Annotation and Issues in Building an English Dependency Treebank

Himani Chaudhry, Dipti M. Sharma
Language Technologies Research Centre,
International Institute of Information Technology,
Gachibowli, Hyderabad, India-500032
himani@iiit.ac.in
dipti@iiit.ac.in

Abstract
The Paninian Grammar framework, given by Panini for his analysis of Sanskrit Language, is finding its extensive application on languages other than Sanskrit, about two thousand five hundred years after its formulation. The work presented in this paper is one such application that extends Paninian Grammar (PG or CPG: Computational Paninian Grammar) to English, a fixed word order language. It presents how CPG can account for English and makes available, a linguistically rich resource in the form of an English Dependency Treebank. At present, 2000 sentences have been annotated as part of this effort, using the Hyderabad Dependency Treebank (HyDT) Annotation Scheme for Indian languages, (modelled on CPG). In the course of this paper we talk about CPG and the annotation scheme used for this work. We then talk about the task of annotation of the English language data per the scheme and how its application to English varies from Hindi. Further, we discuss our handling of some constructions of English, and some anomalies in the language that pose a challenge to the application of this annotation scheme to English, as is.

1 Introduction
The Paninian Grammar model, the oldest in the tradition of Dependency Grammar, treats a sentence as a series of modifier-modified relations. Thus, a sentence has a primary modified, (generally the main verb of the sentence) which is the root of the dependency tree, and its modifiers. The modifiers of the verb participate in the action specified by the verb and their relations with the verb are called karaka. ‘Karaka’ is the name given to the relation subsisting between a noun and a verb in a sentence. (Shastri, 1990) Panini recognises six types of karaka relations (Kiparsky and Stall , 1969): adhikaran ‘location’, apaadaan ‘source’, sampradaan ‘recipient’, karana ‘instrument’, karma ‘theme’, karta ‘agent’. It is important to note that karaka relations are not equivalent to theta roles (although they are mapped sometimes, for the sake of elucidation). While thematic roles are purely semantic in nature karaka relations are syntactico-semantic. (Begum et al., 2008)

In PG the role of case endings or markers such as post-positions and verbal inflections is emphasized. Positions or word order are considered only when necessary, since they contain only secondary information such as emphasis, in free word order languages. Inspired by an inflectionally rich language, PG is very well suited for languages that have a relatively free word order. (Bharti and Sangal, 1993).

However, its previous application to English by Bharati et. al. (1997) proved that PG “is more general than hitherto considered, and can be used to explain not just free word order languages but also positional languages.” Further, Begum et al. (2008) opine that a dependency framework is closer to semantics than phrase structure grammar (PSG) if the dependency relations are judiciously chosen. Given this, more and more research groups have been shifting to dependency analysis, of late.

Thus, extending the annotation scheme based on the CPG model to English, helps capture semantic information along with providing a syntactic analysis. The level of semantics they capture is reflected
in the surface form of the sentences, and is important syntactically. Such a level of annotation makes available a syntactico-semantic interface that can be easy to exploit computationally, for linguistic investigations and experimentation. This includes facilitating mappings between semantic arguments and syntactic dependents.

In their preliminary work, where they proposed a karaka based (CPG based) annotation scheme for English, Vaidya et al. (2009) establish that karaka relations capture some level of local semantics of a verb in a sentence, while also taking cues from the surface level morpho-syntactic information such as vibhakti.

Our work is a step ahead in the direction of application of the CPG model to English, taking forward the preliminary task carried out by (Vaidya et al., 2009). We have adopted the HyDT annotation scheme (Begum et al., 2008) and adapted it to suit the needs of the language wherever required. We discuss the task in its various aspects, in the later sections of the paper. However, it needs a mention here, that this effort culminates, among other things, in formalizing the CPG based annotation scheme for English, and preparation of the initial version of detailed guidelines to facilitate future efforts in English Annotation with CPG.

1.1 Annotation Scheme

Sanskrit grammarians like Panini and later (Tesnière and Fourquet, 1969), have used the idea of the verb being the core of the sentence, binding other elements as dependents by its valencies, in their grammars. The HyDT annotation scheme makes use of this concept of syntactic valencies (Herbst, 1999), (Bethke, 2003) and that the requirements of the verb for its arguments make the starting point of the dependency analysis. We make use mainly of the karaka relations (though some other relations are employed too, (Palmer et al., 2009)) to mark the relations between a verb and the arguments and adjuncts, in a sentence. The notion of karaka incorporates the local semantics of a verb in a sentence, while also taking cues from the surface level morpho-syntactic information. Karaka relations thus enjoy an intermediary status between a morpho-syntactic and semantic level. Thus, in the course of annotation, the arguments and the adjuncts in the sentence are annotated taking into consideration the verb meaning. (Bharati et al., 1995)

The HyDT annotation scheme used for annotation of Hindi and other Indian languages was formulated to facilitate creation of the CPG based Hyderabad Dependency Treebank (HyDT). It is a pilot treebank consisting of dependency annotations for 1865 Hindi sentences. The relations marked under this scheme are of two types - karaka and others (Bhatt et al., 2009). The relationships between various participants in an action (arguments) and the activity/state denoted by the verb are labeled with karaka relations. Relations other than karaka relations, such as purpose, reason, and possession (adjuncts) and also, special constructions like conjunct verbs, coordinating conjunctions and ellipses are taken care of, using the relational concepts prescribed by the annotation scheme by way of annotation labels and tags of various levels of granularity.

This annotation scheme marks inter-chunk relations instead of marking relations between words. A chunk (with boundaries marked) in HyDT, by definition, represents a set of adjacent words in a sentence, which are in dependency relation with each other, and where one of these words is their head. The relations among the words in a chunk are not annotated manually as part of this effort, since these are easily extracted automatically. A rule-based intra-chunk labeler whose accuracy is 96% is used for intra-chunk annotation. Ambati et al. (2010) showed that POS tag information itself is sufficient for marking intra-chunk labels with high accuracy (Ambati et al., 2010).

This paper is structured as follows: In section 2, we outline the related works that have employed CPG to English before. Section 3, describes our construction of the CPG based treebank for English. We broadly classify this section into 2 subsections, to talk about the data and tools used and discuss the current annotation effort. In section 4, we look at the handling of some English constructions under the CPG framework, and the extensions and adap-

Vibhakti is a Sanskrit grammatical term that encompasses post-positionals and case endings for nouns, as well as inflection and auxiliaries for verbs. (Pedersen et al., 2004)

The complete list of current tags can be found here: http://ltrc.iiit.ac.in/MachineTrans/research/tb/k1.pdf
tations that were made in the Annotation tagset. We also discuss the issues encountered during annotation, along with their resolution. In this section, we also see how the annotation of English departs from the annotation of Hindi, at times. Section 5, summarizes the paper and sketches some ideas for future work.

2 Treebanking Efforts and Related Works

“The need for treebanks as an empirical basis for research on the grammar of a language is well established...”(Nivre, 2005). Treebanks therefore, play an increasingly important role in NLP tasks such as parsing and generation etc. Other than this, they can be an indispensable resource for linguistic investigations. For example, they can be used in corpus linguistics to study syntactic phenomenon.

Beginning with the Penn treebank (PTB) (Marcus et al., 1993), treebank annotation has remained an important research area in CL and NLP. (Vaidya et al., 2009) Other notable efforts in treebanking include Prague Dependency Treebank (PDT) for Czech (Hajicova, 1998), and Hyderabad Dependency Treebank (HyDT) (Begum et al., 2008) etc. While PTB is an effort in Phrase structure annotation, PDT and HyDT use dependency based formalisms for annotation. There is increasing interest in the use of dependency parses in NLP applications, as they are considered to be simpler structures which can be computed more rapidly and are closer to the kinds of semantic representations that applications can make immediate use of. (Palmer et al., 2009)

Though originally formulated for Sanskrit, Paninian grammar has found an extensive application to many other languages with different word orders, such as Hindi (Begum et al., 2008), Arabic (Pedersen et al., 2004) and English. (Bharati et al., 1997), one of the first applications of CPG to English, shows that CPG is extensible even to fixed word order languages like English, resulting in an elegant computational grammar for English.

In their approach, Bharati et. al. (1995) generalize the notion of vibhakti to include position information, since English is a positional language. This includes the relative position of the constituent (eg. subject or object position) and the associated preposition. The generalization helps account for various phenomena and constructions of the language through the use of the PG notions of karaka chart, karaka chart transformation, and sharing rules (Bharati et al., 1995). They elaborate on their handling of long distance dependencies such as relative clauses, and also talk about constraint based parsing using CPG.

‘A Karaka Based Annotation Scheme for English’ is another step towards application of CPG to English by Vaidya et. al.(2009). It forms a preliminary work for the current research effort, which elaborates, builds upon, and carries their work forward. As their task was a preliminary one, a total of two annotators worked on 500 Part of Speech (POS) tagged sentences from the Wall Street Journal section of the Penn Treebank. They followed version 1.9 of the HyDT tagset\(^3\) for the annotation. In (Vaidya et al., 2009) they discuss about the tagset used and annotation of certain syntactic constructions of English such as yes-no questions, control verbs etc. using the scheme. They conclude with a comparison of their work with Penn Dependency Treebank and PropBank.

Section 3 discusses the current work in detail and talks about how it differs from the earlier works in this direction.

3 Our Work

3.1 Data and Tools Used

The corpus used for this annotation task consists of a total of 25000 words (around 2000 sentences) manually translated from the currently underway Hindi/Urdu Dependency Treebank Project (Bharati et al., 2009). The annotation is done using Sanchay\(^4\) annotation interface, in Shakti Standard Format (SSF) (Bharati et al., 2006).

3.2 The Task

We first POS tagged and Chunked\(^5\) the corpus automatically and also ran it through a morpholog-

---

\(^3\)http://ltrc.iiit.ac.in/MachineTrans/showfile.php?filename=publications/technicalReports/

\(^4\)Sanchay is an open source platform for working on languages, with components like a text editor with customizable support for languages and encodings, annotation interfaces, etc.

\(^5\)We used the POS and chunk tags from the Penn Treebank tagsets (Marcus et al, 2003) for this level of annotation.
cal analyzer to mark feature structure information at word level. The data was cleaned manually, for any discrepancies, at every stage. It was then manually annotated for dependency relations, using the CPG model based HyDT dependency annotation scheme.

Following the norms of the annotation scheme used, the dependency relations were marked at chunk level (between chunk heads) rather than within chunks, since the annotation emphasizes on depicting the modifier-modified relationships. In addition to the dependency relations, verbal nodes have also been marked with certain additional feature structure information. For instance, in order to handle cases with expletives such as ‘it’, (given a sentences such as ‘It is raining’) the information < stype = expletive__it > is added to the verbal node ‘is’. Since this work takes cues from the preliminary investigations by Vaidya et al. (2009) in this direction, to some extent the annotation carried out for this work falls in line with their work. However, as mentioned earlier, this takes their work further ahead, in more aspects than one.

We discuss handling of constructions and phenomenon of English (section 4.1) such as idioms and make adaptations in the annotation scheme to accommodate annotation of anomaly occurrences in the language (See section 4.2). Our discussion regarding handling certain constructions and some specific cases of English will shed further light on the way this CPG based annotation scheme works. Formalizing the CPG based annotation scheme for English, and preparation of the initial version of detailed guidelines to facilitate future efforts in English Annotation with CPG are other aspects of this work, not covered earlier.

For an insight into the annotation task, let us consider an example:

maa ne raama ko skuul bhejaa
mother ERG Ram ACC school send+past

“Mother sent Ram to school”

‘ne’ and ‘ko’ in the Hindi sentence above are vibhaktis (case markers) that help decide the karaka relations to be assigned to ‘maa’ and ‘raama’.

Given the same sentence in English, ‘Mother sent Ram to school’ we see that it has no surface level cues for vibhaktis (except for the preposition ‘to’). The ‘vibhakti’ information lies in the position of the arguments (‘Mother’ and ‘Rama’ here) in a positional language like English. (Bharati et al., 1997).

The corresponding karaka relations in this sentence are karta (k1) for Mother, karma (k2) for Rama, and adhikarana (k7) for school. While we can see from the example, that the surface level representation of vibhaktis differs in both the languages, the semantic relations captured between the verbs and their corresponding arguments (modifier and modified) are similar because of the similarity in the semantics of the verbs ‘sent’ and ‘bhejaa’.

The notion of vibhakti needs to be extrapolated (extended/generalized) to include the position of the arguments in the sentence, since there are not many surface level realizations in English that correspond to vibhaktis. Through generalizing this notion to accommodate the language facts of English it is possible to ignore the details of word order in the language, so that the assignment of karaka relations can be focused on exclusively (Bharati et al., 1997). Thus, “for English, the karaka relations will take the syntactic cues from the vibhaktis (in their generalized form) and from the semantic relationship between the modifier and modified components. The syntactico-semantic properties of the karaka relations will hereby be preserved.” (Bharati et al., 1997).

4 Constructions Handled

As mentioned earlier, the application of this annotation scheme to English varies from Hindi in certain aspects, given the dissimilarities between the two languages. In this section we take a look at these dissimilarities and at the ‘what and why’ of these dissimilarities that bring forth some annotation issues during the course of this work. We also look

---

6< stype = expletive__it > indicates that this is a sentence that contains expletive of the type ‘it’. The double underscore ‘__’ is read as ‘of the type’, and provides a more fine grained classification of the element to its immediate left. So, ‘stype=expletive__it’ would be read as, the sentence type is expletive ‘of the type ‘it’.

7karta–an argument that is most central to the action described by the verb.

8karma– the argument most desired by the karta.

9adhikarana–location of the action.
at their resolution, that was facilitated by an investigation into the reasons behind these differences and how English encodes information. “The major reason behind the structural differences between English and Hindi is the absence of accusative marker and yes–no marker in English. To compensate for this absence, English resorts to word order. This gives rise to more structural differences between the two languages.” (Bharti and Kulkarni, 2005).

This section reports our handling of some specific constructions of English in 4.1. We also bring out the difference in annotation of such constructions in Hindi, if any. In 4.2 we discuss our annotation of some anomaly occurrences in the language that pose a challenge to the application of CPG to English, as is.

4.1 Some English constructions

4.1.1 Reflexive Pronouns

Sometimes, a reflexive pronoun is used in an ap
tative way to indicate that the person who realizes the action of the verb is also the person who receives the action, whereas at other times reflexives are used to emphasize the subject, and are called emphatic pronouns e.g. himself, itself etc. Our dependency analysis of the above two cases would differ per their role in the sentence, i.e. according to the relation the pronoun has with the other entities in the sentence. In the case of normally occurring reflexive pronouns such as the one in the example: *I see myself in the mirror.* The reflexive pronoun ‘myself’ in this sentence will be labeled ‘k2’ of the verb ‘see’, since it is the ‘karma’ of the verb. Whereas in case of emphatic pronouns the pronoun isn’t the karma of the verb but a modifier of the noun that it goes back to. Thus to handle emphatic pronouns we use the dependency label ‘nmod__emph’, that makes its role in the sentence lucid. The label ‘nmod__emph’ stands for ‘a noun modifier (nmod) of the type ‘emph’ (emphatic pronoun/marker)’. An example sentence for emphatic pronouns would be: *The story had come out in the report of the commission itself.*

As seen in Figure 1, the emphatic marker ‘itself’ is annotated ‘nmod__emph’ of the noun ‘commission’, within the PP ‘of the commission’.

![Figure 1: Example showing emphatic markers.](image)

4.1.2 Idioms

The fact that idioms do not always follow the normal rules of meaning and grammar such as inflection for number or tense etc., and have a grammatical behavior of their own, poses quite a challenge to their handling in dependency analysis given the inconsistencies in their behavior. There are idioms such as ‘Hold your horses’ that accept changes such as inflection, per normal grammar rules, as demanded by the context. For example:

“*Hold your horses,*” I said to my friend.

They were asked to *hold their horses* until further orders from the officer.

Such idioms are termed transparent. Whereas some idioms have tightly bound frozen structures and lose their meaning until they are taken as a single unit, though they consist of multiple words. Also, they don’t generally take any inflection. Such idioms are referred to, as completely opaque/rigid grammatically. It is vital to mention here however, that there can be no clear demarcation as black and white for idioms in terms of their transparency or opacity, since they exist in all possible shades of grey. To facilitate annotation of idiomatic expressions per the HyDT Scheme, to an extent we take decisions regarding their annotation based on their rigidity. For example:

*He had to agree in the face of her strong determination.*

Though ‘in the face of’ appears as an idiom in English dictionaries, in order to annotate this sentence per the HyDT Scheme, we chunk the clause ‘in the face of her determination’ as ‘in the face’ and ‘of her strong determination’ (see Figure 2). This helps retain the HyDT framework while annotating this type of sentences. Further, we mark the
relation between these two chunks as ‘pof-idiom’ so as to retain the information that ‘in the face of’ is an idiom, and is followed by an NP.

Another example following this analysis is: **Pointing the finger at BJP, Vaidya said that those who have termed her resignation as an act, should first resign themselves.** Here, though ‘Pointing the finger at’ is an idiom, we annotate ‘the finger’ as ‘k2’ of ‘pointing’. Retaining the grammar in such cases is possible since these constructions are pletely frozen grammatically. We therefor such constructions more at syntactic level, their grammar. However, in some constrit English this may not be possible, if attain the grammar cause loss of the sense of or the sentence. e.g. According to Shyari trade had gone to the wall.

**Figure 2: Annotation of Idioms-1.**

4.1.3 Expletive Subjects

Expletives are semantically vacuous words meant to fill the syntactic vacancy in a sentence. ‘it’ and ‘there’ are the two commonest expletives of English. Though Expletives are subjects syntactically, they cannot be karta in a sentence since they are semantically vacuous, and karta, though syntactically grounded, entails some semantics too. Since expletives are not found in Hindi, a new label ‘dummy-sub’ was formalized to facilitate annotation of the expletives of English. E.g. **There was no option but to dissolve the house.**

**Figure 4: Example showing there as expletive subject.**

As seen in Figure 4, ‘There’ is a semantically vacuous element that serves to fill the empty subject position in the sentence. The fact that ‘here’ is not the locus of the action in the sentence substantiates that though it fills the subject position, it fails to function as ‘karta’. Therefore, we mark it with a special relation ‘dummy-sub’, which reflects the fact that ‘There’ is a dummy element in the sentence. Further, we also add the information about the expletive construction to the feature structure of the verbal node ‘was’ that the expletive is related to. `<sttype=expletive__there>` indicates that this is a sentence that contains expletive of the type ‘there’. The double underscore ‘__’ is read as ‘of the type’, and provides a more fine grained classification of the element to its immediate left. So, ‘stype=expletive__there’ would be read as, sentence type ‘of the type’ expletive ‘of the type’ there.
4.2 Annotating some Anomaly occurrences in the Lexicon:

4.2.1 ‘Given’

Though ‘given’ occurs, most of the time, as a Past Participial form of the verb ‘give’ or as an adjectival predicate of a copula, in English, it also is a preposition in some cases. When it appears as a past participial it is chunked in the NP, with the proceeding noun, per our annotation scheme, and it needn’t be assigned a separate annotation label. eg. I have answered the given question.

Whereas in the case of its occurrence as an adjectival predicate with a copula, it is labeled ‘k*s’ (samanadhikarana\textsuperscript{10} of the corresponding ‘karaka’ label), depending on the ‘k*’ label (such as \(k1, k2\)) of the corresponding noun preceding the copula. eg. She is given to making snide remarks. In this sentence ‘She’ is ‘k1’ (karta), so ‘given’ will be labeled ‘k1*s’ (karta samanadhikarana ‘noun complement of karta’), per our scheme.

In cases where ‘given’ is a preposition, it is chunked in the NP or VP that immediately follows it. In these cases it doesn’t take arguments, and thus has no direct children either. If we say ‘given something’ we mean ‘taking that thing into account’ or ‘because of that thing’. An example of ‘given’ as a preposition would be:

\textit{It is very easy for a prisoner to go insane given the conditions inside the prisons.}

In the above sentence, ‘given’ is chunked with the Noun Phrase (NP) ‘the conditions’ to form a Preposition Phrase (PP) ‘given the conditions’. This PP is labeled as ‘rh’\textsuperscript{11} of ‘to go’ since the clause ‘given the conditions inside the prisons’ expresses the reason for the prisoners to go insane. The tree rendering of the sentence in 5, makes the relations in the sentence more lucid.

Another example where ‘given’ appears as a preposition is: ‘Given the uncertainty over his future, I was left with little other choice.’

4.2.2 ‘be’ verbs

‘Be’ verbs (is, am, are, was, were) in English, are used as a copula verbs more often than not, as they serve to link the subject of a sentence with a predicate (a subject compliment or an adverbial). e.g. ‘I was confused.’ ‘Was’ here, is one such case. Per CPG in such cases where the copula verb is either a subject compliment or an object compliment, it has the relation samanadhikarna with either the ‘karta’ (in case of a subject compliment) or the ‘karma’ (in case of a object compliment). We use the HyDT labels ‘k1s’ (karta samanadhikarana ‘noun complement of karta’) and ‘k2s’ (karma samanadhikarana ‘object complement’) respectively, to annotate these.

However, in some cases, ‘was’ is found not to be a copula verb, though the resemblance is marked. The second case calls for a different dependency analysis of ‘was’ from its regular analysis as a copular verb. For example: Raw’s counter intelligence and security division was unaware of this fact for a long time.

As seen in Figure 6, the verb ‘was’ in the example above, enables the use of the adjective ‘unaware’ as a predicate. Thus ‘was’ is not a copula verb here. In fact, ‘was unaware’ is a complex predicate and can be substituted with ‘did not know’. Therefore, unlike the cases where ‘was’ is a copula, here ‘unaware’ is labeled ‘pof’ of ‘was’ to indicate that it is a part of the complex predicate.

\textsuperscript{10}The term ‘samanadhikarana’ indicates ‘having the same locus’

\textsuperscript{11}Since the term ‘samanadhikarana’ indicates ‘having the same locus’, ‘karta samanadhikarana’ indicates having the same locus as ‘karta’ and ‘karma samanadhikarana’ indicates having the same locus as ‘karma’
4.2.3 ‘But’

In English most of the times ‘but’ occurs as a Coordinator or a Conjunction, and is annotated with the HyDT dependency label ‘CCP’. eg. The committees were formed but their recommendations were never implemented.

However, in some constructions ‘but’ occurs as a preposition. In these cases we chunk it with the NP or VP following it, and annotation is done accordingly. eg. He said there was no option but to dissolve the house.

In the clause ‘but to dissolve the house,’ in the sentence above, ‘but’ is a preposition. Therefore we chunk it with ‘to dissolve’. (see Figure 8) This chunk has the dependency relation ‘vmod’ with the main verb ‘was’.

Some other examples where ‘but’ is a preposition:

- No one replied but me.
- He ate everything but the meat.
- Any day but tomorrow would be fine.

The preposition kind of ‘but’ is not only an anomaly case in English, its handling during annotation also departs from the annotation of its equivalent of Hindi, ‘lekin’ which is invariably a Coordinator or a Conjunction and is annotated with the dependency label ‘CCP’

5 Conclusions and Future Work

In this paper we have discussed an application of the CPG to English and reported how it can account for English, a fixed word order language. We talked of the annotation scheme and discussed our annotation effort. We also looked at the handling of some English constructions and the extensions and adaptations that were made in the Annotation tagset. Further, we also discussed the issues encountered during annotation, along with their resolution. We thus report our making available a linguistically rich resource by way of an English Treebank that can help reach global semantic relations since it has previously been presented that syntactic parsing facilitates identification of semantic relations (Gildea and Palmer, 2002). Our effort can thus benefit tasks like automatic semantic role labeling. Automatic conversion of existing English treebanks into this scheme and vice-versa is another area that can be explored given our work, since “treebanks available in one formalism can be converted into another either automatically or with minimal human effort by exploiting the similarities and differences between the two.” (Mannem and Chaudhry, 2009). The initial version of the detailed annotation guidelines prepared as part of this work are another resource that will facilitate future efforts in the annotation of English data with CPG or in experiments with its application to other languages. Additionally, this English Treebank, being parallel to the ongoing Hindi-Urdu
Treebank can prove to be of immense help in experimentation on parallel Treebanks.

References