Corpus preparation for low-resource languages and for development of human language technology to analyze or computationally process them is a laborious task, primarily due to the unavailability of expert linguists who are native speakers of these languages and also due to the time and resources required. Bhojpuri, Magahi, and Maithili, languages of the Purvanchal region of India (in the north-eastern parts), are low-resource languages belonging to the Indo-Aryan (or Indic) family. They are closely related to Hindi, which is a relatively high-resource language, which is why we compare them with Hindi. We collected corpora for these three languages from various sources and cleaned them to the extent possible, without changing the data in them. The text belongs to different domains and genres. We calculated some basic statistical measures for these corpora at character, word, syllable, and morpheme levels. These corpora were also annotated with parts-of-speech (POS) and chunk tags. The basic statistical measures were both absolute and relative and were expected to indicate linguistic properties, such as morphological, lexical, phonological, and syntactic complexities (or richness). The results were compared with a standard Hindi corpus. For most of the measures, we tried to match the corpus size across the languages to avoid the effect of corpus size, but in some cases it turned out that using the full corpus was better, even if sizes were very different. Although the results are not very clear, we tried to draw some conclusions about the languages and the corpora. For POS tagging and chunking, the BIS tagset was used to manually annotate the data. The POS-tagged data sizes are 16,067, 14,669, and 12,310 sentences, respectively, for Bhojpuri, Magahi, and Maithili. The sizes for chunking are 9,695 and 1,954 sentences for Bhojpuri and Maithili, respectively. The inter-annotator agreement for these annotations, using Cohen’s Kappa, was 0.92, 0.64, and 0.74, respectively, for the three languages. These (annotated) corpora have been used for developing preliminary automated tools, which include POS tagger, Chunker, and Language Identifier. We have also developed the Bilingual dictionary (Purvanchal languages to Hindi) and a Synset (that can be integrated later in the Indo-WordNet) as additional resources. The main contribution of the work is the creation of basic resources for facilitating further language processing research for these languages, providing some quantitative measures about them and their similarities among themselves and with Hindi. For similarities, we use a somewhat novel measure of language similarity based on an n-gram-based language identification algorithm. An additional contribution is providing baselines for three basic NLP applications (POS tagging, chunking, and language identification) for these closely related languages.
1 INTRODUCTION

Plain text corpus, from the point of view of Computational Linguistics (CL) or Natural Language Processing (NLP), is a systematic collection of text that should have certain properties like representativeness, correctness (if it is for standard written language), and sufficient quantity. For languages on which no NLP work has been performed before, or very little such work exists, creating a good plain text corpus is the foundation for building NLP resources and applications or tools for that language. Plain text corpus is basically a collection of written text or transcribed speeches of a particular natural language stored in electronic or digital form. Linguists and NLP researchers use corpora for statistical analysis, hypothesis testing, word-frequency analysis, and validating linguistic knowledge in a language territory or domain. Once this corpus gets annotation with richer linguistic knowledge about its structure, meaning, usage, and so on, it can act as training data for supervised learning for building tools that solve various NLP problems, such as Part-of-Speech (POS) tagging, chunking, morphological analysis, parsing, and so on. In this article, we describe our efforts to prepare some plain text corpus for three languages of the Purvanchal region of central north-eastern parts of India and to annotate this corpus with POS and chunk tags. These languages are Bhojpuri, Maithili, and Magahi.

These languages are low-resource languages and are phylogenetically related, since they belong to the same (Indo-Aryan, or Indic) language family and also mainly occur in a geographically contiguous area. Globally, these are the first-languages of 50, 12, and 30 million speakers, respectively. The availability of an enormous base of speakers is a good reason for researchers to understand and analyze different properties of these languages and to build language technologies for them. Kumar et al. [40], Singh and Jha [81] began a project by collecting corpora of 1 million words for the Bhojpuri and Magahi languages, of which 89,997 and 63,000 words have been annotated at the POS level, respectively. Simultaneously, at our end, we have collected corpora for the Bhojpuri, Magahi, and Maithili, consisting of 704,543, 424,434, and 267,881 words. We have also carried out manual POS annotation on corpus sizes of 245,482, 171,509, and 208,640 words, respectively, and the chunk-annotated corpus sizes of 60,588, 0, and 10,476 words, respectively, as chunk annotation for Maithili could not be carried out.

To the best of our knowledge, our effort, which began in 2014, was one of the first to create annotated resources for these languages. At that time, there were hardly any publicly available sources of Maithili or Bhojpuri text corpus, and it was the first attempt to build the corpora and annotate them for POS tags and chunks. The project also involved building language processing tools for these languages and, in the end, to build machine translations systems for these languages.
These linguistically annotated basic resources can serve as the base for further developing language technology and tools for these languages such as POS tagger, Chunker, Morphological Analyzer, Morphological Generator, Machine Translation, Cognate Identification, and so on, can be carried out by researchers. This article describes these basic resources for three Purvanchal languages and some basic statistical details about them at the word, the morpheme, the syllable, and the character levels, and their similarities among themselves and with Hindi. These details include some simple measures of complexity and similarity, including a slightly novel similarity measure. Apart from this, we also highlight some issues encountered during the Part-of-Speech and Chunking annotations for all these languages. We have also tried to pay some special attention to the similarities among these languages, as their differences are still debated. The inter-annotator agreement has been calculated through Cohen’s Kappa (Cohen [17]). The (annotated) corpora have been employed to build the POS taggers, Chunkers, and a Language identifier for the three languages.

2 RELATED WORK

India, as is well known, is a multilingual country, with 22 scheduled (constitutionally recognised) languages, written in (at least) 13 different extant scripts and several other variations of them. Apart from the scheduled languages, there are at least 400 plus other languages. Most of the developed countries, especially the ones with a not so diverse language base compared to India, do possess relatively more abundant, easily accessible, and most importantly, good quality corpora. However, Indian languages have relatively little corpora for a large number of languages, even though the degree of resource-scarceness varies. This scarcity severely constraints the language technology development and application of state-of-the-art algorithms, particularly deep learning approaches.

Technology Development for Indian Languages (TDIL)² (Dash [21]), a Government of India department under the Ministry of Electronic and Information Technology, has been actively promoting work on language resource and technology development for Indian languages, particularly on the scheduled languages. As an example of work done as part of this initiative, machine-readable corpora of around 10M tokens of Indian languages was prepared. The languages include English, Hindi, Punjabi, Telugu, Kannada, Tamil, Malayalam, Marathi, Gujarati, Oriya, Bangla, Assamese, Urdu, Sindhi, Kashmiri, and Sanskrit. Of these, Urdu, Kashmiri, and Sindhi use Perso-Arabic scripts for their writing systems, while the remaining languages use scripts derived from the ancient Brahmi script. All these languages belong to either the Dravidian and or the Indo-Aryan families. The preliminary analysis of some of their initial corpora is described in Bharati et al. [7]. The resources and tools or technologies developed for Indian languages and shared with the TDIL are distributed through the Indian Languages Linguistic Data Consortium (IL-LDC).

Effort was made by various research centres working with TDIL to create representative corpora for Indian languages and basic analysis of such corpora was published. For example, the Bangla corpus, that was compiled from various disciplines such as literature, mass communication, science, and so on, and was initially subjected to statistical analysis by Chaudhuri and Ghosh (Chaudhuri and Ghosh [15]) at word, character, and sentence levels on the monolingual corpus comprising 2.5 million words from the Anusaaraka group ⁴ and 150,000 words from their own laboratory. A further collection of the corpus from the Bangla newspaper website (Prothom-Alo) was crawled, considering the balance and representativeness of the corpus. Its comparison with the earlier corpus (Majumder and Arafat [48]) was also reported. In addition to this, to boost the

²http://www.tdil.meity.gov.in.
³http://www.ildc.in/.
⁴https://ltrc.iiit.ac.in/projects/Anusaaraka/anu_home.html.
language technology development for Bangla, Bangla-translated Basic Travel Expression Corpus (BTEC) (Khan et al. [36]) was compiled to develop an English-Bangla speech translation system in the travel domain.

Another corpora analysis under the TDIL program was for the Telugu language, which is a morphologically rich language, much more so than Bhojpuri, Maithili, and Magahi. It is also one of the most spoken languages of India (ranking third in terms of its speakers), and mainly based geographically in the southern states of Andhra Pradesh and Telangana, India. Its characteristics were analyzed using statistical tools on the compiled monolingual corpus of Telugu comprising 39 million words (Kumar et al. [38]). Moreover, lexical categories produced from the morphological analyzer were used as parts of speech due to morphological richness of Telugu language (Badugu [4]).

After Anusaaraka and several other individual machine translation systems, Indian language to Indian language machine translation (IL-ILMT) project was initiated to work on bidirectional machine translation between nine Indian languages translation. In this consortium-based project, 11 Indian institutes participated and designed a system called Sampark. This system was developed by building the required “horizontal” and “vertical” resources, tools and other components. This is a transfer-based hybrid machine translation system that includes knowledge-based and statistical modules. So far, it has publicly made available Hindi↔Urdu, Hindi↔Punjabi, Hindi↔Telugu, Hindi↔Tamil, Hindi↔Bengali, Hindi↔Marathi, Hindi↔Kannada, Tamil↔Telugu, and Tamil↔Malayalam machine translation systems. Our continuing work on building translation systems for languages of the Purvanchal languages can be considered an extension of this project.

Manipuri (Meitei) was added a few years ago to the list of scheduled languages and belongs to the Sino-Tibetan language family. A recent attempt (Singh et al. [79]) was made to create corpora from two local Manipuri newspapers, written in the Bangla script, and it was made compatible with Unicode through manual correction from Bangla OCR. Similar to Manipuri speech corpora (Jha [34]), the collection of data for Automatic Speech Recognition (ASR) from native speaker(s) and the telephonic speech data transcription was carried out.

Over the years, NLP and CL community has continued to work on the development of resources for low resource Indian languages (although work on a few languages like Hindi has been going on for more than two decades) and use them for several language-based applications (Jha [34], Kumar et al. [39], Srivastava et al. [83]). With the growing use of the Internet in the age of digitization, speech technology (Mohan et al. [54]) also came into the picture for Indian languages, and played a crucial role in the development of applications in various domains such as agriculture, healthcare, and other services for ordinary people (Kumar et al. [39]). Researchers outside India have also contributed to building corpora for Indian languages and one such example is the HindEnCorp, a monolingual corpus for the Hindi language and a parallel corpus of Hindi-English comprising 787M and 3.8M tokens, respectively (Bojar et al. [11]).

For the three Purvanchal languages, the first attempt to collect monolingual Bhojpuri corpus was made through a web crawler that includes dialects, different genres, and domains of Bhojpuri (Singh [80]). Further, this corpus is annotated for POS tagging, and a statistical tagger is developed, which allows us to compare with two other Indo-Aryan languages: Hindi and Oriya (Ojha et al. [58], Singh and Jha [81]). A similar attempt (Kumar et al. [40]) was made to collect a monolingual corpus of Magahi language from blogs, stories, magazines, and transcripts of spontaneous

\[http://ilmt.tdil-dc.gov.in/sampark/web/index.php/content\]
\[https://sampark.iiit.ac.in\]
\[https://ufal.mff.cuni.cz/hindencorp\].

ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 20, No. 6, Article 95. Publication date: September 2021.
conversations carried out in public, private, and home domains, which made it possible to prepare a corpus of around 1 million tokens. POS annotation used BIS standard tagset. The tagset consists of 11 coarse-grained tags and around 30 fine-grained tags for each language that it covers. This standard tagset was slightly modified and tested on around 50,000 tokens (with fine-grained tags) with TnT, Memory-based tagger, Maximum-entropy tagger, HMM-based tagger, SVM-based tagger by Kumar et al. [41]. Maithili corpus from web resources and Wikipedia dump was used toward developing a POS tagger, using CRF, on 52,190 words with an accuracy of 82.67% (Priyadarshi and Saha [62]). Apart from this, HMM model was also employed on 10K words with an accuracy of 95.67% (Rishikesh [65]).

3 ABOUT BHOJPURI, MAITHILI, AND MAGAHI

Bhojpuri belongs to the Indo-Aryan (Indic) language family. Its origin is often attributed to a place in Bihar called Ara/Bhojpur. According to the census of India (2011) data, approximately 50,579,447 speakers use it as their first language. Ironically, a language with such a massive user base is yet to have official recognition and attain the status of a scheduled (or officially recognised) language of India, because the government agencies classify Bhojpuri as a dialect/regional variety of Hindi, in particular, of “Eastern Hindi.” The resultant lack of recognition for this language, coupled with low employment opportunities, owing to scarce resources for teaching and learning technology, has brought tremendous disadvantage to Bhojpuri and its speech community. Setting aside the classifications based partially on sociopolitical considerations, it is justifiable to treat Bhojpuri as a living language that has a significant user base, and that is spoken in several countries besides India, including Fiji, Nepal, Surinam, and Mauritius. Moreover, its speakers would outnumber speakers of several officially recognized languages and dialects of India, and also of several European languages. Across India, they are spread with high concentrations in the states of Uttar Pradesh, Bihar, Jharkhand, and Chattisgarh.

Magahi became a distinct language around the tenth century, similar to other New Indo-Aryan (NIA) languages. Grierson classified Magahi under the Eastern group of the outer sub-branch. Scholars like Turner have clubbed the Bihari languages with Eastern and Western Hindi (Masica [51]). Chatterji [13] gave an entirely different classification, where Western Hindi is almost an isolated group, while Eastern Hindi, Bihari, and other languages are clubbed together in the Eastern group. Nevertheless, the proposition, given above, by Grierson [28] is the most widely accepted. Jeffers [33] and others have proposed a similar categorization.

Although Magahi is frequently mislabelled as a dialect of Hindi, the fact is that it is not even part of the same sub-family as Hindi. It was developed from the Maagadhi/Eastern Apbhranksha branch of Maagadhi Prakrit. Other languages developing from this branch are Maithili, Bhojpuri, Bangla, Assamese, and Oriya. This deduction is validated using the Grierson’s classification and is strengthened even via grammars of Hindi and Magahi. According to the features of the languages, Bangla, Oriya, Assamese, Maithili, Bhojpuri, and Magahi are close cognate sister languages, which means that Bhojpuri and Magahi are sister languages. Consequently, Magahi shares many characteristics with the Bhojpuri language, such as a massive user base of 20.7 million speakers. Maithili is their first cousin, and languages like Bangla, Assamese, and Oriya are their second cousins (Verma [87]). Contemporarily, Magahi is spoken chiefly in the states of Bihar, Jharkhand, West Bengal, and Orissa.

Maithili serves as the first language to approximately 30 million speakers, who live in the Indian state of Bihar and South-eastern region of Nepal. It is a highly dialect-oriented language having several variations due to social factors such as gender, education level, caste, and so on. It shows a
relatively stronger influence of Sanskrit and, further, has at least ten varieties in India and at least three regional dialects in Nepal. The social status shows the honorific agreement in this language (Yadava et al. [92]).

**Linguistic Features**

Bhojpuri, Maithili, and Magahi languages have Subject-Object-Verb (SOV) word order. These languages are written in Devanagari and Kaithi scripts. Like many other modern Indo-Aryan languages these languages are also non-tonal languages. Word-formation in these languages is somewhere between synthetic and analytical typology. These languages, unlike Hindi, lack ergativity. More details about linguistic features of these languages are discussed in the section on annotation issues.

Based on the linguistic features, Bengali, Oriya, Assamese, Maithili, Bhojpuri, and Magahi are close cognate languages, which means that Bhojpuri and Magahi are sister languages and Maithili is their first cousin. Languages like Bengali, Assamese, and Oriya are their second cousins. Hindi is a distant relative, even though prolonged close contact (with Hindi being the official language of the India and also of the two states where these languages are spoken the most) has made things a bit cloudy. Some examples of the differences are as follows:

- **Non-ergative Construction:** Standard Hindi is an Ergative language. In Hindi when the subject is marked with [ne] (oblique case) then the transitive verb agrees with the object in terms of person, number and gender, whereas Bhojpuri, Magahi, and Maithili are non-ergative languages. For example:
  
  **Bhojpuri:** mohan kiwAb paDZalas.
  **Hindi:** mohan ne kiwAb paDZI.
  **English Translation:** Mohan read the book.

- **Emphatic Expressions:** In these languages emphatic particles (like certain case markers) are generally merged with nominals and pronominals, whereas in Hindi, we have separate emphatic particles such as [BI], [hI]. For example:
  
  **Bhojpuri:** laikav[o] Ayala rahe.
  **Hindi:** ladZakA [BI] AyA WA.
  **English Translation:** The boy [too] had come.

In this example sentence of Bhojpuri, the emphatic particle [o] is fused with the nominal to show emphasis.

- **Classifiers:** Like Bengali and Oriya, Bhojpuri, Magahi, and Maithili languages also very often use classifiers with numerals. Classifier markers in these languages are: To, Te, go, Ke, goTa, and so on, whereas Hindi and its other dialects do not show this characteristic. For example:
Maithili: eka [goTa] sajjana.
Hindi: eka [human-classifier] sajjana.
English Translation: One [human-classifier] gentlemen.

- **Determiners**: Wide use of determiners is again a unique feature of Bhojpuri, Maithili, and Magahi languages. In these languages determiners can occur with almost all the proper and common nouns and similar to the emphatic markers, determiners also get fused with nominal forms. This constructional feature marks another point of contrast from Hindi.

In Bhojpuri, Maithili, and Magahi languages a word ending with [-a] or [-A] sound will take [-vA] suffix, [-i/-I] will take [-yA] suffix, and [-u] will take an [-A] suffix.

Examples:
Bhojpuri: [XiyavA mAI se kahalas]
Hindi: [beTI ne mAz se kahA]
English Translation: Daughter told the mother.

Bhojpuri, Maithili, and Magahi are partially synthetic languages, as can be seen from the above examples. Morphology of these languages differs significantly from the morphology of Hindi in some aspects. One way to perform a preliminary check about the idiosyncrasies of a language is by calculating or applying some statistical measures on the corpus of a language. Before describing this part, we describe text corpus creation for the three languages in the next section.

4 TEXT CORPUS CREATION OF PURVANCHAL LANGUAGE: BHOJPURI, MAITHILI, AND MAGAHI

For the project on which this study is based, we collected data from web pages and made a collection of corpus under the supervision of linguists and speakers of these languages. These were qualified linguists and native speakers of these languages, because of which they were also familiar with the writing systems used for these languages. This last part about writing is only partially true, because none of these languages is mediums of education anywhere in India and very few people actually read or write in these languages, even though they are spoken by a large number of people. It is partially true, because all these speakers of these three languages are also native speakers of Hindi, which uses the Devanagari script. These speakers often identify themselves as Hindi speakers, because Hindi is their main first language, not only for communication but also for education and official purposes, apart from English, which is still the de-facto first official language of India (although constitutionally, English is the second official language, after Hindi).9 None of these speakers, with rare exceptions, regularly reads or writes in these languages.

We collected raw text in Bhojpuri, Magahi, and Maithili from online newspapers, magazines, and other relevant resources. The collected data was made balanced and representative to some extent by including varied genres such as sports, editorial, politics, science fiction, entertainment, education, and daily news. In magazines, we preferred [pAwI]10 for Bhojpuri and [axrXaga] novel for Maithili. While building the corpora, we avoided data from blogs so that we might be able to maintain the linguistic authenticity of the sources to the Bhojpuri and Maithili languages. Blog posts in these languages often use a high degree of code-mixing. However, significant deficit of

9https://rajbhasha.nic.in/en/constitutional-provisions.
10The items within [] brackets indicate WX notation. The WX (https://en.wikipedia.org/wiki/WX_notation) notation is used throughout this article, as is the usual practice for Indian languages for computational purposes.
data sources for the Magahi language compelled us to obtain its data from various blogs to at least establish an approximate comparison while making an analysis. Most of sources of corpora on the Web that we used are listed in Appendix A.

4.1 Raw Corpus Cleaning and Issues

After extracting the corpus from the Web and other sources, we had to clean it in the usual way (such as removing of unwanted characters, symbols) but also, in some cases, to convert the encoding into Unicode (UTF-8). This is because for most Indian languages, Unicode is still not universally used. For the three Purvanchal languages, many sources, particularly those from printed material, were written using a proprietary encoding that has been popular for Desktop Publishing in Devanagari. Since the encoding converter is not perfect (as the mapping between this encoding and Unicode is not straightforward), there were some errors in conversion to Unicode. Because of this, some ‘noise’ is present in the text corpus like extra punctuation, invalid ASCII characters, foreign languages words, misspelled words, and so on. Part of the cleaning process involved removing such noise as far as possible, given the time and resource constraints. We also removed the code mixed sentences from the crawled data, only excluding sentences with Hindi nouns, after verifying with a language-specific linguist. There are still some errors left that were introduced due to encoding conversion.

Since there were not many significant lexical or structural mistakes, so the corpus cleaning process, on the whole, was not very time-consuming, nor could we afford to spend too much time on it. Part of the problem was the lack of standardisation and the confusability among the three language. We have described this in more detail in the section on language identification. There were also time and resource constraints. We found several minor orthographic issues in the collected corpora of these three languages, such as for [Sa, Ra]. An example of words where such problems were found is [akCara → akRara] in Maithili. The problem in these cases is that the orthographic practices of Hindi are carried over to the concerned languages, causing mistakes, e.g., [akCara] is written as [akSar], which is the Hindi spelling for this word, roughly meaning orthographic syllable. Another common problem is due to the diacritic called Avagrah. Avagrah is frequently used in Bhojpuri, Magahi, and Maithili, and it indicates lengthening of the preceding vowel for our purposes. In many cases, this diacritic was missing, which can change the meaning of a word in these languages. We tried to introduce it in places where it was missing, but this process may not have been rigorous due to the lack of available annotator-time.

Since these languages (previously considered dialects) do not have much standardization, the words are not spelled consistently, adding to the problem of language processing. A word can have multiple spelling variations. Some examples of problems due to these reasons are given below:

- **Orthographic variations:** As mentioned above, words having the same meaning can often be spelled in more than one ways, i.e., there is a lot of orthographic variation. This is due to the lack of standardization for the three languages. This problem is present, to a much lesser extent, even in Hindi. For example, consider one form of variants formed by replacing Anusvara with Anunasik or Anusvara with half-nasal consonants (such as in [samvawa/saMvawa], [samBala/saMBala/saBala]. Examples of other orthographic variants may include [karie/kariye (e/ye)], [karAyI/karAI, uTAyI/uTAI (I/yI)], [bAyI/bAyIZ], [karyavAI/karyavAhI/kAravAhI], [BEyA/BaiA/BaiyA], [vAyasarAya/vAisarAya], and [suwala/sowala].

- **Particle identification:** Particles are functional words that do not have very clear meanings, even as function words. Examples of such particles include [vA, bA] and [mA]
constructions. Bhojpuri and Magahi have two types of roots: one without [vA] and another attached with [vA], e.g., [laikA] (boy) and [lalkavA] occur in masculine and non-masculine words, but in feminine words, [yA] suffix is present. Maithili has a similar type of construction, however, it uses [mA] and [bA] suffixes; for example, [CoramA] (boy) and [CorabA]. It is debatable how these suffixes change the meanings. Strictly speaking, these are suffixes and are not words, so they should not be called particles in the sense Hindi has particles such as [hlI] or [wo]. Still, they pose problems while annotating the corpus.

- Incorrectly spelled words or foreign language words: As some words may not be correct or borrowed from any different language than the chosen three languages for analyses, they are often not be spelled correctly (or consistently) as there is no standard spelling. Standard spelling is a general problem for all three Purvanchal languages even for their native words, but it is particularly problematic for borrowed words. For the subset of data that was annotated, we tried to filter out such problematic words to some extent.

- Numbers, Abbreviations, Symbols: We also had to consider the use of alternative written forms for words such as [SrI, dA.], symbols (*, & ), and numbers (3224.34).

4.2 Frequency Distribution Analysis

Statistical analyses of the corpora can help us understand the language characteristics. In quantitative analysis, we classify different linguistic properties or features of a particular language, count them, and construct more complex statistical models in an attempt to explain what is observed. It allows us to discover and differentiate the genuine from the improbable reflections of phenomena capable of describing the behaviors of a language or a variety of languages. However, the quantitative analysis aims at providing a complete and detailed description of the observed phenomena obtained from the corpus.

It may be of additional interest to the reader that the work described by Kumar et al. [38] involved corpora analyses on many other Indian languages, such as Hindi, Telugu, and so on, one or all of them could be used for comparison with the three primary languages of interest in this article, i.e., Bhojpuri, Maithili, and Magahi. In our work, wherever a comparison is useful and possible, we have augmented our analysis in the following sections by comparing the measures and metrics with the Hindi language, similar to the work by Kumar et al. [38]. A monolingual Hindi corpus11 (Europarl v7), has been used for this comparison.

4.2.1 Unigram. From the corpus, the initial statistical analysis in terms of the most basic units of language can be performed, e.g., word frequency, vocabulary size, and so on. We extracted the total number of sentences, tokens, types, and rare types. Here, the type can be defined as the vocabulary or lexical variety of the language expressed as a number of distinct (unique) tokens available in the corpus. Further, the rare types are those tokens that occur only once throughout the corpus, i.e., with frequency 1. The relationship between the types and tokens is sometimes determined by calculating the type-token ratio (TTR) that shows how often new word forms appear in the corpus (Sarkar et al. [70]). Since the TTR is sensitive to the corpus size, it has further been improved to avoid the effect of corpus size by Covington and McFall [19]. This better measure, namely, moving-average type-token ratio (MATTR), which is less dependent on the sentence length or corpus size, since it works on a sliding window for a fixed length. We have performed the

11 http://www.cfilt.iitb.ac.in/Downloads.html.
Table 1. Unigram Analysis of Purvanchal Languages Compared with Hindi

| Language   | Total Sentences | Total Tokens | Total Types | Rare Types | TTR    | MATTR  | TTR*   |
|------------|-----------------|--------------|-------------|------------|--------|--------|--------|
| Bhojpuri   | 51,374          | 704,543      | 38,680      | 178        | 5.49   | 0.572  | 9.71   |
| Maithili   | 19,948          | 267,881      | 27,043      | 10,597     | 10.09  | 0.587  | 10.09  |
| Magahi     | 31,251          | 424,434      | 35,253      | 12,978     | 8.30   | 0.575  | 9.09   |
| Hindi      | 221,768         | 2,917,271    | 151,516     | 89,436     | 5.19   | 0.516  | 11.40  |

Here * Indicates the TTR ratio over the number of sentences according to the Maithili corpus, which was the smallest. TTR is calculated on this (minimum) corpus size for all four languages, to observe the effect of corpus size on TTR.

Table 2. Percentage of Types Covering \( n \) Percent of the Corpus Where \( n \) Can Be 50, 60, 70, 80, and 90

| Language | 50% | 50%* | 60% | 60%* | 70% | 70%* | 80% | 80%* | 90% | 90%* |
|----------|-----|------|-----|------|-----|------|-----|------|-----|------|
| Bhojpuri | 60.15 | 89.71 | 70.49 | 99.88 | 79.36 | 99.89 | 85.25 | 99.97 | 88.47 | 99.99 |
| Maithili | 66.79 | 66.79 | 76.82 | 76.82 | 85.13 | 85.13 | 89.49 | 89.49 | 93.61 | 93.61 |
| Magahi   | 71.49 | 56.79 | 74.76 | 62.83 | 80.6 | 70.81 | 86.64 | 81.92 | 93.09 | 91.69 |
| Hindi    | 52.4 | 60.27 | 61.67 | 68.46 | 69.15 | 76.06 | 77.44 | 84.41 | 92.49 | 90.56 |

Here * indicates the type coverage over the percent of the corpus according to the Maithili corpus, which was the smallest.

MATTR experiments with 500 window size (throughout this article), as suggested by Covington and McFall [19].

Intuitively, MATTR also roughly indicates the vocabulary size, and therefore, the morphological complexity or richness. This is partly because we are counting over word forms and not lexemes or "root words." We can examine the relationship between the lexical or morphological richness and the TTR and the MATTR values: high TTR and MATTR meaning high lexical/morphological richness, and vice-versa. With reference to these terms in the context of the corpus that we created of three Purvanchal languages, the Table 1 lists the total the number of sentences, tokens, types, rare types, and also the TTR and MATTR corresponding to each language.

From the Table 1, we find that the value of TTR and MATTR comes out to be highest for the Maithili language among the three languages. This indicates that Maithili and Hindi are the most and the least lexically/morphologically rich, respectively, among the four languages. However, a look at the second column tells us that the number of sentences for the Maithili language is the least and quite less in comparison to the remaining three languages. Thus, only from Table 1, we cannot rule out the possibility that the value of TTR (if not MATTR) could have been higher had we created a larger corpus of the Maithili language, comparable to the other three languages.

While this is plausible, the previously available linguistic knowledge about the complexities or richness of the four languages at various linguistic levels does not agree with this observation, based on the last column. This knowledge agrees more with the TTR (without the same corpus size restriction), and the MATTR column. This could be interpreted in two ways. Either that the corpus size should not be the same for all languages for the TTR to be compared for lexical/morphological richness, or that the quantitative approach has some limitations, because intuitively the same corpus size restriction seems reasonable for quantitative analysis. However, MATTR values agree with the linguistic knowledge.

To further explore this inference, we come to our second step of statistical analysis, provided in Table 2, that is, the percentage of types required to cover a certain percentage \( (n) \) of the random
Table 3. The Top Ten Most Frequent Words for the Three Languages in Decreasing Order of Their Frequency (Freq) with Their Relative Frequency (RF) and Cumulative Coverage (CC) Values Compared with Hindi

| Bhojpuri   | Freq | RF  | CC  |
|------------|------|-----|-----|
| [ke]       | 49,700 | 0.070 | 0.070 |
| [meM]      | 14,893 | 0.021 | 0.091 |
| [A]        | 12,990 | 0.018 | 0.109 |
| [se]       | 12,893 | 0.018 | 0.127 |
| [bA]       | 9,386  | 0.013 | 0.140 |
| [kA]       | 8,279  | 0.011 | 0.151 |
| [nA]       | 7,304  | 0.010 | 0.161 |
| [wa]       | 7,256  | 0.010 | 0.171 |
| [ki]       | 5,849  | 0.008 | 0.179 |
| [rahe]     | 5,839  | 0.008 | 0.187 |

| Maithili   | Freq | RF  | CC  |
|------------|------|-----|-----|
| [aCi]      | 7,359 | 0.027 | 0.027 |
| [saBa]     | 4,021 | 0.015 | 0.042 |
| [je]       | 3,797  | 0.014 | 0.056 |
| [nahi]     | 3,318  | 0.012 | 0.068 |
| [o]        | 2,685  | 0.011 | 0.078 |
| [ka]       | 2,281  | 0.008 | 0.086 |
| [me]       | 2,270  | 0.008 | 0.094 |
| [lela]     | 2,128  | 0.007 | 0.101 |
| [gela]     | 2,014  | 0.007 | 0.108 |
| [I]        | 1,906  | 0.007 | 0.115 |

| Magahi     | Freq | RF  | CC  |
|------------|------|-----|-----|
| [ke]       | 29,977 | 0.07 | 0.07 |
| [meM]      | 10,050 | 0.023 | 0.093 |
| [he]       | 7,603  | 0.017 | 0.11 |
| [se]       | 7,521  | 0.017 | 0.127 |
| [hala]     | 6,465  | 0.015 | 0.142 |
| [Au]       | 5,651  | 0.013 | 0.155 |
| [para]     | 3,998  | 0.009 | 0.164 |
| [gela]     | 3,598  | 0.008 | 0.172 |
| [ki]       | 2,591  | 0.006 | 0.178 |
| [I]        | 2,585  | 0.006 | 0.184 |

| Hindi      | Freq | RF  | CC  |
|------------|------|-----|-----|
| [ke]       | 102,401 | 0.0351 | 0.0351 |
| [hE]       | 87,065 | 0.0298 | 0.0649 |
| [meM]      | 80,276 | 0.0275 | 0.0924 |
| [kl]       | 66,900 | 0.0229 | 0.1153 |
| [se]       | 48,278 | 0.0165 | 0.1481 |
| [Ora]      | 47,449 | 0.0163 | 0.1481 |
| [kA]       | 47,017 | 0.0161 | 0.1642 |
| [ko]       | 40,774 | 0.0141 | 0.1782 |
| [para]     | 26,402 | 0.0091 | 0.1977 |
| [BI]       | 25,405 | 0.0087 | 0.2064 |

corpus, where n can be 50, 60, 70, 80, and 90 (100% corpus coverage is already done in Table 1). For instance, the value in the third row and fourth column shows a value of 74.76, which means that of a total of 35,253 types available in the entire corpus for the Magahi language, only 74.76% types (or 26,355 types approximately) are available in the 60% tokens (or 254,660 tokens nearly) chosen randomly from the entire number of tokens (424,434) in the corpus. We have closely followed settings of Bharati et al. [7]. For all values of n, it can be observed that the percentage of corpus coverage is highest for Bhojpuri. Similarly, except for the case when Magahi exceeds Maithili in percentage coverage of types, Maithili consistently exhibits the highest percentage coverage of the types. These results seem to contradict the conclusion from the second last column of Table 2 (TTR on full corpus size) as well as previous linguistic knowledge. However, these results in Table 2 are on the same percentages, not on the same corpus sizes in terms of the number of tokens. Since different languages have different degrees of lexical richness, they will need different amounts of corpora to have the same coverage of types. Therefore, it seems better to compare the MATTR in the second last column of Table 1, rather than the last column.

Taking this analysis further, we summarize in Table 3 the Cumulative Coverage, the Frequency, and the Relative Frequency (or probability) of the occurrence of tokens for a particular language. This table shows the top ten most frequent tokens in the three languages in decreasing order of
their frequencies (or RF). A high value of CC indicates that the probability of occurrence regarding both rank and frequency is high.

We first observe the relationship between Bhojpuri and Magahi using various patterns of data in Table 3. The token [ke] exhibits the highest frequency in both these languages. Further, the CC of the word [ke] is equal in Magahi in comparison to Bhojpuri. Besides, this word does not exist in this list of the Maithili language, indicating that there is a closer association between Magahi and Bhojpuri. Similar to [ke], [meM], [se], and [ki] are on the equally high frequent ranks in Bhojpuri and Magahi. The CCs are again found to be higher in Magahi (0.093, 0.127, 0.178) than Bhojpuri (0.091, 0.127, 0.179). Thus, since phylogenetically related languages have a certain degree of similarity in terms of frequent words by order and rank, the four tokens considered above strengthen an assertion that Bhojpuri and Magahi are genetically related languages.

Additionally, Maithili and Magahi also have interesting patterns in terms of frequency and CC. Consider two of the most frequent words in the list: [I] and [gela]. First, these words do not appear in Bhojpuri’s top ten most frequent words list, but occur at the same rank of occurrence among the two languages. Further, the [gela] word is nearly equally frequent in Magahi and Maithili, and it differs by just one rank between the two languages. The CCs of both of the words in Magahi is higher, being 1.184 and 0.172, while it is 0.115 and 0.108 in Maithili, respectively. Similar to the assertion in the preceding paragraph, our assertion of genetic relations between Maithili and Magahi languages is strengthened. However, it also indicates that perhaps Bhojpuri and Magahi are closer than Magahi and Maithili.

Finally, looking at the total CC for the top ten words, the value is significantly lower for Maithili than it is for Bhojpuri and Magahi, even though Maithili has the least amount of corpus. This value is the highest for Hindi, but that is most probably due to the large size of Hindi corpus and also the fact that Hindi is used much more for academic purposes, thereby having a larger vocabulary. It also supports the linguistic intuition that Maithili is lexically or morphologically richer.

4.2.2 Character. Since three languages, and also Hindi, use the Devanagari script almost in the same way (i.e., writing systems are also similar), comparing statistics about characters used in the corpora can also perhaps give some insights, or at least validate (or invalidate) some linguistic intuition about these languages, we have performed an exercise similar to words at the character level. This is also supported by the fact that the Devanagari script has an almost one-to-one mapping from letters to phonemes. We have considered character $n$-grams (excluding space and special symbols), with $n$ up to 7, as seven $n$-grams of characters seem to capture the identity of a language quite well (Singh [75]) for language identification. Figure 1 lists the TTR and MATTR values for $n$-grams for 1 to 7 levels for the four languages. For character unigrams, the values are nearly zeros for all the languages, because the number of unigrams is quite small (equal to the size of the Devanagari character set), while the total number of character tokens is much larger. This begins to change from bigram upwards. The trends, in terms of numbers, seem to run opposite to those for word-level TTRs. We interpret it as meaning that higher values of character $n$-gram TTR and MATTR indicate greater lexical richness (relative vocabulary size), but lesser morphological complexity, which is in consonance with linguistic intuition. In phonological terms, it could also mean that higher TTR and MATTR for character (or phoneme) $n$-grams is indicative of greater phonological variety. We are not aware of any linguistic intuition about this, but perhaps phonologists familiar with these languages can comment on this better.

4.2.3 Syllable or Akshara. Since there is an almost direct mapping from letters to phonemes in Devanagari, it is also possible to analyze the data in terms of orthographic syllables (which are very similar to phonological syllables). The term Akshara is used for such orthographic syllables. This term has been used since the time of the Paninian grammarians, but in our case, we use a slightly
In this section, we perform a similar exercise as in the preceding two sections but based on the level of orthographic syllables or Aksharas. It has been observed and reported by multiple authors that Akshara is a better linguistic unit than character for many language processing and analysis purposes for Indian languages (Kumar et al. [38], Singh [74]). The grammar for orthographic syllabification is based on the document defining the ISCII standard.\(^\text{13}\)

In general, the syllable structure is defined by the onset-nucleus-coda (consonant-vowel-consonant). The core component of the syllable is the nucleus that is a vowel, and the remaining components—onset and coda—are optional, and are the preceding and the following consonant, respectively. An orthographic syllable for us is an orthographic unit of the Devanagari script (conjuncts included) that may or may not end with a vowel, but does correspond to one or more phonological syllables. Since the position of a syllable in a word (initial, medial or final) is also sometimes important, we consider the numbers of orthographic syllables for each of these positions in Table 4. The TTR and MATTR value for orthographic syllables are given in Table 5.

\[^{12}\url{http://www.tdil-dc.in/tdildcMain/articles/737061Hindi%20Script%20Grammar%20ver%201.4-2.pdf},^{13}\url{http://varamozhi.sourceforge.net/iscii91.pdf}\]
top ten most frequent orthographic syllables in Bhojpuri, Maithili, Magahi, and Hindi languages are given in Table 6.

Observations from Table 4 are as follows. The percentages of syllables that occur in the initial and the final are almost the same in all the four languages. The percentages of syllables that can occur in a medial position is much larger for all languages, which is natural, since the medial part of the word can contain more than one syllables and most words are more than three syllables long. The differences in the percent of syllables that occur in the medial position for each of these languages do tell us indirectly about the complexity (Maddieson [47]). Once again, Maithili has the highest percentage of syllables that occur in the medial position, meaning it has longer and more complex words, in spite of its corpus being the smallest. Hindi also has a large percentage of syllables that occur in the medial, but that is most probably due to the fact that it is used for academic purposes and has a lot of recently borrowed technical or semi-technical long words originating from Sanskrit. Such words are not normally used in spoken languages, whereas the words in the other three languages (as per our corpus) are used mostly in spoken language, since these languages are not used much in the written form, and therefore, for academic purposes. The same also applies to phonological complexities for these languages. In fact, it is difficult to separate phonological complexity from morphological complexity solely based on these statistics.

Table 5 further strengthens the above observation about phonological complexity of the four languages. Earlier TTR and MATTR have been used for morphological complexity while working at the word level (Kettunen [35]). It seems to be a reasonable assumption that if these measures are used at the syllable level, they will similarly reflect phonological complexity, since we are dealing with different kinds of symbolic complexities.

Table 6 is given mainly for reference, although it does not seem to be very informative, except in showing that Hindi is lexically richer than the other languages, which can be explained by the reasons mentioned above.

4.2.4 Morpheme. Since we are interested in relative morphological complexities of the four languages, performing the same exercise as in the previous three sections for morphemes should be more reliably informative. Since we do not have a completely accurate algorithm or grammar for segmenting words into morphemes, we employ the Morfessor tool (Smit et al. [82]) for obtaining morphemes from the corpora. This tool uses a machine learning based algorithm that segments the words into morphemes, based on the corpus provided. Morfessor has two phases: training and decoding. The training data (unannotated word set: \( D_w \)) is used as the input for the training phase, and the log-likelihood cost estimation function learns the theta-model parameter, defined by \( L(D_w, \theta) \). The learned parameter \( \theta \) is used for morphological segmentation of new “test” data \( W \) in the decoding phase:

\[
\text{Training : } \arg \max_\theta L(D_w, \theta),
\]

\[
\text{Decoding : } A = \phi(W; \theta).
\]

We used the Morfessor 2.0\(^\text{14}\) version, with unsupervised learning. Each word in the data is re-written as a sequence of its constituent morphs (hopefully approximating morphemes). This section provides statistics similar to the preceding sections, but at the morpheme level, as given in the Table 7.

Apart from its purpose as a reference, this table provides two main takeaways. Frequencies and relative frequencies (RFs) do not, by themselves, provide much information that is generalisable, although it may be of interest to linguists who want to study these languages. For our

\(^{14}\text{https://morfessor.readthedocs.io/en/latest/}.$
purposes, the first takeaway is very similar to the one from preceding tables. In other words, from the cumulative coverage (CC) values, we can observe the same trend that Maithili has the higher morphological complexity among the three Purvanchal languages. Although Hindi again has a high value, it can be explained in the same way as earlier, i.e., larger and more complex technical or academic vocabulary due to its usage for academic purposes and as a standard language at the national level.

An observation is that the top ten morphemes are very less similar (phonologically and orthographically) for Maithili, as compared to the other languages, which also matches the existing linguistic knowledge about the typological nature of Maithili. The popular knowledge about Maithili says that it is, in fact, closer to Bangla than to Hindi (Chatterji [14]). This suggests that it might be a good idea for future work to compare these three languages with Bangla (for more clarity, please see Section 3).

In fact, there are several common morphemes ([meM], [ke], [se], [na], [nA]) among the top ten for Bhojpuri, Magahi, and Hindi, but none of them occurs in the list for Maithili. This might be the strongest evidence of the common knowledge that Maithili is typologically the most distinct among these four languages.

4.2.5 Zipf’s Curve. Since all languages have a long tail distribution in terms of frequencies and ranks of the words in the vocabulary, in this section, we describe the result of our routine exercise of plotting frequencies versus ranks. According to the Zipf’s law (Piantadosi [59]), the frequency
Table 7. Top-10 Frequent Morphemes Enlisted in Decreasing Order of Frequency/Relative Frequency

| Bhojpuri | Freq  | RF   | CC   |
|----------|-------|------|------|
| [ke]     | 53,802| 0.051| 0.051|
| [wa]     | 15,628| 0.015| 0.153|
| [meM]    | 15,096| 0.014| 0.181|
| [A]      | 14,910| 0.014| 0.195|
| [se]     | 13,871| 0.013| 0.208|
| [na]     | 12,627| 0.012| 0.233|
| [kA]     | 10,820| 0.01  | 0.254|
| [bA]     | 10,377| 0.01  | 0.264|
| [nA]     | 9,903 | 0.009| 0.273|
| [la]     | 8,169 | 0.008| 0.29 |

| Maithili | Freq  | RF   | CC   |
|----------|-------|------|------|
| [ka]     | 23,537| 0.055| 0.055|
| [me]     | 8,788 | 0.021| 0.144|
| [aCi]    | 7,383 | 0.017| 0.161|
| [ke]     | 6,581 | 0.015| 0.176|
| [saz]    | 5,733 | 0.013| 0.189|
| [saBa]   | 5,659 | 0.013| 0.202|
| [je]     | 3,905 | 0.009| 0.232|
| [nahi]   | 3,353 | 0.008| 0.24 |
| [o]      | 3,130 | 0.007| 0.247|
| [para]   | 2,725 | 0.006| 0.253|

| Magahi   | Freq  | RF   | CC   |
|----------|-------|------|------|
| [ke]     | 33,935| 0.048| 0.048|
| [meM]    | 10,145| 0.014| 0.184|
| [la]     | 9,061 | 0.013| 0.223|
| [he]     | 8,694 | 0.012| 0.235|
| [se]     | 8,250 | 0.012| 0.247|
| [na]     | 7,639 | 0.011| 0.28 |
| [hala]   | 6,952 | 0.01 | 0.29 |
| [Au]     | 5,663 | 0.008| 0.298|
| [ho]     | 5,241 | 0.007| 0.305|
| [gela]   | 5,140 | 0.007| 0.312|

| Hindi    | Freq  | RF   | CC   |
|----------|-------|------|------|
| [ke]     | 24,053| 0.028| 0.028|
| [meM]    | 20,465| 0.024| 0.052|
| [he]     | 17,378| 0.02  | 0.072|
| [kl]     | 16,957| 0.02  | 0.092|
| [se]     | 13,457| 0.016| 0.108|
| [kA]     | 12,287| 0.014| 0.136|
| [ko]     | 10,476| 0.012| 0.148|
| [ne]     | 9,457 | 0.011| 0.159|
| [para]   | 7,393 | 0.009| 0.168|

Fig. 2. Zipf’s curve for Bhojpuri, Maithili, Magahi, and Hindi: The plot on the left is for full corpora, while the one on the right is for minimum (Maithili) size

of a word is inversely proportional to its rank. A small proportion of words are used almost all the time, while a vast proportion of words are used rarely.

The Zipf’s curve for Bhojpuri, Maithili, Magahi and Hindi are shown in Figure 2. The first Zipf’s curve is created by looking at the entire corpus, while the second curve follows the minimum number of sentences (Maithili corpus). In this figure, the frequency is in logarithmic scale. The plots for all the four languages are similar as expected and roughly fit the Zipf’s law. It may be
noticed that there are far more rare (frequency one) words in Hindi, which is due to the high usage of academic or technical words borrowed from Sanskrit.

4.2.6 Word length. Yet another parameter that can be used to compare languages based on basic quantitative measures is the word length. Since the postposition marker or case marker (vibhakti) behaves differently in these languages, affecting the word lengths. Sometimes it attaches to the previous word as a morphological suffix, while other times it is a separate word. Word length analysis can give some more evidence in support of or against some existing linguistic knowledge. Also, word length has been an important factor in readability metrics (Kincaid et al. [37]), although our concern is not about document readability, but language complexity. Table 8 shows the word length statistics about the four languages and Figure 3 plots word length versus total words for a certain word length. From the table, the earlier observations about morphological complexity are again strengthened. While the minimum has to be one for all languages, all the other four statistics (maximum, mean, median and standard deviation) have values in the same order as earlier, i.e., highest for Maithili among the three Purvanchal languages and lower for Bhojpuri and Magahi. Hindi has the highest values for the same reasons stated earlier. Bhojpuri, Maithili, and Magahi have almost the same metrics scores, since morphological complexity is higher than Hindi on the restricted corpus. The figure provides even more evidence for this. It also perhaps shows the typological distance of Maithili from the other two Purvanchal languages, with Hindi being a special case due to relatively high standardization and academic usage.

On a side note, it may be seen that the word lengths go up to more than 60 characters. This is after removing some “words” from this plot that were not really words but strings left in the corpus due to incorrect tokenization. Even after removing such non-words, on arranging the words in a descending order according to the word length, we found (as expected) that the longest words were either hyphenated words used for literary purposes, or were from quotations from Sanskrit,
which is morphologically very complex as well as a fusional language. Non-hyphenated native words in all the four languages rarely exceed the word length of 10, as can be seen from the figure. In the figure, dips and a peaks can be seen. On the left side, the peak is indicative of the corpus size, while on the right (for equal corpus size) it is an indication of the effect of the language family. Dips are present for three of the languages, but not for Maithili. From this figure, the following observations can be made:

1. For Hindi, Magahi, and Maithili, single character words are mainly or purely punctuation marks (mostly the period symbol). For Bhojpuri, however, there are some single character function words, so that the frequency of single character types is significantly more for Bhojpuri.
2. For Hindi, Bhojpuri, and Magahi, the first and the highest peak occurs at word length two, since two-character function words are the most common and frequent. For Maithili, however, this peak occurs at three, which is in accord with the Table 7.
3. The dip, which occurs at word length four for three languages, represents the relationship between (and transition from) the function words and the content words. There is no dip in Maithili, because the transition is smooth, as it has many frequent four character function words as well as content words.
4. The second peak occurs either because the said transition is not smooth, or because a transition happens suddenly to the most frequent content words at a certain word length.

4.2.7 Entropy and Perplexity. Information Theory provides us with some measures to further allow us to compare the complexities and to consider the similarities and the differences of the four languages. Such work has a great deal of precedence, also some specifically for Indian languages (Bharati et al. [7]).

In terms of Information Theory, we can see a sentence as being generated from a probabilistic model such as the $n$-gram model, which is still heavily used, in spite of the availability of deep learning models, which give better performance on many problems. This insight has been used for seemingly very different problems such as protein sequencing or DNA sequencing. In such a view, words are generated one after the other, in sequence, based on their $n$-gram probabilities. At each step of the generation of a sentence, when a word is picked up from the vocabulary, there are as many choices as there are words in the vocabulary.

Based on this, we get two very basic, but extremely useful measures. One is entropy, which is for the distribution as a whole, while the other is perplexity, which takes into account the
Table 11. Perplexity of Purvanchal Languages and Hindi

| Language | Unigram | Bigram | Trigram | Unigram* | Bigram* | Trigram* |
|----------|---------|--------|---------|----------|---------|---------|
| Bhojpuri | 1,472.36 | 89.13  | 89.13   | 876.67   | 411.18  | 404.19  |
| Maithili | 1,326.84 | 51.11  | 21.29   | 1327.29  | 51.40   | 21.46   |
| Magahi   | 762.12   | 222.71 | 201.83  | 814.27   | 166.55  | 140.50  |
| Hindi    | 1,610.76 | 134.42 | 99.89   | 1,140.84 | 662.96  | 660.73  |

*The PP over the equal number of sentences according to Maithili CC corpus.

Table 12. Character-level Perplexity of Purvanchal Languages and Hindi

| Language | Unigram | Bigram | Trigram | Unigram* | Bigram* | Trigram* |
|----------|---------|--------|---------|----------|---------|---------|
| Bhojpuri | 26.19   | 17.62  | 10.93   | 26.17    | 17.8    | 11.89   |
| Maithili | 25.4    | 15     | 8.75    | 25.4     | 15      | 8.75    |
| Magahi   | 26.31   | 19.89  | 16.11   | 27.54    | 21.29   | 17.6    |
| Hindi    | 20.38   | 13.03  | 8.91    | 20.69    | 13.97   | 10.61   |

*Indicates the PP over the Equal number of Sentences According to Maithili Corpus.

The perplexity (PP) of a language is a weighted average of the reciprocal of its branching factor. And the branching factor of a language is the number of possible words that can succeed any word, based on the context as available in the corpus. Therefore, it is a mean representative of the possible succeeding words, given a word. We can roughly consider it as the mean branching factor, and therefore, of the syntactic complexity. If the model is a good enough representation of the true distribution for the language, then the PP value indicates syntactic complexity.

We have used the SRILM toolkit\(^{15}\) for obtaining the \(n\)-gram-based statistical model and for calculating perplexity. We estimated the \(n\)-gram language model with backoff smoothing.

From Table 9, we can see (from the values for full corpus size) that the entropy is the lowest for Maithili and the highest for Hindi. Among the three Purvanchal languages, Maithili is the lowest and Bhojpuri the highest. This again indicates higher morphological complexity for Maithili, because it has lower syntactic complexity. The table also indicates high lexical richness for Hindi.

In Table 11, we observe widely varying values for unigrams, bigrams and trigrams, and also widely varying trends as we go from unigrams to trigrams. We also notice that the values for minimum corpus size seem to be more meaningful for perplexity comparison than full corpus size,

\(^{15}\)http://www.speech.sri.com/projects/srilm/.
which intuitively makes sense, since the dataset size has a large effect on the perplexity values. Therefore, we focus on the equal corpus size values in this table. From these, we can observe that the value of perplexity for Bhojpuri becomes less than half when we go from unigram to bigram, but decreases only slightly when we go further to trigrams. This may be interpreted to mean that two word groups (or chunks) and three word groups are almost equally syntactically relevant for Bhojpuri. In the case of Magahi, the value drops steeply from unigrams to bigrams and then does not decrease so much. This is similar to the case of Bhojpuri.

We also calculated character-level perplexity, that we suggest could be indicative of overall phonological (more accurately, orthographic) and morphosyntactic complexity. If that is the case, which can be further investigated, then it also agrees with the earlier observations. In this context, character-level perplexity for words has been related with lexical decision accuracy (Le Godais et al. [45]), although we are not aware of character-level perplexity being used for language complexity.

The trend for Hindi is very similar to Bhojpuri, which matches the linguistic intuition that these two languages have the highest similarity, at least as far as is visible in the written corpus available to us.

5 EXPERIMENTS ON LANGUAGE SIMILARITY: SSNGLMSCORE

Preliminary analyses have shown the potential benefits of building tools for resource-poor languages by leveraging resources from resource-rich languages (here, Hindi) based on similarity (Singh et al. [78], Tiedemann [85], Wu et al. [90], Zoph et al. [99]). The similarity among corpora gives a degree of homogeneity among languages (Porta and Sancho [60]), motivated by their phylogenetic relationships. Hindi, Bhojpuri, Maithili, and Magahi belong to the Indo-Aryan language family, hence they are phylogenetically related languages, and have typological similarity due to the inheritance of certain features from their common ancestor. We have calculated the character-\(n\)-gram-based cross-lingual similarity among these languages. To calculate similarity between two languages, a character-based language model was trained for one language and tested on the other language, and the sentence-wise sum of cross-entropy was obtained. This score should work well as a measure of language similarities, because character-level cross entropy is an established measure of language similarity (as used in language identification, etc.). It has been suggested previously that many NLP problems can be solved either in part of in full by estimation of the right kind of similarity at the right granularity Kumar Singh [42, 43]. A prominent example of this is language identification. The scaled sum of sentence-wise cross-entropy allows us to compare language similarities on a fair ground. These values are listed in Table 13.

5.1 SSNGLMSCore

We used the KenLm language model tool\(^\text{16}\) (Heafield et al. [30]) trained on the corpus using each language (\(SL \in \{\text{Bhojpuri, Magahi, Maithili, Hindi}\}\)) and tested on the corpus of the other three languages \(TL \in \{\text{Bhojpuri, Magahi, Maithili, Hindi}\}\) with the 5-gram and Kneser-Ney smoothing technique.

Let \(m\) sentences exist in a target language (\(tl\)), and the model was trained on the source language (\(sl\)), where \(sl \in SL\), and \(tl \in TL\). The testing scores are scaled in a particular range by the Min-Max scaler algorithm, and we then obtained the score by taking the average mean from a language model in the following way:

\[
SS(sl, tl) = \sum_{tl=1}^{m} p_{sl, tl}(w_n | w_{n-1}).
\]

\(^\text{16}\)https://kheafield.com/code/kenlm/.
Table 13. Character-based n-Gram Corpus
Similarities Based on Cross-Entropy

| Language Pair       | Cross-entropy score |
|---------------------|---------------------|
| Hindi-Bhojpuri      | 3.55                |
| Hindi-Maithili      | 3.27                |
| Hindi-Maithili      | 6.04                |
| Bhojpuri-Maithili   | 7.97                |
| Bhojpuri-Maithili   | 3.88                |
| Maithili-Maithili   | 3.68                |

Here SS stands for Similarity Score based on n-gram language model scores.

\[ MSS_{sl,tl} = \frac{SS_{sl,tl} - \min(SS_{SL,TL})}{\max(SS_{SL,TL}) - \min(SS_{SL,TL})}. \] (4)

Here MSS stands for Min-Max Scaled Similarity Score.

When the \( sl \) and \( tl \) both are the same, then the score will be high. For example, when the language model is trained as well as tested on Bhojpuri, it gives a cross-lingual similarity of 1.

Since, to the best of our knowledge, this measure has not been used before for estimating language similarity, we call this measure the Scaled Sum of N-Gram LM Scores or SSNGLMScore. We have juxtaposed the obtained similarity scores with the Gamallo et al. [26] approach of distance scores, since both scores provide a—supposedly inverse—relationship among languages.

It can be seen from Table 13 that the linguistic intuition about similarities for these languages is corroborated by the values of cross entropy scores, with a couple of caveats. Maithili is the most distant from Hindi, but is close to Bhojpuri, which matches linguistic knowledge. However, one observable discrepancy is that Bhojpuri is almost equally close to Maithili.

So, we look at Table 14, which gives the sequence-based similarity scores. These are symmetric scores, as opposed to the cross entropy scores. Bhojpuri is closest to Maithili and almost equally distant from Hindi and Maithili. Maithili is the most distant from Hindi, but almost equally close to the other two Purvanchal languages. Similar reciprocal patterns have been obtained from Table 15, which calculates the distance among languages.

Magahi is closest to Bhojpuri, then to Maithili, and the least to Maithili. This more or less matches the linguistic knowledge. The less expected closeness between Maithili and Hindi is due to the nature of the data as well as the nature of formal written Magahi, which has been more heavily influenced by Hindi, even though it is expected to be more similar to Maithili than Hindi. This is not an uncommon fact for many languages of northern India. For example, formal written Punjabi is much more similar to Hindi, based on the personal knowledge of one of the authors, than the colloquial and rural Punjabi. Some of this may also be affected by chance, because of taking random samples of 50% corpus size for training and testing. Perhaps n-fold cross validation will give us more reliable estimates. It may also indicate that calculating this score at the sentence level is not sufficient.

From Tables 14 and 15, we have been observed that the Bhojpuri and Magahi are similar language and also have minimum distance than others.

6 BILINGUAL DICTIONARY AND SYNSET

Bilingual lexicon and WordNet (Miller et al. [53]) are crucial linguistic resources for various NLP applications, such as Machine Translation (Gracia et al. [27], Madankar et al. [46], Zhang [98]),
Table 14. Cross-lingual Similarity Score Induced after Applying Character-level Language Model Using SSNGLMScore

| Language Model | Test Language | Bhojpuri | Hindi | Magahi | Maithili |
|----------------|---------------|---------|-------|--------|---------|
| Bhojpuri       | 1             | 0.2398  | 0.3177| 0.1331 |         |
| Hindi          | 0.8676        | 0.2120  | 0.0874|        |         |
| Magahi         |               | 0.9863  | 0.1173|        |         |
| Maithili       |               |         | 0.9543|        |         |

Table 15. Cross-lingual Distance Score Induced after Applying Character-level Language Model Using Perplexity-based Score (Gamallo et al. [26])

| Language Model | Test Language | Bhojpuri | Hindi | Magahi | Maithili |
|----------------|---------------|---------|-------|--------|---------|
| Bhojpuri       | 0             | 0.8571  | 0.4245| 0.6015 |         |
| Hindi          | 0             | 0.8164  | 0.7850|        |         |
| Magahi         | 0             | 0.6825  |        |        |         |
| Maithili       | 0             |         |        |        |         |

Cognate Identification (Ciobanu and Dinu [16], Nakov et al. [56], Nasution et al. [57], Wang and Sitbon [88]), Part-of-Speech Tagging (Fang and Cohn [25], Prabhu and Natarajan [61]), Word-sense Disambiguation (Jain and Lobiyal [31], Vaishnav and Sajja [86]), Bilingual Lexicon Induction (Hangya et al. [29], Nasution et al. [57], Zhang et al. [96]), Word Alignment (Probst and Brown [63]), Cross-language Information Retrieval (Madankar et al. [46], Sharma and Mittal [72]), Word Translation (Lample et al. [44]), and Cross-lingual Word-embedding Alignment (Duong et al. [22], Ruder et al. [67]).

Since this work was directed toward developing Machine Translation systems between Hindi and the three Purvanchal languages, we had undertaken to initiate building WordNets for these languages as well. Since there is already a linked WordNet for many Indian languages, which is called IndoWordNet (Bhattacharyya [10]), in which WordNets of different languages are linked together by mapping synsets (sets of synonyms) of different languages to those of Hindi, i.e., Hindi is used as the Inter-Lingual Index. However, as part of our project, we only created the synsets and mapped them to Hindi. Bilingual lexicon for the three languages are prepared in the usual way, as suitable for machine translation.

The information in a synset includes IDs (corresponding to word meanings in Hindi as given in IndoWordNet), synonyms, lexical categories, and concepts for each entry. The Hindi synset\textsuperscript{17} and bilingual dictionary (Shabdanjali,\textsuperscript{18} an English-Hindi open source e-dictionary) have been used for comparison. The sizes of these two resources are given in Table 16.

The created synset has the highest number of total synonyms and synonyms per word for Bhojpuri. Although Magahi has more cumulative synonyms, Maithili closely surpasses Magahi in synonyms per word. The synonym-statistics is given in Table 17. Each entry in this table correspond to the total number of synonyms per word. For instance, entry in column 1 signify that minimum 1 synonym per word is present in the synset. Similarly, columns 2, 3, and 4 indicate the maximum,

\textsuperscript{17}http://www.cfilt.iitb.ac.in/wordnet/webhwn/.
\textsuperscript{18}http://ltrc.iit.ac.in/onlineServices/Dictionaries/Diet_Frame.html.
Table 16. Cumulative Resource Size of Bilingual Dictionary (Word-Meanings) and Synsets

| Resources          | Bhojpuri | Maithili | Magahi | Hindi  |
|--------------------|----------|----------|--------|--------|
| Bilingual dictionary       | 12,211   | 7,485    | 15,379 | 27,007 |
| Synset                | 37,421   | 8,217    | 10,439 | 39,715 |

Table 17. Distribution of Synonyms Per Word in the Created Synset for the Three Languages Compared with Hindi Synset

| Language | Min | Max | Mean |
|----------|-----|-----|------|
| Bhojpuri | 1   | 101 | 2.93 |
| Magahi   | 1   | 95  | 2.35 |
| Maithili | 1   | 41  | 2.50 |
| Hindi    | 0   | 140 | 2.75 |

and mean of synonyms per word in the corresponding language. It is not clear if any conclusions can be drawn from these statistics alone. However, the data could be useful for studying, say, the degree of ambiguity in these languages.

7 SHALLOW SYNTACTIC ANNOTATION

Annotation enriches the raw corpus with information and allows both linguistic analysis and machine learning. However, annotation by a human is not guaranteed to be perfect. It has been shown that annotation can reflect the annotator’s errors and predilections (Wynne [91]). Therefore, we try to discuss some of the annotation issues that we faced while preparing the annotated corpus.

Syntactic annotation can be very useful for numerous problems in NLP, whether it is Machine Translation, Automatic Speech Recognition, or Information Retrieval. Since this work can be seen as a continuation of the ILMT project (Section 2), we used the standard Bureau of Indian Standards (BIS) tagset (Bharati et al. [8]) for Indian languages that was used in the ILMT project. The BIS tagset consists of 25 POS tags (Table 18) and 11 chunk tags (Table 19) for Bhojpuri, Maithili, and Magahi language. This tagset was used with the understanding that if it can be used for most major Indian languages, it can also be used for the Purvanchal languages. However, for modeling language-specific properties, an extension of this tagset may be required, since, according to some researchers, it is not suitable as it is for all Indian languages (Kumar et al. [41]).

Since we did not go beyond Chunking, our syntactic annotation is basically shallow syntactic annotation, which can be useful for building, for example, shallow parsers for these languages (Sangal et al. [69], Sha and Pereira [71]), or for shallow parsing-based machine translation systems.

Although it is possible to perform annotation in a simple text editor, since the annotated file is stored in a text file, such an approach can lead to easily avoidable errors such as appending operating system specific encoding characters, junk characters, or due to typos. It is, therefore, a common practice to use an annotation tool with a Graphical User Interface that supports a degree of error checking. We used the same tool for such annotation as was used for the ILMT project, namely, Sanchay (Singh [76]). As far as actual file format or representation is concerned, we used the Shakti Standard Format (SSF), which is a robust recursive representation. It allows storing a

---

The chunk annotation was performed for Bhojpuri and Maithili languages only.
Table 18. BIS Tagset Used in POS Annotation

| S.No. | Tag | Description       | S. No. | Tag | Description       |
|-------|-----|-------------------|--------|-----|-------------------|
| 1     | NN  | NOUN              | 14     | QO  | ORDINAL           |
| 2     | NNP | PROPER NOUN       | 15     | SYM | SPECIAL SYMBOL    |
| 3     | PRP | PRONOUN           | 16     | CC  | CONJUNCTS         |
| 4     | VM  | VERB MAIN         | 17     | CL  | CLASSIFIER        |
| 5     | VAUX| VERB AUXILIARY    | 18     | QF  | QUANTIFIER        |
| 6     | DEM | DEMONSTRATIVE     | 19     | INJ | INTERJECTION      |
| 7     | PSP | POST POSITION     | 20     | INTF| INTENSIFIER       |
| 8     | QC  | CARDINAL          | 21     | NEG | NEGATIVE          |
| 9     | JJ  | ADJECTIVE         | 22     | RDP | REDULICATIVE      |
| 10    | RB  | ADVERB            | 23     | ECH | ECHO              |
| 11    | NST | NOUN LOCATION     | 24     | UNK | UNKNOWN           |
| 12    | WQ  | QUESTION WORD     | 25     | UT  | QUOTATIVE         |
| 13    | RP  | PARTICLE          |        |     |                   |

Table 19. BIS Tagset Used in Chunk Annotation

| S. No. | Tag | Description         |
|--------|-----|---------------------|
| 1      | NP  | Noun Chunk          |
| 2      | VGF | Finite Verb Chunk   |
| 3      | VGNF| Non-finite Verb Chunk |
| 4      | VGINF| Infinitival Verb Chunk |
| 5      | VGNN| Gerunds             |
| 6      | JJP | Adjectival Chunk    |
| 7      | RBP | Adverb Chunk        |
| 8      | NEGP| Negatives           |
| 9      | CCP | Conjuncts           |
| 10     | FRAGP| Chunk Fragments    |
| 11     | BLK | Miscellaneous entities |

A wide variety of linguistic and non-linguistic information in one place. The detailed description of the SSF representation can be found in Bharati et al. [9]. We tried to get the same kind of statistics as above for POS tags, pairs of words and POS tags and of chunks. Table 20, 21 show the TTR, entropy and MATTR values for these units, respectively.

Entropy values do not vary much for the three units over the four languages. Only two things are noticeable here from the second column of the Tables 20 and 21. One is that Word_POS_Tag TTR and MATTR are quite low for Hindi and, for the same language, the Chunk Tag TTR and MATTR are very high for Hindi, while POS Tag TTR and MATTR are almost the same for all languages. This seems to indicate that syntactic complexity of the three languages in terms of POS tags is almost the same. The Word_POS_Tag TTR is twice for Hindi that the other three languages, meaning lexicalized POS tag complexity (say, as modelled by HMM) is higher for Hindi. Chunk TTR and MATTR being very high for Hindi indicates much greater syntactic complexity at the chunk tag level, which could partially be explained by the same reason of academic usage but also of larger data size. Hindi is also relatively less morphologically complex, which could suggest it...
has relatively higher syntactic complexity. We do not rely as much on varied data for the three Purvanchal languages as we have for Hindi. This factor might be affecting other measures too.

### 7.1 Annotation Issues

There are many challenges while annotating the corpora of these less-resourced languages. Bhojpuri, Maithili and Magahi are partially synthetic languages, because they often express syntactic relations within word forms through suffixes, as shown in some of the examples given in Section 3. Thus, the use of embedded case markers, emphatic markers, classifiers, determiners, and so on, is frequent in these languages. This is responsible for many problems in annotation.

#### 7.1.1 Embedded Case Markers

It is already mentioned in Section 3 that in many morphological constructions of these languages, case markers get merged with nominals and pronominals, especially locative, instrumental and genitive case markers get fused with nominals and pronominals. The annotated examples listed here are the problematic cases, because the presently assigned tags are unable to capture the full meaning of such nominal and pronominal fused forms. This can be a matter for further work on these languages. Perhaps the tagset also may have to be revised a little to accommodate certain idiosyncrasies of these languages.

#### 7.1.2 Diverse Realizations of Single Token

In Bhojpuri, Maithili, and Magahi, several cases of diverse realization of a single token are found. The token [ke] has the highest different realizations in Magahi and Bhojpuri: 12 and 15 tags, respectively. Whereas in Maithili the token [je] has been assigned 14 tags for different realizations. List of such tokens is long in all these three languages. Mostly these tokens have multiple functional and connotative meanings, making the process of annotation hard and error prone.

#### 7.1.3 Homophones

Like many other Indian languages, Bhojpuri, Maithili and Magahi also have homophone words in abundance. These words look similar, but their POS tags are varied, which may pose a sense of perplexity for the annotator.

Example 1:

**Bhojpuri**: [kula] (JJ) paisave (NN) orA (VM) gaila (VAUX). (SYM)

**Hindi**: [sArA] pEsA hI samApwa ho gayA.

**English Translation**: [All] the money is gone.
Table 22. Annotation Statistics

| Language  | POS annotation | Chunk annotation |
|-----------|----------------|------------------|
|           | # Sentence  | # Token     | # Sentence | # Token    |
| Bhojpuri  | 16,067      | 245,482     | 9,695      | 60,588     |
| Maithili  | 12,310      | 208,640     | 1,954      | 10,476     |
| Magahi    | 14,669      | 171,509     | -          | -          |

Bhojpuri: wohAra (PRP) kOna (WQ) [kula] (NN) bA (VM)? (SYM)
Hindi: wumhArA kOna sA [kula] hE?
English Translation: Which one is your [clan]?

In these example sentences of Bhojpuri, the word [kula] is homophones, each has a different meaning and different grammatical category.

Example 2:

Bhojpuri: laikA (NN) geMxa (NN) ke (PSP) [loka] (VM) lihalasa (VAUX). (SYM)
Hindi: ladZake ne geMxa [lapaka] II.
English Translation: Boy [caught] the ball.

Bhojpuri: sahara (NN) A (CC) gazvaI (JJ) [loka] (NN) meM (PSP) aMwara (NN) bA (VM).
(SYM)
Hindi: Sahara Ora grAmIna [samAja] meM aMwara hE.
English Translation: There is difference between the urban and the rural [society].

In these example sentences of Bhojpuri, the word [loka] is homophonous, with each homophone having a different meaning and different grammatical category.

7.1.4 Amalgamated Emphatic Expressions. In these languages, emphatic particles are generally merged with nominals and pronominals and hence they make annotation task difficult, as mentioned in Section 3.

7.2 Annotated Corpus Statistics

Earlier reported work on Bhojpuri and Magahi had 120k and 63k tokens, respectively, annotated with modified tagset of Bureau of Indian Standards or BIS (Kumar et al. [41], Ojha et al. [58]). Bhojpuri, Maithili, and Magahi being under-resourced languages, researchers have not paid that much attention to these languages, even though a large number of native speakers are available. One earlier work on Maithili POS tool is available online.20 Recently, Priyadarshi and Saha [62] have reported a POS tagger and studied the influence of the language-specific features (Saha and Priyadarshi [68]).

In our work on annotation, the cleaned raw corpus was used for POS tag and Chunk annotation. POS annotated corpus comprises of 245,489 words for Bhojpuri, 208,648 words for Maithili and 171,538 million words for Magahi. Also, Chunk annotation has 9,695 and 1,954 sentences (60,591 and 10,476 tokens, respectively) for Bhojpuri and Maithili. For Hindi, we used the Hindi-Urdu dependency treebank dataset, which comprises 20,866 sentences, which have 436,940 POS tags and 235,426 chunks. The annotation statistics of these language given in Table 22.

The POS tag distribution in the annotated corpus is given in Table 23. The NN, VM, PSP tags covers approximately 50% of Bhojpuri annotated corpus, whereas the same is true for NN, VM, PSP, SYM tags in Maithili and NN, VM, SYM tags in Magahi annotated corpus include. Moreover,

20http://sanskrit.jnu.ac.in/mpost/index.jsp.
Table 23. Frequency-based POS Tag Percentage Ratio of Bhojpuri, Maithili, Magahi, and Hindi Annotated Corpus

| POS Tag | Tag Frequency (Bhojpuri) | Tag Percentage (Bhojpuri) | Tag Frequency (Maithili) | Tag Percentage (Maithili) | Tag Frequency (Magahi) | Tag Percentage (Magahi) | Tag Frequency (Hindi) | Tag Percentage (Hindi) |
|---------|--------------------------|---------------------------|--------------------------|---------------------------|------------------------|------------------------|-----------------------|------------------------|
| NN      | 64,659                   | 26.34                     | 54,971                   | 26.35                     | 42,925                 | 25.03                  | 86,742                | 19.85                  |
| VM      | 38,098                   | 15.52                     | 26,226                   | 12.57                     | 21,678                 | 12.64                  | 46,500                | 10.64                  |
| PSP     | 35,226                   | 14.35                     | 17,012                   | 8.15                      | 17,933                 | 10.46                  | 83,879                | 19.19                  |
| SYM     | 28,993                   | 11.81                     | 21,108                   | 10.12                     | 26,167                 | 15.26                  | 31,445                | 7.1                    |
| PRP     | 16,655                   | 6.78                      | 12,976                   | 6.22                      | 5,840                  | 3.41                   | 18,486                | 4.23                   |
| Vaux    | 10,740                   | 4.38                      | 18,028                   | 8.64                      | 17,094                 | 9.97                   | 28,195                | 6.45                   |
| NNP     | 9,232                    | 3.76                      | 13,381                   | 6.41                      | 4,111                  | 2.4                    | 35,519                | 7.67                   |
| JJ      | 8,781                    | 3.58                      | 9,124                    | 4.37                      | 6,053                  | 3.53                   | 23,380                | 5.35                   |
| CC      | 7,476                    | 3.05                      | 7,600                    | 3.64                      | 4,747                  | 2.77                   | 16,070                | 3.67                   |
| RP      | 4,287                    | 1.75                      | 3,076                    | 1.47                      | 5,247                  | 3.06                   | 7,233                 | 1.65                   |
| NST     | 4,092                    | 1.67                      | 5,148                    | 2.47                      | 4,694                  | 2.74                   | 6,488                 | 1.48                   |
| QF      | 3,590                    | 1.46                      | 3,560                    | 1.71                      | 2,480                  | 1.45                   | 3,425                 | 0.78                   |
| QC      | 3,385                    | 1.38                      | 4,324                    | 2.07                      | 1,780                  | 1.04                   | 7,396                 | 1.69                   |
| NEG     | 3,197                    | 1.3                       | 2,232                    | 1.07                      | 2,140                  | 1.25                   | 3,253                 | 0.74                   |
| WQ      | 1,811                    | 0.74                      | 1,770                    | 0.85                      | 1,221                  | 0.71                   | 724                   | 0.16                   |
| DEM     | 1,429                    | 0.58                      | 2,761                    | 1.32                      | 2,122                  | 1.24                   | 5,442                 | 1.29                   |
| RB      | 1,239                    | 0.5                       | 2,917                    | 1.4                       | 3,500                  | 2.04                   | 1,956                 | 0.44                   |
| RDP     | 722                      | 0.29                      | 320                      | 0.15                      | 205                    | 0.12                   | 365                   | 0.08                   |
| OO      | 513                      | 0.21                      | 654                      | 0.31                      | 571                    | 0.33                   | 742                   | 0.16                   |
| INTF    | 499                      | 0.2                       | 295                      | 0.14                      | 489                    | 0.29                   | 438                   | 0.1                    |
| INJ     | 418                      | 0.17                      | 413                      | 0.2                       | 288                    | 0.17                   | 164                   | 0.03                   |
| CL      | 248                      | 0.1                       | 295                      | 0.14                      | 66                     | 0.04                   | 0                     | 0                      |
| ECH     | 122                      | 0.05                      | 54                       | 0.03                      | 129                    | 0.08                   | 1                     | 0                      |
| UNK     | 63                       | 0.03                      | 394                      | 0.19                      | 28                     | 0.02                   | 205                   | 0.04                   |
| UT      | 7                        | 0                         | 1                        | 0                         | 1                      | 0                      | 0                     | 0                      |

Here only those POS tags of Hindi are considered, which were used for annotation of the three languages.

some tags having a coverage of less than 0.1% are considered as rare tags, such as ECH, UNK, and UT for Bhojpuri, ECH and UT for Maithili, and CL, UNK, and UT for Magahi. These have nearly 0% coverage of POS tag distribution in annotated corpus.

In the chunk tag annotated corpus, the **Noun Phrase (NP)** is the most frequent as compared to other phrases in Bhojpuri and Maithili. Moreover, **Chunk Fragment Phrase (FRAGP)** has the minimum occurrence, as mentioned in Table 24.

We calculated the inter-annotator agreement among the random annotated sets of 962 for Bhojpuri, 578 for Maithili and 381 for Maithili using Cohen’s Kappa (Cohen [17]) to validate the quality of annotation. Cohen’s Kappa inter-annotator score was 0.92, 0.64, and 0.74 for Bhojpuri, Magahi and Maithili, respectively, which indicates that the quality of the annotation and presented schema is usable to different degrees.

### 7.3 Baseline for POS Tagging and Chunking for Purvanchal Languages

Since the objective of this work as to function as the basis for further work on these languages, we finish the section on annotation by describing a reasonably good baseline, which can be treated as the lower baseline for later research on building POS taggers and Chunkers of these languages.

As part of our experiment with this baseline, we compare our results with those for Hindi part of the Hindi-Urdu dependency treebank for automatic annotation. Since we suspect the quality of the data and annotation is not as good as, say, for Hindi, we also consider these experiments as a kind of additional validation that our data may be useful for the purposes for which it was
Table 24. Frequency-based Chunk Tag Percentage Ratio of Bhojpuri, Maithili, and Hindi Annotated Corpus

| Chunk Tag | Tag Frequency (Bhojpuri) | Tag Frequency (Maithili) | Tag Frequency (Hindi) |
|-----------|--------------------------|--------------------------|-----------------------|
| NP        | 35,393                   | 4,321                    | 135,423               |
| VGF       | 16,951                   | 3,069                    | 32,886                |
| CCP       | 3,394                    | 1,125                    | 16,119                |
| VGNF      | 1,669                    | 49                       | 5,250                 |
| BLK       | 1,559                    | 47                       | 21,828                |
| JJP       | 690                      | 555                      | 10,309                |
| NEGP      | 447                      | 347                      | 379                   |
| RBP       | 397                      | 458                      | 2,268                 |
| VGNN      | 54                       | 395                      | 8,344                 |
| VGINF     | 32                       | 108                      | 0                     |
| FRAGP     | 2                        | 2                        | 412                   |

Here only those chunk tags of Hindi are considered, which were used for annotation of the three languages.

Table 25. Effect of Features in POS Tagging’s F\textsubscript{1}-score in Percentage

| Features          | Bhojpuri | Maithili | Magahi | Hindi |
|-------------------|----------|----------|--------|-------|
| Word              | 0.80     | 0.72     | 0.76   | 0.85  |
| +Previous words   | 0.82     | 0.75     | 0.77   | 0.89  |
| +Next words       | 0.85     | 0.75     | 0.78   | 0.93  |
| +Group words      | 0.88     | 0.76     | 0.78   | 0.94  |

made. The idea is that if a reasonably good machine learning algorithm is able to give tolerable predictions (which could be improved in future) using this data, then the data is not useless, in spite of shortcomings.

For the experiments, we have used the traditional state-of-the-art machine learning or probabilistic technique, known as the \textbf{Conditional Random Fields (CRF)}, for sequence labeling. The CRF model provides surprisingly good results, even in the absence of an ample amount of annotated dataset. CRF++ toolkit\textsuperscript{21} with the default feature set and parameters have been used for POS tagging and Chunking. For performing these experiments, we have divided the dataset into 80–20 ratio for training and testing. The reported results for POS tagging and Chunking of each language are evaluated on the testing dataset. The feature-wise obtained results for POS tagging and Chunking in terms of weighted F\textsubscript{1}-score are shown in Tables 25 and 26, respectively. We have used the previous word, next word and contextual word as a feature for POS tagging. Additionally, the contextual POS tag is also considered as a feature for Chunking. The contextual information (both word and POS tag) play an essential role in improving model performance, as shown in these tables.

Furthermore, the obtained results are compared with the Hindi dependency treebank dataset in terms of accuracy, weighted precision, weighted recall, and weighted F\textsubscript{1}-score, which are shown in the Tables 27 and 28.

\textsuperscript{21}https://taku910.github.io/crffpp/.

ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 20, No. 6, Article 95. Publication date: September 2021.
Table 26. Effect of Features in Chunking’s F₁-score in Percentage

| Features          | Bhojpuri | Maithili | Hindi |
|-------------------|----------|----------|-------|
| Word              | 0.79     | 0.64     | 0.92  |
| +Previous words   | 0.83     | 0.66     | 0.94  |
| +Next words       | 0.91     | 0.93     | 0.98  |
| +Contextual tags  | 0.92     | 0.93     | 0.97  |
| +Group words      | 0.92     | 0.94     | 0.99  |

Table 27. POS Tagging Experimental Result in Percentage Using CRF Technique

| Language | Precision | Recall | F₁-score | Accuracy |
|----------|-----------|--------|----------|----------|
| Bhojpuri | 0.89      | 0.89   | 0.88     | 0.89     |
| Maithili | 0.77      | 0.77   | 0.76     | 0.77     |
| Magahi   | 0.8       | 0.78   | 0.78     | 0.78     |
| Hindi    | 0.94      | 0.94   | 0.94     | 0.94     |

Table 28. Chunking Experimental Result in Percentage Using CRF Technique

| Language | Precision | Recall | F₁-score | Accuracy |
|----------|-----------|--------|----------|----------|
| Bhojpuri | 0.92      | 0.92   | 0.92     | 0.92     |
| Maithili | 0.94      | 0.94   | 0.94     | 0.94     |
| Hindi    | 0.99      | 0.99   | 0.99     | 0.99     |

For POS tagging, we get 88% F₁-score Bhojpuri, which is not bad, considering that we used the default template. For Maithili and Magahi, the values are a bit below the 80% mark, which can be explained by the low quantity of data. Compared to Hindi, the results are not very unexpected.

For Chunking, the results for Bhojpuri and Maithili are even better, being above 90%, although much lower than the 99% for Hindi. These values are higher for Chunking than for tagging, because the number of Chunk tags is less than half of the number of POS tags. Moreover, only two Chunk tags (noun phrase and finite verb phrase) cover most of the corpus for all three languages. Also, due to their higher morphological complexity, what would have been a multi-word chunk often becomes a single word chunk in these languages, leading to easier prediction of chunks.

8 AN EXPERIMENT ON LANGUAGE IDENTIFICATION

Language Identification (LI) on text is one of the most basic tasks in NLP, which determines the language of the provided text, where the input text is either a word, a sentence, paragraph or even a document or a set of documents. LI aims to mimic human ability to recognise or identify a language, without actually knowing that language. Multiple approaches have been proposed over the years, that infer the languages without human intervention. According to Simons and Fennig [73], LI approaches should able to discern thousands of human language.

LI is one of the oldest problems in NLP. The text displayed on computing devices is made up of a stream of characters (or glyphs), which are digitised by a particular encoding scheme. Most
troubling for LI is the use of either isomorphic or proprietary encodings to encode text in some languages where there is a lack of standardization or standardization has not yet been fully adapted. Perhaps the most important advance was the Cavnาร et al. [12] method, which has popularised the use of character-level n-gram models for automatic LI.

It is useful in various NLP applications, which include authorship profiling, machine translation, information retrieval, lexicography, and many more. Moreover, the LI system’s output is utilized in the adaptation of boilerplate NLP tools that require annotated data, such as POS tagger, Chunker for resource-poor languages.

Earlier work considers encoding in terms of bytes (Singh [75], Singh and Gorla [77]), character (Baldwin and Lui [5], Damashek [20], Singh [75], Singh and Gorla [77]), character combination (Banerjee et al. [6], Elfardy and Diab [24], Sterneberg [84], Windisch and Csink [89]), morpheme (Anand [3], Marcadet et al. [49], Romsdorfer and Pfister [66]), words (Adouane and Dobnik [1], Rehurek and Kolkus [64]), word combination (Coltekin and Rama [18], Singh [75]), syllables (Murthy and Kumar [55], Yeong and Tan [93]), syntax (Alex [2], Martinc et al. [50]), and chunks (Elfardy et al. [23], You et al. [94]) as features. Even though LI task is considered a solved problem for distant languages (McNamee [52]), it is still challenging to identify closely related languages. Recently, more attention has been paid to similar or closely related languages, such as the Purvanchal languages we deal with in this work.

The more similar the languages are, the more challenging it is to discriminate them or to identify them correctly for an LI system. Instead of n-gram-based models, which did not consider sequences of characters or other units, the approaches to LI for closely related languages are mostly based on sequential algorithms. One of the earlier work on closely related Indian languages identification applied by pair-wise binary classification on Indo-Aryan (Hindi, Marathi, Punjabi, Oriya, Assamese and Bengali) and Dravidian (Tamil, Telugu, Kannada and Malayalam) language families. For the LI system, characters and aksharas are considered as the smallest appropriate unit to apply multiple learning methods (Murthy and Kumar [55], Singh [75]). Recently, VarDial 2018\(^\text{22}\) workshop included a shared task on language identification, which included Indo-Aryan languages, i.e., Hindi, Bhojpuri, Magahi, Braj, and Awadhi. The dataset consisted of 15,000 instances corresponding to each language. HeLI (Jauhiainen et al. [32]) is a character-level, n-gram language model with adaptation technique, where the system provides 0.958% macro F-score (Zampieri et al. [95]).

We used a convolutional neural network (ConvNet) for the LI task at the character level. Earlier, ConvNet has been applied to text classification (Zhang et al. [97]). The ConvNet takes the input of all possible characters used in the script of the language of the corpus (Devanagari for Bhojpuri, Maithili, Magahi and Hindi) to create embeddings of characters. Here, a total of 115 characters extracted from the corpus of all languages, written in Devanagari script include consonants, vowels, special symbols, digits, and so on. The input sentences are quantized lengthwise, which reduces the dimension of the character’s one-hot vector. These vectors or embeddings are passed to the convolutional layer to extract language-specific features and select the optimal features by applying max-pooling over them. The pair of convolution and max-pooling layer is sequentially used six times. Finally, three fully connected layers with a non-linear function, followed by a pooling layer are applied, of which the final fully connected layer defines the number of languages in our training corpus.

The total instances in the dataset are 149,999, created by merging of Bhojpuri, Maithili, Magahi and Hindi monolingual corpus for performing LI using ConvNet. These instances correspond to sentences, of which 50,149, 19,463, 30,520, and 49,873 sentences are from Bhojpuri, Maithili, Magahi and Hindi, respectively. This dataset was divided into 80–20 ratio for training and validation of our

\(^{22}\)http://alt.qcri.org/vardial2018/index.php?id=campaign.
Table 29. Language Identification Scores (%) Obtained by ConvNet

| Language   | Precision | Recall  | F1-score |
|------------|-----------|---------|----------|
| Bhojpuri   | 98.94     | 99.67   | 99.30    |
| Hindi      | 99.58     | 99.92   | 99.75    |
| Magahi     | 99.59     | 98.36   | 98.97    |
| Maithili   | 99.56     | 98.69   | 99.12    |
| Macro avg. | 99.42     | 99.16   | 99.29    |
| Weighted avg. | 99.36   | 99.36   | 99.36    |

The length of input sentences is fixed at 150 (with padding, as is usual while using deep neural networks) before training the model. The initial two pairs are of convolution and max-pooling using 256, 7, and 3 as convolution, kernel, and pooling size. Remaining pairs have exploited 256, 3, and 3 (only on the last pair) as respective values. The units of fully connected layers are 1,024, 1,024, and 4, which are separated by a dropout layer (0.5 as dropout probability) to avoid overfitting of the model. The model was trained for 10 epochs with the minibatch size of 128 and Stochastic Gradient Descent (SGD) optimizer with the learning rate of 0.01.

The obtained results on the testing dataset (3,604 instances, which are different from the training and the validation) data are given in Table 29. We have obtained the 99.36% and 98.70% accuracy on the testing and validation dataset, respectively, by this model.

While we get better language identification results for our experiments than the workshop on Indo-Aryan languages, this may be partly because we tried only with four languages.

Apart from providing a baseline for comparison for future work, this experiment on LI also acts as the last check on the quality of the data. These languages are quite close and almost none of the annotators use them for writing. Also, there is a complete lack of standardization and the boundaries between these languages are sometimes blurred as they are spoken mostly in a contiguous area. Each of them has numerous varieties, which change gradually at least according to the geographical location. It means that on the border areas, it is sometime not clear which language is being spoken, which is a common sociolinguistic phenomenon. Also, the annotators are from all over this area and they are familiar with different varieties of these languages, so there was often disagreement among them. The LI experiments at least shows that the data can be discriminated based on the language it is supposed to be from, using a standard LI algorithm for closely related languages.

9 CONCLUSION AND FUTURE WORK

We collected raw corpora from various sources (as mentioned in the Appendix) that cover different genres and domains for Bhojpuri, Magahi, and Maithili languages. After cleaning their raw corpora, we obtained 51,374, 31,251, and 19,948 sentences, respectively. Of these, 16,067, 14,669, and 12,310 sentences are annotated for Part-of-speech tagging, respectively. We have performed statistical analysis at the character, word, morpheme, and syllable levels to exhibit the agglutinative and word-sharing properties. We have computed the unigram (or word), POS, and POS-word entropy of our corpus for the three languages and the n-gram character-based cross-entropy score among the languages, including Hindi, which show a degree of similarity among languages. Additionally, the manually annotated sentence reliability is estimated through Cohen’s kappa Inter-Annotator Agreement and found to be 0.92, 0.64, and 0.74 for Bhojpuri, Magahi, and Maithili, respectively. The highest inter-annotator agreement score for Bhojpuri indicates that its annotation done by
different linguists is found to be in the fairest agreement. We discussed some of the annotation issues for all the three languages. The POS-tagged data was then annotated for Chunk tags: Bhojpuri (9,695 sentences) and Maithili (1,954 sentences). We built an initial POS tagger, Chunker, and Language identifier tool by using these resources. Also, the bilingual dictionaries from these three Purvanchal languages to Hindi and synsets for IndoWordNet are prepared.

To the best of our knowledge, such corpora and corresponding syntactic annotation using POS tagging and Chunking, the synset (WordNet) and bilingual dictionary preparation is not available so far for these low-resource Purvanchal languages. The collected and annotated corpora can be used as the basis for the creation of further linguistic resources for these languages. These linguistic resources will then be equipped to create machine translation systems for these languages.

The results for the initial POS tagger and Chunker were reasonably good, i.e., good enough to be used as baselines for future work. The language identification results were better than reported earlier and they also confirm that the quality of the data, though problematic, is not completely useless.

APPENDIX

This section contains online sources from where different parts of the data for Purvanchal languages were collected and the example of their annotated sentences.

Bhojpuri sources
http://www.bhojpuria.com/page=sahitya/index.html.
http://www.thesundayindian.com/bh/.
http://anjoria.com/.
http://bhojpurika.com/.
http://anjoria.com/v1/bhasha/bhasha.htm.
http://www.readwhere.com/read/637950.

Magahi sources
http://magahi-sahitya.blogspot.com/default, the above blog taken contained a large fraction of Magahi colloquial (TeMTa) words.

Maithili sources
http://www.esamaad.com/.
http://www.maithilijindabaad.com/.
https://www.bible.com/versions/1302-maint-jiv-at-n-s-at-ndesh.
http://www.videha.co.in/.
http://mithila-maithil-maithili-sites.blogspot.in/.
http://maithili-katha.blogspot.in/.
http://www.apnikahaani.com/2013/08/maithili-story-basaat.html.
http://www.maithilijindabaad.com/?p=3902.
https://hindipatal.wordpress.com/.

REFERENCES
[1] Wafia Adouane and Simon Dobnik. 2017. Identification of languages in Algerian Arabic multilingual documents. In Proceedings of the 3rd Arabic Natural Language Processing Workshop. 1–8.
[2] Beatrice Alex. 2005. An unsupervised system for identifying English inclusions in German text. In Proceedings of the ACL Student Research Workshop. 133–138.
[3] Supriya Anand. 2014. Language identification for transliterated forms of Indian language queries. In Proceedings of the Forum for Information Retrieval Evaluation (FIRE’14).
[4] Srinivasu Badugu. 2014. Morphology-based POS tagging on Telugu. Int. J. Comput. Sci. Issues 11, 1 (2014), 181.
[5] Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies. Association for Computational Linguistics, 229–237.

[6] Somnath Banerjee, Alapan Kuila, Aniruddha Roy, Sudip Kumar Naskar, Paolo Rosso, and Sivaji Bandyopadhyay. 2014. A hybrid approach for transliterated word-level language identification: CRF with post-processing heuristics. In Proceedings of the Forum for Information Retrieval Evaluation. 54–59.

[7] Akshar Bharati, K. Prakash Rao, Rajeev Sangal, and S. M. Bendre. 2000. Basic statistical analysis of corpus and cross comparison among corpora. Technical Report, Indian Institute of Information Technology.

[8] Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma, and Lakshmi Bai. 2006. Anncorra: Annotating corpora guidelines for pos and chunk annotation for Indian languages. In Proceedings of the Annual Language Testing Research Colloquium (LTRC’06). 1–38.

[9] Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma, and Anil Kumar Singh. 2014. SSF: A common representation scheme for language analysis for language technology infrastructure development. In Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT. 66–76.

[10] Pushpak Bhattacharyya. 2010. Indowordnet. In Proceedings of the Language Resources and Evaluation Conference (LREC’10).

[11] Ondrej Bojar, Vojtech Diatka, Pavel Rychlí, Pavel Stranák, Vit Suchomel, Ales Tamchyna, and Daniel Zeman. 2014. HindEnCorp-Hindi-English and Hindi-only corpus for machine translation. In Proceedings of the Language Resources and Evaluation Conference (LREC’14). 3550–3555.

[12] William B. Cavnar, John M. Trenkle et al. 1994. N-gram-based text categorization. In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, Vol. 161175. Citeseer.

[13] Sunita Kumar Chatterji. 1926. The Evolution of Bengali Language. Rupa, Delhi.

[14] Suniti Kumar Chatterji. 1986. The Origin and Development of the Bengali Language, vol. 1. Rupa, Delhi.

[15] B. Chaudhuri and S. Ghosh. 1998. A statistical study of Bangla corpus. In Proceedings of the International Conference on Computational Linguistics, Speech, and Document Processing.

[16] Alina Maria Ciobanu and Liviu Petrisor Dinu. 2013. A dictionary-based approach for evaluating orthographic methods in cognates identification. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP’13). 141–147. Retrieved from https://www.aclweb.org/anthology/R13-1019.

[17] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Edu. Psychol. Measure. 20, 1 (1960), 37–46. https://doi.org/10.1177/001316446002000104

[18] Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear SVMs and neural networks. In Proceedings of the 3rd Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial’16). 15–24.

[19] Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio. Amer. J. Quant. Linguist. 17, 2 (2010), 94–100. https://doi.org/10.1080/09296171003643098

[20] Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. Science 267, 5199 (1995), 843–848.

[21] Niladri Sekhar Dash. 2004. Language corpora: Present Indian need. In Proceedings of the SCALLA Working Conference. 5–7.

[22] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 1285–1295. https://doi.org/10.18653/v1/D16-1136

[23] Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. ALIDA: Identifying code switching in informal Arabic text. In Proceedings of the 1st Workshop on Computational Approaches to Code Switching. 94–101.

[24] Heba Elfardy and Mona Diab. 2013. Sentence-level dialect identification in Arabic. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 456–461.

[25] Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 587–593. https://doi.org/10.18653/v1/P17-2093

[26] Pablo Gamallo, José Ramom Pichel, and liaki Alegria. 2017. From language identification to language distance. Physica A: Stat. Mech. Appl. 484 (2017), 152–162.

[27] Jorge Gracia, Besim Kabashi, Ilan Kernerman, Marta Lanau-Coronas, and Dorielle Lonke. 2019. Results of the translation inference across dictionaries 2019 shared task. In Proceedings of TIAD-2019 Shared Task - Translation Inference Across Dictionaries co-located with the 2nd Language, Data and Knowledge Conference (LDK’19), Leipzig, Germany, May 20, 2019, Vol. 2493. CEUR-WS.org, 1–12.

[28] George Abraham Grierson. 1967. Linguistic Survey of India, vol. III. Motilal Banarsidass. https://dsal.uchicago.edu/books/lsi/.
[53] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, and K. Miller. 2010. WordNet: An online lexical database. *Int. J. Lexicogr.* 3, 4 (2010), 235–244.

[54] Aanchan Mohan, Richard Rose, Sina Hamidi Ghalejehg, and Sririvasan Umesh. 2014. Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain. *Speech Commun.* 56 (2014), 167–180.

[55] Kavi Narayana Murthy and G. Bharadwaja Kumar. 2006. Language identification from small text samples. *J. Quant. Linguist.* 13, 01 (2006), 57–80.

[56] Svetlin Nakov, Preslav Nakov, and Elena Paskaleva. 2009. Unsupervised extraction of false friends from parallel bi-texts using the web as a corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP’09)*. 292–298.

[57] Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2016. Constraint-based bilingual lexicon induction for closely related languages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’16)*. 3291–3298.

[58] Atul Ku Ojha, Pitambar Behera, Srishti Singh, and Girish N. Jha. 2015. Training & evaluation of POS taggers in Indo-Aryan languages: A case of Hindi, Odia and Bhojpuri. In *Proceedings of the 7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. 524–529.

[59] Steven T. Piantadosi. 2014. Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonom. Bull. Rev.* 21, 5 (2014), 1112–1130.

[60] Jordi Porta and José-Luis Sancho. 2014. Using maximum entropy models to discriminate between similar languages and varieties. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties, and Dialects*. 120–128.

[61] Nikhil Prabhu and S. Natarajan. 2019. Extraction of character personas from novels using dependency trees and POS tags. In *Emerging Research in Computing, Information, Communication, and Applications*. Springer, 65–74.

[62] Ankur Priyadarshi and Sujan Kumar Saha. 2020. Towards the first Maithili part of speech tagger: Resource creation and system development. *Comput. Speech Lang.* 62 (2020), 101054.

[63] Katharina Probst and Ralf Brown. 2002. Using similarity scoring to improve the bilingual dictionary for word alignment. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 409–416.

[64] Radim Rehurek and Milan Korkus. 2009. Language identification on the web: Extending the dictionary method. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 357–368.

[65] Rishikesh. 2018. Parts of speech tagger for Maithili language using HMM. *Int. J. Innovat. Adv. Comput. Sci.* 7 (2018), 206.

[66] Harald Romsdorfer and Beat Pfister. 2007. Text analysis and language identification for polyglot text-to-speech synthesis. *Speech Commun.* 49, 9 (2007), 697–724.

[67] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artific. Intell. Res.* 65 (2019), 569–631.

[68] Sujan Kumar Saha and Ankur Priyadarshi. [n.d.]. A study on the importance of linguistic suffixes in Maithili POS tagger development. In *Proceedings of the 7th International Conference on Mining Intelligence and Knowledge Exploration (MIKE’19)*. Lecture Notes in Computer Science, vol. 11987. Springer, 11–20. DOI: 10.1007/978-3-030-66187-8_2.

[69] Rajeev Sangal, Sushma Bendre, Dipti Sharma, and Prashanth Mannem. 2007. Introduction to shallow parsing contest on south asian languages. In *Proceedings of the IJCAI Workshop On Shallow Parsing for South Asian Languages (SPSAL’07)*. 1–8.

[70] A. Sarkar, A. De Roeck, and P. Garthwaite. 2004. Easy measures for evaluating non-English corpora for language engineering: Some lessons from Arabic and Bengali. Technical report, Dept. of Comp., Faculty of Math. and Comp., Open University, Walton Hall, UK.

[71] Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL’03)*. Association for Computational Linguistics, 134–141. https://doi.org/10.3115/1073445.1073473.

[72] Vijay Kumar Sharma and Namita Mittal. 2018. Cross-lingual information retrieval: A dictionary-based query translation approach. In *Advances in Computer and Computational Sciences*. Springer, 611–618.

[73] Gary F. Simons and Charles D. Fennig. 2017. *Ethnologue: Languages of Asia*. SIL International.

[74] Anil Kumar Singh. 2006. A computational phonetic model for Indian language scripts. In *Proceedings of the 5th International Workshop on Writing Systems: Constraints on Spelling Changes*. 1–19.

[75] Anil Kumar Singh. 2006. Study of some distance measures for language and encoding identification. In *Proceedings of the Workshop on Linguistic Distances*. 63–72.
[76] Anil Kumar Singh. 2008. A mechanism to provide language-encoding support and an nlp friendly editor. In Proceedings of the 3rd International Joint Conference on Natural Language Processing.

[77] Anil Kumar Singh and Jagadeesh Gorla. 2007. Identification of languages and encodings in a multilingual document. In Proceedings of the 3rd Web as Corpus Workshop, Incorporating CleanEval: Building and Exploring Web Corpora (WAC’07), Vol. 4. Presses Univ. de Louvain, 95.

[78] Anil Kumar Singh, Kiran Pala, and Harshit Surana. 2008. Estimating the resource adaption cost from a resource rich language to a similar resource poor language. In Proceedings of the Language Resources and Evaluation Conference (LREC’08).

[79] Loitongbam Gyanendro Singh, Lenin Laitonjam, and Sanasam Ranbir Singh. 2016. Automatic syllabification for manipuri language. In Proceedings of the 5th International Conference on Computational Linguistics (COLING’16).

[80] Srishiti Singh. [n.d.]. Web drawn corpus for Bhojpuri. In Proceedings of the Conference on NLP, MGAHV, Wardha.

[81] Srishiti Singh and Girish Nath Jha. 2015. Statistical tagger for Bhojpuri (employing support vector machine). In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI’15). IEEE, 1524–1529.

[82] Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. 21–24.

[83] Brij Mohan Lal Srivastava, Sunayana Sitaram, Rupesh Kumar Mehta, Krishna Doss Mohan, Pallavi Matani, Sandeep-kumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. 2018. Interspeech 2018: Low-resource automatic speech recognition challenge for Indian languages. In Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages. 11–14.

[84] Erik Sterneberg. 2012. Language identification of person names using cascaded SVMs. Bachelor’s Thesis, Uppsala University, Uppsala.

[85] Jörg Tiedemann. 2017. Cross-lingual dependency parsing for closely related languages—Helsinki’s submission to VarDial 2017. In Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial’17). Association for Computational Linguistics, 131–136. https://doi.org/10.18653/v1/W17-1216

[86] Zankhana B. Vaishnav and Priti S. Sajja. 2019. Knowledge-based approach for word sense disambiguation using genetic algorithm for gujarati. In Proceedings of the Conference on Information and Communication Technology for Intelligent Systems (ICTIS’19). Springer, 485–494.

[87] Manindra K. Verma. 1991. Exploring the parameters of agreement: The case of Magahi. Lang. Sci. 13, 2 (1991), 125–143.

[88] Haoxing Wang and Laurianne Sitbon. 2014. Multilingual lexical resources to detect cognates in non-aligned texts. In Proceedings of the 3rd International Joint Conference on Natural Language Processing (ICNLP’14). Association for Computational Linguistics, 1–17. Retrieved from https://aclweb.org/anthology/W14-1026.

[89] Martin Wynne. 2005. Developing Linguistic Corpora: A Guide to Good Practice. Vol. 92. Oxbow Books Oxford.

[90] Yogendra P. Yadava, Oliver Bond, Irina Nikolaeva, and Sandy Ritchie. 2019. The syntax of possessor prominence in Maithili. Prom. Intern. Possess. (2019), 39–79. DOI: 10.1093/oso/9780198812142.003.0002

[91] Yin-Lai Yeong and Tien-Ping Tan. 2011. Applying grapheme, word, and syllable information for language identification in code switching sentences. In Proceedings of the International Conference on Asian Language Processing. IEEE, 111–114.

[92] Jia-Li You, Yi-Ning Chen, Min Chu, Frank K. Soong, and Jin-Lin Wang. 2008. Identifying language origin of named entity with multiple information sources. IEEE Trans. Audio Speech Lang. Process. 16, 6 (2008), 1077–1086.

[93] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Siwun Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, et al. 2018. Language identification and morphosyntactic tagging. The second VarDial evaluation campaign. In Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial@COLING’18), Santa Fe, New Mexico. Association for Computational Linguistics, 1–17. https://aclanthology.org/W18-3901.

[94] Meng Zhang, Haoruo Peng, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Bilingual lexicon induction from non-parallel data with minimal supervision. In Proceedings of the 31st AAAI Conference on Artificial Intelligence.
[97] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In Advances in Neural Information Processing Systems. MIT Press, 649–657.

[98] Yujie Zhang. 2019. Improving performance of NMT using semantic concept of wordnet synset. In Proceedings of the 14th China Workshop on Machine Translation (CWMT’18), Vol. 954. Springer, 39.

[99] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 1568–1575. https://doi.org/10.18653/v1/D16-1163

Received May 2020; revised November 2020; accepted March 2021