==========
Language in India www.languageinindia.com **ISSN 1930-2940 Vol. 18:1 January 2018**
India's Higher Education Authority UGC Approved List of Journals Serial Number 49042
==========

Developing POS Tagset for Dogri

Sunil Kumar, M.A., M.Phil., B.Ed.
Central Institute of Indian Languages

==========

Abstract

Annotated Text Corpora is an important resource for advances in Natural Language Processing (NLP) research and for developing different language technologies. The annotation of corpora is done using a set of tags, which mark the linguistic properties of a word, sentence or discourse. In corpus linguistics the parts of speech tagging is also called as grammatical tagging or word category disambiguation. This is a process of marking up the words in text or corpus as corresponding to a particular part of speech based on both its definition, as well as its context i.e. the relationship with adjacent and related words in phrase, sentence, or paragraph. The corpora annotated with various linguistic information not only form a precious resource for language technologies but also involves large amount of effort and time. Therefore, it is important to create corpora which once created can be used for various purposes. In softwares like Machine Translation, Information Retrieval, speech recognition and other related areas, the significance of large annotated corpora in the present day is widely known. This paper makes an attempt to provide a structure of POS tag set module for Dogri language, one among the languages of Indo-Aryan family.

Key Words: Corpora, Dogri, Part-of-Speech (POS), Tagging, Tagset.

Dogri Language

Dogri is one of the modern Indo-Aryan languages along with Punjabi which have developed tonal contrasts. It has three tones: low / ` / mid / - / and high / ´/. Dogri is a morphologically rich language having the pre-dominant word order of Subject-Object-Verb (SOV) with a flexibility to rearrange the constituents as many Indian languages allow. Nouns are generally inflected for number, gender and case. There are two numbers –singular and plural; two genders-masculine and feminine; and three cases- simple, oblique and vocative. The oblique forms occur when a noun or noun phrase is followed by a

postposition. Nouns are inflected according to their gender and the word final sound. Dogri is a modern Indo-Aryan language spoken primarily in the Jammu and Kashmir state and the adjoining areas of Himachal Pradesh, Punjab and across the border in Sialkot and Shakargarh tehsils presently in Pakistan. As language part of the Census of India 2011 is not available so according to the Census of India 2001 the number of Dogri speakers is 22,82,589.

The History of modern Indo-Aryan languages such as Hindi, Marathi, Gujarati, Assamese, Bengali, Odia, Punjabi and Dogri can be traced to its earlier stages-Old Indo-Aryan language (1500 BC to 600 BC) and Middle Indo-Aryan language (600BC to 1000 AD). The development of Dogri as a language can be divided into the following three stages: Old Dogri (10th to 16th century), Medieval Dogri (16th to 19th Century) and Modern Dogri (19th century to the present). Dogri has its own script namely "dogre akkhar" or "dogre" based on Takri script which is closely related to the Sharada script employed by Kashmiri language (Veena Gupta). This script was the official language of Jammu & Kashmir state during the regime of Maharaja Ranbir Singh (1857-1885 AD) After the independence the state government constituted a committee on 29th October, 1953 headed by Sh. Girdhari Lal Dogra presented a report and accordingly the state government decided to adopt Devnagri as well as Persian script for Dogri and it was incorporated in the State Constitution in 1957. So at present the Devnagri script is mainly used in India and the Nasta'liq form of Perso-Arabic in Pakistan.

Now as one of the recognized languages in the 8th schedule of Indian Constitution, Dogri is trying to compete with any major language of Indian Constitution and to enable them to cope up with the requirements of the future the Government of India, realizing the importance of developing Information Technology tools in regional languages, has involved different universities and IIT's in this field. Department of Information Technology has already taken initiative to provide software tools in Dogri Language

POS Tagging and Indian Languages

Not much work has been carried out in different languages due to unavailability of large annotated corpus. The Indian languages are morphologically rich languages and generating a

standard tagset framework for POS tagging is a challenge. IIT Hyderabad and Baskaran et. al (2008:89) tried to design a common POS tag set framework for Indian languages. Due to varied structure of sentences and grammatical rules, a common tag set for Indian languages is not possible.

What is Parts-of-Speech Tagging (POS Tagging)?

The process of assigning the Part-of-speech label to words in a given context is said to be Part-of-speech (POS) tagging, which is an important aspect of Natural Language Processing. In a sentence it inevitably involves the task of marking each word with its appropriate part of speech. For any Part-of-speech (POS) work, the tag set of the language has to be developed. It should contain the morpho-syntactic features of the language called the sub tags. Different Parts of speech include nouns, pronoun, adjectives, adverbs, verbs, postposition, conjunction and their sub-categories should be covered.

Tag Set

The first step towards developing the computational grammar for any language and basic building block of any NLP works is Part-of-speech (POS) tagging. Hence, POS tagging is not about just providing a tag to token but it encompasses a whole range of grammatical information for that token in the sentence from a particular language. Different languages may have its own Part-of-speech (POS) classification schemes in terms of nouns, pronouns, adjectives, adverbs, verbs, postposition, conjunction etc. So, tag set is a set of defined tags, for example, a set of word categories to be applied to the word tokens of a text.

Types of Tag Set

There are three types of tag sets, namely:

- Flat tag set
- Hierarchical tag set
- Fine grained tag set

Flat tag sets just list down the categories applicable for a particular language without any provision for modularity or feature reusability.

Hierarchical tag sets, on the other hand, are structured relative to one another and offer a well-defined mechanism for creating a common tag set framework for multiple languages

while providing flexibility for customization according to the language or application. Decomposability in a tag set allows different features to be encoded in a tag by separate sub-strings.

Fine grained tag set is the tag set where the minute things are considered and is accurate in syntactic analysis (Vijayalaxmi .F. Patil).

Focus of This Paper

The present paper is based on a hierarchical tag set. The term "hierarchical" means that the categories in that tag sets are structured relative to one another rather than a large number of independent categories. A hierarchical tag set will contain a small number of categories, each category contains a number of types and each type contains attributes, and so on, in a tree-like structure**.**

Note to the Annotators

The annotation will be carried on the sentence level provided by Annotation Tool in a sentence window for tagging. While tagging, the category is assigned on the basis of the grammatical class that a token assumes in a sentence. On the other hand, the type (of a category) is based on function. The attributes (of a type of a category) are based on the form – morphologically visible realization of the morphosyntactic features.

The ILPOSTS-Dogri provides two additional attributes to facilitate annotators with respect to the attribute in case of the following two situations:

1. Not-applicable (0)

In those cases where the given morphosyntactic feature is not applicable or not available.

2. Undecided/doubtful (x)

In case of ambiguity, the annotators can resolve the ambiguity within the given sentential context, and assign the appropriate value. However, if it persists as a case of doubt or in lack of clear, confident judgment, the annotators can mark the value as (x).

Remember

Token: A printed item separated by white space.

(POS) Tag: A POS label given to a token along with its morphosyntactic attributes.

Tag set: A set of defined tags.

Tagging: The process of assigning a tag to a token. Also known as annotation.

Annotation Tool: A tool is used for tagging.

Dogri Tagset

Category

1. **Noun (N)**
2. **Pronoun (P)**
3. **Demonstrative (D)**
4. **Nominal Modifier (J)**
5. **Verb (V)**
6. **Adverb (A)**
7. **Postposition (PP)**
8. **Particle (C)**
9. **Numeral (NUM)**
10. **Reduplication (RDP)**
11. **Residual (RD**)

1. **Noun:** The word that refers to people, animal, object, idea, concept, feeling etc. is a Noun. In Dogri a noun is hosts the attributes like gender, number and case. The types and attributes of a Noun are –

Category	Type	Attributes	Examples
Noun(N)	Common (NC)	Gender, Number, Case, Case marker, Distributive.	जागतें/NC.mas.pl.obl.0.0
	Proper(NP)	Gender, Number, Case, Case marker.	राम/NP.mas.sg.0.0
	Verbal (NV)	Case, Case marker.	पीने/NV.obl.0
	Spatio-temporal (NST)	Case, Case marker, Distributive, Dimension.	बाह्रा/NST.obl.abl.0.prx

2. **Pronoun:** The words which function like a Noun and substitute a noun or Noun phrase are called pronoun. The types and attributes of Dogri pronouns are:

Category	Type	Attributes	Examples
	Pronominal (PPR)	Gender, Number, Person, Case, Case marker, Emphatic, Distributive, Dimension, Honorificity.	असें/PPR.0.pl.1.obl.erg.0.n.n.n
	Reflexive	Gender, Number, Case, Case	आपूं/PRF.0.0.dir.0.0

Pronoun (P)	(PRF)	marker Distributive	
	Reciprocal (PRC)	Case.	इक-दुए/PRC.obl
	Relative (PRL)	Gender, Number, Case, Case marker, Emphatic, Distributive, Honorificity.	जेह्ड़ा/PRL.mas.sg.obl. gen.n.n
	Wh-pronoun (PWH)	Gender, Number, Case, Case marker, Emphatic, Distributive, Honorificity.	कोह्दा/PWH.mas.sg.obl. gen.n.n.n

3. Demonstrative: A demonstrative is form or class of words that is used deictically to indicate a referent's spatial, temporal or discourse location. A demonstrative functions as a modifier of a noun, or a pronoun. In Dogri, the forms are same in demonstratives and demonstrative pronouns, but the only difference is that the demonstrative always followed by a noun or the pronoun. Types and attributes of demonstrative are the following. All types of demonstrative contain the same attributes.

Category	**Type**	**Attributes**	**Examples**
Demonstrative (D)	Absolutive (DAB)	Gender, Number, Dimension, Distributive, Emphatic.	एह्/DAB.0.0.prx.0.n
	Relative Demonstrative (DRL)	Gender, Number, Distributive	जेह्ड़ा /DRL.mas.sg.0
	Wh-demostrative (DWH)	Gender, Number, Distributive	केह्ड़ा /DWH.mas.sg.0

4. Nominal Modifier: Nominal modifier is the category which usually modifies noun and pronoun in the sentence. The types and attributes of nominal modifiers are listed below. An adjective in Dogri inflected for number, gender and case.

Category	**Type**	**Attributes**	**Examples**
Nominal Modifier (J)	Adjective (JJ)	Gender, Number, Case, Distributive	शैल/JJ.0.0.dir.n
	Quantifier (JQ)	Gender, Number Case, Emphatic, Distributive Numeral.	मता/JQ.mas.sg.dir. n.n.0
	Intensifier (JINT)	Gender, Number, Case.	बड़ी/ JINT.fem.sg.dir

5. Verb: A verb usually denotes action ("go", "eat"), occurrence ("to modify" (itself), "to glitter"), or a state of being (survive "live", "stand"). A verb may vary in form according to its tense, aspect, mood and voice. It may also agree with the person, gender, and/or number of some of its arguments (what we usually call subject, object, etc.).v The types and attributes of Dogri verbs are:

Category	Type	Attributes	Examples
Verb (V)	Main Verb (VM)	Gender, Number, Person, Tense, Aspect, Mood, Negation, Finiteness, Honorificity.	जंदा/VM.mas.sg.0.0.pft.dcl.n.fnt.n
	Auxiliary Verb (VA)	Gender, Number, Person, Tense, Aspect, Mood, Negation, Finiteness, Honorificity.	ऐ/VA.0.sg.3.prs.prg.dcl.n.fnt.n

6. Adverb: An adverb is a part of speech that belongs to a group of words that modifies verbs, adjectives, other adverbs, clauses, and sentences. Types and attributes of Dogri adverbs are:

Category	Type	Attributes	Examples
Adverb(A)	Manner (AMN)	Gender, Number, Case.	बल्लें/AMN.0.0.0

7. Postposition: A postposition is a functional word which occurs after the word to point towards that word to show the relationship of that word with the other entity. It occurs after Noun or Pronoun. Gender and Number attributes are present in postposition only for the genitive markers, e.g.dA, de, dI, diyAN and not usually for others. However these attributes will be tagged according to their value, if they are physically marked in the postposition itself. Since case markers are mostly marked by the postposition in Dogri, it is mandatory to tag the case markers in the postpositions. The postpositions are written separate to nouns but it is attached with pronouns.

Category	Type	Attributes	Examples
Post-position(PP)	Case(PP)	Gender, Number, Case marker, Honorificity.	दियां/PP.fem.pl. gen.n

8. Particle: A Particle is a word which doesn't belong to any of the main Part of Speeches. It is indeclinable or uninflected and has important function. The types and attributes are as given below.

Category	Type	Attributes	Examples
Particle (C)	Co-ordinating (CCD)		ते/CCD
	Subordinating (CSB)		जेकर/CSB
	Interjection (CIN)	Gender, Number, Case Marker.	अड़िये/CIN.fem.sg.0
	(Dis)Agreement (CAGR)		नेईं/CAGR
	Emphatic (CEMP)		गै/CEMP
	Topic (CTOP)		ते/CTOP
	Delimitive (CDLIM)		मात्र/CDLIM
	Honorific (CHON)		होरCHON/
	Dedative (CDED)		बारै/CDED
	Exclusive (CEXCL)		बगैर/CEXCL
	Interrogative (CINT)		कीह्/CINT
	Dubitative (CDUB)		खबरै/CDUB
	Similative (CSIM)	Gender, Number, Case.	आंह्गर /CSIM
	Others (CX)	Gender, Number, Case.	आह्ला/CX.mas.sg.dir

9. Numeral: a word referring to a cardinal number (one, two, three, etc) or an ordinal number (first, second, third, etc.).

Category	Type	Attributes	Examples
Numeral (NUM)	Real (NUMR)		1,2,3/NUMR
	Serial (NUMS)		i.ii,iii,iv/NUMS
	Calendric (NUMC)		02-04-2011/NUMC
	Ordinal (NUMO)		1st 2nd /NUMO

10. Reduplication: is a morphological process in which the root or stem of a word or even the whole word is repeated exactly or with a slight change with a meaning change:

Category	Type	Attributes	Examples

Reduplication (RDP)			घर-घर/RDP

11. Residual: Foreign words are those words which are written in any script other than Dogri e.g. 16, building, news etc. But the borrowed words which are written in the script of Dogri don't come under Residuals. E.g. 16, building, news.

Category	Type	Examples
Residual(RD)	Foreign Word (RDF)	प्रसीडैंट/RDF
	Symbol (RDS)	@,$/RDS
	Punctuation (PU)	। , -/PU
	Unknown (UNK)	Officers/UNK
	Not-applicable (0)	(0)
	Undecided/doubtful (x)	(X)

Table1.2 Attributes and their Value

ATTRIBUTE\ SYMBOL	Value\symbol				
GENDER\GEN	Masculine\mas	Feminine\fem			
NUMBER\NUM	Singular\sg	Plural\pl			
PERSON\PER	First\1	Second\2	Third\3		
TENSE\TNS	Present\prs	Past\pst	Future\fut		
CASE\CS	Direct\dir	Oblique\obl			
CASE MARKER\CSM	Ergative\erg	Accusative\acc	Instrumental\ins	Dative\dat	Genitive\gen
	Locative\loc	Ablative\abl	Vocative\voc		
ASPECT\ASP	Simple\smp	Progressive\prg	Perfect\pft		
MOOD\MOOD	Imperative\imp	Optative\opt	Conditional/con	Declarative/dec	
	Obligative/obl	Promisive/pro			
FINITENESS\FIN	Finite\fin	Non-finite\nfn	Infinite\ifn		
DISTRIBUTIVE\DSTB	Yes\y	No\n			
EMPHATIC\EMPH	Yes\y	No\n			
NEGATIVE\NEG	Yes\y	No\n			
HONORIFICITY\HON	Yes\y	No\n			
NUMERAL\NML	Ordinal\ord	Cardinal\crd	Non numeral\nnm		
DIMENSION\DIM	Proximal\prx	Distal\dst			
NEGATIVE\NEG	Yes\y	No\n			

Conclusion

The aim of this tag set is to provide clear instructions for annotating the Dogri corpus. The tag set developed so far is hierarchical in nature as it is divided into main word categories, types of the categories and their sub features or attributes as discussed above. Since hierarchical tag set are more elaborative and comprehensive in nature, consequently

the same tag set can be used at all levels-Pos tagging, chunking, dictionary and morphological analysis.

==========

I thank Dr. B. Mallikarjun, Former Head, Linguistic Data Consortium for Indian Languages (LDC-IL) and Dr. L. Ramamoorthy, Head, Linguistic Data Consortium for Indian Languages (LDC-IL) for providing academic guidance and constant moral support in designing this tag set.

==========

Works Cited

Agnihotri, 2006, *Hindi: An Essential Grammer,* Routledge London and New York

F. Patil, Vijayalaxmi, 2010, *DESIGNING POS TAG SET FOR KANNADA,* Editor, Sharma, Atreyee, *Knowledge Sharing Events,* Linguistic Data Consortium for Indian Languages (LDC-IL), CIIL, Mysore

Gupta, Veena.1995, *Dogri Vyakaran*. Academy of Art, Culture and Language, Jammu

Kachru, Yamuna, 2006. *Hindi.* John Benjamins: Amsterdam/Philadelphia.

Koul, Omkar N, 2008 *Modern Hindi Grammar,* Dunwoody Press

Masica, Colin, 1993. *The Indo-Aryan Languages*. CUP: Cambridge.

Schmidt, Ruth Laila, 1999. *Urdu: An Essential Grammar*. Routledge: London

Sharma, Aryendra, 1994. *A Basic Grammar for Modern Hindi*. Central Hindi Directorate.

==========

Sunil Kumar, M.A., M.Phil., B.Ed.
Senior Resource Person (Academic)
National Translation Mission
Central Institute of Indian Languages

==========

Mysore 570006
Karnataka
India
sk07choudhary@gmail.com