Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Comparative Analysis of Morphological Approaches for Low-Resource Indo-Aryan Languages: Case Studies in Angika, Maithili, and Hindi

Alok Kumar[*1] **and Manisha Kumari Deep**[2]

[1]Department of Computer Science Engineering, YBN University Ranchi
[2]School of Computer Science and IT, YBN University Ranchi

[*]Corresponding author: srm.alokkumar@gmail.com

Abstract- Morphological analysis is important and challenging sub-task in Natural Language Processing (NLP), particularly for morphologically rich Indo-Aryan languages. Yet, some regional languages as Angika and Maithili are still under-resourced due to the absence of annotated corpora and computational tools. This work is a comparative study of morphological analysis in the context of Angika, Maithili and Hindi, covering low-and little-resource scenarios. The work compares rule-based, finite-state transducer and hybrid methods and concentrates on the treatment of inflectional and derivational morphology. Linguistically motivated rules and small lexical re- sources are used for low resource languages, while Hindi is used as a reference language. Rule-based and hybrid models perform more robustly and have better interpretable results in low-resource settings than the purely data-driven models. The study demonstrates the need for LingA (linguistic analyzers) to seamlessly combine linguistic knowledge with computational techniques, in order to develop efficient morpho-analysis tools for Indo-Aryan languages that are still under-represented typologically.

keywords: Morphological Analysis, Low-Resource Indo-Aryan Languages, Finite-State Trans-ducer, Rule-Based NLP, Hybrid Morphological Models

1 Introduction

Morphological analysis forms a basis for most NLP systems, especially for languages which have rich inflectional and derivational structure. It allows com- putational models to recover root forms and grammatical features from the sur- face word forms. This procedure is also crucial to enhance the performance of subsequent tasks such as machine translation, part-of-speech tagging and information retrieval [2].

Indo-Aryan languages show highly complex morphological phenomena of suffixation, compounding and agreement. Standard Hindi, being spoken by a larger number of people and

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

also benefitting from the greater attention that computational research receives in it as well as more annotated resources, fares better when compared to most other Indo-Aryan languages. Lower-resource languages, like Angika and Maithili are spoken by millions of people but suffer from the absence of digital corpora to create statistical language model or other NLP tools to facilitate in-depth analysis for the languages; there still is limitation regarding a very limited amount digital corpus, morphological lexicons and annotated resources available for such languages [3].

Morphological processing is especially difficult in low-resource scenarios where the training set is small and orthographic variation exists. In the latter, word forms often represent more than one grammatical category at once resulting in further ambiguity. In the absence of context-insensitive lexicons, the NLP applications generalize poorly over unseen data, with a resulting performance degradation and coverage [8].

Various methods of morphological analysis have been developed, such as rule based systems, finite-state transducers (FST), statistical models and neural architectures. While data-driven approaches to pretraining have performed well for high-resource languages (Radford et al., 2018), their performance degrades rapidly with less annotated data. However, rule-based and hybrid models tend to produce more robust results in resource- constrained environments, particularly when linguistic knowledge is well incorporated [10].

In this paper morphological approaches for three Indo- Aryan languages Angika, Maithili and Hindi are compared. The Bantu languages are closely related both genealogically and structurally for comparative purposes. Thereby they also substantially differ in availability of datasets, standardisation and computational state-of-the-art. The major goal of this study is to investigate the performance of diverse morphology-based treatments in different resource environments. By comparing performance on languages with varied dataset sizes and levels of annotation, this study seeks out scalable and transferable approaches to low-resource language processing. This work has three main highlights. Firstly, it presents a convenient comparison of morphological strategies in many Indo-Aryan languages. Second, it demonstrates the impact of dataset size and quality on sys- tem performance. Third, it has practical implications for the design of morphological analyzers for low-resource languages.

2 Linguistic Background

The following is an overview of the linguistic structure of Angika, Maithili and Hindi mainly from the point of view morphological structure. As morphological analysis rely heavily on language specific characteristics, knowledge of the grammatical and lexical properties of these languages is crucial for system development and evaluation. The three languages all are members of the Indo-Aryan branch of the Indo-European language family and historically have shared linguistic features, but there are also differences in their morphology, standardization and writing system.

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Angika, Maithili and Hindi are part of the Eastern and Central subgroups of Indo-Aryan division. They are descended from Middle Indo-Aryan varieties and have a fairly elaborate inflectional system. They are mostly suffixing languages and have a strict SOV word order. But their morphological counterparts are realised more simply and invariantly.

Table 1: Linguistic and Dataset Characteristics of Angika, Maithili, and Hindi

Language	Resource Status	Script	Corpus Size and Source	Annotation Level
Angika	Low-resource	Devanagari	~150k tokens collected from folk literature, blogs, and local news text	Mostly unannotated with limited manually verified word lists
Maithili	Medium-resource	Devanagari	~500k tokens extracted from literary texts, Wikipedia articles, and online archives	Partially annotated with morphological tags
Hindi	High-resource	Devanagari	>50 million tokens from publicly available corpora and linguistic repositories	Fully annotated with POS and morphological features

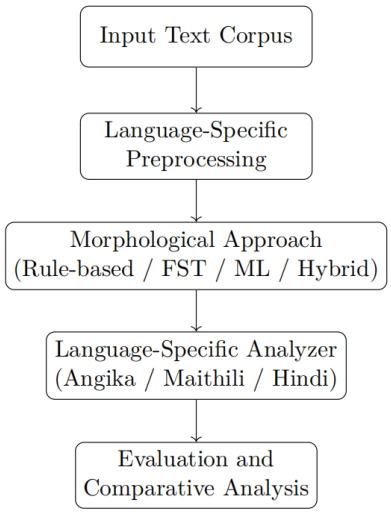

Figure 1: Comparative framework for morphological analysis across Indo-Aryan languages

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Table 2: Genetic and Typological Classification of Selected Languages

Language	Indo-Aryan Subgroup	Word Order	Morphological Type
Angika	Eastern Indo-Aryan	SOV	Highly Inflectional
Maithili	Eastern Indo-Aryan	SOV	Highly Inflectional
Hindi	Central Indo-Aryan	SOV	Moderately Inflectional

2.1 Morphological Features of Angika

Angika is complete with complex polysyllabic words typical of the Eastern Indo-Aryan languages. Its system is overwhelmingly suffixal and agglutinative, with the marking of grammatical relations by inflection and postposition. Gender distinctions are predominantly semantic, and are only partially grammatically marked and activated [9, 6]. Pronouns manifest systematic variation with respect to the person, number and honorificity features, which also index sociolinguistic hierarchies in a speech community. Tense, aspect, mood and the person/honorific of the subject are expressed by complex verb morphology. Verbal agreement is determined more by politeness levels rather than gender and number, a property common with many of the neighbouring Indo-Aryan languages [5]. In Angika, adjectives are as a rule invariable and hardly change form to agree with the noun. Productivity of Derivational Morphology The derivational morphology is productive so that new lexical items can be formed through affixation. Furthermore, replication and compounding may be used to create emphasis, plurality; as well as semantic extension. These morphological properties create certain challenges for computational modeling, especially in the context of low-resource NLP. Some characteristic features are shown in table 3.

Table 3: Examples of Angika Morphological Forms

Surface Form	Root	Morphological Information
लइकन	लइका	Noun, plural
खाइतहलहुँ	खा	Verb, past tense, first person
घरवाला	घर	Noun, derivational suffix

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

2.2 Morphological Features of Maithili

Maithili shows an elaborate and characteristic morphology of the Eastern Indo-Aryan languages. Its morphological type is mostly suffixal, with grammatical relations typically expressed via inflection and/or postpositions. Nouns in Maithili exhibit very little gender marking, and are inflected mostly for number and case. Nominal agreement is limited to a weak degree, while case relationships are frequently indicated by postpositions [9, 6]. The system is quite elaborated in the language and encode person, a number as well as more than one level of honorificity, thereby depicting deep sociolinguistic sensitivity. Verb conjugations are highly complex, and diagram the tense, aspect, mood, person and honorific state of each verb. In contrast to Hindi, AGR in Maithili is mostly an honorific agreement rather than gender or number based, thus honorific is a key grammatical feature [24]. Adjective in Maithili are mostly not number sensitive, they need not agree with nouns, but sometimes can. Derivational morphology is productive, permitting lexicon expansion by means of affixation. The use of compounding and reduplication Interestingly enough, compound words in panic are used for emphasis and plurality that variations arise. These features add to the morphological complexity of Maithili, which is also arduous to deal with for computational processing in low-resource languages. Example features can be found in table 4.

Table 4: Examples of Maithili Morphological Forms

Surface Form	Root	Morphological Information
लइकासभ	लइका	Noun, plural marker
करैत छिथ	कर	Verb, present tense, honorific
पढ़नाइ	पढ़	Verbal noun formation

2.3 Morphological Features of Hindi

Hindi displays a neatly documented and systematic set of morphological phenomena which is characteristic for Central Indo- Aryan languages. Its morphological system is mostly suffixal and many inflections are postpositional, with the expression of grammatical relations handled by postpositions rather than noun-case inflections. Hindi nouns are inflected for gender, number and case, and agree with adjectives and verbs [7, 16]. Pronouns: Pronouns in Hindi inflect for person, number, gender and honorificity and can have significant implications on agreement. The morphology of the verb in Hindi is very highly developed, and is marked for tense, aspect, mood, person, number and gender. The verbal agreement is subject or object marking sensitive, particularly in ergative perfective constructions [4]. Adjectives in Hindi may be either inflecting or non-inflecting, the latter called invariabl e adjectives agreeing with no uns in gender and number. Productive derivational morphology allows the construction of nouns, verbs and

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

adjectives by affixation. On the other hand, compounding and reduplication are often utilized for emphasizing effect, distributive meaning and semantic change. Hindi is also an ideal language for doing comparative and computational morphology work because of its comparably standardised morphological characteristics, textual resources etc. Example features are outlined in table 5.

Table 5: Examples of Hindi Morphological Forms

Surface Form	Root	Morphological Information
लड़कियाँ	लड़का	Noun, feminine plural
खाया था	खा	Verb, past perfect
लिखावट	लिख	Noun, derivational suffix

2.4 Dataset-Oriented Morphological Comparison

This morphological diversity in languages has direct impact on design of datasets and annotation strategies. Angika needs to be carefully validated based on a concordance with manual material because of the variation in orthography. It is able to perform semi-automatic Maithili annotation with the help of rule. Hindi: Unlike most of the previous works, Hindi has complete automatic annotation pipeline. This variance needs to be captured in adaptive morphology. Table 6 lists exemplary features.

Table 6: Morphological Dataset Requirements Across Languages

Language	Major Morphological Challenges	Dataset Size Needed	Annotation Effort
Angika	Dialectal variation, spelling inconsistency	High relative to availability	Manual-heavy

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Maithili	Rich inflection, partial standardization	Medium	Semi-automatic
Hindi	Agreement complexity	Moderate	Fully automatic

3 Review of Existing Morphological Approaches

Morphological analysis is a crucial challenge in NLP for resource-strained languages. In the last few decades, various methods have been developed which address inflectional and derivational morphology in Indo-Aryan languages. These methods can be generally divided into (i) rule-based, (ii) finite-state transducer (FST)-based approaches, (iii) statistical and machine learning techniques, and (iv) hybrid models which join rules with machine learning. This section evaluates critically the current morphological techniques as applied to Angika, Maithili and Hindi, highlighting their advantages/disadvantages and appropriateness for low-resource setting.

3.1 Rule-Based Approaches

Rule-based morphological analyzers rely on handcrafted grammatical rules to generate and analyze word forms. They require detailed knowledge of inflectional paradigms and derivational patterns of the target language. For example, the suffix-based rules can generate plural nouns in Angika, such as: लइका → लइकन (Noun, plural) and घर → घरवाला (Noun, derivational).

Hindi and other Indo-Aryan languages Rule-based morphological analysis has been the workhorse of language processing in Hindi. Agarwal et al [1] designed a rule-based morphological analyzer that treats both inflectional and derivational morphology using handcrafted linguisti- cally motivated rules and exceptions. Rastogi and Khanna [20] presented a rule based morphological analyzer along with corpus knowledge for root form identification and grammatical features such as gender, number, and part of speech. Kumar et al [14] proposed a rule-based morphological analy zers with dictionary and compounding for Hindi nouns, with an extensive tag-set for improved analysis. While these evaluations have shown that the rule-based approaches to morphology are efficient and linguistically transparent, the reliance upon hand-crafted rules and limited scalability of some frameworks call for more flexible morphological structures.

3.2 Finite-State Transducer (FST) Approaches

FST-based analyzers encode morphological structure in terms of a network consisting of states corresponding to roots, suffixes and morphotactic constraints. For Hindi, they are widespread because it is a language with regular morphological structure. For example, the FST can produce all forms of verbs from सर (to eat): खा: खाता है (Verb, present tense, masculine singular) and खाया था (Verb, past perfect).

Finite-State Transducer (FST)–based methods have been often used in the context of

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

morphological analysis due to their formal accuracy as well as computational tractability [19]. FSTs were developed by Beesley and Kart- tunen [2] as a powerful framework for representing morphological pro- cesses using bidirectional lexical to surface mappings. Karttunen [11] also showed that FSTs are appropriate for modeling rule-based morphology by translating linguistic rules into finite-state networks. Follow-up work in other languages also showed that the FST-based analyzers satisfy at least consistency, reversibility and linguistic transparency for morphological processing. Nevertheless, such systems tend to require significant levels of linguistic expertise as well as manual rule crafting, and hence have not been proven to be adaptable for low-resource or under-documented languages. In consequence, recent research focus has been on hybrid architectures that combine FST based methods with rule-based or data-driven methods in order to improve scalability and coverage.

3.3 Statistical and Machine Learning Approaches

Statistical methods, including Hidden Markov Models (HMMs),Conditional Random Fields (CRFs), and neural networks, learn morphological patterns from annotated corpora. For example, a CRF model can predict suffixes and POS tags for Maithili words: लइकासभ → Root: लइका, Morphology: Noun, plural marker and पढ़नाइ → Root: पढ़, Morphology: Verbal noun.

Statistical and machine-learning approaches learn morphological patterns from annotated cor- pora and have become central to modern morphological analysis. Hidden Markov Models (HMMs) provided an early probabilistic framework for sequence modeling and were widely used for tagging and segmentation tasks. [18] Conditional Random Fields (CRFs) extended this work by offering a discriminative sequence-labeling model that relaxes HMM independence assumptions and is well suited for morphological segmentation and labeling. [15] CRF-based analyzers have been success- fully applied to morphologically rich languages (e.g., Japanese), demonstrating strong performance when informative feature sets are available. [13] More recently, neural architectures — especially character-level and windowed LSTM models — learn morpheme boundaries and label sequences end-to-end without heavy feature engineering, improving robustness across typologically diverse languages. [23, 17] For low-resource Indo-Aryan languages (example: Maithili), supervised ML models can predict roots and morphological features given annotated training data (e.g.,→ root:, Noun+plural;→ root:, Verbal noun), but they require sufficiently large and repre- sentative corpora and careful domain adaptation. [17] While statistical and neural methods scale better than handcrafted systems, their effectiveness depends on corpus size, annotation quality, and transferability — which motivates hybrid pipelines that combine linguistic constraints with data-driven learning.

3.4 Hybrid Approaches

Some hybrid approaches 60 combine rule-based linguistic knowledge with statistical and machine learning methods in order to overcome the deficiencies of purely symbolic or purely data- riven systems. Hand made rules guarantee reliable treatment of regular and predictable morphological patterns (like suffixation and inflectional paradigms), whereas statistical or neural IR captures irregular forms as well as unseen variants from a training set. In previous work, we

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

have shown that marrying the FSM based or rule based analyzers with a statistical model increases robustness and coverage for morphologically rich languages [12, 22]. Recent works have demonstrated that rule based frameworks can be 'wrapped' in neural components and utilized to refine morphological predictions and mitigate the error propagation [21]. For low-resource Indo-Aryan languages like Angika, hybrid architectures work best: rule based suffix patterns can accordlingly analyze the plural nouns whereas for generalizing verb inflexions and parsing lexical ambiguity neural models works wells." This balanced integration provides both scaling, linguistic interpretability and better results and hybrid systems is a promising way for sustainable morphological analysis in under resourced languages [17].

Table 7: Comparison of Morphological Approaches for Low-Resource Indo-Aryan Languages

Approach	Languages Tested	Strengths	Limitations
Rule-Based	Angika, Maithili	Interpretability, high linguistic consistency	Labor-intensive, low coverage for dialects
FST-Based	Hindi	Fast, accurate, scalable	Requires detailed lex-icons, low adaptabil- ity for low-resource lan-guages
Statistical / ML	Hindi, Maithili	Adaptable, auto-mates feature ex-traction	Requires large anno-tated corpora, poor un-seen word handling
Hybrid	Angika, Maithili	Combines Strengths of rules and ML Complexity in system integration	

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037
International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

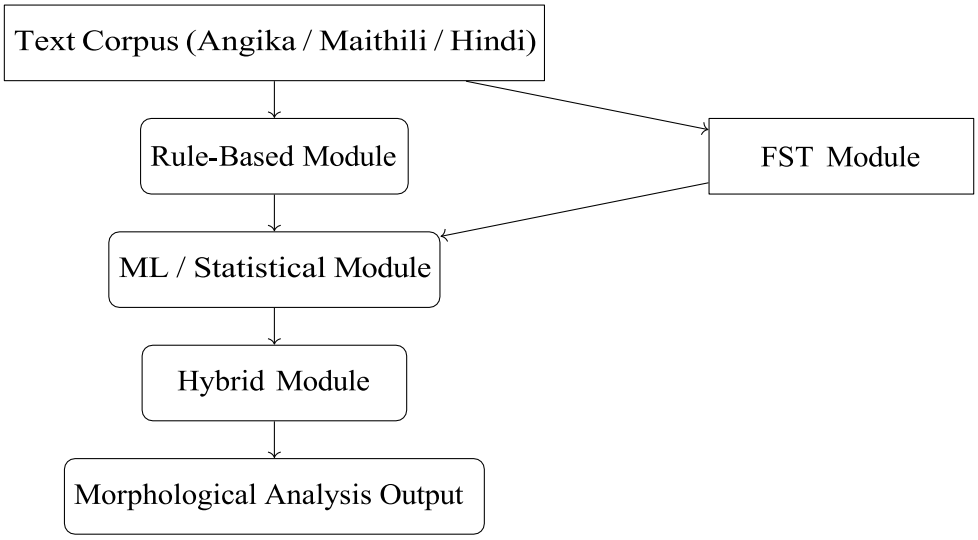

Figure 2: Framework of Morphological Analysis Approaches for Low-Resource Indo-Aryan Languages

4 Issues and Challenges for Maithili and Hindi

While there has been increasing interest in IndoAryan language technologies, Maithili and Hindi continues to struggle with computational development; though the scales of struggles and intensity of challenges that these two languages encounter differ. These challenges are due to resource availability, linguistic diversity, technological infrastructure and institutional support which have direct implications on the design and evaluation of NLP systems.

Maithili is a low-resource language and its digital availability is limited along with the presence of labeled standard linguistic resources. Large, balanced and representative text and speech corpora are rare enough to seriously hinder the performance of supervised or data-driven learning procedures. Orthographic variation has also made the creation of a corpus more problematic: Maithili was traditionally written in the Tirhuta script, but nowadays it is mostly written using Devanagari; as a result two parallel and non-interchangeable traditions coexist. In addition, available annotated datasets for core NLP tasks such as POS-tagging, morphological analy- sis and syntactic parsing are either very small or non-existent which makes systematic evaluation difficult. Regional differences in dialect and thereby additional lexical and morphological variation is however hardly included in such data. The absence of mature language-engineering resources (e.g., high-quality tokenizers, lemmatizers and morphological analyzers) further hinders computational advance. In the speech do main, the lack of well-annotated audio corpora limits both the development of systems for automatic speech recognition and synthesis. These technical bottlenecks are further aggravated by a lack of lexicographic resources, scarcity of trained annotators across years, non-availability of scholarly tagged documents for the extraction and expanding ages in that it requires digitization due to an un optimized OCR support for Tirhuta, compared to other scripts and lesser funding at the institution levels and policy-level prioritization.

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Hindi, which is relatively better endowed, poses its own set of challenges due to the scale, diversity and distributional properties from real-world usage. There is widespread dialectal and register variation among the Hindi-speaking world entails modelling challenges in terms of generalisation across) communities or communicative contexts. There is a high level of code-switching with English and other regional languages in informal and digital communication which results in mixed-script data, non-standard spellings included and complex tokenization. Hindi is linguistically rich with respective to inflectional morphology, compounding, and cliticization that continues to cause big problems for precise segmentation and tagging. Even though there are large corpora, domain specific datasets for legal, medical and low literacy contexts especially in Indian languages are scarce. Assessment methods are frequently disjoint, resulting in consistent reference information and metrics between studies for meaningful results' comparison. User-generated text also brings in more noise such as spelling variations, transliterations, emojis and punctuations making models pretrained on clean text less reliable. In the speech processing field, existing datasets focus on standard varieties with regional accent and colloquial faster-speech being underrepresented. Problems with data licensing, the central availability of resources, bias in training data, and computational requirements to fine-tune large pre-trained language models are additional barriers to inclusive and reproducible analysis.

Collectively, these challenges emphasize that despite Hindi and Maithili falling at different ends of the resource spectrum, each requires carefully designed strategies accounting for linguistic diversity, resource imbalance and real world variability. Solving these issues are critical for building scalable, fair and socially inclusive language technologies for Indo Aryan languages.

5 Discussion

The development paths of Maithili and Hindi morphological analyzers has several specific lessons for Angika researchers who want to develop hilanguage technology. The divergent resource condition of the two languages underscores the importance of neatly curated corpora: while Hindi can make good use from a rather large repository of text and annotated datasets, Maithili has been stifled by paucity and fragmentation. For Angika, this means that initial investment is a good idea in terms of building balanced representative corpora of text and speech with clear denotation of the source, register and preprocessing steps on the way to reproducible slicing-and-dicing since it'll save one relying on fragile heuristics.

Second, orthographic integrity and normalisation rules are underlined as practical requirements for scalable systems. The presence of several scripts and ad hoc Romanization in related Indo-Aryan languages complicated corpus harmonization and model training, so Angika developers should concentrate efforts on a standardization layer (or strong transliteration/normalization modules) compatible with preserving linguistic distinctions while yielding an accessible working single representation for further use.

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

Third, methodological decisions must be friends of hybridity rather than purity. Rule-based and finite-state approaches can be used for precise definition of morphotactic regularities and linguistic restrictions, while statistical and neural models enable also flexibility towards irregular phenomena and distributional generalization. Hybrid pipelines translating high-confidence morphological rules and machine learning of ambiguous or rare patterns balance interpretability with scalability; Angika systems will benefit from this design approach, particularly in the low-annotation regime.

Design and evaluation of annotation practice annotation design and evaluation need to be intentional and conducted in the community. The splintered benchmarks in Hindi make cross-LSA difficult; so to not fall into the same trap, projects on Angika need to converge on annotation schemas, inter-annotator agreement protocols and share a small number of test suites addressing input scenarios like code-switching, dialectal differences and domain shift. Consuming both the guidelines, and our modest yet well-documented testbed publicly will speed up the time for actual life comparison and external adoption.

Fifth, non-textual modalities and problems in digitizations cannot be dismissed. Experiences from Maithili demonstrate that scant OCR support and speech data create compounded downstream bottlenecks for digitization of historic texts and ASR corpus development. Thus, Angika work should include context-dependent OCR training for print traditions, design methodologies for speech corpus development with emphasis on accental diversity and create metadata headers that allows multimodal alignment.

Capacity building and collaborative governance(ecological) that matters, as not much as algorithmic decisions. Sound and sustainable progress in Hindi-MT has frequently come about on the basis of institutional support, large coordinated annotation drives, and shared tool development. Angika research should be pursued in active collaboration with native speaker communities, local institutions, and the open-source community to reach a larger scale of annotation, attend linguistic assumptions which may arise including validation.Datasets generated must be ethically steward. And lastly, ethical and practical safeguards have to be hardwired from the beginning. Problems in licensing, representativeness, and bias experienced during Hindi research demonstrate how technical improvements can be implemented without the benefits necessarily following suit with respect to inclusivity when data lineage, consent, or demographic coverage are not explicitly handled. In the case of Angika, transparent licensing, work with marginalized dialects and bias audits at regular intervals can help guarantee that technologies benefit all users.

6 Conclusion

This paper compared the morphological and computational aspects of Indo-Aryan languages: Angika, Maithili, Hindi in low resource setting. The study showed that despite its highly standardized morphology and list of linguistic resources, Hindi has several standard aspects such

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

as suffixation based morphology, honorificity driven agreement and productive derivational processes which also bring unique challenges to building computational models for Angika and Maithili. The results of our analysis indicate that completely data oriented techniques fail in many cases for low resource languages, whereas rule based and hybrid systems based on linguistic knowledge are more reliable and interpretable solutions. In general, the results highlight the value of linguistically sensitive morphological analysis for low resource Indo Aryan languages and are expected to be useful in resource building, language conservation and downstream NLP tasks.

References

[1] Ankita Agarwal, Shashi Pal Singh, Ajai Kumar, Hemant Darbari, et al. Morphological analyser for hindi-a rule based implementation. *International Journal of Advanced Computer Research*, 4(1):19, 2014.

[2] Kenneth R. Beesley and Lauri Karttunen. *Finite State Morphology*. CSLI Publications, 2003.

[3] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages. *Speech Communication*, 56:85–100, 2014.

[4] Miriam Butt. *The Structure of Complex Predicates in Urdu*. CSLI Publications, Stanford, 1995.

[5] Amit Kumar Chandrana and Neha Garg. Number and gender agreement: A comparative study of angika and maithili. *Anukriti: An International Peer Reviewed Refereed Research Journal*, 11(6):47–52, 2021.

[6] Suniti Kumar Chatterji. *The Origin and Development of the Bengali Language*. George Allen & Unwin, London, 1926.

[7] Suniti Kumar Chatterji. *Indo-Aryan and Hindi*. Motilal Banarsidass, Delhi, 1960.

[8] John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

[9] George Abraham Grierson. *Linguistic Survey of India, Vol. V: Indo-Aryan Languages*. Government of India, Calcutta, 1903.

[10] Nizar Habash. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool, 2010.

[11] Lauri Karttunen. Constructing lexical transducers. *Proceedings of the ACL Workshop on Computational Morphology*, pages 1–10, 1997.

[12] Kimmo Koskenniemi. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD thesis, University of Helsinki, 1983.

[13] Taku Kudo and Yukio Matsumoto. Applying conditional random fields to japanese morpho-

Received: 2025-07-15
Accepted: 2025-09-15
Published Online: 2025-09-30
DOI: 10.70454/IJMRE.2025.05037

International Journal of Multidisciplinary Research and Explorer (IJMRE)
E-ISSN: 2833-7298, P-ISSN: 2833-7301

logical analysis. In *ACL Workshop on Morphological and Phonological Processing*, 2004.

[14] Ishan Kumar, Renu Dhir, Gurpreet S Lehal, and Sanjeev Kumar Sharma. Design of dynamic morphological analyser for hindi nouns using rule based approach. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, 13(6):1152–1157, 2020.

[15] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289. Morgan Kaufmann, 2001.

[16] Colin P. Masica. *The Indo-Aryan Languages*. Cambridge University Press, Cambridge, 1991.

[17] Siddhesh Pawar and Pushpak Bhattacharyya. Neural morphology analysis – a survey. Technical report, CFILT, IIT Bombay, 2022. survey; available as CFILT technical report.

[18] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[19] Raza Rahi, Sumant Pushp, Arif Khan, and Smriti Kumar Sinha. A finite state transducer based morphological analyzer of maithili language. *arXiv preprint arXiv:2003.00234*, 2020.

[20] Mayuri Rastogi and Pooja Khanna. Development of morphological analyzer for hindi. *International Journal of Computer Applications*, 95(17):1–5, 2014.

[21] Teemu Ruokolainen and Mikko Kurimo. Neural network morphological analyzers for highly inflecting languages. In *Proceedings of the Workshop on Computational Morphology and Phonology*, 2016.

[22] Helmut Schmid. Efficient parsing of highly ambiguous context-free grammars with bit vectors. *Proceedings of COLING*, 2004.

[23] Linlin Wang, Zhu Cao, Yu Xia, and Gerard de Melo. Morphological segmentation with win- dow lstm neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[24] Ramawatar Yadav. *A Reference Grammar of Maithili*. Mouton de Gruyter, Berlin, 1996.