Creating Multilingual Parallel Corpora in Indian Languages

Narayan Choudhary
JNU, New Delhi
choudharynarayan@gmail.com

Girish Nath Jha
JNU, New Delhi
girishjha@gmail.com

Abstract

This paper presents a description of the parallel corpora being created simultaneously in 12 major Indian languages including English under a nationally funded project named Indian Languages Corpora Initiative (ILCI) run through a consortium of institutions across India. The project runs in two phases. The first phase of the project has two distinct goals - creating parallel sentence aligned corpus and parts of speech (POS) annotation of the corpora as per recently evolved national standard under Bureau of Indian Standard (BIS). This phase of the project is finishing in April 2012 and the next phase with newer domains and more national languages is likely to take off in May 2012. The goal of the current phase is to create parallel aligned POS tagged corpora in 12 major Indian languages (including English) with Hindi as the source language in health and tourism domains. Additional languages and domains will be added in the next phase. With the goal of 25 thousand sentences in each domain, we find that the total number of words in each of the domains has reached up to 400 thousands, the largest in size for a parallel corpus in any pair of Indian languages. A careful attempt has been made to capture various types of texts. With an analysis of the domains, we divided the two domains into sub-domains and then looked for the source text in those particular sub-domains to be included in the source text. With a preferable structure of the corpora in mind, we present our experiences also in selecting the text as the source and recount the problems like that of a judgment on the sub-domain text representation in the corpora. The POS annotation framework used for this corpora creation has also seen new changes in the POS tagsets. We also give a brief on the POS annotation framework being applied in this endeavor.

Keywords: Corpora Creation; Source Text for Multilingual Parallel Corpus; Parallel Corpus in Indian Languages; Tourism and Health Corpus in Indian Languages; Parts of speech annotation; Annotated Corpora; LRL

1. Introduction

Parallel corpora are of great importance in various natural language processing (NLP) and non-NLP tasks. Starting from a comparative and contrastive linguistic analysis for various linguistic features of the languages concerned to machine translation, there are various use for such a corpus in any given language pair.

India is nation with great linguistic diversity with over 452 individual languages listed by Ethnologue[1]. Out of these, 22 languages are listed as ‘scheduled’ (also sometimes called ‘national’) languages comprising a total of 96.56% of the national population[2]. Hindi is the largest spoken language across India (sharing above 41% of the national population) and also the official language of the Indian state (along with English).

Electronic content came rather late into Indian languages. The importance of corpus studies itself came into fore with the prevalence of e-text. In such a scenario, the corpus study in Indian languages was negligible prior to this century. With the advent of common use of computers, the Indian languages also got some share and e-content gradually started growing in Indian languages. Though Unicode standards in Indian languages has helped grow the content, there is not enough content available that can be used to create parallel corpus in Indian languages.

There have been attempts to develop parallel corpora in Indian languages earlier as well. But none of such corpora have been developed from the scratch and is mostly not publically available for the research community. Barring one exception of the EMILLE parallel corpus (Baker, P. et.al., 2004) of 200 thousand words in three languages in general domain, there is no other parallel corpus made in Indian languages. For the annotated parallel corpus, there are none available in Indian languages. To fill this gap, the Department of Information Technology (DIT), Govt. of India sanctioned a project run through a consortium involving 11 institutions across India (Jha, Girish Nath, 2010). This paper presents a summary of the work carried out under this project. This is an attempt to build a representative and comprehensive corpus of two domains in 12 major scheduled Indian languages. The structure of consortium has been given in the following table with the names of the principle investigator, language(s) and the name of the host institute.

Principle Investigator	Language(s)	Host Institute
Girish N. Jha	Hindi,English, Oriya[3]	JNU, New Delhi
S. Virk	Punjabi	Punjabi Uni., Patiala
M.M. Hussain	Urdu	JNU, New Delhi
Niladri S. Dash	Bangla	ISI, Kolkata
M. A. Kulkarni	Marathi	IITB, Mumbai
Kirtida S. Shah	Gujarati	Guj. Uni., Ahm’bad
Jyoti D. Pawar	Konkani	Goa Uni., Goa
S. Arulmozhi	Telugu	Drav. Uni., Kuppam,
S. Rajendran	Tamil	Tamil Uni.,Thanjavur
Elizabeth Sherly	Malayalam	IIITM-K, Trivandrum

[1] http://www.ethnologue.com/show_country.asp?name=in accessed: 4 September, 2011

[2] as per Census of India, 2001 http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement5.htm accessed 4th September, 2011

[3] Oriya was earlier assigned to Utkal University, Bhubaneswar but now it has been transferred to the consortium head institute of JNU, New Delhi.

Table 1: Summary of the Consortium Structure

These languages represent both the two major language families present in India, namely Indo-Aryan and Dravidian. Being the Associate Official Language (AOL) of India, English, a Germanic language, is also included.

The corpora creation has two principal tasks: creation of the raw parallel aligned text and POS annotation. The translation is done manually by especially trained native speakers of the language in their regions. Annotation is also done manually with no use of available automatic taggers.

For translation there are minimal guidelines with respect to format and structure of the target sentences. The source text is formatted to be one sentence per line and each sentence is given a unique identification (ID) number. The translated text in the target languages are also formatted accordingly i.e. they are one sentence per line and correspond to the sentence ID number of the source text. This ensures that we have the source and the target text aligned as we progress. We do not use any alignment tool for this purpose.

Creating the source text is equivalent to corpus creation. As the source text corpus is domain specific and has limitations with regard to the size each of these domains can grow, a careful selection of the text had to be followed. The two domains of health and tourism are not very prolific ones in Hindi. Most of the works done in these two domains are in English. Therefore finding original text in Hindi in these two domains has been a difficult task. The average of words per sentence (out of a total of 25000 sentences per domain) comes out to be 16. Thus we get a corpus consisting of a total of about 400,000 words in each domain.

2. Creating the Source Text

While it is possible to collect the source text online, it is advisable that one should do this with extra caution when creating an ambitious corpus as presented here, particularly for less resourced languages like the Indian languages. Besides, most of the text over the internet would need editing and proofing (Choudhary, N. 2010). For the source text or the base corpus, we first tried selecting text online. But then we realized that most of the text that was available in Hindi over the internet was translated from English or other languages. Besides, our choice necessarily had to be very eclectic as we were specific about the domain and ensure that proper representation was given to the various sub-domains and genres within the domains. So, we went on to collect text from various other sources e.g. promotional materials published and distributed by government and/or private institutions/agencies. We also selected extracts from books, articles and stories from magazines and newspaper.

2.1. Domain Analysis

To ensure that the diversity of the corpus with regard to various types of genres available in the domain is maintained and gets reflected in the corpus, we did a domain analysis before embarking on the text selection. Both the health and tourism domains are vast topics and collecting text within a specific size in the domain necessitated eclecticism. So, we divided both the domains into several sub-domains and gave a priority to the texts that are more common in use. Therefore some sub-domains have greater representation in the corpus than others. These sub-domains were further divided into other categories of text so that a cap is maintained for each variety of text and an even representation of the domain as a whole gets reflected through the corpus.

2.2. Health Domain

Health domain was divided into a total of 16 sub-domains. These sub-domains were made mainly to capture the different disciplines within the medical arena. No sub-domain was allotted to different genres of medical practice like allopath, ayurveda, acupressure, acupuncture etc. However, these were included in the corpus in a certain proportion with the total of the text. For example a disease, its description and symptoms are given only once as these are common in each of the medical practices. It is the diagnosis and treatment where the difference would be reflected.

As summarized in Table 1 below, the health domain has a total of 419420 words, with the total number of words per sentence being 16.77. The total number of unique words in this domain comes out to be 21446.

Major Domains	Domain Code	No. of Sentences	Percentage
Blood, Heart and Circulation	H1	2192	8.76
Bones, Joints and Muscles	H2	1022	4.09
Brain and Nerves	H3	1792	7.17
Digestive System	H4	2175	8.70
Ear, Nose and Throat	H5	620	2.48
Endocrine System	H6	111	0.44
Eyes and Vision	H7	824	3.30
Immune System	H8	634	2.54
Kidneys and Urinary System	H9	575	2.30
Lungs and Breathing	H10	573	2.29
Oral and Dental	H11	610	2.44
Skin, Hair and Nails	H12	2104	8.42
Female Reproductive System	H13	2099	8.40
Male Reproductive System	H14	325	1.30
Life style	H15	4591	18.37
Miscellaneous	H16	3431	13.73
Pediatrics	H17	1321	5.28
Total		25000	100.00

Table 2: Summary of the Health Domain Corpus

2.3. Tourism Domain

Tourism domain was divided into a total of 17 major sub-domains. These were further divided into categories as per the requirement. For example, pilgrimage was divided into two categories of Indian and extra-Indian, ecotourism was divided into wildlife, hill stations, desert and others. There were also sub-domains that did not have any categories like leisure tourism, medical tourism etc. Table 2 below gives a summary of the tourism corpus. The tourism corpus has a total of 396204 words with a per sentence word average of 15.8. Total number of unique words in the tourism corpus is 28542.

Major Domains	Domain Code	No. of Sentences	Percentage

Pilgrimage	T1	3401	13.60
Ecotourism	T2	6803	27.21
Heritage Tourism	T3	4012	16.05
Adventure Tourism	T4	1843	7.37
Mass Travel and Tour	T5	1576	6.30
Leisure Tour	T6	50	0.20
Medical Tourism	T7	16	0.06
Nautical Tourism	T8	605	2.42
Culinary Tourism	T9	144	0.58
Disaster Tourism	T10	171	0.68
Dark Tourism	T11	16	0.06
Space Tourism	T12	0	0.00
War Tourism	T13	9	0.04
Shopping Tourism	T14	202	0.81
Others	T15	2037	8.15
General Description	T16	4115	16.46
Total		25000	100.00

Table 3: Summary of the Tourism Domain Corpus

3. Data Storage, Maintenance and Dissemination

The Hindi source data collected manually with careful selection criteria in mind was mostly typed out by language editors. Out of the 25 thousand sentences in each of the domains only a meager 1500 sentences or 6% were taken from the internet. The whole of the corpus was first typed into spread sheets on normal PCs by the language editors of the source text. It was further validated by the present authors. Each sentence in the corpus has a unique ID which gets carried forward to each of the translated languages. Thus the alignment is done simultaneously as the translation in each of the languages progresses.

All the data collected and incorporated in the source text are stored with their metadata information which includes various information e.g. the source, number of words selected from the source, names of the authors/copyright holders and their sub-domain details. For the archiving purposes, all the source text is hyperlinked with a scanned image file of the source document from where the text was taken.

The source text is encoded in Unicode. All the translated texts in other languages are also in Unicode. As for the quality of the source or the translated text, we believe this to be the best possible. We say the source text to be the best possible for a corpus because of the following reasons:

a. it is typed by the trained language editors
b. it has been internally and externally validated by language experts

For the translated text, usually we seek only one translation. However, wherever possible, if two or more options are available for a sentence, we encourage the translators to provide those translations as optional ones. Once the text was translated, we evaluated the translation through external evaluators of each of the language pairs and the suggestions/corrections recommended by them have been incorporated in the target text. The whole of the corpus creation process has been supervised and the corpus principally has 0% ‘noise’ in terms of spelling mistakes, wrong character encodings, incorrect translations etc.

Govt. of India has started a data centre (http://tdil-dc.in) The ILCI corpora is in the process of being uploaded to this data centre and will be available for free download as per the Govt. of India guidelines.

4. Parallel Corpus Creation and Alignment

As noted above, the parallel corpora are created simultaneously, in each of the language pairs as the translation progresses. As the source text is created it is electronically sent to the other members of the consortium where the respective translators translate the source text in the respective target languages and send back to us. We achieved the raw parallel sentence-wise aligned corpus in a period of about one and a half years.

5. POS Annotation

We are now in the middle of the second principal task - doing POS annotation of the parallel corpora in each language. The annotations are done manually for each of the languages. Although there are some POS taggers available for some of the Indian languages, their efficacy and standard input/output has been doubtful. Moreover, Indian did not have a common standard till very recently when it got its first national standard in POS annotation through the efforts of BIS and ILCI. There are some POS taggers developed with various accuracies for the languages like Hindi (Shrivastava, M. & Bhattacharya, P., 2008), Telugu (Avinesh, PVS & G. Karthick, 2007), Bengali (Dandpat, S. et.al., 2007) etc. (for a survey report see Kumar, D. & Josan, G.S., 2010.) The Indian Languages Machine Translation (ILMT) project funded by DIT (Govt. of India) claims to have developed several POS taggers, but they are yet to find users in the corpora community.

5.1. POS Tagset

Until recently, there have been two major types of tagsets used for POS annotation of texts in Indian languages. These two include a tagset developed by IIIT Hyderabad (Bharti, A. et.al., 2006) and another one developed under the leadership of Microsoft Research, India (MSRI), known as IL-POST (Baskaran, S. et.al., 2008). The IIIT tagset is a flat tagset based on the Penn tagset (Santorini, B. 1990) with some modifications to suit major Indian languages. The IL-POST tagset is a rather new annotation framework put to use in Indian languages. The IL-POST framework provides for a hierarchical, multi-layered tagset where much of the linguistic information is captured through explicit tags, including the ones that can be possibly identified through a morphological analyzer. Advocates of the IIIT tagset emphasizes that the information that can be extracted through the use of a language specific morphological analyzer should not be marked manually because it would only increase the amount of human labor put to use.

There is no sizeable POS annotated corpus available in any of the Indian languages at present. As POS annotation is a part of this project, the tagset to be used for the corpora of these 12 languages became an issue. Several meetings were held under the aegis of BIS to come to a conclusion. Finally, a POS tagset was agreed upon by the stake-holders. This tagset has come to be known as the BIS parts-of-speech annotation tagset[4].

The BIS Tagset contains the features of the hierarchical tagset. However, it has tags for only first two tiers of linguistic information (POS and their subtypes) and excludes information from tier three onwards as these can be provided by morph analyzers and parsers. Morphological analyzers are available for some of the languages in the group and many more are in the process of being developed. For Hindi, morphological analyzers have been reported from various quarters e.g. (Goyal, V. & Singh Lehal, G. 2008; Bögel, T. et.al., 2007; etc).

5.1.1. Principles for Designing Linguistic Standards for Corpora Annotation

The BIS standard has set the following principles for designing linguistic standards for corpora annotation.

i. Generic Tag Sets

ii. Layered approach

Layer I: Morphology

Layer II: POS <morphosyntactic>

Layer III: LWG

Layer IV: Chunks

Layer V: Syntactic Analysis

Layer VI: Thematic roles/Predicate Argument structure

Layer VII: Semantic properties of the lexical items

Layers VIII, IX... Word sense, Pronoun referents (Anaphora), etc,

iii. Hierarchy within each layer

iv. Extensibility (including the language specific requirements and additional languages)

v. If a tag is redundant for a language, it should be deprecated

vi. ISO 639:3[5] Language code should be used <in metadata>

vii. Follow global guidelines such as EAGLES (Leech, G. & Wilson, A. 1999) where available.

viii. Standards should be mappable to/compatible with existing schemes to and from

ix. Standard is designed to handle wide range of applications and also should support all types of NLP Research efforts independent of a particular technology development approach

x. The scheme should be Annotator friendly.

5.1.2. Super Set of POS Tags

Guided by the principles above, a super set of POS tags for Indian languages has been developed (Appendix I). Tagsets for different Indian languages have been drawn from this super tagset. As can be seen in Appendix I below, there are 11 top level categories. These are further classified into types and subtypes. There are a total of 45 tags in this set. If a language demands further sub-types, the principles above allow that. However, top level categories cannot be changed or new top level categories are not recommended to be added. No individual language has used all of these categories. The tagsets for all the 12 languages have been drawn from this super tagset.

5.2. Manual POS Annotation

The annotation is being done manually by the language experts/native linguists following the annotation guideline prepared for respective languages. There are some languages in the group that are morphologically agglutinating. For such languages direct annotation is not possible and morphological segmentation is required before POS annotation can begin. For such languages e.g. Tamil, Telugu and Malayalam, segmentation is recommended as a pre-processing task before the POS annotation.Additonally, a server-based, access-anywhere, annotation tool is put in place where the annotators can annotate the text in their respective language over the internet. The tool can be accessed here:

http://sanskrit.jnu.ac.in/ilciann/index.jsp

6. Conclusion

In this paper we have presented a description of processes involved in creating the parallel corpora in two specified domains of health and tourism for 12 major Indian languages. We have shown how the source text was created and how the raw corpora in target languages have been created/translated.

We have shown the representation given to different genres of writing within these two domains and tried to show that the source corpus created represents the domains under study in their totality. As the source text is created specifically for parallel corpora development, we have also shown that the process of its creation gives us an aligned corpus by default. Qualitatively, the corpus created is richer than other corpora in terms of lack of noise and integrity. This can be corroborated with the fact the corpus does not have any spelling mistakes, errors of character encoding (as is common in Indian languages written in their native scripts), and that the translations have been verified through external evaluators.

For POS annotation, we are following the latest annotation framework approved by the BIS and the annotated corpora generated through this task will prove to be a great resource in the NLP and related areas of Indian languages in particular and other languages in general.

The process of corpora creation has been though labor intensive, the result so far is worthwhile. Additional Indian languages will be added following the same process. That is the source Hindi text can be translated into any language and then POS tagged. This will give newer pairs of parallel corpora in 12 languages simultaneously.

The chosen two domains are of great importance in itself as both the health and tourism are the focus areas of any government in general and the Indian government in particular. Both the raw corpus and the annotated corpus can be used for various purposes of language engineering and linguistic analysis.

By the end of the project, we expect to achieve one million tagged words in each of the 12 languages because there is some additional data collection in the process of selecting 50 thousand word in each language.

[4] No standard published reference can be given for this tagset as yet. We refer to the document circulated in the consortia meetings. This document was referred as "Linguistic Resource Standards: Standards for POS Tagsets for Indian Languages", ver. 005, August, 2010.

[5] http://www.sil.org/iso639-3/default.asp

References

Avinesh, P. V. S. & G. Karthik. (2007). Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation-Based Learning. In: *Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL)* (2007), pp. 21-24.

Baker, Paul, Andrew Hardie, Tony McEnery, Richard Xiao, Kalina Bontcheva, Hamish Cunningham, Robert Gaizauskas, Oana Hamza, Diana Maynard, Valentin Tablan, Cristian Ursu, B. D. Jayaram, and Mark Leisher. (2004). Corpus Linguistics and South Asian Languages: Corpus Creation and Tool Development. In: *Literary & Linguistic Computing*, 19: 509-524

Baskaran, Sankaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha and K.V. Subbarao. (2008). A Common Parts-of-Speech Tag Set Framework for Indian Languages. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Daniel Tapias (Eds.) *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Bharati, Akshar, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal. (2006). *Anncorra: Annotating Corpora*. LTRC, IIIT, Hyderabad

Bögel, Tina, Miriam Butt, Annette Hautli, and Sebastian Sulger. (2007). Developing a Finite-State Morphological Analyzer for Urdu and Hindi. In: *Proceedings of the Sixth International Workshop on Finite-State Methods and Natural Language Processing*. Potsdam.

Choudhary, N. (2011). Web-drawn corpus for Indian Languages: A Case of Hindi. In: *Proceedings of Information Systems for Indian Languages*. Volume 139, Part 2, 218-223. Springer Verlag.

Goyal, V. & Singh Lehal, G. (2008). Hindi Morphological Analyzer and Generator. In: *Proceedings of the First International Conference on Emerging Trends in Engineering and Technology*.

Kumar, Dinesh & Gurpreet Singh Josan. (2010). Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey. In: *International Journal of Computer Applications*. Vol. 6(5):1–9. Foundation of Computer Science.

Dandapat, Sandipan, Sudeshna Sarkar, Anupam Basu. (2007) Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario. In: *Proceedings of the Association for Computational Linguistic*, pp 221-224

Jha, Girish Nath. (2010). The TDIL Program and the Indian Language Corpora Initiative (ILCI). In: Calzolari, Nicolai et.al. (eds) *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Leech, G. & Wilson, A. (1999). Standards for Tagsets. In: van Halteren, H. (ed.) *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora*. http://www.ilc.cnr.it/EAGLES96/browse.html

Santorini, Beatrice. (1990). *Part-of-speech Tagging Guidelines for the Penn Treebank Project*. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Shrivastava, Manish and Pushpak Bhattacharyya. (2008) Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge. In: *Proceedings of the International Conference on NLP (ICON08)*, Pune, India,

Appendix I: Super Set of POS Tags for Indian Languages

Sl. No.	Category (Category.Type.Subtype)	Label	Annotation Convention
1	1 Noun	N	N
2	1.1 Common	NN	N_NN
3	1.2 Proper	NNP	N_NNP
4	1.3 Verbal	NNV	N_NNV
5	1.4 Nloc	NST	N_NST
6	2 Pronoun	PR	PR
7	2.1 Personal	PRP	PR_PRP
8	2.2 Reflexive	PRF	PR_PRF
9	2.3 Relative	PRL	PR_PRL
10	2.4 Reciprocal	PRC	PR_PRC
11	2.5 Wh-word	PRQ	PR_PRQ
12	3 Demonstrative	DM	DM
13	3.1 Deictic	DMD	DM_DMD
14	3.2 Relative	DMR	DM_DMR
15	3.3 Wh-word	DMQ	DM_DMQ
16	Verb	V	V
17	4.1 Main	VM	V_VM
18	4.1.1 Finite	VF	V_VM_VF
19	4.1.2 Non-finite	VNF	V_VM_VNF
20	4.1.3 Infinitive	VINF	V_VM_VINF
21	4.1.4 Gerund	VNG	V_VM_VNG
22	4.2 Auxiliary	VAUX	V_VAUX
23	5 Adjective	JJ	
24	6 Adverb	RB	
25	7 Postposition	PSP	
26	8 Conjunction	CC	CC
27	8.1 Co-ordinator	CCD	CC_CCD
28	8.2 Subordinator	CCS	CC_CCS
29	8.2.1 Quotative	UT	CC_CCS_UT
30	9 Particles	RP	RP
31	9.1 Default	RPD	RP_RPD
32	9.2 Classifier	CL	RP_CL
33	9.3 Interjection	INJ	RP_INJ
34	9.4 Intensifier	INTF	RP_INTF
35	9.5 Negation	NEG	RP_NEG
36	10 Quantifiers	QT	QT
37	10.1 General	QTF	QT_QTF
38	10.2 Cardinals	QTC	QT_QTC
39	10.3 Ordinals	QTO	QT_QTO
40	11 Residuals	RD	RD
41	11.1 Foreign word	RDF	RD_RDF
42	11.2 Symbol	SYM	RD_SYM
43	11.3 Punctuation	PUNC	RD_PUNC
44	11.4 Unknown	UNK	RD_UNK
45	11.5 Echo-words	ECH	RD_ECH