A POS-Tagged Corpus for Dogri: Development and Annotation using DogriTag

Vipul Saluja
University Department of Computer Science
RD & SH National College
University of Mumbai
Mumbai, 400093, India

Jyotshna Dongardive
University Department of Computer Science
University of Mumbai
Mumbai, 400093, India

ABSTRACT
This paper discusses about the process used to create a linguistically selected and manually annotated Part of Speech (POS) tagged corpus for Dogri. Dogri is a low resource and Indo Aryan language that is spoken in the Indian Union Territory of Jammu and Kashmir and in some regions of Pakistan. Dogri is poorly represented in Natural Language Processing (NLP) despite the sufficient number of speakers and the official recognition. This is due to the absence of resources such as defined tag sets, annotated corpora and annotation specific tools. To fill this gap, a POS tagged Dogri corpus was developed from a domain-specific subset of the Linguistic Data Consortium for Indian Languages (LDC-IL). This corpus has about 25,000 sentences (approximately 400,000 tokens). A specialized web platform named DogriTag was developed that can track audits and make semi-automated tag suggestions, to do the annotation. To check the quality of the annotations, inter annotator agreement analysis was used. The results show a Cohen's Kappa score of 0.89 indicating a lot of agreement. This resource is very important for making NLP tools like POS taggers, syntactic parsers, and morphological analyzers for Dogri. Future work will include adding more tags, using pretrained language models to transfer information between languages, and covering more areas.

Keywords
Dogri language; Part of speech tagging; ILPOSTS; low resource NLP; annotated corpus; Indian languages; inter annotator agreement; linguistic annotation; web-based annotation tool; Indo Aryan languages

1. INTRODUCTION
The availability of linguistically annotated corpora has greatly advanced research and applications in Natural Language Processing (NLP) [1]. These corpora offer the empirical basis for developing and evaluating syntactic and semantic models across various languages. However, for low resource languages like Dogri, a significant gap still exists in annotated data, tools, and standardized tagsets, which limits both computational modeling and formal linguistic analysis [2,3]. Dogri is an Indo Aryan language spoken mainly in Jammu and Kashmir and has official status under the Eighth Schedule of the Indian Constitution [4]. Even though Dogri is being spoken by around 5 million people [5], it still remains underrepresented in digital and computational linguistic resources [6,7]. Unlike high resource languages like Hindi or English there is a lack of large scale annotated Dogri corpora. This hinders the development of robust NLP tools like POS taggers, dependency parsers, and machine translation systems [8,9]. Part of speech tagging assigns each word of a text the proper syntactic tag based on its context. [10]. POS tagging, also known as grammatical tagging, automatically assigns Part of speech tags such as verbs, adjectives, adverbs, nouns, etc. to words in a sentence. The applications of POS tagging include machine translation, word sense disambiguation, question answering parsing, and so on. [6,11]. Despite advancements in POS tagging for many Indian languages, Dogri continues to lack sufficient annotated resources and computational tools. [2,12]. This paper introduces a manually annotated, POS tagged corpus for Dogri, created using a subset of the ILPOSTS framework [13]. The project involved corpus compilation, preprocessing, tagset adaptation, and manual annotation by trained linguists. The annotation quality was validated through Cohen's Kappa [14]. The key contributions of this work include:

- A publicly available, manually annotated POS tagged corpus for Dogri
- A Dogri specific adaptation of the ILPOSTS tagset
- A lightweight, web-based annotation platform optimized for low resource language workflows
- An empirical reliability analysis of linguistic annotation using Cohen's Kappa

The rest of this paper is organized as follows: Section 2 reviews prior research, Section 3 details the corpus and annotation methodology, Section 4 discusses results and linguistic insights, and Section 5 concludes with future directions.

2. LITERATURE REVIEW
In the past, Dogri has been studied mainly from a linguistic point of view. However, there is limited work on Dogri in the field of Natural Language Processing (NLP) [2,9]. This section reviews the earlier work on Dogri's structure and grammar linking it to the challenges of Part of speech tagging. It also points out a major gap — Dogri lacks computational resources like annotated corpora, tagsets, and other processing tools. In order to bridge this gap, it is necessary to build NLP tools that align with Dogri's own linguistic features and style.

2.1 Linguistic Profile of Dogri
Dogri is an Indo Aryan language mainly spoken in Jammu and Kashmir and some parts of Himachal Pradesh and Punjab. It is also spoken in some parts of Pakistan. In 2003, it was included in the Eighth Schedule of the Indian Constitution and is one of the 22 official languages scheduled in the Constitution of India [4]. Since 2020, it has been recognized as the official language of the J&K [15]. Historically, Dogra Akkhar was used to write Dogri which is now replaced by Devanagari.

Despite its rich literary and cultural history, Dogri continues to be classified as a low resource language in computational linguistics. This is due to limited digital corpora, linguistic tools, and annotated datasets available for research and development [6,7].

2.2 Morphosyntactic Features Relevant to POS Tagging

Dogri has several morphosyntactic features that influence the development and performance of POS tagging systems. It has rich inflectional morphology, agglutinative postpositions and a flexible word order. Also, Dogri speakers often engage in code-switching with Hindi, Urdu, and English. These features make tasks like tokenization, syntactic disambiguation, and sequence modeling more complex [8,17].

2.2.1 Nominal and Verbal Morphology

Dogri exhibits rich inflectional morphology across both nominal and verbal forms. Nouns inflect to show gender – masculine or feminine, number – singular or plural, and case – direct, oblique, or vocative. Verbs also inflect according to tense, aspect, and mood. Depending on the context, they sometimes agree with the subject or sometimes with the object. This agreement reflects person, number and gender distinctions [9,18].

Examples:

- *रीता खड़याल ने पंजाबी गीत पेश कीता (Rita Khadyaal presented a Punjabi song)*
- *रीता खड़याल ने गजल पेश कीता (Rita Khadyaal presented a ghazal)*

In the above examples, कीता (kītā) agrees with the masculine noun *गीत (gīt – song)*, while कीती (kītī) agrees with the feminine noun *गजल (ghazal)*.

These morphological characteristics necessitate elaborated tagsets, in most instances demand the incorporation of morphological analyzers in the POS tagging systems [13].

2.2.2 Agglutination

Dogri has postpositions that are affixed to oblique form of nouns. This leads to agglutinative constructions in which several morphemes are combined in various combinations to convey subtle meaning. These structures render it challenging to tokenize and need meticulous segmentation [8].

Example:

- घोषणा दे कन्नै गै सरकार दी रेल दौड़ाई दित्ती। (With the announcement, the government's campaign train was launched).
- जम्मू तवी थमां दरभंगा बश्कार चलङन। (Trains run from Jammu Tawi to Darbhanga).
- श्रेणियें दे कराये च कमी कीती ऐ। (A reduction was made in the fares of all classes).

In these examples, postpositions like कन्नै, थमां, बश्कार, and च follow oblique noun forms. If such expressions are not properly segmented, there is a risk incorrect tagging. This becomes a major challenge for annotating Dogri.

2.2.3 Flexible Word Order

Dogri generally follows a Subject Object Verb (SOV) word order. However, word placement in a sentence can be changed to show focus, topic, or emphasis. This flexibility makes the language more expressive but also complicates POS tagging, as it is often dependent on word position [15,19].

Examples:

- *रीता खड़याल ने पंजाबी गीत पेश कीता।* (Rita Khadyaal presented a Punjabi song).
- *पंजाबी गीत रीता खड़याल ने पेश कीता।* (A Punjabi song was presented by Rita Khadyaal).

Context-aware models that take advantage of both syntactic and semantic cues, in addition to surpassing positional heuristics, are necessary to handle such syntactic variations.

2.2.4 Code-Switching

Code-switching between Hindi, Urdu, and English is a common practice in Dogri. Particularly in informal speech and media content, people frequently mix words from these languages. Due to this reason, developing purely monolingual POS tagging systems is challenging. [2,20].

Examples:

- जे मकबूलियत सिर्फ *Media* दे तैह्त गै होंदी तां साढ़े अज्जे दे शायरें कोल इक थमां बधियै जरी'या हा। (If popularity depended only on media, today's poets would have an easier way to succeed).
- ...शायरी दा Institution इंदे आस्तै चाली-पंजाह् रपेऽ दे सरकारी *Contract* दी लालसा बे-मैह्ने ही निता इंदे समकालीनें दी tragedy ए ऐ। (The desire for a forty or fifty rupee government contract influenced the poetic institutions of his contemporaries).

In the above examples, English words such as Media, Institution and Contract are used within Dogri sentences. They follow Dogri style usage and inflection patterns, blending seamlessly into the local grammatical framework This shows how natural code switching occurs in everyday use as English words adjust to Dogri morphology and syntax. For accurate POS tagging, a system must be able to handle such language mixing which would dynamically recognize the language and apply the appropriate grammatical tags.

2.2.5 Summary

In summary, Dogri's linguistic features like rich inflection, agglutinative structures, flexible word order and frequent code-switching all make POS tagging more complex. Table 1 summarizes these features with examples their impact on POS tagging.

Table 1: Morphosyntactic Features of Dogri and Their Impact on POS Tagging

Feature	Description	Example	Impact on POS Tagging
Inflectional Morphology	Rich noun/verb inflections for gender, number, case, tense, etc.	कीता vs कीती	Requires fine grained morphological tagging
Agglutination	Oblique nouns with bound postpositions	घोषणा दे कन्नै	Complex tokenization; multi token expressions
Free Word Order	Non-rigid sentence structure for emphasis/focus	रीता खड़याल ने पंजाबी गीत पेश	Requires context sensitive models

		कीता vs पंजाबी गीत रीता खड़याल ने पेश कीता	
Code-Switching	Integration of Hindi, Urdu, and English words	Media दे तैह्त, सरकारी Contract दी लालसा	Needs multilingual lexicons and tagging support

2.3 Motivation and Research Gap

Although Dogri's grammar and morphology have been studied descriptively, these insights have rarely been converted into computational resources. Currently there is no extensive, linguistically validated corpus or standardized tagset that encapsulates the language's inflectional richness, syntactic flexibility, and multilingual context. As a result, even a fundamental NLP task like POS tagging cannot be performed reliably.

Previous studies describe the linguistic functioning of Dogri but not its computational modeling. The key gap is the link between the descriptive linguistics and the development of computational resources. This study attempts to bridge that gap by developing a manually annotated POS tagged corpus for Dogri through a web-based annotation platform, DogriTag.

3. CORPUS AND ANNOTATION METHODOLOGY

This section explains the step-by-step process used to manually annotate the POS-tagged corpus for Dogri. The methodology comprised corpus collection, data cleaning and preprocessing, refinement of the tagset, annotation practices, and tool development. It also covers the evaluation of inter-annotator agreement.

3.1 Corpus Source

The data source used for this study is the Dogri corpus dataset developed by the Linguistic Data Consortium for Indian Languages (LDC-IL). A total of about 25,000 sentences comprising nearly 350,000 tokens were selected from the Aesthetics domain of the dataset.

3.2 Preprocessing Steps

The preprocessing stage played an important role in preparing the raw Dogri text for accurate Part of Speech annotation. The preprocessing pipeline included three main stages i.e. script standardization, sentence segmentation, and rule-based tokenization. Figure 1 illustrates the preprocessing pipeline stages showing the stages of script standardization, sentence segmentation and tokenization.

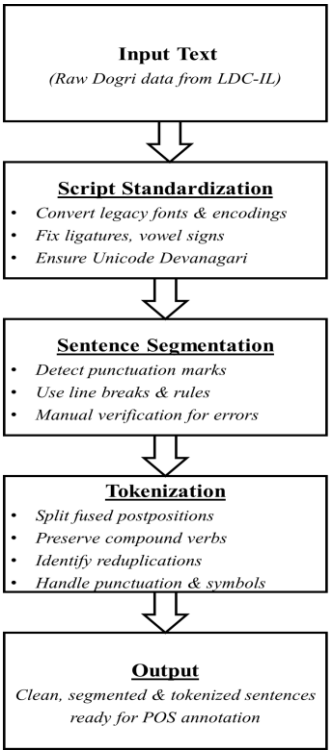

Fig 1: Preprocessing pipeline for Dogri corpus

The steps for preprocessing are discussed below:

1. **Script Standardization:** In script standardization stage, the raw corpus was converted to a standard Unicode compliant Devanagari script. This process removed any inconsistencies caused by legacy fonts and mixed encodings. This also ensured uniform script before any linguistic processing. For example, tokens such as उ'आं and फ्ही containing non-standard apostrophe or half-form markers were replaced with their Unicode equivalents ऊआं and फी. Also forms like खोह्लियै and अड्रैस्स were standardized to खोलियै and ऐड्रेस.

2. **Sentence Segmentation:** In the next stage sentence segmentation was done using a semi-automated method. A rule-based script scanned the text and used punctuation markers and clues like line breaks to detect sentence boundaries. For example, it split text wherever it found "।" or "?" or "!". After the automated pass, the segmented sentences were manually reviewed to correct cases where missing punctuation marks, quotes or abbreviations caused false splitting or merging of sentences.

3. **Tokenization:** The final stage was rule-based tokenization. It involved breaking up of sentences into tokens using linguistic rules based on Dogri-specific grammatical patterns. The tokenizer preserved compound verb forms such as कर ल्ये आ ("did and came"), separated fused postpositions from nouns (e.g., घरेच → घर + एच) and recognized reduplicated expressions like फेर फेर ("again and again") as valid pairs.

All the above three steps transformed the raw Dogri text into a clean and linguistically structured format. The processed corpus was consistent, well-segmented and ready for accurate Part of Speech tagging.

3.3 Design of the Dogri POS Tagset

An important step before starting manual annotation was to design a Dogri POS tagset that could represent grammatical structure of Dogri and be compatible with frameworks used for other Indian languages. The steps followed in developing the tagset are as below:

1. **Reference Framework Selection:** The ILPOSTS (Indian Languages Part-of-Speech Tagset) which already defined POS categories for several Indian languages such as Hindi, Bengali and Tamil it was taken as the base framework.
2. **Customization for Dogri:** The framework was customized to suit Dogri's morphological and syntactic features. Some categories to retained, dropped or added as needed. For example, the postposition (PSP) tag was retained the foreign word (FW) tag was add to capture code switching and categories such as honorifics and classifiers which are common in Dravidian languages were dropped as they do not occur in Dogri.
3. **Tagset Validation:** In order to ensure coverage of all tags the customized tagset was tested on a small subset of the corpus containing around 2000 tokens. The validation checked whether all common grammatical forms were represented, whether any tags were too rare and could be merged and whether any new categories were needed. This validation confirmed that the final sixteen tags covered over 99% of the word types found in the Dogri corpus.
4. **Final Tagset:** The final Dogri POS tagset design included sixteen categories that made it simple to apply while maintaining consistency across the entire corpus.

This finalized tagset as shown in Table 2 became the standard reference for all subsequent manual annotation. Each tag includes a short label to represent the tag, a brief description and examples in both Devanagari and Roman scripts to guide annotators during the tagging process.

Table 2: Customized ILPOSTS POS Tagset for Dogri

Tag	Category	Description	Example (Devanagari)	Example (Roman)
NN	Noun	Common/proper noun	बच्चा	bachcha
JJ	Adjective	Descriptive modifiers	सुन्दर	sundar
VM	Verb Main	Main verb	खाता	khata
VAUX	Verb Auxiliary	Helping verb	रहा	raha
RB	Adverb	Modifies verb/adjective	जल्दी	jaldi
PSP	Postposition	Follows an oblique noun	के	ke
PRP	Pronoun	Personal or demonstrative pronoun	वह	wah
DEM	Demonstrative	Points to a noun	यह	yah
QF	Quantifier	Quantity	कुछ	kuch
		expressions		
CC	Conjunction	Connects clauses/phrase	और	aur
INTF	Intensifier	Degree modifier	बहुत	bahut
INJ	Interjection	Exclamatory words	अरे!	are!
RP	Particle	Discourse or aspectual marker	ही	hi
NEG	Negation	Negative marker	नहीं	nahin
SYM	Symbol	Non-alphabetical symbols	@, #	@, #
FW	Foreign Word	Hindi/Urdu/English borrowings	मोबाइल	mobile

3.4 Annotation Tool: DogriTag

In order to make POS tagging faster and consistent, a lightweight web-based annotation tool called DogriTag was developed. Its interface was designed to be simple and clear and as per the Dogri tagset developed. As shown in Figure 2, the tool displays each token with a dropdown menu containing Part of Speech tags. The tool also provided auto suggestions based on previously annotated token–tag pairs. This feature reduced redundant tagging efforts and ensured greater consistency across similar contexts.

Fig 2: Screenshot of DogriTag Annotation Tool Interface

In order to be compatible with other NLP tools, the tagged data could be exported in both CSV and CoNLL-U formats. DogriTag was found to be both efficient and reliable, ensuring uniform annotations and faster tagging without any loss of accuracy.

3.5 Manual Annotation

The manual annotation was carried out using the DogriTag tool described in Section 3.4. Two annotators trained in Dogri and Hindi tagged the corpus independently. The Dogri tagset as described in Section 3.3 was used as the reference throughout the process. The most suitable POS label was assigned to each token based on its grammatical function and the context in which it was used. In cases of ambiguity, the annotators deliberated and selected a single, consistent tag. This systematic approach ensured sure that the corpus was correctly and consistently tagged, providing a solid foundation for upcoming Dogri NLP research.

3.6 Methodological Summary

This work combined linguistics and computer algorithms to develop a credible Dogri POS-tagged corpus. The workflow commenced by standardization of scripts, segmentation of sentence and Marking with Tokens, and these aided in transforming the raw text into clean and structured data. A Dogri tagset based on the Ilposts frame was modified to reflect the grammatical peculiarities of Dogri, and maintain the system simple and convenient to be annotated with. The DogriTag tool minimized the process by making tagging faster and more reliable. The process of manual annotation was carefully performed to ensure accuracy and consistency within the corpus.

The result is a resource that is linguistically rich and well-structured which forms the foundation of future NLP applications in Dogri. In addition to this, the same methodological framework can be used with other low-resource Indian languages, allowing to reinforce the larger ecosystem of Indic language processing.

4. RESULTS AND DISCUSSION

4.1 Corpus Statistics

The completed POS tagged Dogri corpus consisted of 398,765 tokens of 19,863 sentences that were obtained out of the Aesthetics domain of the LDC-IL Dogri data. A preprocessing process, which involved script standardization, sentence segmentation, and tokenization was conducted. These measures contributed to the production of quality linguistic data, which can be used in downstream POS tagging processes. A subset of 16 POS tags of the ILPOSTS framework was selected to make sure that the tagset covered all the necessary grammatical information, but was not too complex to be manually annotated uniformly. This is among the biggest and well annotated corpus of a low resource Indo Aryan language.

Table 3 shows the overall corpus statistics, including total tokens, sentence count, average sentence length, and other relevant details.

Table 3. Overview of the Annotated Dogri Corpus

Metric	Value
Total Tokens	398,765
Unique Tokens	35,781
Total Sentences	19,863
Average Sentence Length	20.16 tokens
Domain	Aesthetics
Script	Devanagari

4.2 POS Tag Distribution

The frequency distribution of POS tags gives some insights into the linguistic structure of Dogri. Figure 3 and Table 4 shows that commons nouns (NN) constitute 26.4-percent of total tokens, as a result of the noun-heavy syntax structure of Indo Aryan languages. There are also high frequencies of main verbs (VMs) and pronouns (PRPs), adjectives and postpositions. This distribution shows the order of subject object verb (SOV) in Dogri and the use of postpositional case marking.

Table 4: POS Tag Frequency Distribution in the Dogri Corpus

Tag	Description	Frequency	Percentage
NN	Common Noun	105,403	26.4%
VM	Main Verb	79,184	19.8%
PRP	Pronoun	54,290	13.6%
JJ	Adjective	28,317	7.1%
RB	Adverb	19,845	5.0%
VAUX	Auxiliary Verb	17,204	4.3%
PSP	Postposition	16,798	4.2%
QF	Foreign/Code-Switch	2,317	0.6%
Others (8 tags)	—	75,407	18.9%
Total		398,765	100%

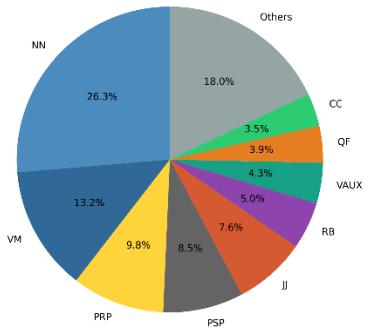

Fig 3: POS Tag Distribution

The low percentage of the auxiliary verbs and adverbs (4.3 and 5.0 respectively) is typical of morphologically rich South Asian languages and in such cases, tense, aspect and modality tend to be represented through suffixation and not the use of auxiliary structure.

4.3 Sample Annotations

Illustrative examples of Dogri sentence annotations are presented in Table 5, where each token is tagged according to ILPOSTS guidelines. The examples highlight common syntactic structures, multi-token verb forms, and noun-adjective agreement. Such examples also guided the training of annotators and the refinement of guidelines.

Table 5: Sample Annotated Sentences from the Corpus

Sentence	Token/POS
अरुणा ने फ्ही बी कोई जवाब नेईं दित्ता अरुणा।	अरुणा/N_NP, ने/V_VA, फ्ही/N_NC, बी/C_CCD, कोई/N_NC, जवाब/N_NC, नेईं/V_VM, दित्ता/V_VM, अरुणा/N_NP, ।/PU_PU
छड़े 12.6 फीसदी भारतीय कौंपनियें दा छंटनी उप्पर बचार : सर्वेक्षण	छड़े/N_NC, 12.6/NUM_NUMR, फीसदी/N_NC, भारतीय/N_NC, कौंपनियें/N_NC, दा/N_NC, छंटनी/N_NC, उप्पर/N_NST, बचार/N_NC, :/PU_PU, सर्वेक्षण/N_NC,

नमीं दिल्ली, 22 फरवरी।	नमीं/N_NST, दिल्ली/N_NC, ,/PU_PU, 22/NUM_NUMR, फरवरी/N_NST, ।/PU_PU
बदलाव उस्सै चाल्ली दे हुंदे न जैसे पैह्ल्ले कीत्ते जंदे हे।	बदलाव/N_NC, उस्सै/N_NC, चाल्ली/N_NC, दे/V_VM, हुंदे/V_VM, न/J_JJ, जैसे/J_JJ, पैह्ल्ले/N_NC, कीत्ते/N_NC, जंदे/N_NC, हे/V_VM, ।/PU_PU
पीटरसन दी धैर्यपूर्ण पारी न इंगलैंड गी संभालेआ किंग्सटन, 5 फरवरी।	पीटरसन/N_NC, दी/P_PPR, धैर्यपूर्ण/J_JJ, पारी/N_NC, न/C_CCD, इंगलैंड/N_NC, गी/N_NST, संभालेआ/N_NC, किंग्सटन/N_NC, ,/PU_PU, 5/NUM_NUMR, फरवरी/N_NST, ।/PU_PU
कुसै शा दूर जाने आस्तै गै में बम्बई थमां दिल्ली अपनी बदली कराई ही।	कुसै/N_NC, शा/N_NC, दूर/N_NC, जाने/V_VM, आस्तै/N_NST, गै/V_VM, में/PP_PP, बम्बई/N_NP, थमां/N_NC, दिल्ली/N_NP, अपनी/P_PPR, बदली/N_NC, कराई/N_NV, ही/A_AMN, ।/PU_PU

4.4 Inter annotator Agreement

In order to determine the consistency of the annotation, as well as to evaluate the reliability of the suggested tagset and guidelines, inter-annotator agreement was computed with the help of Cohen Kappa.

In the first experiment, a sample size of 5,000 tokens which was selected randomly was annotated by two trained annotators. A Cohen's Kappa score of 0.89 indicated a high degree of inter-annotator agreement which confirmed that the tagset used by ILPOSTS is intuitive and operationally viable.

The majority of the conflicts were witnessed in multi token verb structures and utilization of auxiliaries. Table 6 indicates that annotators disagreed on the analysis of the progressive verb phrase "जा रही" as either a main verb (VM) or auxiliary (VAUX).

Table 6: Example of Annotation Disagreement

Token Sequence	Annotator A	Annotator B	Final Decision
ओ कीता ऐ	कीता / VM ऐ / VAUX	कीता / VAUX ऐ / VAUX	कीता / VM ऐ / VAUX

A second round of validation involved the annotation of a random set of 3,000 tokens by two native Dogri-speaking linguists with a Cohen Kappa of 0.8967 and a total agreement rate of 91.6, as shown in Table 7. The most common disagreements were the ones containing such tags as VAUX vs VM, QF (foreign/code-switched words) vs. NN.

Table 7: Cohen's Kappa Score and Agreement rate between linguists

Measure	Value
Tokens Sampled	3,000
Agreement Rate	91.6%
Cohen's Kappa (κ)	0.8967
Most Disagreed Tags	VAUX vs VM, QF vs NN

The agreement scores obtained in the two validation rounds are shown in Figure 4. These results confirm the clarity and applicability of the tagset in the Dogri context. They also confirm the quality and strength of the annotation required in downstream POS tagging work in a low resource language environment.

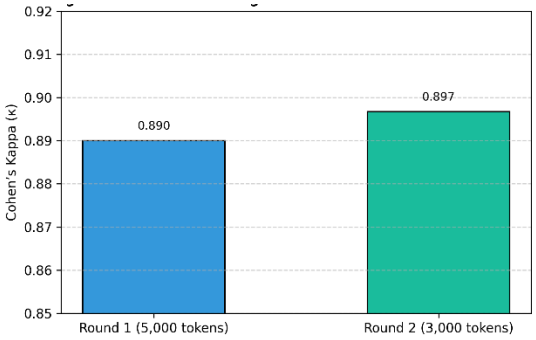

Fig 4: Inter-Annotator Agreement Across Two Validation Rounds

4.5 Annotation Challenges and Linguistic Insights

The annotation phase exposed a number of linguistically important issues, which indicate the complexity of the Dogri syntax and semantics. These lessons were used to improve the annotation guidelines and led to an increased inter-annotator agreement.

4.5.1 Contextual Disambiguation

One of the issues that were recurrent during the annotation was lexical ambiguity. As an example, the word concept "साफ़" (sāf) had several meanings based on its syntactic and semantic contexts. Table 8 demonstrates that the same token may be used as an adjective (J_JJ) or an adverb (A_AMN).

Table 8: Contextual variation in POS tagging of the word "साफ़"

Sentence	Token	POS Tag
ओसने साफ़ कपड़े पाये।	साफ़	J_JJ
ओ साफ़ बोलदा ऐ।	साफ़	A_AMN

4.5.2 Code-Switching

The Dogri speakers often mix Hindi, Urdu and English words in their daily conversation. Therefore, English tokens such as "national" and "students" were found in the corpus. These were labeled as foreign words and labeled with QF. This approch guaranteed a steady tagging as well as enriching the corpus of subsequent studies in the areas of bilingual NLP and sociolinguistics.

4.5.3 Multi-word Expressions and Conjunct Verbs

Dogri mostly uses conjunct verbs, nominal constructions that are used in a semantic unit, but composed of more than one word. Indicatively, e.g. "ध्यान देना" (pay attention) is a verb with the meaning of a single verb, but was syntactically divided in the course of annotation to ensure computational uniformity. Table 9 shows that "ध्यान" was tagged as noun (NN), and "देना" was tagged as the main verb (VM).

Table 9: POS tagging of a conjunct verb expression

Expression	Token	POS Tag
ध्यान देना	ध्यान	NN
	देना	VM

These challenges reveal the delicate balance between Dogri's linguistic richness and its computational representation, offering insights for multilingual POS tagging frameworks.

4.6 Discussion

This annotation resulted in the formation of the Dogri POS-tagged corpus of high quality. It is computationally useful as well as linguistically rich. This can serve as a gold-standard dataset in further Natural Language Processing (NLP) studies. The tagset applied was a reduced though complete version of the ILPOSTS scheme and it had 16 Part of speech categories. This was to ensure a balance between linguistic and annotation consistency. There was high agreement among the annotators and Cohens Kappa indicated a value of 0.8967. This is an indicator of high dependability and intelligibility of guidelines applied.

There are a number of challenges that were noted during the process of annotation: polysemy, code-switching, and multi-word expressions. They were useful in fine-tuning the tagging rules and making it more accurate. The combination of these efforts created a strong and stable resource. It fosters machine learning training of POS tagging, it advances morphosyntactic studies and NLP development of Dogri, and other low resource Indo Aryan languages.

5. CONCLUSION AND FUTURE WORK

This study developed a manually annotated POS-tagged corpus containing more than 400,000 tokens of the Aesthetics domain of the LDC-IL Dogri data. It employed a subset of ILPOSTS tagset. It was done through a careful process of annotation that included customization of tagsets, guideline creation and creation of an annotation tool DogriTag.

High agreement has been observed among the linguists and this is supported by the fact that the Cohen Kappa score of 0.8967 indicates that the guidelines were clear, consistent, and easy to use. Moreover, the tagging rules were enhanced with the aid of error analysis and linguistic review. These developments enriched and improved the corpus in terms of syntax and meaning.

This corpus is a solid base of the NLP research in Dogri. It is applicable in training and benchmarking POS tagging models particularly in low resource language environment. It also contributes to Dogri linguistic record. Meanwhile, it helps calculate research on less-represented Indo Aryan languages.

The corpus may be extended to include additional domains in the future like journalism, law and day-to-day conversations. It is also possible to extend the tagset to more detailed grammatical categories so as to do more detailed syntactic work. In future, semiautomatic and LLM-assisted annotation approaches can be used, which should help to create a corpus in less time, without compromising the quality. The second one is to train and benchmark other POS tagging models with supervised, unsupervised and transfer learning-based models. These will be used to create trustworthy and replicable NLP systems of low resource languages in India such as Dogri.

6. REFERENCES

[1] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O'Reilly Media.

[2] Sreelekha, S., & Bhattacharyya, P. (2020). Low-resource language processing: A review. ACM Computing Surveys, 53(4), 1–34. https://doi.org/10.1145/3397512

[3] Zeman, D. (2008). Reusability of morpho-syntactic annotation across languages. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). European Language Resources Association (ELRA).

[4] Government of India. (2003). The Constitution (Eighty-second Amendment) Act, 2003. Ministry of Law and Justice, Government of India.

[5] Ethnologue. (2022). Dogri language profile. SIL International. https://www.ethnologue.com/language/dgo

[6] Bhattacharyya, P. (2016). IndoWordNet and its application to Indian language processing. IEEE Transactions on Knowledge and Data Engineering, 28(10), 2672–2682. https://doi.org/10.1109/TKDE.2016.2571686

[7] Jha, G. N. (2010). The TDIL program and the Indian language corpora initiative. In Proceedings of the LREC Workshop on Language Resources and Human Language Technologies for Dravidian Languages. European Language Resources Association (ELRA).

[8] Mishra, A., & Jha, G. N. (2012). Creating POS tagger for Maithili using IL-POST tagset. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA).

[9] Sunil Kumar, & Prem Kumar. (2020). Developing a POS tagset for Dogri language. Language in India, 20(1), 186–200.

[10] Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press.

[11] Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed., draft). Pearson. https://web.stanford.edu/~jurafsky/slp3/

[12] Jamatia, A., Das, D., & Bandyopadhyay, S. (2015). Part-of-speech tagging for code-mixed social media text. In Proceedings of the 12th International Conference on Natural Language Processing (ICON-2015). NLP Association of India.

[13] Begum, R., Sinha, R., Husain, S., Dhwaj, A., & Bharati, A. (2008). Developing a syntactic and lexical resource for Indian languages: The Hindi and Urdu Treebank. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). European Language Resources Association (ELRA).

[14] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

[15] Government of India. (2020). The Jammu and Kashmir Official Languages Act, 2020. Ministry of Law and Justice. https://legislative.gov.in/actsofparliamentfromtheyear/jammu-and-kashmir-official-languages-act-2020

[16] Shackle, C. (1979). Problems of Classification in the Indo-Aryan Languages: The Case of Dogri. Transactions of the

Philological Society, 77(1), 191–210.

[17] Bhat, R. A. (2009). Morphosyntactic Properties of Dogri. Journal of South Asian Languages and Linguistics, 1(1), 45–67.

[18] Bharati, A., Sharma, D. M., Sangal, R., & Bai, L. (2006). AnnCorra: POS and Chunk Annotation Guidelines for Indian Languages. IIIT Hyderabad.

[19] Schmidt, R. L. (1999). Urdu: An Essential Grammar. Routledge.

[20] Kumawat, D., & Jain, V. (2015). Parts of Speech Tagging for Dogri Language. International Journal of Computer Applications, 111(14), 16–20.