IEEE
TÜRKİYE SECTION

UBMK'24

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

9th International Conference on Computer Science and Engineering

26-27-28 Ekim (October) 2024 Antalya - Türkiye

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK'2024)

9th International Conference on Computer Science and Engineering

26-28 Ekim 2024 Akdeniz Üniversitesi Antalya Türkiye
26-28 October 2024 Akdeniz University Antalya Türkiye

Telif Hakkı

Bu elektronik kitabın içinde yer alan tüm bildirilerin telif hakları IEEE'ye devredilmiştir. Bu kitabın tamamı veya herhangi bir kısmı yayımcının izni olmaksızın yayımlanamaz, basılı veya elektronik biçimde çoğaltılamaz. Tersi davranışta bulunanlara ABD Telif Hakkı Yasalarına göre ceza uygulanır.

Copyright and Reprint Permission

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. Copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid throught Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For reprint or republication permission, email to IEEE Copyright Manager at pubs-permission@ieee.org

All right reserved. Copyright C 2024

IEEE Catalog Number : CFP24L97-CDR

ISBN : 979-8-3503-6587-0

Additional copies may be ordered from:
Curran Associates, Inc
57 Morehouse Lane Red Hook, NY 12571 USA
Phone: (845) 758 0400
Fax: (845) 758 2633
E-mail: curran@proceeding.com

	ID	Title	Authors	Pages
NLP	1850	HTV-News: A New Dataset with High Novelty Rate for Turkish Text Summarization	Nihal Zuhal Kayalı Sevinç İlhan Omurca	1 - 6
	1853	Sentiment Analysis of Social Network Comments in Uzbek Language	Saboxat Allanazarova	7 - 10
	1861	An Online Platform for Uzbek-Russian and Russian-Uzbek Parallel Corpora	Nigmatova Lolakhon Khamidovna Saidova Mokhira Rasulevna Djuraeva Zulkhumor Radzhabovna Sharipov Sokhib Salimovich Avezov Sukhrob Sobirovich	11 - 16
	1863	Yapım Ekleri ile Özbekçe Sözcük Türetme Deriving Uzbek Words with Derivational Suffixes	Eşref Adalı Khamroeva Shahlo Mirdjonovna	17 - 21
	1864	Database of Grammatical Form and Meaning Correspondence for Uzbek-English Translation Software	Jamoliddinova Odina Rasulovna Muhammedova Saodat Xudoyberdiyevna Khamroeva Shahlo Mirdjanovna Ergasheva Guli Ismoil Qizi Qodirova Madinabonu Murodjon qizi Kholova Muyassar	22 - 26
	1868	Words Speak Louder Than Actions: Decoding Emotions Through NLP	Melda Paksoy Gokhan Bakal	27 - 31
	1872	Özbek Dilinin Pankronik Derleminin Oluşturulması (Kutadgu bilig' Nüshalarına Göre)	Xamroyeva Şahlo Mircanovna Avezova Nigora Safarovna Toştemirova Sitora Bahodir kızı Qodirova Fazilat Şukurovna Atanazarova Shoxida Baxramovna	32 - 37
	1873	Preserving Semantic Integrity in Paraphrasing Texts: Rule-Based Paraphrasing	Zarnigor M. Khayatova Jurakuziyev Nodirbek Kholova Muyassar	38 - 42
	1975	NLLB-Based Uzbek NMT: Leveraging Multisource Data	Nilufar Abdurakhmonova Adkham Zokhirov Mohirdev Mukhammadali Salokhiddinov Anvar Narzullayev Ayrat Gatiatullin	43 - 47
	1976	The Problem of the Archaic Words' Semantic Description in the Alisher Navoi Authorship Corpus	Shuhrat Sirojiddinov Manzura Abjalova Sulton Normamatov Nargiza Gulomova	48 - 53
	1979	Multi-Label Emotion Classification with Fine-Tuned BERT and Contrastive Learning	Ammara Naseem Khan Mehmet Nafiz Aydın	54 - 59
	1885	Algorithm for Aligning Paragraphs and Sentences in Aligner Tool	Elov Botir Boltayevich Khamroeva Shahlo Mirdjonovna Dauletov Adilbek Yusupbayevich Matyakubova Noila Shakirjanovna	60 - 63
	1888	Semantic and Grammatical Issues in Translating Idioms with Automatic Translation Systems	Manzura Abjalova Sarvinoz Sharipova	64 - 69
	1889	Improving Methodology for Selecting Fiction Works for Grades 5-9	Khabibulla Madatov Sapura Sattarova	70 - 73
	1897	Innovating SQL Automation: Evaluating Open-Source Large Language Models with a Dual-Stage Approach for Corporate Data Solutions	Erhan Arslan Eugen Harinda	74 - 79
	1898	Development of a Method for Automatic Extraction of Ontology Entity Names from Natural Language Texts	Sadirmekova Zhanna Elmira Daiyrbayeva Karbozova Indira Bekbolatov Samat	80 - 85
	1899	UZBEKCORPORA.UZ: Creation of Author Corpora on the Platform	Karshiev Abduvali Berkinovich Karimov Suyun Amirovich Tursunov Mukhammadsolikh Sadin ugli Kudratov Iskandar Makhmidov Shirinboy B otirovich	86 - 90
	1901	A Technique for Automatic Extraction of Basis Words: a Case Study on "Uzbek Primary School Corpus"	Khabibulla Madatov	91 - 94

		Shukurla Bekchanov Surayyo Khajibaeva	
1908	Özbek Dili Metinlerindeki Eşgönderge Çözümlemesi Algoritması Algorithm of Coreference Resolution in Uzbek Texts	Elov Botir Boltayeviç Abdisalomova Şahlo Abdimurod kızı Karimova Xurriyat Şaripovna	95 - 100
1909	The Process of Lemmatization and Stemming in the Automatic Morphological Analysis of Uzbek Texts	Manzura Abjalova Eşref Adalı Munojot Adilova	101 - 106
1913	Turkish Question - Answer Dataset Evaluated with Deep Learning	Kadir Tutar Olcay Taner Yıldız	107 - 111
1914	Exploring the Semantic Complexity of Adjective-noun Collocations Between Uzbek and English for Improved Machine Translation	Nilufar Abdurakhmonova Nargiza Shamieva Eşref Adalı	112 - 115
1915	Linguistic Support and Algorithm of Phrase Alignment in Aligner Tool	Xoshimovna Shohida Shahabitdinova Matyakubova Noila Shakirjanovna	116 - 120
1946	Evaluating the Performance of PEFT and LoRA Approaches in Transformer-based Architecture for Handwritten Character Recognition	Pinar Savci Bihter Das Resul Das	121 - 126
1947	Evaluating the Performance of Turkish Automatic Speech Recognition Using the Generative AI-based Whisper Model	Tunahan Gokcimen Bihter Das Resul Das	127 - 131
1961	Özbek Dilinde Kelimelerin Kök ve Köklerini Ayırma Algoritmaları Algorithms for Parsing Roots and Stems of Words in Uzbek Language	Elov Botir Boltayeviç Xusainova Zilola Yuldashevna Umirova Svetlana Mamurjonovna Normamatov Sultonbek A.Şahlo Abdimurod kızı Mahmadiev Shavkatjon	132 - 136
1970	Özbekçe Deyimlerin Dilsel Modellemesinin DDİ Açısından Önemi The Importance of Linguistic Modeling of Uzbek Idioms in NLP	Gülyamova Shahnoza Kahramonovna Nurboyeva Maftuna Vahobcon kızı	137 - 140
1981	IT Service Desk Ticket Classification via Large Language Models	Ezgi Paket Göksu Şenerkek Fatma Betül Akyol Furkan Salman	141 - 146
1998	Enhancing the Quality of Clustering Job Skills Through the Fine-tuning of the Sentence Transformer Model	V. S. Ramazanova A. S. Yerimbetova M. A. Sambetbaeva Zh. B. Sadirmekova S. K. Serikbayeva	147 - 152
2004	Named Entity Recognition from Kazakh Speech	Bauyrzhan Kairatuly Madina Mansurova	153 -156
2005	Sentiment Analysis of Reviews for E-Commerce Applications	Elif Hanife Aydoğan Feyza Yıldırım Okay	157 - 162
2011	LLM and RAG-Based Question Answering Assistant for Enterprise Knowledge Management	Gürkan Şahin Karya Varol Burcu Kuleli Pak	163 - 168
2018	Named Entities Recognition in Kazakh Text by SpaCy NER Models	Nurzhan Mukazhanov Aigerim Yerimbetova Mussa Turdalyuly Bakzhan Sakenov	169 - 174
2022	Efficient Unsupervised Domain Adaptation with	M. Rawhani D. Karaboğa U. Nalbantoğlu A. Baştürk B. Akay	175 - 180
2032	Aligning Sentences for Kazakh-Turkish Parallel Corpora	Diana Rakhimova Eşref Adalı Aidana Karibayeva	181 - 186
2048	Automated-Computer Based Assessment of Free-Text Exam Answers By Transformers	Burak Keskin Melih Günay	187 - 192

	2073	Knowledge-Augmented Large Language Model	Başak Buluz Kömeçoğlu Burcu Yılmaz	193 - 198
	2092	LLMs for Document-level Text Simplification in Turkish Foreign Language Learning	Fatih Bektaş Kutay Arda Dinç Gülşen Eryiğit	199 - 203
	2105	Mining Software Requirements From Turkish Texts:Techniques and Challenges	Serhat Uzunbayir Senem Kumova Metin	204 - 209
	2108	Using LLMs for Annotation and ML Methods for Comparative Analysis of Large-Scale Turkish Sentiment Datasets	Ercan Ezin Rukiye Savran Kiziltepe Murat Karakus	210 - 215
	2122	Sentiment Analysis for Hotel Reviews in Turkish by using LLMs	Ata Onur Özdemir Efe Batur Giritli Yekta Said Can	216 - 220
	2123	Semantic Knowledge Base for the Emotional Coloring Analysis of Kazakh Texts	Banu Yergesh Dinara Kabdylova Tilekbergen Mukhamet	221 - 224
	2140	Adaptive Batch Budget for LLM Inference	Cağrı Yeşil Berhan Turku Ay Funda Ay Ak Öykü Berfin Mercan Oğuzhan Nefesoğlu	225 - 229
	2154	Open Data Processing for Analysis of Behavior of Users in Digital Environment	Nikolai Prokopyev Pavel Ustin Fail Gafarov	230 - 232
	2164	Evaluating Turkish BERT-based Language Models for Effective Customer Feedback Interpretation in CRM	Can İşcan Muhammet Furkan Özara Ahmet Erkan Çelik Akhan Akbulut	233 - 238
	2175	Does Prompt Engineering Help Turkish Named Entity Recognition?	Arda Serdar Pektezol Ahmet Batuhan Ulugergerli Volkan Öztoklu Seniz Demir	239 - 243
	2188	Araneum Uzbecicum: A Gigaword Web-Crawled Uzbek Corpus	Vladimi-r Benko Radovan Garabik Shahlo Khamroyeva	244 - 2149
	2189	Automatic Topic Detection in Large Text Data in Uzbek using Clustering Methods	Umirova Svetlana Kholmuhamedov Bakhtiyor Karimov Suyun Narzieva Mamura	250 - 255
	2193	Machine Learning-Based Synonymous Word	Diana Rakhimova Madina Mansurova Nurakhmet Matanov Amira Yerik	256 - 260
	2194	Enhancing Text-to-SQL Conversion in Turkish: An Analysis of LLMs with Schema Context	Ferhat D emirkıran Ali Kemal Coşkun Yavuz Kömeçoğlu Başak Buluz Kömeçoğlu Ramazan Güven	261 - 265
	2203	Linguistic Knowledge Graph "Turklang" as Universal Model for Linguistic Resources and Tools in Turkic Languages	Ayrat Gatiatullin Nikolai Prokopyev Lenara Kubedinova Nilufar Abdurakhmonova Rustam Burnashev	266 - 270
	2216	Requirements for the Development of a Website Builder with Adaptive Design	Gulmira Bekmanova Banu Yergesh Assel Omarbekova Laura Orynbay Aiganym Bessembayeva Dinara Kabdylova Altanbek Zulkhazhav Beibarys Sultan	271 - 276
AI	**1870**	A Modified Marine Predators Algorithm for Numerical Function Optimization	Resul Özdemir Murat Taşyürek	277 - 281

		Veysel Aslantaş	
1884	Document Classification and Key Information Extraction Using Multimodal Transformers	Mehmet Selman Baysan Furkan Kızılay Ayşe İrem Özmen Gökhan İnce	282 - 287
1926	Employee Turnover Prediction on Synthetic and Real Datasets	Azhar Murzaeva Volkan Illik Kadir Yunus Koc	288 - 292
1937	Data Augmentation in Remote Sensing Image Change Captioning	Orkhan Karimli Ilyas Mustafazade Ali Can Karaca Fatih Amasyalı	293 - 298
1968	Comparative Evaluation of Word2Vec and Node2Vec for Frequently Bought Together Recommendations in E-commerce	Mustafa Keskin Enis Teper Alya Kurt	299 - 303
1971	Disrupting the Retrieval Phase in RAG-Based LLM Chatbots through Input Manipulation	Felix Pitterling Georgii Koshelev Oliver Haddad Rodrigo Teran	304 - 308
1976	Mutation Testing Reinvented: How Artificial Intelligence Complements Classic Methods	Serhat Uzunbayir Kaan Kurtel	309 - 314
1997	Prompting Large Language Models for Aerial Navigation	Emirhan Balcı Mehmet Sarıgül Barış Ata	315 - 320
2006	Açıklanabilir Yapay Zeka Üzerine Bir İnceleme A Review of Explainable Artificial Intelligence	Samed Al Şeref Sağıroğlu	321 - 326
2020	Investigating the Effects of Pre-trained Deep Learning Models and Fusion Techniques on Fruit Segmentation Performance	Esma İbiş Aybars Uğur	327 - 332
2034	Beyin Tümörü Ameliyatlarında Beyin Kayması Probleminin Telafisi: Bilgisayar Destekli Yöntemler Uzerine Bir Derleme Compensating for the Brain Shift Problem in Brain Tumor Surgeries: A Review of Computer-Assisted Methods	Ayşe Gül Eker Meltem Kurt Pehlivanoğlu Nur Banu Albayrak Nevcihan Duru Tolga Turan Dündar	333 - 338
2039	AI-Enhanced Endometrial Cancer Diagnosis System	Raziye Aslıhan Kürkcü Işın Yeşim Yeşilkaya Baylan Erkut Attar Burcu Selçuk Tacha Serif	339 - 343
2040	El Titreme Sinyalinden Transformer Temelli Parkinson Hastalığı Sınıflandırması Transformer Based Parkinson's Disease Classification from Hand Tremor Signal	Murat Atçeken Lütfü Hanoğlu Mehmet Ersin Bitirgen Bahadır K. Güntürk İbrahim Özşeker	344 - 347
2042	Üretken Yapay Zeka Fırsatlar ve Tehditler: Bibliyometrik Analiz Generative Artificial Intelligence Opportunities and Threats: Bibliometric Analysis	Betül Ersöz Şeref Sağıroğlu Halil İbrahim Bülbül	348 - 353
2053	Users Acceptance of Generative Artificial Intelligence Based on ChatBot Use	Ayşe Yeşim Mutlu	354 - 359
2054	MMBAttn: Max-Mean and Bit-wise Attention for CTR Prediction	Hasan Saribas Cagri Yesil Serdarcan Dilbaz Halit Orenbas	360 - 365
2065	Türkçe Otomatik Konuşma Tanıma Sistemleri için Sentetik Veri Üretme Yöntemi Synthetic Data Generation Method for Turkish Automatic Speech Recognition Systems	Hilal Tekgöz Harun Uz Muhammed Murat Özbek Tolga Büyüktanır	366 - 370
2066	A Federated Learning Framework for Classifying the Images in Ultrasonic Nondestructive Testing	Abdulkadir Gulsen Hilal Hacılar	371 - 375

			Burak Kolukısa Burcu Bakir-Gungor	
	2071	AI Assisted Customer Review Sentiment Analysis and Department Classification Tool	Burcu Seçuk Berat Ilgaz Dursun Tacha Serif	376 - 381
	2075	Histopatolojik Görüntülerde Doğru Mitoz Tespiti için Geliştirilmiş Renk Normalleştirme Yöntemi Enhanced Stain Normalization Method for Accurate Mitosis Detection in Histopathological Images	Refik Samet Nooshin Nemati Emrah Hançer Serpil Sak Bilge Ayça Kırmızı	382 - 387
	2110	ATGRUVAE: Reducing Noise and Improving Forecasting Performance in Stock Data	Hüseyin Akkas Burak Kolukisa Burcu Bakir-Gungor	388 - 392
	2113	Evaluating the Impact of Sentiment Analysis on Deep Reinforcement Learning-Based Trading Strategies	Mustafa Etcil Burak Kolukisa Burcu Bakir-Gungor	393 - 398
	2118	PsychSynth: Advancing Mental Health AI Through Synthetic Data Generation and Curriculum Training	Vedanta S P Madhav Rao	399 - 404
	2121	Covid-19 Bilgisayarlı Tomografi Görüntülerinin Segmentasyonu için Topluluk Öğrenmesiyle Birleştirilmiş U-Net Modeli U-Net Model Combined with Ensemble Learning for Segmentation of Covid-19 Computed Tomography Images	Furkan Atlan İhsan Pençe	405 - 410
	2129	Turkish Sign Language Video Generation from Pose Sequences Using a Conditional GAN Model	Feyza Özkan Hatice Kübra Tekin Hacer Yalim Keles	411 - 416
	2131	Statistical Attention Layer for Neural Network Training	Ergun Biçici Hasan Saribaş	417 - 420
	2134	Dudak Şekillerinden Kişilik Analizi: Marifetname ve Açıklanabilir Yapay Zeka Yaklaşımı Personality analysis from lip shapes: Marifetnâme and explainable artificial intelligence approach	Semra Çelebi İbrahim Türkoğlu	421 - 426
	2148	Breast Cancer Detection Using Deep Learning Models	Erkan Akkaş Mustafa Yapar Şeref Sağıroğlu Ali Öter Betül Ersöz	427 - 430
	2191	CGAN Mimarisi ile Kısıtlı Veri Ortamında Kestirimci Bakım Performansının İyileştirilmesi Improving Predictive Maintenance Performance in Limited Data Environment with CGAN	Mehmet Musa Özcan Uğurhan Kutbay	433 - 438
	2207	Integrating the Focusing Neuron Model with N-BEATS and N-HiTS	Şuayb Talha Özçelik Faik Boray Tek	439 - 442
	2217	Advanced Convolutional Neural Networks for Plastic Classification in Recycling Systems	Nazym Alimbekova Ainur Zhumadillayeva Sunggat Aiymbay	443 - 447
ML	**1846**	Semi-Supervised Learning for Sensor-Based Flash Point Prediction in Oil Industry	Mert Sülük Şule Gündüz Öğüdücü	448 - 452
	1847	Robust Stacked Ensemble Model for Lung Cancer Diagnosis	Ayhan Akbas Gonca Buyrukoglu Selim Buyrukoglu	453 - 457
	1856	Kod Kusurlarının Tespitinde Makine Öğrenmesi Tekniklerinin Başarım Analizi The Performance Analysis of Machine Learning Techniques in Code Smell Detection	Arman Yavuz Oya Kalıpsız	458 - 463
	1860	Enhancing Financial Time-Series Analysis with TimeGAN: A Novel Approach	Cemal Öztürk	464 - 467
	1869	From Traditional to Deep: Evaluating Sentiment Analysis Models on a Large-Scale Tweet Dataset	Alisahib Mammadov Gokhan Bakal	468 - 473

1871	Comparison of Feature Selection Methods for Mechanical Properties of Cold Rolled Products in Flat Steel Manufacturing	Didem Bakiler İlme Merve Öper E. Fatih Yetkin	474 - 478
1883	A Machine Learning Approach to Steel Sheet Production Surface Quality	Asena Öztürk Mehmet Nafiz Aydın	479 - 484
1938	NLP Tabanlı Özellik Çıkarımına Dayalı Makine Öğrenme Algoritmaları Kullanılarak SQL Enjeksiyonu Ataklarının Tespit Edilmesi Detection of SQL Injection Attacks Using Machine Learning Algorithms Based on NLP-Based Feature Extraction	Hakan Can Altunay	485 - 489
1940	Enhancing Driver Injury Severity Prediction Using Optimized Oversampling and Feature Selection Techniques	Vahide Nida Kılıç Gizen Mutlu Esra Saraç Çiğdem Acı Murat Özen	490 - 495
1972	Predicting Electric Vehicle Adoption in the EU: Analyzing Classification Performance and Influencing Attributes Across Countries, Gender, and Education Level	Mert Kumbasar Gül Tokdemir Thouraya Gherissi Labben Gurdal Ertek	496- 499
1977	Arabam.com Kurumsal Üyelerinin Satın Alma Davranışlarının Analizi	Dilan Oragaz İsmail Duru	500 - 505
1983	Makine Öğrenmesi ve Derin Öğrenme Modelleri Kullanılarak MR Görüntülerinden Demans Sınıflandırma	Meltem Kurt Pehlivanoğlu Onur Varol Osman Aldemir Ata Emir Uncu Nevcihan Duru	506 - 511
1995	Indoor Localization with GravNetConv and Dynamic Graphs	Mert Bayraktar Alper Ozcan Umit Deniz Ulusar	512 - 516
2023	GPS Spoofing Detection on Autonomous Vehicles with XGBoost	Emre İşleyen Şerif Bahtiyar	517 - 522
2025	SmartSniffer: Predicting Food Spoilage Time with an Electronic Nose-based Gas Monitoring Apparatus Utilizing a Two-Stage Pipeline Model	Pandey Shourya Prasad Sreyas Janamanchi Barath S Narayan Madhav Rao	523 - 528
2064	GNF: Generative Natural Faces Dataset for Face Expression Recognition Task	Cansu Özer Volkan Dağlı Mustafa Kağan Gürkan	529 - 533
2081	Incorporating Knowledge Graph Embeddings into Graph Neural Networks for Sequential Recommender Systems	Kazım Emre Yüksel Susan Üsküdarlı	534 - 539
2095	The Classification of Human and Cognitive Architecture Time Estimation Using Machine Learning Methods	Behiye Şahin Sonay Duman	540 - 543
2101	Optimization of Roadside Assistance Tow Truck Services Provided to Insured Individuals with Traffic and Comprehensive Insurance Policies	Serkan Kırca Fatma Yağmur Erbaş Sedat Cebesoy Buse Özkanat Aysun Yıldırım Ceren Yıldırım	544 - 549
2104	Optimizing Recommendation Systems By Fusion of KNN, Singular Value Decomposition, and XGBoost for Enhanced Performance	Mohammed Basim Mohammed Mohammed Erkut Arıcan	550 - 555
2119	Spoken Accent Detection in English using Audio-Based Transformer Models	Oguzhan Ozturk Hasan Kilimci H.Hakan Kilinc Zeynep Hilal Kilimci	556 - 561
2136	Analysis of Adversarial Training for Resilient Image Recognition Against Different Attacks	Emir Can Karakuş Fırat Kutluhan İslim Ece Gelal Soyak	562 - 567

	2150	Business Process Management Anomaly Detection through Semantic Embedding-Integrated Graph Neural Networks	Teoman Berkay Ayaz Ege Gülce Stanley Hsu Alper Özcan Akhan Akbulut	568 - 573
	2155	Multi-Aspect Anomaly Detection with Graph Neural Networks and Kolmogorov-Arnold Networks in Business Process Management	Teoman Berkay Ayaz Ege Gülce Stanley Hsu Alper Özcan Akhan Akbulut	574 - 579
	2156	Comparison of a Deep Learning and a Hybrid Model for Classification of an Unbalanced Urgent Cases Dataset for Human Faces	Faruk Özgür Neslihan Arıkan Özge Öztimur Karadağ	580 - 585
	2171	Estimating the Manufacturing Cost of a Metal Part from Textual and Geometric Features	Talha Rehman Abid Mert Daloğlu Cem Yıldız Ali Erman Erten Kamer Kaya	586 - 591
	2176	Object Detection in Hyperspectral Images with Unsupervised Domain Adaptation	Sinem Aybüke Şakacı Alp Ertürk Erchan Aptoula	592 - 596
	2182	Machine Learning Approaches to Predict Thyroid Cancer Recurrence: A Comparative Study	Candide Ozturk Ozgur Sagir Ulas Vural	597 - 602
	2197	Yangın Söndürme Süreçlerinde Su Tüketiminin Makine Öğrenmesi ile Tahmini Prediction of Water Consumption in Fire Extinguishing Processes Using Machine Learning Approaches	Emin Ölmez Ahsen Usta Orhan Akbulut	603 - 606
	2202	Synthetic Vibration Data Generation and Fault Classification in CNC Machines Using Transformer GANs and ConvLSTM Networks	Özlem Erbay Batıray Erbay	607 - 612
IoT	**1862**	Design and Implementation of a Management System for Wireless Electronic Combination Locks	Batuhan Kol Metin Bilgin	613 - 618
	1887	Solar IoT: Monitoring the Orientation and Electrical Parameters of the Solar Panel	Hakan Dalkılıç Oğuz Gora	619 - 624
	1973	Multifunctional Smart Spoon for Parkinson's: Stability Enhancement and Diagnostic Tools	Divyansh Singhal Sasi Snigdha Yadavalli Nupur Patil Siddharth Chauhan Madhav Rao	625 - 629
	1988	An AI-Assisted Autonomous IoRT Agent for Smart Spaces	Yakup Kayataş Sanem Kabadayı	630 - 635
	1991	Advanced Detection and Prevention of Sinkhole Attacks in 6TiSCH Networks	Burak Aydın Hakan Aydın Sedat Görmüş	636 - 641
	2000	Comparison of Different Weather Data Acquisition Methods	Emre Evcin Yusuf Murat Erten	642 - 647
	2057	Employing Digital Twin to Forest Fire Management Systems	Bugra Aydin Sema Fatma Oktug	648 - 653
	2085	Distributed Key Value Store for IoT Edge Devices	Burak Aslantaş Elif Nurdan Pektaş Şebnem Baydere	654 - 659
EMB	**2096**	Soft Error Reliability Assessment of TinyML Algorithms on STM32 Microcontroller	Ahmet Selim Karakuş Osman Buğra Göktaş Sadık Akgedik Sanem Arslan	660 - 664
HUM	**1999**	Development of An Algorithm for Converting Json Formats to Xml by Forming its File Data Structure	Aigul Mukhitova Aigerim Yerimbetova Vladimir Barakhnin	665 - 670

			Elmira Daiyrbayeva	
	2130	Facial Emotion Recognition for Imitation in Human-Robot Interaction	İbrahim Yanç Aykan İpek Selma Yılmazyıldız Kayaarma	671 - 676
HPER	**2030**	A Glimpse to Scalable LLM Architectures: Building Real-Time Sentiment Analyser with Kafka and FastAPI	Uzay Çetin Yunus Emre Gündoğmuş	677 - 681
	2165	Parallelization of BitColor Algorithm via Multithreading and GPU for Graph Coloring	Burak Kocausta Gizem Sungu Terci Alp Arslan Bayrakci	682 - 687
BCHN	**1922**	Security Dynamics of Blockchain-Enabled SDN Systems: A Taxonomic Approach	Deniz Dudukcu Murat Karakus	688 - 693
	2056	Comparative Performance Analysis of Ethereum and Optimism Smart Contracts in Health Insurance	Beyhan Adanur Dedetürk Bilge Kagan Dedetürk	694 - 699
BIO	**1907**	Disease Prediction From Human Microbiome by Utilizing Machine Learning	Bahadir Emin Temel Bora Kocapinar Zerrin Isik	700 - 705
	1923	Medikal Görüntülerin Etkili Bir Şekilde Sınıflandırılması için Dikkat Tabanlı Temsil Öğrenimi Attention Based Representation Learning for Effective Classification of Medical Images	Selim Arslan Kemal Polat	706 - 711
	1933	Utilizing Tree-Based Algorithms for Genetic Variant Interpretation	Rumeysa Aslıhan Ertürk Mehmet Baysan	712 - 717
	1978	A Test Bench for Replicating Human Breathing: Evaluating Thermal Effects of N95 Filtering Facepiece Respirator Leaks ' Preliminary Findings	Geoffrey Marchais Barthelemy Topilko Mohamed Arbane Jean Brousseau Clothilde Brochot Yacine Yaddaden Ali Bahloul Xavier Maldague	718 - 723
	1979	Development of Multiphysics Models for the Study of Airflow and Thermal Effects During the Use of Filtering Facepiece Respirators	Barthelemy Topilko Geoffrey Marchais Mohamed Arbane Jean Brousseau Ali Bahloul Xavier Maldague Clothilde Brochot Yacine Yaddaden	724 - 730
	1994	Awareness of Gamification as a Non-Pharmacological Intervention in Sleep Health Research: A Systematic Literature Review	Arife Gülah Erol Murat Y ılmaz Paul M. Clarke	731 - 736
	2003	Optik Çukur Bölütlemesi Segmentation of the Optic Cup	Saadet Aytaç Arpacı Songül Varlı	737 - 741
	2090	Clinically Significant Prostate Cancer Detection and Diagnosis in Bi-parametric MRI with Deep Learning Models	Clinton Binda Asoh-Itambi Mecit Yüzkat Songül Varlı	742 - 744
	1993	Malaria Incidence Prediction Using Climate Factors with Machine Learning Models	Ayoade Adeyemi Cyrille Mesue Njume Goodnews Akindele Nana Aisha Umar Ayodele James Oyejide	748 - 753
CDSec	**1857**	Dual-layered Approach for Malicious Domain Detection	Nadide Bilge Doğan Alp Barış¸ Beydemir Şerif Bahtiyar Umutcan Doğan	754 - 759
	1953	Analysis of the Zero-Day Detection of Metamorphic Malware	Sibel Gulmez Arzu Gorgulu Kakisim Ibrahim Sogukpinar	760 - 765
	1960	White-Box Style Intrusion Detection System Integration into Industrial Metaverses	Yasir Kılıç	766 - 771

			Vahide Nida Kılıç Ali İnan	
	2037	A Practical Investigation of Spear Phishing Spam Emails: Comparative Analysis and Evaluation	Kendrick Kurt Günter Bollens	772 - 777
	2091	Multipurpose Malware Detection System	Mert Gursimsir Cem Ayar Ibrahim Sogukpinar	778 - 782
CRYP	1867	A New Method to Detect Malicious DNS over HTTPS via Feature Reduction	Ali K. Bozkurt Halil E. Aköz Ataberk Taşpınar Şerif Bahtiyar	783 - 788
	1939	Generative Adversarial Networks for Synthetic Jamming Attacks on UAVs	Burcu Sönmez Sarıkaya Şerif Bahtiyar	789 - 794
	1958	Detecting Corruptive Noise Rounds for Statistical Disclosure Attacks	Alperen Aksoy Doğan Kesdoğan	795 - 800
	1996	Future Directions of Cybersecurity in Industrial Internet of Things Through Edge Computing	Tamara Zhukabayeva Lazzat Zholshiyeva Nurdaulet Karabayev	801 - 806
	2061	Resource-Efficient Ensemble Learning for Edge IIoT Network Security against OSINT-based Attacks	Mert İlhan Ecevit Zakire Çukur Muhammed Ali İzgün Noor Ul Ain Hasan Dağ	807 - 812
	2074	Transfer Learning for Phishing Detection: Screenshot-Based Website Classification	Furkan Çolhak Mert İlhan Ecevit Hasan Dağ	813 - 818
	2145	Blok Şifrelerin Karıştırma ve Yayılım Tabakaları için Yeni Bir Analiz Aracı A New Analysis Tool for Confusion and Diffusion Layers of Block Ciphers	Mehmet Ali Demir Meltem Kurt Pehlivanoğlu Pınar Savaştürk Emir Öztürk Muharrem Tolga Sakallı Sedat Akleylek	819 - 824
CVIS	1851	Impact of Image Augmentation on Deep Learning-Based Classification of Granite Tiles	Gaye Ediboglu Bartos Sibel Ünaldı Nesibe Yalçın	825 - 828
	1877	Advanced Facial Expression Classification with CNN-Transformer Integration for Human-Computer Interaction	Ali Azmoudeh Cigdem Altin Gumussoy Hazım Kemal Ekenel	829 - 834
	1936	Word Image Representation at Local and Global Levels Based on Vision Transformers	Baha Edine Harrath Mohamed Mhiri Mohamed Cheriet	835 - 840
	1949	Detecting Duplicate Products in E-Commerce Images Using Siamese Networks	Enis Teper Furkan Eseoğlu Mustafa Keskin	841- 846
	1963	Comparative Analysis of Visual Attribute Tagging Models for Upper-Body Clothing Products	Engin Kaya Mert Yanık	846 - 850
	1986	Development of A Model of Kazakh Sign Language Recognition Based on Deep Learning Method	Aigerim Yerimbetova Bakzhan Sakenov Ulmeken Berzhanova Nurzhan Mukazhanov Elmira Daiyrbayeva Mohamed Othman	851 - 856
	1989	Recognising Kazakh Sign Language with Mediapipe	Aigerim Yerimbetova Diana Kaidina Bakzhan Sakenov Elmira Daiyrbayeva Mussa Turdalyuly Ulmeken Berzhanova	857 - 862
	1990	Ultrason Görüntülerinden Meme Kanseri Teşhisi için Lezyon Tespitli Hibrit Derin Öğrenme Modelleri	Osman Doğuş Gülgün	863 - 868

		Hybrid Deep Learning Models with Lesion Detection for Breast Cancer Diagnosis from Ultrasound Images	Hamza Erol	
	1992	Olumsuz Hava Koşullarında Gemi Tespiti ve Sınıflandırılması Ship Detection and Classification in Adverse Weather Conditions	Yahya İzala Yaşar Becerikli	869 - 874
	2010	Removing Background from Noisy Handwritten Signatures on Banking Documents using GANs	Ege Dinçer Sacide Kalaycı Emre Yurdakul Bilge Köroğlu	875 - 879
	2013	A Faster R-CNN Model for Multi-class Classification and Detection of Land, Air, and Sea Vehicles	Enes Güvelioğlu Çiğdem İnan Acı	880 - 885
	2098	Çelik Hurdasının Sınıflandırılmasında ResNet ve Görüntü Dönüştürücü Tabanlı Modellerin Başarımı	Sefa Temur Levent Karacan	886 - 891
	2117	Advanced Computer Vision Techniques for Reliable Gender Determination in Budgerigars (Melopsittacus undulatus)	Atalay Denknalbant Efe Ilhan Cemalcılar Majid Ahangari Abdussamat Saidburkhan Alireza Zirak Ghazani Erkut Arıcan	892 - 897
	2151	Teslimat Süresi Tahminlerinde Makine Öğrenmesi Modellerinin Yorumlanabilirliği Interpretability of Machine Learning Models in Delivery Time Predictions	Serhat Agit Satıcı Habil Kalkan	898 - 903
	2169	Deep Learning based Order Form Recognition	Enes Alperen Buğaz Orhan Akbulut Aysun Taşyapı Çelebi Uğur Yıldız	904 - 908
	2170	Learning Based Photo Management on Smartphones	Beyza Nur Şenay Orhan Akbulut Aysun Taşyapı Çelebi Uğur Yıldız	909 - 912
	2177	Retinal Disease Classification Using Optical Coherence Tomography Angiography Images	Omer Faruk Aydın Muhammet Serdar Nazlı F. Boray Tek Yasemin Turkan	913 - 918
	2178	Segmentation Based Classification of Retinal Diseases in OCT Images	Öykü Eren F. Boray Tek Yasemin Turkan	919 - 924
	2183	Unsupervised Translation from Shortwave Infrared Images to RGB Images: A Comparative Evaluation	Duygu Tasbas Hacer Yalim Keles	925 - 930
DSCI	**1882**	Automatic Segmentation of Time Series Data with PELT Algorithm for Predictive Maintenance in the Flat Steel Industry	Saygın Kaçar Tuğçe Ballı E. Fatih Yetkin	931 - 936
	1921	Otomobil Kredilerinde Temerrüt Tahmini ve Araç Geri Kazanım Olasılığı Analizi - Bir Segmentasyon Çalışması Default Prediction and Vehicle Recovery Probability Analysis in Auto Loans - A Segmentation Study	Sahin Nicat Anıl Ferdi Kaya	937 - 942
	1955	E-Ticaret Sadakat Programı Müşteri Eğilim Tahmini Customer Propensity Prediction in E-Commerce Loyalty Program	Yunus Emre Gündoğmuş Sinan Keçeci Ege Erdem Emre Rençberoğlu	943 - 946
	1967	Yapay Zeka ve Makroekonomik Göstergeler ile Tüzel Kredilerin Değerlendirilmesi Evaluation of Corporate Loans with Artificial Intelligence and Macroeconomic Indicators	Burak Yüksel Hakkı Berkay Çiçek	947 - 951
	1974	Enhanced Bot Detection on TwiBot-20 Dataset	Mehmet Ali Osman Atik Şevket Umut Çakır Alper Özcan	952 - 956
	1975	Drug-Drug and Drug-Protein Link Prediction on DTINet dataset	Mehmet Ali Osman Atik Yusuf Çelik Alper Özcan	957 - 960
	2029	Perakende Verilerinde Anomali Tespiti ve Döviz Kuru Älilişkisi Üzerine ChatGPT Destekli Yorumlama	Şadi Evren Şeker	961 - 966

		ChatGPT Supported Interpretation on Anomaly Detection in Retail Data and Exchange Rate Relationship	Hatice Nizam-Özoğur	
	2033	Text to SQL Transformation Using LLM: a Comparative Research of T5, Seq2Seq, and SQLNet Models	Zhazira Shaikhiyeva Madina Mansurova Gulshat Amirkhanova	967 - 972
	2076	Sağlık Sigortası Sahiplerinin Davranışsal Analizi ve Kümelenmesi Clustering and Behavioral Analysis of Health Insurance Owners	Omer Sezer Koyuncu Seçil Arslan	973 - 978
	2087	On symbolic Prediction of Time Series for Predictive Maintenance Based on SAX-LSTM	Aykut Güler Tuğçe Ballı E. Fatih Yetkin	979 - 983
	2135	Profiling Driver Behaviors Using AI-Based Methods and Deep Learning Techniques for Improving Road Safety: A Comparative Study of Algorithms	Volkan Oban Mustafa Kaya Güzide Safi İrem Nur Çimen Tubanur Çatak Bulut Karadağ Gökhan Gümüş Aslıhan Çandır Fatih Alagöz	984 - 989
IR	1896	ReRag: A New Architecture for Reducing the Hallucination by Retrieval-Augmented Generation	Robin Koç Mustafa Kağan Gürkan Fatoş T. Yarman Vural	990 -994
	1941	Enhancing Object Detection in Aerial Images Using Transformer-Based Super-Resolution	Aslan Ahmet Haykır İlkay Öksüz	995 - 1000
NET	1985	Proof of Concept Implementation for RSVP TSN Control Plane	Necip Gozuacik	1001 - 1004
	2100	Integrating Blockchain and SDN for Centrality-Aware Virtual Multicast Tree Embedding	Furkan Ayaz Evrim Guler Murat Karakus Davut Hanbay	1005 - 1010
	1969	QoS Aware Routing Approaches in Software Defined Smart Grids	Sedef Demirci	1011 - 1016
	2008	Deep Reinforcement Learning Routing in Mobile Networks	Arif Burak Dikmen Hasari Çelebi	1017 - 1022
RBOT	1942	Endüstriyel Robotik Sistemlerin Güvenlik Doğrulaması Safety Verification of Industrial Robotic Systems	Fatih Furkan Arslan Metin Özkan	1023 - 1028
	2077	EKF Based Localization: Integrating IMU and LiDAR Data in the Hilti SLAM Challenge	Behice Bakır Havvanur Bozömeroğlu Ebu Yusuf Güven	1029 - 1034
SING	1965	Communication (Educational) Kit (HaKi)	Murat Sever Utku Bilgin	1035 - 1038
	2089	Manyetik Parçacık Görüntülemede Sistem Matrisi için Farklı Dalgacık Dönüşümlerinin Seyreklik Seviyesi Karşılaştırması Sparsity Level Comparison of Different Wavelet Transforms for the System Matrix in Magnetic Particle Imaging	Vildan Atalay Aydın	1039 - 1043
	2097	Sparse Channel Estimation For M-QAM-Based Underwater Acoustic Communication Systems	Mhd Tahssin Altabbaa Berkay Tekat Emin Tarik Iseri	1044 - 1048
OTH	1858	The 80/20 Principle in Morphemics-Morphology in the Educational Corpus of the Uzbek Language	Shahlo Khamroeva Bakhtiyor Mengliyev Muyassar Kholova	1049 - 1052
	1904	Gamification as a Tool for Personalized Learning in Inclusive Education	Dilaram Baumuratova Tamara Zhukabayeva Mira Rakhimzhanova	1053 - 1058
	1918	A Metaheuristic Algorithm for the Fixed Charge Transportation Problem	Nermin Kartli	1059 - 1062
	2027	Eğitimde Sürükleyici Teknolojilerin Kullanılması Fırsatlar ve Beklentiler	Atamuratov Rasuljon Kadırjanovich Majıdova Gulhayo Abdırazzoq qızı Bayjonov Furqat Baxramovıch Ongarov Mansurbek Bayrambekovıch	1063 - 1068

			Saydullayev Zafar Erkınovıch	
	2103	Bilgisayar Mühendisliği Öğrencilerinin Perspektifinden Bilişim Hukukunun Güncel Sorunları ve Çözüm Önerileri Current Challenges and Solution Proposals in IT Law from the Perspective of Computer Engineering Students	Sevda Bora Çınar	1069 - 1075
	2200	A Comparison of shcU-Net Based GAN and U-net Based GAN in Adult Dental Segmentation	Gürdal Altundağ Hakan Öcal	1075 - 1080
	1932	Leveraging Quantum Computing and Optimization to Estimate Financial Crashes in Small and Medium-Sized Enterprises	Ege Dincer Berkay Coskuner Ege Bilaloglu Bilge Koroglu	1081 - 1086
SW	**1859**	Investigating The Adoption of International Software Quality Standards in Turkey: A Comprehensive Analysis	Sevgi Koyuncu Tunç	1087 - 1093
	1886	Development of the Functional Structure of the Science and Education Information System	Dauletov Adilbek Yusupbayevich Matyakubova Noila Shakirjanovna	1094 - 1098
	1892	React ve Preact Javascript Çerçevelerinde Karşılaştırmalı Analiz Comparative Analysis on React and Preact Javascript Frameworks	Muhammed Furkan Uygur Nesibe Yalçın	1099 - 1104
	1917	CAGE: A Tool for Code Assessment and Grading	Ümit Kanoğlu Oğuz Kerem Yıldız Hasan Sözer Olcay Taner Yıldız	1115 - 1110
	1957	Extracting Driving Styles from Automotive Sensor Data to Develop Personas	M. Cagri Kaya Tayssir Bouraffa Krzysztof Wnuk	1111 - 1114
	1962	Lojistik Sipariş Dağıtım Entegrasyonu Sürecinde Sipariş Geri Çağırma Süreci Tasarımı ve Yazılım Geliştirmesi Design and Software Development of The Order Recall Progress in The Logistics Order Distribution Integration Process	İklim Barman Ersin Şengül	1115 - 1120
	2009	The Dimension of Green Coding in Software Quality Control Processes	Volkan Abur	1121 - 1126
	2055	Are We Asking the Right Questions to ChatGPT for Learning Software Design Patterns?	Çağdaş Evren Gerede	1127 - 1132
	2060	Optimizing LLVM IR: Transforming Multiplication to Addition for Enhanced Execution Efficiency	Huseyin Karacalı Efecan Cebel Nevzat Donum	1133 - 1138
	2080	Estimation of Software Integration Test Duration via UML Statecharts	Fehim Göler Tolga Ovatman	1139 - 1144
	2093	DIA4M: A Tool to Streamline DevOps Processes of Distributed Cloud-Native Systems	Eren Tarak H. Hakan Kilinc	1145 - 1150
	2111	Software Industry Perception of Academic Collaboration	Deniz Akdur	1151 - 1156
	2139	Görüntü İşlemeyle Doğrulamalı Robotik Test Otomasyon Kullanımı: POS Cihazları Üzerine Uygulama	Miraç Emektar Harun Kadıoğlu Ahmet Efendioğlu Fatih Mehmet Harmancı	1157 - 1161
	2141	VoIP Sistemlerinde Zihin Haritası Tabanlı Test Stratejiler : SIP Pbx Ürünü Üzerine Bir İnceleme Mind Map-Based Testing Strategies in VoIP Systems: A Case Study on SIP Pbx Products	Miraç Emektar Furkan Günaydın Fatih Mehmet Harmancı	1162 - 1167
	2173	A Robust Microservices Framework for Indoor Tracking System Development	Gafur Hayytbayev Kerem Küçük Mahmut Çavur	1168 - 1172
DM	**1927**	Unsupervised Pattern Extraction of Time Series Data for Energy Disaggregation	Şirin Azazi Deveci Melih Günay	1173 - 1178
	1944	Topic Modeling Enhanced Tripartite Graph for Recommendation using Metapaths	Yaren Yılmaz Irem İşlek Şule Gündüz Öğüdücü	1179 - 1184
	1948	Community Detection on Software Library Dependency Graphs using Graph Neural Networks	Şevket Umut Çakır	1185 - 1190

			Mehmet Ali Osman Atik Ümit Deniz Uluşar	
	2190	Enhancing Mesh and Point Cloud Similarity Detection through Geometric Features and ICP	Talha Rehman Abid Mehtap Öklü Cem Yıldız Ali Erman Erten Kamer Kaya	1191 - 1196
	2214	Comparative Analysis and Practical Implementation of Machine Learning Algorithms for Phishing Website Detection	Samad Najjar-Ghabel Shamim Yousefi Payam Habibi	1197 - 1202
	2215	A Technical Analysis and Practical Implementation of Machine Learning Algorithms for Predicting Survival in Breast Cancer Patients	Shamim Yousefi Samad Najjar-Ghabel Hamidreza Shafaei	1203 - 1208
BIG	**1881**	Comparison Between Time Series and Relational Databases	Alpar Türkoğlu Onurcan Ersen Ibrahim Onuralp Yiğit Dincer Unal Hatice Golcuk	1209 - 1212
	1930	A Performance Evaluation Study on a Data Analytics Platform for Emergency Calls	Engin Yakar H. Hakan Kilinc	1213 - 1218
	2079	Adaptive Composite Market Volatility Index (CMVI) for Enhanced Stock Price Forecasting	Rabia Çevik Uğur Barış Özyürek Ali Kanal Vael Kokach Büşra Kocaçınar Oznur Şengel Fatma Patlar Akbulut	1219 - 1223
	2142	Hybrid Deep Learning Framework for Stock Price Prediction Incorporating Technical and Macroeconomic Indicators	Ali Can Turan Vael Kokach Büşra Kocaçınar Oznur Şengel Fatma Patlar Akbulut	1224 - 1228
	2125	Emotion-Aware Multimodal Biometric Identification by using Biosignals	Yekta Said Can Beyzanur Bektan Fatih Alagöz	1229 - 1235
	1854	Özbekçe-Türkçe Otomatik Çeviri Yazılımı için Deyimlerin Veritabanını Teşkil Etmede Karşılaşılan Güçlükler Automatic Translation Software Difficulties in Organizing the Database of Idioms for Uzbek and Turkish	Manzura Abjalova Umida Raşidova Eşref Adalı	1236 - 1240
	2028	Reversible Steganographic System for the Transmission of Personal Medical Data	Elmira Daiyrbayeva Ekaterina Merzlyakova Aigerim Yerimbetova Aigul Mukhitova	1241 - 1246

Araneum Uzbecicum: A Gigaword Web-Crawled Uzbek Corpus

Vladimír Benko
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Batislava, Slovkia
vladimir.benko@juls.savba.sk

Radovan Garabík
Ľ. Štúr Institute of Linguistics Slovak Academy of Sciences Batislava, Slovkia
radovan.garabik@kassiopeia.juls.savba.sk

Shahlo Khamroyeva
Tashkent State University of Uzbek language and literature
Tashkent, Uzbekistan
shaxlo.xamrayeva@navoiy-uni.uz

***Abstract*— We want to introduce a Project of creation of a web-crawled Uzbek corpus. The main steps related to data capture, pre-processing, tokenization, deduplication, PoS tagging and lemmatization are discussed, and some query examples are shown. This article analyzes the problems in the process of creating and lemmatization the corpus of Araneum Uzbecicum. We also show word embeddings created from the Uzbek data. The Word embeddings function serves as a useful tool for researching the semantic relations in the Uzbek language, word context and valence. Also, the paradigm created by the Word embeddings function helps to perfect dictionaries.**

Keywords—Uzbek language, web-crawled corpus, PoS tagging and lemmatization, word embeddings

I. Introduction

After Tatar and Kazakh, Uzbek is the third Turkic language in our corpus. Spoken by approx. 33 million of L1 or L2 speakers in Uzbekistan and several of its neighboring countries, and also being “sufficiently represented” in the Internet, the language is a suitable target for creation of a web-crawled corpus.

In our paper, we would like to introduce a Project aimed at creation of a Gigaword Uzbek corpus that would contain texts in both scripts currently used in Uzbekistan, i.e., modified Latin and modified Cyrillic, respectively. The corpus data is PoS tagged and lemmatized by *Apertium*, processed by the *NoSketch Engine* corpus manager and made available at our web-corpora portal[1].

II. Related Work

In recent years, building web corpora has become a well-established part of linguistics. The Web, teeming as it is with language data, of all manner of varieties and languages, in vast quantities and freely available, is a fabulous linguists' playground. This special issue of Computational Linguistics explores ways in which this dream is being explored.

The use of web corpora in linguistic research has become widespread, so the problem of using web corpora in linguistic research has been studied in depth.

The problems of creating a web corpus using the BootCaT tool for low-resource languages were studied and clear results were obtained.

Several corpora of Turkish languages were developed on the Sketch Engine platform.

“uzb_community_2017” is an Uzbek community corpus based on material from 2017. It contains 663,119 sentences and 9,256,001 tokens [2].

The initial efforts to develop the Uzbek language corpus commenced in Uzbekistan in 2018. While there have been some theoretical articles on the construction of the Uzbek language corpus, the practical developments on the corpus have become more prevalent by 2018. A bibliometric analysis based on Scopus was conducted on corpus linguistics for the period spanning 2017 to 2021 [41]. There is some work about creating a morphological and syntactic tagged corpus for the Uzbek language. Several studies have been conducted on the processing of corpus texts in the Uzbek language [44], [45].

Additionally, several other corpora have been created for the Uzbek language. In particular, the “Educational corpus of the Uzbek language”[3] within the project of the Tashkent State University of Uzbek Language and Literature, the “Uzbek language corpus”[4]; based on the project of the National University of Uzbekistan, the “National corpus of the Uzbek language”. “Alisher Navoi authorship corpus”[5] and the “Diachronic corpus of Uzbek-English newspapers”[6] were developed by independent researchers, are available for free use. Software for the national corpus of the Uzbek language was developed.

Furthermore, it is worth noting the contributions of researchers who have undertaken studies on specific types of corpora. These include articles on the development of the Uzbek author's corpus, the Uzbek-English parallel corpus, Uzbek-Russian parallel corpus and the diachronic corpus of the Uzbek language. There is research in the field of author lexicography and the relationship between author corpora (Anonymous). An algorithm for creating a parallel corpus of the Uzbek and Russian languages was. The theoretical foundations of the Uzbek-English parallel corpus have been developed [42], [43]. A scientific study of the open Uzbek speech corpus and preliminary experiments on speech recognition were conducted.

III. Data for the New Corpus

During the last decade, the methodology and tools for compilation web-crawled corpora has been effectively standardized. In our Project, we were using a collection of tools usually referred to as “Brno Pipeline” available under a

[1] http://aranea.juls.savba.sk/guest/index.html
[2] https://www.sketchengine.eu/uzwac-uzbek-corpus/
[3] https://uzschoolcorpara.uz/
[4] https://uzbekcorpus.uz/
[5] https://alishernavoiykorpusi.uz/
[6] https://mediatextcorpus.com/

FLOSS license[7] . Its main component is SpiderLing[8], a web crawler optimized for downloading textual data. In its latest Version 2.2, it represents a mature, effective and stable tool that is able to get (within our technical infrastructure) as much as two Gigaword of plain text within 24 hours of crawling. The tool also includes a character encoding identification utility; a language detection module that is using samples of texts for the languages to be recognized provided by the user to build the respective language models using character trigram frequencies; and a boilerplate identification and removal tool **..** All the processes work smoothly on the fly, outputting a reasonably clean plain text with only light XML markup encoding some metadata at the document and paragraph level.

To initialize the process, SpiderLing needs a list of seed URLs that are used in the first round of crawling. In our case, the seed list of approx. 2,000 URLs has been obtained by means of the WebBootCat functionality accessible at the Sketch Engine Portal.

The actual crawling was performed during two separate sessions, in December 2020, and March 2024, respectively. The seed URLs for the second one was obtained from the data already downloaded during the first session – the last 100,000 URLs were used, expecting that some of the addresses might meanwhile have changed the contents of the respective web pages, or could be removed during deduplication.

IV. Pre-Tokenization Processing

Though the language identification procedure of SpiderLing works reliably for most languages, it is often not able to discriminate between very close languages using the same script. We therefore perform a secondary language filtering based on character frequencies, making use of the fact that some characters are unique for a particular language. This was also the case of Uzbek that is using the ҳ, қ, ў, and ғ in the Cyrillic script, and o‘, g‘ and ‘ characters or character combinations in the Latin script. The respective thresholds for filtration were established experimentally.

Other filters detect potential mojibake encoding problems appearing as “artifacts” such as *subâ€TMektlarini* or *oÃƒ¢€TMtgan* in the texts – documents with too many suspicious character sequences are identified by means of a series of regular expressions and removed.

V. Tokenization

For various reasons, we perform tokenization as a separate processing step, even in situations when the tagger would prefer to do it as a part of the tagging process. Here again, one of the “Brno” tools is used – Unitok **.**, the universal tokenizer is a Python program using a custom-modified parameter file containing regular expressions for language-specific period-final abbreviations, ordinal numbers, etc. The initial list of Uzbek abbreviations (both in Latin and Cyrillic) presently contains just approx. two dozen items – we hope, however, to be able to expand it after analyzing the processed corpus.

VI. Data Normalization

Now it is necessary to mention one of the peculiarities of the Uzbek language – the simultaneous use of two different scripts, with the speakers being proficient in both of them. This situation is similar to that of the Serbian language, but unlike in Serbia (and Bosnia and Herzegovina), where the digraphia is a result of centuries long development and any of the scripts are widely accepted in almost all situations, official and unofficial, , the de facto digraphia in Uzbekistan is a result of partially unsuccessful transition from Cyrillic to Latin script in early 1990s. According to our observation, the amount of texts in Internet in either of the scripts is about equal, with many websites, including those belonging to government organizations, private companies, or even the Uzbek Wikipedia, provide two versions of their web pages.

We have decided, on one hand, to retain the information on the respective script in our corpus, as we consider it to be an interesting sociolinguistic parameter, and on the other hand to provide transliterated version of all texts, so that users could easily query the corpus. Moreover, this is also necessary as the PoS tagger used expects its input in Latin script.

The other issue is the presence of the “modified letter turned comma” (U+02BB) and “modified letter apostrophe” (U+02BC) characters in the Latin script. Since these characters are absent from most keyboard layouts, they are often replaced by similar ones, such as “left single quotation mark” (U+0218), “apostrophe” (U+0027), or “right single quotation mark” (U+0219), or oven other similarly looking characters, such as an acute or grave accent. As the PoS tagger expects the text to contain canonical forms of these characters only, all these irregularities have to be normalized before further processing.

VII. Deduplication

Web-crawled texts are known to contain a large proportion of duplicate contents at different levels. We have adopted a two-step approach: (partial) duplicates at the document level are removed and not passed to further processing, while paragraph and sentence level deduplication is provided, if required, at the very end of the processing only.

For the document-level deduplication we use a tool called Onion working on the principle of comparing word n-grams from the document processed with those encountered in all previous documents. A document is considered (partially) duplicate if it contains certain proportion of n-grams already seen. Both values, i.e., the length of n-grams to compare and the similarity level is set by the user – we work with 5-grams and 90% threshold as discussed in [45].

VIII. PoS Tagging and Lemmatization

Uzbek unfortunately does not belong to languages with a language model available for the “traditional” multilingual taggers, such as TreeTagger], or at least a lemmatizer like CSTlemma, though we hope the situation might improve in the future.

The baseline morpho-syntactic annotation has been performed by Apertium, that is in fact machine translation tool, yet its morphological module can be used separately for morpho-syntactic annotation. Its main deficiency, for most languages where a language model is available, is a rather small size of the respective morphological lexicon and a limited disambiguation procedure. It is also not able to guess lemmas and/or PoS tags for the out-of-dictionary (OOV) lexical items.

[7] https://corpus.tools/

[8] https://corpus.tools/wiki/SpiderLing

Another problem of using Apertium as a tagger is its “attitude” to data already tokenized. Being a machine translation tool, its lexicon can also contain multi-word expressions than are to be translated as whole units. If such a multi-word unit is encountered in the text, the system decides that a new token to be added, i.e., spoiling the original tokenization that is expected by the overall processing, e.g., to be able to merge the Cyrillic and Latin data. As the Apertiun morphological lexicon is available in a readable form, we have decided to delete all multi-word expressions from it during this initial phase of our Project.

The respective processing steps of tokenization, normalization, PoS tagging and lemmatization are shown in Fig. 1 to 4.

```
<s>
Янги
стратегияни
август
ойига
қадар
Тошкент
шаҳри
миқёсида
тасдиқланиши
режалаштирилмоқда
<g/>
.
</s>
```

Fig. 1. Tokenized sentence in original Cyrillic Script

```
<>
Yangi
strategiyani
avgust
oyiga
qadar
Toshkent
shahri
miqyosida
tasdiqlanishi
rejalashtirilmoqda
<>
.
<>
```

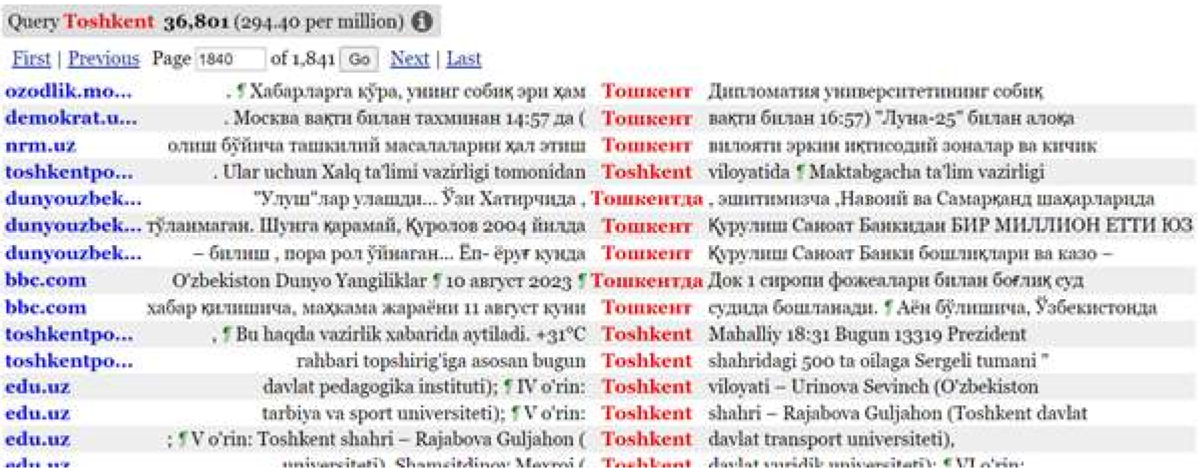
Fig. 2. Normalized (Romanized) and filtered sentence to be sent as input for *Apertium*

```
<>
^Yangi/yangi<n><nom>$
^strategiyani/strategiya<n><acc>$
^avgust/*avgust$
^oyiga/oy<n><px3sp><dat>$
^qadar/qadar<post>$
^Toshkent/Toshkent<np><top><nom>$
^shahri/shahr<n><px3sp><nom>$
^miqesida/ miqyos<n><px3sp><loc>$
^tasdiqlanishi/tasdiqlan<v><iv>
   <ger_abst><px3sp><nom>$
^rejalashtirilmoqda/rejalashtir<v><tv>
   <pass><prog_inf><p3><pl>$
<>
^./.<sent>$
<>
```

Fig. 3. Sentence tagged by *Apertium* in its internal format

TABLE I. FULLY POS TAGGED AND LEMMATIZED SENTENCE

Word	nword	lemma	atag	tag	ztag
<s>					
Янги	Yangi	yangi	Nn	n:nom	1
стратегияни	strategiyani	strategiya	Nn	n:acc	1
август	avgust	avgust	Yy	*	0
ойига	oyiga	oy	Nn	n:px3sp:dat	1
қадар	qadar	qadar	Pp	post	1
Тошкент	Toshkent	Toshkent	Nn	np:top:nom	1
шаҳри	shahri	shahr	Nn	n:px3sp:nom	1
миқёсида	miqyosida	miqyos	Nn	n:px3sp:loc	1
тасдиқланиши	tasdiqlanishi	tasdiqlan	Vb	v:iv:ger_abst:px3sp:nom	1
режалаштирилмоқда	rejalashtirilmoqda	rejalashtir	Vb	v:tv:pass:prog_inf:p3:pl	1
<g/>					
.	.	.	Zz	sent	1
</s>					



Query Toshkent 36,801 (294.40 per million)

First | Previous Page 1840 of 1,841 Go Next | Last

ozodlik.mo...	. ¶ Хабарларга кўра, унинг собиқ эри ҳам	Тошкент	Дипломатия университетининг собиқ
demokrat.u...	. Москва вақти билан тахминан 14:57 да (	Тошкент	вақти билан 16:57) "Луна-25" билан алоқа
nrm.uz	олиш бўйича ташкилий масалаларини ҳал этиш	Тошкент	вилояти эркин иқтисодий зоналар ва кичик
toshkentpo...	. Ular uchun Xalq ta'limi vazirligi tomonidan	Toshkent	viloyatida ¶ Maktabgacha ta'lim vazirligi
dunyouzbek...	"Улуш"лар улашди... Ўзи Хатирчида ,	Тошкентда	, эшитимизча ,Навоий ва Самарқанд шаҳарларида
dunyouzbek...	тўланмаган. Шунга қарамай, Қуролов 2004 йилда	Тошкент	Қурулиш Саноат Банкидан БИР МИЛЛИОН ЕТТИ ЮЗ
dunyouzbek...	– билиш , пора рол ўйнаган... Ёп- ёруғ кунда	Тошкент	Қурулиш Саноат Банки бошлиқлари ва казо –
bbc.com	O'zbekiston Dunyo Yangiliklar ¶ 10 август 2023 ¶	Тошкентда	Док 1 сиропи фожеалари билан боғлиқ суд
bbc.com	хабар қилишича, маҳкама жараёни 11 август куни	Тошкент	судида бошланади. ¶ Аён бўлишича, Ўзбекистонда
toshkentpo...	. ¶ Bu haqda vazirlik xabarida aytiladi. +31°C	Toshkent	Mahalliy 18:31 Bugun 13319 Prezident
toshkentpo...	rahbari topshirig'iga asosan bugun	Toshkent	shahridagi 500 ta oilaga Sergeli tumani "
edu.uz	davlat pedagogika instituti); ¶ IV o'rin:	Toshkent	viloyati – Urinova Sevinch (O'zbekiston
edu.uz	tarbiya va sport universiteti); ¶ V o'rin:	Toshkent	shahri – Rajabova Guljahon (Toshkent davlat
edu.uz	; ¶ V o'rin: Toshkent shahri – Rajabova Guljahon (	Toshkent	davlat transport universiteti).
edu.uz	universiteti). Shamsitdinov Mavrої (	Toshkent	davlat yuridik universiteti); ¶ VI o'rin:

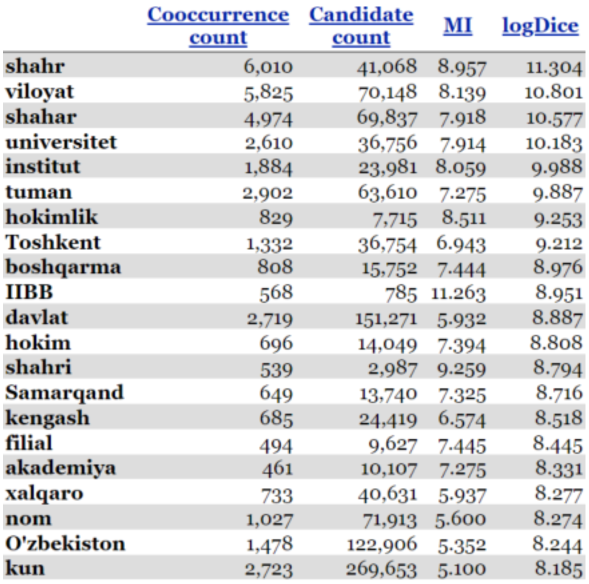
Fig. 4. Querying the Corpus in the “Simple Query” mode

The sentence chosen as an example happened to be in Cyrillic script (Fig. 1), it needed to be converted to Latin script with the XML tags “rasped” so that Apertium would to try to tag them as well (Fig. 2). The Apertium output (Fig. 3) has to be parsed, converted to column format with the original word forms, and XML tags restored to get the format expected by the NoSketch Engine corpus manager (Table I).

The word attribute contains the word form it the original script, nword is its normalized and/or Romanized value, lemma and tag are values assigned by the tagger. Two extra attributes result from the post-PoS processing: atag is the part of speech information expressed in the Araneum Universal Tagset (Anonymous), and ztag displays the result of the morphological lexicon lookup, with the “1” value meaning success, and “0” out of vocabulary (OOV) lexical item. All these corpus attributes are available for querying to corpus users. Moreover, as word, nword, and lemma are present in the “Simple query” CQL definition, user can query either in Latin or Cyrillic script, where queries in Cyrillic script will find Cyrillic occurrences only, while those in Latin script will find occurrences in both scripts, as shown in Fig. 4.

IX. Querying the Corpus

A 125 Megatoken sample of the corpus has already been published at our “Sandbox” Corpora Portal providing the users with full NoSketch Engine functionality. The Fig. 5, for example, shows the most salient collocates of the word “Toshkent” within the “–4 to +5” window sorted by the logDice value.



	Cooccurrence count	Candidate count	MI	logDice
shahr	6,010	41,068	8.957	11.304
viloyat	5,825	70,148	8.139	10.801
shahar	4,974	69,837	7.918	10.577
universitet	2,610	36,756	7.914	10.183
institut	1,884	23,981	8.059	9.988
tuman	2,902	63,610	7.275	9.887
hokimlik	829	7,715	8.511	9.253
Toshkent	1,332	36,754	6.943	9.212
boshqarma	808	15,752	7.444	8.976
IIBB	568	785	11.263	8.951
davlat	2,719	151,271	5.932	8.887
hokim	696	14,049	7.394	8.808
shahri	539	2,987	9.259	8.794
Samarqand	649	13,740	7.325	8.716
kengash	685	24,419	6.574	8.518
filial	494	9,627	7.445	8.445
akademiya	461	10,107	7.275	8.331
xalqaro	733	40,631	5.937	8.277
nom	1,027	71,913	5.600	8.274
O'zbekiston	1,478	122,906	5.352	8.244
kun	2,723	269,653	5.100	8.185

Fig. 5. Most salient collocates of “Toshkent”.

X. Word Embeddings

Word embeddings is a method that assigns multidimensional vectors to individual words. Pioneered by the word2vec, it soon became an indispensable part of many NLP applications. Although nowadays mostly replaced by contextual embeddings, word embeddings are still used for many purposes (e.g in lexicography), because they capture semantic and grammatical properties of individual words.

We provide vector embeddings models for most of the languages of the Aranea family of corpora, the newest addition is the set of Uzbek embeddings trained on the Araneum Uzbecicum corpus with the GenSIM framework. The models use a context window 7 words wide and the skip-gram training algorithm, with dimensionality 200.

The model is accessible through a user-friendly web interface[9] that can be used to access and query the models and demonstrate several common ways of using the models. The interface is built upon an assumption that the vectors reflect semantic properties of the words and the position (and distance) of the vectors in the embedding space corresponds to the position in an abstract “semantic space” of the language. The interface is geared towards lexicographers and linguists without experience in NLP. The models use automatic detection of bigrams, these are treated as regular words (i.e. a vector is assigned to them) and can be used in queries, with the underscore character _ (U+005F LOW LINE) joining the components of the bigram.

The basic information provided is the list of words corresponding to vectors with small angular distance from the given word. Instead of cosine similarity, we define “semantic difference” as √(1–cos2 θ) = sin θ, being close to zero for near-synonyms and close to one for unrelated words. We found out this better aligns with the expectation of our users. An example of the list of near synonyms is at Fig. X2. Then there is a visualization of the embedding vectors, we are using ISOMAP dimensionality reduction to display the results as a 2-dimensional, 3-dimensional or (for those users able to mentally imagine four-dimensional Euclidean space) 4-dimensional map, presented as a perspective projection into 2-dimensional picture, as well as two different word cloud representations. There is also a link to a raw dimensionality reduced data in Gnuplot forma, giving the users ability to pan, zoom and rotate the graphs, or use the data in other statistical software.

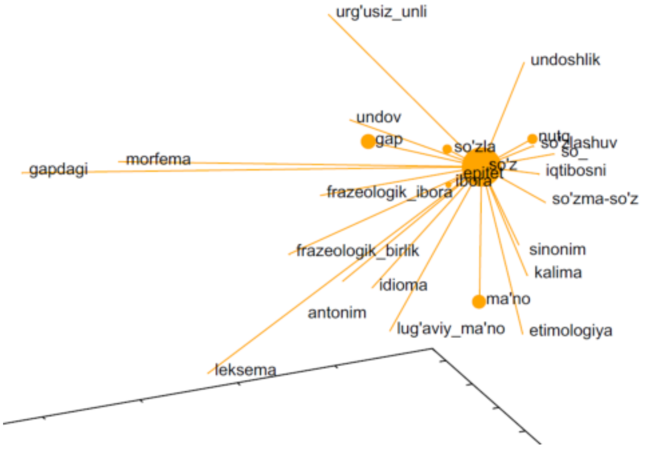
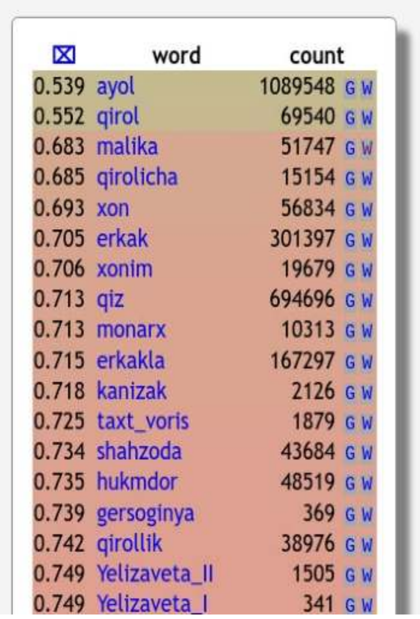


Fig. 6. Three dimensional visualization of word embedding vectors, in the vicinity of the lemma *so'z*.

[9] https://www.juls.savba.sk/semä/?lang=uz

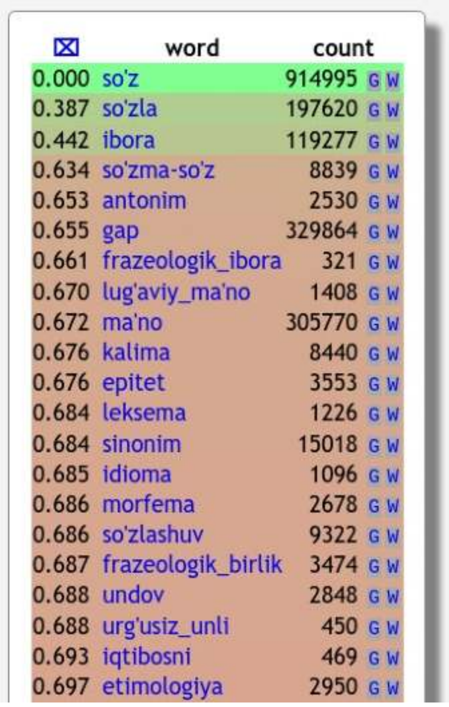
There are three different models available: the model trained on the combination of part of speech and lemmas, the model trained on word forms, and a FastText model The models are also provided as downloadable datasets in text GenSim format.



☒	word	count	
0.539	ayol	1089548	G W
0.552	qirol	69540	G W
0.683	malika	51747	G W
0.685	qirolicha	15154	G W
0.693	xon	56834	G W
0.705	erkak	301397	G W
0.706	xonim	19679	G W
0.713	qiz	694696	G W
0.713	monarx	10313	G W
0.715	erkakla	167297	G W
0.718	kanizak	2126	G W
0.725	taxt_voris	1879	G W
0.734	shahzoda	43684	G W
0.735	hukmdor	48519	G W
0.739	gersoginya	369	G W
0.742	qirollik	38976	G W
0.749	Yelizaveta_II	1505	G W
0.749	Yelizaveta_I	341	G W

Fig. 8. List of nearest vectors to the result of an arithmetic expression qirol – er + ayol.

The links to external resources are (together with the hyperlink symbols): link to the Aranem Uzbecicum Minus corpus; link to the Google search for the word, restricted to the *.uz* top level domain; link to English language Wiktionary entry for the word. This allows quick verification of the usage of the word, mostly for lexicographic purposes.



☒	word	count	
0.000	so'z	914995	G W
0.387	so'zla	197620	G W
0.442	ibora	119277	G W
0.634	so'zma-so'z	8839	G W
0.653	antonim	2530	G W
0.655	gap	329864	G W
0.661	frazeologik_ibora	321	G W
0.670	lug'aviy_ma'no	1408	G W
0.672	ma'no	305770	G W
0.676	kalima	8440	G W
0.676	epitet	3553	G W
0.684	leksema	1226	G W
0.684	sinonim	15018	G W
0.685	idioma	1096	G W
0.686	morfema	2678	G W
0.686	so'zlashuv	9322	G W
0.687	frazeologik_birlik	3474	G W
0.688	undov	2848	G W
0.688	urg'usiz_unli	450	G W
0.693	iqtibosni	469	G W
0.697	etimologiya	2950	G W

Fig. 7. List of nearest vectors to the lemma *so'z*. The columns contain sine similarity, lemma, absolute frequency in the corpus, link to the *Araneum Uzbecicum Minus* corpus, link to Google Search, link to Wiktionary.

The interface supports two other modes of operation: if there are two words (separated by a space) in the input, it displays the sine similarity between them.

The other mode gives access to a basic vector arithmetics, supporting addition and subtraction. This allows easy evaluation of the well known vector transfer (the famous king – man + woman = queen example, see [14] with many applications in linguistic and sociolinguistic research. In Fig. 8. We provide an Uzbek equivalent, the nearest vectors to the expression qirol - er + ayol correspond to the words malika and qirolicha, meeting our expectations.

XI. Conclusion and Further Work

Based on the feedback of the early users, we would like to perform further processing in the foreseeable future as follows:

• As the corpus contains some amount of non-Uzbek (Mostly English and Russian) text fragments, a sentence-level language filtration using the methodology suggested by [28] will be applied.

• Sketch and term grammars for the Sketch Engine will be crated.

• Based on the Apertium morphological lexicon, we'll try to train rules for the CSTlemma lemmatizer, so that lemmas for the OOV lexical items could be guessed.

Besides the availability of our new corpus for on-line querying at our Corpus Portal, we can also provide the processed corpus source data (for non-commercial purposes) under an open license. We hope this this might improve the situation with availability of large-scale language resources for the Uzbek language.

The full Gigaword version of the new corpus will be published on-line in a foreseeable future.

Acknowledgment

This work was, in part, financially supported by the Slovak VEGA Grant Agency for Science, Project No. 2/0016/21.

References

[1] N. Abdurakhmonova, U.Tuliyev, A. Gatiatullin. Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz // 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, pp. 1-4, (2021).

[2] S. S. Avezov Development of a Parallel Corpus of the Uzbek and Russian Languages // Vital Annex: International Journal of Novel Research in Advanced Sciences (IJNRAS) Volume: 01, Issue: 05, (2022).

[3] V.Baisa & V.Suchomel. Turkic language support in Sketch Engine. // 3rd International Conference on Computer Processing in Turkic Languages (TURKLANG 2015).

[4] A.Barbaresi. Challenges in web corpus construction for low-resource languages in a post-BootCaT world. 6th Language & Technology Conference, Less Resourced Languages special track, Poznan, Poland. pp. 69-73 (2013).

[5] Baroni, M., & Bernardini, S. BootCaT: Bootstrapping Corpora and Terms from the Web. In Proceedings of LREC 2004, pp. 1313-1316 (2004).

[6] V. Benko Aranea: Yet Another Family of (Comparable) Web Corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th In-ternational Conference, TSD 2014,

Brno, Czech Republic, September 8-12, 2014. Pro-ceedings. LNCS 8655. Springer International Publishing Switzerland (2014)

[7] G. Radovan: Word Embedding Based on Large-Scale Web Corpora as a Powerful Lexicographic Tool. In: Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje 46, br. 2 (2020): 603-618

[8] B.Jongejan and H.Dalianis. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In Proceedings of the Joint Con-ference of the 47th Annual Meeting of the ACL and the 4th International Joint Confer-ence on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, pp. 145-153 (2009).

[9] T.Khanna, J. N. Washington, F. M.Tyers, S.Bayatlı, D. G.Swanson, T. A.Pirinen, I.Tang, and Alòs i Font, H. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. Machine Translation (2021)

[10] A.Kilgarriff, V.Baisa, J. Bušta, M.Jakubíček, V.Kovář, J.Michelfeit, P.Rychlý, and V.Suchomel. The Sketch Engine: Ten Years On. Lexicography 1(1):7–36 DOI: 10.1007/s40607-014-0009-9 (2014).

[11] N.Ljubešić and F. Klubička. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. Proceedings of the 9th Web as Corpus Workshop (WaC-9). Gothenburg, Sweden (2014).

[12] R.McDonald, J.Nivre, Y.Quirmbach-Brundage, Z.Goldberg, D.Das, K.Ganchev, K.Hall, S.Petrov, H. Zhang, 0.Täckstrom, C.Bedini, Bertomeu Castelló, N., Lee, J. Universal Dependency Annotation for Multilingual Parsing. In Proceedings of ACL (2013).

[13] J.Michelfeit, J.Pomikálek, V.Suchomel. Text Tokenisation Using unitok. In 8th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU, pp. 71-75 (2014).

[14] T.Mikolov, K ai Chen, G.Corrado, J. Dean, 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. Université de Montreal. Scottsdale.

[15] T.Mikolov, E.Grave, P.Bojanowski, Ch.Puhrsch, A.Joulin. 2018. Advances in Pre-Training Distributed Word Representations. Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association. Miyazaki.

[16] M.Musaev, S.Mussakhojayeva, I.Xujayorov, Y.Khassanov, M.Ochilov, A.Varol. USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. (2021)

[17] P.Paikens. Deep Neural Learning Approaches for Latvian Morphological Tagging. In Human Langiage Technologies – The Balitc Perspective (2016).

[18] J. M. Patel. Introduction to Common Crawl Datasets. In Getting Structured Data from the Internet, pp 277–324. Apress (2020)

[19] J.Pomikálek. Removing boilerplate and duplicate content from web corpora. PhD the-sis, Masaryk University, Faculty of informatics, Brno, Czech Republic (2011).

[20] M. O. Rabin (1981). Fingerprinting by Random Polynomials. Center for Research in Computing Technology, Harvard University. Tech Report TR-CSE-03-01 (1981).

[21] R.Řehůřek, P.Sojka. Software framework for topic modelling with large corpora. 2010. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

[22] P.Rychlý. Manatee/Bonito – A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2007. p. 65–70 (2007).

[23] R.Schäfer and Bildhauer. Building Large Corpora from the Web Using a New Efficient Tool Chain. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 486–493 (2012).

[24] H.Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester (1994).

[25] D.Spoustová. J.Hajič, J.Raab, M. Spousta. Semi-Supervised Training for the Averaged Perceptron POS Tagger. In Proceedings of the 12th Conference of the Eu-ropean Chapter of the ACL (EACL), pages 763-771, Athens, Greece, March. Associa-tion for Computational Linguistics (2019).

[26] M.Straka, J.Hajič, J.Straková. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evalua-tion (LREC), Portorož, Slovenia (2016)

[27] V.Suchomel, J.Pomikálek, Efficient Web Crawling for Large Text Corpora. In Adam Kilgarriff, Serge Sharoff. Proceedings of the seventh Web as Corpus Workshop (WAC7). Lyon, pp. 39–43 (2012).

[28] V. Suchomel. Discriminating Between Similar Languages Using Large Web Corpora in the Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN (2019).

[29] M.Sharipov, J.Mattiev, J.Sobirov, R.Baltayev. Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language // https://ceur-ws.org/Vol-3315/paper10.pdf (2022)

[30] A.S. Sobirovich (2022). Development of a parallel corpus of the uzbek and russian languages. Vital Annex: International Journal of Novel Research in Advanced Sciences, 1(5), 152–155.

[31] M. Tursunov (2023). Software of the national corpus of the uzbek language. International Journal of Advance Scientific Research, 3(10), 190-199.

[32] M.S. Tursunov (2022). Development of software for a text corpus in uzbek // Native Languages and Cultures in the Modern Changing World. №1, 62-70 p.

[33] M.Volk (2002). Using the web as corpus for linguistic research. 10.5167/uzh-20339.

[34] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[35] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[36] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[37] K. Elissa, "Title of paper if known," unpublished.

[38] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[39] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[40] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[41] B. R. Mengliyev, S.Hamroyeva, O.Abdullayeva (2023). Scopus-based bibliometric analysis on corpus linguistics for the period of 2017-2021. In *E3S Web of Conferences* (Vol. 413, p. 03008). EDP Sciences.

[42] B.Mengliyev, S.Shahabitdinova, S.Khamroeva, S.Gulyamova, A.Botirova. (2021). The morphological analysis and synthesis of word forms in the linguistic analyzer. *Journal of Language and Linguistic Studies*, *17*(1), 558-564.

[43] K. R.Abdurasulovich, M. B. Rajabovich. (2019). The Role of the Parallel Corpus in Linguistics, the Importance and the Possibilities of Interpretation. *International Journal of Engineering and Advanced Technology*, *8*(5), 388-391.

[44] E.B.Boltayevich, S.S.Samariddinovich, K.S.Mirdjonovna, E.Adali, X.Z. Yuldashevna. Pos Taging of Uzbek Text Using Hidden Markov Model. UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering, 2023, P. 63–68.

[45] E.B.Boltayevich, E.Adali, K.S.Mirdjonovna, X.Z.Yuldashevna, N.Uktamboy o'g'li, X.The Problem of Pos Tagging and Stemming for Agglutinative Languages (Turkish, Uyghur, Uzbek Languages). UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering, 2023, P. 57–62.