PART-OF-SPEECH TAGGING AND ITS APPLICATIONS

SRSTI 17.07.13

Xonzoda Akramova
Student, Jizzakh State Pedagogical University,
Republic of Uzbekistan, Jizzakh
ORCID: https://orcid.org/0009-0009-5282-2819
E-mail: xonzodaakramova55@gmail.com
Hakima Abdullajonova (Botirova) Abdukodir kizi
Scientific supervisor, Jizzakh State Pedagogical University,
Republic of Uzbekistan, Jizzakh
E-mail: hakimabotirova9@gmail.com
ORCID:0000-0001-6418-4599

Abstract

This article investigates the theoretical foundations, methodologies, and applications of part-of-speech (POS) tagging within the framework of corpus linguistics. POS tagging, as a fundamental stage of natural language processing (NLP), enables computational systems to identify and classify words into grammatical categories, thereby providing a structural basis for syntactic and semantic analysis. The study reviews the evolution of POS tagging approaches—from rule-based to stochastic, hybrid, and deep learning models—and emphasizes their role in linguistic annotation and empirical language research. Using corpus-based examples, the research explores how accurate POS tagging enhances tasks such as parsing, information retrieval, text classification, and machine translation. The discussion highlights challenges specific to morphologically rich and low-resource languages, including Kazakh and Uzbek, and outlines strategies for building effective language resources. The study concludes that POS tagging is not only a technical process but also a methodological instrument for linguistic inquiry, linking computational technology with linguistic theory. The findings contribute to both applied NLP and

theoretical corpus linguistics by demonstrating that well-designed POS annotation schemes can reveal underlying grammatical regularities and patterns of language use.

Keywords: part-of-speech tagging, corpus linguistics, natural language processing, linguistic annotation, morphologically rich languages, text analysis, machine learning, computational linguistics, Kazakh language, Uzbek language.

Introduction

The rapid development of computational linguistics and the increasing availability of large electronic corpora have transformed linguistic research from a primarily theoretical discipline into an empirical, data-driven science. At the core of corpus linguistics lies the systematic analysis of authentic language data, facilitated by annotation layers that enrich the corpus with linguistic information. Among these layers, part-of-speech (POS) tagging plays a fundamental role, serving as a bridge between raw text and higher-level linguistic analysis.

Part-of-speech tagging refers to the process of assigning grammatical categories—such as noun, verb, adjective, or adverb—to each token in a text. This process, though conceptually simple, requires sophisticated modeling of linguistic regularities and statistical dependencies. In the context of corpus linguistics, POS tagging provides essential metadata that enables frequency analysis, collocation studies, syntactic parsing, and semantic modeling.

The significance of POS tagging extends beyond theoretical linguistics; it underpins numerous applications in natural language processing (NLP), including information retrieval, sentiment analysis, machine translation, and automatic speech recognition. Thus, an understanding of POS tagging techniques and their integration within corpus-based frameworks is critical for both linguists and computational scientists.

The objective of this article is to examine the theoretical principles, methodological developments, and practical applications of POS tagging in corpus linguistics. The study also addresses the challenges of implementing POS taggers for morphologically complex and low-resource languages, focusing on the Turkic linguistic context, particularly the Kazakh and Uzbek languages.

Materials and Methods

This research adopts a corpus-linguistic methodology, emphasizing empirical observation over intuition. A corpus in this context refers to a structured, electronically stored collection of texts representing real language use. The analysis draws on various types of corpora—balanced, specialized, and parallel—used in prior linguistic and computational studies.

Corpus data enable quantitative evaluation of tagging performance and linguistic validity. The study examines established annotated corpora such as the British National Corpus (BNC), Penn Treebank, Universal Dependencies (UD) corpora, and smaller Turkic-language datasets developed by academic and national research projects.

Methodological Approach

The article employs a comparative analytical method to review and synthesize findings from different POS tagging paradigms:

1. Rule-based tagging (linguistic grammar and dictionary lookup),
2. Stochastic tagging (Hidden Markov Models and conditional probabilities),
3. Hybrid models combining rules and probabilities,
4. Neural approaches employing deep learning (BiLSTM-CRF, Transformer-based models).

Evaluation metrics such as accuracy, precision, recall, and F-measure are discussed in relation to tagging quality. Additionally, the article considers error analysis as a diagnostic tool for improving annotation consistency in corpora.

Data and Tools

Examples and data illustrations are drawn from existing linguistic resources and open-source NLP frameworks, including NLTK, SpaCy, and Stanza, which provide standard POS tagging pipelines. For morphologically rich languages, morphological analyzers and lemmatizers (e.g., Apertium-based modules and UDPipe) are considered essential tools for preprocessing.

Literature Review

The development of POS tagging has undergone several key phases since the mid-20th century. Early computational linguistics, represented by pioneers such as Zellig Harris and Noam Chomsky, established the theoretical groundwork for syntactic classification. However, practical tagging systems emerged in the 1960s and 1970s with the introduction of automatic text analysis for English.

Rule-Based Taggers

The first generation of taggers, such as the Greene and Rubin TAGGIT system (1971), relied on handcrafted grammatical rules and extensive lexicons. These systems achieved reasonable accuracy for limited domains but required significant manual effort and linguistic expertise. Later, the ENGCG (English Constraint Grammar) introduced constraint-based disambiguation, a technique later adapted for many European languages.

Statistical Taggers

A major shift occurred with the rise of stochastic models in the 1990s. The Hidden Markov Model (HMM) became the dominant framework for probabilistic POS tagging, as demonstrated in the Penn Treebank Tagger (Charniak, 1997) and Brill Tagger (1995), which combined rule induction with probabilistic learning. These models achieved tagging accuracies above 95 % for English and inspired parallel efforts for other languages.

Hybrid and Neural Approaches

In the 2010s, the rise of machine learning and deep neural networks transformed POS tagging. Neural architectures—particularly bidirectional LSTM (BiLSTM) models combined with Conditional Random Fields (CRF)—captured contextual dependencies beyond the capabilities of HMMs. More recently, Transformer-based models (e.g., BERT, RoBERTa, XLM-R) achieved state-of-the-art accuracy across multilingual corpora (Devlin et al., 2019).

POS Tagging in Corpus Linguistics

From a corpus-linguistic perspective, POS tagging serves as a gateway to grammatical annotation and quantitative syntax. Leech and Wilson (1999) emphasized that tagged corpora enable frequency-based grammatical studies and cross-linguistic comparison. The Universal Dependencies (UD) initiative further standardized POS and morphological annotation schemes, facilitating multilingual linguistic research.

POS Tagging for Turkic and Low-Resource Languages

Languages such as Kazakh, Uzbek, Kyrgyz, and Tatar present unique challenges due to agglutinative morphology and rich inflectional paradigms. The absence of large, annotated corpora and morphological analyzers complicates automatic tagging. Research by Mukushev et al. (2020) and Kadirova (2021) demonstrates that transfer learning and character-level embeddings significantly improve tagging accuracy for Turkic languages.

Thus, current trends in POS tagging for low-resource languages involve leveraging multilingual pretrained models and corpus-based annotation transfer from typologically similar languages.

Results and Discussion

1. Theoretical and Practical Relevance

The analysis confirms that POS tagging remains an indispensable component of corpus-based linguistic research and natural language processing. The integration of accurate tagging facilitates the extraction of syntactic and lexical patterns, allowing scholars to explore language variation, register, and diachronic change.

In computational applications, POS tagging provides the structural foundation for tasks such as syntactic parsing, named entity recognition, sentiment classification, and machine translation. For instance, POS-tagged corpora improve the alignment quality in bilingual datasets, thereby enhancing translation accuracy.

2. Comparative Evaluation of POS Tagging Models

Empirical studies show that rule-based systems typically achieve accuracies of 85–90 %, while statistical HMM-based taggers reach 94–96 %. Neural models trained on large, annotated corpora exceed 97 % accuracy for high-resource languages such as English or Chinese. However, accuracy decreases substantially (often below 90 %) for low-resource languages without sufficient training data.

The key to improving tagging accuracy in these contexts lies in combining linguistic insight with machine learning efficiency. Hybrid approaches—such as rule-enhanced neural networks—capitalize on morphological constraints while adapting to contextual variability.

3. Corpus Design and Annotation Quality

The design of the corpus significantly affects tagging outcomes. Balanced corpora, representing diverse text genres and registers, allow taggers to generalize across contexts. Annotation consistency, achieved through detailed tagset guidelines and inter-annotator agreement checks, ensures reliability and reproducibility.

For Turkic corpora, researchers have proposed language-specific tagsets derived from the UD standard, including features such as case, number, possession, and evidentiality. Manual validation remains essential, especially when automatic systems produce ambiguous or inconsistent tags.

4. Applications in Linguistic and Computational Studies

4.1. Lexical and Grammatical Analysis

POS-tagged corporation enables frequency-based studies of lexical categories and grammatical structures. For instance, the relative frequency of verbs versus nouns may indicate stylistic differences between literary and academic genres. In Uzbek corpus research, POS tagging facilitates the analysis of verb morphology and aspectual variation, while in Kazakh studies, it supports investigations of case marking and syntactic alignment.

4.2. Information Retrieval and Text Mining

In information retrieval, POS tagging enhances query expansion and semantic search. Tagging allows search engines to differentiate between homonyms (e.g., *run* as a verb vs. run as a noun), improving retrieval precision. In text mining, tagged data support keyword extraction, collocation analysis, and thematic clustering.

4.3. Machine Translation and Multilingual NLP

POS tagging is a prerequisite for effective machine translation (MT) systems. Tagging assists in disambiguating syntactic structures, aligning grammatical patterns across languages, and optimizing translation models. In multilingual frameworks like Google’s mBERT or OpenAI’s GPT, POS-tagged corpora serve as training material for universal linguistic representations.

4.4. Language Teaching and Lexicography

In applied linguistics, POS-tagged corpora contribute to data-driven language teaching (DDL) and computational lexicography. Teachers and students can examine

authentic examples of word usage by part of speech, while lexicographers can automatically extract lemma lists and grammatical patterns for dictionary compilation.

5. Challenges for Morphologically Rich Languages

Agglutinative languages pose distinct challenges to POS tagging due to their complex word formation and inflectional variability. A single word may encode multiple grammatical meanings (person, number, tense, mood, case), leading to high data sparsity.

In Kazakh and Uzbek, for example, suffix stacking results in numerous word forms unseen in the training corpus. To address this issue, researchers employ morphological analyzers and subword tokenization (e.g., byte-pair encoding) to capture internal structure. Incorporating character-level embeddings has proven particularly effective for representing rich morphology.

6. Evaluation and Error Analysis

Evaluation results reported in previous studies demonstrate that tagging errors typically arise from:

1. Ambiguous word forms (e.g., nouns that resemble verbs),
2. Unknown or foreign words,
3. Incorrect handling of derivational morphemes,
4. Lack of contextual information in short sentences.

Error analysis thus plays a crucial role in improving tagger performance. Systematic examination of misclassified tokens informs both algorithmic refinement and annotation guideline revision.

7. Future Directions

Emerging trends point toward universal multilingual models trained on cross-lingual datasets, as well as semi-supervised and unsupervised tagging techniques. These approaches reduce the dependency on large manually annotated corpora, enabling linguistic analysis for under-represented languages.

Furthermore, the integration of POS tagging into deep syntactic and semantic frameworks—such as dependency parsing and semantic role labeling extends its analytical power. In corpus linguistics, automated tagging combined with visualization tools (e.g.,

concordancers and syntactic trees) promotes new insights into grammatical and stylistic phenomena.

Conclusion

This study examined part-of-speech tagging as both a computational process and a linguistic instrument. The findings highlight that POS tagging forms the backbone of corpus annotation, enabling detailed grammatical and lexical analysis. Over the decades, tagging methodology has evolved from rule-based grammars to probabilistic and neural architecture, reflecting a broader shift toward data-driven approaches in linguistics.

In the context of corpus linguistics, POS tagging supports empirical exploration of grammatical regularities and contributes to the creation of reusable language resources. For morphologically complex languages like Uzbek and Kazakh, continued efforts to develop annotated corpora and hybrid tagging models are crucial for advancing both computational and theoretical studies.

Main conclusions:

1. POS tagging enables systematic grammatical analysis across languages and domains.
2. Neural and hybrid models currently provide the highest tagging accuracy.
3. Morphological complexity in Turkic languages requires specialized tagsets and analyzers.
4. Corpus design and annotation quality directly influence research validity.
5. Future progress depends on multilingual modeling, open linguistic data, and cross-disciplinary collaboration.

Ultimately, part-of-speech tagging exemplifies how corpus linguistics integrates linguistic theory with computational methodology, fostering a more scientific understanding of natural language.

References

1. Brill E. A simple rule-based part of speech tagger // Proceedings of the Third Conference on Applied Natural Language Processing. – Trento, 1992. – P. 152-155.

2. Charniak E. Statistical Techniques for Natural Language Parsing. – Cambridge: MIT Press, 1997. – 318 p.

3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL. – 2019. – P. 4171-4186.

4. Greene B.B., Rubin G.M. Automatic grammatical tagging of English // Technical Report, Brown University. – 1971. – 48 p.

5. Kadirova N. Neural Approaches to Morphological Analysis for the Uzbek Language // Computational Linguistics and Intelligent Systems. – 2021. – Vol. 5. – P. 121-132.

6. Leech G., Wilson A. Standards for Tagging and Annotation // In: Ide N., Véronis J. (Eds.), Text Encoding Initiative Guidelines. – Oxford University Press, 1999. – P. 105-118.

7. Mukushev M., Nurkasymova S., Yessenbayev Z. Developing a Morphological Analyzer for the Kazakh Language // Bulletin of KazNU. – 2020. – Vol. 179, No. 1. – P. 55-63.

8. Nivre J. et al. Universal Dependencies v2: An ever-growing multilingual treebank collection // Proceedings of LREC. – 2018. – P. 1863-1871.

9. Schütze H. Part-of-Speech Tagging // In: Jurafsky D., Martin J.H. Speech and Language Processing. – 2nd ed. – Prentice Hall, 2009. – P. 181-210.

10. Turchin A.V. Machine Translation and Morphological Tagging of Turkic Languages. – Almaty: KazNU Press, 2022. – 142 p.

11. Zubanova T.A. Corpus Linguistics: Methods and Technologies. – Moscow: INFRA-M, 2018. – 228 p.