International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 18 (2018) pp. 13911-13918

Identifying Syntactic Transformations through POS Tag Distribution in Two Languages

Sonal Khosla[1], Dr. Haridasa Acharya[2]

[1]*Research Scholar, Symbiosis International (Deemed University), Pune, Maharashtra, India.*

[2]*Professor, Allana Institute of Management Sciences, Pune, Maharashtra, India.*

Abstract

Natural Language Processing is emerging fast over the past few years making it important to have quick solutions for building knowledge structures required for NLP applications which depends on the syntax of the languages. With languages having a common origin and common vocabulary a lot of efforts can be reduced if the differences and similarities in their syntactic structure are identified. Hence the focus of the current study is on comparison of grammatical features between Hindi and Marathi by studying the frequency distribution of the PoS tags in the two languages. We have also used Medical Articles and Reports parallel under Hindi and Marathi to study the effect of type of document along with the change in language on the grammatical features of the two languages.

Keywords: POS, Syntax, Syntactic, Hindi, Marathi

INTRODUCTION

In the field of text mining syntactic similarity and relations are important and are heavily exploited in developing NLP applications. The problem of measuring of similarity between two short segments has become increasingly important for many tasks(Kaur, 2015). Algorithms based on sentence similarity and syntactic structure were proposed by Li and Li (2015).

Traditionally, syntax is defined as a taxonomy of different types of syntactic structures assuming that the sentence is a series of syntactic units. On the basis of their similar syntactic behaviour, words can be grouped together into classes, known as syntactic or grammatical category (Manning &Schutze, 1999). A grammatical category is also known as "Part of Speech" (POS). A part of speech is assigned to a word on the basis of its appearance in a phrase (group of words). Grammatical tagging, commonly known as POS tagging is a function of relationship between adjacent words. Hence is a far stronger representative of syntactic transformations rather than simple syntax. If two languages have similar POS tagging systems then strength of inherent similarity in their grammatical structures can be assumed to be more.

We believe that knowledge of similarity and dissimilarity in the POS tag distribution of words in a parallel dataset is an indicator of similarity in language features. There can be innumerable ways of exploiting the similarity depending on the context and purpose. We consider parallel documents as translations of each other hence we presume that semantic similarity is in-built (Figure 1). Effect of this on how POS tags are distributed should be interesting, and hence a study of the distributions can be a good way to get an insight into syntactic characteristics**.** Such studies are not unknown, but evidence of such study in respect of Hindi and Marathi are rare.

RELATED WORK

Studies by Shih, Chiang & Tien (2000) has aimed to discover interlanguage features of Taiwanese learners of English using two Corpus (British National Corpus and the Taiwanese Learner Corpus of English by studying the frequency distribution of the POS tags.

Pak and Paroubek (2010) have used distribution of POS tags to analyse positive and negative sentiment in twitter messages. The study has also been extended on subjective and objective tags. An objective text tends to contain more common and proper nouns, while authors of subjective texts use more often personal pronouns. Spencer and Uchyigit (2012) is another study where distributions of POS tags have been put to effective use. While building a parallel corpus on Arabic-Spanish-English, Samy et. al. (2006) have put POS tag distribution to effective use. They have found that the percentage of nouns was highest in Arabic, while in the case of Spanish it is Prepositions, followed by Nouns and Articles. Campbell and Johnson (2001) have made syntactic comparisons within text from medical and non-medical corpora and found that medical corpora has less syntactic complexities as compared to non-medical corpora.

Ramanand et.al. (2007) is the only paper that was found where analysis of the distribution of POS categories in Indian languages were effectively used in building a Wordnet.

PROPOSED WORK

What is Syntax?

A **sentence** is ***a structured set*** of tokens (words/punctuation marks/numerals/dates etc.) having a structural arrangement providing meaning in itself. The structure is helpful in **encoding** the information, discovering relationships or transformations between words. A good account of supporting

theory can be found in Bender (2013) andManning and Schutze (1999).
The grammatical (**Part of Speech)** category is given to each word in a sentence, based on the role that is played by the word in relationship with adjacent words, phrase, sentence or paragraph (Rathod and Govilkar, 2015) as shown in Example 1. The basic structure of a sentence is defined in terms of Subject, Object and Verb. Hindi and Marathi follow the SOV (Subject-Object-Verb) Pattern. Each Subject-Object-Verb could be a word or a group of words. Hence a sentence structure could be defined as:

Sentence = Subject Phrase + Object Phrase + Verb Phrase . (1)

This structure can be further defined as a combination of a Verb Phrase and a Noun Phrase. A verb phrase is a predication of an action and consists of a verb. A Noun Phrase is a reference to an object and consists of a noun. Other phrasal categories may include Adjective Phrase, Adverb Phrase and so on. The POS categories are of two types: Open and Closed.

1. Open (Nouns, Adjectives, Adverbs and Verbs):

These classes can be compared cross linguistically and are generally common in all languages.

2. Closed (Pronouns, Determiners, Postpositions etc.):

These classes are more language specific.

Some categories which hold across languages help in designing common tools. Some do not hold across languages. Categories which are language specific are important in designing differentiators. Hence the first point of interest is to determine which are the ones which hold,and which are the ones which do not.

Hindi and Marathi share a common POS tag set as defined by Bureau of Indian Standards (BIS). The BIS POS tag is a superset containing tags for both languages. Set is defined at two levels. The tags are common across all Indian languages at Level 1, however there are language specific tags at Level 2. From the machine learning point of view when taggers are trained for specific languages the process would be identical and the language specific aspects are taken care of by the training sets. Since the training sets provided by us are 'parallel' (Figure 1), it is safe to assume that the resultant statistical distributions contain language specific characteristics built into them.

Here is a sample tagged output.

Example 1:

Hindi: दमे/NN के/PSP दौरे/NN से/PSP मरीज/NN की/PSP मौत/NN भी/RP हो/VM सकती/VAUXहै/VAUX

Marathi:तीव्र/JJ स्वरुपाच्या/NN दमा/NN अटॅकमधे/JJ लोकांचा/NN जीव/NN जाऊ/VM शकतो/VAUX

Example 2:

Hindi: पहले\QTO ओ.आर.एस.\NNP का\PSP घोल\NN पिलायें\VM,\PUNC इसके\RPP बाद\NST बच्चे\NN को\PSP किसी\DMI शिशु\NN रोग\NN विशेषज्ञ\NN से\PSP शीघ्र\RB दिखायें\VM ।\PUNC

Marathi: सर्वप्रथम\RB ओआरएसचे\NN मिश्रण\NN पाजावे\VM ,\PUNC त्यानंतर\PR शिशूला\NN कुठल्याही\DMQ बाल\NN रोगतज्ञाकडे\NN लवकर\RB दाखवावे\VM .\PUNC

It can be safely deducted from above discussion that statistical properties of POS tag distributions across parallel sources should be indicative of similarity in grammars of the two languages. However, the similarity holds better in Example 2.

Experimental Datasets and Methodology

The datasets used in the current study have been taken from two sources:

Dataset I (DS01)

The DS01 dataset has been manually collected from two web sources. First, **http://vikaspedia.in/InDG** is a multilingual website dedicated for providing knowledge and ICT (Information and Communication Technology) based knowledge products and services. It is developed as part of the national level initiative by the Ministry of Electronics and Information Technology (MeitY), Govt. of India executed by Centre for Development of Advanced Computing, Hyderabad. It has contents in multiple languages including Hindi (http://hi.vikaspedia.in/health) and Marathi (http://mr.vikaspedia.in/marathi). Articles related to different topics in the medical field are available mostly as parallel content in both the languages. The second, **http://Healthinfotranslations.org/** is a multilingual website that provides resources in health care. It has content in multiple languages including only Hindi, but not Marathi. So translations in Marathi with the help of human translators was obtained.

Dataset II (DS02)

The second dataset is taken from the ILCI Project of the Govt. of India. The Hindi-Marathi Health text corpus developed under the TDIL(Technology Development for Indian Languages) Mission of the Govt. of India [http://tdil-dc.in/index.php?option=com_vertical&parentid=58&lang=en] is used. It is a collection of 25000 POS tagged sentences stored in Excel. The sentences are POS tagged under the BIS (Bureau of Indian Standards) POS (Part of Speech) Tagset. DS02 was already tagged manually.

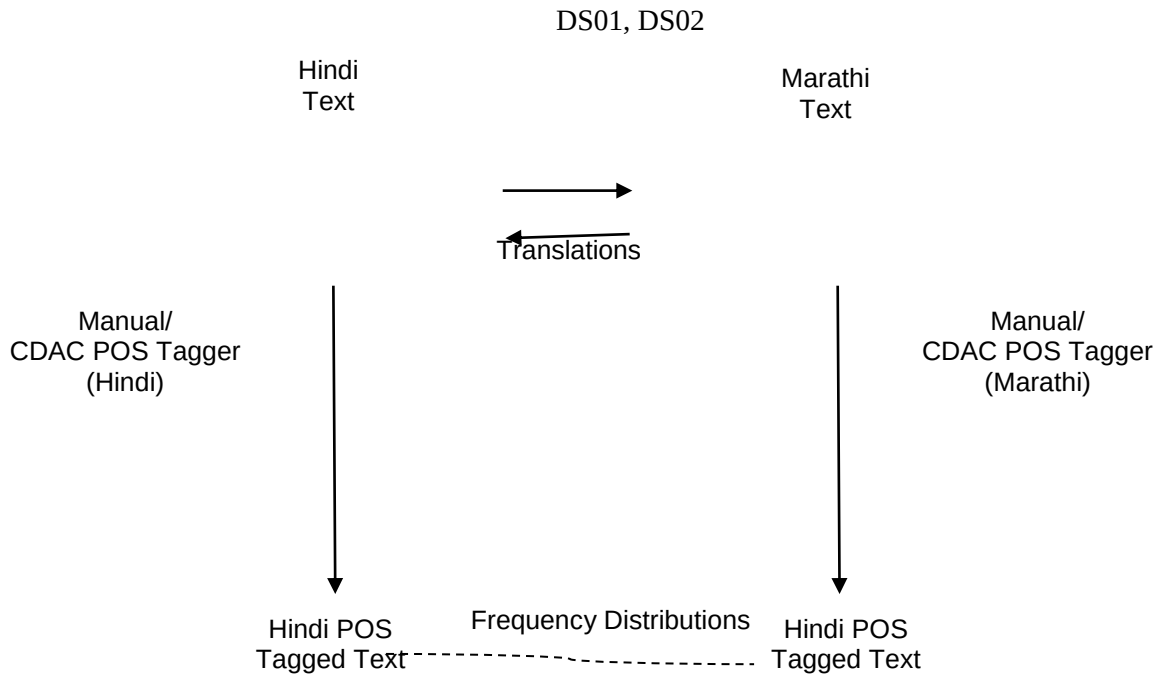

Figure 1: The parallel Tagging in similar Context

Experiments were conducted on the POS tagged sentences on the datasets to examine:

- Differences in typical distribution of POS Tags in documents of the two languages
- Differences in distribution of POS Tags which characterize the different types of documents. The frequency distributions of the POS tags in both the scenarios were obtained and Chi Squared tests (2x2 = Two Languages x Two document types) were used to compare the probability distributions. Correlations were also used to compare the relative frequencies (percentages) of the distributions.

A significant Chi-Squared indicate dependency of document types and languages with respect to the POS tag under consideration. The discussion would be relevant only when the 'tags hold' across the languages. There is an extended tag set which includes both language specific tag sets as subsets. The real positive fact is that the size of the intersection (ones that hold across) is almost equal to the whole set.

POS Tagging Methodology

Not many POS taggers available for Hindi and Marathi. There are taggers for Hindi but not for Marathi. The CDAC POS tagger is available for both Hindi and Marathi developed by CDAC, Mumbai (http://kbcs.in/tools.php) and was used for tagging DS01 to maintain uniformity in the tagging process. It is a statistical Tagger and has a trained file (with extension ". tagger") developed with the Stanford POS tagger based on Maximum Entropy Approach (Dwivedi &Malakar, 2015). While training the pos tagger, the tagger works on the hidden Markov model. Though the base-tool is same, two separate taggers become available to users once the training has been done on two independent data sets.

It follows the BIS POS tagset which is a universal set for many Indian languages, with some language specific distinctions (http://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf]. The BIS tagset is hierarchical and has 11 main categories at Level 1 (Dholakia &Yoonus, 2014). Language specific features are defined at level 2.

DS02 was already manually tagged and made available by TDIL. Upon looking at the tagged data in DS01(obtained through CDAC POS tagger) and DS01(manually tagged) it was concluded that the CDAC POS tagger can be safely used for tagging DS01. Even the results in the next section confirmed the same.

RESULTS AND DISCUSSIONS

Overall the distribution of tags is similar in DS01 and DS02, in spite of the fact that DS01 is trained on a corpus and tagged, while DS02 is manually tagged. It may be concluded that the choice of tagger either Manual or Digital (trained) has not made any significant difference in the tagged output. Though no literature is available to that discusses the efficiency of the tagger, however we can infer from the results.

Distribution of POS tags in Hindi and Marathi

Table 1 gives the distribution of the POS tags in both the Datasets DS01 and DS02. The distribution of POS tags provides ample evidences regarding the gross syntactic similarities. A closer look into changes in pmf (probability mass function) within the Level 1 shows that distributions are almost identical. However, change in language peculiar features (through the process of translation may lead to changes of frequencies within groups of tags, effectively implying that at Level 2 differences start appearing. We conclude

Overall the tag distribution is almost identical indicating strong similarity.

Correlations between POS tag distributions were evaluated taking percentages of tags as variates. Values were

DS01: Correlation (Hindi, Marathi) = 0.915

DS02: Correlation (Hindi, Marathi) = 0.870

Similarly, within languages the values were

Hindi: Correlation(DS01, DS02) = 0.952

Marathi: Correlation(DS01, DS02) = 0.984

All the values are highly significant.

Table 1: Distribution of POS tags in DS01 and DS02 summarized to Level 1

	Hindi (22771)	Marathi (17991)	Hindi (447189)	Marathi (332502)
Nouns	38.12	49.61	34.06	41.23
Pronouns	2.93	4.02	3.04	4.49
Demonstratives	1.69	2.48	1.64	1.57
Verbs	17.58	17.56	19.89	20.45
Adjective	4.63	6.26	4.69	7.23
Adverb	0.17	1.26	0.61	1.44
Conjunctions	4.73	6.06	4.04	5.91
PSP	14.92	0.00	16.07	0.02
Particles	1.58	0.42	2.94	0.74
Quantifiers	0.00	0.00	0.13	0.06
Residuals	0.00	0.00	0.13	0.06
Symbols	5.59	0.97	0.08	0.09
Punctuation	2.35	6.21	9.17	10.67
Unknown	0.00	0.00	0.00	0.00
Echo	0.00	0.00	0.05	0.01

Hindi uses postpositions (PSP) rather than prepositions for case marking and auxiliaries. In Marathi, postpositions are added to the word preceding it. It also adds suffixes to root form of words to build inflected form of words (Thompson, 2014, Dhore et.al, 2013). Our results confirm this observation. The PSP 's like के, भे or है are removed in Marathi and added as a morphological phenomenon. Hence ***PSP's in Hindi do not find any translation equivalents in Marathi.*** This is supported by statistical results as well as works of earlier researchers (Thompson, 2014, Dhore et.al, 2013).

It can be seen that 34% of the words in Hindi and 41% of the words in Marathi belong to the Noun Category in DS02. The nouns considered are of three types: Common Nouns (NN), Proper Nouns (NNP) and Noun Locations (NST). There is a relative increase in the percentage of nouns in Marathi since the number of tokens in Marathi is less due to formation of Compound words.

A noun in Marathi appears in different inflected forms as compared to Hindi. The inflections in Hindi nouns is only due to number and case (Singh &Sarma, 2010, Ramanathan and Rao, 2003) while in Marathi, inflections in nouns is due to Gender, Number and Case (Tidke, Binayaka, Patil and Sugandhi, 2013,Dolamic& Savoy, 2010). For single words in Hindi, multiple forms appear in Marathi. Like मधुमेह in Hindi appears in a single form, while in Marathi it appears as मधुमेच्या, मधुमेहमुळे etc. Similarly, the word आहार in Hindi has various forms in Marathi आहाराची, आहाराचे, आहारात, आहारातील, आहारातून. Hence due to these grammatical differences (Table 2), the differences in distribution is seen at Level 2

Table 2: Reasons for inflected forms in Hindi and Marathi

Difference in forms Gender, Number & Case in Hindi and Marathi

	Hindi	Marathi
Number	Singular, Plural	Singular, Plural
Case	Direct, Oblique	Nominative, Accusative, Instrumental, Dative, Ablative, Genitive, Locative, Vocative
Gender	Masculine, Feminine	Masculine, Feminine, Neuter

The count of nouns at Level 1 are almost similar in Hindi and Marathi, but the count of Proper nouns in Hindi is greater than Marathi as seen at Level 2. ***A Proper noun gets converted to Common noun while doing translation from Hindi to Marathi*** as seen in Example 2 and Example 3.

Example 3:

Hindi: मधुमेह|NNPके|PSPमरीजों|NNको|PSPतरह|NN -|SYM तरह|RDPके|PSPअनुभव|NNहोते|VMहैं|VAUX

Marathi:मधुमेह|NNअसलेल्या|VMलोकांना|NNविविध|JJप्रकारची| NNलक्षणे|NNदिसतात|VM.|PUNC

Example 4

Hindi:खाने\VM के\PSP बाद\NST मुँह\NN साफ\JJ करें\VM ।\PUNC

Marathi:जेवणानंतर\NN तोंड\NN स्वच्छ\JJ करावे\VM.\PUNC

It is worth noting that few of the NNs in Example 1 and Example 2 get converted into an Adjective (JJ) resulting in an increase in JJ in Marathi as compared to Hindi (Table 3).

Also, worth noting is that 59% and 69% of the POS tags in Hindi and Marathi respectively belong to the Nouns, Adjectives and Verb categories (Table 1). The Residuals consist of the tags like Unknown characters, punctuation characters which do not add any content to the data and are used mainly for grammatical correctness. However, these tags bring internal changes in Nouns, Adjectives and Verbs. A certain tag can appear as a noun in one instance while it may appear as a verb in the other.

The British National Corpus of English has the following percentage distribution of the major POS categories: Noun (24%), Verb (15%), Adverbs (5%) and Adjectives (8%). The difference in the percentage between Noun and Verb is 9% (Shih et. al., 2000). In comparison to the results obtained, the percentage of nouns is quite high in the case of DS01 and DS02 in both the languages. Also, the difference between the verbs and nouns is quite high as compared to English. In DS01, the difference in Nouns & Verbs in Hindi and Marathi are 20.54% and 32.05% respectively. In DS02, the difference in Nouns & Verbs in Hindi and Marathi are 14.17% and 20.78% respectively.

Table 3: Chi-Square Values for POS tag distribution in DS01 and DS02

POS Tag	Dataset	Hindi	Marathi	CHI-Sq value	Prob	POS Tag	Dataset	Hindi	Marathi	CHI-Sq value	Prob
Nouns						Pronouns					
NN	DS01	7889	8658	113.37	0.00	PRP	DS01	543	496	1.52	0.218
	DS02	141213	130682				DS02	5548	5490		
NNP	DS01	583	79	12.71	0.0003	PRF	DS01	82	16	1.80	0.180
	DS02	7302	1530				DS02	1729	489		
NST	DS01	209	188	12.445	0.0003	PRL	DS01	0	161	104.0	0.00
	DS02	3792	4894				DS02	5432	8284		
Adjectives						PRC	DS01	0	1	1.355	0.244
JJ	DS01	1054	1126	2.59	0.11		DS02	55	40		
	DS02	20975	24047			PRQ	DS01	42	49	0.006	0.937
Adverbs							DS02	545	625		
RB	DS01	39	226	51.76	0.00	PRI	DS01	0	0	NA	NA
	DS02	2728	4801				DS02	307	0		
Demonstratives						Verbs					
DMD	DS01	240	24	7.97	0.005	VM	DS01	2001	2548	15.98	0.000
	DS02	5364	289				DS02	47842	53926		
DMR	DS01	12	387	41.15	0.00	VINF	DS01	0	0	NA	NA
	DS02	703	4161				DS02	0	4		
DMQ	DS01	8	36	2.31	0.13	VING	DS01	0	0	NA	NA

POS Tag	Dataset	Hindi	Marathi	CHI-Sq value	Prob	POS Tag	Dataset	Hindi	Marathi	CHI-Sq value	Prob
	DS02	312	775				DS02	0	0		
DMI	DS01	124	0	NA	NA	VAUX	DS01	2003	612	5.95	0.014
	DS02	960	0				DS02	41088	14086		
Conjunctions						Postpositions					
CCD	DS01	821	881	7.36	0.006	PSP	DS01	3397	0	2.88	0.81
	DS02	11302	13893				DS02	71856	61		
CCS	DS01	257	209	0.25	0.620	Residuals					
	DS02	6764	5765			RDF	DS01	0	0	NA	NA
Particles							DS02	600	190		
RPD	DS01	173	5	9.94	0.0016	Symbols					
	DS02	7965	874			SYM	DS01	1273	175	272.87	0.000
INJ	DS01	0	0	NA	NA		DS02	374	296		
	DS02	7	13			Punctuations					
INTF	DS01	43	25	0.48	0.83	PUNC	DS01	536	1118	292.37	0.000
	DS02	1537	845				DS02	40991	35468		
NEG	DS01	144	46	0.799	0.005	Unknown					
	DS02	3652	716			UNK	DS01	0	0	NA	NA
QTF	DS01	332	431	0.109	0.74		DS02	5	3		
	DS02	7387	9830			Echo					
QTC	DS01	771	420	126.93	0.000	ECH	DS01	0	0	NA	NA
	DS02	7214	7883				DS02	209	42		
QTO	DS01	19	32	0.231	0.630						
	DS02	578	845								

Even though there are morphological differences in Hindi and Marathi grammatical rules governing construction of verbs (Deoskar, 2006), the count of Verbs is almost same at level 1. The Main Verb in Hindi is further subdivided into Non-finite, Infinitive and Gerund categories, but this is not done in Marathi. Hence the differences are visible at Level 2.

The results of the Chi-Square values indicate that the distribution of the POS tags in PUNC, QTC, NN, NNP and NST are significantly different. The NST's in Hindi are more as compared to Marathi. As could be seen from Example 2 and Example 4, the NST gets converted exist as independent words in Hindi while they become a part of the compound word in Marathi.

Distribution of POS tags in different types of Documents

The Dataset DS01 consists of two types of documents, i.e. Medical Articles and Medical Reports. The aim of this experiment was to study how the distribution of the POS tags is affected by the type of document.

As could be seen from the results, the percentage of Nouns in the Reports in Hindi and Marathi are 43% and 57%

respectively. While in the case of articles it is 34% and 42% respectively. The count of nouns is more in Reports as compared to Articles. **The reports contain more named entities like name of persons, diseases, medicines etc.** The Reports taken here are pathological reports and discharge summaries hence the effect. Table 4 shows that except for PRFs within pronouns most of the tags show dependency on type of documents.

What is also worth noting is that though there is a decrease in the count of NN's in DS01, there is an increase of NN's in DS02 on the other hand. Due to smaller sentences and not much syntactic complexity in DS02, there is no change from NNP to NN as in in the case of DS01.

Table 4: Chi Squared values of POS tags in DS02

Level 1	Level 2						Level 1	Level 2					
			Articles	Reports	Chisq	Prob				Articles	Reports	Chisq	Prob
Nouns	NN	Hindi	4341	3548	23.8	0	Demonstratives	DMD	Hindi	150	90	18.85	0
		Marathi	4436	4222					Marathi	4	20		
	NNP	Hindi	187	396	15.33	0		DMR	Hindi	10	2	3.22	0.07
		Marathi	43	36					Marathi	222	165		
	NST	Hindi	120	89	8.74	0.0031		DMQ	Hindi	8	0	7.43	0.006
		Marathi	80	108					Marathi	17	19		
Pronouns	PRP	Hindi	362	181	21.34	0		DMI	Hindi	52	72	NA	NA
		Marathi	394	102					Marathi	0			
	PRF	Hindi	61	21	0.21	0.64	Postpositions	PSP	Hindi	2004	1393	NA	NA
		Marathi	11	5					Marathi	0	0		
	PRL	Hindi	0	0	NA	NA	Particles	RPD	Hindi	130	43	0.61	0.8
		Marathi	123	38					Marathi	4	1		
	PRC	Hindi	0	0	NA	NA		INJ	Hindi	0	0	NA	NA
		Marathi	0	1					Marathi	0	0		
	PRQ	Hindi	40	2	1.58	0.21		INTF	Hindi	36	7	2.27	0.13
		Marathi	43	6					Marathi	17	8		
	PRI	Hindi	0	0	NA	NA		NEG	Hindi	99	45	2.316	0.128
		Marathi	0	0					Marathi	26	20		
Verbs	VM	Hindi	1401	600	16.49	0	Quantifiers	QTF	Hindi	257	75	1.168	0.279
		Marathi	1640	908					Marathi	319	112		
	VINF	Hindi	0	0	NA	NA		QTC	Hindi	179	539	16.7	0
		Marathi	0	0					Marathi	150	262		
	VING	Hindi	0	0	NA	NA		QTO	Hindi	9	10	0.844	0.358
		Marathi	0	0					Marathi	11	21		
	VAUX	Hindi	1284	719	0.958	0.327	RDF	RDF	Hindi	0	0	NA	NA
		Marathi	379	233					Marathi	0	0		
Adjective	JJ	Hindi	572	482	3.97	0.046	SYM	SYM	Hindi	745	528	50.26	0
		Marathi	563	563					Marathi	151	24		
Adverb	RB	Hindi	33	6	5.4	0.02	PUNC	PUNC	Hindi	363	173	362.6	0
		Marathi	149	77					Marathi	1108	10		
Conjunctions	CCD	Hindi	499	322	7.39	0.006	UNK	UNK	Hindi	0	0	NA	NA
		Marathi	478	403					Marathi	0	0		
	CCC	Hindi	184	73	13.21	0	ECH	ECH	Hindi	0	0	NA	NA
		Marathi	179	30					Marathi	0	0		
									Total	13309	9462		

CONCLUSIONS

The following conclusions have been made:

1. *Overall the tag distribution is almost identical indicating strong similarity in the grammar of the two languages,*
2. *The count of Nouns changes with domain specificity, so do the verbs of Type VM change from Hindi to Marathi as a result of translation. The Chi-Square values show that the change is not arbitrary but is dependent on the type of the document.*
3. *PSP's in Hindi do not find any translation equivalents in Marathi.*
4. *When Articles vs Reports comparison is considered Chi-squared values are significant in all cases except PRF's under pronouns.*
5. *A Proper Noun gets converted to a Common Noun as a result of translation from Hindi to Marathi.*

REFERENCES

[1] Kaur, A., 2015, "A Novel Approach for Syntactic Similarity Between Two Short Text," International Journal of Scientific & Technology Research, 4(06), 2277-8616.

[2] Li, X., and Li, Q., 2015, "Calculation of sentence semantic similarity based on syntactic structure," Mathematical Problems in Engineering, Article ID 203475. Vol. 2015.

[3] Manning, C. D., andSchütze, H, 1999, "Foundations of statistical natural language processing", Vol. 999, Cambridge: MIT press.

[4] Shih, R. H., Chiang, J. Y., and Tien, F., 2000, "Part-of-speech Sequences and Distribution in a Learner Corpus of English", In ROCLING.

[5] Pak, A., andParoubek, P., 2010, "Twitter as a corpus for sentiment analysis and opinion mining," In LREc, Vol. 10.

[6] Spencer, J., andUchyigit, G., 2012, "Sentimentor: Sentiment analysis of twitter data," In Proceedings of European conference on machine learning and principles and practice of knowledge discovery in databases, p.p. 56-66.

[7] Samy, D., Sandoval, A. M., Guirao, J. M., and Alfonseca, E., 2006, "Building a Parallel Multilingual Corpus (Arabic-Spanish-English)," In Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC.

[8] Campbell, D. A., and Johnson, S. B., 2001, "Comparing syntactic complexity in medical and non-medical corpora," In Proceedings of the AMIA Symposium p.p. 90, American Medical Informatics Association.

[9] Ramanand, J., Ukey, A., Singh, B. K., and Bhattacharyya, P., 2007, "Mapping and Structural Analysis of Multi-lingual Wordnets," IEEE Data Eng. Bull., 30(1), 30-43.

[10] Bender, E. M., 2013, "Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax," Synthesis Lectures on Human Language Technologies, 6(3), 1-184.

[11] Rathod, S., andGovilkar, S., 2015, "Survey of various POS tagging techniques for Indian regional languages," (IJCSIT) International Journal of Computer Science and Information Technologies, IJCSIT, 6(3), 2525-2529.

[12] Dwivedi, P.K., andMalakar, P.K., 2015, "Hybrid approach based POS tagger for Hindi language," International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), p.p. 63-68, ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online).

[13] Dholakia, P. and Yoonus, M., 2014, "Rule Based Approach for the Transition of Tagsets to Build the POS Annotated Corpus," International Journal of Advanced Research in Computer and Communication Engineering, 3 (7).

[14] Thompson, I., 2014, "About World languages: Hindi," http:// aboutworldlanguages.com/ hindi, July, 2014.

[15] Dhore, M. L., Dhore, R. M., and Rathod, P. H., 2013, "Transliteration by orthography or phonology for Hindi and Marathi to English: case study," International Journal on Natural Language Computing (IJNLC), Vol, 2.

[16] Singh, S., andSarma, V. M., 2010, "Hindi noun inflection and Distributed Morphology," In Proceedings of the International Conference on Head-Driven Phrase Structure Grammar (pp. 307-321).

[17] Ramanathan, A., and Rao, D. D., 2003, "A lightweight stemmer for Hindi," In the Proceedings of EACL.

[18] Tidke, C., Binayakya, S., Patil, S., andSugandhi, R., 2013, "Inflection Rules for English to Marathi Translation," International Journal of Computer Science and Mobile Computing, IJCSMC, 2(4), 7-18.

[19] Dolamic, L., and Savoy, J., 2010, "Comparative study of indexing and search strategies for the Hindi, Marathi, and Bengali languages," ACM Transactions on Asian Language Information Processing (TALIP), 9(3), 11.

[20] Deoskar, T., 2006, "Marathi light verbs," In Proceedings from the Annual Meeting of the Chicago Linguistic Society, Vol. 42, No. 2, pp. 183-198, Chicago Linguistic Society.