BIS Annotation Standards With Reference to Konkani Language

Edna Vaz [1], Shantaram V. Walawalikar[2], Dr.Jyoti Pawar[3], Dr.Madhavi Sardesai[4]
(1) Goa University, Taleigao-Goa.
edna.vaz22@gmail.com,goembab@hotmail.com,jyotidpawar@gmail.com,
madhavikonkani@gmail.com

ABSTRACT

The Bureau of Indian Standards (BIS) Part Of Speech (POS) tagset has been prepared for the Indian Languages by the POS Tag Standardization Committee of Department of Information Technology (DIT), New Delhi, India. The BIS POS tagset aims to ensure standardization in the POS tagging of all the Indian Languages. It has been used for POS tagging in the Indian Languages Corpora Initiative (ILCI) project which has developed parallel annotated corpora consisting of 25000 sentences each from the tourism and the health domain for 11 Indian Languages.

In this paper we present some challenges encountered while using the BIS POS tagset for Konkani, a morphologically rich Indian Language, along with the possible solutions to overcome these challenges.

KEYWORDS: Part Of Speech Tagging, POS, Department of Information Technology, DIT, Bureau of Indian Standards tagset, BIS tagset, Indian Languages Corpora Initiative, ILCI, Konkani, Natural Language Processing, NLP.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 145–152,
COLING 2012, Mumbai, December 2012.

1. Introduction

Annotated text corpora are a basic and a very useful resource for researchers in Natural Language Processing (NLP) for developing various language technologies. The annotation of corpora is done using a set of tags defined for this purpose. The BIS tagset is a set of tags evolved by the POS Tag Standardization committee appointed by the DIT to standardize and streamline the process of POS tagging in Indian languages.

A need for standard tagset along with guidelines for using it arose because a lot of researchers were working independently following the tags of their own choice to mark the POS within and across languages. This affected the reusability of the tagged data amongst researchers. Hence, in order to facilitate interoperability an exercise was made to have a consensus on the style and characteristics of POS tagging, so as to arrive at a common standard for tagging in Indian Languages. This led to the standardized BIS tagset for POS tagging for Indian Languages. Several meetings of experts in the field were held to decide on the tagset and all groups engaged in research in NLP were given standard guidelines for annotation. This paper aims at shedding light on the peculiarities of the Konkani language that have posed challenges in tagging corpora using the standard BIS tagset. The paper is organized as follows – section 2 briefly introduces the BIS tagset and the Konkani Language and the challenges encountered while tagging using the BIS POS tagset for Konkani are presented in section 3 which is followed by section on conclusion and future work.

2. BIS tagset and Konkani Language

The BIS POS tagset is prepared keeping in view the comments of experts in the area of NLP and Language Technology (LT) for Indian languages. This tagset is an important step taken by DIT to ensure that NLP practitioners involved in tagging follow a common tagset while tagging various corpora. The tagset initially consisted of 38 tags. These tags were then modified after taking inputs from linguists, computer scientists and language experts. More details of the BIS tagset can be found in Chaudhary, 2010. The BIS tagset is a commendable effort. But the process of tagging was, at times, quite challenging as it was different from the conventional style of tagging (for example, the case of marking adverbs of place and location as NSTs (locative nouns) .

Konkani is an Indo-European (Indo-Aryan) language evolved from Sanskrit. It is a morphologically rich Indian language (Almeida, 1989). It is one of the twenty two languages included in the Eighth Schedule of the Indian Constitution. It is spoken by the people of the state of Goa, in some parts of Maharashtra, Karnataka and in some pockets of Kerala. It is influenced and enriched by various other languages like Marathi, Kannada, Malayalam, Hindi, Arabic, Persian, Portuguese and English. It is the official language of Goa with Devanagari as the officially recognized script. It is also written in Roman, Kannada, and Malayalam scripts (Walawalikar, et.al. 2010).

2.1 BIS POS tagset for Konkani

The BIS tagset was used for tagging the Konkani ILCI corpus consisting of a total of 50000 sentences (730330 tokens). The main objective of the ILCI project was to develop standard quality parallel annotated corpora for 11 Indian languages including English language to promote NLP research for Indian Languages (Chaudhary, 2010).

The following is the tagset for Konkani prepared in line with the BIS tagset.

SI	**Category**			**Label**	**Annotation Convention**		
	Top Level	**Subtype (level1)**	**Subtype (Level2)**			**Hindi**	**Konkani**
1	**Noun**			N	N	संज्ञा	नाम
1.1		Common		NN	N__NN	जातिवाचक संज्ञा	जातीवाचक नाम
1.2		Proper		NNP	N__NNP	व्यक्तिवाचक संज्ञा	व्यक्तीवाचक नाम
1.3		Nloc		NST	N__NST	देश-काल-सापेक्ष संज्ञा	थळ-काळ-सापेक्ष नाम
2	**Pronoun**			PR	PR	सर्वनाम	सर्वनाम
2.1		Personal		PRP	PR__PRP	पुरूषवाचक सर्वनाम	पुरूश सर्वनाम
2.2		Reflexive		PRF	PR__PRF	निजवाचक सर्वनाम	आत्मवाचक सर्वनाम
2.3		Relative		PRL	PR__PRL	संबंधवाचक सर्वनाम	संबंदी सर्वनाम
2.4		Reciprocal		PRC	PR__PRC	पारस्परिक सर्वनाम	एकमेकी सर्वनाम
2.5		Wh-word		PRQ	PR__PRQ	प्रश्नवाचक सर्वनाम	प्रस्नार्थी सर्वनाम
2.6		Indefinite		PRI	PR_PRI		अनिश्चीत सर्वनाम
3	**Demonstrative**			DM	DM	संकेतवाची	दर्शक
3.1		Deictic		DMD	DM__DMD		
3.2		Relative		DMR	DM__DMR	संबंधवाचक संकेतवाची	संबंदी दर्शक
3.3		Wh-word		DMQ	DM__DMQ	प्रश्नसूचक संकेतवाची	प्रस्नार्थी दर्शक
		Indefinite		DMI	DM_DMI		अनिश्चीत दर्शक

4	**Verb**			V	V	क्रिया	क्रियापद
4.1		Main		VM	V__VM	मुख्य क्रिया	मुखेल क्रियापद
4.1.1			Finite	VF	V_VM_VF	परिमित क्रिया	पूर्ण क्रियापद
4.1.2			Non-finite	VNF	V_VM_VNF		अपूर्ण क्रियापद
4.1.3			Infinitive	VINF	V_VM_VINF	अपरिमित क्रिया	सादारण रूप
4.2		Gerund		VNG	V_VM_VNG	क्रियावाचक संज्ञा	क्रियावाचक नाम
4.4		Auxiliary		VAUX			
4.4.1			Finite		V_VAUX_VF	पालवी क्रिया	पालवी पूर्ण क्रियापद
4.4.2			Non-Finite		V_VAUX_VNF		पालवी अपूर्ण क्रियापद
5	**Adjective**			JJ		विशेषण	विशेशण
6	**Adverb**			RB		क्रियाविशेषण	क्रियाविशेशण
7	**Postposition**			PSP		परसर्ग	संबंदी अव्यय
8	**Conjunction**			CC	CC	योजक शब्द	जोड अव्यय
8.1		Co-ordinator		CCD	CC_CCD	समानाधिकरण	समानाधीकरण जोड अव्यय
8.2		Subordinator		CCS	CC_CCS	आश्रित	आश्रीत जोड अव्यय
9	**Particles**			RP	RP	अव्यय	अव्यय
9.1		Default		RPD	RP_RPD	सामान्य	सरभरस अव्यय
9.2		Classifier		CL	RP_CL	वर्गक	वर्गक
9.3		Interjection		INJ	RP_INJ	विस्मयादि बोधक	उमाळी अव्यय

9.4		Intensifier		INTF	RP__INTF	तीव्रता-बोधक	तिव्रकारी अव्यय
9.5		Negation		NEG	RP__NEG	नकारात्मक	न्हयकारी अव्यय
10	**Quantifiers**			QT	QT	संख्यावाची	संख्यादर्शक
10.1		General		QTF	QT__QTF	सामान्य	सामान्य
10.2		Cardinals		QTC	QT__QTC	गणनावाची	संख्यावाचक
10.3		Ordinals		QTO	QT__QTO	क्रमवाची	क्रमवाचक
11	**Residuals**			RD	RD	अवशिष्ट	हेर
11.1		Foreign word		RDF	RD_RDF	विदेशज	विदेशी
11.2		Symbol		SYM	RD_SYM	चिह्न	कूरू
11.3		Punctuation		PUNC	RD_PUNC		विरामकूरू
11.4		Unknown		UNK	RD__UNK	अज्ञात	अनवळखी
11.5		Echowords		ECH	RD__ECH	प्रतिध्वनि-शब्द	पडसादी उतरां

TABLE 1 – The BIS tagset for Konkani

3. Issues in tagging using the standardized BIS tagset

POS tagging is an ongoing process and we may need to modify the tagset to accommodate newer findings, etc. Some tags have been modified whereas some are still being discussed. For example, in the initial stages of BIS POS discussions, it was decided that names of cities, institutions, organizations, people and months were to be tagged as Proper Nouns (NNP) whereas names of medicines, diseases, flowers, animals and seasons were to be taken as Common Nouns (NN). This decision was later refined. Blood Cancer (a specific type of a Cancer) was tagged as NNP and Cancer (a general category of diseases) was tagged as NN. Decisions taken may have to be modified and documenting these will help us trace the line of our path. So also this documentation will help newcomers in this area to foresee problems which may arise in tagging corpora using this tagset. It is important to remember that flexibility in the tagset is for the purpose of refinement and accuracy in the tagging process and not for disclosing our earlier 'inappropriate' decisions. We know that words in a language have the capacity to function differently when they appear in a sentence. For example, in रांदिल्लें जेवण (rAMdilleM jevaNa) '**cooked food**', रांदिल्लें functions as a modifier whereas रांदिल्लें in हांवें दनपारां जेवण रांदिल्लें (hAveM danapArAM jevaNa rAMdilleM) ('I had **cooked** food in the afternoon') functions as a verb. The present tagset is used to tag words according to their lexical rather than syntactic function. The lacunae if any in this practice may come to light only when we move to higher levels of NLP.

Some of the minor challenges arising from the usage of the proposed BIS tagset while POS tagging the ILCI Corpora for Konkani language could be placed under the following main heads:

3.1 **Word Sense Disambiguation Challenges**

The hyphen is marked as punctuation in the BIS tagset, but in Konkani it conveys different information on different occasions. In some cases, a pair of two words joined by a hyphen is a compound word whereas in other cases it may be just a noun phrase. Various senses conveyed by the hyphen need to be separated and structurally identified as specific compounds. This would facilitate word sense disambiguation.

Examples:

- बरें-वायट bareM-vAyaT ('good-bad'). In this example, the hyphen marks a pair of antonyms. In isolation, the words in this pair function as adjectives whereas the pair as a whole functions as a noun.
  For example, in तें एक बरें चली teM eka bareM chalI ('she is a good girl'), and तें एक वायट चली teM eka vAyaT chalI ('she is a bad girl', बरें bareM 'good' and वायट vAyaT 'bad' both function as adjectives. However, in ताका बरें-वायट समजना tAkA bareM-vAyaT samajanA ('he does not know what is good and what is bad' i.e. 'he does not know his good'), the pair बरें-वायट bareM-vAyaT functions as noun.
- The hyphen in the pair बायल-मनीस bAyal manIsa ('woman-human') marks a specifier-specified relationship. In Konkani, मनीस manIsa 'human being' is a gender neutral term. The following sentences illustrate this point:
  1. तो एक बरो मनीस to ek baro manIs ('he is a good human being')
  2. ती एक बरी मनीस tI ek barI manIs ('she is a good human being')

We see that in both the above sentences मनीस 'human being' remains the same. However, the term can be specified by other gender specific nouns such as दादलो dAdlo (masculine) 'man', चली chalI (feminine/neuter) 'girl/daughter', बायल bAyl (feminine) 'woman' when it occurs in compounds such as दादलो मनीस ('man-human'), बायल मनीस ('woman-human'), चली मनीस ('girl-human'). For example, in दादल्या मनशाक तें कळचेंना dAdlya manashAka teM kaLacheMnA ('a male human (i.e. a man) will not understand this'), dAdlo 'man' specifies the masculine gender of मनीस manIsa 'human being'. Thus the role of hyphen cannot be ignored in the compounds of the above type.

- In भाट-बेंस bhAT- beMsa ('property-assets'), the hyphen is used to mark a 'synonymic compound'. Such an occurrence of a hyphen is very common in Konkani. For example, भांगर-शिंगर, धन-दौलत etc. These synonymous pairs are mostly a combination of a native and a foreign language word. For example, in भाट-बेंस, the first word is a native word whereas the second one is a word from Portuguese.
- The hyphen in किताब-ए-हिंद kitAba-e-hiMda ('kitab-e-hind') is used to highlight a foreign title.
- The hyphen in 10-20 means 'to' (i.e. it conveys the inclusion of numerals from 10 to 20). The hyphen does occur in other expressions containing numerals. However, it conveys a different sense. For example in the Konkani sentence ताका मारपाक सात-आठ मनीस आशिल्ले (tAkA mArapAka sAta-ATh manIs Ashille) 'to hit him seven-eight people were there' (i.e. 'around seven to eight people were present to hit him'), the phrase सात-आठ is joined by the hyphen. Here, the hyphen expresses a sense of 'unsurety' within the mind of the speaker about number of people that were present before him.

- The phrase साडे-सात sADe-sAta means 'half past seven'. Such an usage of the hyphen can be found in other **temporal expressions** such as in सवाय-आठ savAya-ATh (quarter past eight), etc. It is the hyphen that helps to mark this phrase.

3.2 **Punctuation**

Marking an inverted comma as punctuation is misleading in some cases. For example, in 'कॉंग्रेस पार्टी' चो kA.Ngresa pArTI cha.o ('of Congress Party'), चो is a suffix of the noun phrase 'Congress Party'. This phrase is put in inverted commas as it is a foreign language phrase. Thus marking inverted commas in such cases as punctuation is not proper.

3.3 **Tagging of spatio-temporal adverbs as NST**

All nouns in Konkani undergo oblique formation before they take a case suffix.
राम rAm becomes रामा rAmA before it takes any suffix like -क ka, -न na, -चो cho, etc. However, the spatio-temporal adverbs take suffixes before undergoing oblique formation. E.g. भायर+लो bhAyara + lo> भायलो bhAyalo 'of outside', सकयल +च्यान sakayala +chyAna > सकयल्यान sakayalyAna 'from below'.
Moreover, these adverbs take only the above mentioned suffixes. Combinations like भायर +क bhAyara +ka 'to the one outside' भायर +आंत bhAyara+AMta 'in the one outside' are not permitted in the language. If these adverbs are tagged as nouns (NSTs) then they must be treated as a special category of nouns.

3.4 **Recognition of frozen expressions**

Some expressions in language function in a certain way. खरें म्हणल्यार khareM mhaNalyAra (Truly/Really speaking) खरें KhareM in Konkani means 'truth' whereas म्हणल्यार mhaNalyAra means 'having told' which is a quotative.
However, when these two words come together, they always function like an adverb. Tagging these words individually would serve no purpose. Some more examples in the same category are listed below -
सांगपाचें म्हणल्यार sAMgapAcheM mhaNalyAra "Of-telling if told" means 'actually'.
तशें पळेल्यार tasheM paLayalyAra "that way if seen" also means 'actually'.

3.5 **Negation**

One has to be careful in the treatment of negation in Konkani. Negation found in Hindi is not all the same in Konkani. In Hindi, negation is mainly a syntactic process whereas in Konkani, it is a morphological one. Examples illustrating this point are given below:

Hindi:

3.5.1 a नहीं. मैं नहीं आउंगा nahIM. maiM nahI AuMgA ('**No**. I will not come')

3.5.1 b नहीं. हम नहीं आयेंगे nahIM. hama nahIM AyeMge ('**No**. We will not come')

Konkani:

3.5.2 a ना. हांव येवंचो ना nA. hAMva yevaMcho nA ('**No**. I will **not** come')

3.5.2 b ना. आमी येवंचे नात nA. AmI yevaMche nAt ('**No**. We will **not** come')

So also in Hindi, negation occurs at all places as a particle (in the sentences 3.5.1 a and 3.5.1 b), whereas in Konkani the first word in 3.5.2 a and 3.5.2 b is a particle and the second one in 3.5.2b is a finite verb.

Conclusion and Future Work

A fruitful comparison between languages (for example between different grammatical categories) has been possible because of the BIS tagset. The tagset has helped in bringing to our minds the differences existing between Indian Languages. For example, Konkani words marked as NSTs (a subcategory of nouns) under the BIS tagset do not have an oblique form whereas other subcategories of nouns in Konkani always have an oblique form. However, the tag NST seems to work fairly well in other Indian languages.

Differences between languages which may pose **challenges in Indian Language–Indian Language translation can be predicted** to a fair extent with this tagset. For example, postpositions like Hindi का never occur separately in Konkani. One may be tempted to translate राम का rAm kA (of Ram) as राम चो rAm -cho which actually should be रामाचो rAmAcho.

- **solutions proposed for the issues faced:**

The BIS tagset has no doubt prepared the ground for fruitful annotations. It is up to the experts of each language to examine this tagset closely and suggest necessary refinements pertaining to their language. It may not be possible to include new tags to handle current issues but we could deal with them keeping the standardized tagset intact. We propose following suggestions to deal with some of the above mentioned issues.

The hyphen sometimes functions as an integral part of compounds and phrases. Marking it as a punctuation mark will mean ignoring the subtle information that it conveys in these occurrences. We feel that the role of the hyphen has to be recognized as it would not only help in word sense disambiguation (as in the case of बरें-वायट bareM-vAyaT 'good-bad') but also bring to light certain peculiarities of Konkani (as in the case of भाट-बेंस bhAT- beMsa 'property-assets' where the synonymic compound is formed with a word from Portuguese).

Frozen expressions, for example can be marked with a special tag at POS level itself so that there is no wastage of time in unnecessary work, later. For example, in खरें म्हणल्यार KhareM mhaNalyAra 'really/truly speaking' tagging individual words would serve no purpose. Instead the token **'**खरें म्हणल्यार**'** could be marked as **FE** i.e. frozen expression. The appropriate category (i.e. adverb) could be marked at the next level of NLP.

While recognizing the importance of standardization in the POS tagging of all the Indian Languages, we also feel that we should be careful in not to lose on the grammatical peculiarities of individual languages as it may have an adverse effect on the later stages of NLP.

References

Almeida, Matthew. (1989). *A Description of Konkani*. Miramar, Goa: Thomas Stephens Konkani Center.

Chaudhary. Narayan et.al. (2010). *ILCI Parts of Speech guidelines document*.

Jawaharlal Nehru University.

Walawalikar et.al. (2010) *Experiences in Building the Konkani Wordnet using the Expansion Approach* in Proceedings of the 5th Global Wordnet Conference.