Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

CSI4001
Natural Language Processing

Winter Semester 2024 -25

J – Component Final Report

Under Prof. Sharmila Banu K

Generating a Lexicon of 1 lakh words for the Nepali Language with its Word Pronunciation (English Approx.), Grammatical Meaning, Part of Speech Syntactic Meaning and its Semantic Information

-Samay Raj Adhikari
21MID0181

Introduction:

Natural Language Processing (NLP) is a unique set of challenges to work with low-resources languages like Nepali. Unlike English, Nepali does not have a treasure of annotate corpora, comprehensive dictionary or morphological analysis such as equipment and dataset. But these are things that work effectively the NLP system. One of the most important pieces is a solid list of root or base words - the main form of words that help in understanding the meaning, performs stems or lemmatization, and clean the noise text. Nepali is a morphological language, which means that a single word can be carried on many forms in addition to prefixes, suffixes and other modifies. Without a good root word list, it is easy for the model to be confused with all these variations. This is why the creation of a large, smart-generated set of root-like words is such an important step-it lays the foundation for everything else in the pipeline.

Problem Statement:

One of the biggest obstacles in the construction of NLP systems for Nepali is the lack of a reliable root word corpus. Most available datasets are either too small or do not account for rich word formation patterns found in language. Nepali words often change form due to affixes, pluralization, honorifics, and compounding- making it becomes difficult for the model to identify different versions of the same root. Without a strong set of base words to refer to, tasks like stemming, lemmatization, or even basic word matching become less accurate. This creates a gap in performance for systems that rely on clean, normalized text. So, the core problem is simple: we need a well-thought-out, scalable way to generate a large and diverse list of Nepali root words to support foundational NLP tasks.

Use case of my project:

The generated Nepali lexicon with detailed pronunciation, grammatical, syntactic, and semantic information serves as a valuable resource for a variety of Natural Language Processing (NLP) applications. It can be utilized in language model training, particularly for BERT-like models, to improve NLP tasks in Nepali, a low-resource language. Researchers and developers can leverage the lexicon for comparative studies between languages using the Devanagari script, such as Nepali and Hindi. It also acts as a foundational resource for building models focused on text normalization, stemming, and lemmatization. Additionally, it supports those interested in NLP and language modelling research for Nepali, offering insights into language structure, morphology, and linguistic features that are critical for more advanced computational linguistics tasks.

Methodology:

The technique for generating the Nepali lexicon includes a multi-step technique designed to create a diverse and linguistically rich useful resource. Initially, a centre set of 20,000 Nepali root phrases is accumulated. This is observed via systematic enlargement using a set of predefined guidelines that account for pluralization, affixation, compound phrase technology, and reduplication.

These guidelines ensure that the lexicon captures a huge range of phrase variations, that's essential for processing the morphologically rich Nepali language. The lexicon is then further superior with part-of-speech (POS) classification, grammatical, syntactic, and semantic information thru a mixture of rule-based totally techniques and linguistic heuristics. Each word is mapped to its corresponding POS, grammatical which means, syntactic function, and semantic records, making an allowance for a complete understanding of the word's utilization.

Additionally, replica and inappropriate entries are filtered out to make certain statistics excellent. The very last lexicon is then stored in more than one codecs, consisting of CSV and HTML, for clean get admission to and usage in NLP packages.

This lexicon may be used for numerous responsibilities, which include model schooling for language processing, move-language comparisons, and linguistic studies, ultimately contributing to the development of extra correct and green NLP structures for Nepali

Rule Used:

In this project, rule-based logic played a central role in generating linguistic annotations for each word in the lexicon. Rather than relying on pre-trained models or external libraries, a custom set of handcrafted rules was developed to determine the part of speech (POS) based on common Nepali suffixes (e.g., words ending in "ने" or "नु" were identified as verbs, while those ending in "ता" or "पन" were treated as nouns).

Similarly, fixed mappings had been used to companion every POS with its corresponding grammatical meaning, syntactic function, and semantic class. This approach allowed for steady, interpretable tagging and ensured scalability across all one 1,00,000 entries. By embedding linguistic information into the common sense, the machine completed a dependent and wise manner of enriching each word without the need for huge annotated datasets.

Errors faced:

During the development of lexicon, many challenges faced, especially when generating an early set of 20,000 root words. The process of scrapping and ensuring that the words were formed correctly, initially more complicated than anticipated. Several adjustments were required to refine the scraping process and recognize and remove valid original words accurately.

During this phase, several adjustments were made in the code, including changes to handle the variations and ensured that the original words follow the morphological rules of the Nepali language.
Additionally, the task of stemming and transliterating Nepali words reflects significant obstacles. The initial translation did not accurately map the Nepali script for its Romanized counterpart, causing discrepancies. There were several rounds of purification to ensure that the accent was represented correctly in the English connect, especially for complex characters and combinations.
These issues were resolved by testing them against a wide range of words by constantly updating the mapping rules and to achieve the desired output. Despite these challenges, the recurrence of the process allowed for continuous improvement, leading to the stronger and more accurate lexicon that fulfils the requirements for NLP functions.

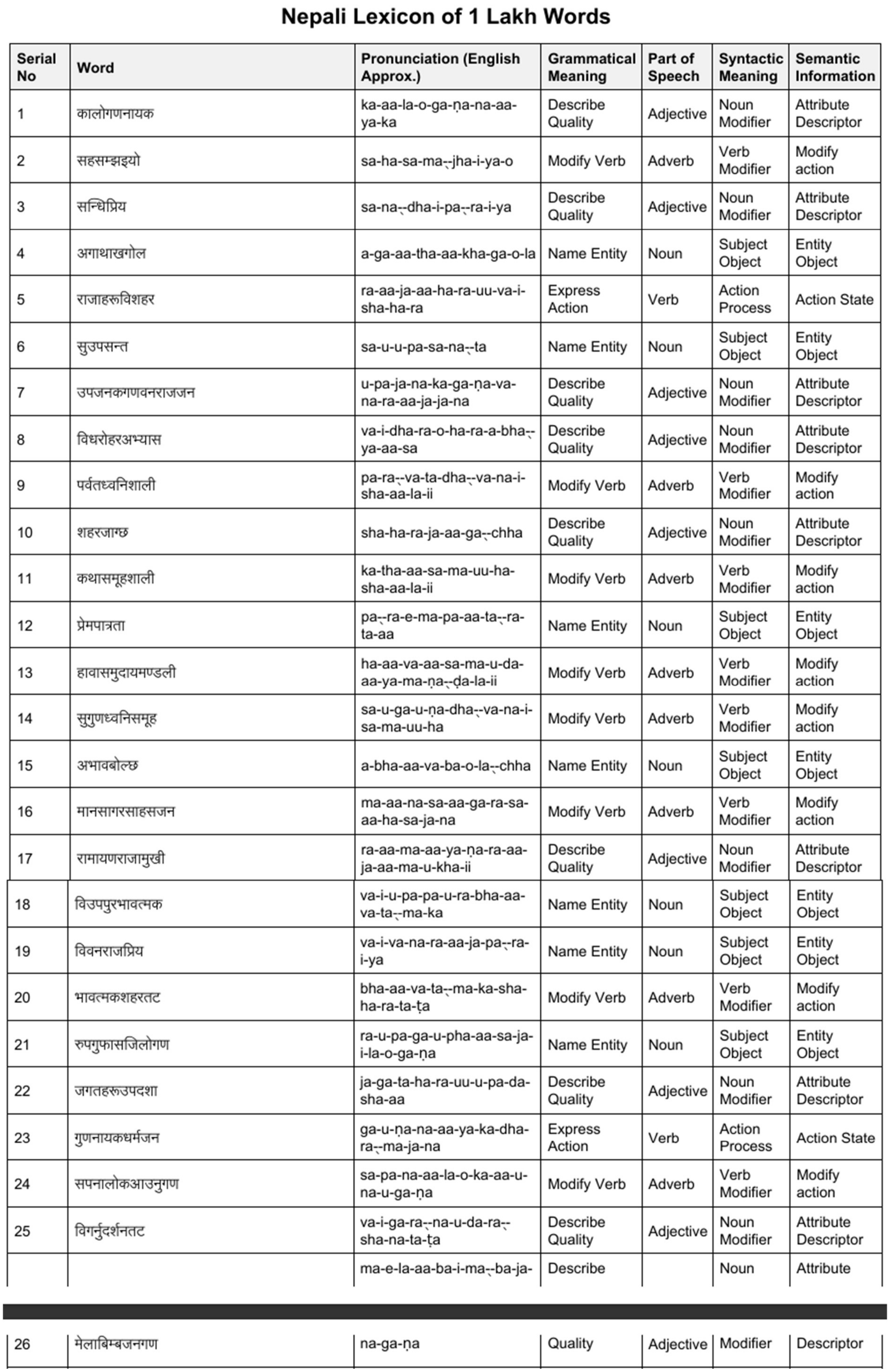
Constraints/Limitations:

The system uses rule-based POS tagging based on Nepali suffixes like **"ने"**, **"ता"**, or **"पन"**, but this leads to misclassification for irregular forms. Transliteration from Devanagari to Roman script uses a static map (e.g., **"क"** → **"ka"**, **"त्र"** → **"tra"**), but fails for conjuncts and diacritics, causing mismatches in pronunciation. Since there's no morphological stripping, affixed or compound forms like **"विद्यालयमा"** aren't reduced to **"विद्यालय"**. Words that don't match any rule are tagged randomly, which injects noise. Generating the initial 20k root words was difficult due to limited scraping accuracy, requiring multiple corrections and code adjustments. The lack of context handling also means words like **"काम"** (work vs. deed) can't be disambiguated.

Conclusion to my project:

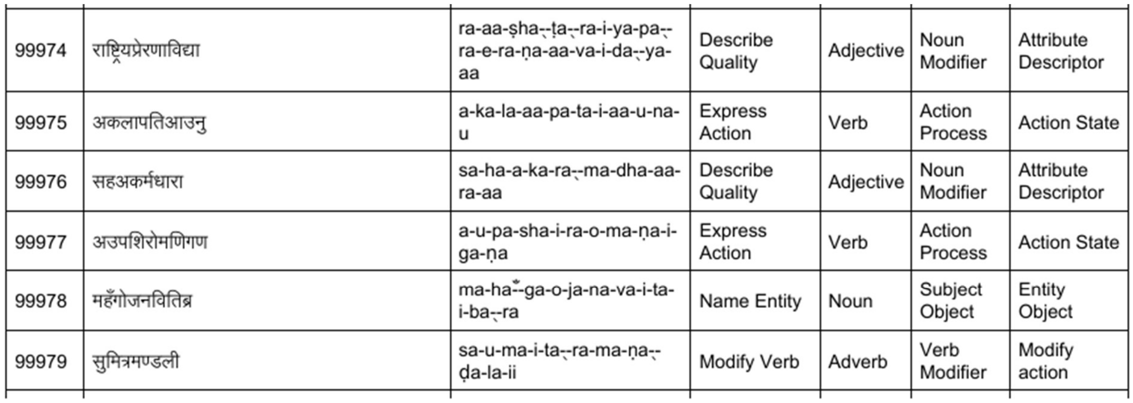
The project is an important step to further NLP resources for Nepali, which is a low-resources yet morphologically rich language. By generating a high-quality lexicon of 1 lakh words- that is each annotated with each English-approximate pronunciation, grammatical meaning, part of speech, syntactic role, and semantic information-This gives a strong foundation for a wide range of language technologies.
Despite the challenges in the word extraction and accurate translation, repetition refinement led to a strong and scalable resource. This lexicon can support future development of Nepali language modelling, linguistic research, and cross-script or cross-language, especially within the Devanagari-based languages.
It provides valuable ground truth for researchers, teachers and developers, which is considered more accessible and better in the field of computational linguistics to Nepali.

Some snapshots of my project, the top 26 and the last 26 words from my project.

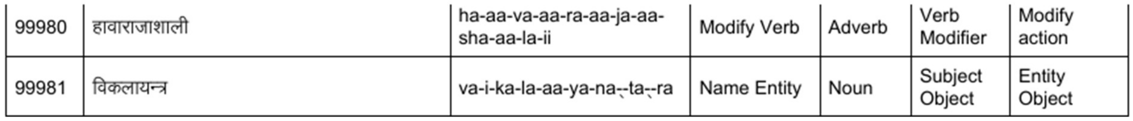


Nepali Lexicon of 1 Lakh Words

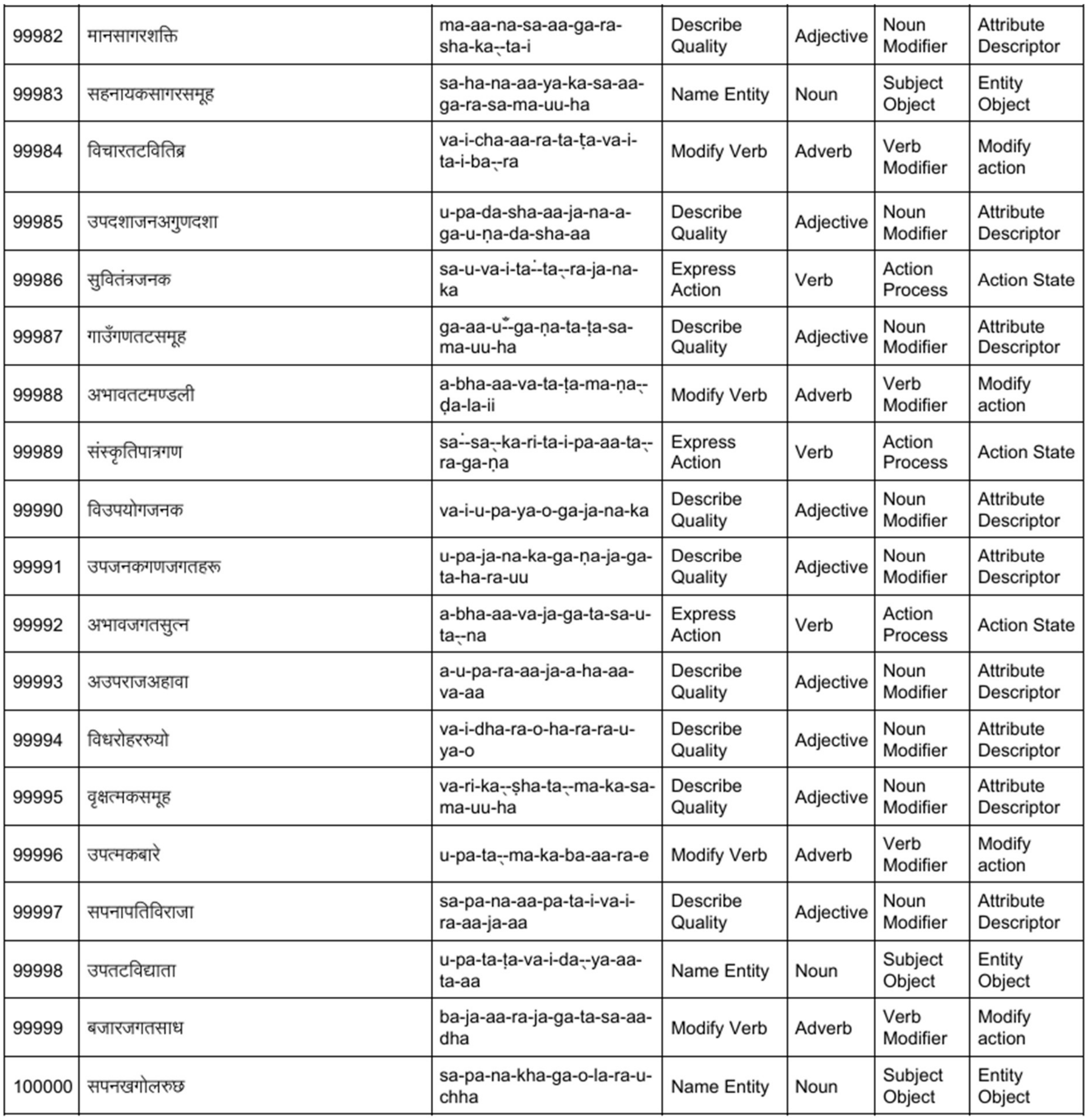
Serial No	Word	Pronunciation (English Approx.)	Grammatical Meaning	Part of Speech	Syntactic Meaning	Semantic Information
1	कालोगणनायक	ka-aa-la-o-ga-ṇa-na-aa-ya-ka	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
2	सहसम्झइयो	sa-ha-sa-ma-्-jha-i-ya-o	Modify Verb	Adverb	Verb Modifier	Modify action
3	सन्धिप्रिय	sa-na-्-dha-i-pa-्-ra-i-ya	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
4	अगाथाखगोल	a-ga-aa-tha-aa-kha-ga-o-la	Name Entity	Noun	Subject Object	Entity Object
5	राजाहरूविशहर	ra-aa-ja-aa-ha-ra-uu-va-i-sha-ha-ra	Express Action	Verb	Action Process	Action State
6	सुउपसन्त	sa-u-u-pa-sa-na-्-ta	Name Entity	Noun	Subject Object	Entity Object
7	उपजनकगणवनराजजन	u-pa-ja-na-ka-ga-ṇa-va-na-ra-aa-ja-ja-na	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
8	विधरोहरअभ्यास	va-i-dha-ra-o-ha-ra-a-bha-्-ya-aa-sa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
9	पर्वतध्वनिशाली	pa-ra-्-va-ta-dha-्-va-na-i-sha-aa-la-ii	Modify Verb	Adverb	Verb Modifier	Modify action
10	शहरजाग्छ	sha-ha-ra-ja-aa-ga-्-chha	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
11	कथासमूहशाली	ka-tha-aa-sa-ma-uu-ha-sha-aa-la-ii	Modify Verb	Adverb	Verb Modifier	Modify action
12	प्रेमपात्रता	pa-्-ra-e-ma-pa-aa-ta-्-ra-ta-aa	Name Entity	Noun	Subject Object	Entity Object
13	हावासमुदायमण्डली	ha-aa-va-aa-sa-ma-u-da-aa-ya-ma-ṇa-्-ḍa-la-ii	Modify Verb	Adverb	Verb Modifier	Modify action
14	सुगुणध्वनिसमूह	sa-u-ga-u-ṇa-dha-्-va-na-i-sa-ma-uu-ha	Modify Verb	Adverb	Verb Modifier	Modify action
15	अभावबोल्छ	a-bha-aa-va-ba-o-la-्-chha	Name Entity	Noun	Subject Object	Entity Object
16	मानसागरसाहसजन	ma-aa-na-sa-aa-ga-ra-sa-aa-ha-sa-ja-na	Modify Verb	Adverb	Verb Modifier	Modify action
17	रामायणराजामुखी	ra-aa-ma-aa-ya-ṇa-ra-aa-ja-aa-ma-u-kha-ii	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
18	विउपपुरभावत्मक	va-i-u-pa-pa-u-ra-bha-aa-va-ta-्-ma-ka	Name Entity	Noun	Subject Object	Entity Object
19	विवनराजप्रिय	va-i-va-na-ra-aa-ja-pa-्-ra-i-ya	Name Entity	Noun	Subject Object	Entity Object
20	भावत्मकशहरतट	bha-aa-va-ta-्-ma-ka-sha-ha-ra-ta-ṭa	Modify Verb	Adverb	Verb Modifier	Modify action
21	रुपगुफासजिलोगण	ra-u-pa-ga-u-pha-aa-sa-ja-i-la-o-ga-ṇa	Name Entity	Noun	Subject Object	Entity Object
22	जगतहरूउपदशा	ja-ga-ta-ha-ra-uu-u-pa-da-sha-aa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
23	गुणनायकधर्मजन	ga-u-ṇa-na-aa-ya-ka-dha-ra-्-ma-ja-na	Express Action	Verb	Action Process	Action State
24	सपनालोकआउनुगण	sa-pa-na-aa-la-o-ka-aa-u-na-u-ga-ṇa	Modify Verb	Adverb	Verb Modifier	Modify action
25	विगर्नुदर्शनतट	va-i-ga-ra-्-na-u-da-ra-्-sha-na-ta-ṭa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
26	मेलाबिम्बजनगण	ma-e-la-aa-ba-i-ma-्-ba-ja-na-ga-ṇa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor



99974	राष्ट्रियप्रेरणाविद्या	ra-aa-ṣha-̖-ṭa-̖-ra-i-ya-pa-̖-ra-e-ra-ṇa-aa-va-i-da-̖-ya-aa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99975	अकलापतिआउनु	a-ka-la-aa-pa-ta-i-aa-u-na-u	Express Action	Verb	Action Process	Action State
99976	सहअकर्मधारा	sa-ha-a-ka-ra-̖-ma-dha-aa-ra-aa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99977	अउपशिरोमणिगण	a-u-pa-sha-i-ra-o-ma-ṇa-i-ga-ṇa	Express Action	Verb	Action Process	Action State
99978	महँगोजनवितिब्र	ma-ha-̐-ga-o-ja-na-va-i-ta-i-ba-̖-ra	Name Entity	Noun	Subject Object	Entity Object
99979	सुमित्रमण्डली	sa-u-ma-i-ta-̖-ra-ma-ṇa-̖-ḍa-la-ii	Modify Verb	Adverb	Verb Modifier	Modify action



99980	हावाराजाशाली	ha-aa-va-aa-ra-aa-ja-aa-sha-aa-la-ii	Modify Verb	Adverb	Verb Modifier	Modify action
99981	विकलायन्त्र	va-i-ka-la-aa-ya-na-̖-ta-̖-ra	Name Entity	Noun	Subject Object	Entity Object



99982	मानसागरशक्ति	ma-aa-na-sa-aa-ga-ra-sha-ka-̖-ta-i	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99983	सहनायकसागरसमूह	sa-ha-na-aa-ya-ka-sa-aa-ga-ra-sa-ma-uu-ha	Name Entity	Noun	Subject Object	Entity Object
99984	विचारतटवितिब्र	va-i-cha-aa-ra-ta-ṭa-va-i-ta-i-ba-̖-ra	Modify Verb	Adverb	Verb Modifier	Modify action
99985	उपदशाजनअगुणदशा	u-pa-da-sha-aa-ja-na-a-ga-u-ṇa-da-sha-aa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99986	सुवितंत्रजनक	sa-u-va-i-ta-̇-ta-̖-ra-ja-na-ka	Express Action	Verb	Action Process	Action State
99987	गाउँगणतटसमूह	ga-aa-u-̐-ga-ṇa-ta-ṭa-sa-ma-uu-ha	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99988	अभावतटमण्डली	a-bha-aa-va-ta-ṭa-ma-ṇa-̖-ḍa-la-ii	Modify Verb	Adverb	Verb Modifier	Modify action
99989	संस्कृतिपात्रगण	sa-̇-sa-̖-ka-ri-ta-i-pa-aa-ta-̖-ra-ga-ṇa	Express Action	Verb	Action Process	Action State
99990	विउपयोगजनक	va-i-u-pa-ya-o-ga-ja-na-ka	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99991	उपजनकगणजगतहरू	u-pa-ja-na-ka-ga-ṇa-ja-ga-ta-ha-ra-uu	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99992	अभावजगतसुत्न	a-bha-aa-va-ja-ga-ta-sa-u-ta-̖-na	Express Action	Verb	Action Process	Action State
99993	अउपराजअहावा	a-u-pa-ra-aa-ja-a-ha-aa-va-aa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99994	विधरोहररुयो	va-i-dha-ra-o-ha-ra-ra-u-ya-o	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99995	वृक्षत्मकसमूह	va-ri-ka-̖-ṣha-ta-̖-ma-ka-sa-ma-uu-ha	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99996	उपत्मकबारे	u-pa-ta-̖-ma-ka-ba-aa-ra-e	Modify Verb	Adverb	Verb Modifier	Modify action
99997	सपनापतिविराजा	sa-pa-na-aa-pa-ta-i-va-i-ra-aa-ja-aa	Describe Quality	Adjective	Noun Modifier	Attribute Descriptor
99998	उपतटविद्याता	u-pa-ta-ṭa-va-i-da-̖-ya-aa-ta-aa	Name Entity	Noun	Subject Object	Entity Object
99999	बजारजगतसाध	ba-ja-aa-ra-ja-ga-ta-sa-aa-dha	Modify Verb	Adverb	Verb Modifier	Modify action
100000	सपनखगोलरुछ	sa-pa-na-kha-ga-o-la-ra-u-chha	Name Entity	Noun	Subject Object	Entity Object

Author	**YEAR OF PUBLICATION**	**METHODOLOGY / OBSERVATIONS**	**LIMITATIONS**
Pooja Rai, Sanjay Chatterji	December 2022 (Annotation Projection-based Dependency Parser Development for Nepali)	Developed the first Nepali dependency parser using cross-lingual annotation projection from a Bengali treebank. Constructed a Nepali-Bengali parallel corpus. Trained parsers using MaltParser and Neural network-based models. Achieved 81.2 UAS, 73.2 LA, and 66.1 LAS. Also built a CRF-based POS tagger.	Limited treebank size, cross-lingual projection dependency, performance variance with predicted POS tags, and absence of benchmarking with Universal Dependencies or multilingual parsers.
Shushanta Pudasaini, Sunil Ghimire, Prabhat Ale, Aman Shakya, Prakriti Paudel, Basanta Joshi	January 2024 (Application of Nepali Large Language Models to Improve Sentiment Analysis)	Leveraged open-source Nepali pre-trained Large Language Models (LLMs) for sentiment analysis on YouTube data. Extracted sentence embeddings and evaluated multiple models, achieving an F-score of 0.88. Released dataset and code publicly. Compared transfer learning approaches across machine/deep learning models.	Performance limited to YouTube domain; lacks generalizability to other domains like e-commerce. Domain-specific fine-tuning required. Challenges in capturing contextual sentiment. Ethical concerns related to bias, privacy, and misuse.
Pooja Rai, Sanjay Chatterji, Byung-Gyu Kim	August 2023 (Deep Learning-based Sequence Labeling Tools for Nepali)	Developed POS tagger and chunker for Nepali using BI-LSTM-CRF and other LSTM-based models with character and word embeddings. Achieved 99.20% and 98.40% accuracy respectively. Introduced the first statistical chunker for Nepali. Included a CRF-based baseline for optimal feature identification.	Focused only on shallow parsing (POS tagging and chunking); does not include higher-level syntactic or semantic parsing. Future scope includes integrating grammatical rules and building a full treebank/parser. Limited exploration of real-world deployment scenarios.
Samarjeet Borah, Upali Choden, Nermit Lepcha	2017 (Design of a Morph Analyzer for Non-Declinable Adjectives of Nepali Language)	Developed a Morph Analyzer (MA) for non-declinable adjectives in Nepali using a finite state grammar approach. The system transliterates Nepali to Roman script, tags adjectives based on suffixes/prefixes, and achieves 76% accuracy for prefix tagging and 63% for suffix tagging. Also creates a dictionary to prevent redundant tagging.	Focuses only on non-declinable adjectives; other nominals were not considered. Accuracy affected when multiple nouns appear around adjectives. Needs integration with a noun tagging module for better precision.

AUTHOR	YEAR OF PUBLICATION	METHODOLOGY / OBSERVATIONS	LIMITATIONS
Latesh Malik, Sonakshi Goyal, Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, Ondrej Krejcar	2024 (Devanagari Character Recognition: A Comprehensive Literature Review)	Provided an extensive review of handwritten Devanagari character recognition (HDCR). Covered the evolution from early template and feature-based methods to advanced machine learning and deep learning techniques (CNNs, RNNs). Highlighted hybrid approaches and efforts to develop annotated datasets. Offered a timeline and analysis of various techniques used in HDCR.	No prior dedicated review on HDCR was available. Challenges still include handwriting variability, dataset scarcity, noise, and achieving real-time performance.
Nobal B. Niraula, Saurab Dulal, Diwa Koirala	2022 (Linguistic Taboos and Euphemisms in Nepali)	Conducted a corpus-based study of offensive language in Nepali. Identified 18 categories of linguistic taboos and 12 types of euphemisms. Created a dataset of over 1,000 offensive terms from 7,000 social media posts. Baseline experiments were performed using rule-based and machine learning techniques for offensive language detection.	Offensiveness is socio-cultural and context-specific, making generalization difficult. Current models offer only baseline performance. Further work needed in ontology construction, fine-grained classification, and advanced ML techniques.
Manish K. Sharma, Bidhan Bhattarai	MLC 2017, February 24–26, 2017 (Optical Character Recognition System for Nepali Language Using ConvNet)	Developed a CNN-based OCR system using Keras and Theano. Utilized both real-world and synthesized datasets. Achieved 99.31% accuracy recognizing basic and compound characters. Implemented histogram-based localization and grayscale feature extraction.	Limited dataset, especially for compound characters. System may underperform with varied fonts or noisy data. Needs improved labeling, larger datasets, edge detection techniques, and language model integration for disambiguation.
Ashish Pradhan, Archit Yajnik	ISEEIE 2021, February 19–21, 2021 (Probabilistic and Neural Network Based POS Tagging of Ambiguous Nepali text: A Comparative Study)	Compared Hidden Markov Model (HMM) and General Regression Neural Network (GRNN) based POS taggers for Nepali. Used TDIL corpus and implemented using Python, Java, and NLTK. Achieved 100% accuracy for known words, ~60% for ambiguous ones, and 85.36% (GRNN) for unknown non-ambiguous words.	GRNN becomes unstable with training data >7000 words. HMM is more scalable but depends on tagged corpora. GRNN has slower runtime in Java. No mention of integration with context-aware models.

AUTHOR	**YEAR OF PUBLICATION**	**METHODOLOGY / OBSERVATIONS**	**LIMITATIONS**
Bal Krishna Bal	ACL-IJCNLP 2009, August 6–7, 2009 (Towards Building Advanced Natural Language Applications- An Overview of the Existing Primary Resources and Applications in Nepali)	Presented an overview of primary NLP resources and applications developed for Nepali. Highlighted approaches used in existing systems and their coverage. Emphasized the foundation laid for building advanced applications like SMT, NER, QA, IR, and IE.	Limited advancement beyond foundational resources. Many systems are still in early stages. Emphasis needed on refining existing tools and extending them to support more complex applications.
Mir Ragib Ishraq, Nitesh Khadka, Asif Mohammed Samir, M. Shahidur Rahman	ACM TALLIP, December 2021 (Towards Developing Uniform Lexicon Based Sorting Algorithm for Three Prominent Indo-Aryan Languages)	Developed a uniform lexicon-based sorting algorithm for Bengali, Nepali, and later Hindi by analyzing character-level similarities among these Indic languages. The algorithm was tested on 30,000+ words from each language and achieved 100% accuracy and good efficiency.	Focused primarily on sorting; other NLP applications like transliteration or categorization are only suggested as future work. Requires linguistic expertise from multiple languages to scale to broader Indic NLP tasks.
Bipesh Subedi, Prakash Poudyal	NLPIR 2022, December 16–18, 2022, Bangkok, Thailand (WordEmbedding in Nepali Language using Word2Vec)	Implemented Word2Vec using Gensim on Nepali health news data. Used CBOW and Skip-Gram models to generate word embeddings, showing promising semantic similarity between Nepali words. Data was scraped from five news portals.	Dataset limited to health news domain. Lacks stemming and other preprocessing steps that could improve performance. Needs more diverse and larger datasets for general applicability.
Ann Irvine, Chris Callison-Burch	2017 (Computational Linguistics, Vol. 43, No. 2) (A Comprehensive Analysis of Bilingual Lexicon Induction)	Analyzed bilingual lexicon induction across 25 languages including Nepali using various monolingual signals like contextual, topical, and orthographic similarity. Proposed a discriminative model that significantly outperforms traditional MRR and MCCA models. Achieved 42% accuracy vs. 15% by MCCA.	Performance still dependent on data quality and seed dictionary size. Gains diminish with larger corpora. Results most promising when Wikipedia data is used; less consistent with newswire data. Model is linear; future work could explore non-linear models.

References:

[1] P. Rai and S. Chatterji, "Annotation Projection-based Dependency Parser Development for Nepali," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 2, Art. no. 36, pp. 1–19, Dec. 2022, doi: 10.1145/3542696.

[2] S. Pudasaini, S. Ghimire, P. Ale, A. Shakya, P. Paudel, and B. Joshi, "Application of Nepali Large Language Models to Improve Sentiment Analysis," in *Proc. 2024 7th Int. Conf. on Computers in Management and Business (ICCMB)*, Singapore, Jan. 2024, pp. 1–7, doi: 10.1145/3647782.3647804.

[3] P. Rai, S. Chatterji, and B.-G. Kim, "Deep Learning-based Sequence Labeling Tools for Nepali," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 8, Art. no. 211, pp. 1–23, Aug. 2023, doi: 10.1145/3606696.

[4] S. Borah, U. Choden, and N. Lepcha, "Design of a Morph Analyzer for Non-Declinable Adjectives of Nepali Language," in *Proc. 2017 Int. Conf. on Machine Learning and Soft Computing (ICMLSC)*, Ho Chi Minh City, Vietnam, Jan. 2017, pp. 1–6, doi: 10.1145/3036290.3036307.

[5] L. Malik, S. Goyal, S. Arora, D. Bhattacharjee, M. Nasipuri, and O. Krejcar, "Devanagari Character Recognition: A Comprehensive Literature Review," *IEEE Access*, vol. 12, pp. 1–20, Jan. 2025, doi: 10.1109/ACCESS.2024.3520248.

[6] N. B. Niraula, S. Dulal, and D. Koirala, "Linguistic Taboos and Euphemisms in Nepali," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 6, Art. no. 129, pp. 1–26, Nov. 2022, doi: 10.1145/3524111.

[7] M. K. Sharma and B. Bhattarai, "Optical Character Recognition System for Nepali Language Using ConvNet," in *Proc. 2017 Int. Conf. on Machine Learning and Computing (ICMLC)*, Singapore, Feb. 24–26, 2017, pp. –, doi: 10.1145/3055635.3056635.

[8] A. Pradhan and A. Yajnik, "Probabilistic and Neural Network Based POS Tagging of Ambiguous Nepali Text: A Comparative Study," in *Proc. 2021 Int. Symp. on Electrical, Electronics and Information Engineering (ISEEIE)*, Seoul, South Korea, Feb. 19–21, 2021, pp. –, doi: 10.1145/3459104.3459146.

[9] B. K. Bal, "Towards Building Advanced Natural Language Applications: An Overview of the Existing Primary Resources and Applications in Nepali," in *Proc. 7th Workshop on Asian Language Resources, ACL-IJCNLP*, Suntec, Singapore, Aug. 6–7, 2009, pp. 165–170.

[10] M. R. Ishraq, N. Khadka, A. M. Samir, and M. S. Rahman, "Towards Developing Uniform Lexicon Based Sorting Algorithm for Three Prominent Indo-Aryan Languages," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, no. 3, Art. 57, Dec. 2021, pp. 20. https://doi.org/10.1145/3488371.

[11] B. Subedi and P. Poudyal, "Word Embedding in Nepali Language using Word2Vec," in *2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2022)*, Bangkok, Thailand, Dec. 16-18, 2022. ACM, New York, NY, USA, pp. 5. https://doi.org/10.1145/3582768.3582799.

[12] A. Irvine and C. Callison-Burch, "A Comprehensive Analysis of Bilingual Lexicon Induction," *Computational Linguistics*, vol. 43, no. 1, pp. 1-34, 2017, doi: 10.1162/COLI_a_00284.