	1861	723
17.	Qishloq xo‘jaligi	0	0	1751	1167	0	45	141	0	17	15
18.	Qissa	1729	711	8874	8974	0	502	4	202	571	46
19.	Qonun va qarorlar	1	0	3308	3228	0	62	723	0	207	958
20.	Rassomchilik	12	29	4457	3576	0	409	270	2	157	97
21.	Roman	80	141	1618	1623	382	188	24	213	51	1

ISSN 3030-3907

22.	San’at	129	201	12086	7852	372	1309	711	122	0	824
23.	Siyosat	7	7	673	704	45	30	31	0	19	14
24.	Sport	18	76	2421	2750	20	218	336	1	114	97
25.	Tabobat	1	6	412	385	7	30	22	0	24	7
26.	Tarix	5	333	6159	4286	2961	715	355	4	187	286
27.	Texnologiya	7	78	372	713	49	11	30	0	23	45
28.	Tibbiyot	0	214	4243	5330	418	35	2277	0	191	113
29.	Urush	10	40	947	1034	8	229	32	6	46	4
30.	Zoologiya	1	432	2476	4009	237	28	254	2	196	493
	jami	**3492**	**7479**	**113909**	**125823**	**14162**	**9366**	**15571**	**178**	**8211**	**8989**

Xulosa

O‘zbek tili punktuatsiya korpusini ishlab chiqish tabiiy tilni qayta ishlash va avtomatik punktuatsiya tizimlarini rivojlantirish uchun muhim jarayondir. To‘g‘ri yig‘ilgan va annotatsiya qilingan korpus yuqori sifatli NLP modellarini ishlab chiqishga yordam beradi. Punktuatsiya korpusimiz tinish belgilarining qo‘llanilishini o‘rganish va avtomatlashtirish uchun tuzilgan maxsus matnlar to‘plamidan iborat. Bu korpus mashinada o‘qiladigan formatda bo‘lib, avtomatik punktuatsiya tiklash, nutqni matnga aylantirish, grammatik xatolarni tekshirish va mashinali tarjima kabi jarayonlarda muhim rol o‘ynaydi. Bu korpus sun’iy intellekt va tabiiy tilni qayta ishlash (NLP) sohasida katta foyda keltirishi mumkin.

Ishlab chiqilgan korpusimizda o‘zbek tilining 30 ta turdan iborat 1 642 860 ta so‘z va 137 848 ta gapdan iborat. Ushbu ishning davomi sifatida keljakda o‘zbek tili matnlarini punktuatsion tahlil qilish masalasini sun’iy intellekt modellaridan foydalanib yechish va yaxshiroq aniqlikdagi natijalarni olish maqsad qilingan.

FOYADALANILGAN ADABIYOTLAR RO‘YXATI

[1] M. A. , M. N. , A. D. Mahmudov N., *O‘ZBEK TILI ME’YORLARI (Punktuatsiya)*, vol. 1. Tashkent: Zamin nashr, 2021.

[2] J. Pokrywka, “Punctuation Prediction for Polish Texts using Transformers,” Feb. 2024. doi: 10.48550/arXiv.2410.04621.

[3] P. Żelasko, P. Szymański, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, “Punctuation Prediction Model for Conversational Speech,” Feb. 2018, pp. 2633–2637. doi: 10.21437/Interspeech.2018-1096.

[4] M. S. Sharipov, H. S. Adinaev, and E. R. Kuriyozov, “Rule-Based Punctuation Algorithm for the Uzbek Language,” in *International Conference of Young*

ISSN 3030-3907

Specialists on Micro/Nanotechnologies and Electron Devices, EDM, 2024, pp. 2410 – 2414. doi: 10.1109/EDM61683.2024.10615061.

[5] M. Bayraktar, B. Say, and V. Akman, “An analysis of English punctuation: the special case of comma,” International Journal of Corpus Linguistics, vol. 3, Jul. 1998, doi: 10.1075/ijcl.3.1.03bay.

[6] D. Hardt, “Comma checking in Danish,” 2001.

[7] U. Salaev, E. Kuriyozov, and C. Gomez-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: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138700420&partnerID=40&md5=bf476cd74317f06577dd0548c5c600d6

[8] A. M. Abdurashetona and I. O. Ismailovich, “Methods of Tagging Part of Speech of Uzbek Language,” in Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021, 2021, pp. 82 – 85. doi: 10.1109/UBMK52708.2021.9558900.

[9] A. M. Abdurashetona and U. Mokhiyakon, “Software Features and Linguistic Features of Uzbek Synonymizer,” in Proceedings - 7th International Conference on Computer Science and Engineering, UBMK 2022, 2022, pp. 171 – 175. doi: 10.1109/UBMK55850.2022.9919447.

[10] B. Mengliyev, S. Shahabitdinova, S. Khamroeva, S. Gulyamova, and A. Botirova, “The morphological analysis and synthesis of word forms in the linguistic analyzer,” Journal of Language and Linguistic Studies, vol. 17, no. 1, pp. 558 – 564, 2021, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103797274&partnerID=40&md5=8a4052419f721c3c734f8fe1c48984ec

[11] K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of Karakalpak language stop words,” Data Brief, vol. 48, 2023, doi: 10.1016/j.dib.2023.109111.

[12] M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, ‘Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language’, CEUR Workshop Proceedings, vol. 3315, pp. 93–98, 2022.

[13] D. Mengliev, E. Akhmedov, V. Barakhnin, Z. Hakimov, and O. Alloyorov, “Utilizing Lexicographic Resources for Sentiment Classification in Uzbek Language,” Jan. 2023, pp. 1720–1724. doi: 10.1109/APEIE59731.2023.10347765.

[14] K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of stopwords extracted from Uzbek texts”, Data in Brief, vol. 43, 2022.

ISSN 3030-3907