IEEE
TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editör Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

IEEE TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

Media type	Part Number	ISBN	Online ISSN
XPLORE COMPLIANT	CFP25L97-ART	979-8-3315-9975-1	2521-1641
CD-ROM	CFP25L97-CDR	979-8-3315-9974-4	

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK'2025)

10th International Conference on Computer Science and Engineering

17-18-19 Eylül 2025 -İstanbul-Türkiye
17-18-19 September 2025 - İstanbul-Türkiye

IEEE Catalog Number : CFP25L97-ART

ISBN : 978-8-3315-9975-1

Additional copies mey 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

UBMK’2025’ye Hoşgeldiniz
Welcome to UBMK’2025

Sevgili Katılımcılar:

UBMK uluslararası nitelikli konferans serisi, 1990 yılından beri düzenli olarak yapılmakta olan Bilgisayar Mühendisliği Bölüm Başkanları toplantılarında alınan bir kararla on yıl önce başlamıştır. Konferansın 10.su IEEE-UBMK-2025 bu yıl 17-18-19 Eylül, 2025 günlerinde İstanbul Teknik Üniversitesinin ev sahipliğinde düzenlemiştir.

IEEE-UBMK-2025 konferansına bu yıl Almanya, Amerika Birleşik Devletleri, Azerbaycan, Fransa, Irak, İngiltere, İsveç, İtalya, Kanada, Kazakistan, Kırım, Kırgızistan, Rusya, Özbekistan, Tataristan, Taylant, Ürdün ve Türkiye’den 610 dolayında bildiri gönderilmiş ve bu bildiriler Türk ve yabancı 250 hakem tarafından değerlendirilmiştir.

Her bildiri en az iki hakem tarafından incelenmiş ve uzlaşma olmadığı durumlarda üçüncü bir hakemin değerlendirmesine başvurulmuştur. Bildiri başına düşen ortalama hakemlik 2,3 olmuştur. Bu değerlendirmelerin sonunda 327 bildirinin sözlü olarak sunulması uygun bulunmuştur. Kabul edilen ve sunulan bildiriler içerik ve kalite ölçünlerini sağlaması durumunda IEEE Xplore’da yayımlanacaktır.

Konferans çalışmalarında, Bilgisayar Mühendisliği Bölüm Başkanları Danışma Kurulu olarak görev almışlardır. Bildirilerin değerlendirilmesi Bilim Kurulu üyeleri tarafından yapılmıştır. Konferansın düzenlenmesi ise Yürütme Kurulunun önerileri doğrultusunda, Düzenleme Kurulu tarafından yapılmıştır.

Son olarak, konferansın başarılı bir şekilde yürütülmesi için tüm olanaklarını sunan İstanbul Teknik Üniversitesi Rektörü Sayın Prof. Dr. Hasan Mandal’a teşekkür ediyoruz. Ayrıca Düzenleme Kuruluna, bildirileri titizlikle değerlendiren Bilim Kurulu Üyelerine ve değerli araştırmalarının sonuçlarını bilişim camiası ile paylaşan bildiri sahiplerine teşekkürlerimizi iletiriz.

Prof. Dr. Eşref ADALI
UBMK-2025 Konferans Başkanı ve Bildiri Kitabı Editörü

Dear Participants:

The UBMK international conference series started nine years ago with a decision taken at the Computer Engineering Department Heads (BMBB) meetings, which have been held regularly since 1990. The 10th edition of the conference, UBMK'25, was held this year on October 17-18-19, 2025, hosted by İstanbul Technichal University.

This year, approximately 610 papers were submitted to the IEEE-UBMK-2025 conference from Germany, the United States, Azerbaijan, France, Iraq, the United Kingdom, Sweden, Italy, Canada, Kazakhstan, Crimea, Kyrgyzstan, Russia, Uzbekistan, Tatarstan, Thailand, Jordan, and Turkey, and these papers were evaluated by 250 Turkish and foreign referees.

Each paper was evaluated at least by two referees, and in cases where there was no consensus, a third referee was consulted. At the end of these evaluations, 327 papers were accepted for oral presentation. Accepted and presented papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

During the conference, Heads of Information Engineering Departments took part in the Advisory Board. The evaluation of the papers was made by the members of the Scientific Committee. The conference was organized by the Organizing Committee in line with the recommendations of the Executive Committee.

Finally, we would like to thank İstanbul Technical University Rector Prof. Dr. Hasan Mandal for his continued support for the success of the conference. In addition, we would like to thank the Organizing Committee, the Scientific Committee Members who carefully evaluated the papers, and the owners of the papers who shared the results of their valuable research with the informatics community.

Prof. Dr. Esref ADALI
UBMK'25 Conference Chair and Proceedings Editor

Düzenleyenler / Organizer

Destekleyenler / Sponsors

Methods of Morphological and Syntactic Analysis for the Uzbek Language

Elov Botir Boltayevich
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
elov@navoiy-uni.uz

Abdullayeva Oqila Xolmo'minovna
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
abdullayevaoqila@gmail.com

Abdullayeva Firuza Sharipovna
Dept. of Social and Political Sciences
Bukhara State University
Bukhara, Uzbekistan
sharipovna14@gmail.com

Alavutdinova Nadira Ganiyevna
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
alavutdinovanodira@gmail.com

***Abstract*—The linguistic features of the Uzbek language - complex agglutinative morphology, free word order, and limited resources - necessitate a specialized approach and thorough research in the application of morphological and syntactic methods. Within the framework of the study, morphological analysis methods and syntactic analysis methods are reviewed based on scientific sources. Each section presents the existing advantages and disadvantages, experience of their use in the Uzbek language, as well as a comparative analysis with foreign languages. Rule-based methods, statistical models (HMM, CRF, etc.), Neural network-based approaches (BiLSTM-CRF, seq2seq) of morphological analysis in the Uzbek language are discussed, and the results are given in examples and percentages. It is shown that syntactic parsing is implemented using dependency and constituency parsing analysis methods. The issue of building a UD treebank for the Uzbek language with SOV order is considered. The impact of complex morphological structure and free word order in sentences on the construction of parsers is highlighted. As a result of the studied approaches, the issue of building hybrid parsers, integrating them with morphological analysis and assigning grammatical categories of words to the parser is raised. Also, the development of neural constituency parsers based on neural networks and the effectiveness of the results obtained from them are analyzed.**

Keywords—natural language processing (NLP), POS tagging, parser, syntactic parsing, treebank.

I. Introduction

Linguistic knowledge relevant to NLP should include lexical, syntactic, semantic, and pragmatic features. Each feature represents linguistic information in a different way. For example, a lexical feature might include knowledge about the basic word-level components (e.g., morphemes) and their inflectional forms; a syntactic feature might include knowledge about how words or word combinations in a given language form sentences; semantics might include knowledge about what meaning words or word combinations have in sentences and their semantic valence; and pragmatics might include knowledge about changes in the focus of speech in conversation, and about interpreting the meaning of a sentence in a given context.

In contemporary linguistics, modern techniques for text annotation have been widely developed, along with various tagging tasks. The sections that follow focus on the methods of morphological and syntactic analysis.

II. Literature Review

Morphological analysis serves to automatically identify the internal structure of words, including their roots, affixes, and grammatical meanings. This task is particularly complex in agglutinative languages such as Uzbek, where a single word may contain numerous suffixes that convey various grammatical functions. For example, the word "kutubxona-lar-i-ga-mi" consists of the lemma kutubxona (library), the plural suffix -lar, third-person possessive -i, dative case marker -ga, and the interrogative particle -mi [1]. To analyze such rich morphology, both rule-based linguistic frameworks and statistical models are employed. The following section explores the primary approaches to morphological analysis in detail: rule-based systems, statistical models, neural network-based approaches, as well as the application of Finite-State Transducers (FSTs).

In rule-based approaches, the morphological structure of words is analyzed using a set of linguistic rules and lexicons developed by linguists. For example, in Uzbek, such systems include lists of parts of speech and the affixes that can attach to them, along with rules governing the sequence of affixation [2]. Typically, rule-based morphological analyzers utilize a dictionary of root forms and a comprehensive list of affixes, with explicit rules that define how these elements can combine.

The primary advantage of manually crafted systems lies in their precision and consistency, as they often yield linguistically accurate results. For Turkic languages, including Uzbek, researchers have emphasized the effectiveness of finite-state transducers (FSTs) for automatic morphological analysis [3].

However, a key limitation of rule-based systems is their inability to handle out-of-vocabulary words, such as neologisms, foreign names, or domain-specific terminology

and their limited capacity to disambiguate homonymous forms in context. In Uzbek, for example, the auxiliary verb “qilmoq” frequently forms compound constructions such as “yordam qildi” (he helped) or “ta’bir qilmoq” (to express). A rule-based analyzer would treat such expressions as separate words but may fail to capture their combined semantic meaning. Similarly, rule-based part-of-speech (POS) taggers rely primarily on affixes and neighboring words, without adequately modeling broader context, which can lead to errors in certain cases [4].

Despite these limitations, rule-based methods are of particular importance for low-resource or under-resourced languages in the early stages of computational processing. They are often used to generate initial annotations for training more advanced models, such as statistical or neural network-based approaches (see Fig. 1).

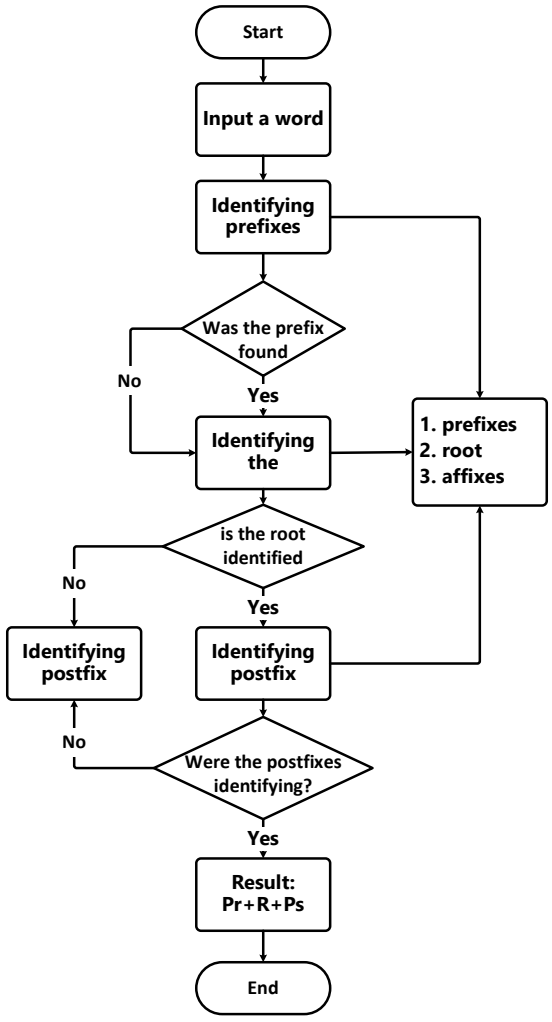

Fig. 1. Root Identification Process in Rule-Based Systems

Statistical or data-driven models perform morphological analysis and tagging by learning from large amounts of annotated texts. Among these, models such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) have been widely used. According to the literature, the introduction of HMM [5] and CRF [6] models has significantly improved the accuracy of POS taggers [4]. In HMM models, probabilities are calculated based on word sequences to determine the most likely grammatical tag for each word, while CRF models maximize the conditional probability across the entire sequence. These models, when trained on annotated corpora, derive statistical properties from the data and make automatic, rule-free decisions. Such approaches have also been tested for Uzbek. For instance, Elov et al. [7] applied the HMM model to small-scale Uzbek texts and demonstrated a working prototype with promising results.

However, a key limitation of statistical models is their reliance on large and balanced datasets. In the case of Uzbek, large-scale annotated corpora have only recently become available; previously, such resources were insufficient. A multi-head attention-based POS tagger proposed by Murat and Ali [8] outperformed earlier BiLSTM, CNN, and CRF-based models by 4.1%, achieving 79.74% accuracy. While this surpasses the performance of rule-based systems, it still falls short of the levels attained by deep learning-based models.

Neural networks enable the learning of abstract features from large datasets and allow for the development of more accurate models. In particular, the Bi-directional LSTM + CRF (BiLSTM-CRF) architecture has proven highly effective in sequence labeling tasks, including part-of-speech and morphological tagging [11]. BiLSTM captures contextual dependencies from both the left and right of a word, creating internal state vectors, while the CRF layer selects the most likely tag sequence for the entire sentence. This model has achieved accuracy rates of up to 97% in English [12], and applying deep learning to Uzbek has also shown significant improvements. For example, a monolingual BERT-based POS tagger has achieved an average accuracy of 97.8% [13]. Deep learning models are now being applied to Uzbek as well. Notably, Elov et al. [14] developed the UzbMorphAnalyzer system, which offers an integrated solution for stemming, lemmatization, and POS tagging in Uzbek. This system combines traditional FST components with probabilistic models to provide comprehensive morphological analysis.

Syntactic analysis, or parsing, identifies the grammatical relationships between words in a sentence. There are two primary and widely used approaches to syntactic parsing: constituency parsing and dependency parsing.

The aim of constituency parsing is to break down a sentence into hierarchical phrase structures according to grammatical rules. This type of parsing relies on formal grammatical frameworks such as the Chomsky hierarchy and is particularly effective for languages with relatively fixed word order, such as English (SVO structure), where grammatical relations are largely expressed through word order. However, in free word-order languages like Uzbek, where the same idea can be expressed in multiple syntactic orders, traditional phrase structure grammars become more complex and less efficient (see Fig. 2). The flexibility in word placement increases the ambiguity and reduces the effectiveness of purely structure-based parsing approaches.

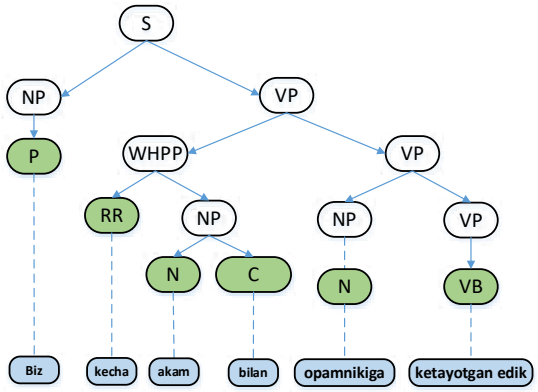

Fig. 2. Tree Diagram of Constituency Parsing in Uzbek

Unlike constituency parsing, dependency parsing focuses on the direct grammatical relations between words in a sentence. Instead of using explicit phrase nodes, this approach connects each word to another "head word". For example, in the sentence "Bolalar daftarga yozdi" ("The children wrote in the notebook"), yozdi ("wrote") is the predicate (main verb/root), Bolalar ("children") is linked as the subject (nsubj), and daftarga ("in the notebook") is linked as the object (obj). In dependency parsing, each arc in the tree is labeled with a grammatical relation type such as nsubj, obj, advmod, etc. One major advantage of this approach is its relative independence from word order. Regardless of the syntactic order, the dependency tree retains the same relational structure. As a result, dependency parsing is considered more robust for free word order and morphologically rich languages [1].

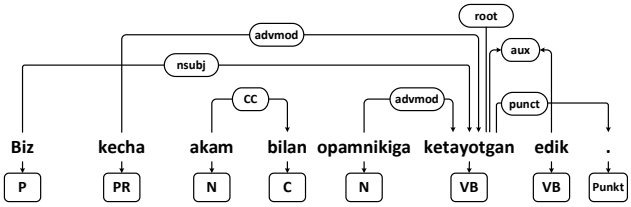

Fig. 3. Example of a Dependency parsing in Uzbek

Syntax of the Uzbek language relies on dependent structures built around content words, and the verb typically appears at the end of the sentence. Thus, dependency grammar provides a practical framework for formalizing syntax of the Uzbek language (see Fig. 3). Currently, there is no fully developed constituency grammar for Uzbek (such as CHELG or HPSG), but small-scale corpora annotated in dependency format have been initiated. Below we will discuss this approach using the UD format and the Uzbek treebank.

Universal Dependencies (UD) is an international project that offers a single, agreed-upon dependency annotation format for different languages. In UD, each word is assigned a universal POS tag (UPOS), a set of morphological features, and a dependency relation linking it to its head word. For example, labels such as nsubj (nominal subject), obj (object), and obl (oblique modifier) are used consistently across all languages. UD also defines a shared inventory of part-of-speech categories (e.g., noun, verb, adjective) and morphological features such as number, tense, and case. The goal of the UD format is to create treebanks in the same scheme for all languages, without having to invent a separate formalism for each language, thereby facilitating multilingual research.

The basic word order of the Uzbek language is considered to be SOV (Subject–Object–Verb). That is, in typical sentences, the possessive comes at the beginning, and the participle comes at the end (Bugun men universitetga keldim - here "men" is S, "universitetga, bugun" is O, "keldim" is V). However, in practice, the syntax of the Uzbek language is very free: words can be rearranged in a sentence according to the emphasis of a certain meaning, and the meaning of the sentence is preserved (often only the actual parts are stressed). For example, "Bugun men universitetga keldim" (simple form); "Universitetga men bugun keldim" may be used for emphasis or stylistic purposes. Such free word order is also reflected in morphological indicators - the suffixes of nouns indicate what role they play in the sentence (whether they are determiners, complements, or cases).

The flexible word order in a sentence affects syntactic analysis in several ways. From the perspective of constituency parsing, it is difficult to identify phrases in Uzbek based on strict rules, because, the subject does not always precede the verb, or a modifier can follow the word it modifies (e.g., kitob yaxshi, yaxshi kitob). Therefore, a traditional CFG grammar may require a large number of rules and exceptions. Dependency parsing, on the other hand, addresses this through direct relations: for instance, in the sentence "Bugun men universitetga keldim", regardless of the word order, keldim is the main verb, men is linked to it as the subject (nsubj), and universitetga as an oblique modifier (obl).

In Uzbek, the omission of sentence constituents (ellipsis) occurs frequently, particularly the subject or predicate, because the verb form clearly indicates person and number. For example, "Keldim" ("I came") is a sentence where the subject "men" ("I") is omitted. In such cases, the parsing process assumes that the implicitly omitted element is connected to the main verb. The parser attempts to infer this from context, but it may also misinterpret it.

So far, when looking at the results of parsing experiments conducted for the Uzbek language, as expected, the outcomes are lower compared to English. For example, training the Stanza neural parser on a small training set of 400 sentences resulted in only 52% LAS F1 score on the test set. This can be compared to a similar model trained on the English Penn Treebank, which achieves around 90% LAS. Of course, a major part of this gap is due to the limited amount of training data, and partly due to the language's inherently free structure. Therefore, to improve syntactic parsing in Uzbek, the following measures are proposed:

1) increasing the size of the treebank with texts from more diverse domains;
2) integrating with morphological analysis by feeding grammatical categories of words into the parser;

applying transfer learning from multilingual models (using models trained on related languages such as Turkish or Kazakh).

III. METHODS

Each method and approach discussed above has its own strengths and limitations. Below, we summarize them in the form of a general conclusion, with particular attention to the specific characteristics of the Uzbek language.

1. Rule-based methods: Advantages – They are linguistically consistent and precise; relatively few errors

occurre and are prepared for grammatically irregular cases (as such rules are predefined by experts). In Uzbek, this approach allows for the analysis of even incomplete sentences to a certain extent (e.g., spelling and grammar checkers based on linguistic rules). Disadvantages – Developing rules requires a large amount of manual labor; the system must be regularly updated when new words or exceptions appear; and most importantly, it lacks contextual understanding. For instance, a rule-based POS tagger assigns tags based only on affixes and ignores the broader sentence context, which often leads to errors around ambiguous or functional words.

2. Statistical and classical ML methods (HMM, CRF, etc.): Advantages - instead of manually building rules where there is information, the model learns itself, which means it adapts faster and easier; solves complex boundary cases through probabilities (for example, whether the word "yuz" is a number or a human limb - it gives a bias towards the case that occurs more often in the corpus). Disadvantages - require a lot of data; if the corpus is small, the statistical model will not be perfect or will make more mistakes than rule-based methods. In the early stages of Uzbek, such models were slower due to the lack of a sufficient corpus. For example, the HMM-based POS tagger was able to achieve 75% accuracy, while the rule-based tagger showed 90% in its internal test [4]. The advantage of CRF is that it can incorporate various features, for example, it works better than the rule that if the 2 letters at the end of a word are "-im", then it is probably a possessive sign, because - im can also be a declension form (it considers the rules of possessive: -m and -im; declension: -m together).

3. Deep learning models (BiLSTM, CNN, Transformer): Advantages – automatically models a very large amount of linguistic connections, that is, it reflects rules that were not entered manually in its analysis. Therefore, it achieves the highest results: for example, in English, POS labeling reached 97% accuracy, close to human level [11]. In Uzbek, BERT-based models also significantly outperformed previous approaches (91% accuracy in labeling, parsing or other tasks are also expected). Neural models fully take into account the context, for example, taking into account the previous and next words, the meaning of the entire sentence, and excludes inappropriate labels – this is a difficult task for rule-based or traditional statistical models. Disadvantages – require a lot of information and resources: a large fixed corpus or a trained model (pretrained) is required for training. Another one is the difficulty of explanation: it is difficult to explain why the model made such a decision (lack of transparency). Another problem in practice is that if the data is uneven (for example, some constructs are very rare), the model will learn them poorly.

4. FST and rule-based morphological analysis: Advantages – Provides all possible analyses completely and accurately; it is an indispensable tool for morphologically rich languages (especially for languages like Turkish, Finnish, and Korean). In Uzbek as well, an FST analyzer can segment the grammatical structure of any given word (if it exists in the lexicon, or even generate analyses based on word formation rules). Disadvantages – It does not resolve ambiguity: if there are multiple formally similar analyses for a word, it outputs all of them, but determining which one is correct is deferred to a later stage. Additionally, if the FST database is not regularly updated, it cannot recognize new jargon or dialect words. However, these limitations are not sufficient reasons to abandon FST for morphological analysis, since statistical models can be used to resolve such issues. Training such models on an infinite combination of affixes is practically difficult. Therefore, in many systems, the initial analysis is performed using FST, followed by ranking or disambiguation using machine learning models.

5. Dependency parsing and Constituency parsing: The advantage of dependency parsing is that it is independent, adaptable to free word order, supported by a universal standard like UD that incorporates experience from many languages, and is convenient for transitioning to semantic analysis (predicate-argument structure is easily identifiable). The advantage of constituency parsing is that it aligns with traditional syntactic analysis (e.g., modifier + head is treated as a single phrase, whereas in dependency parsing, they are two separate relations); it is also easier to explain in some linguistic theories and manuals that are based on constituency. In Uzbek, as mentioned above, the dependency-based approach has proven effective and has been adopted.

6. Language-specific challenges of Uzbek: As discussed above, Uzbek is morphologically rich and syntactically flexible. On one hand, this complicates analysis (many components, many variants), but on the other hand, it simplifies certain solutions (e.g., case markers help identify sentence constituents). Another challenge in Uzbek is the coexistence of Cyrillic and Latin scripts: texts may appear in either script. This must be considered when building models, for example, a special version of UzBERT was developed to include Cyrillic texts. Additionally, the lack of resources and their fragmentation, such as the absence of large corpora, the recent emergence of a treebank, and the lack of semantically annotated corpora, also present significant obstacles.

IV. RESULTS AND DISCUSSION

There are not many scientific experiments conducted on morphological and syntactic analysis in the Uzbek language, but those that exist show the effectiveness of one or another approach. This section presents some important results and indicators. Accuracy, F1-measure (combination of accuracy and completeness), perplexity (for language models), etc. are used to compare models and methods. Some experimental results are presented in the table I below:

TABLE I. ACCURACY OF LANGUAGE MODELS

Task/resources	Method/Model	Measurement (metric)	Result
Identifying part of	Rule-based method (UzbekTagger)	Accuracy	82.2%

speech (POS, uz)	HMM/CRF model (baseline)	Accuracy	87.5%
	Neyron (BiLSTM+attention, 2024)	Accuracy	89.7%
	BERT fine-tuning (Elov)	Accuracy	97.8%
Dependency parsing (uz)	Stanford Stanza (UD 500 sentences)	LAS (F1)	8.9

The BERT-based POS tagging model developed by B. Elov achieved 97.8% accuracy, significantly outperforming the rule-based system. This improvement was largely due to the newly annotated corpus of 500 sentences, previously, the lack of data had given rule-based approaches an advantage. Now that evaluation has become possible for the first time, shortcomings have been identified: for example, the rule-based tagger makes decisions based only on the immediately adjacent word, whereas the BERT model considers the entire sentence and assigns tags more accurately.

As for dependency parsing, current results remain low–52% LAS, but this is based on training the model with only 400 sentences. For comparison, parsers trained on English treebanks with 40,000 sentences achieve over 90% LAS. Therefore, increasing the dataset size in Uzbek is likely to significantly reduce errors. Another observed phenomenon is that in Uzbek parsing results, label accuracy (identifying the type of relation) is relatively high, while head accuracy (determining which word the relation connects to) is lower. This indicates that the parser often correctly predicts the relation type (e.g., that it is a subject), but errs in identifying the head word, most likely due to the language's free word order.

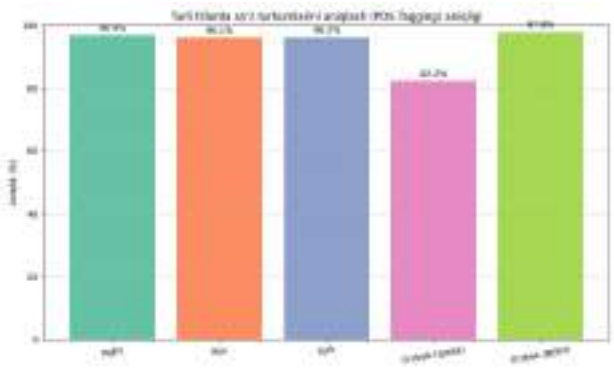

Fig. 4. Accuracy of word segmentation (POS tagging) in different languages

Automatic word class recognition in English, Russian, and Turkish is quite high (around 96–97%) [11]. In Uzbek, the previously used rule-based approach showed an accuracy of 82.2% in independent tests [15]. The newly trained UzBERT-based model achieved a significant improvement, reaching 97.8%. This graph clearly shows the effectiveness of the models that learn the context and the suffix in the Uzbek language.

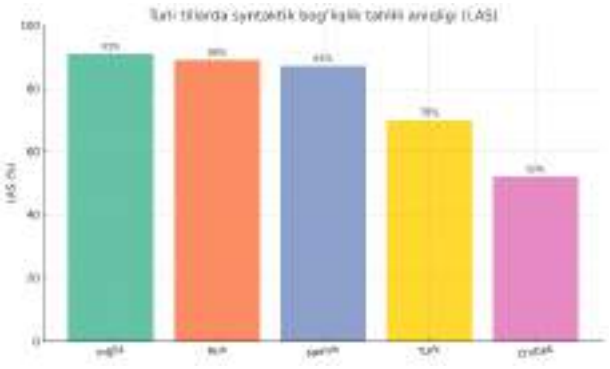

Fig. 5. Accuracy of syntactic dependency analysis (LAS) in foreign languages and Uzbek

In English, due to the presence of large treebanks, connection parsing achieves 90% accuracy (StanfordNN parser, 2017) [16]. In Russian, results around ~80% have also been shown based on the large SynTagRus corpus. Although Turkish is a free-form language, it has a medium-sized treebank and an LAS of around 70% has been observed (CoNLL 2017 collective results). In Uzbek, the treebank is very small (500 sentences), so far LAS of 52% is observed, i.e., very low accuracy. These figures show that the size of the resource and syntactic complexity have a significant impact on the parsing result: it is currently difficult to create an accurate parsing model in languages with free-form word order and small corpuses, but significant improvements are expected with increasing corpus size.

V. STUDY OF THE PROBLEM

Comparison with foreign languages: English, Turkish and Russian

When evaluating NLP approaches for Uzbek, it is useful to compare it with the examples of English, Turkish, and Russian. English is an analytical language with simple morphology and a large amount of resources; Turkish is an agglutinative language very close to Uzbek, with a free word order and resources that have increased in recent years; Russian is an inflectional language with moderately rich morphology, relatively free word order, and at the same time has a large treebank and corpus resources.

Morphological analysis and tagging: While in English morphological analysis grammatical forms are mainly limited to determining the tense of verbs or the plural form of nouns, in Turkish and Uzbek each word can be complicated by a chain of successive suffixes. For this reason, in English, regular morphological analysis can be solved even with simple regex rules or with a common dictionary (e.g. WordNet lemmas). In Turkish, the FST analyzer standard created by Oflazer (1994) is still the main tool - it is said to correctly cover 98-99% of the inputs. Such an FST appeared in Uzbek recently and its coverage has not yet been fully evaluated, but in theory it should be close to Turkish. For morphological analysis in Russian, there are regular+statistical analyzers, such as pymorph2, which find a word in a dictionary and give a suitable paradigm; its accuracy is higher than 95%. Also, in Russian, the Neural morphological tagger (UDPipe model) showed 98% accuracy (POS+feats). So, due to the abundance of resources and the simplicity of the language, morphological analysis becomes more complicated in the order English > Russian > Turkish > Uzbek. However, it is convenient to exchange experiences

between Turkish and Uzbek - for example, if morphological disambiguation based on CRF in Turkish is 96%, it is assumed that it can be achieved in Uzbek as well [9].

Parts of speech (POS) and morphological tagging: In English, there is a Penn Treebank with millions of tokens for POS tagging, and state-of-the-art models achieve 97.5% accuracy [11]. Even simple HMM models exceed 95%. In Russian, too, POS tagging has reached 96-97% thanks to SynTagRus and other corpora (the StanfordNLPCore Russian model is above 96% [16]). In Turkish, there is a large Bosphorus treebank and IMST treebank, but due to the complex morphology, it is more important to identify the full set of morphological tags, not POS - the latest results are around 93-94% (for example, 96% with the transformer model). In Uzbek, 500 sentences have been tagged for POS so far, and the first neural tagger achieved 97.8%, but the accuracy of full morphological tags (agreement, number, possession, etc.) is even lower (it can be around 70-80%, approximately). Here it seems that the corpus size is the decisive factor: English/Russian is limited to hundreds of thousands of words, Turkish to tens of thousands, and Uzbek to a few thousand tokens. If Uzbek POS corpora exceed the 100k token limit in the coming years, the tagger accuracy may also approach 95%.

Syntactic parsing: In English, constituency parsing is currently 95% F1 (PTB, 2021), dependency parsing LAS is 92% (stanfordnlp, 2018). In Russian, LAS has reached 88-90% based on the UD treebank (SynTagRus 50k sentences) (neuron model, 2020). There are several treebanks in Turkish (10k sentences), in which LAS is 75-80%. In addition, it has been shown that accuracy increases when morphological units are parsed as basic units instead of words in Turkish. A treebank in Uzbek is currently being developed, and not many experiments have been conducted yet, but the 52% mentioned above is the minimum model result. In parsing, English and Russian are much more advanced due to the large size of the created corpora, the strictness of word order in sentence construction, and long-term research experiences. The small size of the corpora, free word order, and complex morphology in Turkish and Uzbek complicate the model. But it is worth noting that the approaches in Turkish are worth trying out in Uzbek, since they are structurally and typologically very close. For example, a parser prepared for the Turkish treebank can be applied to Uzbek text and then fine-tuned to the Uzbek tone (transfer learning). Similarly, there was an experience in building an Uzbek WordNet based on the Turkish WordNet (BalkaNet).

VI. Conclusion

In conclusion, the experiments carried out show that the development of language technologies initially requires a lot of work and resources, but once mastered, many results can be achieved in a short time. The results of using methods developed in different languages vary depending on the nature of the language. When comparing Uzbek with English, Russian and Turkish, the accuracy of POS labeling results in percentages is different depending on the consistency of word order in sentences, the presence of large-scale corpora, long-term experience, the development of artificial intelligence-based models (for example, models such as UDPipe, RuBERT, RuGPT3), the presence of multi-disciplinary analyzed dictionaries, WordNets. Current research on the Uzbek language is gradually moving towards the development of deep hybrid models. It can be observed from the analysis results that neural approaches in the Uzbek language have a significant advantage over traditional methods. As a result of the research, we found that the BERT model achieved 97.8% accuracy in POS tagging, which is much higher than the rule program. The Uzbek language is one of the languages with its own historical development, with a complex structure and rich structure. The corpus of the existing Uzbek language database based on the Krill and Latin alphabets in the Uzbek language serves to solve problems in NLP by performing initial processing stages on the collected texts. In the near future, we can hope to see artificial intelligence systems that can freely understand, analyze and synthesize texts in the Uzbek language. The morphological and syntactic methods that we analyzed are the foundation of these systems.

References

[1] N. Abdurakhmonova, “Dependency parsing based on Uzbek Corpus,” in *Proc. Int. Conf. on Language Technologies for All (LT4All)*, 2019.

[2] K. S. Mirdjanovna, “Finite State Machine Model for Uzbek Language Morphological Analyzer,” in *2021 6th Int. Conf. on Computer Science and Engineering (UBMK)*, Sep. 2021, pp. 395–400.

[3] E. Adalı, *Türkçe Doğal Dil İşleme*, Ankara, Turkey: Akçağ Yayınları, 2020.

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

[5] J. Kupiec, “Robust part-of-speech tagging using a hidden Markov model,” *Comput. Speech Lang.*, vol. 6, no. 3, pp. 225–242, 1992.

[6] P. Awasthi, D. Rao, and B. Ravindran, “Part of speech tagging and chunking with HMM and CRF,” in *Proc. NLPAI ML Contest*, 2006.

[7] E. B. Boltayevich, S. S. Samariddinovich, K. S. Mirdjonovna, E. Adalı, and X. Z. Yuldashevna, “POS tagging of Uzbek text using hidden Markov model,” in *2023 8th Int. Conf. on Computer Science and Engineering (UBMK)*, Sep. 2023, pp. 63–68.

[8] A. Murat and S. Ali, “Low-resource POS tagging with deep affix representation and multi-head attention,” *IEEE Access*, 2024.

[9] H. Özer and E. E. Korkmaz, “Transmorph: A transformer based morphological disambiguator for Turkish,” *Turk. J. Electr. Eng. Comput. Sci.*, vol. 30, no. 5, pp. 1897–1913, 2022.

[10] S. Meftah, N. Semmar, and F. Sadat, “A neural network model for part-of-speech tagging of social media texts,” in *Proc. LREC 2018 - 11th Int. Conf. on Language Resources and Evaluation*, May 2018.

[11] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF models for sequence tagging,” *arXiv preprint arXiv:1508.01991*, 2015.

[12] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a large annotated corpus of English: The Penn Treebank,” *Comput. Linguist.*, vol. 19, no. 2, pp. 313–330, 1993.

[13] E. B. Boltayevich, E. Adalı, K. S. Mirdjonovna, A. O. Xolmo'Minovna, X. Z. Yuldashevna, and X. N. Uktamboy O'g'li, “The Problem of POS Tagging and Stemming for Agglutinative Languages (Turkish, Uyghur, Uzbek Languages),” in *2023 8th Int. Conf. on Computer Science and Engineering (UBMK)*, Sep. 2023, pp. 57–62.

[14] E. B. Boltayeviç, X. Z. Yuldashevna, U. S. Mamurjonovna, N. S. Ermamatovich, A. Ş. A. Kızı, and M. Shavkatjon, “Algorithms for Parsing Roots and Stems of Words in Uzbek Language,” in *2024 9th Int. Conf. on Computer Science and Engineering (UBMK)*, Oct. 2024, pp. 126–130.

[15] L. Bobojonova, A. Akhundjanova, P. Ostheimer, and S. Fellenz, “BBPOS: BERT-based Part-of-Speech Tagging for Uzbek,” *arXiv preprint arXiv:2501.10107*, 2025.

[16] https://github.com/stanfordnlp/CoreNLP/issues/480