A Computational Treatment of Differential Case Marking in Malayalam

Abstract

Case is often treated as an uninteresting part of computational processing (both parsing and generation). In the mainly free word order South Asian languages, case plays a theoretically well established role in syntactic and semantic processing. Case is used not only to help identify grammatical relations (e.g., ergatives indicate subjects), but also contributes significantly to the semantic analysis of a clause. This paper extends Butt and King's (2001) computational treatment of case in the Indo-European language Urdu to the Dravidian language Malayalam. The data from Malayalam confirms Butt and King's general approach, by which case markers add requirements about the syntactic structure and the semantic analysis of the clause via individual lexical entries. In particular, the paper proposes a computational treatment of the expression of modality via differential subject case marking.

1 Introduction

Computational treatments of case for South Asian languages that are also theoretically valid tend to be few and far between. Butt and King (2001) propose a computational theory of case which recognizes the fact that case plays an active role in the construction of syntactic and semantic analyses. Case is used not only to help identify grammatical relations (e.g., ergatives indicate subjects), but also contributes significantly to the semantic analysis of a clause.

This paper extends Butt and King's (2001) computational treatment of case in the Indo-European language Urdu to the Dravidian language Malayalam. Butt and King use Lexical-Functional Grammar's (LFG) inside-out functional uncertainty (IO-FU) for a treatment of case. IO-FU has been used in LFG for an analysis of anaphora (Dalrymple 1993) and, in an explicitly computational setting, for the more efficient processing of tagged text. The effect of IO-FU is that a given element may require its context to have a certain structure or analytical configuration (i.e., Urdu pronouns such as *us kaa/kii/kee* can require that they not be bound to a subject) from the "inside out". That is, restrictions pertinent for the final clausal analysis are stated in a bottom-up, rather than top-down approach. The implementation is part of the ParGram (parallel grammar) development effort and uses the XLE development platform (Butt et. al. 1999).

The data from Malayalam confirms Butt and King's general approach to case, by which case markers add requirements about the syntactic structure and the semantic analysis of the clause via individual lexical entries. Unlike in Urdu, where the case markers are clitics and therefore merit individual lexical entries under almost any theoretical approach, Malayalam uses morphological case marking. However, this surface difference does not impinge on the underlying analysis since LFG's f(unctional)-structure representation abstracts away from the morphosyntactic surface differences (e.g., Bresnan 2001). In terms of the computational implementation, the integration of a finite-state morphological analyzer (Beesley and Karttunen 2003), provides access to tags like `+Dat` or `+Acc`. These tags are treated as sublexical items within the XLE grammar development platform used for the Urdu grammar explored by Butt and King and can therefore be associated with lexical entries, just as the Urdu case clitics are.

The paper is organized as follows. The first section lays out the pertinent data, comparing Urdu and Malayalam differential subject marking. This is followed by a section on the treatment of case within LFG and a summary of Butt and King's general approach. Section **??** extends this approach to a treatment of Malayalam's morphological case and section **??** concludes the paper.

2 Constructive Case

Within formal linguistic theory, case has traditionally been thought of in terms of a contrast between structural and inherent case, where inherent case is stipulated in the lexical entry of the verb. In computational terms, this division is easy to encode: structural case is the default case and is part of the grammar, while inherent case, which deviates from the nominative-subject/accusative-object pattern, is included in the lexical entries of individual verbs. In languages such as Georgian, where the choice of the case marker on subjects (nominative, ergative, or dative) depends on the particular tense/aspect morphology of the verb, the case disjunctions have to be encoded in terms of these morphological distinctions as well.

Butt and King (1991, 2001) present a number of case alternations in Urdu/Hindi and argue that they are governed by regular semantic alternations,

rather than by idiosyncratic requirements of individual verbs. While these semantic alternations are also sensitive to structural and morphological conditions, they are not determined exclusively by them. Butt and King (2001, 2003) therefore argue for a three way division in the case system. In addition to structural and quirky case, they posit semantic case and allow case markers themselves to play an active role in the construction of the syntactic and semantic analysis of a clause. As we will see, this is necessary to account for case alternations in Urdu and Malayalam, including the appearance of non-nominative subjects.

This alternative approach to case requires a complex interaction between semantic features, argument structure, grammatical functions, and phrase structure. Lexical-Functional Grammar (LFG) provides the possibility of such a complex interaction via its system of mutually constraining levels of representation or *projections*. Due to its mathematically constrained nature, various computational implementations of LFG exist (`http://clwww.essex.ac.uk/LFG/`). We base this paper on the XLE grammar development platform (see Butt et al. 1999 for an overview and further references).

2.1 Case Alternations

Semantic correlates with alternations in case marking seem to be the rule in South Asian languages and have been firmly established for Urdu/Hindi (e.g., Blake 2001, T. Mohanan 1994, Butt and King 1991, 2001). One of the best known alternations is on objects of transitive verbs which can appear in the nominative or the accusative, as in (1). When the object is accusative, it must be specific (T. Mohanan 1994; Butt 1993; Enç 1991, de Hoop 1996).

(1) a. ram=ne jiraf dekʰ-i
Ram=Erg giraffe.F.Nom see-Perf.F.Sg
'Ram saw a/some giraffe.' Urdu

b. ram=ne jiraf=ko dekʰ-a
Ram=Erg giraffe.F=Acc see-Perf.M.Sg
'Ram saw the (particular) giraffe.' Urdu

In this paper we concentrate on case alternations on subjects. These alternations show a combination of semantic and structural effects. First consider the ergative, whose use in many constructions is correlated with volitionality (Tuite, Agha, and Graczyk 1985; Butt and King 2001). This is seen with certain perfective unergatives whose subject can be either nominative (2a) or ergative (2b).

(2) a. ram kʰãs-a
Ram.M.Nom cough-Perf.M.Sg
'Ram coughed.' Urdu

b. ram=ne kʰãs-a
Ram.M=Erg cough-Perf.M.Sg
'Ram coughed (purposefully).' Urdu

Another interesting alternation that correlates with volitionality or control over an action concerns noun-verb complex predicates and modal readings with infinitives. The sentences in (3) illustrate an alternation with noun-verb complex predicates (T. Mohanan 1994). Here the case alternation interacts with a difference in the choice of light verb: agentive 'do' vs. unaccusative 'come'. The dative *ko* in (3b) marks a goal or experiencer in the manner of psych predicates, while the ergative *ne* marks agentivity or volitionality in (3a), thus confirming the semantic correlation between the ergative case and volitionality.

(3) a. nadya=ne kɑhani yad
Nadya.F.Sg=Erg story.F.Sg.Nom memory
k-i
do-Perf.F.Sg
'Nadya remembered the story (actively).' Urdu

b. nadya=ko kɑhani yad
Nadya.F.Sg=Dat story.F.Sg.Nom memory
a-yi
come-Perf.F.Sg
'Nadya remembered the story (the story came to Nadya).' Urdu

Furthermore, in a departure from the split-ergative pattern in which ergative case is tied to the presence of perfect morphology, Urdu allows the ergative to appear with an infinitive in combination with a present or past form of *ho* 'be'. This construction shows a systematic alternation between ergative and dative subjects, which coincides with a difference in modality, as shown in (4).

(4) a. nadya=ne zu ja-na hɛ
Nadya.F=Erg zoo.M.Loc go-Inf is
'Nadya wants to go to the zoo.' Urdu

b. nadya=ko zu ja-na hɛ
Nadya.F=Dat zoo.M.Loc go-Inf is
'Nadya wants/has to go to the zoo.' Urdu

In this infinitive construction, the ergative is the marked form and entails a subject who has control over the action. The dative is the unmarked form or elsewhere case: the dative subject may or may not have control over the action, the precise interpretation depends on the context (Bashir 1999). For a detailed LFG analysis of this construction see Butt and King (2001).

Strikingly, the same type of pattern emerges in the genetically unrelated Dravidian language Malayalam, pointing to the fact that the use of case as documented and analyzed for Urdu/Hindi is not confined to one language or to one language family, but is characteristic of a larger crosslinguistic pattern that needs to be dealt with correctly and efficiently.

Consider the alternation in (5) between a nominative and a dative subject with the Malayalam modal clitic/suffix *aṇam*. The only difference between (5a) and (5b) is the case marking on the subject. When the subject is nominative, as in (5a), the modal interpretation is a 'must', but not a 'want' one. In contrast, the same verb form with the same accusative marked object but with a dative subject results in a volitional 'want' reading. This basic semantic alternation is similar to the one seen for Urdu in (4), though the surface morphosyntax differs considerably.

(5) a. amma kuṭṭiye aḍik'k'-aṇam
mother.Nom child.Acc beat-want
'Mother must beat the child.' Malayalam

b. ammak'k'ə kuṭṭiye aḍik'k'-aṇam
mother.Dat child.Acc beat-want
'Mother wants to beat the child.' Malayalam

A similar contrast between internal and external control over the action is seen with the permissive in (6) in which the verb has the modal suffix *-aam*. In (6a) the verb takes a nominative subject and the result is a possibility reading. In contrast, in (6b) the same verb appears with a dative subject and the result is one of externally granted permission.

(6) a. avan var-aam
he.Nom come-may
'He may come.' (possibility) Malayalam

b. avanə var-aam
he.Dat come-may
'He may come.' (permission) Malayalam

A similar alternation between nominative and dative marking on subjects in seen in (7). Note that unlike the contrast in (5) and (6) this difference in interpretation correlates with a difference in verb form. Despite the fact that the data in (7) do not represent a minimal pair, the contrast is useful to illustrate a general South Asian tendency by which datives mark both concrete and abstract goals (genitives are also possible, cf. Bengali): (7a) can be interepreted literally as "this is known to me", whereby the "to me" is an abstract goal/location.

(7) a. enik'k'o itə aṛiy-aam
I-Dat this know-Modal
'I know this.' (state of knowledge) Malayalam

b. iaan itə aṛiññu
I-Nom this know-Past
'I know this.' (came to know) Malayalam

We close this section with just one more subject case marking alternation. The examples in (8) are instances of what has been labeled "(dis)ability passives" in grammars (e.g., Glassman 1976, Van Olphen 1980) on Urdu/Hindi. In contrast to the Urdu example in (9), where no subject alternation is possible, Malayalam permits a subject alternation. The alternation this time is not between nominative and dative, but between a nominative and an instrumental. Just as with the nominative/dative alternations seen so far, the alternation forms a minimal pair in which only the case marking on the subject differs. This minimal difference in case marking also results in a semantic difference. The use of the dative subject indicates a temporary inability, while the instrumental subject, as in the Urdu examples, signals a dispositional property of the subject that holds true over long periods of time.

(8) a. meeri-k'k'ə paaḍaan kazhiy-illa/patt-illa
Mary-Dat sing.Inf be.able-neg/can-neg
'Mary cannot sing.' (she is unable to sing for now) Malayalam

b. meeri-ekkoṇḍə paaḍaan kazhiy-illa/patt-illa
Mary-Inst sing-Inf be-able-neg/can-neg
'Mary cannot sing.' (she could never sing/she is too lazy to sing) Malayalam

(9) ʊs=se cɑl-a nahĩ jaega
Pron=Inst walk-M.Sg not go.Fut.M.3.Sg
'She/he can't possibly walk.' (in the context of a broken leg)
(Glassman 1976:275) Urdu

To conclude, the generalization indicated by the data presented in this section is that semantic factors are closely linked with alternative case realization possbilities. Therefore, an analysis which requires brute-force listing of case assignments via lexical stipulation is unfeasible. Instead, an approach is needed whereby case markers can play an active role in the construction of the syntactic and semantic analysis of the clause. This general conclusion is further supported by work on Australian languages (Nordlinger 1998). The next section describes LFG's *Constructive Case* approach to case that has been developed on the basis of data from South Asian languages and Australian languages. This general sketch of the analysis is followed by a description of the implementation necessary for Malayalam.

3 Case in LFG

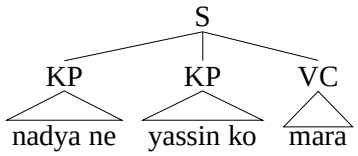
In LFG, information from different components combines to constrain one another and produce a consistent and coherent analysis of a given clause (see Bresnan 2001, Dalrymple 2001 for a recent overview of LFG). The differing modules of grammar (e.g., grammatical functions, semantics, and phonological information) are encoded in terms of projections from lexical entries and phrase structure rules, which in turn encode syntactic and morphological constituency. This is illustrated informally in (11) for the Urdu sentence in (10).

A sentence like (10) has two syntactic structures associated with it. The first is a phrase structure tree, referred to as the c(onstituent)-structure.[1] LFG avoids the use of traces. The c-structure therefore closely reflects the actual string and contains a faithful representation of linear order and constituency information. The grammatical functions are encoded in the f(unctional)-structure as an attribute value matrix (AVM).

(10) nadya=ne yassin=ko mar-a
Nadya.F=Erg Yassin.M=Acc hit-Perf.M.Sg
'Nadya hit Yassin.'

[1]Because Urdu case markers are clitics and hence semi-independent elements, we use the notion of KP (KasePhrase) to encode case-marked noun phrases. The VC stands for "verbal complex".

(11) a. Constituent-structure:

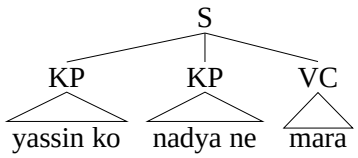


b. Functional-structure:

$$\begin{bmatrix} \text{PRED} & \text{'hit<SUBJ, OBJ>'} \\ \text{TENSE} & \text{PRES} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Nadya'} \\ \text{CASE} & \text{ERG} \\ \text{GEND} & \text{FEM} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Yassin'} \\ \text{CASE} & \text{ACC} \\ \text{GEND} & \text{MASC} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

Note that different word orders of (10), which are possible since Urdu is a "free" word order language, will have different c-structures but identical f-structures; the correlated differences in discourse-functions can be encoded in a separate projection (King 1997) or in the f-structure. Thus the c-structure in (12) also corresponds to the f-structure in (11b).

(12)



3.1 Theoretical Approach

Case phenomena have been extensively analyzed in LFG. Of particular interest here is the idea of Constructive Case, proposed by Nordlinger (1998). The basic idea behind Constructive Case is that constituents with case morphology can define the larger syntactic context in which they appear. This is accomplished via *inside-out functional uncertainty* constraints (Halvorsen and Kaplan 1988; Dalrymple 1993; Andrews 1996) that are associated with the cases. Consider the Wambaya example in (13).

(13) galalarrinyi-ni gini-ng-a dawu
dog.I-ERG 3SG.MASC.A-1.O-NFUT bite

bugayini-ni
big.I-ERG
'The big dog bit me.' Nordlinger (1998:96) Wambaya

In (13), *galalarrinyi-ni bugayini-ni* 'big dog' is a discontinuous constituent, but both parts of the constituent are marked with ergative case. Under Nordlinger's analysis, the ergative case itself specifies that it is an ergative and that it must be part of a subject for the clause to be grammatical. This is outlined in the lexical entry in (14); the first line indicates that the noun phrase has ergative case, while the second states that it must be a subject. This second line is an instance of an inside-out functional uncertainty constraint.[2]

(14) ni $(\uparrow \text{CASE}) = \text{ERG}$
$(\text{SUBJ} \uparrow)$

Together with the predicate value for 'dog' supplied by the noun, this entry for the case marker results in the (simplified) syntactic f-structure in (15) for *galalarrinyi-ni* 'dog-Erg'.

(15) $\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'dog'} \\ \text{CASE} & \text{ERG} \end{bmatrix} \end{bmatrix}$

This structure can then be unified with the representation projected by the adjective (adjunct) *bugayini-ni* 'big-Erg' in (16) to give a coherent analysis of the subject of the clause, as in (17).

(16) $\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{CASE} & \text{ERG} \\ \text{ADJUNCT} & \begin{bmatrix} \text{PRED} & \text{'big'} \end{bmatrix} \end{bmatrix} \end{bmatrix}$

(17) $\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{CASE} & \text{ERG} \\ \text{PRED} & \text{'dog'} \\ \text{ADJUNCT} & \begin{bmatrix} \text{PRED} & \text{'big'} \end{bmatrix} \end{bmatrix} \end{bmatrix}$

[2]We do not discuss the details of the LFG formalism here; these can be found in Dalrymple 2000 and references therein. Basically, the up arrows (↑) encode mappings between nodes of the phrase structure tree and the functional-structure. The '↑' refers to the particular AVM that the phrase structure node in question corresponds to. So, in the examples in this paper, the '↑' refers to the functional-structure of the noun phrase containing the case marker. For example, in (14) the '↑' refers to the AVM with PRED 'dog' in it in (15); thus, the first line of (14) states that this part of the functional-structure contains the pair CASE ERG, as is seen in (15), while the second states that this part of the functional-structure is contained within the SUBJ of the next bigger AVM, as is also seen in (15).

As can be seen from these examples, Constructive Case allows the case markers to play an active role in the clause. Not only do they assign case, but they can also specify information about the syntactic environment in which they occur, e.g., attaching to subject nominals in the example of the Wambaya ergative.

In addition, case markers can also add semantic information about parameters such as volitionality, modality, aspectual affectedness, specificity or partitivity (Butt 2004). An example of such semantic case is the Urdu ergative, which is associated with volitionality or agentivity not only in Urdu, but in a range of languages crosslinguistically (cf. (**??**), (**??**), Butt 2004). A further example of such a case is the dative: the dative in Urdu and South Asian languages in general is associated with a goal argument (Verma and Mohanan 1990). Possible entries for these cases are shown in (18).

(18) a. ko $(\uparrow \text{CASE})=\text{DAT}$
$(\text{GOAL} \uparrow_{a-str})$
$\{ (\text{OBJ}_{theta} \uparrow) \mid (\text{SUBJ} \uparrow) \}$

b. ne $(\uparrow \text{CASE})=\text{ERG}$
$(\text{AGENT} \uparrow_{a-str})$
$(\text{SUBJ} \uparrow)$

The first line of (18a,b) associates the case clitics with the dative and ergative case, respectively. The second line states semantic correspondence at argument structure, goals for datives and agents for ergatives.[3] The third line states the inside-out functional constraints on the outer f-structure of the case marker: ergatives require a subject, while datives can appear with either subjects or thematic (secondary/indirect) objects.

Case in LFG is not necessarily associated with a particular phrase structure position; in fact, this is the least common way case is assigned (see Butt and King 1999, 2003 for some discussion). Instead, the relevant notion is the grammatical function of the case marked noun phrase. In our model, the case does not itself assign the grammatical function but instead helps to characterize the grammatical functions via wellformedness conditions. That is, the information contributed by the case marker provides further constraints on the grammatical functions. This is an important point as case markers

[3]See footnote **??** on the meaning of the up arrows; in this paper, subscripted arrows refer to projections other than the functional-structure, namely the argument-structure.

do not define grammatical functions. Conversely, grammatical functions do not exclusively determine the case marking. Parsing must therefore take into account a complex interaction between a range of different types of information.

3.2 Computational Implementation

The XLE system (see Butt et al. 1999; Butt et al. 2002 for a description and further references) is a platform for large-scale LFG grammar development. XLE comprises interfaces to finite-state preprocessing modules for tokenization and morphological analysis (Karttunen et al. 1992; Kaplan and Kay 1994; Kaplan et al. 2004), as well as an efficient parser and generator for LFG grammars (Maxwell and Kaplan 1991, 1993; Shemtov 1996; Kaplan and Wedekind 2000).

This section briefly describes the existing Urdu implementation before moving on to an extension to Malayalam (section **??**). In particular, we present an analysis of the contrast in (19).

(19) a. nadya=ne zu ja-na hɛ
Nadya.F=Erg zoo.M.Loc go-Inf is
'Nadya wants to go to the zoo.' Urdu

b. nadya=ko zu ja-na hɛ
Nadya.F=Dat zoo.M.Loc go-Inf is
'Nadya wants/has to go to the zoo.' Urdu

Recall that in this infinitive construction, the ergative is the marked form and entails a subject who has control over the action. The dative is the unmarked form or elsewhere case: the dative subject may or may not have control over the action, the precise interpretation depends on the context. Bashir (1999) suggests that the relevant semantic contrast lies in the difference between *internal* vs. *external* control. Internal control entails that the subject has control over the action and can choose whether to perform it or not. External control indicates that the subject has no ability to choose whether or not to perform a given action: the control over the performance of the action is imposed from the outside.

In the current Urdu implementation, the difference between the two sentences in (19) is therefore encoded via a semantic feature named SEM-PROP (semantic property) within the f-structure (part of the SUBJ f-structure). As can be seen in (20) and (21), the analysis of the ergative version contains the semantic information that control over the action is internal. As part of a more sophisticated semantic analysis, this information leads to the inference of a 'want' type of modality.

(20)

```
"nAdyA nE zU jAnA hE"
PRED    'hO<[0:nAdyA] [32:jA]>'
SUBJ    PRED     'nAdyA'
        NTYPE    NSEM [PROPER [PROPER-TYPEname]]
                 NSYN proper
        SEM-PROP CONTROL internal, SPECIFIC +
      0 CASE erg, GEND fem, NUM sg, PERS 3
XCOMP   PRED  'jA<[0:nAdyA] [17:zU]>'
        SUBJ  [0:nAdyA]
        OBL   PRED     'zU'
              CHECK    [_NMORPH obl]
              NTYPE    NSEM [COMMON count]
                       NSYN common
              SEM-PROP LOCATION inherent
     17       CASE loc, GEND masc, PERS 3
        CHECK [_NMORPH nom]
   32   CASE nom, GEND masc, NUM sg, PASSIVE -, PERS 3, VFORM inf
CHECK   [_VMORPH [_MTYPE inf]]
LEX-SEM GOAL +
TNS-ASP MOOD indicative, TENSE pres
52 CLAUSE-TYPEdecl, PASSIVE -, STMT-TYPEdecl, VTYPE main
```

(21)

```
"nAdyA kO zU jAnA hE"
PRED    'hO<[0:nAdyA] [32:jA]>'
SUBJ    PRED     'nAdyA'
        NTYPE    NSEM [PROPER [PROPER-TYPEname]]
                 NSYN proper
        SEM-PROP SPECIFIC +
      0 CASE dat, GEND fem, NUM sg, PERS 3
XCOMP   PRED  'jA<[0:nAdyA] [17:zU]>'
        SUBJ  [0:nAdyA]
        OBL   PRED     'zU'
              CHECK    [_NMORPH obl]
              NTYPE    NSEM [COMMON count]
                       NSYN common
              SEM-PROP LOCATION inherent
     17       CASE loc, GEND masc, PERS 3
        CHECK [_NMORPH nom]
   32   CASE nom, GEND masc, NUM sg, PASSIVE -, PERS 3, VFORM inf
CHECK   [_VMORPH [_MTYPE inf]]
LEX-SEM GOAL +
TNS-ASP MOOD indicative, TENSE pres
52 CLAUSE-TYPEdecl, PASSIVE -, STMT-TYPEdecl, VTYPE main
```

In contrast, the analysis of the dative variant shows *underspecification* for the property SEM-PROP CONTROL. This underspecification is indicated by the absence of such a feature in (21). The modal interpretations that can be inferred by the semantics is therefore that both the 'want' and the 'must' type of modality are possible semantic interpretations in this case.

The fact that a modal reading is to be inferred at all falls out of the particular f-structure analysis assumed in both (20) and (21). The verb *ho* 'be' is treated along the lines of a typical modal verb in the analyses above, it takes a subject and an infinitival complement (XCOMP), the subject of the matrix 'be' ('Nadya') controls the subject of the infintival. This is a typical structure for modals, as is evidenced by the English *Nadya wants to go*, where the subject 'Nadya' controls the subject of the infinitival *to go*.

Since the only overt difference between the two sentences in (19) is the case marker (dative vs. ergative), the information whether the subject has internal control over the action or not comes out of the lexical entry for the case marker and no other place.

The complete entry for the Urdu dative/accusative case marker *ko* is shown in (22).

When this case marker is used as a dative (`option 1`), the dative can be associated either with subjects (SUBJ, `option 1a`)) or indirect objects (OBJ$_{theta}$, `option 1b`) in the f-structure. Both of these possibilities correspond to a goal argument at argument structure. Here, the fact that it must be a goal is represented as a call to a template (@GOAL). Templates are implementational devices in LFG used to capture generalizations in the lexicon and grammar. For example, it would be possible to detail all lexical information for each verb entry. However, this would lead to maintenance problems, as well as increase the chances of making typographical errors. So, instead a template is created containing all the relevant information, and the lexical entries call the template. See Butt et al. 1999, Butt et al. 2003 for more on templates in XLE.[4] As an accusative (`option 2`), the *ko* denotes semantic specificity or definiteness (cf. (**??**)) and is restricted to objects (OBJ).

(22) ko K * { (↑CASE) = dat `option 1`
{ (SUBJ↑) `option 1a`
|(OBJth ↑) } `option 1b`
@GOAL
|(↑CASE) = acc `option 2`
(OBJ↑)
(↑SEM-PROP SPECIFIC) = +}.

The complete entry for the ergative case marker *ne* is shown in (23). As can be seen, it is much simpler than the entry for the dative/accusative *ko* in (**??**). This is primarily because ergatives can only appear on subjects.

(23) ne K * (↑CASE) = erg
(SUBJ↑)
@VOLITION

However, the entry as presented in (23) is deceptively simple, because of the template call to @VOLITION in line 3. This template governs the rather complex relationship between ergative case, perfect aspect and volitional semantics (internal control) that exists in Urdu. This relationship has been well documented (e.g., Davison 1999) and will not be discussed here beyond noting that the @VOLITION template as implemented contains a rather complex bundle of interdependencies.

4 Case in Malayalam

Now consider the Malayalam alternation repeated in (24), in which a dative subject alternates with a nominative, resulting in different interpretations of the modality.[5]

(24) a. avan var-aam
he.Nom come-may
'He may come.' (possibility) Malayalam

b. avanə var-aam
he.Dat come-may
'He may come.' (permission) Malayalam

Unlike in Urdu, in Malayalam the cases are not clitics, but affixes on the nouns. This poses no problem from an implementational standpoint because of the integration of a finite-state morphology for Malayalam. Finite-state morphologies (Beesley and Karttunen 2003, Butt et al. 1999) associate surface forms with canonicalized stems (lemmata) and a series of tags encoding the relevant morphological information. For example, the subject pronouns in (24) would be associated with the lemmata and tags in (25).

(25) a. avan ↔ avan +Pron +3 +Sg +Nom

b. avanə ↔ avan +Pron +3 +Sg +Dat

These tags are assigned lexical entries and combined via sublexical rules in XLE (Kaplan et al. 2004). A possible sublexical rule for Malayalam pronouns is shown in (26).

(26) PRON ⟶ PRON_STEM
PRON_SFX
PERS_SFX
NUM_SFX
CASE_SFX.

What is of interest to us here is the lexical entry for the case tags +Nom and +Dat. As with the Urdu dative and ergative case clitics, these case tags are associated with both argument-structure

[4] Remember that in section **??** we argued that in theoretical analyses of these constructions the goal is represented in argument-structure, not functional-structure. It is possible to implement an argument-structure in XLE (Butt and King 2001). However, for ease of grammar maintenance, we encode these thematic role restrictions in the f-structure. The template expansion for (22) is shown in (i).

(i) GOAL = (({ SUBJ| OBJtheta } ↑) LEX-SEM GOAL) = +

[5] The discussion in this section is based on data from Jayaseelan 1999, 2001, Madhavan 1997, Hany Babu 1997.

and functional-structure information. The relevant entries are shown in (27). As in Urdu, the abstract entries for case provide information about the value of the CASE feature, make a connection to thematic argument structure information, and contribute inside-out functional constraints as to the possible grammatical functions this case marker can appear on. Note that the nominative has no particular argument structure specification because nominatives (as in many languages crosslinguistically) can be associated with a large range of thematic roles (cf. Blake 2001, Butt 2004).

(27) a. +Dat CASE_SFX
(↑CASE)=dat
@GOAL
{ (SUBJ↑) | (OBJtheta ↑) }

b. +Nom CASE_SFX
(↑CASE)=nom
{ (SUBJ↑)
@INTERNAL-CTRL
| (OBJ↑) }

The entry for the nominative also includes information that is meant to feed into a subsequent sophisticated semantic analysis. The template call to INTERNAL-CTRL ensures that when the nominative marks a subject of an active sentence, this participant is interpreted as having internal control over the action ((↑SEM-PROP CONTROL) = internal). The entry for the dative has no such specification. This is because the dative can be used in a range of subtle shades of meaning, as evidenced by the data presented in section **??**. As in Urdu, therefore, the dative is left unspecified for the CONTROL feature. The precise semantics involved in the dative subject sentences in section **??** must be computed from the fact that there is a dative subject in conjunction with the particular lexical semantics of the verb/modal that is involved.[6] The Malayalam data can thus be dealt with along the same lines as the Urdu data. This is linguistically satisfying, as it points to a case marking strategy that is an areal characteristic for South Asian languages. It is also computationally satisfying because similar implementation strategies can guide the development of grammars for genetically diverse languages.

Before presenting the f-structure analyses for (**??**), we briefly discuss the sublexical rules needed for an analysis of the modal suffix, which is also formed in the morphological domain, like the case markers, rather than being realized as an independent lexical item. The analysis of the modals is relatively simple because the FST morphological analysis described above also applies to the verbal domain in our example. A possible analysis for *varaam* is shown in (28) with sample lexical entries in (29) and the relevant sublexical rule in (30).

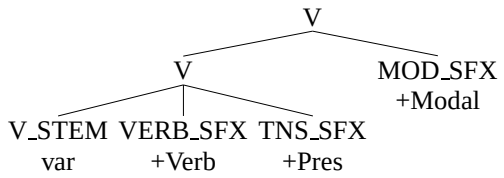
(28) varaam ↔ var +Verb +Pres +Modal

(29) a. var V_STEM (↑PRED)='var<SUBJ>'
b. +Verb VERB_SFX
c. +Pres TNS_SFX (↑TENSE)=pres
d. +Modal MOD_SFX
(↑PRED)='aam<SUBJ,XCOMP>'

(30) V ⟶ { V_STEM `option 1`
VERB_SFX
TNS_SFX
| V: (↑XCOMP)=↓ `option 2`
MOD_SFX }

This sublexical rule is more complicated than the one for the pronoun. The first option states that a verb can consist of a verb stem with tense marking. This would be the usual case. However, verbs with modal suffixes are more complicated and require the second option. Here, the modal suffix *aam* is the head of the V, which in turn takes the `option 1` expansion of V as its XCOMP.

(31)



Given the entries for case and the verbal sublexical rules, the f-structure analysis of the examples in (**??**) follow quite straightforwardly. The modal *aam* 'may' is analyzed as taking a subject and an XCOMP in both (32) and (33). As mentioned previously, this is a standard LFG analysis for modals. The analysis of the semantic difference in the alternation in

[6] An in-depth discussion of this interaction goes beyond the scope of this paper, but there are many more interesting things to be said about the distribution of nominative vs. dative vs. instrumental subjects. Unaccusative verbs like 'die' or 'fall', for example, are incompatible with dative subjects in the construction in (**??**). This has nothing to do with internal/external control, but follows from the fact that unaccusative verbs have theme arguments. This is incompatible with the requirement of a goal argument in the lexical entry of the dative ((27a)).

(**??**) is as follows. The semantics of 'may' modality is in principle compatible with both internal and external control. The preferred and even default interpretation of subjects crosslinguistically, however, is as actors, i.e., as participants which have internal control over an action. Languages also contain a set of verbs for which this does not apply. Well known examples are psych verbs (e.g., 'fear'), verbs of sensation (e.g., 'hear', 'see'), or a subclass of the modal verbs. In order to mark the deviance in the possible interpretation of the subject as not having internal control over an action, languages can avail themselves of differing strategies. A very common strategy is to mark the subject as special, usually by some type of non-nominative case. A popular crosslinguistic choice for pysch verbs and experiencer verbs in general is the dative.

(32)
$$\begin{bmatrix} \text{PRED} & \text{'aam<SUBJ,XCOMP>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{CASE} & \text{dat} \end{bmatrix} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{'var<SUBJ,XCOMP>'} \\ \text{SUBJ} & [\ ] \end{bmatrix} \\ \text{TNS-ASP} & \begin{bmatrix} \text{MOOD} & \text{indicative} \\ \text{TENSE} & \text{pres} \end{bmatrix} \\ \text{LEX-SEM} & [\text{GOAL} \ +] \\ \text{STMT-TYPE} & \text{decl} \end{bmatrix}$$

(33)
$$\begin{bmatrix} \text{PRED} & \text{'aam<SUBJ,XCOMP>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{CASE} & \text{nom} \\ \text{SEM-PROP} & [\text{CONTROL} \quad \text{internal}] \end{bmatrix} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{'var<SUBJ,XCOMP>'} \\ \text{SUBJ} & [\ ] \end{bmatrix} \\ \text{TNS-ASP} & \begin{bmatrix} \text{MOOD} & \text{indicative} \\ \text{TENSE} & \text{pres} \end{bmatrix} \\ \text{STMT-TYPE} & \text{decl} \end{bmatrix}$$

In the particular alternation in (**??**), the dative is therefore used as the canonical indicator of a possible situation of non-internal control on the part of the subject. It is the "normal" choice in conjunction with the modal semantics of 'may'. On the other hand, the nominative is associated with internal control over an action. This is not the normal choice and gives rise to a more restricted reading, namely, the possiblity reading, as it is up to the subject to decide whether to perform the action or not. In the permission reading, by contrast, control over the action is external in the sense that permission must be granted for the action to occur, a semantics that is more in line with the use of the dative case.

This difference in the semantics that are to be inferred at a later stage in the analysis can be computed from the presence of the CONTROL feature in the f-structure analysis of the example with the nominative subject ((33)), and its absence in the f-structure analysis of the example with the dative subject in (32). Because this difference in semantic interpretation is triggered by the minimal difference in the case marking on the subject, an analysis like the one presented here, which encodes the difference squarely in the lexical entry of the case markers would seem to be desirable.[7]

5 Conclusion

In this paper, we have presented a theoretical approach to non-nominative subjects in Malayalam which involves a "constructive" treatment of case. Under our analysis, the case markers themselves are specified for structural and semantic information. This information interacts with information specified in other parts of the grammar (primarily the verbal lexical entries) in order to produce wellformed analyses. The existence of several semantically motivated case alternations in Malayalam points to the need for incorporating semantic information into any approach to Malayalam case marking. Under our analysis, Malayalam dative and instrumental cases can be seen as *semantic cases* in the sense that they help express semantically motivated alternations.

The paper also showed that this theoretically motivated approach to case is computationally viable in the sense that the resulting analyses are constrained in just the right way. We discussed the integration of semantically based features into the implementation and showed how this information representation could facilitate the formulation of a non-stipulative and generally applicable account of case-marking and semantically conditioned case alternations. The account is both linguistically satisfying and computationally feasible.

[7] Note that the f-structures in (32) and (33) have been simplified for expository purposes.

References

Avery Andrews. 1996. Semantic case-stacking and inside-out unification. *Australian Journal of Linguistics*, 16:1–55.

Elena Bashir. 1999. The Urdu and Hindi ergative postposition *ne*: Its changing role in the grammar. In Rajendra Singh, editor, *The Yearbook of South Asian Languages and Linguistics*, pages 11–36. Sage Publications, New Delhi.

Barry Blake. 2001. *Case*. Cambridge University Press, Cambridge. Second Edition.

Joan Bresnan and Annie Zaenen. 1990. Deep unaccusativity in LFG. In Katarzyna Dziwirek, Patrick Farrel, and Errapel Mejias-Bekandi, editors, *Grammatical Relations: A Cross-Theoretical Perspective*, pages 45–57. CSLI Publications, Stanford.

Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell, Oxford.

Miriam Butt and Tracy Holloway King. 1991. Semantic case in Urdu. In Lisa Dobrin, Lynn Nichols, and Rosa M. Rodriguez, editors, *Papers from the 27th Regional Meeting of the Chicago Linguistic Society*, pages 31–45.

Miriam Butt and Tracy Holloway King. 2001a. Non-nominative subjects in Urdu: A computational analysis. In *Proceedings of the International Symposium on Non-nominative Subjects*, pages 525–548, Tokyo. ILCAA.

Miriam Butt and Tracy Holloway King. 2001b. The status of case. In Veneeta Dayal and Anoop Mahajan, editors, *Clause Structure in South Asian Languages*. To appear.

Miriam Butt and Tracy Holloway King. 2003. Case systems: Beyond structural distinctions. In Ellen Brandner and Heike Zinsmeister, editors, *New Perspectives on Case Theory*, pages 53–87. CSLI Publications, Stanford.

Miriam Butt, Tracy Holloway King, Maria-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. CSLI Publications, Stanford.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Miriam Butt, Martin Forst, Tracy Holloway King, and Jonas Kuhn. 2003. The feature space in parallel grammar writing. In *Proceedings of the ESSLLI 2003 Workshop on Ideas and Strategies for Multilingual Grammar Development*, pages 9–16.

Miriam Butt. 2004. *Theories of Case*. Cambridge University Press, Cambridge. To Appear.

Mary Dalrymple. 1993. *The Syntax of Anaphoric Binding*. CSLI Publications, Stanford.

Mary Dalrymple. 2001. *Lexical Functional Grammar*. Academic Press, New York. Syntax and Semantics Volume 34.

Alice Davison. 1999. Ergativity: Functional and formal issues. In Michael Darnell, Edith Moravcsik, Frederick Newmeyer, Michael Noonan, and Kathleen Wheatley, editors, *Functionalism and Formalism in Linguistics, Volume I: General Papers*. John Benjamins, Amsterdam.

Helen de Hoop. 1996. *Case Configuration and Noun Phrase Interpretation*. Garland, New York.

Mürvet Enç. 1991. The semantics of specificity. *Linguistic Inquiry*, 22(1):1–25.

Eugene Glassman. 1976. *Spoken Urdu*. Nirali Kitaben, Lahore.

Per-Kristian Halvorsen and Ronald M. Kaplan. 1988. Projections and semantic description in Lexical-Functional grammar. In *Proceedings of the International Conference on Fifth Generation Computer Systems (FGCS-88)*, pages 1116–1122, Tokyo, Japan. Reprinted in Mary Dalrymple, Ronald M. Kaplan, John Maxwell, and Annie Zaenen, eds., *Formal Issues in Lexical-Functional Grammar*, 279–292. CSLI Publications, Stanford. 1995.

M.T. Hany Babu. 1997. *The Syntax of Functional Categories*. Ph.D. thesis, CIEFL, Hyderabad.

John T. Maxwell III and Ron Kaplan. 1991. A method for disjunctive constraint satisfaction. In M. Tomita, editor, *Current Issues in Parsing Technology*, pages 173–190. Kluwer Academic Publishers, Boston.

John T. Maxwell III and Ron Kaplan. 1993. The interface between phrasal and functional constraints. *Computational Linguistics*, 19:571–590.

K.A. Jayaseelan. 1999. *Parametric Studies in Malayalam Syntax*. Allied Publishers, New Delhi.

K.A. Jayaseelan. 2001. Malayalam non-nominative subjects. In *Proceedings of the International Symposium on Non-nominative Subjects*, Tokyo. ILCAA.

Ronald Kaplan and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan, editor,

The Mental Representation of Grammatical Relations, pages 173–281. The MIT Press, Cambridge, Massachussetts. Reprinted in Mary Dalrymple, Ronald M. Kaplan, John Maxwell, and Annie Zaenen, eds., *Formal Issues in Lexical-Functional Grammar*, 29–130. CSLI Publications, Stanford. 1995.

Ronald Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20:331–378.

Ronald Kaplan and Jürgen Wedekind. 2000. LFG generation produces context-free languages. In *Proceedings of COLING 2000, Saarbrueken*, pages 141–148.

Ronald Kaplan, John T. Maxwell III, Tracy Holloway King, and Richard Crouch. 2004. Integrating finite-state technology with deep LFG grammars. In *Proceedings of the ESSLLI Workshop on Combining Shallow and Deep Processing for NLP*.

Ronald M. Kaplan. 1995. Three seductions of computational psycholinguistics. In Mary Dalrymple, Ronald M. Kaplan, John Maxwell, and Annie Zaenen, editors, *Formal Issues in Lexical-Functional Grammar*, pages 339–368. CSLI Publications, Stanford.

Lauri Karttunen, Ron Kaplan, and Annie Zaenen. 1992. Two level morphology with composition. In *Proceedings of COLING 92, Nantes*, volume I, pages 141–148.

Tracy Holloway King. 1997. Focus domains and Information-Structure. In *Proceedings of the LFG 1997 Conference*. CSLI On-line Publications.

P. Madhavan. 1997. Null accusative case and Malayalam causatives. Manuscript, Paper presented at the NULLS Seminar at the University of Delhi, 1997.

Tara Mohanan. 1994. *Argument Structure in Hindi*. CSLI Publications, Stanford.

Rachel Nordlinger. 1998. *Constructive Case: Evidence from Australian Languages*. CSLI Publications, Stanford.

Hadar Shemtov. 1996. Generation of paraphrases from ambiguous logical forms. In *Proceedings of COLING 1996*, pages 919–924.

Kevin Tuite, Asif Agha, and Randolph Graczyk. 1985. Agentivity, transitivity, and the question of active typology. In W. H. Eilfort, P. D. Kroeber, and K. L. Peterson, editors, *Papers from the Parasession on Causatives and Agentivity at the 21st Regional Meeting of the Chicago Linguistic Society*, pages 252–270.

Herman Van Olphen. 1980. *First-Year Hindi Course*. Department of Oriental and African Languages and Literatures, University of Texas, Austin, Texas, Austin.

M. K. Verma and K.P.Mohanan, editors. 1990. *Experiencer Subjects in South Asian Languages*. CSLI Publications, Stanford.