Divergences in English-Hindi Parallel Dependency Treebanks

Himani Chaudhry, Himanshu Sharma and **Dipti Misra Sharma**
Language Technologies Research Centre
International Institute of Information Technology-Hyderabad
India
{himani,himanshu.sharma}@research.iiit.ac.in
{dipti}@iiit.ac.in

Abstract

We present, here, our analysis of systematic divergences in parallel English-Hindi dependency treebanks based on the Computational Paninian Grammar (CPG) framework. Study of structural divergences in parallel treebanks not only helps in developing larger treebanks automatically, but can also be useful for many NLP applications such as data-driven machine translation (MT) systems. Given that the two treebanks are based on the same grammatical model, a study of divergences in them could be of advantage to such tasks, along with making it more interesting to study how and where they diverge. We consider two parallel trees divergent based on differences in constructions, relations marked, frequency of annotation labels and tree depth. Some interesting instances of structural divergences in the treebanks have been discussed in the course of this paper. We also present our task of alignment of the two treebanks, wherein we talk about our extraction of divergent structures in the trees, and discuss the results of this exercise.

1 Introduction

Treebanks play an increasingly important role in computational linguistics, and with the availability of a number of treebanks of various languages now, studies based on parallel treebanks are one of the directions application/use of treebanks has taken. "Such resources could be useful for many applications, e.g. as training or evaluation corpora for word and phrase alignment, as training material for data-driven MT systems and for the automatic induction of transfer rules" (Hearne et al., 2007) and so on. However, though recent years have seen an increasing interest in research based on parallel corpora, "surprisingly little work has been reported on parallel treebanks." opine Volk et al. (2004). "A parallel treebank comprises syntactically annotated aligned sentences in two or more languages. In addition to this, the trees are aligned on a sub-sentential level." (Tinsley et al., 2009)

In this paper we report our study on parallel English and Hindi dependency treebanks based on the CPG model. The annotation labels used to mark the relations in the example trees here (as also in the treebanks) conform to the dependency annotation scheme given by Begum et al. (2008). An adaptation of this scheme was subsequently used for the English treebank, as reported in (Chaudhry and Sharma, 2011)

We detail here, how we make use of the existing Hindi dependency treebank and its parallel English dependency treebank, to study systematic divergences in the treebank pair, given that both of these treebanks use the same dependency grammar formalism. We sought to find here, the types and reasons for these differences. We find that the two treebanks diverge mainly from two aspects:

- Stylistic
- Structural

A good example of stylistic variation or translator preference, from our data would be:

(1) kendrIya sarkAr-ke anek varishtha netA bhI mojUd the.

kendriiya	*sarkaar-ke*	*anek*	*varishtha*
ruling	party-of	many	senior
netaa	*bhii*	*mojuud*	*the*
leaders	also	in-attendance	were

'A number of senior leaders from the ruling party were also in attendance.'

Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 33–40,
Prague, August 27–30, 2013.

The Hindi sentence in example (1), has been translated as '*A number of senior leaders from the ruling party were also in attendance.*' in our corpus. While another possible (more regular/natural) translation of the sentence would be:

A number of senior leaders from the ruling party were also present.

Stylistic divergences can be due to preferred translations in the language or due to the lexical choice of the translator, (or even translator's preference for a specific type of constructions). This said, though study of stylistic divergence can help recognize preferred constructions in a given language, this would need copies of translations by multiple translators to perform exercises such as inter-translator agreement. Since our data has only one translator, analysis of stylistic divergences is beyond the scope of the work we report here.

Structural divergence, thus, is the focus of our study here, as it abounds in these treebanks and brings forth interesting examples of divergences between the two treebanks. We discuss some occurrences of it in our data. Further, we aim to see if some systematic patterns of divergence could be arrived at, in the treebanks, through a comparative study of the structures of their trees. However, since this is a work in progress, we are yet to sum up any such generalizations, and we do not include them here.

The remainder of this paper is organized as follows: Section 2 gives some background on the data, the annotation scheme and the methodology we used for our study. In Section 3 we take a look at the dissimilarities in the two treebanks, and discuss our investigations into the reasons behind them. Section 4 presents our observations. Section 5 presents the task of our alignment of the two treebanks, where in we talk about our extraction of divergent structures in the trees, and discusses the results of this exercise. And, in Sections 6, we conclude and sketch the possibilities for some future work in this direction.

2 Methodology

2.1 The Data

The data for this study comprises a set of parallel English and Hindi dependency treebanks. A small section (25000 words) of the Hindi dependency treebank (Bhatt et al., 2009) (being developed at IIIT-H, under the Hindi-Urdu Treebank (HUTB) project) was translated to English to form a parallel corpus. The English treebank used here (reported in (Chaudhry and Sharma, 2011)), has been developed on these translations and has 1143 sentences annotated using the dependency annotation scheme modeled on the CPG framework (Begum et al., 2008) (as also the Hindi treebank used here).

2.2 The Annotation Scheme

As mentioned earlier, the annotation scheme used for the creation of the two parallel dependency treebanks (English and Hindi) is based on CPG, a dependency grammar model proposed by Bharati et al. (1995). This annotation scheme, developed for Hindi and other Indian languages, by Begum et al. (2008) was later applied to English first by Vaidya et al. (2009) and then, by Chaudhry and Sharma (2011) to develop their English dependency treebank (used for this work). Paninian Grammar assigns 'karaka' (verb argument relations) to arguments in a sentence, based on the relationship they have with the verb. "karaka relations are syntactico-semantic (or semantico-syntactic) relations between the verbals and other related constituents in a sentence." (Bharati et al., 1995). There are six basic karakas, namely *adhikarana* 'location', *apaadaan* 'source', *sampradaan* 'recipient', *karana* 'instrument', *karma* 'theme', *karta* 'agent'. It must be noted, that though the first four karakas (as listed here) can be roughly mapped to their thematic role counterparts, *karma* and *karta* tend to be different from 'theme' and 'agent' respectively". (Begum et al., 2008) Further, the annotation directly represents the relations between a syntactic head and its arguments and adjuncts (that is, its dependents or modifiers) in a sentence/clause. It is noteworthy, that the main verb is taken to be the central and binding element of the sentence, and is therefore, the root node of a dependency tree, per the annotation scheme. However, there can be exceptions to this, such as in the cases of co-ordination, where a co-ordinating conjunct co-ordinates sentences/clauses that do not have dependencies over/with each other. For example:

'Ram ate an apple and Ravi drank milk.'

Here, the two verbs 'ate' and 'drank' are the root nodes for their respective sentences, and these

two are then co-ordinated by the co-ordinating conjunct 'and', which is taken as the head of the entire co-ordinated structure.

Futher, two types of relations are marked under this scheme–karaka and others. (Bhatt et al., 2009). Relations other than karakas, such as purpose, reason, and possession (adjuncts) and also, non-dependency relations as in co-ordination and light verb constructions etc., too are therefore, taken care of, using the relational concepts prescribed by this annotation scheme. *Table 1* provides information about the relation labels (from the two treebanks) referred to, in this work.

The dependency relations are marked at inter-chunk level, instead of marking relations between words. So, function words are attached to (chunked with) their lexical heads. Per this scheme, a chunk (with boundaries marked), by definition, represents a group of adjacent words in a sentence, which are in dependency relation with each other, and where one of them is their head (Mannem et al., 2009).

2.3 Procedure

For the purpose of a detailed comparative study of the two treebanks, about 700 sentence pairs from the two treebanks were manually aligned at sentence level and the trees were then aligned automatically. "A *sentence pair* is a pair of sentences which are translations of each other, and the dependency trees for the two sentences in a sentence pair form a *tree pair*." (Georgi et al., 2012)

After this, various instances of the dependency relations in the parallel sentences were automatically extracted for the study. We then (manually) compared the tree pairs as regards their similarities and contrasts. The comparisons were made not just for their spans as complete trees, but also at the level of their subtrees. Given a sentence pair, we first observed the entire tree spans for potential divergences. And, if they were divergent, we looked further on, to find where they diverged, followed by how much they diverged, and why. This has been discussed in detail in section 5. We sought to find what type of divergences they were. We talk in detail of these aspects the two treebanks were compared on, in Section 3.

3 Divergence Types

The two treebanks were considered 'divergent' if the paralle trees fell under any of the following:

- Differences in the construction (structure)
- Difference in relations marked (on the parallel sentences)
- Difference in tree depth
- Difference in the frequency of annotation labels

3.1 Difference in construction

Changes in lexical category of a word of one language and its counterpart in the other, lead to Categorial divergence visible in the data. '*It suffices.*' would be translated in Hindi as '*yaha kAfI hE.*' (It sufficient is). While the word '*suffices*' is realized as the main verb in English it is an adjectival modifier '*kAfI*' (sufficient) in the phrase '*kAfI hE*', in Hindi. *Figure 1* shows the divergent trees for the sentence pair.

(2) Hindi: 'yaha kAfI hE.'

yaha *kaafii* *hE*
It sufficient is
English: 'It suffices'

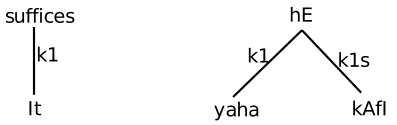

Figure 1: Example showing categorial divergence.

Event verbs of English such as '*flagged*' or '*flagged off*' may not have Hindi equivalents. In such cases they are substituted with description/descriptive phrases such as '*jhandI dikhA kar ravAnA kiyA*', as seen in *figure 2*, for the sentence pair:

(3) Hindi: 'unhone tren ko jhandI dikhA kar ravAnA kiyA.'

unhone *tren* *ko* *jhandii* *dikhaa* *kar*
He train to flag show do
ravaanaa *kiyaa.*
send did
English: 'He flagged off the train.'

Label Name	Relation Name	Description
k1	karta	Doer/agent/subject.
k1s	karta samaanaadhikarana	Noun complement of karta.
k2	karma	Object/patient.
pof-phrv	Phrasal verb	Part of units in phrasal verb constructions.
vmod	Verb modifier	General verb modification.
k3	karana	Instrument that helps achieve the action/activity.
k4a	anubhava karta	Experiencer.
k7	vishayaadhikarana	Abstract location in time or place.
r6	shashthi	The genitive/possessive relation between nouns.
nmod__emph	emphatic marker	noun modifier of the type emphatic marker.
k7p	deshaadhikarana	Place/Location.
fragof	Fragment-of	Relation to link elements of a fragmented chunk.
k5	apaadaana	A point of separation/departure from source.
ccof	Conjunct-of	Co-ordination and sub-ordination.
k7t	samayaadhikarana	Location in time.
nmod	Noun modifier	General noun modification, including participles.
pof	Part-of relation	Part of units such as light-verb+noun.
r6-k1	karta of noun in 'part-of' relation	Karta of noun in light-verb+noun construction.
r6-k2	karma of noun in 'part-of' relation	Karma of noun in light-verb+noun construction.
rs	Relation samaanaadhikarana	Noun complement/elaboration.
sent-adv	Sentential Adverb	Adverbial expression with a sentence in its scope.

Table 1: Description of Dependency Relations.

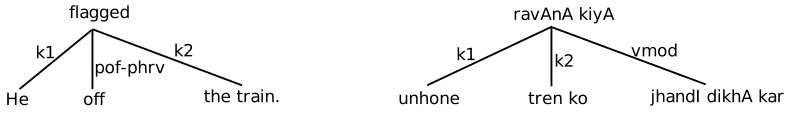

Figure 2: Categorial divergence due to Event verbs (Example (3)).

3.2 Difference in relations marked

We see that the frequency of the core arguments (such as 'karta', 'karma' and thus, the labels pertaining to them) does not vary much, between the two languages, since these are mandatory arguments for both of the languages, and must be present. However, these relations (and their labels) may not always match for all of the trees, of the two treebanks, since there are instances where a word that is a mandatory argument in one language data may realize differently in the other. This happens in the case of other arguments too. For example, 'preposition-stranding' in English is another reason for difference in dependency relations marked on parallel trees. This is because preposition-stranding is specific to English, and is

not found in Hindi, which has postpositions that are required to follow the noun they are associated with. Prepositions of English are different from Hindi postpositions which seldom occur discontiguous with the noun they relate with, and never due to movement. Occasional examples that one comes across, of a Hindi postposition separated from its noun, are due to translational choices or due to some additional information about the noun (in written texts). Thus, Hindi doesn't have the phenomena of 'stranding'. An example of this kind of divergence would be:

(4) Hindi: 'jUn kaun-sI dukAn mein gayI?'

juun kaun-sii dukaan mein gayii?
June which shop in go+PAST
English: 'Which shop did June go to?'

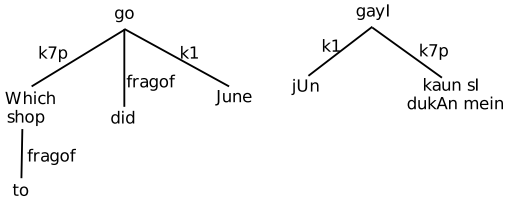

Figure 3: Divergence due to 'preposition-stranding' in English.

In example (4), while in English the preposition '*to*' will have the relation 'fragment-of' (annotated with the label 'fragof') with the noun phrase (NP) '*which shop*', to indicate that though separated from it, the preposition is related to the NP. It may be noted that noun within the NP of a Preposition Phrase (PP) is considered the head of the phrase, in our analysis. In its Hindi counterpart, the postposition '*mein*' will be part of the NP preceding it, and doesn't need to be annotated separately. Also, the auxiliary '*did*' will be a 'fragof' of the verb '*go*' because the auxilliary '*did*' and the verb '*go*' are discontiguous here. While in Hindi the verb is a single word expression. Thus, as seen in *figure 3*, the English tree has extra relations marked in it, making the two trees divergent.

Null subject divergence is another major aspect leading to divergences in the two treebanks. In Hindi the subject of a sentence is left to be implicit many times, since Hindi allows dropping of the subject where it is obvious. This is not so with English. Being a positional language English encodes much information in the subject (even object) position, hence the subject can't be dropped. For an insight into subject dropping in Hindi, let us look at example (5) in (*figure 4*)

(5) Hindi: '(tum) kyA kar rahe ho?'

(tum) kyaa kar-rahe-ho?
(you) what do+CONT+be+PRES
English: 'What are you doing?'

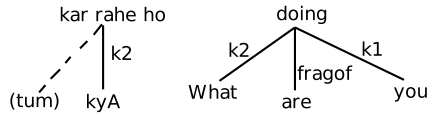

Figure 4: Example of null subject divergence.

Thus, while it is possible to ask someone '*kyA kar rahe ho?*' in Hindi, it is ungrammatical to ask '*what (are) doing?*' in English, dropping the subject, in such sentences. Divergence is bound to creep in, between the trees of two parallel sentences, in terms of dependency relations as well as labels, for such instances.

3.3 Difference in tree depth

Varying relations (not just their number, but their types too) affect the depth of trees from one language to the other. For instance, the presence of modifier-modified relations such as 'nmod' (noun modifier) or fragmented chunks (depicted with the label 'fragof' in our annotation), in the sentences of one language, and their absence in the parallel sentence in the other, can cause such divergences. This leads to a difference in the depths of the two trees, as is evident from the trees in *figure 3* of example (4).

3.4 Difference in the frequency of annotation labels

We also automatically extracted the relation labels from the parallel dependency treebanks, and studied their instances in the data, based on their high frequency or paucity in either of the language's treebank (The automatic extraction and its results are discussed in section 5). In cases where we found consistency in divergence patterns we investigated further to analyze what lay beneath their surfaces.

Tag/Label	Relation Name	English Count	Hindi Count
ccof	Conjunct-of	1161	730
k1	karta	1206	1032
k1s	karta samaanaadhikarana	153	160
k2	karma	873	1040
k7	vishayaadhikarana	460	282
k7p	deshaadhikarana	272	200
k7t	samayaadhikarana	396	297
nmod	Noun modier	256	296
pof	Part-of relation	781	63
r6	shashthi	935	284
r6-k1	karta of a noun in a lightverb+noun construction	53	04
r6-k2	karma of a noun in a lightverb+noun construction	151	14
r6v	Genetive relation with verb	04	0
rs	Relation samaanaadhikarana	45	0
sent-adv	Sentential Adverb	44	68
vmod	Verb modier	218	175

Table 2: A comparative Dependency Relations count.

For instance, the frequency of the 'part-of' relation label, 'pof', in Hindi and paucity of the same in English (as seen in *Table 2*) point to the fact that Hindi abounds in complex predicates, where as English has few instances of them. As mentioned earlier in the discussion, the noun components of conjunct verbs are annotated with the label 'pof' to convey that that noun has a 'part-of' relation with the verb it is attached to. Another relation label that needs mention here, is 'r6v'. While there are instances of this in the Hindi side of the data, there are none in English. The reason being, this is a relation that attaches to the verb, though not a karaka relation. It indicates a sense of possession, so it is given the tag 'r6v', where 'r6' indicates a possession relation, and 'v' indicates that this relation is marked with the verb. There are no instances of this relation in the English data as this type of realization wasn't encountered in English. The relation tag hasn't therefore, been included in the annotation scheme for English, as of now.

4 Observations

English and Hindi being significantly divergent, we came across varied instances of diversities in the two treebanks. The instances of English manner-motion verbs we came across in the data seemingly indicate regular divergence in that English has the tendency to pair up with satellite prepositions such as 'into' in the expression 'danced into', to form manner-motion verbs. Whereas, Hindi resorts to using separate verbs for manner and motion to represent the action as a whole. For example, 'He broke into the house.' would translate as 'vah zabardastI ghar main ghusA.' (he forcefully home-in enter). Another example for this would be, 'She danced into the room.' which translates as 'vah nAchte hue kamare main ghusI.' (she dance+cont+manner room-in enter).

Another divergence is that English induces *expletives* to fill the canonical subject position in a sentence, in the absence of a logical subject. However, Hindi can conveniently drop the subject as and when. An example for this would be:

(6) Hindi: 'bAhar bArish ho rahI hE.'

baahar *baarish* *ho-rahii-hai.*
outside rain be+PRES+CONT
English: 'It is raining outside.'

The examples show that the two sentences diverge syntactically, since the Hindi sentence has no equivalent for 'it', here. However, our annotation scheme licenses incorporation of semantic information along with syntactic analysis (being syntactico-semantic). This said, if we delve a little into their semantics, we see that the dissimilar-

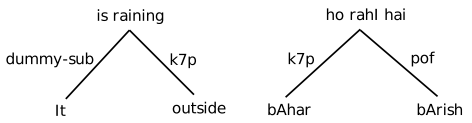

Figure 5: Observation: No 'karta'

ity isn't as pronounced. Expletive 'it', though in the subject position in the sentence, is annotated 'dummy-subject' of the verb. Thus there isn't a *karta* in the English sentence, as also is the case for its Hindi counterpart.

5 Automatic Tree Comparison and Results

In this section, we discuss the structural comparison between the two treebanks, and its results.

5.1 Comparison Criteria

The basis of this comparison is the divergence in the tree structures and in the labels. For any given pair of source (English) sentence, target (Hindi) sentence and the respective Source Sentence Dependency Tree (sTree) and Target Sentence Dependency Tree (tTree), the comparison criteria was:

- Full Structure Match
- Partial Structure Match

5.1.1 Full Structure Match

We consider it a full structure match if the full structure of the sTree matches the tTree. Starting with the ROOT Node, the child nodes in a sub-tree of an sTree are matched with the child nodes of the corresponding sub-tree in the tTree. This process is repeated recursively for each node. If there is a full structure match, then the number of chunk nodes in the sTree matches the number of chunk nodes in the tTree and the tree structure is exactly similar. There are 15 sentences where the structure of the sentences is similar in both the languages.

5.1.2 Partial Match

Partial match between sTree and tTree is calculated on the basis of:

1. Argument/Arc Match for a given node: To see if a particular node has a fixed number of arguments in both the languages.
2. Particular Label Match: If a particular node (event) demands an argument with a particular label, then the label is bound to occur in both of the languages. For Example, if an event X has a 'k1' in its demand-frame for an sTree, and the construction and the lexical choice of words imply that 'k1' should occur in the tTree as well, then there is a potential positive case for Label Match, regardless of the lexical items assigned to the label in the tree pair.
3. Both, Argument and Label Match

5.2 Results

In this section, we take a look at the results of Structural Comparison. For partial sub-tree matching, we calculated the number of sub-trees that have the same number of arguments from a set of possible subTrees.
In our data, 113 sub-trees (*Same Argument Count*) out of 215 (*Total Sub-trees*) were found satisfying the criteria.

In the calculation of Labelled Accuracy, three types of statistics were calculated. "*Common Labels*" gives the number of labels that were shared by the aligned node in both, the source (S) and the target (T) language (L). "*Source Unique Labels*" shows the number of labels owned only by the SL that were not present in the TL. While "*Target Unique Labels*" shows the number of labels present in the TL, but not present in the SL.
Their values for our data are:
$CommonLabels = 371$
$SourceUniqueLabels = 209$
$TargetUniqueLabels = 219$

6 Conclusion and Future Works

In this paper we looked at the divergences in the CPG based English and Hindi parallel treebanks. The English treebank varies from its Hindi counterpart in certain aspects, (in spite of being based on the same grammatical model, and using a quite similar annotation scheme) given the dissimilarities between the two languages. The treebank pair was compared and contrasted based on differences in constructions, relations marked, frequency of annotation labels and tree depth. The tree pairs were considered divergent if their differences fell under one of the criteria above.

Further, we investigated into the reasons behind these divergences. Though we have calculated the

extent of divergences in the treebanks, at this point we do not make any generalizations about them. Our observations and their classifications regarding these treebanks can provide insights into improvement of algorithms used for NLP tasks, especially Machine Translation.

Also, as a future work, stylistic divergences between parallel treebanks can be an interesting subject of study, with the availability of data suited to the needs of this task.

Acknowledgments

We gratefully acknowledge the provision of the useful resource by way of the Hindi Treebank developed under HUTB, of which the Hindi treebank used for our research purpose is a part, and the work for which is supported by the NSF grant (Award Number: CNS 0751202; CFDA Number: 47.070). Also, any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for indian languages. In *Proceedings of IJCNLP*.

Akshar Bharati, Vineet Chaitanya, Rajeev Sangal, and KV Ramakrishnamacharyulu. 1995. *Natural language processing: a Paninian perspective*. Prentice-Hall of India New Delhi.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189. Association for Computational Linguistics.

Himani Chaudhry and Dipti M Sharma. 2011. Annotation and issues in building an english dependency treebank.

Ryan Georgi, Fei Xia, and W Lewis. 2012. Measuring the divergence of dependency structures cross-linguistically to improve syntactic projection algorithms. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12), Istanbul, Turkey. European Language Resources Association (ELRA)*.

Mary Hearne, John Tinsley, Ventsislav Zhechev, and Andy Way. 2007. Capturing translational divergences with a statistical tree-to-tree aligner.

Prashanth Mannem, Himani Chaudhry, and Akshar Bharati. 2009. Insights into non-projectivity in hindi. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 10–17. Association for Computational Linguistics.

John Tinsley, Mary Hearne, and Andy Way. 2009. Exploiting parallel treebanks to improve phrase-based statistical machine translation. In *Computational Linguistics and Intelligent Text Processing*, pages 318–331. Springer.

Ashwini Vaidya, Samar Husain, Prashanth Mannem, and Dipti Misra Sharma. 2009. A karaka based annotation scheme for english. In *Computational Linguistics and Intelligent Text Processing*, pages 41–52. Springer.

Martin Volk and Yvonne Samuelsson. 2004. Bootstrapping parallel treebanks. In *Proc. of Workshop on Linguistically Interpreted Corpora (LINC) at COLING*.