Animacy Annotation in the Hindi Treebank

Itisree Jena, Riyaz Ahmad Bhat, Sambhav Jain and Dipti Misra Sharma
Language Technologies Research Centre, IIIT-Hyderabad, India
{itisree|riyaz.bhat|sambhav.jain}@research.iiit.ac.in, dipti@iiit.ac.in

Abstract

In this paper, we discuss our efforts to annotate nominals in the Hindi Treebank with the semantic property of animacy. Although the treebank already encodes lexical information at a number of levels such as morph and part of speech, the addition of animacy information seems promising given its relevance to varied linguistic phenomena. The suggestion is based on the theoretical and computational analysis of the property of animacy in the context of anaphora resolution, syntactic parsing, verb classification and argument differentiation.

1 Introduction

Animacy can either be viewed as a biological property or a grammatical category of nouns. In a strictly biological sense, all living entities are animate, while all other entities are seen as inanimate. However, in its linguistic sense, the term is synonymous with a referent's ability to act or instigate events volitionally (Kittilä et al., 2011). Although seemingly different, linguistic animacy can be implied from biological animacy. In linguistics, the manifestation of animacy and its relevance to linguistic phenomena have been studied quite extensively. Animacy has been shown, cross linguistically, to control a number of linguistic phenomena. Case marking, argument realization, topicality or discourse salience are some phenomena, highly correlated with the property of animacy (Aissen, 2003). In linguistic theory, however, animacy is not seen as a dichotomous variable, rather a range capturing finer distinctions of linguistic relevance. Animacy hierarchy proposed in Silverstein's influential article on "animacy hierarchy" (Silverstein, 1986) ranks nominals on a scale of the following gradience: *1st pers* > *2nd pers* > *3rd anim* > *3rd inanim*. Several such hierarchies of animacy have been proposed following (Silverstein, 1986), one basic scale taken from (Aissen, 2003) makes a three-way distinction as *humans* > *animates* > *inanimates*. These hierarchies can be said to be based on the likelihood of a referent of a nominal to act as an agent in an event (Kittilä et al., 2011). Thus higher a nominal on these hierarchies higher the degree of agency/control it has over an action. In morphologically rich languages, the degree of control/agency is expressed by case marking. Case markers capture the degree of control a nominal has in a given context (Hopper and Thompson, 1980; Butt, 2006). They rank nominals on the continuum of control as shown in (1)[1]. Nominals marked with Ergative case have highest control and the ones marked with Locative have lowest.

$$Erg > Gen > Inst > Dat > Acc > Loc \quad (1)$$

Of late the systematic correspondences between animacy and linguistic phenomena have been explored for various NLP applications. It has been noted that animacy provides important information, to mention a few, for anaphora resolution (Evans and Orasan, 2000), argument disambiguation (Dell'Orletta et al., 2005), syntactic parsing (Øvrelid and Nivre, 2007; Bharati et al., 2008; Ambati et al., 2009) and verb classification (Merlo and Steven-

[1]Ergative, Genitive, Instrumental, Dative, Accusative and Locative in the given order.

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 159–167,
Sofia, Bulgaria, August 8-9, 2013.

son, 2001). Despite the fact that animacy could play an important role in NLP applications, its annotation, however, is not usually featured in a treebank or any other annotated corpora used for developing these applications. There are a very few annotation projects that have included animacy in their annotation manual, following its strong theoretical and computational implications. One such work, motivated by the theoretical significance of the property of animacy, is (Zaenen et al., 2004). They make use of a coding scheme drafted for a paraphrase project (Bresnan et al., 2002) and present an explicit annotation scheme for animacy in English. The annotation scheme assumes a three-way distinction, distinguishing Human, Other animates and Inanimates. Among the latter two categories '$Other\ animates$' is further sub-categorized into Organizations and Animals, while the category of '$Inanimates$' further distinguishes between concrete and non-concrete, and time and place nominals. As per the annotation scheme, nominals are annotated according to the animacy of their referents in a given context. Another annotation work that includes animacy for nominals is (Teleman, 1974), however, the distinction made is binary between human and non-human referents of a nominal in a given context. In a recent work on animacy annotation, Thuilier et al. (2012) have annotated a multi-source French corpora with animacy and verb semantics, on the lines of (Zaenen et al., 2004). Apart from the manual annotation for animacy, lexical resources like wordnets are an important source of this information, if available. These resources usually cover animacy, though indirectly (Fellbaum, 2010; Narayan et al., 2002). Although a wordnet is an easily accessible resource for animacy information, there are some limitations on its use, as discussed below:

1. *Coverage*: Hindi wordnet only treats common nouns while proper nouns are excluded (except famous names) see Table 1. The problem is severe where the domain of text includes more proper than common nouns, which is the case with the Hindi Treebank as it is annotated on newspaper articles.

2. *Ambiguity*: Since words can be ambiguous, the animacy listed in wordnet can only be used in presence of a high performance word sense disambiguation system. As shown in Table 2, only 38.02% of nouns have a single sense as listed in Hindi Wordnet.

3. *Metonymy or Complex Types*: Domains like newspaper articles are filled with metonymic expressions like courts, institute names, country names etc, that can refer to a building, a geographical place or a group of people depending on the context of use. These words are not ambiguous per se but show different aspects of their semantics in different contexts (logically polysemous). Hindi wordnet treats these types of nouns as inanimate.

Nominals in HTB	*Hindi WordNet*	*Coverage*
78,136	65,064	83.27%

Table 1: Coverage of Hindi WordNet on HTB Nominals.

HTB Nominals with WN Semantics	*Single Unique Sense in Hindi WordNet*
65,064	24,741 (38.02%)

Table 2: Nominals in HTB with multiple senses

Given these drawbacks, we have included animacy information manually in the annotation of the Hindi Treebank, as discussed in this work. In the rest, we will discuss the annotation of nominal expressions with animacy and the motivation for the same, the discussion will follow as: Section 2 gives a brief overview of the Hindi Treebank with all its layers. Section 3 motivates the annotation of nominals with animacy, followed by the annotation efforts and issues encountered in Section 4. Section 5 concludes the paper with a discussion on possible future directions.

2 Description of the Hindi Treebank

In the following, we give an overview of the Hindi Treebank (HTB), focusing mainly on its dependency layer. The Hindi-Urdu Treebank (Palmer et al., 2009; Bhatt et al., 2009) is a multi-layered and multi-representational treebank. It includes three levels of annotation, namely two syntactic levels and one lexical-semantic level. One syntactic level is a dependency layer which follows the CPG (Begum

et al., 2008), inspired by the Pāṇinian grammatical theory of Sanskrit. The other level is annotated with phrase structure inspired by the Chomskyan approach to syntax (Chomsky, 1981) and follows a binary branching representation. The third layer of annotation, a purely lexical semantic one, encodes the semantic relations following the English PropBank (Palmer et al., 2005).

In the dependency annotation, relations are mainly verb-centric. The relation that holds between a verb and its arguments is called a $kaṛaka$ relation. Besides $kaṛaka$ relations, dependency relations also exist between nouns (genitives), between nouns and their modifiers (adjectival modification, relativization), between verbs and their modifiers (adverbial modification including subordination). CPG provides an essentially syntactico-semantic dependency annotation, incorporating $kaṛaka$ (e.g., agent, theme, etc.), non-$kaṛaka$ (e.g. possession, purpose) and other (part of) relations. A complete tag set of dependency relations based on CPG can be found in (Bharati et al., 2009), the ones starting with 'k' are largely Pāṇinian $kaṛaka$ relations, and are assigned to the arguments of a verb. Figure 1 encodes the dependency structure of (5), the preterminal node is a part of speech of a lexical item (e.g. NN,VM, PSP). The lexical items with their part of speech tags are further grouped into constituents called chunks (e.g. NP, VGF) as part of the sentence analysis. The dependencies are attached at the chunk level, marked with '$drel$' in the SSF format. k1 is the agent of an action (खाया 'eat'), whereas k2 is the object or patient.

(5) *संध्या ने सेब खाया ।*
Sandhya-*Erg* apple-*Nom* eat-*Perf*
'Sandhya ate an apple.'

<Sentence id = "1">

Offset	Token	Tag	Feature structure
1	((	NP	<fs name='NP' drel='k1:VGF'>
1.1	संध्या	NNP	<fs af='संध्या,n,f,sg,3,o,0,0'>
1.2	ने	PSP	<fs af='ने,psp,,,,,'>
	))		
2	((	NP	<fs name='NP2' drel='k2:VGF'>
2.1	सेब	NN	<fs af='सेब,n,m,sg,3,d,0,0'>
	))		
3	((	VGF	<fs name='VGF'>
3.1	खाया	VM	<fs af='खा,v,m,sg,any,,yA,yA'>
	))		

</Sentence>

Figure 1: Annotation of an Example Sentence in SSF.

Despite the fact that the Hindi Treebank already features a number of layers as discussed above, there have been different proposals to enrich it further. Hautli et al. (2012) proposed an additional layer to the treebank, for the deep analysis of the language, by incorporating the $functional\ structure$ (or f-structure) of Lexical Functional Grammar which encodes traditional syntactic notions such as subject, object, complement and adjunct. Dakwale et al. (2012) have also extended the treebank with anaphoric relations, with a motive to develop a data driven anaphora resolution system for Hindi. Given this scenario, our effort is to enrich the treebank with the animacy annotation. In the following sections, we will discuss in detail, the annotation of the animacy property of nominals in the treebank and the motive for the same.

3 Motivation: In the Context of Dependency Parsing

Hindi is a morphologically rich language, grammatical relations are depicted by its morphology via case clitics. Hindi has a morphologically split-ergative case marking system (Mahajan, 1990; Dixon, 1994). Case marking is dependent on the aspect of a verb (progressive/perfective), transitivity (transitive/intransitive) and the type of a nominal (definite/indefinite, animate/inanimate). Given this peculiar behavior of case marking in Hindi, arguments of a verb (e.g. transitive) have a number of possible configurations with respect to the case marking as shown in the statistics drawn from the Hindi Treebank released for MTPIL Hindi Dependency parsing shared task (Sharma et al., 2012) in Table 3. Almost in 15% of the transitive clauses, there is no morphological case marker on any of the arguments of a verb which, in the context of data driven parsing, means lack of an explicit cue for machine learning. Although, in other cases there is a case marker, at least on one argument of a verb, the ambiguity in case markers (one-to-many mapping between case markers and grammatical functions as presented in Table 4) further worsens the situation (however, see Ambati et al. (2010) and Bhat et al. (2012) for the impact of case markers on parsing Hindi/Urdu). Consider the examples from

(6a-e), the instrumental *se* is extremely ambiguous. It can mark the instrumental adjuncts as in (6a), source expressions as in (6b), material as in (6c), comitatives as in (6d), and causes as in (6e).

	K2-Unmarked	*K2-Marked*
K1-Unmarked	1276	741
K1-Marked	5373	966

Table 3: Co-occurrence of Marked and Unmarked verb arguments (core) in HTB.

	ने/ne (Ergative)	को/ko (Dative)	से/se (Instrumental)	में/meN (Locative)	पर/par (Locative)	का/kaa (Genitive)
k1(agent)	7222	575	21	11	3	612
k2(patient)	0	3448	451	8	24	39
k3(instrument)	0	0	347	0	0	1
k4(recipient)	0	1851	351	0	1	4
k4a(experiencer)	0	420	8	0	0	2
k5(source)	0	2	1176	12	1	0
k7(location)	0	1140	308	8707	3116	19
r6(possession)	0	3	1	0	0	2251

Table 4 : Distribution of case markers across case function.

(6a) *मोहन ने चाबी से ताला खोला ।*
Mohan-*Erg* key-*Inst* lock-*Nom* open
'Mohan opened the lock with a key.'

(6b) *गीता ने दिल्ली से सामान मंगवाया ।*
Geeta-*Erg* Delhi-*Inst* luggage-*Nom* procure
'Geeta procured the luggage from Delhi.'

(6c) *मूर्तिकार ने पत्थर से मूर्ति बनायी ।*
sculptor-*Erg* stone-*Inst* idol-*Nom* make
'The sculptor made an idol out of stone.'

(6d) *राम की श्याम से बात हुई ।*
Ram-*Gen* Shyaam-*Inst* talk-*Nom* happen
'Ram spoke to Shyaam.'

(6e) *बारिश से कई फसलें तबाह हो गयीं ।*
rain-*Inst* many crops-*Nom* destroy happen-*Perf*
'Many crops were destroyed due to the rain.'

(7) *चिड़िया दाना चुग रही है ।*
bird-*Nom* grain-*Nom* devour-*Prog*
'A bird is devouring grain.'

A conventional parser has no cue for the disambiguation of instrumental case marker *se* in examples (6a-e) and similarly, in example (7), it's hard for the parser to know whether 'bird' or 'grain' is the agent of the action 'devour'. Traditionally, syntactic parsing has largely been limited to the use of only a few lexical features. Features like POS-tags are way too coarser to provide deep information valuable for syntactic parsing while on the other hand lexical items often suffer from lexical ambiguity or out of vocabulary problem. So in oder to assist the parser for better judgments, we need to complement the morphology somehow. A careful observation easily states that a simple world knowledge about the nature (e.g. living-nonliving, artifact, place) of the participants is enough to disambiguate. For Swedish, Øvrelid and Nivre (2007) and Øvrelid (2009) have shown improvement, with animacy information, in differentiation of core arguments of a verb in dependency parsing. Similarly for Hindi, Bharati et al. (2008) and Ambati et al. (2009) have shown that even when the training data is small simple animacy information can boost dependency parsing accuracies, particularly handling the differentiation of core arguments. In Table 5, we show the distribution of animacy with respect to case markers and dependency relations in the annotated portion of the Hindi Treebank. The high rate of co-occurrence between animacy and dependency relations makes a clear statement about the role animacy can play in parsing. Nominals marked with dependency relations as k1 '$agent$', k4 '$recipient$', k4a '$experiencer$' are largely annotated as $human$ while k3 '$instrument$' is marked as $inanimate$, which confirms our conjecture that with animacy information a parser can reliably predict linguistic patterns. Apart from parsing, animacy has been reported to be beneficial for a number of natural language applications (Evans and Orasan, 2000; Merlo and Stevenson, 2001). Following these computational implications of animacy, we started encoded this property of nominals explicitly in our treebank. In the next section, we will present these efforts fol-

lowed by the inter-annotator agreement studies.

		Human	Other-Animates	Inanimate
k1	ने/ne (Erg)	2321	630	108
	को/ko (Dat/Acc)	172	8	135
	से/se (Inst)	6	0	14
	में/me (Loc)	0	0	7
	पर/par (Loc)	0	0	1
	का/kaa (Gen)	135	2	99
	ϕ (Nom)	1052	5	3072
k2	ने/ne (Erg)	0	0	0
	को/ko (Dat/Acc)	625	200	226
	से/se (Inst)	67	0	88
	में/me (Loc)	2	0	6
	पर/par (Loc)	5	0	37
	का/kaa (Gen)	15	0	14
	ϕ (Nom)	107	61	2998
k3	ने/ne (Erg)	0	0	0
	को/ko (Dat/Acc)	0	0	0
	से/se (Inst)	2	0	199
	में/me (Loc)	0	0	0
	पर/par (Loc)	0	0	0
	का/kaa (Gen)	0	0	0
	ϕ (Nom)	0	0	20
k4	ने/ne (Erg)	0	0	0
	को/ko (Dat/Acc)	597	0	13
	से/se (Inst)	53	0	56
	में/me (Loc)	0	0	0
	पर/par (Loc)	0	0	0
	का/kaa (Gen)	0	0	0
	ϕ (Nom)	7	0	8
k4a	ने/ne (Erg)	0	0	0
	को/ko (Dat/Acc)	132	0	8
	से/se (Inst)	4	0	2
	में/me (Loc)	0	0	0
	पर/par (Loc)	0	0	0
	का/kaa (Gen)	1	0	0
	ϕ (Nom)	56	0	1
k5	ने/ne (Erg)	0	0	0
	को/ko (Dat/Acc)	0	0	0
	से/se (Inst)	7	0	460
	में/me (Loc)	0	0	1
	पर/par (Loc)	0	0	0
	का/kaa (Gen)	0	0	0
	ϕ (Nom)	0	0	2
k7	ने/ne (Erg)	0	0	0
	को/ko (Dat/Acc)	4	0	0
	से/se (Inst)	3	0	129
	में/me (Loc)	0	1977	1563
	पर/par (Loc)	66	0	1083
	का/kaa (Gen)	0	0	8
	ϕ (Nom)	5	0	1775
r6	ने/ne (Erg)	0	0	0
	को/ko (Dat/Acc)	0	0	0
	से/se (Inst)	1	0	0
	में/me (Loc)	0	0	0
	पर/par (Loc)	0	0	0
	का/kaa (Gen)	156	80	605
	ϕ (Nom)	13	3	25

Table 5: Distribution of semantic features with respect to case markers and dependency relations [a].

[a] k1 'agent', k2 'patient', k3 'instrument', k4 'recipient', k4a 'experiencer', k5 'source', k7 'location', r6 'possession'

4 Animacy Annotation

Following Zaenen et al. (2004), we make a three-way distinction, distinguishing between $Human$, $Other\ Animate$ and $In\text{-}animate$ referents of a nominal in a given context. The animacy of a referent is decided based on its sentience and/or control/volitionality in a particular context. Since, prototypically, agents tend to be animate and patients tend to be inanimate (Comrie, 1989), higher animates such as humans, dogs etc. are annotated as such in all contexts since they frequently tend to be seen in contexts of high control. However, lower animates such as insects, plants etc. are annotated as '$In\text{-}animate$' because they are ascribed less or no control in human languages like inanimates (Kittilä et al., 2011). Non-sentient referents, except intelligent machines and vehicles, are annotated as '$In\text{-}animate$' in all contexts. Intelligent machines like robots and vehicles, although, lack any sentience, they possess an animal like behavior which separates them from inanimate nouns with no animal resemblance, reflected in human language as control/volitionality. These nouns unlike humans and other higher animates are annotated as per the context they are used in. They are annotated as '$Other\ animate$' only in their agentive roles. Nominals that vary in sentience in varying contexts are annotated based on their reference in a given context as discussed in Subsection 4.2. These nominals include country names referring to geographical places, teams playing for the country, governments or their inhabitants; and organizations including courts, colleges, schools, banks etc. Unlike Zaenen et al. (2004) we don't further categorize '$Other\ Animate$' and '$In\text{-}animate$' classes. We

don't distinguish between *Organizations* and *Animals* in '$Other\ \ Animate$' and *Time* and *Place* in '$In\text{-}animates$'.

The process of animacy annotation in the Hindi Treebank is straight forward. For every chunk in a sentence, the animacy of its head word is captured in an '$attribute\text{-}value$' pair in SSF format, as shown in Figure 3. Hitherto, around 6485 sentence, of the Hindi Treebank, have been annotated with the animacy information.

<Sentence id = "1">

Offset	Token	Tag	Feature structure
1	((	NP	<fs name='NP' drel='k1:VGF' **semprop='human'**>
1.1	मोहन	NNP	<fs af='मोहन,n,m,sg,3,d,0,0'>
1.2	ने	PSP	<fs af='ने,psp,,,,,,' name='ने'>
	))		
2	((	NP	<fs name='NP2' drel='k4:VGF' **semprop='other-animate'**>
2.1	बिल्ली	NN	<fs af='बिल्ली,n,f,sg,3,d,0,0'>
2.2	को	PSP	<fs af='को,psp,,,,,,' name='को'>
	))		
3	((	NP	<fs name='NP3' drel='k3:VGF' **semprop='inanimate'**>
3.1	बोतल	NN	<fs af='बोतल ,n,f,sg,3,d,0,0'>
3.2	से	PSP	<fs af='से,psp,,,,,,'>
	))		
4	((	NP	<fs name='NP4' drel='k2:VGF' **semprop='inanimate'**>
4.1	दूध	NN	<fs af='दूध,n,m,sg,3,d,0,0'>
	))		
5	((	VGF	<fs name='VGF'>
5.1	पिलाया	VM	<fs af='पिला,v,m,sg,any,,yA,yA'>
	))		

</Sentence>

Figure 3: Semantic Annotation in SSF.

(8) *मोहन ने बिल्ली को बोतल से दूध*
Mohan-*Erg* cat-*Dat* bottle-*Inst* milk-*Nom*
पिलाया ।
drink-*Perf*
'Mohan fed milk to the cat with a bottle.'

In the following, we discuss some of the interesting cross linguistic phenomena which added some challenge to the annotation.

4.1 Personification

Personification is a type of meaning extension whereby an entity (usually non-human) is given human qualities. Personified expressions are annotated, in our annotation procedure, as $Human$, since it is the sense they carry in such contexts. However, to retain their literal sense, two attributes are added. One for their context bound sense (metaphorical) and the other for context free sense (literal). In example (9), $waves$ is annotated with literal animacy as $In\text{-}animante$ and metaphoric animacy as $Human$, as shown in Figure 4 (offset 2).

<Sentence id = "1">

Offset	Token	Tag	Feature structure
1	((	NP	<fs name='NP' drel='k7p:VGF' >
1.1	सागर	NNC	<fs af='सागर,n,m,sg,3,d,0,0'>
1.2	तट	NN	<fs af='तट,n,m,sg,3,d,0,0'>
1.3	पर	PSP	<fs af='पर,psp,,,,,,'>
	))		
2	((	NP	<fs name='NP2' drel='k1:VGF' **semprop='inanimate' metaphoric='human'**>
2.1	लहरें	NN	<fs af='लहरें,n,f,pl,3,d,0,0'>
	))		
3	((	VGF	<fs name='VGF'>
3.1	नाच	VM	<fs af='नाच,v,any,any,any,,0,0'>
3.2	रही	VAUX	<fs af='रही,v,f,sg,any,,ya,ya'>
3.3	हैं	AUX	<sf AF='हैं,v,any,pl,1,,he,he'>
	))		

</Sentence>

Figure 4: Semantic Annotation in SSF.

(9) *सागर तट पर लहरें नाच रही हैं ।*
sea coast-*Loc* waves-*Nom* dance-*Prog*
'Waves are dancing on the sea shore.'

4.2 Complex Types

The Hindi Treebank in largely built on newspaper corpus. Logically polysemous expressions (metonymies) such as $government$, $court$, $newspaper$ etc. are very frequent in news reporting. These polysemous nominals can exhibit contradictory semantics in different contexts. In example (10a), $court$ refers to a $person$ (judge) or a $group\ of\ persons$ (jury) while in (10b) it is a $building$ (see Pustejovsky (1996) for the semantics of complex types). In our annotation procedure, such expressions are annotated as per the sense or reference they carry in a given context. So, in case of (10a) $court$ will be annotated as $Human$ while in (10b) it will be annotated as $In\text{-}animante$.

(10a) *अदालत ने मुकदमे का फैंसला*
court-*Erg* case-*Gen* decision-*Nom*
सुनाया ।
declare-*Perf*
'The court declared its decision on the case.'

(10b) मैं अदालत में हूँ ।
I-*Nom* court-*Loc* be-*Prs*
'I am in the court.'

4.3 Inter-Annotator Agreement

We measured the inter-annotator agreement on a set of 358 nominals (∼50 sentences) using Cohen's kappa. We had three annotators annotating the same data set separately. The nominals were annotated in context i.e., the annotation was carried considering the role and reference of a nominal in a particular sentence. The kappa statistics, as presented in Table 6, show a significant understanding of annotators of the property of animacy. In Table 7, we report the confusion between the annotators on the three animacy categories. The confusion is high for '$Inanimate$' class. Annotators don't agree on this category because of its fuzziness. As discussed earlier, although '$Inanimate$' class enlists biologically inanimate entities, some entities may behave like animates in some contexts. They may be sentient and have high linguistic control in some contexts. The difficulty in deciphering the exact nature of the reference of these nominals, as observed, is the reason behind the confusion. The confusion is observed for nouns like organization names, lower animates and vehicles. Apart from the linguistically and contextually defined animacy, there was no confusion, as expected, in the understanding of biological animacy.

$Annotators$	κ
ann1-ann2	0.78
ann1-ann3	0.82
ann2-ann3	0.83
Average κ	0.811

Table 6: Kappa Statistics

	Human	*Other-animate*	*Inanimate*
Human	71	0	14
Other-animate	0	9	5
Inanimate	8	10	241

Table 7: Confusion Matrix

5 Conclusion and Future Work

In this work, we have presented our efforts to enrich the nominals in the Hindi Treebank with animacy information. The annotation was followed by the inter-annotator agreement study for evaluating the confusion over the categories chosen for annotation. The annotators have a significant understanding of the property of animacy as shown by the higher values of Kappa (κ). In future, we plan to continue the animacy annotation for the whole Hindi Treebank. We also plan to utilize the annotated data to build a data driven automatic animacy classifier (Øvrelid, 2006). From a linguistic perspective, an annotation of the type, as discussed in this paper, will also be of great interest for studying information dynamics and see how semantics interacts with syntax in Hindi.

6 Acknowledgments

The work reported in this paper is supported by the NSF grant (Award Number: CNS 0751202; CFDA Number: 47.070). [2]

References

Judith Aissen. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory*, 21(3):435–483.

B.R. Ambati, P. Gade, S. Husain, and GSK Chaitanya. 2009. Effect of minimal semantics on dependency parsing. In *Proceedings of the Student Research Workshop*.

B.R. Ambati, S. Husain, J. Nivre, and R. Sangal. 2010. On the role of morphosyntactic features in Hindi dependency parsing. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 94–102. Association for Computational Linguistics.

R. Begum, S. Husain, A. Dhwaj, D.M. Sharma, L. Bai, and R. Sangal. 2008. Dependency annotation scheme for Indian languages. In *Proceedings of IJCNLP*. Citeseer.

Akshar Bharati, Samar Husain, Bharat Ambati, Sambhav Jain, Dipti Sharma, and Rajeev Sangal. 2008. Two semantic features make all the difference in parsing accuracy. *Proceedings of ICON*, 8.

[2] 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.

A. Bharati, D.M. Sharma, S. Husain, L. Bai, R. Begum, and R. Sangal. 2009. AnnCorra: TreeBanks for Indian Languages Guidelines for Annotating Hindi TreeBank (version–2.0).

R.A. Bhat, S. Jain, and D.M. Sharma. 2012. Experiments on Dependency Parsing of Urdu. In *Proceedings of TLT11 2012 Lisbon Portugal*, pages 31–36. Ediçes Colibri.

R. Bhatt, B. Narasimhan, M. Palmer, O. Rambow, D.M. Sharma, and F. 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.

Joan Bresnan, Jean Carletta, Richard Crouch, Malvina Nissim, Mark Steedman, Tom Wasow, and Annie Zaenen. 2002. Paraphrase analysis for improved generation, link project.

Miriam Butt. 2006. The dative-ergative connection. *Empirical issues in syntax and semantics*, 6:69–92.

N. Chomsky. 1981. Lectures on Government and Binding. *Dordrecht: Foris*.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Praveen Dakwale, Himanshu Sharma, and Dipti M Sharma. 2012. Anaphora Annotation in Hindi Dependency TreeBank. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 391–400, Bali,Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2005. Climbing the path to grammar: A maximum entropy model of subject/object learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 72–81. Association for Computational Linguistics.

R.M.W. Dixon. 1994. *Ergativity*. Number 69. Cambridge University Press.

Richard Evans and Constantin Orasan. 2000. Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, pages 154–162.

Christiane Fellbaum. 2010. *WordNet*. Springer.

A. Hautli, S. Sulger, and M. Butt. 2012. Adding an annotation layer to the Hindi/Urdu treebank. *Linguistic Issues in Language Technology*, 7(1).

Paul J Hopper and Sandra A Thompson. 1980. Transitivity in grammar and discourse. *Language*, pages 251–299.

Seppo Kittilä, Katja Västi, and Jussi Ylikoski. 2011. *Case, Animacy and Semantic Roles*, volume 99. John Benjamins Publishing.

A.K. Mahajan. 1990. *The A/A-bar distinction and movement theory*. Ph.D. thesis, Massachusetts Institute of Technology.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet-a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India*.

Lilja Øvrelid and Joakim Nivre. 2007. When word order and part-of-speech tags are not enough – Swedish dependency parsing with rich linguistic features. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 447–451.

Lilja Øvrelid. 2006. Towards robust animacy classification using morphosyntactic distributional features. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 47–54. Association for Computational Linguistics.

Lilja Øvrelid. 2009. Empirical evaluations of animacy annotation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. volume 31, pages 71–106. MIT Press.

M. Palmer, R. Bhatt, B. Narasimhan, O. Rambow, D.M. Sharma, and F. Xia. 2009. Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

J. Pustejovsky. 1996. The Semantics of Complex Types. *Lingua*.

Dipti Misra Sharma, Prashanth Mannem, Joseph van-Genabith, Sobha Lalitha Devi, Radhika Mamidi, and Ranjani Parthasarathi, editors. 2012. *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*. The COLING 2012 Organizing Committee, Mumbai, India, December.

Michael Silverstein. 1986. Hierarchy of features and ergativity. *Features and projections*, pages 163–232.

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur.

Juliette Thuilier, Laurence Danlos, et al. 2012. Semantic annotation of French corpora: animacy and verb semantic classes. In *LREC 2012-The eighth international conference on Language Resources and Evaluation*.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M Catherine O'Connor, and Tom Wasow. 2004. Animacy Encoding in English: why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 118–125. Association for Computational Linguistics.