PoS to UPoS Conversion and Creation of UPoS Tagged Resources for Assamese Language

Kuwali Talukdar and Shikhar Kumar Sarma
Department of Information Technology, Gauhati University, India
kuwalitalukdar@gmail.com, sks001@gmail.com

Abstract

This paper addresses the vital task of transitioning from traditional Part-of-Speech (PoS) tagging to Universal Part-of-Speech (UPoS) tagging within the linguistic context of the Assamese language. The paper outlines a comprehensive methodology for PoS to UPoS conversion and the creation of UPoS tagged resources, bridging the gap between localized linguistic analysis and universal standards. The significance of this work lies in its potential to enhance natural language processing and understanding for the Assamese language, contributing to broader multilingual applications. The paper details the data preparation and creation processes, annotation methods, and evaluation techniques, shedding light on the challenges and opportunities presented in the pursuit of linguistic universality. The contents of this research have implications for improving language technology in the Assamese language and can serve as a model for similar work in other regional languages. Mapping of standard PoS tagset applicable for Indian languages to that of the primary categories of the UPoS tagset is done with respect to the Assamese language lexical behaviour. Conversion of PoS tagged text corpus to UPoS taged corpus using this mapping, and then utilizing a Deep Learning based model trained on such a dataset to create a sizable UPoS tagged corpus, are presented in a structured flow. This paper is a step towards a more standardized, universal understanding of linguistic elements in a diverse and multilingual world.

1 Introduction

Assamese language, a rich and diverse member of the Indian linguistic landscape, has been a subject of interest in recent years in the field of natural language processing (NLP) and linguistic analysis. An integral component of language analysis is Part-of-Speech (PoS) tagging, a fundamental linguistic process that classifies words into grammatical categories. In the context of Assamese, the PoS tagging has predominantly relied on the BIS (Bureau of Indian Standards) PoS tagset (Bureau of Indian Standards.(2021) for Indian languages, which has served as a valuable framework for linguistic analysis.

The BIS PoS tagset for Indian languages, designed to be applicable across numerous Indian languages, has played a pivotal role in linguistic research and language technology development for Assamese. It has provided a structured foundation for categorizing and understanding the grammatical elements of this language. However, as language technology continues to evolve, there is an increasing need to align with universal standards such as the Universal PoS (UPoS) tagging system (Marie-Catherine et al., 2021). UPoS tagging facilitates cross-linguistic comparisons and fosters interoperability in NLP applications. This paper embarks on the task of transitioning from the conventional BIS PoS tagset to UPoS tagging for the Assamese language. This transition not only reflects the evolving landscape of linguistic analysis but also opens doors to more comprehensive and standardized linguistic resources. It aims to harmonize Assamese language analysis with

broader linguistic universality while respecting the unique intricacies and characteristics that define Assamese. In doing so, it addresses the need for a PoS tagging system that balances local relevance and global linguistic standards. The research presented here not only contributes to the advancement of language technology for Assamese but also serves as a model for similar efforts in other regional languages.

Part-of-Speech tagging, also known as PoS tagging, involves the assignment of grammatical categories (e.g., nouns, verbs, adjectives) to individual words in a text. This process is crucial in understanding the structure and meaning of a sentence. Accurate PoS tagging is the cornerstone of many NLP applications and linguistic analysis. It enables computers to comprehend the syntactic and semantic properties of a language. The accurate tagging of parts of speech (PoS) is a fundamental aspect of natural language processing (NLP) and linguistic analysis. It forms the basis for various language technology applications, including machine translation, sentiment analysis, and information retrieval etc. While PoS tagging has been extensively developed and standardized for many major languages, languages with fewer resources, such as Assamese, face challenges in this regard.

2 Challenges in PoS Tagging for Low Resource Language-Assamese

Assamese is a language spoken in the Indian state of Assam and neighboring regions. It boasts a rich linguistic heritage with its own set of grammatical rules, dialects, and nuances. However, the development of linguistic resources for Assamese, such as standardized PoS tagging, has been relatively limited compared to more widely spoken languages. One of the primary challenges in PoS tagging for Assamese is the absence of a standardized and widely accepted PoS tagset. While major languages like English and Hindi have established PoS tagsets, Assamese lacks a universally recognized set of PoS tags. The BIS superset, designed to be a common primary tagset across the Indian languages, is yet to be customized for Assamese, and this has forced the NLP researcher to confine all works limited to the BIS superset for Indian languages without having the Assamese linguistic features embedded. This poses a hindrance to linguistic research, language technology development, and the interoperability of NLP tools for Assamese. While PoS tagging has seen extensive development and standardization for major languages like English, challenges persist for languages with fewer linguistic resources. Low-resourced languages, such as Assamese, often face different difficulties:

i. Lack of Standardization: Many low-resourced languages lack standardized PoS tagsets, making it challenging to develop linguistic resources and NLP tools.
ii. Resource Scarcity: There is a scarcity of annotated data, including PoS tagged corpora, which are essential for training PoS taggers.
iii. Unique Linguistic Features: Low-resourced languages often have unique linguistic features and dialectical variations, complicating the development of a one-size-fits-all tagging system.

3 Quest for Universality - UPoS Tagging

To address the challenges posed by the absence of a standardized and inclusive PoS tagset for Assamese, this research embarks on the journey of transitioning from traditional PoS tagging to Universal Part-of-Speech (UPoS) tagging. UPoS tagging follows universal annotation standards that aim to provide a common framework for tagging parts of speech across languages. The transition to UPoS tagging is driven by the goal of achieving linguistic universality while retaining the unique linguistic characteristics of Assamese. The current endeavor seeks to develop UPoS tagged resources that are compatible with global linguistic standards, making Assamese more accessible for linguistic analysis, cross-linguistic research, and the development of language technology applications. This research paper aims to achieve the following objectives:

i. PoS to UPoS Conversion: Develop the mapping of PoS tagset to the UPoS tagset, and conversion of existing PoS tagged data in Assamese to UPoS tagging, aligning the language with universal standards.
ii. Creation of UPoS Tagged Resources: Create UPoS tagged linguistic resources, a structured multidomain text corpora annotated with Assamese UPoS tagset, to facilitate linguistic analysis and language technology development.

iii. Evaluation of Resources: Conduct a rigorous evaluation of the UPoS tagged resources to assess their accuracy, effectiveness, and reliability.
iv. Implications and Future Directions: Explore the broader implications of the transition to UPoS tagging for Assamese and outline potential future directions for research and development.
This research endeavors to bridge the gap in linguistic resources for Assamese and contribute to the broader landscape of linguistic analysis and language technology.

4 Background and Related Works

Part-of-Speech tagging, often abbreviated as PoS tagging, is a fundamental task in natural language processing (NLP). It involves the annotation of words in a text with their corresponding grammatical categories, such as nouns, verbs, adjectives, and adverbs. PoS tagging serves as a critical component of various language technology applications and linguistic analysis. By assigning grammatical labels to words, computers can discern the syntactic and semantic structure of sentences, enabling them to understand language more effectively. In the context of PoS tagging, a tagset is a predefined collection of PoS tags, each representing a specific grammatical category. Different languages often have their own unique tagsets, which can make it challenging to develop unified NLP tools and linguistic resources.

Accurate PoS tagging is essential for several reasons:
i. Language Understanding: PoS tags help in disambiguating word meanings. For example, the word "bank" can refer to a financial institution or the notion of dependency, depending upon on what sense the word is used in the sentence. By assigning different tags (e.g., "NN" for noun and "VB" for verb), the meaning becomes clear in the context.
ii. Syntactic Parsing: PoS tagging aids in syntactic parsing, which is crucial for understanding the grammatical structure of sentences. This, in turn, supports various NLP tasks, including machine translation and information retrieval.
iii. Semantic Analysis: By identifying the parts of speech, it becomes easier to perform semantic analysis, extracting information about the roles words play in a sentence. This is invaluable for tasks like sentiment analysis and question-answering systems.
Historically, the Brown Corpus tagset (Francis et al., 1979), the Penn Treebank PoS Tagset (Mitchell et al., 1993), and the IBM Watson project simplified PoS tagsets (https://www.ibm.com) were the fundamental standardizations in the English language tagset evolution. The British National Library BNC Corpus thereafter, was a huge collection of texts tagged with a rather elaborate larger PoS tagset (http://www.natcorp.ox.ac.uk). Indian language PoS tagset evolves with multiple efforts, including Indian Language Machine Translation (ILMT) project, IIIT Hyderabad, and CIIL Mysore initiatives. A convergence has been seen lateron with unification and standardizing effort by the Technology Development in Indian Languages (TDIL) program of Ministry of Communication and Information Technology, Govt. of India (TDIL). The evolution of Indian Languages PoS tagset has got stability with a published standard by the Bureau of Indian Standards (BIS) through an inclusive and structured methodology (Bureau of Indian Standards.(2021)). This standard PoS tagset consists of annotations for 11 Core Categories, and 45 subcategories, applicable as a superset for all Indian languages, which is added with language specific subset for few major languages. BIS tagset has taken the dominating role in Indian Languages PoS tagging endeavors in recent years, and multiple experiments for almost all Indian Languages could be traced. (Kabir et al., 2016); (Alam et al., 2016); (Pooja et al., 2023); (Ovi et al., 2022); (Anbukkarasi and Varadhaganapathy, 2021); (Sunita et al., 2022); (Das et al., 2023).

Among the major Indian Languages, North Eastern languages are relatively new for NLP research works. In recent years, efforts have been intensified by different researchers and groups. Preliminary resources have been created, including corpus, Wordnets etc. (Shikhar et al., 2012); (Anup et al., 2014); (Jumi et al., 2016); (Bhuyan and Sarma, 2018). Experimentations for various NLP tasks have been undertaken including PoS tagging (Barman et al., 2013); (Pathak et al., 2022); (Kuwali Talukdar et al., 2023) with conventional as well as Machine

Learning Techniques. More complex modelling like Machine Translation, (Baruah et al., 2014) Information Retrieval, Neural Machine Translation (Ahmed et al., 2023) also have been observed with limited scopes and performances.

5 Methodology for Conversion

Methodology employed for the conversion of Part-of-Speech (PoS) tagging to Universal Part-of-Speech (UPoS) tagging in the context of the Assamese language is outlined here. It serves as the foundation for the subsequent chapters that describe the creation and evaluation of UPoS tagged resources.

5.1 Data Collection

PoS to UPoS conversion process heavily relies on the quality and representativeness of the linguistic data used. In this section, we outline our data collection efforts and preprocessing steps to ensure the suitability of the dataset. For our research, we considered a diverse range of linguistic data in Assamese, including texts from various genres, sources, and timelines. The selection of such a comprehensive dataset was critical to ensure that the converted UPoS tagging system would be robust and applicable to a wide array of linguistic contexts. We have taken the GUIT Assamese corpus as the source for creating the final UPoS tagged resource, which is a structured and inclusive corpus of Assamese text, with a timeline of 100 years of evolutionary Assamese Text, and covers a wider range of domains. The corpus also is representative in nature, as the sources are from a variety of written forms of publications and documents. For the reference data NPLT Assamese dataset with BIS tags have been considered. In addition to gathering textual data, we undertook preprocessing steps to eliminate noise and enhance data quality. These steps involved tasks such as sentence segmentation, filtering unwanted segments, removing extra long and foreign segments, and tokenization. By doing so, we aimed to create a clean and reliable dataset for the subsequent PoS to UPoS conversion process.

5.2 Annotation Process

The conversion from traditional PoS tagging to Universal PoS (UPoS) tagging in Assamese demanded a meticulous and precise annotation process. Annotation process involved DL based techniques, with manual validation. Linguistic experts were involved in manually validating a subset of the dataset, creating a gold standard for the conversion process. This manual validation was complemented by automated tools, which aided in the conversion of PoS tags to UPoS tags at a larger scale.

5.3 UPoS Tagset for Assamese

The Universal Dependencies (UD) tagset, which has gained popularity for its universality and compatibility with a wide range of languages has been considered. UD tagset offers the advantage of cross-linguistic comparison and compatibility with global standards, making it an ideal choice for our UPoS conversion. The UD tagset, though universal in nature, was adapted to suit the specific linguistic characteristics of Assamese. This involved creating language-specific mapping to ensure a seamless transition from the existing BIS PoS tagging system to the UPoS tags defined by the UD tagset. The UPoS annotations against 17 core lexical categories as prescribed in the universal dependencies are shown in Table 1.

5.4 Software and Tools

The conversion of PoS tagging to UPoS tagging required the development of custom software and tools to streamline the annotation process. A specialized annotation tool has been developed as part of the work, that allowed linguists to view and edit annotations. The tool has provisions for both annotations and validations. Additionally, we employed natural language processing tools to automate parts of the conversion process. These Python based tools were programmed to recognize common PoS patterns and map them to the corresponding UPoS tags, reducing the manual workload and improving efficiency.

6 Creation of UPoS Tagged Resources

Here, the details of the process of creating Universal Part-of-Speech (UPoS) tagged resources for the Assamese language is presented. Building upon the methodology outlined in the previous chapter, the focus now shifts to converting existing Part-of-Speech (PoS) tagged data into UPoS format and leveraging this transformed data to develop invaluable linguistic resources.

Open class words		Closed class words		Other	
Tags	**PoS Categories**	**Tags**	**PoS Categories**	**Tags**	**PoS Categories**
ADJ	adjective	ADP	adposition	PUNCT	punctuation
ADV	adverb	AUX	auxiliary	SYM	symbol
INTJ	interjection	CCONJ	coordinating conjunction	X	other
NOUN	noun	DET	determiner		
PROPN	proper noun	NUM	numeral		
VERB	verb	PART	particle		
		PRON	pronoun		
		SCONJ	subordinating conjunction		

Table 1. UPoS core categories with annotations

1

6.1 Data Conversion

The conversion process is at the heart of our resource creation endeavor. To make Assamese linguistics compatible with universal standards, we meticulously undertaken transition from traditional PoS tagging to UPoS tagging. This involves the conversion of the existing PoS tagged data, which serves as the linguistic foundation for the Assamese language. The data conversion process is a meticulous task that calls for careful consideration of the nuances within Assamese. It requires mapping PoS tags to their corresponding UPoS tags, ensuring that linguistic elements are accurately represented. The mapping of BIS tagset to the 17 UPoS core categories is shown in Table 2.

A customized Python module was run over the BIS tagged initial 30000 sentences, with the mapping of BIS-UPoS as the implementing target. The process pipeline involves look-up and replace, filtering and then validation. The system stages are shown in the Figure 1. The impact of this data conversion is profound. By creating a comprehensive UPoS tagged preliminary dataset, we lay the groundwork for more standardized linguistic analysis, and this standardized preliminary data is then used to power a GRU based deep learning model, which in turn, has been used to run on the gold standard 40000 GUIT Assamese corpus to create the UPoS tagged Assamese Resource.

6.2 Development of Tagged Data

The converted UPoS tagged preliminary data, enriched in linguistic information as per UPoS defined lexical information, becomes the cornerstone for developing the targeted linguistic resources. This base corpus, comprising a collection of 22340 sequences of UPoS tagged texts, are now instrumental in training the machine learning model. The baseline model from a previous experiment in the same lab, with the model hyperparameters as defined in Table 3, has been used to train with the 22340 sequences. This is a GRU based deep learning model, and the system used the data divisions for training, validation, and testing as detailed in Table 4. Table 5 presents the GRU Model performances showing Accuracy, Precision, Recall, and F1 score. The model accuracy of 94.38% with F1 score of 94.56 are considered optimum performed model as this is the best model. The details of the model training are not in the scope of the current resource creation work, and has been incorporated from a related work in the same lab by the same group. This GRU based DL model is now used to tag the GUIT Assamese corpus for all the 40000 untagged sequences. This is thereafter subjected to linguistic evaluation and validation process. Linguistic evaluation and validation were counted for 20% of the total sequences amounting to 8000. The sequences are of diverse sequence length, and these were picked up from the raw corpus through an automated Python based script so that the multiple domains equi-distribution as well as multiple-sequence-length inclusion could be achieved. The domainwise chunks and percentage distribution is shown in Table 6. Table 7 presents the corpus statistics of sequences based on Token-Length. A python script has been configured to filter necessary sequences of length range 5 to 35 tokens.

Core UPoS	BIS Categories
1. Noun (NOUN)	**Noun (N)** Common Noun (NN) Verbal Noun (NNV) Nloc (NST)
2. Proper noun (PROPN)	**Proper Noun (NNP)**
3. Verb (VERB)	**Verb (V)** Verb Main (VM) Verb finite (VF) Verb non-finite (VNF) Verb infinite (VINF) Gerund (VNG) Verbal (VN)
4. Pronoun (PRON)	**Pronoun (PR)** Personal Pronouns (PRP) Reflexive Pronoun(PRF) Relative Pronoun (PRL) Reciprocal Pronoun (PRC) Interrogative Pronoun/WH-word (PRQ) Indefinite Pronoun (PRI)
5. Adjective (ADJ)	**Adjective (JJ)**
6. Adverb (ADV)	**Adverb (RB)**
7. Adposition (ADP)	**Postposition (PSP)**
8. Auxiliary Verb (AUX)	**Auxiliary Verb(VAUX)** Finite (VF) Non finite (VNF) Infinitive (VINF) Participle Noun (VNP)
9. Coordinating conjunction (CCONJ)	**Conjunction (CC)** Coordinator (CCD)
10. Subordinating conjunction (SCONJ)	Subordinator (CCS)
11. Interjection (INTJ)	Interjection (INJ)
12. Determiner (DET)	**Demonstrative (DM)** Deictic (DMD) Relative (DMR) Wh-word (DMQ) Indefinite (DMI)
13. Numeral (NUM)	**Quantifiers (QT)** General (QTF) Cardinal (QTC) Ordinal (QTO)
14. Particle (PART)	**Particle (RP)** Particle Default (RPD) Classifier (CL) Intensifier (INTF) Negation (NEG)
15. Punctuation (PUNCT)	Punctuation (PUNC)
16. Symbol (SYM)	Symbol (SYM)
17. Others(X)	Residuals (RD) Foreign word (RDF) Unknown (UNK) Echowords (ECH)

Table 2. Mapping of BIS PoS to UPoS

7 Evaluation of UPoS Tagged Resources

The effectiveness and accuracy of these resources are essential to their utility in linguistic analysis, research, and natural language processing. To assess the quality and accuracy of the UPoS tagged resources, we employ error analysis through linguistic validation. Two parallel linguistic validation were performed on the 20% randomized chunk of data, and errors detected as well as corrections suggested have been recorded. The linguistic validation with expert linguists has shown almost nearby figures to the DL based model accuracy, substantiating the accuracy of trained model's prediction behaviour. While the GRU model's accuracy was 94.38%, the errors detected by the linguists have been recorded as a 6.24% and 7.08% respectively for Validator 1 and Validator 2.

Measuring inter-annotator agreement is an essential component of our evaluation process. This metric assesses the level of agreement among linguistic experts in their efforts to validate UPoS tagged data. Here, the methodology for measuring inter-annotator agreement involves having two linguistic experts independently check the same dataset. The results of this study provide valuable insights into the consistency and reliability of the UPoS tagging system. Inter-annotator agreement studies reveal the extent to which experts agree on UPoS tags and identify areas of disagreement. These findings help us understand the system's robustness and identify potential areas for improvement. Inter annotator (validator) agreement has been excellent with 0.84% initial disagreement, and then reduces to nil through discussions and mutual agreement. Central to this linguistic evaluation process was the concept of a "gold standard" Assamese UPoS tagged data resource, which has been achieved. This gold standard dataset serves as a benchmark for further NLP research, particularly in training DL models with higher performance of accuracy, so that high quality tagged corpus could be created without having requirements of linguistic validation. High accuracy tagging capability of models are critical in various NLP applications where PoS tagging is integrated as a critical part in the NLP pipelines, like machine translation, sentiment analysis etc.

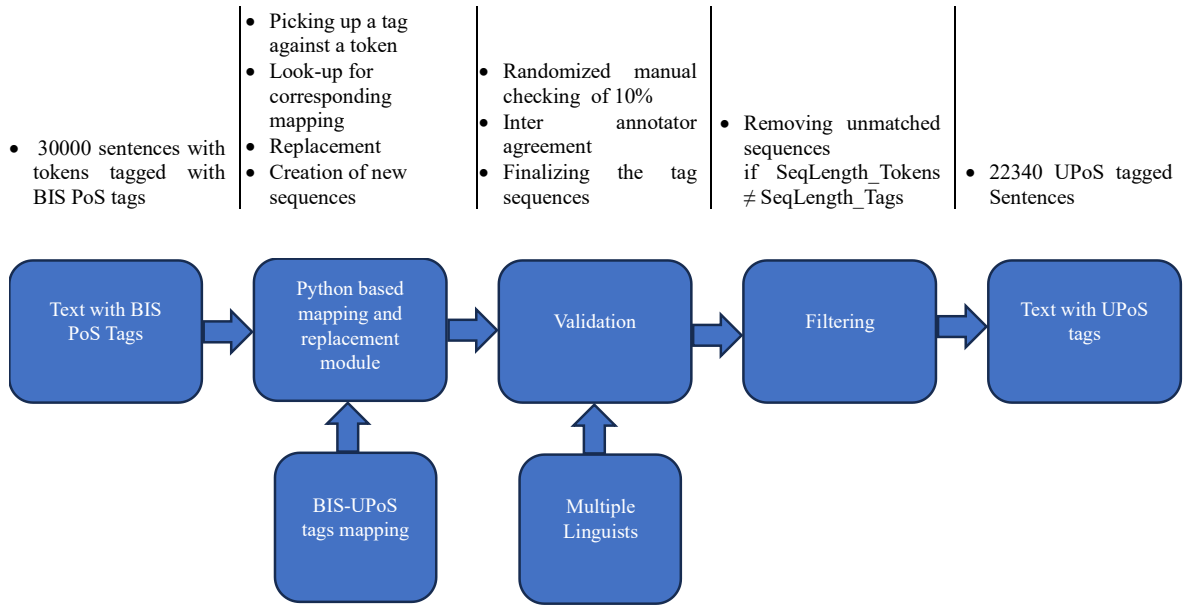

Figure 1. Python based conversion system process flow

Models	Embedding Layer	Model Layer	Dense Layer
GRU	Input dim = vocab size, Output dim = 300, Input length = 100	GRU Layer = 64 Cell	36 no. of classes

Table 3. Model Hyperparameters

Models	Training	Validation	Testing
GRU	17980 sequences 80.5%	2238 sequences 10%	2122 sequences 9.5%

Table 4. Dataset divisions for system level Training, Validation, and Testing

Models	Accuracy	Precision	Recall	F1
GRU (UPoS)	94.38%	95.44	93.70	**94.56**

Table 5. Accuracy, Recall, Precision and F1 score of the trained model

Domain	Frequency	% Frequency
Media	9400	23%
Learned Materials	8400	21%
Literature	8000	20%
Tourism	8000	20%
Science and Technology	6200	16%

Table 6. Domain wise distribution

Token Length	Frequency	% Frequency
5 to 10	11542	29
11 to 20	18345	46
21 to 30	8076	20
30 to 35	2037	5

Table 7. Frequency distribution

8 Contribution and Resource Accessibility

The transition from PoS to UPoS tagging in Assamese and the creation of UPoS tagged resources are pivotal steps in advancing linguistic analysis and language technology. This work contributes to the ongoing journey of language standardization and holds the promise of broader applications in multilingual systems and linguistic research.

By aligning Assamese with UPoS standards, the language becomes more adaptable to multilingual systems, fostering interoperability and collaboration. As we reflect on the accomplishments of this research, we are reminded that the journey towards language standardization is ongoing. The creation of UPoS tagged resources marks a significant milestone, but there are further horizons to explore. Key findings from this research include-successful conversion from PoS to UPoS tagging, aligning Assamese with universal linguistic standards, and creation of UPoS tagged resources. Another contribution of the current work has been the rigorous evaluation measures, combining quantitative and qualitative analyses, to assess the accuracy and reliability of the UPoS tagging system, resulting the resource trusted as high quality.

For the benefit of the research community and language technology developers, it is paramount that these UPoS tagged resources are made widely accessible. Resource accessibility is the cornerstone of collaborative research and development. By making these linguistic resources openly available, we foster a community of researchers and developers who can leverage them for various applications. The UPoS tagged Assamese text corpora is now sharable, and available for the research community. The open nature of these resources ensures transparency, fosters innovation, and paves the way for advancements in language technology. The resources created in this endeavor unlock opportunities for Assamese to participate in the global linguistic landscape and benefit from cross-linguistic analysis and interoperability.

9 Conclusion

The research embarked on the task of transitioning from traditional Part-of-Speech (PoS) tagging to Universal Part-of-Speech (UPoS) tagging in Assamese, with the primary aim of achieving linguistic universality while retaining the unique linguistic characteristics of the language. It presents a tangible outcome in the form of UPoS tagged resources, which are invaluable assets for linguistic analysis, research, and the development of language technology. The adoption of UPoS tagging in Assamese signifies a leap forward in linguistic analysis. It aligns Assamese with global linguistic standards, making cross-linguistic comparisons and research more accessible. The UPoS tagged resources, created as a result of this conversion, enhance the quality and depth of linguistic analysis in Assamese. Researchers and linguists can now explore Assamese with a newfound level of universality and precision. This advancement opens doors to deeper insights into the language's structure, grammar, and semantics, and also on the study of linguistic variations across different languages. The impact of UPoS tagging is not limited to linguistic analysis; it extends to the realm of language technology. With UPoS tagged resources at their disposal, language technology developers can design more sophisticated and accurate natural language processing applications for Assamese. Machine translation, sentiment analysis, information retrieval, and various other language-based technologies can benefit from the standardized linguistic data. As we move forward, the path is illuminated by the understanding that linguistic diversity and universality can coexist, enriching our global linguistic tapestry. Future research and development in this field are poised to unlock deeper insights into Assamese, enhance language technology, and contribute to the wider world of linguistic analysis. This research stands as a testament to the importance of maintaining a balance between linguistic universality and language-specific relevance. The journey towards linguistic universality is ongoing. The creation of UPoS tagged resources for Assamese represents a significant step, but it is not the final destination. The road ahead holds promise and beckons for further exploration.

References

Bureau of Indian Standards.(2021) “Linguistic Resources-POS Tag Set for Indian Languages-Guidelines for Designing Tagsets and Specification.” www.bis.gov.in, www.standardsbis.in

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. Computational Linguistics, 47(2):255–308.

Francis, W. N. and Kucera, H. Brown Corpus Manual. , Department of Linguistics, Brown University, Providence, Rhode Island, US (1979).

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the penn treebank. Comput. Linguist. 19, 2 (June 1993), 313–330.

https://www.ibm.com/docs/en/wca/3.5.0?topic=analytics-part-speech-tag-sets

http://www.natcorp.ox.ac.uk/docs/gramtag.html

TDIL (Technology Development Indian Languages) programme, Ministry of Communication and Information Technology, Govt. of India, Unified Parts of Speech (POS) Standard in Indian Languages, https://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf

A.K. Barman, J. Sarmah and S. K. Sarma, "POS Tagging of Assamese Language and Performance Analysis of CRF++ and fnTBL Approaches," 2013 UKSim 15th International Conference on Computer Modelling and Simulation, Cambridge, UK, 2013, pp. 476-479, doi: 10.1109/UKSim.2013.91.

Pathak, Dhrubajyoti, et al. “AsPOS: Assamese Part of Speech Tagger Using Deep Learning Approach.” 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Dec. 2022. Crossref, https://doi.org/10.1109/aiccsa56895.2022.1001793

Kuwali Talukdar and Shikhar Kumar Sarma, "Parts of Speech Taggers for Indo Aryan Languages: A critical Review of Approaches and Performances," 2023 4th International Conference on Computing and Communication Systems (I3CS), Shillong, India, 2023, pp. 1-6, doi: 10.1109/I3CS58314.2023.10127336.

Shikhar Kr. Sarma, Himadri Bharali, Ambeswar Gogoi, Ratul Deka, and Anup Kr. Barman. 2012. A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges. In Proceedings of the 10th Workshop on Asian Language Resources, pages 21–28, Mumbai, India. The COLING 2012 Organizing Committee.

Anup Barman, Jumi Sarmah, and Shikhar Sarma. 2014. Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System. In Proceedings of the Seventh Global Wordnet Conference, pages 256–261, Tartu, Estonia. University of Tartu Press.

Baruah, Kalyanee & Das, Pranjal & Hannan, Abdul & Sarma, Shikhar. (2014). Assamese-English Bilingual Machine Translation. International Journal on Natural Language Computing. 3. 10.5121/ijnlc.2014.3307.

Jumi Sarmah, Shikhar Kumar Sarma, "Survey on Word Sense Disambiguation: An Initiative towards an Indo-Aryan Language", International Journal of Engineering and Manufacturing(IJEM), Vol.6, No.3, pp.37-52, 2016.DOI: 10.5815/ijem.2016.03.04

M. P. Bhuyan and S. K. Sarma, "Automatic Formation, Termination & Correction of Assamese word using Predictive & Syntactic NLP," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 544-548, doi: 10.1109/CESYS.2018.8724023.

M. A. Ahmed, K. Kashyap and S. K. Sarma, "Pre-processing and Resource Modelling for English-Assamese NMT System," 2023 4th International Conference on Computing and Communication Systems (I3CS), Shillong, India, 2023, pp. 1-6, doi: 10.1109/I3CS58314.2023.10127567.

Kabir, M.F., Abdullah-Al-Mamun, K., & Huda, M.N. (2016). Deep learning based parts of speech tagger for Bengali. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 26-29.

Alam, F., Chowdhury, S. A., & Noori, S. R. H. (2016). Bidirectional LSTMs — CRFs networks for bangla POS tagging. 2016 19th International Conference on Computer and Information Technology (ICCIT). https://doi.org/10.1109/ICCITECHN.2016.7860227

Pooja Rai, Sanjay Chatterji, and Byung-Gyu Kim. 2023. Deep Learning-based Sequence Labeling Tools for Nepali. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22, 8, Article 211 (August 2023), 23 pages. https://doi.org/10.1145/3606696

J. A. Ovi, M. A. Islam and M. R. Karim, "BaNeP: An End-to-End Neural Network Based Model for Bangla Parts-of-Speech Tagging," in IEEE Access, vol. 10, pp. 102753-102769, 2022, doi: 10.1109/ACCESS.2022.3208269.

S. Anbukkarasi and S. Varadhaganapathy, Deep Learning based Tamil Parts of Speech (POS) Tagger,Bulletin of the Polish Academy of Sciences Technical Sciences, Vol. 69(6), 2021, Article number: e138820 DOI: 10.24425/bpasts.2021.138820

Sunita Warjri, Partha Pakray, Saralin A. Lyngdoh, and Arnab Kumar Maji. 2021. Part-of-Speech (POS) Tagging Using Deep Learning-Based Approaches on the Designed Khasi POS Corpus. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21, 3, Article 63 (May 2022), 24 pages. https://doi.org/10.1145/3488381

Das, A., Choudhury, B., Sarma, S.K. (2023). POS Tagging for the Primitive Languages of the World and Introducing a New Set of Universal POS Tagging for Sanskrit. In: Fong, S., Dey, N., Joshi, A. (eds) ICT Analysis and Applications. Lecture Notes in Networks and Systems, vol 517. Springer, Singapore. https://doi.org/10.1007/978-981-19-5224-1_3