Nepalese Linguistics, Journal of Linguistic Society of Nepal; ISSN: 0259-1006(Print)

TAGGING ANGIKA CORPUS USING BIS SCHEME: A PRELIMINARY STUDY

JYOTI KUMARI
Department of Linguistics, Banaras Hindu University.
jyoti@bhu.ac.in

(Received: 03 July 2025; revised: 07 Oct., 2025; accepted: 5 Nov., 2025; published: 26 Nov., 2025)

Angika is an Eastern Indo-Aryan language spoken mainly in the southeastern regions of Bihar, Jharkhand and in some areas of Nepal. Angika is a Low-resource language due to the absence of linguistic resources and NLP tools.. The primary challenge for developing NLP tools for the Angika language is the lack of corpora. In this context, the BIS POS Tagset for Indian languages has been adopted to facilitate Part-of-Speech (POS) tagging for Angika. Part-of-Speech (POS) tagging is a fundamental task in Natural Language Processing (NLP) that involves assigning grammatical categories, such as nouns, verbs, adjectives, and adverbs, to words in a text. This article aims to explore the application of the BIS POS Tagset for Angika.

Keywords: Angika, NLP, POS Tagset, low-resource language

1. Introduction

Angika is an Eastern Indo-Aryan language spoken mainly in the southeastern regions of Bihar and Jharkhand. Angika is also spoken in some areas of Nepal. Historically rooted in the ancient Anga region, Angika is a language with a rich cultural heritage, deeply intertwined with the history and traditions of its speakers. In Bihar, Angika is spoken in the districts of Araria, Katihar, Purnia, Kishanganj, Madhepura, Saharsa, Supaul, Bhagalpur, Banka, Jamui, Munger, Lakhisarai, Begusarai, Sheikhpura and Khagaria. In Jharkhand, Angika is spoken in the districts of Sahebganj, Godda, Deoghar, Pakur, Dumka, Giridih and Jamtara. In West Bengal, Angika is spoken in the Malda and the Uttar Dinajpur. In Nepal, it is mainly spoken in Morang and Sunsari (Regmi, 2017).

Angika is a language with deep historical roots and it is deeply connected to the ancient Anga kingdom, one of the sixteen powerful Mahajanapadas in ancient India. This language shares similarities with nearby languages like Maithili, Magahi, Bengali, and Bhojpuri. It also reflects the shared cultural and linguistic heritage of the region. However, Angika has its own unique sound system, vocabulary, and grammar that make it distinct from these other languages.

In the past, Angika was written in a special script called 'Anga Lipi.' Over time, the Kaithi script became more common, but today, most people write Angika using the Devanagari script, which is also used for Hindi.

British linguist George Abraham Grierson once classified Angika as a dialect of Maithili in his Linguistic Survey of India (1903). However, today Angika speakers believe that their language is independent and is different from Maithili. In recent times, Angika has received some recognition. For example, it has been recognized as a "Second State Language" in the Indian state of Jharkhand since 2018. In Nepal, Angika was officially recognized as a separate language for the first time in the Nepal census 2001. In the 2021 Language Census of Nepal, Angika is listed as one of the mother tongues spoken in the country. However, Angika is not included in the 8th Schedule of the Indian Constitution, which means it does not receive the same level of recognition and support as some other Indian languages. Additionally, Angika is not widely taught in schools, and its use in formal education and the media is very limited.

Like many regional languages, Angika faces challenges such as the dominance of larger languages like Hindi and English, as well as the effects of urbanization and globalization. Despite

these challenges, there is a growing movement among Angika speakers to preserve their language and cultural heritage. Cultural organizations and academic researchers are increasingly interested in studying and documenting Angika. As a language with a rich oral tradition, Angika remains an important part of the cultural identity of its speakers. Studying and documenting Angika is crucial for understanding the linguistic diversity of the Indo-Aryan languages and ensuring that this historically significant language continues to thrive.

Developing NLP tools for Angika faces challenges due to the lack of annotated data, large corpora, and limited funding and interest. For low-resource languages like Angika, creating a POS tagged corpus is crucial as it forms the base for developing NLP tools. A POS tagset also helps document the language's structure, aiding in its preservation. Additionally, it supports linguistic research by allowing deeper studies of Angika's grammar and meaning. These efforts help promote Angika and encourage younger generations to continue using and preserving their language.

Until now, no significant effort has been made to annotate the Angika corpus, which is a crucial step in developing NLP tools and resources for the language. To address this, the BIS POS Tagset for Indian languages has been adopted, enabling effective Part-of-Speech (POS) tagging for Angika. The resulting POS Tagset for Angika consists of 31 tags.

The paper is divided into five sections. Section 2 gives a brief review of related work. Section 3 deals with the POS tagset for Angika. Section 4 deals with Tagging Angika Corpus Using BIS POS Tagset. Lastly, Section 5 concludes the article.

2. Review of literature

In this section, reviews of relevant research work related to POS tagging and Angika is discussed.

Bharati et al. (2006) provides guidelines for annotating linguistic data, specifically for Part-of-Speech (POS) tagging and chunking, in Indian languages. These guidelines were developed as part of the AnnCorra project at the International Institute of Information Technology, Hyderabad. POS tagging involves labeling each word in a sentence with its grammatical category, like noun, verb, or adjective. Chunking is the process of grouping words into phrases, such as noun phrases or verb phrases. The authors created a standardized set of rules and labels to ensure consistency in how different Indian languages are annotated. This standardization helps researchers and developers create accurate language processing tools, like POS taggers, that can work across multiple languages. The guidelines are designed to handle the unique features of Indian languages, such as their rich morphology and free word order. By following these guidelines, annotated corpora (large collections of text) can be created in a way that is consistent, making it easier to develop and compare language processing tools for different Indian languages.

Baskaran et al. (2006) explored the creation of a common POS tagset for Indian languages, which present challenges due to their diverse grammar and scripts. The authors proposed a standardized tagset adaptable to different languages, promoting cross-linguistic research and the development of NLP tools. The framework handles Indian languages' complex syntax and rich morphology, using both universal and language-specific tags. The paper details the methodology, including tag selection and hierarchical organization, and tests the framework on several languages, proving its flexibility. The authors highlight the benefits of improved tool interoperability and easier data integration, contributing to standardization in computational linguistics for Indian languages.

Antony and Soman (2011) explored the existing literature on Parts-of-Speech (POS) tagging methods for Indian languages. It reviews various approaches and techniques used in POS tagging, including rule-based, statistical, and hybrid methods. The authors discussed the challenges that are unique to Indian languages, such as their complex morphology and diverse linguistic features. They have also highlighted the progress made in the field and identified the gaps and the areas for future research. The paper also serves as

a comprehensive overview of how POS tagging has been approached for Indian languages, offering insights into the effectiveness of different methods.

Dalal et al. (2012) discusses the development of a feature-rich Part-of-Speech (POS) tagger specifically designed for Hindi which is a morphologically rich language. Morphologically rich languages have complex word structures that makes POS tagging very challenging. The authors have described their experience in building a POS tagger that incorporates various linguistic features such as morphology, context, and syntactic information. They have experimented with different machine learning models and feature combinations that improve tagging accuracy. The results showed that their approach significantly enhanced the performance of the POS tagger for Hindi, They have provided valuable insights for handling other morphologically rich languages.

Chandra, Kumawat, and Srivastava (2014) explored different tagsets used for Part-of-Speech (POS) tagging in Indian languages. A tagset is a collection of labels used to annotate words in a text with their grammatical categories, like nouns, verbs, or adjectives. The authors review several tagsets specifically designed for Indian languages, comparing their structures, complexities, and suitability for different languages. They also evaluate the performance of these tagsets in POS tagging tasks across various Indian languages, highlighting which tagsets are more effective and why. The article provides insights into how the choice of a tagset can impact the accuracy and efficiency of POS tagging in the diverse linguistic landscape of India.

Kumawa and Vinesh (2014) reviewed various tagging techniques, including rule-based systems, statistical models, and machine learning approaches. They have evaluated the performance of these methods based on factors such as accuracy, computational efficiency, and ease of implementation. The article provides a comparative analysis of these approaches, helping to identify which methods are most suitable for different contexts and languages. The findings offer guidance for selecting and improving POS tagging techniques.

Rathod and Govilkar (2015) provides a survey of various Part-of-Speech (POS) tagging techniques used for Indian regional languages. They have reviewed different approaches, including rule-based, statistical, and hybrid methods. They also highlighted the challenges specific to Indian languages, such as rich morphology, free word order, and the lack of large annotated corpora. They have concluded the article by discussing the effectiveness of different techniques and there is a need for further research and development to improve POS tagging accuracy for Indian languages.

Regmi (2017) presented a report, A sociolinguistic survey of Angika, focused on its usage and status in Nepal. The author has investigated how Angika is spoken and perceived by its speakers by examining the factors like language transmission, community attitudes, and the influence of other languages on Angika. This report also provides insights into the challenges faced by Angika in maintaining its linguistic identity in a multilingual context.

Tosha and Dwivedi (2017) focused on the relationship between Angika folksongs and the physical environment, arguing that both are experiencing a decline. The authors discuss how the degradation of the environment in regions where Angika is spoken has impacted the tradition of folk songs, which often reflect the natural surroundings. They also advocated for the preservation of both the cultural and environmental heritage of the Angika-speaking community.

Priyadarshi and Saha (2018) focused on the development of the first Part-of-Speech (POS) tagger for Maithili, a language spoken in Bihar. The authors detailed the process of creating linguistic resources necessary for this task, such as a POS-tagged corpus, which involved annotating a significant amount of Maithili text with grammatical categories. They have also described the system development for the POS tagger that includes the selection and implementation of machine learning techniques tailored to the specific characteristics of Maithili. The article also highlighted the challenges encountered due to the limited existing resources

and linguistic complexities of Maithili. The development of this POS tagger represents a significant step forward in natural language processing (NLP) for Maithili, facilitating further computational research and applications for this language.

Das, Sarmah, and Sharma (2019) focused on developing a Part-of-Speech (POS) tagger for the Khasi language. They have used the Hidden Markov Model (HMM). Since Khasi is a low-resource language with very limited digital resources, the authors created a POS tagger using statistical method, HMM,. They have trained the model on a small dataset and then tested its accuracy. The results showed that the HMM-based POS tagger worked well for Khasi. The authors also noted that more work is needed to improve its performance with larger datasets.

Suman, Jyoti and Sujeet (2023) explored the historical script used for writing in Angika, which is spoken in parts of Bihar, India. The authors have also traced the evolution of the Angika script, examining its origins, changes over time, and its significance in preserving the language's cultural heritage. The article also highlights the challenges faced in reviving and maintaining the traditional script in the modern digital age.

3. Development of a POS Tagset for Angika

Parts-of-Speech (POS) tagging is a fundamental task in Natural Language Processing (NLP). It involves assigning grammatical categories, such as nouns, verbs, adjectives, and adverbs, to words in a text. The Bureau of Indian Standards (BIS) POS Tagset is a standardized set of POS tags designed to annotate Indian languages, designed by Indian Language Corpora Initiative(ILCI). The BIS POS Tagset for Indian Languages has been adopted to facilitate POS tagging for the Angika Corpus. This POS Tagset for Angika consists of 31 tagsets. For each POS tagset relevant examples from Angika have been provided, ensuring that the tagset aligns with the linguistic characteristics of the language. The description of the tagset is given below.

3.1. Common Noun: A general name for a person, place, thing, or idea.

Examples: घोॅर ‘house’, आँख ’eye’, भोज ‘party’, हटिया ’market’ etc.

3.2. Proper Noun: A specific name for a person, place, or thing.

Example: सत्तन ‘sattan’, पल्टन ‘paltan’, पूरब ’east’, etc.

3.3. Nloc: Nouns that refer to locations or time.

Examples: आगु ‘front’, पिछू ‘back’, ऊपर ‘uppar’, नीचा ‘lower’, बीच ‘middle’, etc.

3.4. Personal Pronoun: A word that replaces a specific person or thing.

Examples: हम्मे ‘I’ , तोयँ ‘you’, ऊ ‘he/She’, etc.

3.5. Reflexive Pronoun: A word that reflects back to the subject.

Examples: आपनऽ ’you’, आपन्हैं ‘own’, स्वं ‘myself’, खुद्दे ‘myself’, etc.

3.6. Relative Pronoun: A word that relates to a noun previously mentioned.

Examples: जेऽ ‘who’, सेऽ ‘which’, etc.

3.7. Reciprocal Pronoun: A word showing a mutual relationship.

Examples: एक-दोसरो ‘each other ’, आपस ‘each other’, etc.

3.8. Wh-word Pronoun: A word used to ask questions.

Examples: किऽ ‘what’, केऽ ‘who’, etc.

3.9. Indefinite Pronoun: A word that refers to an unspecified person or thing.

Examples: कोय ‘someone’, etc.

3.10. Deictic Demonstrative: A word that points to something specific.

Examples: है ‘this’, etc.

3.11. Relative Demonstrative: A word that points to something mentioned earlier.

Examples: जे ‘which’, etc.

3.12. Wh-word Demonstrative: A word used in questions to refer to something specific.

Examples: के ’who’, कोन ‘who’, केना ‘how’, etc.

3.13. Indefinite Demonstrative: A word pointing to something non-specific.

Examples: कोय ‘some one’, कुछु ‘some’, etc.

3.14. Main Verb: The action or main verb in a sentence.

Examples: खाय ‘to eat’, नहाय ‘to bath’, कहि ‘to say’, गेलै ‘to go’, छै ‘is’, छेलै ‘was’, etc.

3.15. Auxiliary Verb: A helping verb that supports the main verb.

Examples: होय ‘is’, रहै ‘is being’, etc.

3.16. Adjective: A word that describes a noun.

Examples: बढ़िया ‘good’, सच्चा ‘truthful’, etc.

3.17. Adverb: A word that describes a verb, adjective, or other adverbs.

Examples: जल्दी ‘fast’, धीरे ‘slow’ , हिन्ने ‘here’, उन्ने ‘there’, etc.

3.18. Postposition: A word that comes after a noun to show a relationship, similar to prepositions in English.

Examples: सँ, कऽ, पे, कॅ, etc.

3.19. Conjunction: A word that joins clauses, sentences, or words.

Examples: केन्ह की ‘that is why’, आरो ‘and’, आरू ‘and’, etc.

3.20. Particles: Small function words that don't fit neatly into other categories.

Examples: या ‘or’, etc.

3.21. Interjection: A word that expresses sudden emotion.

Examples: अहा, वाह ‘wow’ etc.

3.22. Intensifier: A word that emphasizes another word.

Examples: बहुत ‘very’, बेसी ‘very’, खूब ‘very’, etc.

3.23. Negation: A word that negates or reverses meaning.

Examples: नै ‘not’, नाय ‘not’, मत ‘not’, etc.

3.24. General Quantifiers: Words that show quantity in a general sense.

Examples: बहुत ‘very’ , जादा ‘very’, बेसी ‘very’, खूब ‘very’, कुछु ‘some’, etc.

3.25. Cardinal Quantifiers: Words that express exact numbers.

Examples: एक ‘one’, दू ‘two’, तीन ‘three’, etc.

3.26. Ordinals: Words that show the position or order of something.

Examples: पहिलो ‘first’, दुसरो ‘second’, तीसरो ’third’, etc.

3.27. Foreign Word: A word borrowed from another language.

3.28. Symbol: A character or sign used to represent something.

Examples: $, , *, (,), etc.

3.29. Punctuation: Marks used in writing to separate sentences or clarify meaning.

Examples: ., : ; ।, -. ?, etc.

3.30. Unknown: A word or symbol whose meaning is unclear or not recognized.

3.31. Echo Words: Words formed by repeating a sound or syllable, often to create a playful or expressive effect.

Examples: जेना-तेना, अनाप-सनाप, गप-शप, etc.

Table 1: *Proposed Parts of Speech Tagset for Angika*

	Category	Label	Examples

1	Common Noun	NN	घोॅर, आँख,, भोज, हटिया
2	Proper Noun	NNP	सत्तन, पल्टन, पूरब
3	Nloc (Noun denoting spatial and temporal expressions)	NST	आगु, पिछू, ऊपर, नीचा, बीच,
4	Personal Pronoun	PRP	हम्मे, तँऽ, ऊ
5	Reflexive Pronoun	PRF	आपनऽ, आपन्हैं, स्वं, खुद्दे
6	Relative Pronoun	PRL	जेऽ, सेऽ
7	Reciprocal Pronoun	PRC	एक-दोसरो, आपस
8	Wh-word Pronoun	PRQ	किऽ, केऽ
9	Indefinite Pronoun		कोय
10	Deictic Demonstrative	DMD	हिन्ने, उन्ने
11	Relative Demonstrative	DMR	जे
12	Wh-word Demonstrative	DMQ	के, कोन, केना
13	Indefinite Demonstrative	DMI	के, कोय, कुछु
14	Main Verb	VM	खाय, नहाय, कहि, गेलै, छै, छेलै
15	Auxiliary Verb	VAUX	होय, रहै
16	Adjective	JJ	बढ़िया, सच्चा
17	Adverb	RB	एकाएक, जल्दी, धीरे
18	Postposition	PSP	सँ, कऽ, पे, कॅ
19	Conjunction	CC	केन्ह की, आरो, आरू
20	Particles	RP	या
21	Interjection	INJ	अहा, वाह
22	Intensifier	INTF	बहुत, बेसी, खूब,
23	Negation	NEG	नै, नाय, मत
24	General Quantifiers	QTF	बहुत, जादा, बेसी, खूब
25	Cardinals Quantifiers	QTC	एक, दू, तीन
26	Ordinals	QTO	पहिलो, दुसरो, तीसरो
27	Foreign word	RDF	A word written in script other than the script of the original text
28	Symbol	SYM	$, , *, (,)
29	Punctuation	PUNC	., : ; ।, -. ?
30	Unknown	UNK	Unknown Words
31	Echo Words	ECH	जेना-तेना, अनाप-सनाप,गप-शप

4. Tagging Angika Corpus Using BIS POS Tagset

I have collected an Angika corpus from *Kharsup*, an Angika book, and am applying the BIS POS tagset for manual annotation of the text. The manual tagging process presents various

challenges, such as ensuring accurate representation of nuanced linguistic features and handling potential ambiguities unique to Angika. I am still working on overcoming these obstacles to achieve a precise and consistent tagging structure.

5. Conclusion

Developing a POS tagset for the Angika language based on the BIS POS Tagset for Indian Languages is an important step in supporting this low-resource language. By carefully matching the tagset to Angika's unique grammar, I have provided a foundation for future research and the development of NLP tools. The lack of resources for Angika has made it difficult to build such tools, but this work will help to fill that gap. POS tagging is a key task in NLP, and this tagset will make it easier to analyze and work with Angika texts, helping to preserve and promote the language in the digital world.

References

Rathod, S., & Govilkar, S. (2015). Survey of various POS tagging techniques for Indian regional languages. *International Journal of Computer Science and Information Technologies (IJCSIT), 6*(3), 2525-2529.

Das, K. L., Sarmah, S., & Sharma, U. (2019). Identification of POS Tag for Khasi language based on hidden Markov Model POS Tagger. *International Journal of Engineering and Advanced Technology (IJEAT), 8*(5), 1565-1570.

Suman, C., Kiran, J., & Kumar, S. (2023). Script of Angika: A historical writing system. *Analysis: A Peer Review Research Journal of Language and Human Development, 1*(1). *Talent Publication.* https://doi.org/10.13140/RG.2.2.28407.19366

Regmi, A. (2017). *A sociolinguistic survey of Angika*. Report submitted to Linguistic Survey of Nepal (LinSuN). Central Department of Linguistics, Tribhuvan University, Kathmandu, Nepal.

Tosha, M., & Dwivedi, R. R. (2017). Angika folksongs and physical environment: A critical perspective on parallel decline. *Journal of Indian Folkloristics, 20*(1), 89-102.

Dalal, A., Nagaraj, K., Swant, U., Shelke, S., & Bhattacharyya, P. (2007). Building feature-rich pos tagger for morphologically rich languages: Experience in Hindi. ICON.

Chandra, N., Kumawat, S., & Srivastava, V. (2014). Various tagsets for Indian languages and their performance in part of speech tagging. 5th IRF International Conference.

Sankaran, B., Bali, K., Bhattacharya, T., Bhattacharyya, P., Jha, G. N., Rajendran, S.. & Sobha, L. (2008). Designing a common POS-tagset framework for Indian languages. Proceedings of the 6th workshop on Asian language resources.

Bharati, A., Sharma, D. M., Bai, L., & Sangal, R. (2006). AnnCorra: Annotating Corpora guidelines for POS and chunk annotation for Indian languages. *LTRC, International Institute of Information Technology, Hyderabad.*

Antony, P. J., & Soman, K. P. (2011). Parts of speech tagging for Indian languages: A literature survey. *International Journal of Computer Applications, 34*(8), 8-14.

Kumawat, D., & Jain, V. (2014). POS tagging approaches: A comparison. *International Journal of Computer Science and Information Technologies, 5*(1), 1434-1436.

Priyadarshi, A., & Saha, S. K. (2018). Towards the first Maithili part of speech tagger: Resource creation and system development. *Proceedings of the 6th International Conference on Big Data Analytics (BDA)* (pp. 245-252).