Relative Clauses in Hindi and Arabic: A Paninian Dependency Grammar Analysis

Mark Pedersen * †
† School of Information Technology & Electrical Engineering
University of Queensland
St. Lucia 4072, Australia
markp@itee.uq.edu.au

Domenyk Eades ‡
‡ Department of English,
Faculty of Arts,
Sultan Qaboos University,
Muscat, PC 123 Oman
domenyk@squ.edu.om

Samir K. Amin and Lakshmi Prakash ∗
∗Faculty of Applied Science,
Sohar University,
Sohar, PC 311 Oman
s.amin@soharuni.edu.om,
l.prakash@soharuni.edu.om

Abstract

We present a comparative analysis of relative clauses in Hindi and Arabic in the tradition of the Paninian Grammar Framework (Bharati et al., 1996b) which leads to deriving a common logical form for equivalent sentences. Parallels are drawn between the Hindi co-relative construction and resumptive pronouns in Arabic. The analysis arises from the development of lexicalised dependency grammars for Hindi and Arabic that have application for machine translation.

1 Introduction

(Kruijff, 2002) notes that Dependency Grammar has its roots in Panini's grammar of Sanskrit (350/250 BC) and also the work of early Arabic grammarians (Kitab al-Usul of Ibn al-Sarraj, d. 928). Among the recent activities in Dependency Grammar, (Bharati et al., 1996b) have established a computational approach to Indian languages which they call the Paninian Grammar Framework (PGF).

Bharati et. al. (ibid) suggest that PGF is extensible to other languages, including fixed word order languages such as English. In considering a Machine Translation system for Hindi-Arabic, and given the availability of a PGF-style parser for Hindi (Pedersen, 2001), we have sought to establish the suitability of PGF for Arabic.

In the following sections we will briefly describe the general PGF-inspired parsing framework, and then contrast the analysis of Hindi and Arabic relative clauses within this framework. In particular, we examine parallels between the Hindi co-relative construction and resumptive pronouns in Arabic, and demonstrate how a common logical interpretation can be given to syntactic variations of relative clauses in both languages.

2 Dependency Grammar in The Paninian Grammar Framework

2.1 Background

The modern formulation of Dependency Structure is frequently attributed to (Tesnière, 1959). It is interesting to note that among contemporary proponents of dependency structure, there are those, such as (Hudson, 1984), who maintain a general principle of projectivity for their dependency structures, and devise additional means of coping with discontinuity when it arises. Others, such as (Mel'čuk, 1988), (Bharati et al., 1996b) and (Covington, 1990) allow non-projective dependency structures and rely upon a separate means of linearisation. Most recently, (Debusmann and Duchier, 2003) have presented a new formulation of dependency grammar which generalises multi-stratal approaches to an n-dimensional formalism of interacting dependency graphs.

The Paninian Grammar Framework proposed by (Bharati et al., 1996b) is particularly aimed at treating heavily inflected free word order languages such as Hindi and other Indian languages. Like Hindi, Arabic is heavily inflected with overt case marking of nouns, noun-verb agreement, as well as incorporation of pronominals into verb forms. Although Arabic has a word order (VSO) that is more fixed compared to Hindi (canonically SOV but with significant word order freedom), there is also significant word order variation found in nominal and topicalised sentences, thus making alternate word orders such as SVO quite common.

2.2 The Paninian Grammar Framework

PGF has two levels of representation that mediate between an utterance and its meaning: the *vibhakti* level and the *karaka* level. Figure 1 shows the relationship between the levels of representation in Paninian Grammar.

Let us take the surface level of of Figure 1 to represent the level of tokenised and morpho-

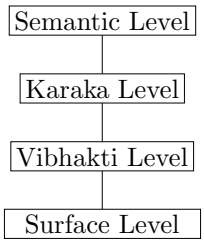


Figure 1: Levels of representation in PGF

logically analysed input to a syntactic-semantic parsing process. Vibhakti is a Sanskrit grammatical term that encompasses postpositionals and case endings for nouns, as well as inflection and auxiliaries for verbs. The vibhakti level groups words together according to explicit case endings and other inflectional markers. Vibhakti word groups are necessarily contiguous and typically have a fixed word order, but vibhakti groups can generally appear in any order without affecting the karaka relations which connect them.

The karaka level serves to relate the lexical elements of a sentence according to their syntactic functions, as derivable from their vibhakti. In terms of more familiar linguistic theory, we can generally equate the vibhakti level with morphotactics and the karaka level with syntactico-semantic functions (e.g. theta relations). The semantic level shown in figure 1 is indicative of a purely semantic representation or logical form, the details of which are beyond the scope of this paper.

Panini recognises six type of karaka relations (Kiparsky and Staal, 1969):

- *karta* (agent)
- *karma* (patient)
- *karan* (instrument)
- *sampradan* (recipient)
- *apaadaan* (point of departure, or cause)
- *adhikaran* (locality)

It is important to note that karaka relations differ from typical theta relations in that they tend to have more fluid definitions rather than fixed syntactico-semantic roles. For instance, the *karta* relation, which may generally be equated to the *agent* theta role, is more precisely defined as "the most independent of all participants" (Bharati et al., 1996b, p. 187). The *karta* relation can equally be applied to the subject of all of the following sentences, although *karta* should not be confused with the purely syntactic role of *subject*:

(1) (a) Mohan opened the lock.
(b) The key opened the lock.
(c) The lock opened.

Karaka relations under consideration here are *karta*, as described, and *karma*, the principle goal of the *karta* (roughly, *patient*). Karaka relations do not map exactly on to the typical semantic relation categories common in contemporary Western linguistic theory, largely because of the differing conceptions of semantic roles underpinning Paninian logical form[1].

A further difference between karaka relations and typical theta roles is that, traditionally, vibhakti is the primary means of identifying which karaka relations may apply[2]. This characteristic lends itself well to heavily inflected languages with free word order. A simple default karaka chart (adapted from (Bharati et al., 1996b)) showing the mapping between vibhakti and karaka is given in Table 1.

Karaka	*Vibhakti*	*Presence*
karta	NOM (∅) num & gen agr.	mandatory
karma	ACC (ko or ∅)	mandatory

Table 1: Default Karaka chart

PGF elegantly handles active-passive transformations and compound sentences through a system of karaka transformations, which change the default vibhakti required by a differently conjugated verb for a given karaka. If we were to use the passive form of *khaataa - khayaa* , the vibhakti required for the karta karaka becomes *ne*, as illustrated in (2).

[1]Paninian semantics makes use of the Indian systems of philosophy and logic. Among these, Navya Nyaya logic was of particular significance. For more detail on Navya Nyaya, see (Ingalls, 1951) and (Matilal, 1968).

[2]Panini's grammar of Sanskrit asserted that every karaka relation in an utterance must have a phonetic realisation (Bharati et al., 1996b, p. 187), rather than via an intermediate syntactic role or sentential position. In this sense, there is a tighter binding between karaka relations and the surface-level syntax than would normally be seen in a typical theta role analysis.

Despite its orientation toward inflected languages, some work has been done on generalising the notion of *vibhakti* to include linear position where this has syntactic significance (Bharati et al., 1996a). The generalisation of vibhakti to account for word order essentially introduces a facility analogous to topological fields (as found in Topological Dependency Grammar (Duchier and Debusmann, 2001) and DACHS (Bröker, 1998)) to PGF. It is important to note however that these word order constraints are treated in the same way as other vibhakti (morphological constraints), and do not use a separate representation. In this sense, PGF does not attempt to separate linear precedence from immediate dominance at a formal level.

In the following sections we explore how both morphology and word order constraints in Arabic can be captured through a mapping of vibhakti to karaka relations.

3 Relative Clauses in Hindi and Arabic

To briefly summarise salient features of Hindi and Arabic:

- Hindi word order is relatively free
- Arabic word order is more fixed, but topicalisation and nominal sentence forms effectively license significant permutations in word order
- both have complex morphology, but
- most case marking in Hindi uses both inflection and post positions
- whereas Arabic generally uses inflection only

3.1 Data

In comparing relative clauses, let us consider the data shown in (2) and (3)[3]. The sentences given are intended to represent the same semantic content, but give different emphases through topicalisation[4]. It should be noted that some utterances, such as 2(b), would only be used in specific context and don't represent the normal speech pattern. Nevertheless all are considered to be grammatical by native speakers and follow in the same pattern as data presented by (Dwivedi, 1994).

3.2 PGF analysis

The immediately observable difference between the Hindi and Arabic is that the Hindi data permits topicalisation primarily through changes in word order and a limited number of changes in vibhakti. The syntactic structure of 2(b) and (c) are isomorphous modulo word order. In 2(a), *machli* is internal to the relative clause, and hence *vo* is pronominal rather than demonstrative. This requires two changes at the vibhakti level: *khayi thi* agrees in gender with the explicitly present female object, rather than the male subject, because of the ergative construction[5]; and because of its position as a syntactic object, the ergative marking on *machli* is dropped.

In contrast, topicalisation in Arabic requires a variety of different syntactic constructions and associated changes in vibhakti, since the word order is not as free. In 3(a) we find the equivalent Arabic sentence to 2(a) preposes the subject of the main verb, which is a permissible variation on the standard VSO word order.

However in 3(b), the sentence must be reconstructed to allow *Zayd* to occupy a topic position. We can neither extract Zayd from the relative clause, nor have the relative clause as the topic, as it appears in 2(b). Therefore, *Zaydun ʾakala ʾal-samakah* becomes the main clause of the sentence.

For the final topicalisation, Arabic takes on the nominal sentence form, in which there is no main verb. Instead, *ʾal-ḍifdaʿ* acts as the predicate, to which nested relative clauses are attached. Analyses for the Hindi and Arabic samples are shown in 4(a)-(c) and 5(a)-(c) respectively.

For the purposes of this analysis, we bor-

[3] In 3(a) and (c), the resumptive pronouns *-hu* and *-ha* are glossed as *-3.M* and *-3.F*, meaning third person pronoun masculine and third person pronoun feminine, respectively. Throughout the examples, gender marking on verbs indicates agreement with the relevant argument.

[4] For the purposes of this paper we use the term topicalisation broadly to refer to the phenomenon of lexical "movement" to a sentence initial position without distinguishing between various types of such movement, such as those discussed by (Gambhir, 1981), (Dwivedi, 1994) and (Butt and Holloway-King, 1997).

[5] In the other example sentences, the agreement reverts to the default male gender, since the direct object has been extracted.

(2) (a) *jo* *machli* *Zayd ne* *khayi thi* *vo* *maindek ko* *khayaa tha*
REL fish.F Zayd-ERG eat-PAST.F CO-REL frog.M-ACC eat-PAST.M
The fish that was eaten by Zayd had eaten a frog.

(b) *jo* *Zayd ne* *khayaa tha* *vo* *machli ne* *maindek ko* *khayaa tha*
REL Zayd-ERG eat-PAST.M CO-REL fish.F-ERG frog.M-ACC eat-PAST.M
The fish *that was eaten by Zayd* had eaten a frog.

(c) *maindek ko* *vo* *machli ne* *khayaa tha* *jo* *Zayd ne* *khayaa tha*
frog.M-ACC CO-REL fish.F-ERG eat-PAST.M REL Zayd-ERG eat-PAST.M
The fish that was eaten by Zayd had eaten *a frog.*

(3) (a) *ʾal-samakah* *allaty* *ʾakala-hā* *Zayd-un* *ʾakalat* *ʾal-ḍifdaʿ*
DEF-fish.F REL.F ate.M-3.F Zayd-NOM ate.F DEF-frog.M
The fish that was eaten by Zayd, had eaten a frog.

(b) *Zayd-un* *ʾakala* *ʾal-samakah* *allaty* *ʾakalat* *ʾal-ḍifdaʿ*
Zayd-NOM ate.M DEF-fish.F REL.F ate.F DEF-frog.M
Zayd ate the fish that had eaten the frog.

(c) *ʾal-ḍifdaʿ* *ʾalladhy* *ʾakalat-hu* *ʾal-samakah* *allaty* *ʾakala-hā* *Zayd-un*
DEF-frog.M REL.M ate.F-3.M DEf-fish.F REL.F ate.M-3.F Zayd-NOM
The frog had been eaten by the fish that Zayd ate.

row the term *avachchedak* (limiter) from Navya Nyaya logic to express the relationship between the relative clause and the noun being modified, which is the *avachchinna* - "the limited". From a PGF perspective, the data is explained as follows.

3.3 Discussion

3.3.1 Hindi Analysis

In Hindi, ergative marking of the subject is required by the *yaa* form of the verb, as shown earlier, except in the case of pronominal *vo* which is not explicitly marked with *ne*. The subject for both clauses is thus clearly identified (either it is marked with *ne* or it is the pronominal *vo*). If word order were completely free, both the main verb and the complement would be candidate heads for the ergatively marked verbs. For 2(a) and 2(b), the co-relative *vo* introduces a projectivity constraint which removes the ambiguity. In 2(a), the pronominal co-relative *vo* takes the place of the object of the relative clause *machli*. In 2(b), the object of the relative clause is absent (the typical constituency analysis would say that it leaves a trace), but is marked by the demonstrative co-relative *vo* in the main clause, establishing the connection to the relativiser *jo*. Both of these represent the left-adjoined form of the relative clause.

The relationship between the relativiser *jo* and the co-relative *vo* is implicit, since the direct relationship between *jo* and the modified noun merely marks the noun, in the case of 2(a), as the *avachchinna* - the item limited by the relative clause. In practical terms, the presence of an argument marked as *avachchinna* is propagated up to the head of the relative clause, marking it as the *avachchedak* limiter, meaning that the verb cannot take anything other than the co-relative *vo* as its head. This analysis is in keeping with the requirement that *jo* must have a matching *vo* in the sentence.

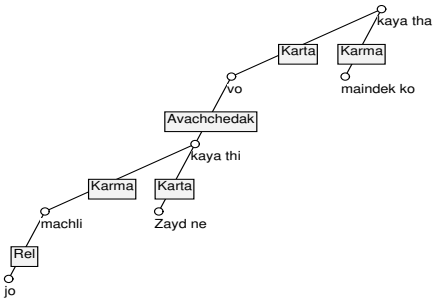
The same analysis applies when the modified noun is not explicitly present in the relative clause. In this case, *jo* still signifies the presence of an *avachchinna* argument, but it is pronominal, and the property of being *avachchinna* is conferred upon the modified noun by *vo*. In this way, there is a pleasing symmetry between *vo* and *jo* in that when they depend on a noun, they act as demonstratives, marking the noun as *avachchinna*, and when appearing independently, they act as pronouns in place of the *avachchinna* noun.

In 2(c), we have the right adjoined form of the relative clause. Additionally, the typical SOV word order of the main clause has been altered to OSV through topicalising the object, *maindek*. The co-relative *vo* continues to assert a projectivity constraint on the relative clause.

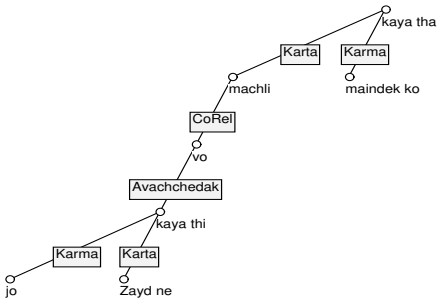
A further constraint on right-adjoined relative clauses is that they may not contain the explicit subject. This constraint is enforced via the karaka relation between the *avachchedak*-marked verb and the pronominal co-relative *vo*[6]

[6]The demonstrative co-relative *vo* marks an explicit

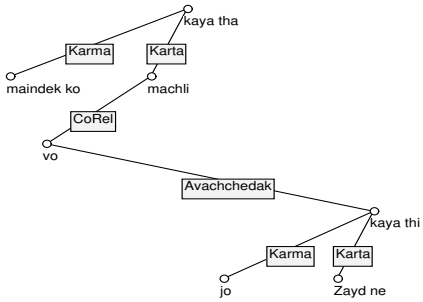
(4) (a)



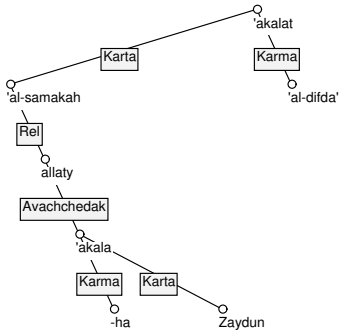
(b)



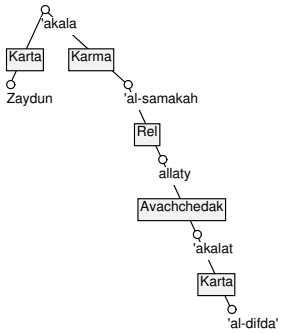
(c)



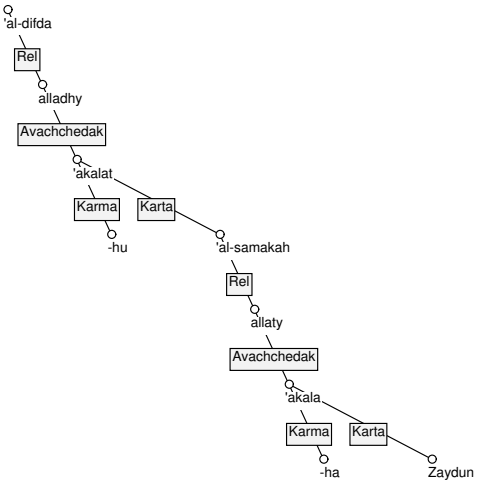
(5) (a)



(b)



(c)



requiring *jo* to always precede *vo*[7], thus ensuring that the pronominal *vo* never precedes its referent.

noun, while the pronominal *vo* does not, and hence we treat these separate lexical entries.

[7]This is an example of a generalisation of the karaka chart to include word order constraints.

3.3.2 Arabic Analysis

To account for the Arabic data, we first must establish the relevant vibhakti to karaka mapping. The syntactic role of inflection in Arabic is well defined (cf. (Fischer, 2002)), and the relevant rules (for the examples under consideration) can

be summarised as follows:

- Nominative case: (*-un* or *-u*) marks the subject
- Verbs agree in number and gender with the subject
- Where a verbal argument is absent, a pronominal affix (e.g. *-hā*) is attached to the verb, which will agree in number and gender with the missing item.
- Relativisers (e.g. *allaty*) agree in gender with the noun being modified.

Given that the *karta* relation does not necessarily require an explicit subject, in absence of explicit marking of the subject, the relationship can still be derived from gender agreement[8]. Furthermore, the subject will always precede the object. In the sense that vibhakti can be generalised to include word order (Bharati et al., 1996a), we may also include this condition as part of the karaka chart (see Table 2). As in Hindi, the relativisers impose projectivity on their dependents.

Karaka	*Vibhakti*	*Presence*
karta	(NOM (-uN/-u) or gen agr). and precedes object (if present)	mandatory
karma	ACC (-aN/-a)	optional

Table 2: Default Karaka chart for Arabic

This set of rules is sufficient to account for the grammaticality of the data given here, but does not account for the semantic relationship between the three sentences. In this respect, the karaka relations have so far only illustrated their connection to the purely syntactic level of analysis. While we do not wish to regress to some kind of transformational account of the surface forms, it is desirable to illustrate that the karaka relations provide sufficient abstraction to permit the formulation of a common semantic representation.

One way to approach this is to argue that relativisers 'mediate' the appropriate karaka relation from the complement to the noun being modified. This is essentially a process of unifying the embedded resumptive pronoun in the complement with the relevant external argument. To make this more explicit, we have separated the resumptive pronoun from the complement in the PGF analysis and shown the karaka relation that exists between them.

Thus in 3(a), *allaty* mediates a *karma* relation between *al-samakah* and *ʾakala-hā*. In 3(c), this connection is repeated in the relative clause, and *ʾalladhy* mediates a *karma* relationship between *al-ḍifdaʿ* and *ʾakalat-hu*. In 3(b), there is no explicit resumptive pronoun, since this feature is only used for a missing object. Instead, gender and number agreement between *al-samakah* and *akalat* means that the *karta* relationship is obtained via *allaty*. This mediation of karaka relations is illustrated by a typical feature structure unification diagram shown in (6).

Even though the feature structures differ in terms of argument structure, they can be given an equivalent semantic interpretation in terms of lambda abstraction:

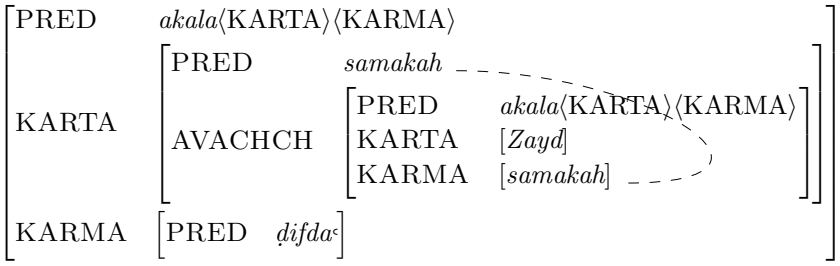
- We can represent 6(a) as: $ate(Zayd, \lambda x[\mathit{fish}'(x) \wedge ate(x,\mathit{frog})]$
- Likewise, 6(b) is represented by: $ate(\lambda x[\mathit{fish}'(x) \wedge (Zayd, x)],\mathit{frog})$
- Both of these can be reduced to: $\lambda x[\mathit{fish}'(x) \wedge ate(Zayd, x) \wedge ate(x,\mathit{frog})]$
- 6(c) merely asserts an additional variable, $\lambda y[ate(Zayd, \lambda x[\mathit{fish}'(x) \wedge ate(x,\mathit{frog}'(y))])]$, which can be factored out, since it is the identity function of *frog*, thus leaving 6(c) with the same interpretation as 6(a).

The same mediation process can be applied to the Hindi data, using the same argument. Here, the relative and (pronominal) co-relative guide the unification. The karaka relation between the (pronominal) relativiser *jo* and the *avachchedak* verb is mapped onto the *avachchinna* noun marked by *vo* or directly onto the pronominal co-relative *vo*, as the case may be. In this sense, the pronominal relative or co-relative operates in the same was the resumptive pronouns in Arabic.

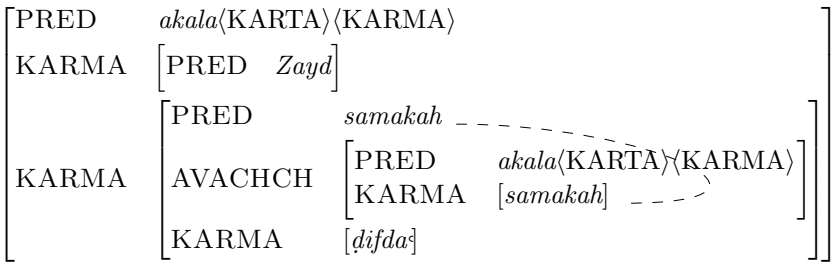
After unification, all three Hindi sentences share the same feature structure, shown in 7, and hence further analysis to demonstrate their semantic equivalence is not needed.

[8]This is also true in Hindi.

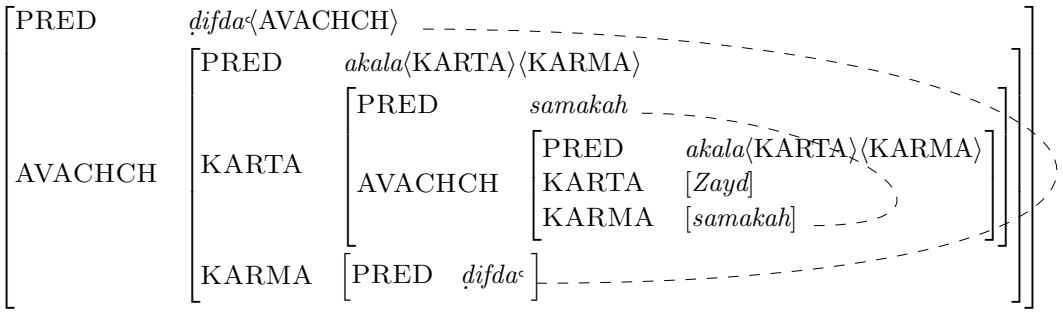
(6) (a)



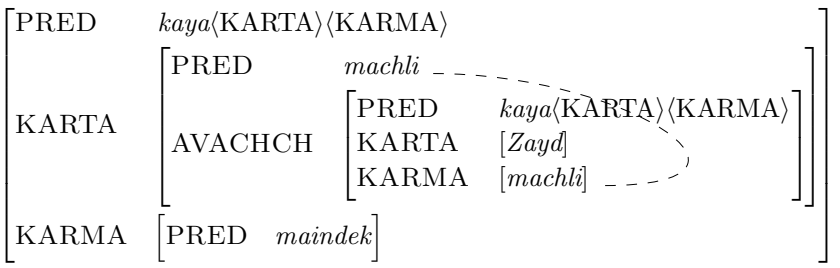
(b)



(c)



(7) (a)



4 Conclusions and Future Work

We have sketched an application of the Paninian Grammar Framework to Arabic, and outlined an approach to deriving a common logical form for equivalent sentences in Hindi and Arabic. In particular, the PGF analysis given here highlights the similarity of resumptive pronoun affix in Arabic to the co-relative in Hindi.

This is a first step toward developing comprehensive Paninian dependency grammars of Arabic and Hindi, with a view to applying the grammars to machine translation. Further development of the grammar is necessary before it is clear that PGF is capable of handling all the requirements of a MT system. In particular, the suitability of PGF for generation needs to be explored, particularly with respect to generating appropriate word orders in the target language.

In terms of on-going development of PGF as a theory for computational linguistics, (Bharati et al., 1996b) and (Pedersen, 2001) have compared PGF to Lexical Functional Grammar (Bresnan and Kaplan, 1982) and Lexicalised Tree Adjoining Grammar (Joshi, 1987) with encouraging results. However a systematic comparison of PGF to more recent dependency grammar formalisms, such as DACHS and TDG, has not yet

been done. Given the strong parallels with recent work in these formalisms (cf. (Kruijff and Duchier, 2003)), such an investigation is now essential.

5 Acknowledgements

Our thanks go to Professor Rajeev Sangal and the Language Technology Research Centre staff at the International Institute of Information Technology, Hyderabad, for their ongoing support of our efforts in applying the Paninian Grammar Framework, Petr Pajas for assistance with the TrEd diagramming tool used for laying out the dependency diagrams, and to Professor Joachim Diederich and the workshop review panel for their helpful comments during the preparation of this paper.

References

A. Bharati, M. Bhatia, V. Chaitanya, and R. Sangal. 1996a. Paninian Grammar Framework Applied to English. Technical Report TRCS-96-238, CSE, IIT Kanpur.

A. Bharati, V. Chaitanya, and R. Sangal. 1996b. *Natural Language Processing - A Paninian Perspective*. Prentice Hall of India, New Delhi.

J. Bresnan and R. Kaplan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In J. Bresnan, editor, *The Mental Representation of Grammatical Relations*. MIT Press.

N. Bröker. 1998. Separating surface order and syntactic relations in a dependency grammar. In *Proceedings of COLING-ACL '98*, pages 174–180.

Miriam Butt and Tracy Holloway-King. 1997. Null elements in discourse structure. In K. V. Subbarao, editor, *Papers from the NULLS Seminar*. Motilal Banarasidas, Delhi.

M. Covington. 1990. Parsing Discontinuous Constituents in Dependency Grammar. *Computational Linguistics*, 16(4), December.

Ralph Debusmann and Denys Duchier. 2003. A meta-grammatical framework for dependency grammar. Technical report, Universität des Saarlande, Saarbrücken, Germany.

Denys Duchier and Ralph Debusmann. 2001. Topological dependency trees: A constraint-based account of linear precedence. In *Meeting of the Association for Computational Linguistics*, pages 180–187.

V. Dwivedi. 1994. Topicalization in Hindi and the correlative construction. In Miriam Butt, Tracy Holloway King, and Gillian Ramchand, editors, *Theoretical Perspectives on word order in South Asian languages*. CSLI Publications, Stanford, CA.

Wolfdietrich Fischer. 2002. *A Grammar of Classical Arabic*. Yale University Press, New haven and London.

V. Gambhir. 1981. *Syntactic Restrictions and Discourse Functions of Word Order in Standard Hindi*. PhD Thesis, University of Pennsylvania, Philadelphia.

R. Hudson. 1984. *Word Grammar*. Basil Blackwell, 108 Cowley Rd, Oxford OX4 1JF, England.

D. Ingalls. 1951. *Materials for the Study of Navya-Nyaya Logic*. Harvard University Press, Cambridge.

A. K. Joshi. 1987. An Introduction to Tree Ajoining Grammars. In A. Manaster-Ramer, editor, *Mathematics of Language*. John Benjamins, Amsterdam.

P. Kiparsky and J. F. Staal. 1969. Syntactic and Semantic Relations in Panini. *Foundations of Language*, 5:84–117.

Geert-Jan M. Kruijff and Denys Duchier. 2003. Information structure in topological dependency grammar. In *EACL 2003, 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 219–226.

Geert-Jan M. Kruijff. 2002. Formal and computational aspects of dependency grammar: History and development of dg. Technical report, ESSLI2002.

B.K. Matilal. 1968. *The Navya-Nyaya Doctrine of Negation*. Harvard University Press, Cambridge.

I. A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University Press of New York.

M. Pedersen. 2001. *Usability Evaluation of Grammar Formalisms for Free Word Order Natural Language Processing*. PhD thesis, School of Computer Science and Electrical Engineering, University of Queensland, Brisbane, Australia.

L. Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.