MODELING NOMINAL PREDICATIONS IN HINDI/URDU

Doctoral thesis for obtaining the academic degree
Doctor of Philosophy

submitted by

Sebastian Sulger

at the

University of Konstanz
Faculty of Humanities
Department of Linguistics
MODELING NOMINAL PREDICATIONS IN
HINDI/URDU

Doctoral thesis for obtaining the academic degree
Doctor of Philosophy

submitted by

Sebastian Sulger

at the

University of Konstanz
Faculty of Humanities
Department of Linguistics
Date of the oral examination: June 22nd, 2015
First referee: Prof. Dr. Miriam Butt
Second referee: Tracy H. King, Ph.D.
Third referee: Prof. Dr. Frans Plank
Modeling Nominal Predications in Hindi/Urdu

Abstract

The identification and classification of nominal predicators and their arguments is a notorious problem in natural language processing (NLP). Semantic reasoning, information retrieval, question answering and other applications can benefit greatly from a successful treatment of nominal predication (Meyers et al., 2004b). Overall, a considerable amount of work in NLP focuses on the identification and annotation of verbal predication and arguments thereof, and there is less research on types and identification of nominal predicates and their arguments. This thesis is a contribution of the latter type. It focuses on the description and analysis of nominal arguments in the South Asian language Hindi/Urdu. The analysis is couched within the theory of Lexical-Functional Grammar (LFG, Bresnan, 2001; Dalrymple, 2001) and is implemented in a computational grammar, the Urdu ParGram grammar (Butt and King, 2007), which is a part of the ParGram (“Parallel Grammar”) project on parallel LFG grammar engineering (Butt et al., 1999a, 2002, 1999b). The implementation makes use of the grammar development platform XLE (Crouch et al., 2015).

Different types of case-marked nominal arguments in Hindi/Urdu are examined: genitive, locative as well as instrumental arguments. Among these, the genitive is a special case marker in that it features morphological agreement with the head noun (Butt and King, 2004b). Each of these argument types is discussed in detail regarding the case marking strategies employed, their general linear order within the noun phrase, the selection by distinct types of head nominals, their overall functional behavior as well as binding properties.

All types of nominal arguments exhibit scrambling and can occur outside of the noun phrases they are licensed in; this is attributed to the tight correlation between the case marking and the thematic role realized by the argument as well as, in the case of the genitive, the morphosyntactic agreement between the case marker and the head noun. In addition, all types of nominal arguments may regularly undergo a process of argument suppression, which results in the argument not being realized, but existentially bound (Barker, 1995); an analysis of these cases in terms of pronominal drop is considered, but rejected. The thesis also reviews possessive and locative clauses, which are analyzed as intransitives and copula clauses, respectively. These novel analyses are shown to better account for the functional behavior, binding patterns, as well as overall structural paradigms of these clauses compared to previous analyses.

The thesis also includes a discussion of noun-verb complex predicates in Hindi/Urdu. Here, nouns (referred to as “nominal hosts”) and verbs (also called “light verbs”) form a single predicate with a single set of grammatical functions, whereas the argument structure is complex, as the nominal host may itself contribute arguments to the overall
predication (Mohanann, 1994). While the construction in and of itself is theoretically comparatively well-understood, the combinatorial possibilities between the nominal host and the light verb are not; some hosts occur with several light verbs, others occur with subsets, while still others occur with a single light verb only. Two corpus studies are discussed that aim at 1) uncovering the constraints on combining hosts with light verbs and 2) creating a lexical resource that can serve as input to NLP applications. One such application is the Urdu ParGram grammar, where it is shown how the results from the corpus studies can be translated into templates that model the combinatorial patterns in terms of statistical (dis)preferences.

From the point of view of grammar development, the thesis argues that a unified account of genitive arguments, as currently employed in ParGram, cannot be maintained. Instead, the thesis proposes to use a more detailed approach that can successfully account for the observed patterns. This connects to the issue of parallelism in ParGram. Conventions developed within the ParGram grammars are extensive and dictate the form and possible values of the features used in the grammars as well as the type of analysis chosen for a particular construction. Grammar writers are in principle only allowed to abandon parallelism if maintaining it would be at the cost of misrepresenting the linguistic facts (Butt and King, 2007, Butt et al., 1999b, King et al., 2005). Being faithful to the facts of nominal predication in Hindi/Urdu entails abandoning the ParGram analysis for possessives, which does not distinguish between different types of possessives (Dipper, 2003).

On the other hand, the implementation profits from the detailed linguistic analysis that is applied within ParGram and uses an array of notational instruments in XLE that model the generalizations in an accurate manner. The complete grammar (in its most recent version, as of the time of submission of the thesis) is included on the CD-ROM attached to this document; the implementation can also be tested using the online INESS platform for treebanking and LFG grammar testing (Rosén et al., 2012a, b).

1The INESS homepage is located at: http://iness.uib.no
Modeling Nominal Predications in Hindi/Urdu

ZUSAMMENFASSUNG


Alle Typen dieser Argumente zeigen Scrambling, wodurch sie außerhalb der Nominalphrase erscheinen können, in der sie lizensiert sind; diese Eigenschaft wird der eigenen Beziehung zwischen der Kasusmarkierung und der thematischen Rolle, die durch das Argument realisiert wird, zugeschrieben, sowie (im Falle des Genitiv) der morphologischen Kongruenz zwischen Kasusmarkierer und Kopfnomen. Darüber hinaus sind alle Arten von substantivischen Argumenten dem Prozess der Argumentunterdrückung unterworfen, was ihre Nichtrealisierung und gleichzeitige existenzielle Bindung zur Folge hat (Barker, 1995); eine Analyse dieser Fälle mit Berufung auf Pronominal-Drop wird in Betracht gezogen, jedoch abgelehnt. Die Dissertation behandelt des Weiteren Possessivsätze sowie Lokativsätze, welche als Intransitive respektive Kopulasätze analysiert werden. Es wird gezeigt, dass diese neuen Arten der Erschließung dieser Satzarten ihr funktionales Verhalten, ihre Bindungsmuster sowie ihre generellen strukturellen Paradigma besser erklären als vorige Analysen.


Andererseits profitiert die Implementierung von der detaillierten linguistischen Analyse, wie sie innerhalb des ParGram-Projekts realisiert wird, und benutzt ein Aufgebot verschiedener Notationen in XLE, die die Generalisierungen adäquat modellieren. Die komplette Grammatik (in ihrer aktuellen Version, zum Zeitpunkt der Einreichung der Dissertation) ist auf der CD-ROM enthalten, die diesem Dokument beigefügt ist; die Implementation kann außerdem online mittels der INESS-Plattform für Baumbanken und LFG-Grammatik-Erprobungen getestet werden (Rosén et al., 2012a, b).

2Die INESS-Webseite befindet sich hier: http://iness.uib.no
Acknowledgments

Writing this thesis would not have been possible without the help and encouragement of the people around me. This is not something I write just to be nice — it literally would not have happened.

First, I would like to thank my advisor, Miriam Butt, for her support throughout my studies. When I was hired by Miriam as an assistant (this was back in 2007), I knew that I was fond of languages and that I liked computers. Miriam drew my interest towards grammar writing, which enabled me to work on language using computers, in a way that is theoretically well-founded. In addition, she acquired funding for three different projects which I have been a part of over the years, and put me in the position of pursuing exciting research in different directions. Her comments on drafts of this work were invaluable. I have no idea where I would be without her believing in me.

Tracy Holloway King, the second member of my dissertation committee, has also read drafts of the thesis and provided many comments, both regarding the soundness of the analyses as well as the computational implementation. During several Skype meetings, she patiently discussed the data, annotations and general grammar development issues with me. She happily wrote up her comments in amazing detail. She also sent me a great pancake recipe once, which I still make use of from time to time.

The third member of my committee is Frans Plank, who I got to know at a very early stage in my studies, when he taught the segment on morphology as part of the Introduction to Linguistics; I believe this was in the winter term of 2003. Several years later, in 2012, I had the opportunity of assisting him in organizing a workshop on the genitive case, which brought together a nice group of people from all over the globe; many ideas that went into the chapter on the Hindi/Urdu genitive materialized during that workshop.

I have greatly profited from discussions with many colleagues from Konstanz and elsewhere. Rajesh Bhatt has continuously provided feedback on my ideas during his visits to Konstanz. Stefano Quaglia has commented extensively in exchanges in Konstanz as well as during a conference trip to Indonesia. Tafseer Ahmed as well as Ghulam Raza have given much input; the two of them have also served throughout as native speaker consultants. Ashwini Vaidya has carried out a corpus study on complex predicates with me, which I ended up incorporating in one of the chapters.

My work environment has been amazing. Office mates are more than a nice-to-have
bonus. They are an integral part of one’s well-being. Tina Bögel, Melanie Seiss and, later, Farhat Jabeen have been great at inquiring about my sanity at the right times, and knowing when to shut up. For me, this makes all the difference.

The native speakers that have provided me with the Hindi/Urdu data are, in alphabetical order: Qaiser Abbas, Tafseer Ahmed, Rajesh Bhatt, Miriam Butt, Farhat Jabeen, Asad Mustafa, Ghulam Raza and Ashwini Vaidya. I am deeply thankful for all their help, which they provided in countless e-mails, Skype calls, or direct elicitations, often at very short notice when there was a deadline and I needed some input at the last minute. Joachim Kleinmann has been of great assistance on the technical side, be it with answering hardware questions or figuring out server administration issues.

My friends in Konstanz have been there for me when I needed some much-needed diversion; they have always been entertaining, during movie visits, dinners, soccer practices, and when sharing a beer at the pub. The professional philosophers among them tried to keep in-depth discussions of moral philosophy to a minimum during lunch breaks. Thanks go to Iris Bräuning, Wolfgang Egner, Paul Ehrhardt, Manuel Fallmann, Simon Hopp, Samuel Kis, Marc Knöbelspies, Brian Leahy, Stefan Leber, Thomas Mayer, Alexander Schmidt, Florian Schönhuber, Kyle Wohlmut.

The Urdu ParGram Group in Konstanz has developed into quite a bit more than just a work group. Miriam, Tina, Nette (Annette Hautli-Janisz) and I share friendships that I think none of us were hoping for when we first met. Together, we have suffered when deadlines approached, when papers were rejected, when hangovers were bad. On the bright side, we have had great times at many conferences all around the world, and we always managed to put in a few extra days to enjoy some time off on road trips through Turkey, Australia or California. Thank you, guys.

I am grateful for having a family that has always supported me, during all these years. My trips home to the other side of the lake have been little islands of peace, where I could relax and not think about the distress of academic life. My heartfelt thanks for all of their moral as well as financial aid. Also, I extend a big thank you to my in-laws in Bavaria, for being good friends, for being curious, for being critical. Those weekends in Schöffau have been wonderful.

Needless to say, I leave the most important person for the end. Steffi has proven incredibly patient when it became obvious that I needed more time to finish; she never held it against me, but instead encouraged me to keep going, and supported me in every way imaginable. That’s not to be taken for granted, since she was the person who suffered the most, especially towards the end. For that, I owe her my deepest gratitude.
# Contents

1 Preface
   1.1 Introduction ............................... 1
   1.2 Nominal Predicates ........................ 2
   1.3 Nominal Arguments in Hindi/Urdu ............ 3
   1.4 Constraining Nominal Arguments ............. 7
   1.5 The Linear Order of Nominal Arguments ........ 8
   1.6 Possessive and Locative Clauses ............. 10
   1.7 Noun-Verb Complex Predicates ............... 11
   1.8 Structure of the Thesis ................. 13

2 Background
   2.1 Introduction ................................ 15
   2.2 Nominal Predicates & Arguments .......... 15
      2.2.1 A Brief Overview of the Theoretical Linguistics Literature .... 16
      2.2.2 Semantic Role Labeling ................... 19
      2.2.3 Summary ................................. 20
   2.3 Hindi/Urdu ............................... 21
      2.3.1 Configurationality ..................... 21
      2.3.2 Case .................................. 22
      2.3.3 Postpositions ......................... 25
      2.3.4 Agreement ............................. 26
      2.3.5 Information Structure ............... 27
   2.4 Lexical-Functional Grammar ............... 28
      2.4.1 General Description and Architecture Assumed .... 29
      2.4.2 C-Structure and F-Structure ........ 30
      2.4.3 Mapping Between C- and F-Structure .... 31
      2.4.4 An Inventory of Grammatical Functions .... 34
      2.4.5 F-Structure Wellformedness Conditions ... 35
         2.4.5.1 Completeness .................... 36
         2.4.5.2 Coherence .................... 36
         2.4.5.3 Consistency ................ 37
   2.5 XLE .................................... 38
3.2.2.1 Payne (1995) ........................................... 80
3.2.2.2 Spencer (2008) .................................. 81
3.2.3 Possessive Pronouns .................................. 83
3.2.4 The Distribution of the Genitive .................. 84
3.3 Related Work in LFG Theory ......................... 85
  3.3.1 English ........................................... 85
  3.3.2 Hungarian ....................................... 88
  3.3.3 Welsh ............................................ 90
3.4 Testing for Subjecthood in Hindi/Urdu ............ 92
  3.4.1 Reflexive Binding ................................ 92
  3.4.2 Pronominal Coreference ......................... 93
  3.4.3 Control ......................................... 95
3.5 Hindi/Urdu Genitive Arguments .................... 97
  3.5.1 Constituent Properties .......................... 97
     3.5.1.1 Coordination ................................ 97
     3.5.1.2 Linear Order ............................... 98
     3.5.1.3 Stacking ................................... 101
     3.5.1.4 Summary ................................... 102
  3.5.2 Functional Properties ............................ 103
     3.5.2.1 Semantic Unrestrictedness ................. 103
     3.5.2.2 Testing for Subjecthood inside NPs ......... 104
     3.5.2.3 Common Nouns .............................. 105
     3.5.2.4 Event Nouns ................................ 109
     3.5.2.5 Relational Nouns ......................... 115
     3.5.2.6 Proper Nouns & Pronouns ................... 117
  3.5.3 Pro-Drop or Argument Suppression? ............. 118
  3.5.4 Arguments or Adjuncts? .......................... 122
  3.5.5 Summary ......................................... 123
3.6 Genitive Scrambling .................................. 124
  3.6.1 General Description ............................. 125
  3.6.2 Preferences and Constraints ................. 127
     3.6.2.1 Local Attachments are Preferred ........... 127
     3.6.2.2 Scrambling and Case ....................... 128
     3.6.2.3 Scrambling from Complement Clauses ...... 129
     3.6.2.4 No Scrambling out of Adjuncts .......... 129
     3.6.2.5 No Scrambling from Deep Within .......... 130
  3.6.3 A Comparison to Turkish ....................... 131
  3.6.4 Summary ......................................... 133
3.7 The Genitive in Possessive Clauses ................. 135
  3.7.1 Constituent Properties ........................ 136
     3.7.1.1 Coordination ............................... 136
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3.2.3</td>
<td>Event Nouns</td>
<td>192</td>
</tr>
<tr>
<td>5.3.2.4</td>
<td>Relational Nouns</td>
<td>194</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Pro-Drop or Argument Suppression?</td>
<td>196</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Arguments or Adjuncts?</td>
<td>196</td>
</tr>
<tr>
<td>5.3.5</td>
<td>Summary</td>
<td>199</td>
</tr>
<tr>
<td>5.4</td>
<td>Scrambling</td>
<td>199</td>
</tr>
<tr>
<td>5.4.1</td>
<td>General Description</td>
<td>200</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Preferences and Constraints</td>
<td>202</td>
</tr>
<tr>
<td>5.4.2.1</td>
<td>Local Attachments are Preferred</td>
<td>202</td>
</tr>
<tr>
<td>5.4.2.2</td>
<td>Scrambling and Case</td>
<td>203</td>
</tr>
<tr>
<td>5.4.2.3</td>
<td>Scrambling from Complement Clauses</td>
<td>203</td>
</tr>
<tr>
<td>5.4.2.4</td>
<td>No Scrambling out of Adjuncts</td>
<td>203</td>
</tr>
<tr>
<td>5.4.2.5</td>
<td>Scrambling from Deep Within</td>
<td>204</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Summary</td>
<td>205</td>
</tr>
<tr>
<td>5.5</td>
<td>The Locative in Locative Clauses</td>
<td>205</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Constituent Properties</td>
<td>207</td>
</tr>
<tr>
<td>5.5.1.1</td>
<td>Coordination</td>
<td>207</td>
</tr>
<tr>
<td>5.5.1.2</td>
<td>Linear Order</td>
<td>209</td>
</tr>
<tr>
<td>5.5.1.3</td>
<td>Summary</td>
<td>213</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Functional Properties</td>
<td>213</td>
</tr>
<tr>
<td>5.5.2.1</td>
<td>Reflexive Binding</td>
<td>213</td>
</tr>
<tr>
<td>5.5.2.2</td>
<td>Pronominal Coreference</td>
<td>215</td>
</tr>
<tr>
<td>5.5.2.3</td>
<td>Control</td>
<td>217</td>
</tr>
<tr>
<td>5.5.2.4</td>
<td>Summary</td>
<td>219</td>
</tr>
<tr>
<td>5.6</td>
<td>Information Structure</td>
<td>219</td>
</tr>
<tr>
<td>5.7</td>
<td>Conclusion</td>
<td>222</td>
</tr>
</tbody>
</table>

6 LFG Analysis & XLE Implementation of the Locative & Instrumental | 225
| 6.1 | Introduction | 225 |
| 6.2 | Locative Inversion | 225 |
| 6.2.1 | Theoretical Explanation | 226 |
| 6.2.2 | Computational Implementation | 228 |
| 6.3 | The Copula Analysis | 229 |
| 6.3.1 | The Single-Tier Analysis | 230 |
| 6.3.2 | The Double-Tier Analyses | 232 |
| 6.3.2.1 | The XCOMP Analysis | 233 |
| 6.3.2.2 | The PREDLINK Analysis | 237 |
| 6.3.3 | Choosing an Analysis for Hindi/Urdu Locative Clauses | 240 |
| 6.3.3.1 | Agreement | 240 |
| 6.3.3.2 | Predicate Categories | 241 |
| 6.3.3.3 | Cross-Linguistic Appeal | 241 |
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>LFG’s correspondence architecture</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>Simplified c-structure</td>
<td>31</td>
</tr>
<tr>
<td>2.3</td>
<td>C-structure with f-structure annotations</td>
<td>33</td>
</tr>
<tr>
<td>2.4</td>
<td>Partial f-structure</td>
<td>34</td>
</tr>
<tr>
<td>2.5</td>
<td>Partial f-structure</td>
<td>34</td>
</tr>
<tr>
<td>2.6</td>
<td>Unified f-structure</td>
<td>34</td>
</tr>
<tr>
<td>2.7</td>
<td>XLE parsing architecture (adapted from Butt et al., 1999b)</td>
<td>39</td>
</tr>
<tr>
<td>2.8</td>
<td>Packed f-structure example</td>
<td>42</td>
</tr>
<tr>
<td>2.9</td>
<td>ParGram f-structure for English sentence</td>
<td>53</td>
</tr>
<tr>
<td>2.10</td>
<td>ParGram f-structure for German sentence</td>
<td>53</td>
</tr>
<tr>
<td>2.11</td>
<td>ParGram f-structure for German possessive</td>
<td>60</td>
</tr>
<tr>
<td>2.12</td>
<td>HPSG derivation tree for Carl’s son laughed</td>
<td>63</td>
</tr>
<tr>
<td>2.13</td>
<td>HPSG derivation tree for The son laughed</td>
<td>64</td>
</tr>
<tr>
<td>2.14</td>
<td>Sublexical c-structure for Hindi/Urdu proper noun</td>
<td>68</td>
</tr>
<tr>
<td>2.15</td>
<td>F-structure for Hindi/Urdu proper noun</td>
<td>69</td>
</tr>
<tr>
<td>2.16</td>
<td>Sublexical c- and f-structure for Hindi/Urdu personal pronoun</td>
<td>70</td>
</tr>
<tr>
<td>2.17</td>
<td>Hindi/Urdu c- and f-structure for KP larke = ne ‘boy.m.sg.OBL = erg’</td>
<td>72</td>
</tr>
<tr>
<td>2.18</td>
<td>Hindi/Urdu c- and f-structure for KP larke ‘boy.m.sg’</td>
<td>72</td>
</tr>
<tr>
<td>2.19</td>
<td>Hindi/Urdu c- and f-structure for (92)</td>
<td>74</td>
</tr>
<tr>
<td>3.1</td>
<td>POSS function for possessors</td>
<td>85</td>
</tr>
<tr>
<td>3.2</td>
<td>SUBJ function for English possessors</td>
<td>87</td>
</tr>
<tr>
<td>3.3</td>
<td>ADNOM function for English possessors</td>
<td>88</td>
</tr>
<tr>
<td>3.4</td>
<td>SUBJ function for Hungarian possessors</td>
<td>89</td>
</tr>
<tr>
<td>3.5</td>
<td>ADNOM function for Hungarian possessors (Chisarik and Payne, 2003)</td>
<td>89</td>
</tr>
<tr>
<td>3.6</td>
<td>POSS function for Welsh possessors</td>
<td>91</td>
</tr>
<tr>
<td>3.7</td>
<td>Examples for f-command relations</td>
<td>94</td>
</tr>
<tr>
<td>3.8</td>
<td>C- and f-structure for Low Saxon possessive pronoun construction</td>
<td>120</td>
</tr>
<tr>
<td>3.9</td>
<td>C- and f-structure for Low Saxon possessive pronoun construction</td>
<td>121</td>
</tr>
<tr>
<td>3.10</td>
<td>F-command relations in possessive clauses</td>
<td>140</td>
</tr>
<tr>
<td>3.11</td>
<td>Pronominal coreference in possessive clauses, analysis 1</td>
<td>141</td>
</tr>
<tr>
<td>3.12</td>
<td>Pronominal coreference in possessive clauses, analysis 2</td>
<td>141</td>
</tr>
</tbody>
</table>
3.13 Full clause f-structure, reflexive binding .............................. 146
4.1 Hindi/Urdu NP c-structure, common noun head ....................... 159
4.2 Hindi/Urdu NP f-structures, common noun head ....................... 160
4.3 Hindi/Urdu NP c-structure, relational noun head ....................... 160
4.4 Hindi/Urdu NP f-structures, relational noun head ....................... 161
4.5 Hindi/Urdu NP c-structure, event noun head I .......................... 162
4.6 Hindi/Urdu NP f-structures, event noun head I .......................... 163
4.7 Hindi/Urdu NP c-structure, event noun head II .......................... 164
4.8 Hindi/Urdu NP f-structure, event noun head II .......................... 164
4.9 Hindi/Urdu NP f-structure, animate noun head ......................... 166
4.10 Hindi/Urdu c-structure, genitive scrambling ........................... 169
4.11 Hindi/Urdu f-structure, genitive scrambling ........................... 169
4.12 Hindi/Urdu c-structure, possessive clause ............................. 171
4.13 Hindi/Urdu f-structure, possessive clause ............................. 171
4.14 Hindi/Urdu c-structures, possessive clause ............................ 172
4.15 Hindi/Urdu f-structure, possessive clause ............................ 173
4.16 Hindi/Urdu f-structure, possessive clause ............................ 173
4.17 ParGram f-structure for German possessive .......................... 174
6.1 Linking analysis for predicative locatives ............................... 226
6.2 F-structure for (2) .................................................. 227
6.3 Optional classification for locative inversion (Kibort, 2007) ......... 227
6.4 Linking analysis for inverted locatives ................................. 227
6.5 F-structure for (3) .................................................. 228
6.6 Copula construction analyses in LFG ................................ 229
6.7 Single-tier analysis for Japanese (Dalrymple et al., 2004a, p. 191) 230
6.8 Open double-tier analysis of French copula ............................. 234
6.9 Open double-tier analysis of French raising verb ....................... 235
6.10 Clashing open double-tier analysis with divergent subjects ........ 236
6.11 (Simplified) f-structure for (19) (Butt et al., 1999b, p. 70) .......... 237
6.12 Closed double-tier analysis of French copula .......................... 238
6.13 Well-formed closed double-tier analysis of She seems in a bad mood 239
6.14 Closed double-tier analysis without surface copula in Arabic ....... 240
6.15 Hindi/Urdu NP c- and f-structure, picture noun ...................... 246
6.16 Hindi/Urdu NP c-structure, relational noun ........................... 248
6.17 Hindi/Urdu NP f-structure, relational noun ........................... 248
6.18 Canonical locative clause c-structure ................................. 252
6.19 Canonical locative clause f-structure ................................. 253
6.20 Inverted locative clause c-structure ................................. 253
6.21 Inverted locative clause f-structure ................................. 254
7.1 Constituency of N-V CPs (Mohanan, 1994, p. 203) .............................. 258
7.2 Maximal projections in N-V CPs (Mohanan, 1994, p. 211) ...................... 259
7.3 Simple restriction example, adapted from Dalrymple (2001), p. 162 .... 270
7.4 First 4 lines of harvested raw text Urdu corpus .................................... 275
7.5 Visualization of the relative frequencies in Table 7.2 ............................ 281
7.6 Visualization of the top of Cluster 2 (left) and Cluster 0 (right) .......... 282
7.7 Visually prominent pattern of a false positive ........................................ 282
7.8 Visualization of the top of Cluster 0 ...................................................... 283
7.9 Choosing the best value for k ............................................................... 292
7.10 Visualization for k = 5 clusters .......................................................... 293
7.11 C-structure for N-V CP ................................................................. 299
7.12 F-structure for N-V CP ................................................................. 299
7.13 Incorrect c-structure for (38a) ................................................................. 300
7.14 Incorrect f-structure for (38a) ................................................................. 301
# List of Tables

2.1 Overview of Hindi/Urdu case clitics ................................. 23  
2.2 Possible inflections on Hindi/Urdu nominals ......................... 23  
2.3 Identification of GFS in LFG by means of $\pm \sigma$ and $\pm r$ .......... 35  
2.4 LFG vis-à-vis XLE notations ........................................... 41  
3.1 Possible inflections of Hindi/Urdu genitive case clitic $k$- .............. 79  
3.2 Possessive pronouns in Hindi/Urdu ..................................... 83  
3.3 Genitive KP types in the Hindi/Urdu NP ................................. 124  
4.1 Hindi/Urdu noun type annotation ....................................... 155  
4.2 Hindi/Urdu sublexical categories for nouns ............................. 158  
5.1 Overview of Hindi/Urdu locative and instrumental case clitics ........... 178  
5.2 Locative/instrumental KP types in the Hindi/Urdu NP .................. 199  
6.1 Copula analysis types in LFG ............................................. 242  
6.2 Locative/instrumental KP overview with example head nominals ...... 247  
7.1 Classes of nouns identified by Ahmed and Butt (2011) .................... 263  
7.2 Relative frequencies of co-occurrence of nouns with light verbs ........ 277  
7.3 Overview of the results — Corpus Study 1 ................................ 283  
7.4 Size distribution in Hindi corpus used ................................. 286  
7.5 POS tagged Hindi corpus excerpt ....................................... 287  
7.6 Hindi seed noun list excerpt ............................................. 288  
7.7 Hindi bigram matches: absolute frequencies ........................... 290  
7.8 Hindi bigram matches: relative frequencies ........................... 290  
7.9 Feature value examples for clustering ..................................... 291  
7.10 Semantic coherence values for $k = 5, \ldots, 9$ for clustering algorithms GVM and $k$-means .................................................. 292  
7.11 Feature values at “semantic centroids” for $k = 5$ ....................... 294  
7.12 Relative frequencies for all clusters for $k = 5$ .......................... 295
# List of Abbreviations

The list of glossing abbreviations used in this thesis (in alphabetical order):

<table>
<thead>
<tr>
<th>Gloss</th>
<th>Meaning</th>
<th>Gloss</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First person</td>
<td>M</td>
<td>Masculine gender</td>
</tr>
<tr>
<td>2</td>
<td>Second person</td>
<td>N</td>
<td>Neuter gender</td>
</tr>
<tr>
<td>3</td>
<td>Third person</td>
<td>NEG</td>
<td>Negation</td>
</tr>
<tr>
<td>ACC</td>
<td>Accusative case</td>
<td>NOM</td>
<td>Nominative case</td>
</tr>
<tr>
<td>ART</td>
<td>Article</td>
<td>OBL</td>
<td>Oblique</td>
</tr>
<tr>
<td>CAUS</td>
<td>Causative</td>
<td>PART</td>
<td>Participle</td>
</tr>
<tr>
<td>DAT</td>
<td>Dative case</td>
<td>PAST</td>
<td>Past tense</td>
</tr>
<tr>
<td>DECL</td>
<td>Declarative</td>
<td>PERF</td>
<td>Perfective aspect</td>
</tr>
<tr>
<td>ERG</td>
<td>Ergative case</td>
<td>PL</td>
<td>Plural number</td>
</tr>
<tr>
<td>F</td>
<td>Feminine gender</td>
<td>POSS</td>
<td>Possessive pronoun</td>
</tr>
<tr>
<td>FOC</td>
<td>Focus</td>
<td>PRES</td>
<td>Present tense</td>
</tr>
<tr>
<td>FUT</td>
<td>Future tense</td>
<td>PRO</td>
<td>Non-overt pronominal</td>
</tr>
<tr>
<td>GEN</td>
<td>Genitive case</td>
<td>PROG</td>
<td>Progressive aspect</td>
</tr>
<tr>
<td>IMP</td>
<td>Imperative</td>
<td>PRON</td>
<td>Pronoun</td>
</tr>
<tr>
<td>IMPF</td>
<td>Imperfective aspect</td>
<td>SBJ</td>
<td>Subject</td>
</tr>
<tr>
<td>INF</td>
<td>Infinitive</td>
<td>SBJV</td>
<td>Subjunctive</td>
</tr>
<tr>
<td>INST</td>
<td>Instrumental case</td>
<td>SG</td>
<td>Singular number</td>
</tr>
<tr>
<td>LOC</td>
<td>Locative case</td>
<td>TOP</td>
<td>Topic</td>
</tr>
</tbody>
</table>

1.1 INTRODUCTION

The identification and classification of nominal predicates and their arguments is a notorious problem in natural language processing (NLP). Semantic reasoning, information retrieval, question answering and other applications can benefit greatly from a successful treatment of nominal predication (Meyers et al., 2004b). Overall, a considerable amount of work in NLP focuses on the identification and annotation of verbal predication and arguments thereof, and there is less research on types and identification of nominal predicates and their arguments. This thesis is a contribution of the latter type. It focuses on the description and analysis of nominal arguments in the South Asian language Hindi/Urdu. The analysis is couched within the theory of Lexical-Functional Grammar (LFG, Bresnan, 2001, Dalrymple, 2001) and is implemented in a computational grammar, the Urdu ParGram grammar (Butt and King, 2007), which is a part of the ParGram (“Parallel Grammar”) project on parallel LFG grammar engineering (Butt et al., 1999a, 2002, 1999b). The implementation makes use of the grammar development platform XLE (Crouch et al., 2015).

Most techniques that aim to deal with nominal arguments use manual annotation of a corpus; the annotated resource can then be used to derive features for the automatic identification of nominal arguments in unseen texts (e.g., Meyers et al., 2004a). Such an approach has several requirements. First, substantial time and effort have to be spent on the annotation of the corpus. Second, the approach works best if there are enough features that can discriminate different types of nominal arguments based on morphological or syntactic criteria, e.g., case marking or word order. Third, in many languages,
nominal arguments do not have to occur next to the noun that selects them; often, how-
ever, the arguments are only annotated as such if occurring next to their selecting noun.
A classifier that is trained on a resource where argument and predicate noun are always
adjacent will obviously perform poorly on non-adjacent dependencies. Compare this to
the annotation of verbal arguments, where it is obvious that they may not occur next to
their predicator.

The approach described in this thesis uses a linguistically well-motivated computa-
tional grammar that features detailed functional annotation and a range of efficient LFG-
specific instruments to automatically identify and correctly analyze nominal arguments.
Different types of nominal arguments in Hindi/Urdu are discussed in detail: genitive (i.e.,
arguments bearing genitive case), locative as well as instrumental arguments. For all of
these argument types, I discuss their morphosyntactic properties as well as their func-
tional behavior. A focus is on the varied realization of nominal arguments as different
grammatical functions (GF) in the NP, which has not been described in previous litera-
ture. As a further contribution, noun predicates that form part of noun-verb complex
predicates (N-V CPS) are examined in detail. Here, the focus is on the empirical prop-
erties of these nouns, in particular their combinatorial possibilities with different light
verbs. Finally, I develop a treatment in terms of a large-scale computational grammar,
the Urdu ParGram grammar (Butt and King, 2007), which encodes the (c)onstituent-
and f(unctional)-structure of nominal predication. The implementation pays tribute to
two separate findings. First, genitive and locative/instrumental nominal arguments oc-
cupy different syntactic functions, an observation which is crucial in order to explain
their syntactic behavior. Second, regarding the N-V CPS, the implementation makes use
of an empirically-motivated system that models the statistical preferences of nominal
predicates to combine with light verbs.

1.2 Nominal Predicates

It is a widely accepted view that not all nominals can take arguments. For example, it is
generally acknowledged that proper nouns do not select arguments, while nouns derived
from verbs select arguments based on the verbs from which they are derived (Chomsky,
1970). Another noun class that is generally assumed to take arguments is the class
Common nouns are usually represented as a mixed class where some take arguments
and others don’t. For example, Meyers et al. (2004a), developing the computational
resource NomBank for English, cite a list of 17 different common noun classes taking
arguments and note that common nouns not belonging to any of these classes do not
take arguments.

The task of defining what exactly constitutes a nominal argument is not straightforward.
The traditional distinction known from verbal predication between arguments
and adjuncts holds only to a certain degree. The distinction indicates that arguments
are those phrases that have to be realized in order to satisfy the verb’s valency requirements, while adjuncts are entirely optional. That is, arguments are a part of the lexical semantics of the predicate, and generally have to be realized overtly in the syntax for the clause to be felicitous. Verbs are generally taken to be more restrictive than nouns with respect to the realization of their arguments (e.g., Grimshaw, 1990). For example, transitive verbs require exactly two arguments, while intransitive verbs require exactly one argument. The English transitive verb destroy requires two arguments, a subject (the agent of the event) and an object (the patient of the event). The nominal predicate destruction features the same two arguments, which can reasonably be argued to be a direct result of it being derived from destroy; nevertheless, an NP headed by destruction may feature both (1a–b) or only a single one of its arguments (1c–d). (1e), in which the NP features no arguments at all, is also grammatical in certain contexts: (1c) is, for example, perfectly acceptable in a context where it is already established what was destroyed by whom.

(1)  a. the city’s destruction by the Romans
    b. the Romans’ destruction of the city
    c. the city’s destruction/the destruction of the city
    d. the Romans’ destruction/the destruction by the Romans
    e. The destruction took ages.

Thus, a central issue in the identification of nominal arguments is that it is often hard to distinguish them from adjuncts. An approach that puts nominal arguments on a par with verbal arguments misses out on the fact that in many cases, nominal arguments are not realized. In fact, this is an observation that turns out to be true across languages (Grimshaw, 1990, Rappaport, 2006, among others).

1.3 NOMINAL ARGUMENTS IN HINDI/URDU

With respect to Hindi/Urdu, the distinction between nominal arguments and nominal adjuncts is even more blurred. First, event nouns as well as relational nouns come with lexically-specified arguments. These are realized with either genitive, locative or instrumental case-marked complements. The NPs in (2) are examples of genitive complements modifying event nominals. (3) illustrates the use of genitive complements with relational nouns. Note that there is a small class of event nouns in Hindi/Urdu that may occur with two distinct genitive arguments, as in (2b). As will be discussed in Chapter §3, the Hindi/Urdu genitive case exhibits an agreement pattern whereby it agrees in gender, number and morphological case form with the head nominal.

Languages with pronoun-dropping form an exception: dropped pronominals also qualify as arguments, even though they are not realized overtly.
(2)  
a. ʃɘhɑr=ki  
city.M.SG = GEN.F.SG destruction.F.SG  
‘the city’s destruction’

b. ʃɘhɑr=ki tɑbahi  
city.M.SG = GEN.F.SG destruction.F.SG  
‘the destruction of the city’

(3)  
a. ʃɘhɑr=ki taβʰɐ  
city.M.SG = GEN.F.SG destruction.F.SG  
‘the destruction of the city’

b. ʃɘhɑr=ki taβʰɐ  
city.M.SG = GEN.F.SG destruction.F.SG  
‘the destruction of the city by youngsters’

(4)  
a. ʃɘhɑr=ki  
city.M.SG = GEN.F.SG destruction.F.SG  
‘the city’s destruction’

b. ʃɘhɑr=ki  
city.M.SG = GEN.F.SG destruction.F.SG  
‘the destruction of the city by youngsters’

(5)  
a. ʃɘhɑr=ki  
city.M.SG = GEN.F.SG destruction.F.SG  
‘the city’s destruction’

b. ʃɘhɑr=ki  
city.M.SG = GEN.F.SG destruction.F.SG  
‘the destruction of the city by youngsters’

(4a,b) illustrates the use of locative case-marked arguments on event (4a) and relational nouns (4b). There may be at most one locative complement for such nouns, and my survey shows that there are indeed no Hindi/Urdu nouns allowing for more than one locative case-marked argument. The survey also shows that the usage of locative arguments in general is more restricted and occurs with a smaller number of event and relational nouns. The same applies to instrumental arguments, two examples of which are shown in (5).

(5a,b) illustrates the use of locative case-marked arguments on event (5a) and relational nouns (5b). There may be at most one locative complement for such nouns, and my survey shows that there are indeed no Hindi/Urdu nouns allowing for more than one locative case-marked argument. The survey also shows that the usage of locative arguments in general is more restricted and occurs with a smaller number of event and relational nouns. The same applies to instrumental arguments, two examples of which are shown in (5).

Hindi/Urdu event and relational nouns also occur without any arguments, as in the example in (6). Following proposals by Barker (1995) and (in part) Laczkó (2009), I argue that argument suppression occurs with such nouns; here, the suppressed argument, although not overtly realized, is existentially bound in the argument structure of the nominal, with the effect that even though the argument(s) are not realized syntactically, their existence is still implied in the lexical semantics of the noun. (6) shows argument suppression for an event nominal, but the operation extends to relational nouns as well.
1.3. NOMINAL ARGUMENTS IN HINDI/URDU

(6) tabahi ek hafte = tak rah-i
destruction.F.SG.NOM one week.M.SG.OBL = LOC.TO remain-PERF.F.SG
‘The (city’s/Its) destruction lasted for a week.’

I apply diagnostics showing that the genitive arguments are either subjects or objects, while the locative and instrumental arguments map onto thematically-restricted obliques. Previous analyses either have not focused on distinguishing these different types of nominal arguments, or have not worked out the correct generalizations in terms of grammatical-functional assignment. The distinction is important to draw since all of these argument types occur on a very regular basis in the language and warrant a thorough and empirically supported treatment.

A complication is that all types of nouns can further be modified by genitive case-marked phrases that show the distinct behavior of adjuncts, as in (7).

(7) lakṛi = ki mez
wood.F.SG = GEN.F.SG table.F.SG
‘table made of wood’

Locative and instrumental phrases, on the other hand, never occur as adjuncts inside NPs, contrary to what can be observed in other languages such as English or German. That is, the Hindi/Urdu NPs in (8) are ungrammatical, while their English and German counterparts in (9–10) are grammatical.

(8) a. * [mez = par kitab]NP
table.F.SG = LOC.ON book.F.SG
‘book on the table’

b. * [mukan = mē cuha]NP
house.M.SG = LOC.IN rat.M.SG
‘rat in the house’

(9) a. the rat in the house

b. the book on the table

(10) a. die Ratte im Haus
the.F.SG.NOM rat.F.SG.NOM in.the.N.SG.DAT house.N.SG
‘the rat in the house’

b. das Buch auf dem Tisch
‘the book on the table’

The picture to be painted of nominal arguments in Hindi/Urdu is even more complex, though. Common nouns, which are usually not assumed to predicate, do occur with both the aforementioned genitive (11) and locative arguments (12). That is, while arguments are not specified in the lexical semantics of such nouns, they appear with both genitive as well as locative complements that behave like subjects and obliques, respectively, as
indicated by several diagnostics. An example of a common noun modified by a genitive argument is given in (11a). The semantics of such genitive phrases are not constrained to a purely possessive reading, but extend to different thematic relations; for example, (11a,b) can also denote a creator/creation relation. The behavior of these phrases is symptomatic of what Barker (1995) describes as extrinsic possession. Syntactic tests indicate that the genitive phrases indeed behave like subjects. I argue that Hindi/Urdu makes use of a lexical rule that augments common nouns enabling them to select subjects, in a fashion similar to what was described for English by Bresnan (2001) and for Hungarian by Chisarik and Payne (2003).

(11) a. nina=ka makan
   Nina.F.SG = GEN.M.SG house.M.SG
   ‘Nina’s house’ (e.g., the house constructed by, owned by Nina)

   b. nina=ki kitab
   Nina.F.SG = GEN.F.SG book.F.SG
   ‘Nina’s book’ (e.g., the book written by, owned by Nina)

Examples such as (12) indicate that there are in fact common nouns that select a thematically-restricted oblique realized by a locative phrase. Closer examination of this class reveals that it correlates with the picture noun class well-known in the syntactic literature on English (e.g., Ross, 1967, Soames and Perlmutter, 1979). It turns out that picture nouns in Hindi/Urdu place tight restrictions on the case of their depictum complement; specifically, they require their depictum complement to bear the locative case marker pɑr ‘on’. Genitive complements occurring with picture nouns, as was shown in (11b), are never interpreted as depictum obliques, but as subjects or adjuncts.

(12) a. mɑntiq=pɑr kitab
    logic.M.SG = LOC.ON book.F.SG
    ‘book about logic’

   b. gandʰi=pɑr film
    Gandhi.M.SG = LOC.ON film.F.SG
    ‘film about Gandhi’

Of course, all common nouns may also be modified by genitive phrases that must be treated like adjuncts. In fact, there is no overall syntactic constraint on the number of genitive phrases that can occur inside an np, contrary to what is known from languages such as English or German. (13) contains two examples of common nouns modified by genitive adjuncts.

(13) a. lɑkṛi=ki surx rang=ki mez
    ‘red table made of wood’
b. do sal=ki surx rang=ki kitab
‘two year old red book’

1.4 Constraining Nominal Arguments

Regarding the genitive phrases, their different functional behavior results in a large number of GF assignments and, ultimately, in a significant amount of ambiguity.\(^3\) For an example from Hindi/Urdu illustrating the issue, consider (14). Given my account of genitive functional assignment in Hindi/Urdu, the NP in (14) yields a number of analyses. The head nominal mez ‘table’ is a common noun that may be optionally augmented to subcategorize for a SUBJ, given what was seen above. Other than that, the nominal may be modified by attributive genitives that are adjuncts. Since the order of the individual genitive types is free within the NP (see the discussion below), the above NP receives a total of four different analyses, as indicated in (14b–e).

(14)

a. makan=ki lakri=ki surx rang=ki
mez
table.F.SG
‘the house’s red table made of wood’
b. [makan=ki]\text{\textsubscript{\textsc{adjunct}}} [lakri=ki]\text{\textsubscript{\textsc{adjunct}}} [surx rang=ki]\text{\textsubscript{\textsc{adjunct}}} mez
c. [makan=ki]\text{\textsubscript{\textsc{subj}}} [lakri=ki]\text{\textsubscript{\textsc{adjunct}}} [surx rang=ki]\text{\textsubscript{\textsc{adjunct}}} mez
d. [makan=ki]\text{\textsubscript{\textsc{adjunct}}} [lakri=ki]\text{\textsubscript{\textsc{subj}}} [surx rang=ki]\text{\textsubscript{\textsc{adjunct}}} mez
e. [makan=ki]\text{\textsubscript{\textsc{adjunct}}} [lakri=ki]\text{\textsubscript{\textsc{adjunct}}} [surx rang=ki]\text{\textsubscript{\textsc{subj}}} mez

Disambiguating between (14b–e) is not a matter of syntax, but belongs in the domain of (lexical) semantics. Given the identical morphosyntactic realization of argument and adjunct genitives, syntax cannot know whether makan=ki is more likely than lakri=ki to be an argument of mez; compare this to languages like English, for example, where there are two distinct realizations of possessive relations that differ widely in the semantic roles they may encode (see, for example, Adger (2013) as well as the works collected in Börjars et al. (2013)). Likewise, syntax cannot know whether makan=ki can or cannot in principle be an argument of mez, and it cannot know all the adjuncts that might modify mez. Knowledge like this is indeed part of large-scale lexical ontologies such as WordNet.

\(^2\)The discussion in this paragraph excludes the locative/instrumental complements, since they uniformly map onto a thematically restricted oblique GF and thus ambiguity is not an issue in their analysis. See the discussion in §1.3 above.
1.5. THE LINEAR ORDER OF NOMINAL ARGUMENTS

(Adeeba and Hussain, 2011, Ahmed and Hautli, 2010, Bhattacharyya, 2010, Fellbaum, 1998), and listing such information in the lexicon of a computational grammar is both conceptually incorrect as well as inefficient.

The possibilities of nominal predication are far from “everything goes”, though. It was already mentioned above that the usage of locative and instrumental arguments is confined to the realization of thematically-restricted obliques, and that all nouns that allow such arguments are either event or relational nouns or members of the picture noun class. Moreover, even though the rules of genitive selection seem much less constrained, there are in fact combinations that are impossible. This can be captured by making use of the detailed feature system of the Urdu ParGram grammar (King et al., 2005). The following restrictions apply to genitive arguments:

- When a genitive argument appears with an animate head noun, the genitive argument itself must also be animate.
- Proper nouns and pronouns never take arguments.
- Pronouns are excluded from modification by genitive adjuncts.
- Genitive adjuncts may appear with proper nouns, but this configuration is dispreferred if other readings are possible in the given context.

When designing a computational grammar that can efficiently parse nominal predicates and their arguments, it is crucial to implement strategies that reduce the ambiguity in genitive argument insertion. The implementation I propose makes use of features that encode animacy as well as nominal type; here, I agree with recent work that encoding animacy as a lexical property can help syntactic parsing (e.g., Jena et al., 2013).

1.5 THE LINEAR ORDER OF NOMINAL ARGUMENTS

Regarding the word-order properties of genitive and locative/instrumental nominal arguments, it is shown that NP-internally, the genitive and locative/instrumental arguments as well as the genitive adjuncts occur in a variety of positions in relation to other constituents within the NP. However, when occurring NP-internally, they do not occur after the NP head; this is in compliance with the general head-final configuration of Hindi/Urdu. Nominal arguments and adjuncts can, in addition, freely be scrambled outside of the NP they are licensed in. Consider the examples in (15–16). In the b. examples, the genitive arguments have been scrambled to the right, outside of the NP they are licensed in.

(15) a. ram = ka dɔst ay-a
Ram.M.SG = GEN.M.SG friend.M.SG.NOM come-PERF.M.SG
‘Ram’s friend came.’ (Butt and Zinsmeister, 2009)
1.5. THE LINEAR ORDER OF NOMINAL ARGUMENTS

b. dost ay-a ram = ka
friend.M.SG.NOM come-PERF.M.SG Ram.M.SG = GEN.M.SG
‘Ram’s friend came.’  
(Butt and Zinsmeister, 2009)

(16) a. garį nadya = ne us = ki
car.F.SG.NOM Nadya.F.SG.OBL = ERG PRON.3.SG.OBL = GEN.F.SG.NOM
bazar = mê dekʰ-i
market.M.SG.OBL = LOC.IN see-PERF.F.SG
‘His/her car, Nadya saw in the market.’  
(Bögel and Butt, 2013, p. 301)
b. kitab tum = ne kis = ki xarid-i?
book.F.SG.NOM you = ERG who.SG.OBL = GEN.F.SG buy-PERF.F.SG
‘Whose book did you buy?’  
(Bögel and Butt, 2013, p. 301)

Locative/instrumental scrambling is also attested in the literature. In (17), for example, there is a locative phrase that needs to be analyzed as an argument of kitab ‘book’. Such examples are mentioned by Raza and Ahmed (2011a), who do not, however, provide an analysis of nominal argument scrambling outside of the NP.

(17) a. nida = ne məntiq = par ek kitab xarid-i
Nida.F.SG = ERG logic.M.SG = LOC.ON one book.F.SG.NOM buy-PERF.F.SG
hɛ
be.PRES.3.SG
‘Nida has purchased a book on logic.’  
(Raza and Ahmed, 2011a)
b. məntiq = par nida = ne ek kitab xarid-i
logic.M.SG = LOC.ON Nida.F.SG = ERG one book.F.SG.NOM buy-PERF.F.SG
hɛ
be.PRES.3.SG
‘Nida has purchased a book on logic.’  
(Raza and Ahmed, 2011a)

I discuss nominal argument scrambling in detail and compare it to other types of scrambling in Hindi/Urdu, such as quantifier float and relative clause extraction. Given the similarities between these operations across languages, it is in fact not surprising that nominal argument scrambling occurs, especially given the free word order of Hindi/Urdu.

To deal with the linear order of nominal arguments, I invoke two separate techniques. First, I show how the LFG shuffle operator (discussed by e.g., Dalrymple, 2001) can efficiently model the free word order inside the Hindi/Urdu NP. In addition, the f-precedence operator, implemented in XLE as head-precedence, ensures that the nominal arguments may not occur to the right of their nominal heads. Second, in order to deal with scrambled nominal arguments, I define functional uncertainty paths (Kaplan and Zaenen, 1995) that correctly assign the genitive GFS to their heads.
1.6 POSSESSIVE AND LOCATIVE CLAUSES

A fundamental difference between genitive and locative arguments is their predicational context: all types of nouns may be complemented by genitive arguments (whether this process is driven by the nouns’ lexical properties or by a form of augmentation via lexical rules), while there is only a small class of nouns that may be complemented by locative arguments (the event, relational and picture nouns discussed above in §1.3). In Chapter §3, I develop a view that treats the genitive as a strictly nominal case; that is, genitive phrases are always licensed inside a nominal domain (as subjects, adjuncts or obliques). Locatives, on the other hand, may be licensed inside a nominal domain, if the noun’s lexical entry permits such a licensing (or if the noun can be augmented to select a locative, in the case of picture nouns). Nouns that do not allow locative arguments, for example common nouns to be located somewhere in space, need a verbal element to realize the locative (Chapter §5).

There are a couple of consequences that directly follow from this view. First, if it is true that the genitive is a strictly nominal case and is licensed exclusively in the nominal domain, I need to provide an explanation of why certain genitive phrases seem to behave like clausal subjects in what I will refer to as possessive clauses. Two examples are given in (18). Mohanan (1994), for example, has argued that the genitive phrases in such clauses should be treated as the subjects of the sentences. However, it turns out that the relevant genitives in examples such as (18) fail one of the clausal subject tests proposed in the literature for Hindi/Urdu. I therefore reanalyze possessive clauses as one-place predicates (“existentials”): a complex NP selected by an intransitive verb ho ‘be’. The genitive is not a clausal subject, but occupies a GF inside the NP domain.

(18)

a. ram = ka mɑkan hɛ
   Ram.M.SG = GEN.M.SG house.M.SG.NOM be.PRES.3.SG
   ‘Of Ram is a house.’ = ‘Ram has a house.’

b. tale = ki cabi nɑhĩ hɛ
   lock.M.SG.OBL = GEN.F.SG key.F.SG.NOM not be.PRES.3.SG
   ‘Of the lock is not a key.’ = ‘The lock has no key.’

Second, regarding locative clauses, the above assumptions again make specific predictions. When locatives are licensed inside the nominal domain, locative clauses are predicted to predicate as existentials, like the possessive clauses in (18). Examples of this type of locative clause are given in (19). When the locatives are not licensed inside the nominal domain (as in the case of common nouns located in space), the clauses must predicate in a different way. Examples of this type of locative clause are shown in (20).

(19)

a. mɑntiq = par kitab hɛ
   logic.M.SG = LOC.ON book.F.SG be.PRES.3.SG
   ‘About logic is a book.’
b. nina=par  bʰarɔsa  he
Nina.F.SG = LOC.ON trust.M.SG be.PRES.3.SG
‘In Nina is trust.’

(20) a. mez=par  cuha  he
table.F.SG = LOC.ON rat.M.SG be.PRES.3.SG
‘On the table is a rat.’ = ‘There is a rat on the table.’
b. kɑmre=mẽ  admi  he
room.M.SG.OBL = LOC.IN man.M.SG be.PRES.3.SG
‘In the room is a man.’ = ‘There is a man in the room.’

Both of these predictions are in fact borne out by evidence collected in this thesis. It turns out that the locatives in (19) do not behave as clausal subjects, but must be treated as obliques embedded inside an NP. This contrasts with the behavior of the locatives in (20), which pass all subject tests, and hence must be treated as true locative subjects.

1.7 NOUN-VERB COMPLEX PREDICATES

In Hindi/Urdu, nominal predicators are not confined to the nominal domain (i.e., to NPs), but may also take part in what is known as complex predication (CP). When combined with a verb in noun-verb complex predicates (N-V cp), the noun and the verb form a single predicational unit. The verb, which is usually referred to as the light verb, dictates the case marking on the subject, determines agreement patterns, carries information about tense and aspect, and adds information about agentivity vs. experiencer predicates (i.e., whether the subject carries an agent or an experiencer role in the predication); however, it features impoverished lexical semantics (Butt, 2003, 2010). The noun, on the other hand, is the main predicational element of the construction, featuring fully-fledged lexical semantics and argument structure (Ahmed and Butt, 2011, Butt et al., 2012, Mohanan, 1994).

In example (21) below (taken from Butt et al. (2012)), it can be seen that a single noun may be combined with different light verbs, giving rise to different semantics. The noun involved in the N-V CP is yad ‘memory’. The light verb may range from kar ‘do’ (21a) to ho ‘be’ (21b) to hu- ‘become’ (21c).

(21) a. lɑṛki=ne  kohani  yad  k-i
girl.F.SG = ERG story.F.SG memory.F.SG do-PERF.F.SG
‘The girl remembered a/the story.’ (lit. ‘The girl did memory of a/the story.’)
b. lɑṛki=ko  kohani  yad  he
girl.F.SG = DAT story.F.SG memory.F.SG be.PRES.3.SG
‘The girl remembers/knows a/the story.’ (lit. ‘Memory of a/the story is at the girl.’)
c. lɑṛki=ko kahani yad hu-i
       girl.F.SG = DAT story.F.SG memory.F.SG be.PART-PERF.F.SG
       ‘The girl came to remember a/the story.’ (lit. ‘Memory of a/the story became to be at the girl.’)

In all of the examples above, the noun and the verb form a single predicational element. kahani ‘story’ is thematically licensed by the noun yad ‘memory’, but it is not realized as a genitive, locative or instrumental, as would be typical for arguments of nouns. Rather, kahani functions as an object of the joint predication. In (21a) the noun yad ‘memory’ is combined with the light verb kar ‘do’. In this case the subject must be ergative and overall reading is one of an agentive, deliberate remembering. The difference between (21b) and (21c) is one of eventive vs. stative, so that in (21b), lɑṛki ‘girl’ is already taken to be in the state of remembering a/the story (and not actively entering a state of remembering it). In (21c) the light verb is the participial form of ho ‘be’ and essentially means ‘become’.

The main mechanisms of N-V CP formation are theoretically well-understood (e.g., Davison, 2005, Mohanan, 1994). What is less well understood is how exactly the predicate nominal and the light verb combine to form an N-V CP. As indicated by recent research on the matter (e.g., Ahmed and Butt, 2011), it is not the case that all nominals are compatible with all light verbs; rather, different nominals display selectional preferences regarding their possible combinations with light verbs. With respect to the three light verbs kar ‘do’, ho ‘be’, hu- ‘become’, Ahmed and Butt (2011) identify three distinct classes. Nouns such as yad ‘memory’ in (21) are compatible with all three light verbs kar ‘do’, ho ‘be’, hu- ‘become’. Nouns patterning with tamir ‘construction’ as in (22) do not form CPs with ho ‘be’ as well as hu- ‘become’; (22b–c) are ungrammatical.

(22) a. bilal=ne mɑkan tamir ki-ya
       ‘Bilal built a/the house.’ (lit.’Bilal did construction of a/the house.’)  (Ahmed and Butt, 2011)

b. * bilal=ko mɑkan tamir he
       ‘Bilal built a/the house.’ (lit. ‘Construction of a/the house is at Bilal.’) (Ahmed and Butt, 2011)

c. * bilal=ko mɑkan tamir hu-a
       ‘Bilal built a/the house.’ (lit. ‘Construction of a/the house became to be at Bilal.’) (Ahmed and Butt, 2011)

N-V CP formation is highly productive in Hindi/Urdu (Ahmed and Butt, 2011, Mohanan, 1994). CPs are in fact encountered frequently in general language use, as well as
With respect to implementing the combinations, the frequency and productivity of N-V CPs mean that it is not feasible to simply construct a static list of possible combinations. What seems to be at the heart of N-V CP formation is that there are distinct groups of nominals that behave alike. Ahmed and Butt (2011) provide an initial discussion, but they limit themselves to 45 nouns and a set of three light verbs.

Thus, one chapter of this dissertation is devoted to the combinatory possibilities of nominals with light verbs. The contribution is exploratory: it consists of two distinct corpus studies that are aimed at providing input to an implementation of N-V CPs by means of the Urdu ParGram grammar. As it turns out, the empirical possibilities can be modeled successfully in an XLE implementation by making use of so-called Optimality Theory (OT) marks (Frank et al., 1998), constraints inspired by Optimality Theory (Prince and Smolensky, 2004) that model statistical preferences. I demonstrate how these marks can be used efficiently to prefer a CP analysis over a non-CP analysis for a given N-V combination, or vice versa if a noun is found to rarely combine with a light verb in a productive way.

1.8 Structure of the Thesis

The thesis is structured as follows. Chapter §2 provides relevant background information. It discusses the main features of Hindi/Urdu and introduces the framework of LFG, which provides the theoretical backdrop for the thesis. The chapter also XLE, the grammar development platform used to implement the nominal predication patterns, as well as the ParGram project, the collaborative computational grammar writing effort which includes the Urdu ParGram Grammar. A description of the main features of that particular grammar is given at the end of the chapter.

Chapter §3 is a detailed discussion of the Hindi/Urdu genitive case. Genitive nominal arguments are examined in detail with respect to their constituent properties as well as functional behavior. It is shown that possessive clauses behave like one-place predicates with a single subject embedding a genitive-marked GF. Chapter §4 presents an LFG analysis that advocates a strictly nominal domain for the genitive. The analysis is implemented in XLE; I explain all the relevant technical details of the implementation.

Chapter §5 proceeds accordingly with the locative and instrumental case. After a general introduction to locative and instrumental case marking in Hindi/Urdu, the chapter discusses locative/instrumental nominal arguments and the ways in which they are assigned, again taking into account their linear (c-structural) as well as functional (f-structural) properties. Data from locative clauses are also presented. The LFG analysis of the locative/instrumental pays tribute to the fact that the locative and instrumental...

---

3There is in fact a range of different elements that may combine with verbs to form CPs: verbs, nouns, prepositions, adjectives all occur in CPs (Butt et al., 2012). Regarding frequencies of occurrences, Vaidya et al. (2014) report that ~37% of predicates have been annotated as some kind of CP in the Hindi/Urdu Treebank (HUTB, Bhatt et al., 2009, Palmer et al., 2009). This figure encompasses all kinds of CPs.
in Hindi/Urdu are non-nominal, except where licensed on arguments of certain nominal classes. The analysis as well as XLE implementation of the locative and instrumental are described in Chapter §6.

Finally, Chapter §7 is concerned with N-V CPS. After providing details about the construction itself and its theoretical assessments, the chapter presents two distinct corpus studies carried out by myself and colleagues. Note that these studies have previously been published as Butt et al. (2012) and Sulger and Vaidya (2014). A separate part of that chapter describes the implementation of N-V CPS; the implementation directly incorporates the empirical results of the second corpus study (Sulger and Vaidya, 2014).

Chapter §8 recalls the main contributions of the thesis, a discussion, as well as a conclusion. The actual implementation in terms of the Urdu ParGram Grammar (in its most recent version, as of the time of submission of the thesis) is included on the CD-ROM attached to this document; the implementation can also be tested using the online INESS platform for treebanking and LFG grammar testing (Rosén et al., 2012a,b).

---

4The INESS homepage is located at http://iness.uib.no.
2.1 INTRODUCTION

This chapter provides background for the remainder of the thesis. §2.2 defines the main topic of the dissertation, nominal predication and arguments thereof. The section discusses relevant literature, both from the theoretical linguistics domain as well as from the natural language processing (NLP) domain. §2.3 introduces the language this thesis is concerned with, Hindi/Urdu, and presents some of its main grammatical features. §2.4 briefly describes the theoretical backdrop for this thesis, Lexical-Functional Grammar (LFG). §2.5 is a description of XLE, a development platform and parser for computational LFG grammars; the section also discusses the notational conventions within XLE that I make use of throughout the thesis. XLE is used by the ParGram (Parallel Grammar) project, which is presented in §2.6, with a focus on ParGram’s main features relevant for this thesis. Since my work is situated within the context of the Urdu ParGram grammar, I proceed to introduce the design and the main ingredients of that particular grammar in §2.7. §2.8 concludes the chapter.

2.2 NOMINAL PREDICATES & ARGUMENTS

Nominal predicates and their arguments have been a hotly debated topic in the theoretical linguistics community for some time. However, as stated in the introductory chapter, there is not a lot of work in the NLP community devoted to identifying, annotating, or parsing nominal predicators and their arguments. This section sums up the most important literature on the subject from two different angles: theoretical linguistics and
semantic role labeling. The general ParGram take on nominal predication and nominal arguments is presented in §2.6.4; the ParGram approach is compared to the approach taken in HPSG grammar engineering in §2.6.5. The survey on the theoretical linguistics literature given here does not include related work in LFG; I discuss related theoretical work in LFG on possessive structures in §3.3.

2.2.1 A Brief Overview of the Theoretical Linguistics Literature

There is no universally accepted theory of nominal arguments in theoretical linguistics, neither in the syntax literature, nor in the semantics literature. Pre-theoretically speaking, there are some generalizations about nominal argument structure that seem to hold across languages and that seem to be accepted throughout the literature. Influential work from theoretical linguistics is summed up in this section; the works are concerned with the optionality of nominal arguments, the types of nouns eligible for argument structure, as well as distinguishing nouns with argument structure from nouns without one.

The majority of researchers have claimed that there are crucial differences in the way nominals select their arguments compared to the way verbs select theirs. In particular, while verbal arguments are generally required for the clause to be felicitous, arguments of many nouns are in fact optional. Ungrammatical examples of verbs missing arguments are given in the a. examples of (1–4); grammatical examples for the corresponding derived nominals missing their arguments are given in the b. and c. examples.

(1)  a. *The Romans destroyed.
    b. The destruction by the Romans was horrific.
    c. The destruction took ages.

(2)  a. *Carl bought.
    b. Carl’s buy was smart.
    c. The buy was smart.

(3)  a. *Examined the horse.
    b. The examination of the horse was thorough.
    c. The examination was thorough.

(4)  a. *Bill translated.
    b. Bill’s translation persuaded the author.
    c. The translation persuaded the author.

These examples have led many researchers to assume that, in contrast to verbal arguments, all arguments of nouns are in fact optional (Barker and Dowty, 1993, Dowty, 1989, Higginbotham, 1983, Rappaport, 2006). Grimshaw (1990), in opposition to this position, argues that certain nouns, which she refers to as ‘complex event nominals’, take
obligatory arguments, and further that many deverbal nouns are systematically ambiguous between argument-taking complex event nominals and non-argument taking ‘result’ or ‘simple event’ nominals. For example, the nouns *felling* and *destroying* are argued by Grimshaw (1990) to only have a complex event reading, and thus have obligatory objects, like their verbal counterparts; see the examples in (5).

(5)  

a. the felling *(of the trees)  

b. the destroying *(of the city)  

Grimshaw (1990) provides a battery of tests that are designed to distinguish argument-taking complex event nominals from non-argument-taking result nominals. She argues, for example, that the adjectives frequent and constant force the complex event reading of singular nouns, and that only result nominals pluralize; in addition, she states that a, one, and that occur only with result nominals, and that the is the only determiner which can occur with both. Ultimately, Grimshaw (1990) argues that what sets apart the two types of nominals is the way they select their external argument: while result nominals select the referential argument (r), complex event nominals select the event argument (ev) (cf. Williams, 1981); the selection of the referential external argument leads to a referential (result) reading, and the selection of the event argument leads to a complex event reading. For positions similar to that of Grimshaw (1990), see Zubizarreta (1987) as well as Lebeaux (1986) (and references therein).

The distinction made by Grimshaw (1990) has since been challenged; for example, Zaenen and Goldberg (1993) provide the examples in (6), arguing that there are in fact contexts in which indefinite complex event nominals do allow indefinite determiners (6a), and complex nominals occur with the demonstrative that fairly freely (6b).

(6)  

a. A careful cleaning of the pool by professionals was performed.  

b. Who performed that careful examination of the body?  

Borer (2013) discusses the syntax of derivational affixes in general, and the role these affixes play in nominalizations in particular. Working in a generative framework which Borer has termed the “eXoSkeletal Model”, she rejects the idea that lexical items have a syntactic category. Rather, it is their distribution that imposes a syntactic category on them (Borer, 2003, 2005a,b, 2013). For example, roots occurring below nominal projections become nominal in that specific structure only by distribution. In other words, it is the structural context which renders the root “equivalent” to a noun. Thus, the notion of equivalence is to be interpreted distributionally, not featurally (Borer, 2005a,b), and the computational workload is shifted from the morphology into the syntactic component.

---

\(^1\) Borer thus opposes lexicalist frameworks, which generally assume that terminal nodes in the syntactic tree are morphologically complete words. LFG, for example, adopts the “Lexical Integrity Principle” which assumes that “syntactic transformations do not play a role in word formation” (Bresnan, 1978). Different but ultimately equivalent formulations of the same principle can be found in e.g., Asudeh et al. (2008), Bresnan.
Regarding nominalization, Borer (2013) argues that derivational affixes merge with their roots in the syntax. Like Grimshaw (1990), Borer (2013) distinguishes complex event nominals and result nominals; she, however, refers to the former as “argument structure nominals” (AS-nominals), and to the latter as “referential nominals” (R-nominals). Although many deverbal nouns are ambiguous between the two types of readings, there are very different syntactic properties attached to both of them: AS-nominals are syntactically richer in that they display properties usually associated with event structure, i.e., they have an argument structure with agent-oriented and other specific modifiers. Under Borer’s account, the nominal derivational affix (e.g., English -ation) takes a root complement which the affix defines as “equivalent” to a verb. Thus, both types of nominals contain a root which has become equivalent to a verb by means of the structure it is part of. The two types of nominals differ, however, in that AS-nominals contain a verbal extended projection, whereas R-nominals do not. Most previous proposals on AS-nominals include the intuition that they must contain a source of verbal properties. In contrast with those proposals, Borer (2013) argues that the source of the eventive interpretation is not the nominal affix or the noun itself (in a lexicalist view), but functional structure which is contained below the nominal affix. She also rejects the idea that an embedded, listed verb gives rise to the eventive interpretation as, under her view, the structure contains a root rather than a verb.

A view related to Borer’s is given by Alexiadou (2001). Here, event/AS-nominals are called “process nominals”; they are assumed to include both nominal as well as verbal projections (cf. the concept of structural equivalence put forward by Borer (2005a,b)). The verbal projections (Voice as well as Aspect) express eventivity as well as viewpoint aspect. Result nominals lack the verbal projections; consequently, process nominals behave like nouns externally, while internally, they display verbal properties. Similar accounts can be found in, e.g., Schoorlemmer (1995), van Hout and Roeper (1998) as well as Sichel (2010).

Barker (1995) is relevant in that he works out an account of possessive arguments of relational and non-relational nouns. Under his account, verb-derived event nominals such as description or kinship terms such as daughter intrinsically (i.e., lexically) select for a possessor, while common nouns like book or firetruck do not intrinsically do so, but nevertheless allow modification by extrinsic possessors. Some examples for intrinsic/lexical possessives are given in (7); extrinsic possessives are in (8).

(7) Lexical Possessives (Barker, 1995, p. 8)
  a. John’s purchase Derived nominals
  b. John’s child Kinship terms
  c. John’s nose Body part terms

The principle extends to inflectional morphology and especially derivational morphology, which is assumed to apply before inflection in the lexicon.
2.2. NOMINAL PREDICATES & ARGUMENTS

d. the table’s top  Generalized part/whole relations
e. the woman’s pen pal  Arbitrary relational nouns

(8) Extrinsic Possessives (Barker, 1995, p. 8)
a. John’s cat
b. John’s yogurt
c. John’s firetruck

Both of these selection processes do not per se, as acknowledged by Barker (1995), put any constraints on the thematic role encoded by the possessor; however, in the case of intrinsic possession, the specific thematic relation depends on inherent qualities of the head noun. According to Barker (1995), relational nouns are semantically two-place predicates, while non-relational nouns are one-place predicates; a noun such as top denotes a two-place relation between the top and the entity that the top is a part of, but a noun such as cat does not specify any such special relation. What complicates matters is that nouns seem to be lexically ambiguous between an intrinsic reading and an extrinsic reading. For example, on one reading of (9a), children enforces a two-place relationship with the possessor, John, in which a kinship relation is entailed; on one reading of (9b) (i.e., a reading where John is responsible for children at a day-care center), no such relation is entailed. The issue here is that nouns such as children nevertheless entail the existence of a second argument (in this case, a parent).

(9) a. John’s children have children of their own now.
   b. John’s children try hard to be good when he gets that look.

Barker (1995) accounts for this by assuming argument suppression: Even though the related entity is not expressed overtly, it is continued to be entailed by the relational noun; in addition, the relational noun remains eligible for extrinsic possession, as in the case of (9b). My own syntactic account of genitive arguments in Hindi/Urdu bears similarities to the account by Barker (1995) (see Chapter §3).

2.2.2 Semantic Role Labeling

Semantic role labeling (srl) focuses on the annotation of arguments as well as the thematic relations to their heads. A major research effort focused on identifying and annotating nominal predicates and their arguments is NomBank (Gerber et al., 2009; Meyers, 2008; Meyers et al., 2004a,b,c,d). The NomBank project is closely related to the PropBank project (Kingsbury et al., 2002; Palmer et al., 2005), which is concerned with annotating verbal predicate-argument relations. The goal of NomBank is to mark sets of arguments that co-occur with nominals in the PropBank corpus (which is identical to the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993)).

One resource developed by NomBank are nominal frame files that include lexical entries for nouns; these lexical entries identify the various arguments of the given head
noun with the argument roles. The argument roles define the semantic relation that holds between the argument and the head noun. Like the PropBank argument roles, NomBank's argument roles are essentially the same as the initial relations of Relational Grammar (Blake, 1990, Perlmutter and Postal, 1984). For example, agents tend to be classified as ARG0 (RG's initial subject), patients and themes tend to be classified as ARG1 (RG's initial object) and indirect objects of all kinds tend to be classified as ARG2. Argument roles may be different across the senses of a noun; e.g., the noun claim comes with three different senses ASSERT, SEIZE and SUE, and the argument roles across these senses differ (Meyers et al., 2004b). In addition to nominal arguments, NomBank also identifies possible adjunct types that may co-occur with noun senses.

NomBank is collaborating closely with PropBank in that the argument role definitions are kept consistent across parts of speech. For example, NomBank is using PropBank’s frame files for annotating nominalizations of verbs, such that, e.g., the noun decision will receive the same argument-role list as the verb decide. NomBank is a manually annotated resource. A major claim of the work around NomBank is that a statistical classifier can be trained on the resource using machine learning techniques, and argument roles of nouns occurring in unseen data can then be automatically classified. Following the release of the NomBank corpus, there have been experiments applying verbal SRL techniques to the nominal domain (e.g., Liu and Ng, 2007).

While SRL is an important task for applications like Information Extraction and Question Answering, it does not help in explaining the heterogenous syntactic behavior of nominal predicates and arguments. Thus, the grammar engineering approach taken in this thesis differs in that it models nominal predication from the bottom up, looking in detail at the morpho-syntactic behavior of different classes of nominals, and cutting across the argument labels identified by NomBank (and its predecessor, PropBank).

2.2.3 Summary

To summarize, I have provided a brief overview of the (non-LFG) theoretical linguistics literature on nominal arguments; I have also discussed recent work from the area of semantic role modeling. It remains an open question in how far the suggestions from theoretical syntax can be translated into computational models; in particular, the ideas by Borer (2013) and others in the same vein concerning the syntactic underspecification of categories (and their definition by means of structural context) are hard to picture in an NLP setting. I myself am in the domain of computational grammar development, subscribing to the lexicalist worldview of LFG; the highly lexicalized theory of extrinsic vs. intrinsic possession put forward by Barker (1995) provides a semantic account that is compatible with the ideas presented in this dissertation.
2.3 Hindi/Urdu

The two languages Hindi and Urdu are so closely related that many researchers in linguistics treat them as a single language, Hindi/Urdu.\footnote{A traditional term for referring to the unity of the two languages is Hindustani.} The Urdu ParGram Project (see §2.7) has been developing a broad-coverage computational grammar which is aimed at handling both Hindi and Urdu at the same time. Differences between Urdu and Hindi are mainly in the script (Urdu uses a version of Arabic script, while Hindi uses Devanagari) as well as in the vocabulary. There are further minor differences in the phonology as well as in the derivational morphology. The bulk of this thesis looks at data from Hindi/Urdu, and the analyses developed throughout the thesis are applicable to both languages. Whenever some piece of data exclusively applies to one of the two languages (e.g., in cases where a sentence may only be understood by speakers of Urdu), it will be indicated along with the data.

2.3.1 Configurationality

The canonical word order in Hindi/Urdu is subject, object, verb (sov). The major constituents in the sentence may, however, scramble. Thus, while sentences will often be in the canonical sov order, all other orders are also in principle possible. See (10) for an example.

\begin{itemize}
  \item a. nadya=ne yasin=ko kɪtab d-i
      ‘Nadya gave Yassin the book.’
  \item b. nadya=ne kɪtab yasin=ko d-i
      ‘Nadya gave Yassin the book.’
  \item c. kɪtab nadya=ne yasin=ko d-i
      ‘Nadya gave Yassin the book.’
  \item d. yasin=ko kɪtab nadya=ne d-i
      ‘Nadya gave Yassin the book.’
  \item e. yasin=ko nadya=ne kɪtab d-i
      ‘Nadya gave Yassin the book.’
\end{itemize}

The sentence in (10a) is in the canonical word order: subject, indirect object, direct object, verb. All other orders (with the exception of the verb-initial ordering) are also in fact possible, as illustrated by the examples (10b–e).\footnote{10 does not include an example where the verb does not occur in final position. Even though the canonical position of the verb is clause-final, material can follow the verb, given the appropriate information-structural configuration. See §2.3.5.} Researchers such as Butt\footnote{\cite{butt1995}}...
and Mohanan (1994) have thus assumed the general clause structure to be a flat one as depicted in the rule in (11), modeling free word order, i.e., non-configurationality.

(11) $s \rightarrow np^*, v$  
(Butt, 1995, p. 21)

As the analyses presented in this thesis are couched in the LFG framework, I will follow the suggestions made by Butt (1995), working within the same theory. Thus, I take (11) as a basis for examining the general clause structure of Hindi/Urdu.

### 2.3.2 Case

As discussed above, Hindi/Urdu displays fairly free word order. At the same time, case is a major component of its grammar. The identification of the grammatical functions (GF) in a clause is largely due to a complex system of morphological case marking.

Hindi/Urdu case marking is taken to fall into three major categories: case clitics, postpositions, and inflections on word forms (Butt, 1995; Butt and King, 2004b; Mohanan, 1994). Table 2.1 is based on Butt and King (2004b); it provides an overview of the case clitics, along with the GFs that they may occur with. The table shows that there are seven distinct overt case clitics, realizing seven distinct cases, occurring with seven different GFs. In addition, nominative case is indicated by the absence of case marking. It can also be seen that the genitive case marker $k$- is the only case marker which inflects; according to Butt and King (2004b), this is due to the fact that it developed from an inflecting participial (see also §3.2).

When arguments or adjuncts appear with a case marker, they always inflect to an oblique form; furthermore, when they appear as locatives with a zero locative case marker, they also inflect to the oblique. When they occur in nominative case (and hence are not followed by a case clitic), they never inflect. Table 2.2 shows the possible inflections on feminine (typically ending in -i) and masculine (typically ending in -a) nouns. The table is based on Butt (1995), extended to include feminine nominals.

For an example of how this system works, consider the sentences in (12). $dek^h$ ‘see’ is a transitive verb selecting a subject and an object; the sentence in (12a) is in the perfective

---

4 A couple of short remarks about (11) are necessary. First, the Kleene star (*) allows zero or infinitely many NPs as daughters of $s$; this part of the rule makes sure that all possible NPs that may occur in a clause are covered by the rule. Second, the comma operator, called the shuffle operator in LFG/XLE, results in the shuffling of the individual parts of the right-hand side of the rule (np* and v) in any possible order (Dalrymple, 2001, p. 99). Third, (11) merely illustrates the basic clause structure of Hindi/Urdu and thus ignores other constituents such as PP (prepositional phrase), ADVP (adverbial phrase) or CP (complementizer phrase), which may be more complicated regarding their linear order within the clause. LFG as a theoretical framework is introduced in §2.4; its implementation in XLE is discussed in §2.5. The shuffle operator is discussed in §2.5.3.6.

5 There has been a debate whether the case markers in Table 2.1 should in fact be considered clitics, or whether they fall into a different grammatical category; for a different take, see, for example, the discussion in Spencer (2005). For the purposes of this thesis, I assume the position of Butt and King (2004b) and Mohanan (1994), who present abundant evidence that the case markers indeed need to be treated as clitics.
aspect, and in this case, the subject must bear ergative case. (12b) encodes exactly the same meaning as (12a), while the linear order of subject and object is different. The case on the object is governed by at least specificity and animacy (Butt, 1993, Butt and King, 2004b, Dayal, 2003, Masica, 1991): animate objects typically require the accusative (12a–b), while non-specific inanimate objects occur in the nominative (12c) and specific inanimate objects occur in the accusative (12d). (12e) contains an example of a zero-marked locative (gɪrje ‘to the church’).

(12)  
\begin{enumerate}
\item \texttt{lar\k{a}=ne kʊ\texttt{tte}=ko d\texttt{ek}^{h-a}}
\texttt{boy.M.SG.OBL = ERG dog.M.SG.OBL = ACC see-PERF.M.SG}
‘The boy saw the dog.’
\item \texttt{kʊ\texttt{tte}=ko lɑ\texttt{rke}=ne d\texttt{ek}^{h-a}}
\texttt{dog.M.SG.OBL = ACC boy.M.SG.OBL = ERG see-PERF.M.SG}
‘The boy saw the dog.’
\item \texttt{lar\k{a}=ne maka\texttt{n} d\texttt{ek}^{h-a}}
\texttt{boy.M.SG.OBL = ERG house.M.SG.NOM see-PERF.M.SG}
‘The boy saw a house.’
\item \texttt{lar\k{a}=ne maka\texttt{n}=ko d\texttt{ek}^{h-a}}
\texttt{boy.M.SG.OBL = ERG house.M.SG = ACC see-PERF.M.SG}
‘The boy saw the house.’
\end{enumerate}
e. lɑṛka ḡirje ja rah-a ḍe
   ‘The boy is going to the church.’

The ergative case marker is optional with unergative verbs in Hindi/Urdu. 6 (13a) is an example with the unergative verb hans ‘laugh’. When the ergative does occur, as in (13b), the sentence expresses volitionality on part of the subject (Butt and King, 2004b, Davison, 1999, Mohanan, 1994).

(13) a. lɑṛka hans-a
   boy.M.SG.NOM laugh-PERF.M.SG
   ‘The boy laughed.’

b. lɑṛke=ne hans-a
   boy.M.SG.OBL =ERG laugh-PERF.M.SG
   ‘The boy laughed (purposefully).’

Observations such as the specificity alternation (12c–d) and the volitionality pattern (13) have been taken as evidence for proposing semantic case marking for Hindi/Urdu (Butt, 1995, Butt and King, 2004a,b). The key idea behind semantic case is that case on a particular NP can be determined by examining the semantic regularities across predicates that accompany the case marking. Another defining characteristic of semantic case is that it is typically subject to only appearing on certain grammatical functions; for example, in Hindi/Urdu, the ergative case marker appears only on subjects. This is formally similar to the notion of constructive case proposed by Nordlinger (1998) for Australian languages, which use case to identify grammatical functions.

At the same time, Hindi/Urdu also makes use of two other case marking strategies: structural case (Butt and King, 2004b, Davison, 1999, Mahajan, 1990) as well as quirky case (Butt and King, 2004b). Generally speaking, structural case is not related to semantic regularities triggered by the case marking, but instead is tied solely to grammatical function or phrase structure position. Structural case is also often an “elsewhere” case, applied only if other strategies are not applicable. In Hindi/Urdu, an example of structural case is the nominative; e.g., subjects as well as objects receive nominative case if nothing else assigns case to them (Butt and King, 2004b). (12c) contains a non-specific object which in this elsewhere case is assigned nominative case; the non-volitional subject lɑṛka ‘boy’ in (13a) is nominative, too.

Finally, “quirky”/idiosyncratic case marking is determined exclusively by the lexical entry of the predicate. This type of case marking is exceptional to the system in that it can be predicted neither based on semantic effects nor by looking at the structural

---

6 The unergative/unaccusative distinction among intransitive verbs is well known (Bresnan and Zaenen, 1990, Hale and Keyser, 1993, Levin and Hovav, 1995, Perlmutter, 1978, among others), and I will not go into detail here. Basically, the subjects of unergative verbs can be argued to behave similarly to the agentive subjects of transitive verbs, while the subjects of unaccusative verbs behave similarly to the patient objects of transitive verbs.
configuration of the NP (Butt and King, 2004b). An example of quirky case marking is given in (14). Here, the transitive verb la ‘bring’ occurs in the perfective; recall that in this case, the subject should bear ergative case. Still, the subject of la is required to be nominative. Nothing in the structure or the semantics of the clause hints at this.

(14) nadya kitab la-yi
    Nadya.FSG.NOM book.FSG.NOM bring-PERF.FSG
    ‘Nadya brought a book.’

(Butt and King, 2004b, p. 25)

See Zaenen et al. (1985) for a general discussion of quirky case marking and its application to data from Icelandic. As will be seen in Chapters §3 and §5, nominal predicates in Hindi/Urdu occur with structural/default case marking as well as quirky case marking.

2.3.3 POSTPOSITIONS

Butt and King (2004b, p. 18–19) state that because Hindi/Urdu case clitics attach post-nominally, they have been called postpositions in previous accounts of the language. Hindi/Urdu, however, also makes use of a distinct set of postpositions. Examples of postpositions are shown in (15). (16a) and (16b) illustrate their use; in both sentences, the postpositions occur on adjuncts.

(15) ke picʰe ‘behind’        ke pele ‘before’
    ke nícʰe ‘under’          ke pas ‘next to/near’
    ke upar ‘over’            ke satʰ ‘with’
    ke andar ‘inside’         ke lye ‘for’
    ke samne ‘in front of’    ke taraf ‘in the direction of’
    ke age ‘in front of (further along)’ ke bad ‘after’

(16) a. ləɾka girje = [ke         pas] hans-a
    boy.M.SG.NOM church.M.SG.OBL = GEN.M.SG.OBL near laugh-PERF.M.SG
    ‘The boy laughed near the church.’

b. nina yasin = [ke          lye] stor gay-i
    Nina.FSG.NOM Yassin.M.SG = GEN.M.SG.OBL for store.M.SG go-PERF.F.SG
    ‘Nina went to the store for Yassin.’

All of the items in (15) are composed of ke, the oblique form of the genitive case marker k-, and a postposition. The genitive is invariant for most of the postpositions; native speakers do not associate it with genitive case but take it to be part of the postposition. Butt and King (2004b) also give the example in (17), showing that some of the postpositions allow inflection reminiscent of nominals and establish a link to an originally genitive construction (cf. Masica, 1991, p. 234–235) who gives the diachronic derivations ke picʰe ‘behind’ < (ka) picʰa ‘(its) rear’ as well as ke andar ‘inside’ < (ka) andar ‘(its) inside’, also arguing for a nominal origin of the postpositions).
2.3. HINDI/URDU

(17) mɛ̃=ne bɪlli=ka picʰa kr-ya
    I=ERG cat.F.SG =GEN.M.SG.NOM behind.M.SG.NOM do-PERF.M.SG
    ‘I went after the cat.’ (Butt and King, 2004b, p. 18)

In contrast to the case clitics, the postpositions may also appear by themselves as in (18). Thus, Butt and King (2004b) argue that they differ in both form and distribution from the case clitics.

(18) upɑr ao
    up come.IMP
    ‘Come up!’ (Butt and King, 2004b, p. 18)

Moreover, the case clitics may all mark GFS, while noun phrases marked with the postpositions in (15) only occur on adjuncts and OBLs (Butt and King, 2004b, Mohanan, 1994). Therefore, there is also a functional difference between phrases marked by case clitics on the one hand and phrases marked by postpositions on the other hand.

2.3.4 AGREEMENT

Verbal agreement in Hindi/Urdu is usually clause-bound; this means that the domain of verbal agreement is the clause, and that agreement may not cross clause boundaries. The basic pattern is the following. The verb only agrees with unmarked GFS. If the subject is unmarked, the verb agrees with it. If the subject is not unmarked, but the object is, then the verb agrees with the object. If the subject and the object are not unmarked, the verb agrees with any remaining unmarked GF in the clause (such as a secondary object, oblique, etc.). Finally, if there is no unmarked GF in the clause, then the verb defaults to masculine singular agreement. Thus, the verb agrees with the highest ranked unmarked GF in the clause; otherwise, it has default marking (Butt and King, 2004b, Mohanan, 1994).

For examples of how this system works, consider the examples in (19). In (19a), the subject is unmarked and the verb agrees with it. In (19b), there is no unmarked argument in the clause, which is why the verb bears default (masculine singular) agreement features. In (19c), the verb agrees with the (non-specific, thus nominative) object, the only unmarked argument in the clause. (19d) has an unmarked subject and an unmarked object; in such cases, the verb always agrees with the subject.

(19) a. nina hans-i
    Nina.F.SG.NOM laugh-PERF.F.SG
    ‘Nina laughed.’

7There are in fact exceptions to the generalization that agreement is clause-bound. For example, long-distance agreement in Hindi/Urdu involves a verb agreeing with a constituent inside the verb’s clausal complement, as discussed by (Bhatt, 2005). See also Davison (2014) and Butt (2014) for further discussions of agreement patterns.
2.3. HINDI/URDU

b. \text{nina=ne  ṭopi=ko  \text{dekʰ-a}}
   \text{Nina.F.SG = ERG  hat.F.SG = ACC  see-PERF.M.SG}
   ‘Nina saw the hat.’

c. \text{nina=ne  ṭopi  \text{dekʰ-i}}
   \text{Nina.F.SG = ERG  hat.F.SG.NOM  see-PERF.F.SG}
   ‘Nina saw a hat.’

d. \text{nina  kʊṭṭa  \text{dekʰ-ti  hɛ}}
   \text{Nina.F.SG.NOM  dog.M.SG.NOM  see-IMPF.FEM.SG  be.PRES.3.SG}
   ‘Nina looks at a dog.’

The example in (20) shows the clause-boundedness of agreement: even though the main clause’s subject \text{Nina} is feminine, the embedded clause’s verb displays masculine agreement, as the embedded clause’s subject is masculine.

(20)  

a. \text{nina  hans-i}
   \text{Nina.F.SG.NOM  laugh-PERF.F.SG}
   ‘Nina laughed.’

b. \text{nina  kah-ti  hɛ  [kɪh  lɑṛka  hans-a]}
   \text{Nina.F.SG.NOM  say-IMPF.FEM.SG  be.PRES.3.SG  that  boy.M.SG.NOM  laugh-PERF.M.SG}
   ‘Nina says that the boy laughed.’

2.3.5 INFORMATION STRUCTURE

\text{Butt and King} (1996) propose to analyze the free word order of languages such as Hindi/Urdu or Turkish as base-generated possibilities (i.e., alternatives that are not derived by movement from a deeper representation) that directly reflect different information structures. From a purely syntactic point of view, the variation in the word order is optional such that it is not motivated by any syntactic principle; rather, it is motivated by semantic factors (e.g., definiteness, specificity, givenness) as well as information-structural configurations (e.g., topic, focus). \text{Butt and King} (1996) postulate that the placement of arguments directly encodes their function in discourse. Using the two features \text{[±new]} and \text{[±prominent]} previously identified by \text{Choi} (1999a), they identify four different discourse functions: topic (\text{[−new]}, \text{[+prominent]}), focus (\text{[+new]}, \text{[+prominent]}), background (\text{[−new]}, \text{[−prominent]}) and completive information (\text{[+new]}, \text{[−prominent]}). Topic and focus information is treated as prominent, which sets them apart from completive and backgrounded information. Completive and focus are regarded as information which is new to the discourse. While the focus function fills the informational gap between speaker and hearer, and is thus \text{[+prominent]}, completive information is \text{[−prominent]}: it is not primarily important to the current discourse.
2.4. LEXICAL-FUNCTIONAL GRAMMAR

Butt and King (1996) make the following generalizations about the way discourse functions are tied to position. In Hindi/Urdu, the first position in an utterance is interpreted as a topic (T in (21a)), while the focused element of an utterance is required to appear in immediately preverbal position (F in (21b)). Background information occurs postverbally (BACK in (21c)). Completive information, just like topic and focus, occurs preverbally, but is not licensed in any particular position. Thus, completive information can be thought of as new, non-prominent information occurring preverbally. For example, in (22b) (an answer to the question in (22a)), the object tɔfi ‘toffee’ occupies the completive discourse function (COMP in (22b)).

The examples in (21) and (22) have been adapted from Butt and King (1996).

(21)  a. [hɑssɑn=ko]T nadya=ne tɔfi d-i
Hassan.M.SG = DAT Nadya.F.SG = ERG toffee.FSG.NOM give-PERF.F.SG
To Hassan Nadya gave toffee.’

b. nadya=ne hassan=ko [tɔfi]F d-i
Nadya.F.SG = ERG Hassan.M.SG = DAT toffee.FSG.NOM give-PERF.F.SG
‘Nadya gave toffee to Hassan.’

c. [tin hi cɪtʰiɣɑ]F dal-i hɛ
three only letter.F.PL.NOM put-PERF.F.SG be.PRES.3.PL
[us=ne]BACK PRON.3.SG.OBL = ERG
‘Only three letters (she) has sent.’

(22)  a. nadya kɑhɑ=se a ruh-i hɛ
Nadya.F.SG.NOM where=LOC.FROM come remain-PERF.F.SG be.PRES.3.SG
‘Where is Nadya coming from?’

b. nadyaT to abhi [tɔfi]COMP [bɑzar=mɛ]F
Nadya.F.SG.NOM indeed just.now toffee.FSG.NOM market.M.SG = LOC.IN
xɑrid ruh-i tʰ-i
buy remain-PERF.F.SG be.PAST-F.SG
‘Nadya was just buying toffee AT THE MARKET.’

This concludes my general introduction to Hindi/Urdu. I will be concerned with more specific investigations into the language, concerning the syntax of nominal predicates and their arguments as well as noun-verb complex predicates, in Chapters 3, 5 and 7.

2.4 LEXICAL-FUNCTIONAL GRAMMAR

The syntactic analysis of the data discussed in this thesis as well as the computational implementation will make use of the theoretical framework of LFG. In this section, the main features of LFG are described.

Butt and King (1996) assume that the completive reading is the default interpretation of discourse entities, i.e., when the requirements for assigning topic, focus or background are not met.
2.4. LEXICAL-FUNCTIONAL GRAMMAR

2.4.1 General Description and Architecture Assumed

One of the theoretical requirements of generative grammar is that it does not suffice to describe language purely in terms of the ultimately visible constituents, which Chomsky called “surface structure” (Chomsky, 1957). Rather, one also needs a somewhat “deeper” level of representation, to represent the implicit knowledge of language. Chomsky thought of this deeper level as a “deep” structure that represents the core semantic relations in a clause. From this abstract level, multiple operations or “transformations” derive the overt surface structure. In this separation of levels, Chomsky saw the explanation for language acquisition; deep structure is argued to be innate to humans, with the result that babies just need to learn the idiosyncratic transformations that are necessary to form grammatical sentences in their native language (the surface structures).

In the second half of the 20th century, a set of new theories were developed. In these novel theories, the focus was on parallel structures rather than serial transformations. A main objective of these theories was to construct typologically valid models of language; comparison of structures across individual languages was a main priority. Because of the typological aspirations of these theories, they abandoned a number of aspects of Chomsky’s original theories. For example, they did not make use of derivations, but instead viewed different levels of language as parallel to and dependent on one another. Restrictions in between the levels could be realized using constraints.

LFG is one of these non-transformational, constraint-based theories. It was developed in the 1980s by Joan Bresnan and Ron Kaplan (Kaplan and Bresnan, 1982) and has since received much attention in both theoretical syntax and natural language processing (NLP). LFG does not posit configurationality (the idea that language makes use of a fixed word order) as a universal fact of language. For that reason, it has adherents among syntacticians working from a typological perspective, who argue that it is a more plausible universal model of language than Chomsky’s frameworks (Bresnan, 2001, Dalrymple, 2001, Kaplan and Bresnan, 1982). It is a popular framework in the NLP community due to its strong mathematical basis, making it formally viable and efficiently parseable (Crouch et al., 2015, Dalrymple et al., 1995, Maxwell III and Kaplan, 1995).

The general LFG grammar architecture used in theoretical work is the one in Figure 2.1. LFG has to be thought of as a grammar architecture with multiple parallel levels of representation; the individual levels are in principle all present in a given analysis, and through so-called functional projections, also referred to as correspondence functions, the levels are mutually constraining (Bresnan, 2001, Kaplan and Bresnan, 1982, Nordlinger and Bresnan, 2011). For that reason, the architecture has been referred to as the Correspondence Architecture (Asudeh, 2006, 2012, Kaplan, 1995). For example, information-structure, projected off the syntactic tree (the c-structure) via the \( \iota \) function, may constrain the semantic-structure via the \( \iota^\sigma \) function (Butt and King, 2000, King, 1997). Furthermore, the \( \phi \) function (see the next section) is defined as the composition

\footnote{Figure 2.1 was adapted from Asudeh (2012).}
of the $\alpha$ (mapping from the c-structure tree to the argument structure) and $\lambda$ (mapping from the argument structure to the functional structure) functions.\(^\text{10}\)

![Figure 2.1: LFG's correspondence architecture](image)

Note however that this thesis is a contribution in LFG grammar development. I define computational grammar rules that can automatically identify nominal predicates and annotate their arguments accordingly. As such, not all the representational levels in Figure 2.1 will be made use of in this thesis. In fact, LFG grammar development is usually only concerned with modeling c(onstituent)-structure and f(unctional)-structure.\(^\text{11}\) The work in this thesis will rely only on these central levels of LFG, c-structure and f-structure, to be described below.

### 2.4.2 C-STRUCTURE AND F-STRUCTURE

LFG posits different levels of syntactic structure, which are parallel to each other, and at the same time may enforce constraints on one another (Bresnan, 2001; Kaplan and Bresnan, 1982; Nordlinger and Bresnan, 2011). At the outset of the theory, LFG consisted of two interdependent levels, c-structure and f-structure. These two levels were taken to adequately represent a native speaker’s syntactic knowledge; however, there are fundamental differences between the two structures. While f-structure describes the relation between predicates and their arguments and adjuncts (basically, a form of dependency structure), c-structure models the surface word order and phrase configurations by which the predicates are expressed (Butt, 1995; Kaplan and Bresnan, 1982; Nordlinger and Bresnan, 2011). This points to f-structure representing a deeper description of language, “deep” in the sense that it represents cross-linguistically valid generalizations about the

\(^{10}\)The terms “(functional) projection” and “correspondence function” are used largely in the same sense in the LFG architecture; I use the latter term in the descriptions below.

\(^{11}\)There are also exceptions where levels of representation other than c- and f-structure are used in implementation. For example, Bögel et al. (2009) make explicit use of p(rosodic)-structure to encode the Urdu Ezafe construction and its particular prosodic properties. Sulger (2009a) describes an LFG analysis and implementation of Irish clefting by means of i(nformation)-structure. Butt et al. (1996) use a structure not present in Figure 2.1, m(orphological)-structure, to encode different morphological wellformedness conditions on auxiliary stacking; the individual grammatical contributions of the auxiliaries, on the other hand, are encoded at f-structure.
principles of syntax, while c-structure represents the language-specifics, such as word order and phrase structure. In Figure 2.2, an example of an English c-structure is given. This c-structure can be produced by the rewrite rule set in (23). Phrase structure rules encode dominance and precedence relations, labeling syntactic entities (constituents and lexical items) using the usual types of symbols (Bresnan, 2001, Butt, 1995, Kaplan and Bresnan, 1982). Each rule has an input (left side) and an output (right side). Optional phrases are indicated using brackets. (23d)–(23g) are rules of lexical insertion which insert lexical symbols into the tree. Note that the rewrite rules constitute an unordered set and could thus appear in any order; they are given in this particular order in (23) only for the sake of illustration.

(23) a. $S \rightarrow NP \ VP$
    b. $NP \rightarrow (DET) \ N$
    c. $VP \rightarrow V \ (NP)$
    d. $N \rightarrow Nina$
    e. $N \rightarrow boy$
    f. $V \rightarrow saw$
    g. $DET \rightarrow the$

Figure 2.2: Simplified c-structure

2.4.3 Mapping Between C- and F-Structure

As mentioned above, f-structures give “deep” descriptions of cross-linguistically invariant syntactic principles; f-structures abstract away from surface properties of different languages such as linear word order, to represent generalizations at deeper levels, such as

\[12\] LFG uses a version of the syntactic rewrite rules first introduced by Chomsky (1957) to define c-structures.
argument structure, tense/aspect configurations, etc. (Bresnan, 2001, Butt et al., 1999b, Dalrymple, 2001, among others). For representing f-structures, the format of attribute value matrices (AVM) is used. AVMs describe the functional nature of syntactic valency (hence the term functional-structure): verbal elements, for example, are essentially regarded as functions that apply to arguments to form semantically well-formed statements (Bresnan, 2001, Butt, 1995, Dalrymple, 2001). The idea of predicates resembling functions applying to their arguments has a long history in natural language semantics; in fact, f-structures have previously been shown to be equivalent to quasi-logical forms (Cahill et al., 2007, van Genabith and Crouch, 1996).

For mapping between the levels of c-structure and f-structure, LFG has defined the φ function. This function relates all nodes in the c-structure tree to corresponding elements in the f-structure. Arguments (e.g., subject, object) in LFG terms are called grammatical functions (GF), since they too can be considered functions: applying a function to the attribute SUBJ gives rise to the value Nina in Figure 2.6 (Dalrymple, 2001). The φ function makes sure that the NP ‘Nina’ in the English tree in Figure 2.2 is mapped onto the SUBJ GF in Figure 2.6, since it is this function that is filled by that particular NP. Conversely, φ also enforces that the NP ‘the boy’ is mapped onto the OBJ function. The mapping function φ is instantiated in LFG using annotations directly on the rewrite rules. The set of rewrite rules in (24), which includes f-structure annotations, is a modified version of the one in (23).

\[
\begin{align*}
\text{(24) } & \text{a. } S \rightarrow \text{ NP } \text{ VP} \\
& \quad (\uparrow \text{SUBJ}) = \downarrow \quad \uparrow = \downarrow \\
& \text{b. } \text{NP} \rightarrow \left( \begin{array}{c}
\text{DET} \\
\uparrow = \downarrow
\end{array} \right) \text{ N} \\
& \quad \uparrow = \downarrow \quad \uparrow = \downarrow \\
& \text{c. } \text{VP} \rightarrow \text{ V} \left( \begin{array}{c}
\text{NP} \\
\uparrow = \downarrow
\end{array} \right) \\
& \quad (\uparrow \text{OBJ}) = \downarrow
\end{align*}
\]

↑ and ↓ are metavariables that represent f-structures; the annotation (↑ SUBJ) = ↓ means the following: take all f-structural information annotated on the current node (↓) and project it upward to the SUBJ f-structure of the mother node’s f-structure ((↑ SUBJ)). The annotation ↑ = ↓ means that all information on the current node is projected upwards without embedding it into some f-structure within the mother node’s f-structure. If no f-structure annotation is given at all, the default annotation ↑ = ↓ is applied.

In (25), the lexical entries needed for the example sentence Nina saw the boy are given. The lexical entry for saw in (25a) specifies two features: a pred feature and a tense feature. The tense feature receives the value past. The pred feature is special: its value is a semantic form. In (25a), the semantic form value of the pred feature for the f-structure annotated on the terminal node V is ‘see<(↑ SUBJ), (↑ OBJ)>’. Likewise, (25b) specifies the semantic form for the f-structure annotated on N to be ‘Nina’. The single quotes enclosing a semantic form indicate that its value is unique: for example, each
instance of use of the word *Nina* gives rise to a uniquely instantiated occurrence of the semantic form ‘Nina’ (Dalrymple, 2001, Section 2.3.2). Figure 2.3 shows the complete c-structure tree for *Nina saw the boy* with f-structure annotations on the nodes.

(25) a. saw \(v\) \((\uparrow \text{PRED}) = \text{‘see}<(\uparrow \text{OBJ}), (\uparrow \text{OBJ})\text{>’}\)  
\((\uparrow \text{TENSE}) = \text{past}\)

b. Nina \(n\) \((\uparrow \text{PRED}) = \text{‘Nina’}\)

c. boy \(n\) \((\uparrow \text{PRED}) = \text{‘boy’}\)

d. the \(\text{DET}\) \((\uparrow \text{DEF}) = +\)

In each rule, the f-structures from lower structures are unified. For example, in (24a), the NP and VP are simply connected up at c-structure; at the same time, the f-structure of the NP and the f-structure of the VP are unified. The f-structure for the NP *Nina* in Figure 2.3 is given in Figure 2.4, while the f-structure for the VP is given in Figure 2.5. The final f-structure for the sentence *Nina saw the boy*, the unification of Figures 2.4 and 2.5, is given in Figure 2.6. LFG can be described as a unification-based grammar theory (Bresnan, 2001, Dalrymple, 2001, Kaplan and Bresnan, 1982): functional information coming from lexical heads travels upwards through the tree and is unified at the top level in the f-structure. Further discussion on the operation of unification can be found in e.g., Shieber (1986).
2.4. LEXICAL-FUNCTIONAL GRAMMAR

2.4.4 AN INVENTORY OF GRAMMATICAL FUNCTIONS

For representing a given argument at f-structure, LFG assumes a set of universally available GFS as shown in (26).

(26) Subject, Object, OBJθ, COMP, XCOMP, OBLIQUEθ, ADJUNCT, XADJUNCT

The labels OBJθ and OBLIQUEθ represent sets of functions that are tied to specific semantic roles. The θ subscript identifies the semantic role that is instantiated with the argument. For example, OBLLOC, forming a part of the thematically restricted OBLθ group, realizes the semantic role of a locative.13

13GFS such as SUBJ and OBLloc are thus regarded as incarnations of semantic arguments. As such, f-structures do not directly represent the argument structure of a predicate; e.g., the f-structure for Reagan picked flowers. would not make reference to the verb pick subcategorizing for an agent Reagan and a patient flowers. LFG includes a separate theory called mapping theory explaining how arguments of a given predicate (which is part of lexical semantics) map onto grammatical functions (syntax). For this thesis, mapping theory is of only minor importance, and I will make only indirect reference to it, the only exception being the short discussion of locative inversion in §6.2.1, where all relevant concepts are introduced on the spot.

Seminal papers advancing mapping theory in LFG are Bresnan and Kanerva (1989), Bresnan and Zaenen (1990), Bresnan and Moshi (1990); for a thorough discussion of LFG’s mapping theory as well as additional references, see, e.g., Chapter 5 of Butt (2006).
2.4. LEXICAL-FUNCTIONAL GRAMMAR

Dalrymple (2001) discusses several natural subgroups of the GFS in (26). First, the 
governable GFS SUBJ, OBJ, OBJ\(_\theta\), COMP, XCOMP and OBL\(_\theta\) can be subcategorized for by a
predicate, while the non-governable (modifier) functions ADJUNCT and XADJUNCT cannot
be subcategorized (e.g., not selected for by a verb as one of its arguments). Note also
that since all GFS are subject to the uniqueness requirement (meaning that they can only
occur once in any given f-structure), governable GFS are unique f-structures, while non-
governable GFS are features that have sets as their values. This reflects the fact that there
can only be a single instance of a particular governable GF per predicate (e.g., there can
only be a single OBJ for a particular verb), while there may be several instances of non-
governable GFS per predicate (e.g., there may be several ADJUNCTS for a particular verb).

Second, the core arguments SUBJ, OBJ, OBJ\(_\theta\) (also referred to as terms) can be distin-
guished from the family of the oblique functions OBL\(_\theta\) (also called nonterms) by ways of
their crosslinguistic behavior with respect to anaphors and control.

Third, the SUBJ and OBJ function are taken to be semantically unrestricted, which means
that they are not tied to realizing a particular semantic role, while the OBJ\(_\theta\) and OBL\(_\theta\)
functions are semantically restricted (as alluded to by the \(\theta\) sign next to them).

A final distinction made by Dalrymple (2001) is between open (indicated by the initial
X on the GF label) and closed GFS. Open GFS such as the clausal XCOMP function have
their subject controlled by an external argument, e.g., in control verbs such as English
want in Peter wants a new car.

Bresnan and Kanerva (1989) in addition distinguish objective from non-objective GFS.
According to them, OBJ and OBJ\(_\theta\) are objective functions, i.e., functions that intuitively
share properties that are commonly associated with objects across languages, while SUBJ
and OBL\(_\theta\) are non-objective in that they do not come with those properties. The features
of semantic restrictedness (\(\pm r\)(stricted)) and objectivity (\(\pm o\)(jective)) can be used as a
means to consistently identify the four clausal GFS SUBJ, OBJ, OBJ\(_\theta\) and OBL\(_\theta\) as in Table
2.3.

<table>
<thead>
<tr>
<th></th>
<th>(-r)</th>
<th>(+r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-o)</td>
<td>SUBJ</td>
<td>OBL(_\theta)</td>
</tr>
<tr>
<td>(+o)</td>
<td>OBJ</td>
<td>OBJ(_\theta)</td>
</tr>
</tbody>
</table>

Table 2.3: Identification of GFS in LFG by means of \(\pm o\) and \(\pm r\)

2.4.5 F-STRUCTURE WELLFORMEDNESS CONDITIONS

Three general well-formedness conditions are enforced on all f-structures: completeness,
coherence and consistency (Dalrymple, 2001, Kaplan and Bresnan, 1982). These condi-
tions are briefly described in the following three paragraphs.
2.4. LEXICAL-FUNCTIONAL GRAMMAR

2.4.5.1 Completeness

The sentence in (27a) is ill-formed. The PRED feature in the lexical entry in (25a), repeated here as (27b), defines the predicate see to govern two separate GFs, namely a SUBJ and an OBJ. However, (27b) does not include the second argument. LF predicts this ill-formedness by the Completeness requirement as stated in (28).\(^{14}\)

(27)  
\[
\begin{align*}
\text{a.} & \quad *\text{Nina saw.} \\
\text{b.} & \quad \text{saw} \quad \text{(↑ PRED)} = \langle \text{↑ OBJ}, (↑ \text{OBJ}) \rangle \\
& \quad (↑ \text{TENSE}) = \text{past}
\end{align*}
\]

(28) Completeness:

An f-structure is locally complete if and only if it contains all the governable GFs that its predicate governs. An f-structure is complete if and only if it and all its subsidiary f-structures are locally complete.

The f-structure for (27a) is not locally complete: it includes only a single governable GF, namely a SUBJ Nina, but an additional OBJ GF is governed by its predicate. Since Completeness is concerned with predicates governing GFs, it is always checked on the PRED features of f-structures.

2.4.5.2 Coherence

On the other hand, the Coherence principle in LF rules out additional GFs which are not specified in the governable GF list of a predicate. Assume the lexical entry in (29b) for laughed, specifying a single SUBJ GF. In the f-structure for the sentence in (29a), there is an additional NP which is not listed in the PRED feature inside the lexical entry for laughed. The Coherence principle as defined in (30) effectively rules out any GFs that are not specified in the GF list of the predicate.

(29)  
\[
\begin{align*}
\text{a.} & \quad *\text{Nina laughed Nadya.} \\
\text{b.} & \quad \text{laughed} \quad \text{(↑ PRED)} = \langle (↑ \text{SUBJ}) \rangle \\
& \quad (↑ \text{TENSE}) = \text{past}
\end{align*}
\]

(30) Coherence:

An f-structure is locally coherent if and only if all the governable GFs that it contains are governed by a local predicate. An f-structure is coherent if and only if it and all its subsidiary f-structures are locally coherent.

The f-structure for (29a) is not locally coherent. Assuming that the additional NP Nadya is mapped onto a governable GF, that GF is not governed by a local predicate: the PRED

\(^{14}\)The definitions in all of these paragraphs are taken from Dalrymple (2001, p. 37), who follows Kaplan and Bresnan (1982).
feature in (29b) includes only a SUBJ which is already filled. Again, Coherence is concerned with predicates and their governable GFS; thus, it is checked on the PRED feature listing these GFS.

2.4.5.3 CONSISTENCY

The Consistency condition ensures that attributes in f-structures will always have only a single value, not more. To see Consistency at work, consider English verbal agreement; the finite verb has to agree with the subject in person and number. The lexical entry for sees in (31b) defines the subject’s number attribute as singular, while the lexical entry for boys in (31c) specifies its number attribute as plural. Now consider the ungrammatical sentence in (31a) with failed verbal agreement. The f-structure for this sentence is not consistent: when the functional information in (31c) is projected up to the SUBJ f-structure (as part of the subject NP), the values for the feature NUMBER coming from (31b) and (31c) clash. Consistency is defined as in (32).

(31)  
   a. *The boys sees Nina.
   b. sees   V  (↑ PRED) = ‘see<(↑ SUBJ), (↑ OBJ)>’
           (↑ TENSE) = present
           (↑ SUBJ NUMBER) = sg
           (↑ SUBJ PERS) = 3
   c. boys   N  (↑ PRED) = ‘boy’
           (↑ NUMBER) = pl
           (↑ PERS) = 3

(32) Consistency:

In a given f-structure a particular attribute may have at most one value.

During unification, matching values of features are unified, as discussed above. For example, boy is specified for singular number and third person; these feature values will unify with the corresponding feature values specified in (31b), and the consistent f-structure for The boy sees Nina will include a single NUMBER and a single PERS feature in the SUBJ f-structure, specified for the values sg and 3, respectively.\(^\text{15}\)

This concludes my general overview of the LFG framework. Throughout the chapters to follow, I will make additional reference to LFG-specific constructs and principles.

\(^\text{15}\) PRED features constitute an exception to the general mechanism of unification. Values of PRED are uniformly treated as semantic forms (see §2.4.3), and semantic forms are uniquely instantiated. Thus, PRED features are never unified, and more than a single PRED feature in a given f-structure will result in inconsistency.
2.5 XLE

2.5.1 General Description

XLE is a powerful NLP tool which is used to parse and generate text using computational grammars based on the LFG framework. It is used for writing and testing small toy grammars which aim at modeling certain syntactic phenomena within LFG, but also for large-scale processing of larger amounts of text using broad-coverage grammars (Butt et al., 1999b, Kaplan and King, 2003, Kuhn, 1998). In particular, XLE forms the basis for the grammar development effort within the ParGram project (§ 2.6). It comes with a graphical user interface, is written in C and runs on Linux, Solaris and Mac OS X machines (Crouch et al., 2015). Many c- and f-structures that appear in this thesis were produced by XLE; the syntactic analyses presented in this thesis was implemented using the XLE system. Crouch et al. (2015) is a documentation for XLE.

Two techniques make XLE an efficient parser for computational LFG grammars. First, XLE pays careful attention to the interface between phrasal and functional constraints (essentially the φ function described in § 2.4.3). It combines a chart parser which processes the phrasal constraints, then uses the result of the chart parse to decide which functional constraints to process (Maxwell III and Kaplan, 1995). Thus, XLE takes advantage of the mostly context-free nature of languages by employing a chart parser as a context-free backbone. This feature ensures that feature unification can be efficiently calculated (Crouch et al., 2015, Maxwell III and Kaplan, 1996). The second key technique regards ambiguity, where XLE uses contexted unification to merge multiple feature structures together into a single, packed feature structure (King et al., 2000, Maxwell III and Kaplan, 1991).

Phrasal and functional description in terms of c- and f-structure, respectively, is the main objective of XLE; most computational grammars developed using XLE confine themselves to specifying constraints for these two levels of representation (Butt et al., 1999b). XLE further provides extensions for pre-/postprocessing as well as a system for rewriting f-structure files into output terms or facts. The preprocessing extensions include a facility for integrating additional modules such as tokenizers and morphological analyzers (Kaplan et al., 2004). F-structure rewriting (also referred to as transfer) has been used extensively in applications built on top of XLE, such as machine translation (Avramidis and Kuhn, 2009, Frank, 1999), question-answering (Bobrow et al., 2007) or semantic reasoning (Crouch and King, 2006, Zarrieß, 2009). The overall parsing architecture of XLE is depicted in Figure 2.7 (adapted from Butt et al., 1999b). Figure 2.7 excludes the transfer system which is a separate module operating on f-structure output.

According to Figure 2.7, XLE uses sets of LFG grammar rules, one or more LFG lexicons, as well as a set of finite-state machines (tokenizers and morphological analyzers) for parsing and generating sentences. While Figure 2.7 proceeds in the parsing direction, generation is the inverse of the process, starting with an analysis in terms of the
Figure 2.7: XLE parsing architecture (adapted from Butt et al., 1999b)
2.5. XLE

f-structure and ending with output in terms of sentences. Grammar writers can further extend the pipeline by including more finite-state tools such as guessers or script normalizers (“Other transducers” in Figure 2.7).16

2.5.2 Grammar Code Example

An exemplary XLE grammar rule set and XLE lexicon is given in (33) and (34), respectively. The XLE notation is very similar to the theoretical LFG notation (cf. the rule set in (24)). The metavariables ↑ and ↓ are represented using the caret sign (^) and the exclamation mark, respectively. In the rules, the star sign (*) marks categories that may appear zero or more times; the operator is well-known from regular expression logic as the Kleene star. In the lexicon, the * sign has a different meaning; here, * is a so-called morphcode. Morphcodes determine where the word is looked up in the grammar. If the morphcode is a star (*), XLE will not use an external morphology module to look up the word. If it is anything other than a star (normally, XLE is used in this case), a morphology module defined elsewhere in the grammar is used — i.e., the word matches morphemes that come out of the morphology (Grouch et al., 2015, Kaplan et al., 2004). Lastly, the functional annotation ! $ has the result that the feature structure provided by the annotated phrase (the PP in (33b) and (33c)) is put in a set notation with the indicated identifier (ADJUNCT in the examples). The equivalent LFG notation is ↓ ∈. Table 2.4 gives an overview of the XLE notation vis-à-vis traditional LFG notation.17

(33) a. S --> NP: (^ SUBJ) = !
   (! CASE) = nom;
   VP.
b. VP --> V
   (NP: (^ OBJ) = !
    (! CASE) = acc)
   PP*: ! $ (^ ADJUNCT).
c. NP --> (D)
   N
   PP*: ! $ (^ ADJUNCT).
d. PP --> P
   NP: (^ OBJ) = !
   (! CASE) = acc.

16 Transducers are two-sided finite-state machines with a lower side (usually referred to as the surface/input side) and an upper side (usually referred to as the analysis/output side). Transducers have proven helpful in dealing with different aspects of NLP, such as tokenization, morphological analysis etc. (Beesley and Karttunen, 2003,Jurafsky and Martin, 2008, among others). Arcs labeled FSM in Figure 2.7 indicate that the output of the particular module is a finite-state machine (FSM); internally, much of XLE’s processing is carried out on finite-state machines, which are highly efficient computational models.

17 The table is taken from a page within the XLE documentation at http://ling.uni-konstanz.de/pages/xle/doc/notations.html#NoA.
<table>
<thead>
<tr>
<th>LFG notation</th>
<th>XLE-equivalent</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>↑</td>
<td>~</td>
<td>f-structure metavariable</td>
</tr>
<tr>
<td>↓</td>
<td>!</td>
<td>f-structure metavariable</td>
</tr>
<tr>
<td>=</td>
<td>=</td>
<td>defining equality</td>
</tr>
<tr>
<td>= c</td>
<td>=&lt;c&gt;</td>
<td>constraining equality</td>
</tr>
<tr>
<td>∈</td>
<td>$</td>
<td>set membership</td>
</tr>
<tr>
<td>¬</td>
<td>~</td>
<td>negation (complementation)</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
<td>existential constraint (standard notation)</td>
</tr>
<tr>
<td>∃ d</td>
<td>d</td>
<td>existential constraint (Sadler notation)</td>
</tr>
<tr>
<td>&lt;-</td>
<td>&lt;-</td>
<td>off-path constraint</td>
</tr>
<tr>
<td>→</td>
<td>-&gt;</td>
<td>off-path constraint</td>
</tr>
<tr>
<td>⊑</td>
<td>&lt;&lt;</td>
<td>subsumption (subsumes)</td>
</tr>
<tr>
<td>⊒</td>
<td>&gt;&gt;</td>
<td>subsumption (is subsumed by)</td>
</tr>
<tr>
<td>{ a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>(a)</td>
<td>{ a }</td>
<td>optional f-structure constraint</td>
</tr>
<tr>
<td>symbol_</td>
<td>symbol_</td>
<td>instantiation</td>
</tr>
<tr>
<td>not used</td>
<td>&lt;h</td>
<td>head precedence</td>
</tr>
<tr>
<td>not used</td>
<td>&gt;h</td>
<td>head precedence</td>
</tr>
<tr>
<td>not used</td>
<td>&lt;s</td>
<td>scope relation</td>
</tr>
<tr>
<td>not used</td>
<td>&gt;s</td>
<td>scope relation</td>
</tr>
<tr>
<td>not used</td>
<td>$&lt;h&lt;s</td>
<td>surface adjunct scope</td>
</tr>
<tr>
<td>not used</td>
<td>$&lt;h&gt;s</td>
<td>surface adjunct scope</td>
</tr>
</tbody>
</table>

**Table 2.4:** LFG vis-à-vis XLE notations

(34) a. Mary N * (^ PRED) = 'Mary'
     (∼ NUMBER) = sg.

b. girl  N * (^ PRED) = 'girl'
     (∼ NUMBER) = sg.

c. telescope  N * (∼ PRED) = 'telescope'
     (∼ NUMBER) = sg.

d. the  D * (∼ SPEC DET PRED) = 'the'
     (∼ SPEC DET DET-TYPE) = def.

e. with  P * (∼ PRED) = 'with<(∼ OBJ)>'

f. saw  V * (∼ PRED) = 'see<(∼ SUBJ)(∼ OBJ)>'
     (∼ TENSE) = past.

Using the rule set in (33) in combination with the set of lexical entries in (34), XLE is able to process the ambiguous English sentence Mary saw the girl with the telescope.
and produces the packed f-structure in Figure 2.8. This structure is packed in the sense that it includes both of the two distinct readings of the sentence and displays the choice space; each choice is given a named variable (Crouch et al., 2015; King et al., 2000) and is visualized in the user interface using colors. Here, the choice is in the attachment of the PP adjunct with the telescope, which may either modify the object NP the girl or the whole VP saw the girl. The first choice is implemented in the NP rule in (33c), while the second choice is implemented in the VP rule in (33b). 

XLE labels these choices a2 and a1, respectively; in the packed f-structure, either of the two solutions can be selected by the user, which disambiguates the sentence and causes XLE to display the selected c- and f-structure.

"Mary saw the girl with the telescope"

![Figure 2.8: Packed f-structure example](image)

2.5.3 Notational Conventions

In this section, I discuss several notational features within XLE that are of central importance for the remainder of this dissertation. These features are constraining equations (as opposed to defining equations), existential constraints, inside-out function application (as opposed to the standard outside-in function application), the shuffle operator for free word order, meta-categories and an extension of XLE to include an optimality theory (OT) structure (via so-called OT marks). All of these notations are also used in theoretical LFG, but since I will be concerned with the treatment of nominal predicates and their arguments from the perspective of grammar writing, it is more suitable to explain these concepts from the XLE (i.e., implementational) point of view.

The analysis of the determiner the reflects the ParGram analysis of determiners in general as involving a SPEC feature. For a general introduction to the ParGram project, refer to §2.6. The SPEC feature is explained in §2.6.4. In comparison to the LFG notation in (24–25), the annotations in (33–34) were also enriched by NUMBER and CASE features, yielding a more expressive f-structure.
2.5.3.1 Constraining Equations

Standard defining equations in XLE look as in (35a). As the name of the notation suggests, defining equations are used to equate the two sides of the = sign. For example, if there is an attribute to the left of =, and a value to the right of =, the attribute is defined to bear that value. The lexical entry for the headword Mary, repeated from (34a) in (35b), includes two defining equations: (˘ PRED) = 'Mary' as well as (˘ NUMBER) = sg.

\[(35) \quad \begin{align*}
    & a. \ d1 = d2 \\
    & b. \ Mary \ N * (˘ PRED) = 'Mary' \\
    & \quad (˘ NUMBER) = sg.
\end{align*}\]

Constraining equations differ in their interpretation. These are not used for assigning values to attributes, but instead require that the given equality be defined elsewhere. The smaller-case letter c is used for constraining equations, as in (36a). A standard application for constraining equations is agreement. In English, verbal agreement is enforced between the subject and the verb. Thus, the lexical entry for the English third person singular verb form sees might look as in (36b), which defines the TENSE feature to bear the value pres (for present tense), and further checks on the subject’s PERS and NUM features to bear the values 3 and sg, respectively.\(^{19}\)

\[(36) \quad \begin{align*}
    & a. \ d1 =c d2 \\
    & b. \ sees \ V * (˘ PRED) = 'see<(˘ SUBJ)(˘ OBJ)>' \\
    & \quad (˘ TENSE) = pres \\
    & \quad (˘ SUBJ PERS) =c 3 \\
    & \quad (˘ SUBJ NUM) =c sg.
\end{align*}\]

In practical LFG grammar writing, it is important to keep in mind the difference between defining equations and constraining equations, as they can result in formally distinct parses. I will be using both types of equations in the course of the dissertation.\(^{20}\)

2.5.3.2 Existential Constraints

One can define an attribute by itself on an annotation, i.e., without providing a value for it. In XLE, this is called an existential constraint, and simply looks as in (37a). An

\[\text{\textsuperscript{19}In this particular case, if lines three and four of (36b) were defining equations, the process of unification would rule out any non-third person singular subjects, and the computational result would essentially be the same: only third person singular subjects would be compatible with this verb form. However, that alternative solution would be linguistically implausible in that it would essentially state that the verb assigns the subject’s number and person features. Moreover, there may be other cases of agreement (or different linguistic phenomena) where the notational difference does bear on the formal outcome, in that the features which are assigned may turn up on f-structures which are themselves not marked for these features.}\]

\[\text{\textsuperscript{20}Documentation on the equality notation can be found as part of the XLE documentation: } \text{http://ling.\texttt{uni-konstanz.de/pages/xle/doc/notations.html\#N4.2.2}}\]
existential constraint is solved if the designated attribute has any denotation in the f-structure associated to the annotated node. An example of an existential constraint would be to require all English sentences to be tensed; this is achieved by altering the rule in (33a) as in (37c), where the existential constraint (ˆ TENSE) annotated on the VP node requires that there be such a feature (but does not further specify the feature’s value). Intuitively, negative existential constraints require the indicated feature to not appear in the annotated node’s f-structure.21

(37) a. d  
b. ~d  
c. S --> NP: (ˆ SUBJ) = !  
    (! CASE) = nom;  
    VP: (ˆ TENSE).

2.5.3.3 HEAD PRECEDENCE

To enforce restrictions on c-structure precedence over functional categories, grammar writers can use the head precedence notation (<h or >h). Consider (38).

(38) a. f1 <h f2  
b. f2 >h f1  

(38a) is true if and only if f1 and f2 have heads and the head of f1 precedes the head of f2 in the c-structure.22 >h is the inverse of <h, so that (38a) and (38b) mean exactly the same thing. Since f-precedence (Zaenen and Kaplan, 1995) is currently not implemented in XLE, head precedence can be used as a substitute if both f-structures have heads.23

2.5.3.4 INSIDE-OUT FUNCTION APPLICATION

Outside-in function applications in XLE annotations typically include an f-structure metavariable (ˆ or !), and a path of attributes that is traversed to the right in the f-structure. The path may be arbitrarily long. For example, in the first line of the lexical entry for the English determiner the (repeated from (34d) in (39)), the path starts from the f-structure denoted by ˆ (here, the f-structure corresponding to the D node), moves to the right along the attributes specified (SPEC, DET, PRED) and assigns a value (‘the’) to the thus defined complex f-structure SPEC DET PRED.

(39) the D * (ˆ SPEC DET PRED) = 'the'  
(ˆ SPEC DET DET-TYPE) = def.

---

21See also the XLE documentation at http://ling.uni-konstanz.de/pages/xle/doc/notations.html#N4.1.4.

22Here, the “head” of an f-structure is the constituent where the f-structure’s PRED is instantiated.

23See also http://ling.uni-konstanz.de/pages/xle/doc/notations.html#N4.1.4.
A grammar writer can ‘inverse’ this process of path traversal using inside-out function application. Inside-out function application notations give a path of attributes to be traversed to the left, followed by an f-structure metavariable. Again, the f-structure path may be arbitrarily long. For example, the lexical entry in (40) for the French attributive adjective belle ‘beautiful’ (inflected for feminine singular from its citation form beau) includes two constraining equations with inside-out function applications. In the second line, the inside-out path starts from the f-structure denoted by $^\sim$ (here, the ADJUNCT set element corresponding to the A node) and moves to the left along the attribute specified (here, along ADJUNCT, as the adjective is a member of an ADJUNCT set). In other words, the inside-out path points from the embedded f-structure of the adjective adjunct set element to the f-structure containing the overall adjunct set. Then, the constraint uses the usual outside-in strategy to check the GEND attribute of the inside-out path. In sum, both of the constraints are satisfied only if $^\sim$ is a member of ADJUNCT contained inside an f-structure that has GEND fem and NUM sg.

(40) belle A * ($^\sim$ PRED) = 'beau'
     ((ADJUNCT $^\sim$) GEND) =c fem
     ((ADJUNCT $^\sim$) NUM) =c sg.

Outside-in function application is useful for describing or constraining f-structures that are a part of the currently annotated node. Inside-out function application, on the other hand, is useful for describing or constraining f-structures that the currently annotated node are a part of.

2.5.3.5 FUNCTIONAL UNCERTAINTY

Some phenomena in natural languages such as topicalization, relative clauses and wh-questions introduce long distance dependencies. Those are constructions where a constituent can be arbitrarily distant from its governing predicate. An example is the wh-question in (41), where the wh word at the beginning of the clause is an object of the predicate see embedded in a complement clause further down in the structure.

(41) What did she fear she would have to see tomorrow?

In order to express such constraints involving unbounded embeddings, LFG uses functional equations with paths through f-structures written as regular expressions. Such equations are referred to as functional uncertainty (FU) equations and are implemented in XLE. (42) is an example; the functional uncertainty path annotated on the NP in the first line, ($^\sim$ {XCOMP|COMP}* OBJ) enables objects to topicalize out of an arbitrarily long path consisting of XCOMPs and/or COMPs (occurring in any order), so that sentences such as The Aston Martin I thought she wanted you to buy. can be parsed.

---

\textsuperscript{24} See also the XLE documentation at http://ling.uni-konstanz.de/pages/xle/doc/notations.html#N.4.1.58.

\textsuperscript{25} Cf. the Kleene star in the annotation.
The regular operators available for the FU path language are drawn from an XLE-wide set of regular operators: these are concatenation, disjunction, one-or-more (+), zero-or-more (Kleene *), etc. The range of operators allows for a precise formulation of the nature of the dependency.

2.5.3.6 Shuffle Operator

The shuffle operator (already mentioned in §2.3.1) is notated using the comma (","). The operator can be used in rules to connect up nodes that may occur in any linear order with respect to each other. For example, the S rule in (43) is satisfied by strings that contain a subject NP and a VP in either order.

\[
(43) \quad S \rightarrow NP: (^{\text{SUBJ}}) = !, \\
\text{VP}.
\]

The shuffle operator is frequently used for modeling free word order. Note that the compiled XLE grammar can become quite large if many elements are connected using the shuffle operator. This is due to the fact that XLE produces all possible orders, and the number of possible orders is factorial to the number of elements. The operator was first suggested within HPSG by Reape (1994) in his analysis of German word order and subsequently taken up by LFG theory (p.100 Dalrymple, 2001).

2.5.3.7 Meta-Categories

Using a meta-category enables grammar writers to encode cross-categorial generalizations in XLE. In English, for example, one may want to associate the open-complement function XCOMP with any of the constituents AP, PP, NP, or VP. Thus, one could write the VP rule as shown in (44).

\[
(44) \quad S \rightarrow (NP: (^{\text{XCOMP|COMP}})* \text{OBJ}) = ! \\
(\text{^TOPIC} = !) \\
\text{NP: (^{\text{SUBJ}}) = !} \\
(\text{!* CASE} = \text{nom}; \\
\text{VP}.
\]

---

26Document for the implementation of FU in XLE is at [http://ling.uni-konstanz.de/pages/xle/doc/notations.html\#N].

27More information on the shuffle operator is given here: [http://ling.uni-konstanz.de/pages/xle/doc/notations.html\#N].

28This example was taken from the XLE documentation on meta-categories at [http://ling.uni-konstanz.de/pages/xle/doc/notations.html\#N].
To model the common functional association of the nodes with XCOMP, one could choose to rewrite the rule as in (45a), pointing to another c-structure node XP, which rewrites to either of the original constituents (45b). However, this notation results in an XP node in the c-structure: an extra, presumably unwanted and otherwise unmotivated constituent.

(45) a. \[ \text{VP} \rightarrow \text{V} \\
\quad (\text{NP: } (^\_\text{OBJ}) = !) \]
\[ \quad (\text{XP: } (^\_\text{XCOMP}) = !) \]

b. \[ \text{XP} \rightarrow \{\text{AP}|\text{PP}|\text{NP}|\text{VP}'\} \].

Instead, the grammar engineer can specify XP as a meta-category, and the abbreviated version of the VP rule would still produce the original tree. A meta-category is notated differently from a standard category by a very small change in the notation of the rule: The right and left sides of the rule are separated by an equal-sign instead of a rewriting arrow, as in the example in (46).

(46) \[ \text{XP} = \{\text{AP}|\text{PP}|\text{NP}|\text{VP}'\} \].

Thus, meta-categories work like placeholders, in that they represent a means of collecting in a single place a specification of category sequences that would otherwise be repeated in several rules; they can be used to express generalizations over many grammatical constructions without adding unmotivated c-structure nodes.

2.5.3.8 LOCAL VARIABLES

Local variables are variables whose scope is limited to the annotation where the variable is defined.\[29\] Using such variables, the grammar writer can make repeated references to a single entity (e.g., a path to an f-structure). When an FU path is involved, local variables can be used to impose several constraints on a single instantiation of the path. An example involving the rule from (42) is shown in (47).

\[\text{\[29\]The notation is also referred to as} \text{local name} \text{in the XLE documentation; see} \text{http://ling.}\]
\[\text{uni-konstanz.de/pages/xle/doc/notations.html#N4.i.6}\]
(47)  S -->(NP: (^{XCOMP|COMP}* OBJ) = %PATH
    %PATH = !
    (%PATH SPEC)
    (%PATH GEND) = c fem
    (^ TOPIC) = !)
NP: (^ SUBJ) = !
    (! CASE) = nom;
VP.

%PATH is a local variable defined as a single instantiation of the FU path defined. As such, all of the constraints below the variable's definition will be enforced on the instantiation. Thus, this particular annotation would state that only feminine objects with a specifier are allowed to be topicalized. ³²

2.5.3.9 C-STRUCTURE METAVARIABLES

In addition to the f-structure metavariables ! (the f-structure of the currently annotated node) and ^ (the f-structure embedding the f-structure of the currently annotated node), there are c-structure metavariables that can be used to denote nodes in the c-structure tree and enforce constraints on categories. These notations can be helpful when testing for the presence or absence of nearby nodes. (48) shows c-structure metavariables that are commonly used.

(48)  a. *
    b. M*
    c. LS*
    d. RS*

When (48a) appears in a schema (e.g., an annotation), it denotes the c-structure node matching the annotated category. (48b) matches the mother node of the annotated category. LS* matches the left sister node, RS* the right sister node.

2.5.3.10 OT MARKS

In XLE, grammar developers can model statistical generalizations using special marks that were inspired by Optimality Theory (OT, Prince and Smolensky, 2004). On top of the classical constraint system of existing LFG grammars (formulated in terms of feature

³⁰An equivalent annotation could not be stated without a local variable, since each of the constraints on the NP could be enforced on a different instantiation of the FU path.

³¹The f-structure metavariables ! and ^ are exactly equivalent to the terms f::* and f::M*, where f:: signifies the c-structure to f-structure correspondence function otherwise known as φ in LFG. One can also denote the weight of a c-structure node by using the *WEIGHT designator. See http://ling.uni-konstanz.de/pages/xle/doc/notations.html#N for a more complete documentation on c-structure metavariables.
constraints stated on the f-structure), a separate projection, o-structure, determines a preference ranking on the set of analyses for a given input sentence. A relative ranking is specified for the constraints that appear at o-structure, and this ranking serves to determine the winner among the competing candidates. The constraints are also referred to as OT marks and are overlaid on the existing grammar in such a way that they do not fundamentally alter the basic tenets of LFG theory. OT marks, and the statistical preferences they encode, are usually derived from either corpus analysis or native speaker introspection (Frank et al., 1998).

OT marks can be added in the appropriate place in the grammar (e.g., in a lexicon entry, template or rule) to punish or prefer a certain analysis over competing ones. For example, (49) states that Mark1 is a member of the optimality projection.

(49) ... Mark1 $ o::* ... 

The order of preference of OT marks over one another needs to be specified in the configuration section of the grammar in a list called OPTIMALITYORDER. An example list is given in (50). The list defines the priority of the marks relative to one another. In this case Mark4 is the most important, and Mark1 is the least important. Marks that have a + sign in front of them are preference marks. The more preference marks that an analysis has, the better. All other marks are dispreference marks (the fewer, the better).

(50) OPTIMALITYORDER Mark4 Mark3 +Mark2 +Mark1.

Given the ranking in (50), an analysis which includes the preference mark in (49) will be preferred over a competing analysis which includes any of the dispreference marks. A separate list for generation can be defined under the heading GENOPTIMALITYORDER. This makes sense since grammar writers may want to formulate grammars to be more permissive in parsing, but rather strict in generation, and so both the number and priority of OT marks may differ between the two applications.

2.5.4 Templates

Templates are of central importance for writing LFG grammars with XLE and are very useful for defining generalizations that hold across families of lexical entries and/or rules. Templates, and the idea of factoring out redundancies and generalizations that apply across lexical entries are in fact a central idea in LFG. XLE’s functional templates, as well as lexical rules, provide mechanisms for capturing generalizations that would otherwise be distributed throughout the lexicon and grammar. The notations allow for the precise control of lexical and syntactic behavior that explicit computational tests

\footnote{For more information on the implementation of OT marks in XLE, including examples of preference calculations if multiple marks interact, refer to the XLE documentation at \url{http://ling.uni-konstanz.de/pages/xle/doc/xle.html#SEC4.}}
of large-scale LFG grammars require \cite{Crouch2015,King2005}\footnote{I will make use of templates throughout those parts of this thesis that are concerned with the implementation of Hindi/Urdu nouns and their predicational properties. I introduce the basic concept of templates and their notations in XLE in this section.} I will make use of templates throughout those parts of this thesis that are concerned with the implementation of Hindi/Urdu nouns and their predicational properties. I introduce the basic concept of templates and their notations in XLE in this section.

2.5.4.1 TEMPLATES IN LEXICAL ENTRIES

Redundancies in the lexicon can be factored out by means of functional templates. These work like macros in that they permit the name assigned to a complex expression to be used in place of that expression in a lexical entry. For example, many English nouns are like girl and telescope in that they have a \texttt{PRED} feature (the value of which is a semantic form, enclosed in single quotes) and are singular \footnote{For theoretical takes on the concept of templates in LFG, see, e.g., Dalrymple et al. (2004b) and Asudeh et al. (2009)}\footnote{Detailed documentation on the notation of templates in lexical entries is given here: \url{http://ling.uni-konstanz.de/pages/xle/doc/notations.html#N}}. One could then choose to define a template as in (51a), which here is called \texttt{NSG} (for \textit{noun singular}). (51a) is called a parametrized functional template in that it takes an argument, \texttt{P}, which is substituted as the value of the \texttt{PRED} feature (i.e., the semantic form). In the lexical entries, the template can be called by using the \texttt{@} sign, followed by a parenthetic expression with the name of the template and a sequence of values to be substituted for the parameters in the template definition (here, \texttt{P}).

\begin{enumerate}
\item \texttt{NSG}(_\texttt{P}) = (\^ \texttt{PRED}) = '_\texttt{P}'
\item \texttt{NUMBER} = sg.
\item girl N * @\texttt{NSG} girl.
\item telescope N * @\texttt{NSG} telescope.
\end{enumerate}

Another example of a template is given in (52a). This template models transitive verbs in that it defines a \texttt{PRED} feature whose value is a semantic form selecting for a \texttt{SUBJ} as well as an \texttt{OBJ}. The template can then be called up by transitive verbs as in (52b–c).

\begin{enumerate}
\item \texttt{TRANS}(_\texttt{P}) = (\^ \texttt{PRED}) = '_\texttt{P}<(\^ \texttt{SUBJ})(\^ \texttt{OBJ})>'
\item saw V * @\texttt{TRANS} see
\item baked V * @\texttt{TRANS} bake
\end{enumerate}

(51) \begin{enumerate}
\item \texttt{NSG}(_\texttt{P}) = (\^ \texttt{PRED}) = '_\texttt{P}'
\item \texttt{NUMBER} = sg.
\item girl N * @\texttt{NSG} girl.
\item telescope N * @\texttt{NSG} telescope.
\end{enumerate}

2.5.4.2 TEMPLATES IN RULES

Functional templates can also be called in rules. For example, the rule in (33b) could be rewritten as in (53c), where the second line adds an \texttt{NP} node to the c-structure and, in the node’s functional annotation, calls the \texttt{ACC-OBJ} template in (53a), which puts the
2.6. NOMINALS & THE PARGRAM PROJECT

functional information coming from the NP node in an OBJ f-structure of the above f-structure. Similarly, the third line in (53a) adds zero or more PP nodes in the c-structure and, via the ADJ template (53b), defines the f-structure corresponding to the PP node as a member of the ADJUNCT set.

(53)  a. ACC-OBJ = (^ OBJ) = !
      (! CASE) = acc.
  b. ADJ = ! $ (^ ADJUNCT).
  c. VP --> V
      (NP: @ACC-OBJ)
      PP*: @ADJ.

The functional templates in (53a) and (53b) are called non-parametrized templates because they do not make use of parameters, contrary to the parametrized NSG template in (51a). Both template types may in principle be called in lexicon entries as well as rules.

2.6 NOMINALS & THE PARGRAM PROJECT

The implementation work presented in this thesis is developed within the Urdu ParGram grammar, which is a computational LFG grammar that forms part of the ParGram (“Parallel Grammar”) project. I provide a description of the ParGram project in this section.

2.6.1 GENERAL DESCRIPTION

The ParGram project is a loose alliance among computational grammar writers around the world within the framework of LFG. The grammars developed in this project are parallel in the sense that they adhere to a commonly defined set of analyses and features and that they are guided by common linguistic principles (Butt et al., 1999a, 2002, 1999b, King et al. 2005, Sulger et al., 2013). In this way, ParGram aims at testing the LFG formalism for its universality and coverage limitations to see how far parallelism can be maintained across a larger array of languages. From a parsing perspective, the grammars are considered deep in the sense that they provide a linguistically informed parsing of sentences. With LFG as their theoretical basis, they not only provide a phrase structure for a specific sentence, but also register GFS (SUBJ, OBJ, etc.) and grammatical features from a commonly defined set (e.g. case, mood, noun type, voice, etc.) (King et al. 2005, Sulger et al., 2013).

35Detailed documentation on the notation of templates in rules is given here: http://ling.uni-konstanz.de/pages/xle/doc/notations.html#notations.
36The ParGram homepage is at: http://pargram.b.uib.no/
37As opposed to “deep parsing”, “shallow” parsers are used to merely annotate superficial phrase structure; an example for this type of parsing is finite-state chunking. For a differentiation, see e.g. Jurafsky and Martin (2008).
2.6. NOMINALS & THE PARGRAM PROJECT

One goal of ParGram is to abstract away as much as possible from surface phenomena to the common properties across languages; the assumption is that by taking this interlingua-type of approach, applications such as machine translation can be improved significantly (Frank, 1999). Another goal is the improvement of applications that can benefit from a deep approach to NLP. Researchers have used ParGram grammars in applications such as question answering (Bobrow et al., 2007), text summarization (Crouch et al., 2004, Riezler et al., 2003), computer-assisted language learning (Fortmann and Forst, 2004, Khader, 2003, Khader et al., 2004) and others. Initiated with only three European languages (English, French and German), ParGram is continuously growing; broad-coverage grammars exist for English (Butt et al., 1999a, King et al., 2003), German (Butt et al., 1999a, Dipper, 2003), French (Butt et al., 1999a), Japanese (Masuichi et al., 2003), Norwegian (Rosén et al., 2005) and Polish (Patejuk and Przepiórkowski, 2012); smaller-scale grammars are under development for a variety of languages, including Chinese (Fang and King, 2007), Greek (Samaridi and Markantonatou, 2014), Hindi/Urdu (Bögel et al., 2009, Butt and King, 2007), Hungarian (Laczkó, 2014a,b,c, Laczkó and Rákosi, 2010), Indonesian (Arka, 2012), Malagasy (Dalrymple et al., 2006), Turkish (Çetinoglu, 2009, Çetinoglu and Oflazer, 2006), Welsh (Mittendorf and Sadler, 2006), among others. The grammars in the ParGram project are all implemented using the XLE grammar development platform.

2.6.2 PARALLELISM

Parallelism is of central importance to ParGram. The conventions developed within the ParGram grammars are extensive and dictate the form and possible values of the features used in the grammars as well as the type of analysis chosen for a particular construction. These conventions are shared throughout ParGram via shared templates and common feature declarations that describe the feature space (King et al., 2005). Grammar writers are in principle only allowed to abandon parallelism if maintaining it would be at the cost of misrepresenting the language (Butt and King, 2007, Butt et al., 1999b, King et al., 2005).

Figures 2.9 and 2.10 display ParGram analyses of the transitive sentence in (54); i.e., the figures show how the computational LFG grammars for English and for German, respectively, analyze transitive sentences by means of f-structure. (54) is part of the recently developed treebank resource PARGRAMBANK, which features parallelized and (partially) phrase-aligned ParGram structures for ten languages from six language families (Sulger et al., 2013); PARGRAMBANK is continuously growing.

(54) Der Fahrer startet den Traktor.

‘The driver starts the tractor.’

I refrain from giving the c-structures here, since they are fully language-specific; parallelism only concerns the f-structures, since they abstract away from the languages’ idiosyncracies.
"The driver starts the tractor."

Figure 2.9: ParGram f-structure for English sentence

"Der Fahrer startet den Traktor."

Figure 2.10: ParGram f-structure for German sentence

Aside from a few differences (e.g., German has a much more elaborate case system than English), the two f-structures are identical. They both feature a main predicator PRED (see/sehen) which selects two GFS, a SUBJ and an OBJ. A minor difference is the fact that in the German grammar, a grammaticized discourse function TOPIC (Bresnan, 2001, Dalrymple, 2001) is implemented which is identified with the preverbal GF; here, this is the SUBJ.

39English and German both feature elaborate determiner systems. In English, articles and quantifiers pattern similarly in that they occur in the first position in the NP and exclude any other article or quantifier modifying the same noun (64b). These elements have been termed specifiers by syntacticians following the intuition that they serve to “specify” the noun (instead of simply modifying it) (Butt et al., 1999b).

(i) * The a/every dog barks. adapted from Butt et al. (1999b), p. 101
The common templates and common feature space of ParGram are distributed in the form of two files, `common.features.lfg` as well as `common.templates.lfg`. Grammar projects which are new to ParGram can orient themselves with the help of these files in order to be parallel with the rest of ParGram (King et al., 2005). (55a) shows an excerpt from `common.features.lfg` for the feature `VTYPE`, which differentiates between different types of verbs (copular, main, modal and raising verbs). The notation -> $ \{ \ldots \}$ specifies that the value of the feature may be any one of the atomic values inside the curly braces. (55b) is an example template from `common.templates.lfg` showing the template `VTYPE`, which can be called from a grammar as in line 2 of (55c) to annotate the `VTYPE` feature (in this case with the value `main`). Line 3 in (55c) calls the common template `TENSE` to annotate the verb form for the value `past` (one of the possible values for the `TENSE` feature defined in the common features file).

(55)  

a. `VTYPE`: -> $ \{ \text{copular} \ \text{main} \ \text{modal} \ \text{raising} \}$.  
b. `VTYPE(_val) = (\text{\textasciitilde} \text{VTYPE}) = _val`.  
c. `saw V * (\text{\textasciitilde} \text{PRED}) = 'P<(\text{\textasciitilde} \text{SUBJ})(\text{\textasciitilde} \text{OBJ})>'`  
\(\oplus(\text{VTYPE main})\)  
\(\oplus(\text{TENSE past}).\)

In addition to the common features and common templates, grammar writers can choose to define particular features, values or templates as they see fit for their language/grammar, as long as there are principled linguistic reasons behind it. All ParGram grammars include a feature declaration which lists all features used across the grammar and their possible values. This lets the grammar writer assert constraints on possible feature values or structures: if a given feature is annotated with some value that is not given for that feature in the feature declaration, XLE will print a warning.\(^{40}\)

In German, the situation is slightly different in that the elements corresponding to English quantifiers `some`, `every` etc. are inflected as adjectives and may occur alongside articles, as in (ii). At c-structure, they are thus not analyzed as specifiers, but as quantifying adjectives in the German ParGram grammar.

(ii) Ein jeder Hund bellt.  
‘Every dog barks.’

The common syntactic and semantic behavior of specifiers was captured in the ParGram grammars by positing a `SPEC` feature occupied by determiners, quantifiers and other nominal modifiers (`SPEC DET`, `SPEC QUANT` etc.). The structures in Figures 2.9 and 2.10 thus list a `SPEC DET` within `SUBJ` and `OBJ`. Thorough descriptions of the treatment of determiners and quantifiers within ParGram can be found in Butt et al. (1999b) as well as Dipper (2003) for German. The `SPEC` feature in its usage for encoding nominal arguments is described in §2.6.4.

\(^{40}\)Feature declarations are discussed in the XLE documentation at [http://ling.uni-konstanz.de/pages/xle/doc/notations.html#N](http://ling.uni-konstanz.de/pages/xle/doc/notations.html#N).
2.6.3 Nominal Features in ParGram

Nouns and their predicational properties are of central importance for this thesis. My analysis of different noun types and the way they select different GFS is embedded in the Urdu ParGram grammar (which will be presented in §2.7). This section discusses the main ingredients of the ParGram grammars concerning nouns; here, the focus is on the different syntactic and semantic features (and their possible values) attributed to different noun types as part of the ParGram common feature declaration.

2.6.3.1 NTYPE

NTYPE registers the syntactic and semantic type of the noun (Butt et al., 1999b). All nouns in the ParGram grammars are encoded with an NTYPE feature. NTYPE is a complex feature that has as its value two other features, NSYN and NSEM, as shown in (56), taken from the ParGram common feature declaration file (common.features.lfg). The notation -> << [...] specifies that the feature's value is itself a feature.

(56) NTYPE: -> << [ NSEM NSYN ].

The division between NSYN and NSEM is intuitive in that NSYN describes the nominal properties that pertain to the noun's syntax, and NSEM describes the noun's semantic properties. These two features are in turn described in the following subsections.41

2.6.3.2 NSYN

The NSYN feature specifies the syntactic type of the noun. The declaration of NSYN is given in (57). The distinction made by NSYN is basic in that it only distinguishes between common nouns, proper nouns, and pronouns. This broad distinction may be relevant for constraining some analyses in the grammars, but it is important to note that different types of common nouns may exhibit patterns quite different from one another (e.g., mass nouns vs. count nouns, etc.). This is also true of pronouns (e.g., reflexive pronouns vs. personal pronouns vs. possessive pronouns, etc.) as well as proper nouns (e.g., locations vs. person names vs. organization, etc.). Examples for English nouns which are annotated with NSYN common by the English grammar are given in (58a); different kinds of pronouns, showing up with NSYN pronoun, are in (58b); and nouns with NSYN proper are shown in (58c).

Pronouns, which are of minor importance for this thesis (as they are generally unable to take arguments, see §2.2 below), have both an NTYPE and a PRON-TYPE feature. Here, NTYPE has only an NSYN feature valued pronoun (pronouns lack the NSEM feature), and the PRON-TYPE feature is valued depending on the type of the pronoun (pers for personal pronouns, poss for possessive pronouns, etc.) (Butt et al., 1999b, p. 75–81). See the f-structure below generated by the English ParGram grammar for the personal pronoun he.

```
PRED 'he'
NTYPE [NSYN pronoun]
1 CASE nom, GEND-SEM male, HUMAN +, NUM sg, PERS 3, PRON-TYPE pers
```
2.6. NOMINALS & THE PARGRAM PROJECT

(57) NSYN: \( \rightarrow \{ \text{common pronoun proper} \} \).

(58) a. NTYPE NSYN common: all common nouns: table, coffee, love, destruction, mother
    b. NTYPE NSYN pronoun: all pronouns: she, him, herself
    c. NTYPE NSYN proper: all proper nouns: San Francisco, CSLI Publications, John

As will be shown in Chapters 3 and 5, the NSYN feature is useful when constraining which types of nouns may take arguments, and which may not. Modeling the general finding that neither pronouns nor proper nouns can predicate (see § 2.2), the grammar can be constrained in a principled manner to allow only those nouns to take arguments which are functionally annotated as NSYN common.

2.6.3.3 NSEM

The NSYN feature offers only a broad distinction in terms of syntactic type. ParGram grammar writers can further specify nouns using the NSEM feature. NSEM subsumes semantic features of nouns; these are usually features that are useful in constraining syntactic constructions, but they may just also pass information on to applications. (59) shows the feature declaration of NSEM. Again, as shown by the notation, the possible values for NSEM are themselves features: COMMON (§ 2.6.3.4), NUMBER-TYPE, PROPER as well as TIME.

(59) NSEM: \( \rightarrow \llbracket \text{COMMON NUMBER-TYPE PROPER TIME} \rrbracket \).

Of these sub-features, COMMON is of importance for this dissertation, since this is the feature (apart from NSYN, see § 2.6.3.2 above) which will serve to control for argument selection by the nominal. The other sub-features of NSEM are attributed to nouns that do not take arguments (see § 2.2), and so I do not discuss these features here.

2.6.3.4 NSEM COMMON

The NSEM COMMON feature distinguishes different kinds of common (non-proper) nouns. The feature is defined with the range of atomic values given in (60). Some examples of the different types of common nouns are given in (61).

(60) COMMON: \( \rightarrow \{ \text{count gerund mass measure partitive} \} \).

(61) a. NTYPE NSEM COMMON count: all count nouns: table, war, mother
    b. NTYPE NSEM COMMON gerund: gerund nouns derived from verbs: (her) doing (the exercise)
    c. NTYPE NSEM COMMON mass: mass nouns: coffee, water
    d. NTYPE NSEM COMMON measure: measure nouns: (two) kilos, (ten) miles

42 The NUMBER-TYPE, PROPER as well as TIME sub-features of NSEM are described in the ParGram Starter Notes as part of the XLE documentation at http://ling.uni-konstanz.de/pages/xle/doc/PargramStarterGrammar/starternotes.html.
2.6. NOMINALS & THE PARGRAM PROJECT

As I will discuss in Chapters 3 and 5, the ParGram feature space for nominal types is useful for realizing some of the constraints that come alongside nominal predication and argument selection in Hindi/Urdu. However, to analyze the full pattern of nominal predication in the Urdu ParGram grammar, extensions to this feature space are necessary. In particular, the noun types currently subsumed under the specification `NSEM COMMON count` have to be differentiated further in order to model the correct generalizations and constraints.

### 2.6.4 NOMINAL PREDICATES & ARGUMENTS IN PARGRAM

Butt et al. (1999b) in their survey of grammar development in the ParGram project include a discussion on nonverbal subcategorization, acknowledging the fact that predicates other than verbs may select for arguments. Butt et al. (1999b) give the examples in (62) from German and English: in (62a), the German adjective `stolz` ‘proud’ subcategorizes for an `OBL`, while in (62b) the English preposition `in` is analyzed in terms of subcategorizing for an `OBJ`.

(62)  

  The driver is proud of her tractor.
  
  Adjective: (`\text{PRED}`) = 'stolz<(`\text{OBL}`)>'  
  (Butt et al., 1999b, p. 45)

- b. The driver parked the tractor in [the garage].  
  Preposition: (`\text{PRED}`) = 'in<(`\text{OBJ}`)>'  
  (Butt et al., 1999b, p. 45)

Butt et al. (1999b) further mention the linguistic debate about nominal argument structure, i.e., to what extent nominals may select for argument GFS and what these GFS should be called (see also the discussion in §2.2). Confining themselves to possessive nominal arguments, they state that the issue regarding nominal argument structure arises from pairs such as those in (63), where the nominal modifiers of `destruction` in (63a) correspond to the verbal arguments of `destroy` in (63b).

(63)  

- a. NP: the Romans’ destruction of the city  
  (Butt et al., 1999b, p. 46)

- b. S: the Romans destroyed the city  
  (Butt et al., 1999b, p. 46)

Butt et al. (1999b) advocate a treatment in terms of either an `ADJUNCT` or a `SPEC` feature. The `SPEC` feature is the same feature that handles determiners, quantifiers and the like (cf. footnote 29, §2.6.2). The idea is to abstract away from the actual semantic relation between the noun and its modifiers and give GFS that are as general as possible; establishing the semantic connection between the noun and its modifiers would then be
left to a semantics module. However, Butt et al. (1999b) in principle remain open to a treatment that analyzes derived nouns such as destruction as retaining the verb’s exact subcategorization frame; e.g., destruction then would be assigned a SUBJ and an OBJ, and the grammar would then have to ensure that modifiers such as of the city and the Romans’ are mapped onto the correct GFSs. Nevertheless, Butt et al. (1999b) maintain that the general ParGram analysis of nominal arguments has been to treat them as adjuncts or (non-argument) specifiers to the head noun. They argue that the correct place to analyze the semantic relations between the nominal head and its modifiers in LFG is not at the level of f-structure, but rather at the level of a(rgument)-structure. Thus, in the example in (63), the common property of destruction and destroy would be that both of them have an agent and a patient role in their a-structure, while the mapping of these roles onto GFS is different across the parts of speech of the predicates: the roles map onto the SUBJ and OBJ function if the predicate is verbal, and onto ADJUNCT or SPEC if the predicate is nominal. The mapping between a- and f-structure would be realized via one of the variants of Mapping Theory (Bresnan and Kanerva, 1989, Bresnan and Moshi, 1990, Bresnan and Zaenen, 1990, Kibort, 2007, among many others). Working in a computational grammar development setting, Butt et al. (1999b) do not elaborate on what this mapping would look like.

Concerning the difference between ADJUNCT and SPEC, the idea is that the SPEC feature is reserved for prenominal possessives; the ADJUNCT function, on the other hand, should be used for postnominal of-possessives. The intuition behind this is that the prenominal possessive is incompatible with other types of prenominal modifiers, such as determiners or quantifiers, and so a specifier cannot be preceded by another specifier; examples from English are shown in (64). Butt et al. (1999b) suggest a SPEC-TYPE feature that registers the type of specifier as well as a SPEC-FORM feature with the surface form of the particular specifier used. Examples for values of SPEC-TYPE are in (65).

(64)  a. The/a/every/Kim’s small dog barks.  
      b. * The a/every/Kim’s dog barks.

(65)  Examples for SPEC-TYPE values (Butt et al., 1999b, p. 102):

<table>
<thead>
<tr>
<th>SPEC-TYPE value</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>indef</td>
<td>a dog</td>
</tr>
<tr>
<td>quant</td>
<td>every dog</td>
</tr>
<tr>
<td>neg</td>
<td>no dog</td>
</tr>
<tr>
<td>poss</td>
<td>my/John’s dog</td>
</tr>
</tbody>
</table>

Butt et al. (1999b) do not mention relational nouns or common nouns in their discussion; as was seen in §2.2, there are good reasons for assuming an argument structure for both of these types of nouns. In addition, as will be seen in Chapters 3 and 5, diagnostics

43In LFG, the principle of inconsistency blocks the occurrence of two separate values for a single feature; thus, more than one assignment of a value to SPEC will be denied (unless the values are identical, in which case they unify; or the features are PRED features, which never unify.)
show that different nominal arguments across noun types display behavior that is hard to explain by simply assuming an adjunct or specifier analysis. Dipper (2003) in her examination of the German ParGram grammar includes an updated discussion of the SPEC feature. Under this newer conception, SPEC is used as a complex feature with sub-classifying features for different types of specifiers or quantifying expressions. Depending on the type of the quantifying expression, it ends up under different f-structures. This effectively removes the need for the SPEC-TYPE, since the type of specifier can be read off the sub-classifying f-structures inside SPEC. Dipper (2003) lists the specifier sub-types in (66) as part of the German ParGram grammar. Note further that under this newer ParGram analysis, each specific SPEC sub-feature includes its own PRED feature occupied by a semantic form provided by the specifier word. Therefore, the SPEC-FORM feature is also no longer needed; if the semantics needs to make reference to the specific form of specifier, it can be read off the PRED feature.

(66) Specifier sub-types for German (Dipper, 2003, p. 184):

<table>
<thead>
<tr>
<th>Sub-classifying f-structure</th>
<th>Specifier type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEC DET</td>
<td>Articles</td>
</tr>
<tr>
<td>SPEC POSS</td>
<td>Possessives</td>
</tr>
<tr>
<td>SPEC DEM</td>
<td>Demonstratives</td>
</tr>
<tr>
<td>SPEC QUANT</td>
<td>Indefinites</td>
</tr>
<tr>
<td>SPEC INT</td>
<td>Interrogatives</td>
</tr>
<tr>
<td>SPEC NUMBER</td>
<td>Cardinals</td>
</tr>
</tbody>
</table>

Dipper (2003)’s description mirrors the now standard ParGram analysis of specifiers in general and possessives in particular: prenominal possessives are in SPEC POSS, postnominal ones are ADJUNCTS. In the most recent version of the German grammar that I have access to, this feature space has been reduced to the one in (67), where SPEC DET and SPEC DEM have been merged into SPEC DET, and SPEC INT has been merged into the other features of SPEC.

(67) Specifier sub-types for German (2009 grammar):

<table>
<thead>
<tr>
<th>Sub-classifying f-structure</th>
<th>Specifier type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEC DET</td>
<td>Articles/demonstratives</td>
</tr>
<tr>
<td>SPEC POSS</td>
<td>Possessives</td>
</tr>
<tr>
<td>SPEC QUANT</td>
<td>Indefinites</td>
</tr>
<tr>
<td>SPEC NUMBER</td>
<td>Cardinals</td>
</tr>
</tbody>
</table>

An example f-structure for the grammatical (68a) is shown in Figure 4.17. Ungrammatical combinations of several types of determiners (e.g., articles, demonstratives) as in (68b) are nicely accounted for by assuming a single SPEC DET function to represent all of these types. Additional constraints have to be applied in the grammar to rule out

44 The German grammar in fact uses the special function ADJ-GEN for postnominal possessives for efficiency reasons (Dipper, 2003, p. 182).
combinations of different types of SPEC features, for example, there must be a constraint ruling out occurrences of SPEC DET together with SPEC POSS as in (68c).

(68) a. Karls Hund
   Karl.M.SG.GEN dog.M.SG.NOM
   ‘Karl’s dog’

   b. * dieser der Hund
   this.M.SG.NOM the.M.SG.NOM dog.M.SG.NOM

   c. * der Karls Hund
   the.M.SG.NOM Karl.M.SG.GEN dog.M.SG.NOM

"Karls Hund"

Figure 2.11: ParGram f-structure for German possessive

The prior discussion in Butt et al. (1999b) dates from a period where ParGram included only French, English as well as German. The discussion of the nominal system in general, and the specifier system and its realization via SPEC in particular, is therefore based on these languages. The intuitive appeal and empirical generalizations of the SPEC feature are undeniable given the data in (64) and (68). However, the problems mentioned above remain, even under the more detailed account of Dipper (2003). Across languages, there is ample evidence that nominal arguments need to be distinguished using different types of GFS. As far as I am aware, none of the ParGram grammars have worked out a more detailed account of nominal predication that takes these distinctions into account.

Developed on top of the English ParGram grammar, the PARC BRIDGE system (Bobrow et al., 2007) uses an abstract knowledge representation (AKR) for question-answering. In the BRIDGE pipeline, a semantics module first produces a flattened representation that normalizes f-structures from XLE output. The module maps GFS into semantic roles, drawing on various lexical resources such as WordNet (Fellbaum, 1998) and VerbNet (Kipper-Schuler, 2006). Then, a second set of mapping rules produces AKRs that are descriptions of the concepts under discussion, and a contextual structure that captures the nested structure of contexts. They also specify for each concept whether it is instan-
tiable in the relevant contexts or not; for example, it can infer from the text passage The reporter failed to discover that three young men were killed in the attack on Ryad. that the answer to Did anyone die in the attack on Ryad? must be YES, because the killing event
becomes instantiable at the top context by means of being in the *that* complement of *discover* (Bobrow et al., 2007).

The BRIDGE system is relevant in the present discussion because, in the first step of the pipeline, the BRIDGE system needs to normalize syntactic dependencies across categories, e.g., it needs to transform deverbal nouns and adjectives into their underlying verbal form. This is necessary for the system to be able to assign the correct semantic roles from VerbNet to complements of derived nouns and adjectives (Bobrow et al., 2007, Gurevich et al., 2008). The roles then have to be mapped to the syntactic modifiers of a given noun, which may be in SPEC POSS or in the ADJUNCT set, based on the morphosyntactic realization of the modifier. Such a mapping may be straightforward in cases where a particular morphosyntactic realization is tied to a particular (set of) semantic roles; the mapping will be non-deterministic (i.e., ambiguous) if a language does not have such ties, or if a language does not make use of distinct morphosyntactic configurations of nominal complements at all (such as Hindi/Urdu). For such languages, at least part of the issue could be resolved if the original f-structure parse included a more detailed account of nominal predication: cross-category normalization to verbal forms would still be required, but the semantic role lookup would be simplified if the nominal’s GFS would match the corresponding verbal GFS more closely.

### 2.6.5 A COMPARISON TO HPSG GRAMMAR ENGINEERING

Pollard and Sag (1994), laying the theoretical foundations for HPSG, note that there is a question whether to treat possessives as specifiers or as subjects. Pollard and Sag (1994) provide the examples in (69).

(69) Possessives – Subjects or Specifiers? (Pollard and Sag, 1994, p. 374)

a. Pat’s tasteless portrait of Bush as a mendicant created quite a stir. (NOUN)

b. Pat’s tasteless portrayal of Bush as a mendicant created quite a stir. (DEVERBATIVE NOUN)

c. Pat’s tasteless portraying of Bush as a mendicant created quite a stir. (NOMINAL GERUND)

d. Pat’s tastelessly portraying Bush as a mendicant created quite a stir. (VERBAL GERUND)

e. Pat tastelessly portraying Bush as a mendicant would create quite a stir. (VERBAL GERUND)

f. Pat tastelessly portrayed Bush as a mendicant. (FINITE VERB)

The distinction made by Butt et al. (1999b) in terms of verbal arguments linking to SUBJ/OBJ functions and nominal arguments linking to SPEC/ADJUNCT functions is similar

---

45This particular step could be removed only if a resource becomes available resembling a “VerbNet for nominals”. Here, the information encoded by NomBank (2.2.2) may be too coarse.
to the one put forward by Pollard and Sag (1994), where nominal modifiers including possessives are also allocated a SPEC feature, while the verbal counterparts are subjects or objects. If this distinction is accepted, Pollard and Sag (1994) state that it is nevertheless hard to draw the line in some cases. Intuitively, based on the level of attachment of the argument, Pat’s in (69a–c) may be considered a specifier, while Pat in (69e–f) may be considered a subject; (69d) then is a borderline case. Pollard and Sag (1994) suggest that one could decide on the basis of categorial membership (not unlike the approach suggested by Butt et al. (1999b)): if nominal gerunds count as nouns, and verbal gerunds count as verbs, then Pat’s in (69d) must be treated as a subject. However, if one takes the possessive -s as the deciding morphosyntactic factor, then it must be a specifier, so that verbal gerunds take either subjects or specifiers (69e). Pollard and Sag (1994) suggest that, at least for English and German, there are good reasons for settling on the specifier analysis of nominal modifiers. One deciding factor in German is agreement: possessive modifiers inflect like other specifiers (e.g., articles, demonstratives) and observe the same pattern of agreement with the head noun. In addition, in both languages, the distributional patterns can be accounted by assuming a single specifier function; if one was to assume a subject and a specifier feature for nouns, this explanatory appeal would be lost, and data such as (64)/(68) would have to be explained otherwise. Pollard and Sag (1994) do, however, admit that one can go the alternative route for other languages for various reasons. For example, in Welsh, non-possessive specifiers precede the noun, while possessives follow the noun as well as trigger clitic doubling (if they are pronominal); thus, for Welsh, there are distributional reasons against treating non-possessive and possessive specifiers alike. In Hungarian, possessor-possessed agreement exactly mirrors subject-verb agreement, and one can consider this an argument from agreement to treat possessives as subjects. See also the LFG analyses for Hungarian and Welsh possessors in §3.3.

The English Resource Grammar ERG is a broad-coverage HPSG grammar for English (Baldwin et al., 2004, Copestake and Flickinger, 2000, Flickinger, 2002); its main output formats are HPSG derivation trees, labeled constituent trees, minimal recursion semantics (MRS) as well as elementary dependency structures (EDS, conceived as a reduction of the full MRS structure). The ERG can be tested using an online interface for parsing. In the derivation tree for Carl’s son laughed. (Figure 2.12), Carl’s is treated as a specifier (cf. the label ‘sp-hd’ on Carl’s).

The online ERG interface further produces the EDS in (70) for Carl’s son laughed. EDSs were conceived as a reduction of full minimal recursion semantics (MRS) formulae to a variable-free form, i.e. a semantic dependency graph. Basically, EDSs are dependency graphs where nodes are labeled with semantic predicates (e.g., proper_q or _dog_n_1) and arcs are labeled with semantic argument roles (e.g., ARG1, ARG2, ...). Here, ARG1, ARG2 etc. are semantic role labels that resemble those used in semantic role labeling (see

---

46MRS (Copestake et al., 2005) is the semantic formalism used within computational HPSG grammars.
47The online interface is located at http://erg.delph-in.net/logon.
What is of importance here is that son is treated as a relational noun, as can be seen from its semantic predicate _son_n_of. Note, however, that son itself does not select for arguments (it has an empty argument frame), but the possessive connection between Carl and son is established only by ways of the possessive -s.

(70) HPSG Extended Dependency Structure for Carl’s son laughed:

```
e3:
-1:proper_q 0:4 [BV x6]
x6:named 0:4 ("Carl")
-2:def_explicit_q 4:6 [BV x11]
e14:poss 4:6 [ARG1 x11, ARG2 x6]
x11:_son_n_of 7:10 []
e3:_laugh_v_at 11:19 [ARG1 x11]
```

Figure 2.13 and (71) show the derivation tree and EDS for The son laughed. Note that the determiner the is analyzed as a specifier (just like the possessor in Figure 2.12), which is in line with the theoretical account given by Pollard and Sag (1994).

(71) HPSG Extended Dependency Structure for The son laughed:

```
e3:
-1:_the_q 0:3 [BV x6]
x6:_son_n_of 4:7 []
e3:_laugh_v_at 8:16 [ARG1 x6]
```
2.6. NOMINALS & THE PARGRAM PROJECT

Moving away from English, it has proven difficult to find out to what extent the typological difference hinted at by Pollard and Sag (1994) has been implemented in large-scale HPSG grammars. Grammars for languages other than English are not available online, and thus testing is an issue.

2.6.6 SUMMARY

I have presented the main traits of the ParGram project and the ways ParGram encodes nominal elements in its feature space. The section further lays out the basic ways in which ParGram deals with nominal arguments, and I have also provided a short comparison with how HPSG handles these issues.

It is important to stress the parallelism aspect of ParGram: dissimilar structures or features to encode the same meaning are only allowed if maintaining equal structures would be at the cost of misrepresenting a language. Similar ideas concerning parallelism are evolving in HPSG grammar engineering as well (Bender, 2014, Fokkens, 2011). The analysis of nominal modifiers in general, and nominal arguments in particular, turns out to be a classic example where parallelism simply cannot be maintained: while it may make sense for some languages to encode different modifiers under SPEC, this does not inform us in any way about the distinct syntactic behavior that nominal arguments exhibit across languages. In particular, as will be seen in Chapters 3 and 5, mapping all Hindi/Urdu nominal argument modifiers to some kind of SPEC feature would be at the cost of misrepresenting the behavior of these phrases.
2.7 THE URDU PARGRAM GRAMMAR

The Urdu ParGram grammar makes up the implementation context of this thesis. Therefore, in this section, after some general remarks about the relationship between the Urdu ParGram grammar and the ParGram project, I introduce the general design of the grammar and discuss its main components.

After the ParGram project was established with the English, French and German grammars, Japanese (Masuichi et al., 2003) and Hindi/Urdu were the first Asian languages to be added. While it was found that many of the conventions developed within ParGram in terms of analyses and features could be extended in a straightforward fashion to the Urdu grammar, ParGram has also profited from the inclusion of Hindi/Urdu through the formulation of new standards of analysis and implementations of formal devices that became necessary to adequately model some of the language’s constructions (Butt and King, 2007). Examples for these constructions are Hindi/Urdu correlatives (Butt et al., 2007), the ezafe construction (Bögel and Butt, 2013; Bögel et al., 2008) as well as complex predicates (Butt, 1995; Butt et al., 2003, see also Chapter §7).

Like all the other ParGram grammars, the Urdu ParGram grammar is implemented using xle. It consists of the three modules described below: a transliterator, a morphology, and an LFG grammar including a lexicon. These modules may in principle be used independently, but in xle, they can be integrated as a pipeline at parsing time (see §2.5). Being part of this pipeline implies that the output of one module forms the input of the next module in succession. The input of the first module is running Urdu text, while the output of the last module is a complete LFG analysis in terms of c- and f-structure.

2.7.1 TRANSLITERATION

As noted in §2.3, Urdu and Hindi are very closely related and syntactically very similar. The Urdu ParGram grammar has been designed so as to parse both Urdu and Hindi input. However, a major difference between the two languages concerns their scripts: while Urdu uses a version of Arabic script, Hindi uses Devanagari. A transliterator placed before the actual XLE parser maps the Unicode Arabic Urdu script to an ASCII transliteration scheme that is a common abstraction from both Hindi and Urdu scripts (Bögel, 2012; Malik et al., 2010); a second transliterator mapping from the Unicode Devanagari script of Hindi to the abstraction scheme is planned for the future. The transliteration scheme has been developed with close attention to the idiosyncrasies of the two scripts. For example, in Urdu, short vowels are realized using diacritics on consonants; however, these diacritics are rarely written in common Urdu texts, since the meaning of the words is often retrievable either from the context or from the rest of the letters in the string (Malik, 2006). To solve this issue, the transliterator matches its output which overgenerates possible short vowel combinations against an ASCII word list extracted from the grammar’s morphological analyzer to find the correct form for each word (Bögel, 2012).

The ASCII transliteration scheme is largely based on Glassman (1986) and described
in some detail in Malik et al. (2010). The transliteration module uses the Xerox Finite-State Toolkit (XFST, Beesley and Karttunen, 2003). An example transliteration is given in (72), where the first line shows the original Arabic Urdu script, the second line the ASCII transliteration, the third line the standard transliteration used in the linguistic literature, the fourth line the glosses and the fifth line the translation. The ASCII transliteration is the input to the morphological analyzer.

(72) نینا کو یاسین سے پیار بی‍‍ے

nInA kO yAsIn sE pyAr he

nina=ko yasin=se pyar he

Nina.F.SG = DAT Yassin.M.SG = INST love.F.SG.NOM be.PRES.3.SG

‘Nina is in love with Yassin.’

2.7.2 Morphological Analyzer

Like many other ParGram grammars, the Urdu grammar uses a manually-built finite-state morphological analyzer (FSMA) implemented using XFST (Beesley and Karttunen, 2003) to handle morphological analysis (i.e., reading in a given word form and returning its lemma as well as a set of morphological features). The FSMA can then be connected up to the actual LFG grammar using a flexible morphology-syntax interface implemented in XLE. Bögel et al. (2009) and, in more detail, Bögel et al. (2007) describe some of the issues with respect to developing an FSMA for Hindi/Urdu. An example analysis using the Hindi/Urdu FSMA is shown in (73). Given the input string hansI ‘(she) laughed’, the perfective feminine singular form of the verb hans ‘laugh’, the FSMA produces the lemma as well as a sequence of morphological tags. The string to the right of the arrow sign in (73) thus forms the input to the LFG grammar module.

(73) hansI ↔ hans+Verb+Perf+Fem+Sg

2.7.3 LFG Grammar

The Hindi/Urdu LFG grammar module can be described as composed of three parts: 1) sublexical rules combined with a morphological lexicon, 2) a syntactic lexicon and 3) phrasal rules. The sublexical rules and the morphological lexicon are used to connect the morphological analyzer to the grammar proper, realizing the computational LFG morphology-syntax interface as described by Kaplan et al. (2004). 48
2.7.3.1 Morphological Lexicon & Sublexical Rules

These two components realize the computational LFG morphology-syntax interface as described by Kaplan et al. (2004). A morphological tag lexicon lists all tags that are produced by the FSMA. Most of the lexicon entries are annotated with morphosyntactic or functional information that is projected into the f-structure or constrains other features there. For example, (74) shows the morphological lexicon entry for the tag +Name, which is a tag of the sublexical category NAME. There is morphosyntactic as well as lexical-semantic information attached to the tag: the first line points to the template PROPER-TYPE given in (75a), which introduces the value name for the feature PROPER-TYPE embedded in the feature complex NTYPE NSEM PROPER (that is, when the template is called, its argument _val will be substituted by the argument in the template call in (74), name). The other templates referred to in (74) are given in (75b–c). Different tags may belong to the same sublexical category; for example, the +Masc and +Fem tags both belong to the GEND category, as can be seen in (76).

(74) +Name NAME XLE @(PROPER-TYPE name)
     @(NSYN proper)
     @(NUM sg).

(75) a. PROPER-TYPE(_val) = (\^ NTYPE NSEM PROPER PROPER-TYPE) = _val.
    b. NSYN(_val) = (\^ NTYPE NSYN) = _val.
    c. NUM(_val) = (\^ NUM) = _val.

(76) a. +Masc GEND XLE @(GEND masc).
    b. +Fem GEND XLE @(GEND fem).

The entry for the morphological tag +Noun, on the other hand, is much less complex. The entry is given in (77) and simply registers the fact that the tag is of the sublexical category N-T (which stands for noun tag).

(77) +Noun N-T XLE.

In addition to the morphological tag lexicon, the sublexical rules ensure that the analyses produced by the FSMA can be parsed as the appropriate lexical category nodes. For example, the output string shown in (73), yAsIn+Noun+Name+Masc, can be parsed as the lexical category node N by the rule in (78a). The rule states that the lexical category N consists of a noun stem and three further morphological tags belonging to the sublexical categories N-T (see (77)), NAME (74) as well as GEND (76). The actual sublexical rules used in the Urdu grammar are more complicated (since they have to account for many different types of nouns/verbs/etc., which may come with different morphological tags.

---

49 The @ sign marks template calls; templates may or may not take arguments. Documentation on template calls in XLE is given as part of the XLE documentation at http://ling.uni-konstanz.de/pages/xle/doc/notations.html#N.
attached to them), but the basic design looks as in (78a). Note that there is a template call attached to the sublexical category of the stem (NOUN-S), which assigns third person to a noun using the common template in (78b) (all nouns are in the third person).

(78) a. $N \rightarrow \text{NOUN-S: } @(\text{PERS 3);}$
    $N-T$
    $\text{NAME}$
    $\text{GEND}.$

b. $\text{PERS}(_P) = (^ \text{PERS}) = _P.$

In addition to the rule in (78), the morphological guesser in (79) ensures that any item in the morphology may be analyzed as a noun stem (NOUN-S), provided the tag sequence attached to the item matches the noun sublexical rule. The special variable $\%\text{stem}$ is instantiated to the headword -unknown and, ultimately, to the noun stem provided by the FSMA. The entry in (79) further points to the NOUN template given in (80a), which assigns values to the PRED and PERS features via the templates in (80b–c). Since $\%\text{stem}$ is called as the argument of the NOUN template in (79), $\%\text{stem}$ will replace $_P$ in the application of the NOUN template.

(79) -unknown $\text{NOUN-S XLE @(NOUN } \%\text{stem).}$

(80) a. $\text{NOUN}(_P) = @(\text{PRED } _P)$
    b. $\text{PRED}(_P) = (^ \text{PRED}) = _P.$

Figure 2.14 shows the sublexical c-structure for the Hindi/Urdu proper noun yasin, constructed based on the morphological analysis provided by the FSMA in combination with the sublexical rule in (78); Figure 2.15 shows the resulting f-structure, populated by information retrieved from the FSMA and the morphology-syntax interface.

Figure 2.14: Sublexical c-structure for Hindi/Urdu proper noun

Many lexical items can be adequately analyzed using morphological tags, sublexical rules like the one in (78) and lists of guessers similar to the one in (79); in fact, most of the lexicon in the Urdu grammar resides solely in the morphology (Butt and King, 2007) and can be parsed by the grammar by means of the morphology-syntax interface described

---

50 The special -unknown lexical entry thus matches stems analyzed by the FSMA which are not given explicit entries elsewhere in the lexicon. The -unknown entry is discussed in the XLE documentation at [http://ling.uni-konstanz.de/pages/xle/doc/notations.html#clm4.]
above. The resulting system is modular in that different FSMAs may be plugged into the pipeline; then, the grammar writer can choose to define a sequence of FSMAs which are ordered in preference over one another for a more differentiated lookup (Kaplan et al., 2004).

2.7.3.2 Lexicon

One of the main advantages of using an FSMA in grammar development is that the syntactic lexicon can be kept reasonably small (Butt et al., 1999, Kaplan et al., 2004). The idea here is that there is a partition of work: FSMAs are highly efficient and much more suitable for modeling natural language morphology than full form lexicons (Beesley and Karttunen, 2003). FSMAs, however, have nothing to say about the predicational nature of some lexical items; these items fulfill predicational syntactic functions in that they select GFSs as their arguments. Prime examples for predicators are verbs. In XLE grammars, the predicational information pertaining to these words is stored in a separate lexicon file for the grammar to pick up.

Some of these items are morphologically complex, such as verbs. In this case their morphological analysis is provided by the FSMA in combination with the morphology-syntax interface as described above, and in addition, their predicational function is listed together with their stems in the syntactic lexicon. For example, the FSMA integrated in the Urdu ParGram grammar produces the morphological analysis in (81) for the masculine singular perfective form of the verb hɑns ‘laugh’, which is hɑnsa ‘(he) laughed’. The (simplified) sublexical rule used for parsing tag sequences attached to verbal items is given in (82).

(81) hɑnsA ↔ hɑns+Verb+Perf+Masc+Sg
(82) V --→ V-S
    V-T
    {IMPERF | PERF}
    GEND
    NUM.

The XLE lexicon lookup model as well as the morphology-syntax interface is explained in detail in the XLE documentation at [http://ling.uni-konstanz.de/pages/xle/doc/notations.html#clm](http://ling.uni-konstanz.de/pages/xle/doc/notations.html#clm).
The verbal lexical entry for the verb hans ‘laugh’ is shown in (83); the entry, carrying the sublexical category V-S, gets picked up by the sublexical rule for verbs in (82). Together with the appropriate tag lexicon entries (which are not shown here), the rule parses the string hans+Verb+Perf+Masc+Sg as a V node. At the same time, (83) invokes the template V-SUBJ_unerg (84), which assigns a subcategorization frame to the verb. The frame specifies that the verb selects a subject function as its sole argument. Again, the variable %stem refers to the head word of the entry, hans. Since %stem is called as the argument of the template, it will replace _P in the instantiation of the template; thus, for this particular instantiation: hans = %stem = _P.

(83) hans V-S XLE @(V-SUBJ_unerg %stem).

(84) V-SUBJ_unerg(_P) = (¬ PRED) = 'P<(¬ SUBJ)>'.

Other items are not morphologically complex and thus do not require an analysis in the morphology. The Hindi/Urdu personal pronoun mɛ̃ ‘I’, for example, has the entry in (85). The first line in (85) points to the PRONOUN-PERS template given in (86a). That template provides a value for the PRED feature and calls the NSYN template with the value pronoun as well as the PRON-TYPE template with the value pers. (85) also assigns number and person values.

(85) meN PRON * @(PRONOUN-PERS meN)
    @ (NUM sg)
    @ (PERS 1).

(86) PRONOUN-PERS(_F) = @(PRED _F)
    @ (NSYN pronoun)
    @ (PRON-TYPE pers).

Using this system of templates, xle produces the c- and f-structure in Figure

"meN"

CS 1: PRON
    meN
    [PRED 'meN'
      NTYPE [NSYN pronoun]
      1 NUM sg, PERS 1, PRON-TYPE pers]

Figure 2.16: Sublexical c- and f-structure for Hindi/Urdu personal pronoun

For another example, consider the lexical entry for the Hindi/Urdu ergative marker ne in (87). The first line in (87) points to a template that annotates the volitionality

52The NUM, PERS, PRED, NSYN as well as PRON-TYPE templates are defined as part of the ParGram common.templates.lfg file and simply assign the given values to features identical to the template names; e.g., the template call @(PRON-TYPE pers) assigns the value pers to the feature (¬ PRON-TYPE). Note further that (85) carries a * morphcode, which means that xle will not use an external morphology module to look up the word. Morphcodes were explained in §2.5.2.
2.7. THE URDU PARGRAM GRAMMAR

associated with the ergative as described in §2.3.2, while the template call in the second line projects a CASE feature with the value \texttt{erg}. The third line is an inside-out existential constraint (see §2.5.3.4) stating that nE’s f-structure must itself be part of a SUBJ f-structure. This models the generalization that the ergative only occurs on subjects (Butt and King, 2004b, Mohanan, 1994, p. 121).

\begin{equation}
\text{nE} \quad \text{K} \ast \text{VOLITION} \\
\quad \Theta(\text{CASE} \text{ erg}) \\
\quad \left(\text{SUBJ }^\cdot\right).
\end{equation}

2.7.3.3 Phrasal Rules

The Urdu ParGram grammar currently includes 70 syntactic rules that build up phrasal categories. Here, I confine myself to describe only those rules which will be of central importance for the rest of this thesis. These are the KP, NP as well as the S rules.

First, note that the implementation of case in the grammar follows the analysis put forward by Butt and King (2004b). The basic idea is that the case clitics in Table 2.1 are of the lexical category K; they are functional heads of a KP (kase phrase) that takes an NP as its complement. For Butt and King (2004b), it is this head-complement relationship that accounts for the oblique form of the NP. In addition to providing a case feature, the category K is able to provide additional information associated with case, such as GF assignment and semantically relevant features (e.g., volitionality in the case of the ergative). The relevant part of the KP rule is shown in (88); the second line in (88) adds the K category and uses a CHECK feature to constrain the nominal morphology to be oblique.54 This results in the c- and f-structures in Figure 2.17.55

\begin{equation}
\text{KP} \rightarrow \text{NP} \\
\quad (\text{K: } (^ \text{ CHECK } \_\text{NMORPH}) = \text{c obl}).
\end{equation}

Regarding bare NPs, i.e., those NPs that do not carry any overt case marking and no oblique inflection, Butt and King (2004b) mention that they distribute syntactically like KPs with overt case marking. Since LFG does not make use of empty categories, they argue that the K head is not projected in bare NPs, but that instead, bare NPs always project a KP node; thus, the K node is optional in (88). As mentioned above, the oblique form of the nominal is assumed to be a result of an overt K’s requirements towards its

\begin{footnotesize}
\begin{itemize}
\item[53] The VOLITION and CASE templates are not shown here. VOLITION is a proprietary template within the Urdu grammar, while CASE is part of the ParGram common templates.
\item[54] The CHECK feature is used throughout the ParGram project for features that do not have any repercussions in the semantics, and are thus not needed for further processing down the NLP pipeline (i.e., for semantic reasoning). Features in CHECK typically include features/values necessary to check on various well-formedness conditions, such as morphological form, which is what it is used for here.
\item[55] The f-structure in Figure 2.17 is in fact not a valid f-structure, since the inside-out existential constraint (\texttt{SUBJ }^\cdot\right), a part of the lexical entry of the ergative case clitic in (87), is not satisfied, as the f-structure is not embedded inside a SUBJ.
\end{itemize}
\end{footnotesize}
complement; if the K is not overt, no oblique form occurs on nominals in bare NPs. A parse example for a bare NP node projecting a KP is shown in Figure 2.18.

The (simplified) NP rule prior to the reanalysis and implementation presented in Chapters 3 and 5 is shown in (89a). The first disjunct handles personal pronouns which are full NPs, such as $mɛ\hat{\imath}$ ‘I’; the second disjunct points to the meta-category Nadj, which handles modified nouns. Nadj is shown in (89b). Here, the first line attaches one or more OBL KPs; this is how non-genitive nominal arguments are currently treated (i.e., arguments bearing locative or instrumental case). In the third line of (89b), one or more KPposs category is attached; this is a category especially for genitive-marked KPs, and thus all genitive nominal arguments are attached to the NP here. As will be discussed in §3, Hindi/Urdu allows multiple genitive phrases with a single nominal. Each KPposs node is universally annotated as being an item in a set inside the SPEC POSS function. The SPEC POSS function is a function used by ParGram grammars for genitive nominal arguments; I will return to this in §2.6.4. In the current implementation, all nodes attached under the Nadj meta-category (and thus under NP in the actual c-structure) are concatenated using the shuffle operator (see §2.5.3.6), which has the result that any possible linear ordering of the nodes can be parsed by the grammar. I will return to this rule in Chapters 3 and 5.

(89)  a. $NP \rightarrow \{ PRON
         | Nadj \}$. 

Figure 2.17: Hindi/Urdu c- and f-structure for KP $\text{larke}=\text{ne}$ ‘boy.M.SG.OBL=ERG’

Figure 2.18: Hindi/Urdu c- and f-structure for KP $\text{larka}$ ‘boy.M.SG’
b. Nadj = KP*: (^ OBL) = !

KPposs*: ! $ (^ SPEC POSS)

(\text{^ NUM}) = (! CHECK _NUM)

(\text{^ GEND}) = (! CHECK _GEND)

(\text{^ CHECK _NMORPH}) = (! CHECK _CMORPH)

, AP*: ! $ (^ ADJUNCT)

, N.

The KPposs rule is shown in (90). KPposs rewrites to either an NP with a Kposs category, a special category for the genitive case clitics, or to a possessive pronoun.\footnote{Possessive pronouns are discussed in Chapter §3.3.2.3.} An example entry for kA, the masculine singular nominative (non-oblique) form of the genitive case marker, is shown in (90b). The CHECK features realize agreement: they are equated with the NP’s NUM, GEND as well as CHECK _NMORPH features in (89b).

(90)

\begin{align*}
\text{a. KPposs} & \rightarrow \{ \text{NP: (! CHECK _NMORPH) = c obl; Kposs}
| \text{PRONposs: } \text{^} = !
\text{\@ (CASE gen) } \}.

\text{b. kA Kposs * @ (CASE gen)}
\end{align*}

(\text{91a}) shows the clausal rule currently implemented in the grammar. It rewrites to five different meta-categories, S_unerg, S_unacc, S_trans, S_cop as well as S_cp. Of these, S_unacc and S_cop will be looked at in detail in Chapters 4 and 6, respectively. S_cp will be discussed as a part of the implementation in Chapter §7 on noun-verb complex predicates. At this stage, I show only the S_unerg meta-category in (91b). S_unerg attaches a subject KP, points to a template @ CL-ADJUNCTS (a template not shown here which adds in sentence-level adjuncts) as well as the verbal complex category VCmain. The category VCmain then adds the verb to the tree. Note that the lexical entry of the verb provides a value for the feature (\text{^ LEX-SEM VERB-CLASS}), which is checked by the rule in (91b). Using these rules, the grammar produces the c- and f-structures in Figure 2.19 for the sentence in (92).\footnote{The grammar further includes rules that build up structures above the S node, cf. the ROOT and Sadj categories in the c-structure in Figure 2.19. These are mainly used for handling punctuation and further sentential adverbials, and are not of importance for this dissertation. Note further that the CHECK features are not displayed in this particular f-structure, and that S_unerg is not shown in the c-structure as it is a metacategory (§2.5.3.7).}
2.7. THE URDU PARGRAM GRAMMAR

(91) a. $S \rightarrow \{S_{\text{unerg}}|S_{\text{unacc}}|S_{\text{trans}}|S_{\text{cop}}|S_{\text{cp}}\}$.  
    b. $S_{\text{unerg}} = \text{KP}: (\neg \text{SUBJ}) = !$  
       @CL-ADJUNCTS  
       ,  
       VCmain: ($\neg \text{LEX-SEM} \ \text{VERB-CLASS}) = c \ \text{unerg}.$

(92) lar$\acute{\text{k}}$a hans-a  
boy.M.SG.NOM laugh-PERF.M.SG  
‘The boy laughed.’

<table>
<thead>
<tr>
<th>CS 1: ROOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sadj</td>
</tr>
<tr>
<td>S</td>
</tr>
<tr>
<td>KP VCmain</td>
</tr>
<tr>
<td>NP V</td>
</tr>
<tr>
<td>N hansA</td>
</tr>
<tr>
<td>lar$\acute{\text{k}}$A</td>
</tr>
</tbody>
</table>

Figure 2.19: Hindi/Urdu c- and f-structure for (92)

The decision to use meta-categories for different types of clauses was made for modularity reasons as well as reasons pertaining to grammar maintenance. Consider the rule in (91b). The rule explicitly controls for the number of main clause GFS, viz. only a single SUBJ, and it checks for the type of verb on VCmain. The other meta-categories in (91a) have definitions where the number and type of GFS subcategorized are different, and VERB-CLASS is checked for a different value. An alternative way of implementing a clausal rule would be to attach arbitrarily many KP nodes in a single rule and leave the VERB-CLASS feature underspecified. There is a distinct advantage to the current implementation depicted in (91b): an S rule combining different clausal types quickly becomes unwieldy, since particular functional categories may only turn up in a specific clausal environment. Enforcing such constraints is much easier in the modular approach that was taken here; in addition, the current grammar is easier to debug and maintain.
2.8  Summary

This chapter provided background for the remainder of the dissertation. I have discussed the main subject of the dissertation, nominal predicates and their arguments, in §2.2, and presented relevant literature in theoretical linguistics and NLP. The chapter went on to describe some of the main features of Hindi/Urdu: word order, case, postpositions, agreement as well as information structure. The chapter then gave a general introduction into LFG and its implementation via XLE. The section on XLE paid attention to the usage of special notations which are of importance for this dissertation. The ParGram project was introduced in §2.6; here, I put a focus on the ParGram features used for describing nominals, as well as on the state of the art of dealing with nominal predication and arguments in ParGram, which I have already argued is not capable of dealing with the typological variation in nominal predication and nominal arguments. Additional arguments that point in the same direction will be presented in the chapters to come. I also included a brief comparison to the way HPSG handles nominal predication. Finally, §2.7 has discussed the main ingredients in the Urdu ParGram grammar: the transliterator, the morphology and the LFG grammar proper. The next chapter turns to a discussion of Hindi/Urdu genitive case and its usage in possessive clauses and presents an implementation of possessives couched within the Urdu ParGram grammar.
3

Genitive Nominal Arguments

3.1 INTRODUCTION

This chapter focuses on Hindi/Urdu genitive nominal arguments. The chapter takes a detailed look at the behavior of the genitive in Hindi/Urdu NPs, discussing its constituent as well as functional properties. It turns out that an XLE treatment in terms of a SPEC POSS feature does not do justice to the syntactic behavior of possessives. Instead, their behavior warrants a more differentiated treatment in terms of three different grammatical functions (GF): ADJUNCT, SUBJ as well as OBJ. Possessive nominal arguments occur with a variety of nominals in Hindi/Urdu: common nouns and relational nouns can take a SUBJ possessor argument, while event nouns can take a SUBJ argument as well as an OBJ. All of these may in addition take one or more non-argument (ADJUNCT) genitive phrase(s). Even though possessor selection is relatively free, two selectional restrictions apply. First, possessive arguments of animate common nouns need to be animates themselves. Second, proper nouns as well as pronouns do not select any arguments.

Within the NP constituent, the linear order of nominal modifiers (including possessor arguments) is quite free, with the constraint that the possessor argument may not occur before its head. I show that the shuffle operator in conjunction with a head precedence constraint can be used to encode the generalizations. Possessives do, however, exhibit scrambling possibilities whereby they may be scrambled outside of the NP constituents they are licensed in. Genitive scrambling is not entirely free; the operation is examined in detail to arrive at generalizations that directly inform an implementation in XLE. The implementation makes use of functional uncertainty that annotates possessives to attach to a GF, which may be embedded inside other GFS; the above selectional restrictions
apply. The agreement pattern observed in genitives can be dealt with using inside-out functional constraints.

The chapter is structured as follows. §3.2 introduces the Hindi/Urdu genitive case, including a discussion of the theoretical status of the case marker and general remarks concerning its distribution. Related LFG work on possessives in other languages is surveyed in §3.3. Prompted by the common analysis of possessives in terms of a SUBJ argument, or a function related to SUBJ, §3.4 reviews subjecthood diagnostics available for Hindi/Urdu. §3.5 then proceeds to review the use of the genitive in NPs and discusses its constituent as well as its functional properties. §3.6 takes a detailed look at the operation of genitive scrambling and several restrictions that apply to it. §3.7 discusses the use of the genitive in possessive clauses. §3.8 provides a novel explanation for the binding patterns observed within Hindi/Urdu NPs, §3.9 discusses issues with respect to ambiguity, and §3.10 concludes the chapter. Note that the actual analysis as well as implementation is presented in Chapter §4.

3.2 THE HINDI/URDU GENITIVE CASE

3.2.1 GENERAL DESCRIPTION

The genitive case in Hindi/Urdu is realized using the clitic k-, which is attached to a possessor NP.1 This clitic is of the same type as the other case clitics discussed in §2.3.2, with one exception: it agrees in number, gender and morphological form (nominative or oblique) with a head noun, the possessum. For the feminine, there is morphological syncretism in that a single form ki is used throughout the feminine inflectional pattern. For the masculine, there is syncretism between the singular oblique and plural nominative and oblique. Table 3.1 shows the complete pattern of the clitic.

Generally, within NPs, the modifying possessor phrase comes first, then the possessum, thus conforming to the general Hindi/Urdu head-final pattern as discussed in §2.3.1. In (1-3) the a. examples are the only grammatical examples, displaying the correct agreement pattern.

1The terms possessor and possessum are highly problematic. Hindi/Urdu, just like other languages, encodes many more semantic relations than just possession by using the genitive case. Thus, whenever I use the terms possessor and possessum throughout this chapter, I use them simply as shortcuts in lieu of more complicated phrases such as genitive-marked modifier phrase or noun modified by a genitive-marked modifier phrase. See also the discussion in §3.5.2.1.

2See §3.5 for a more detailed discussion of the genitive in NPs.
3.2. THE HINDI/URDU GENITIVE CASE

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number</th>
<th>Inflection</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine</td>
<td>Singular</td>
<td>Nominative</td>
<td>ka</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oblique</td>
<td>ke</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>Nominative</td>
<td>ke</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oblique</td>
<td>ke</td>
</tr>
<tr>
<td>Feminine</td>
<td>Singular</td>
<td>Nominative</td>
<td>ki</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oblique</td>
<td>ki</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>Nominative</td>
<td>ki</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oblique</td>
<td>ki</td>
</tr>
</tbody>
</table>

Table 3.1: Possible inflections of Hindi/Urdu genitive case clitic k-

The fact that the genitive inflects is not surprising once its historical origin is taken into account. Both Beames (1872–1879) and Kellogg (1893) analyze the genitive as having arisen from kṛita ‘done by’, the Sanskrit past participle of kri ‘do’, as in (4).

(4) Sanskrit kṛita > Prakrit kerita > keriai > modern Urdu/Hindi k-

As the original participle kṛita ‘done by’ inflected for agreement, the currently used case marker does so, too. See also the accounts given by Butt and King (2004b) as well as Bögel and Butt (2013).

3.2.2 IS THE HINDI/URDU GENITIVE A CASE?

The fact that the Hindi/Urdu genitive agrees with its head noun has spawned a discussion about its status as a case marker. Generally, in the view espoused by those that challenge
its status as a case, the Hindi/Urdu genitive is better treated as deriving attributive adjective phrases (e.g., Bloch, 1965). In this subsection, the discussion is briefly reviewed and evidence is given to support the claim that the Hindi/Urdu genitive must be regarded as a case clitic.

3.2.2.1 Payne (1995)

Payne (1995) discusses the two alternatives, case marker vs. deriving attributive adjective phrases, and mentions the most obvious argument in favor of the adjectival hypothesis: the agreement pattern of the genitive marker parallels that of adjectives in Hindi/Urdu. However, Payne (1995) further mentions several distributional arguments against the adjectival hypothesis. Two of these arguments are cited here. First, the basic order of elements in Hindi NPs is as in (5); an example is given in (6). The general order identified by Payne (1995) is in accordance with the standard NP word order as given in Hindi and Urdu grammars such as Kachru (2006) and Schmidt (1999).

(5) Possessor — Determiner — (Numeral) Adjective — Noun

(6) \( \text{ʃɑrmɪla} = \text{ki} \ vʊh \ do \ lal \ kɪtabẽ \)
\( \text{Sharmila.F.SG = GEN.F.PL. those two red book.F.PL} \)

‘those two red books of Sharmila’s’

This goes to show that there are two separate positions in the NP template, one for possessors (i.e., genitive-marked modifiers), the other for attributive adjectives; in the canonical NP-internal word order, determiners occur in between possessors and adjectives. This does not mean, however, that variations in the word order are not possible; §3.5.1.2 reviews possible alternations in the linear order of NP-internal elements. The account by Payne (1995) further does not discuss the possibility of multiple genitives, which I examine in this chapter.

Second, possessors in Hindi/Urdu may be governed by a distinct set of postpositions, such as nicʰe ‘under’, picʰe ‘behind’, etc. An overview of these complex postpositions, taken from Butt and King (2004b), was given in (15) in §2.3.3 and is repeated below in (7). In all the examples in (7), ke is the oblique form of the masculine singular genitive case marker ka. (8) illustrates the use of one of the postpositions.

\(^3\)On a side note, it is not at all clear that Hindi/Urdu has a part of speech that correlates with the “determiner” part of speech of other languages (Kachru, 2006). Those functional items that correspond to determiners in the classical sense are homophonous with demonstratives and indefinite pronouns. Thus, the term “determiner” is used here in a general way and subsumes demonstratives, indefinite pronouns, and other types of quantifiers.
3.2. THE HINDI/URDU GENITIVE CASE

(7)  
kepicʰe  ‘behind’  
ke pele  ‘before’  
ke nicʰe  ‘under’  
ke pas  ‘next to/near’  
ke upar  ‘over’  
ke satʰ  ‘with’  
ke andar  ‘next to/near’  
ke liye  ‘for’  
ke samne  ‘in front of’  
ke taraf  ‘in the direction of’  
ke age  ‘in front of (further along)’  
ke bad  ‘after’

(8)  
kuṭṭa  dog.
mez=ke  table.
m. sg. nom
picʰe  behind
hɛ  be.
pres. 3. sg
‘The dog is behind the table.’

Again, as noted by Payne (1995), this sets apart phrases headed by the genitive case marker from adjectives. Postpositions such as nicʰe ‘under’ or picʰe ‘behind’ never appear on adjectives; see also the discussion on postpositions in §2.3.3. Thus, the form as well as the distribution of possessors are different from adjectives. For these reasons, amongst others, Payne (1995) arrives at the conclusion that k- is a genuine genitive case marker.

3.2.2.2 Spencer (2008)

In addition to the two alternatives presented so far (genuine case marker vs. formative deriving adjective phrases), Spencer (2008) argues for a third alternative. In his view, Hindi/Urdu lacks a genitive case. k- is not a case marker; it is also not an adjectival derivational formative either, for the reasons enumerated above. However, Spencer notes that the final conclusion drawn by Payne (1995) (that k- must be a case, since it cannot be an adjective-deriving formative) is not valid, in that the marker can still be in construction with an NP without being a case. He calls the construction the “possessum-agreement” construction: the k- marker is attached to NPs and gives that NP the external agreement morphosyntax of an adjective; in their internal syntax, however, the k- marked phrases retain nearly all the properties of NPs.

The account of Spencer (2008) is problematic. Spencer treats possessors as having split properties: external agreement morphosyntax of an adjective, internal makeup of an NP. Clauses such as (9a), where the order of possessor and possessum may be inverted as in (9b), are analyzed by Spencer assuming ellipsis of the head noun in (9b). That is, (9b) is derived by ellipsis from (9c) (Spencer, 2008, p. 16).

(9)  
a. ʊs  stri=ka  yeh makan  hɛ  
pron.3.sg.obl  woman.f.sg  =  gen.m.sg  this  house.m.sg.nom  be.pres.3.sg
‘Of that woman is this house.’ = ‘That woman owns this house.’

b. yeh makan  us  stri=ka  hɛ  
this  house.m.sg.nom  pron.3.sg.obl  woman.f.sg  =  gen.m.sg  be.pres.3.sg
‘This house is of that woman.’ = ‘This house is that woman’s.’

4Cross-linguistically, pre- and postpositions often govern a specific case, so it is not surprising from a typological standpoint that Hindi/Urdu postpositions require the genitive case.

5I discuss the genitive in possessive clauses in §3.7.
3.2. THE HINDI/URDU GENITIVE CASE

c. yeh mākan ʊs stri=ka
   this house.M.SG.NOM PRON.3.SG.OBL woman.F.SG = GEN.M.SG
mākan  hɛ
house.M.SG.NOM be.PRES.3.SG
   ‘This house is the house of that woman.’ = ‘This house is that woman’s house.’

An account involving ellipsis, however, gives rise to two requirements. The first requirement demands that the account lay out a theory of what parts of the NP are elided, and what parts are not. Consider the example in (10). Is the adjectival adjunct nāyi ‘new’ elided together with the head noun — i.e., is (10a) derived from (10b)? This cannot be the case, since (10a) does not mean ‘This new book is Ram’s new book,’ but rather ‘This new book is Ram’s.’ If the adjunct is not elided, one needs to legislate against it, providing an explanation of why it is not.

(10) a. yeh nāyi kītab ram=ki  hɛ
   ‘This new book is of Ram.’ = ‘This new book is Ram’s.’

b. yeh nāyi kītab ram=ki nāyi
kītab  hɛ
book.F.SG.NOM be.PRES.3.SG
   ‘This new book is the new book of Ram.’ = ‘This new book is Ram’s new book.’

A second requirement is related to the semantics of questions. In Hindi/Urdu, the wh-word in question formation usually remains in situ. One may formulate the question in (11a) asking for the predicational NP ram=ki kītab ‘Ram’s book’ in the response in (11b). If the head noun kītab ‘book’ were elided in (11c), one would expect (11c) to also work in response to (11a), which is not the case: (11c) does not answer the question in (11a) in an adequate manner. This presents evidence against an ellipsis analysis; under an analysis involving ellipsis, it is not clear why (11c) should be unacceptable.

(11) a. yeh kītab  kya  hɛ
   this book.F.SG.NOM what be.PRES.3.SG
   ‘What is this book?’

b. yeh kītab ram=ki kītab  hɛ

c. ?? yeh kītab ram=ki  hɛ
   this book.F.SG.NOM Ram.M.SG = GEN.F.SG be.PRES.3.SG
   ‘This book is of Ram.’ = ‘This book is Ram’s.’

\[6\] I thank Rajesh Bhatt for pointing out these two requirements to me.
3.2. THE HINDI/URDU GENITIVE CASE

Spencer (2008) does not present any further evidence in fulfillment of the two mentioned requirements. Ruling out an ellipsis analysis in turn counters Spencer’s approach to the genitive in that it shows that the k-marked phrases do not retain the properties of NPs. I therefore conclude that the Hindi/Urdu genitive is best analyzed as a case clitic along the lines of the other Hindi/Urdu case clitics (as in §2.3.2), and possessors are kps headed by that particular clitic, which is of category k, following Butt and King (2004b).

3.2.3 POSSESSIVE PRONOUNS

Hindi/Urdu also makes use of possessive pronouns. An overview of some of the forms (mera ‘my’ and hamara ‘our’) is shown in Table 3.2. “Antecedent Features” refers to the features of the antecedent; “Modified Noun Features” refers to those features of the modified noun which the pronoun agrees with. Other possessive pronouns include the second person tera ‘your (sg.)’ and tumhara ‘your (pl.)’. The third person pronouns encode a proximal/distal distinction (Kachru, 2006, p. 62); proximal forms are iska/is=ka ‘her/his/its’ and inka/in=ka ‘their’, distal forms are uska/us=ka ‘her/his/its’ and unka/un=ka ‘their’.

<table>
<thead>
<tr>
<th>Antecedent Features</th>
<th>Modified Noun Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Number</td>
</tr>
</tbody>
</table>

| | 1 | Singular | Masculine | Singular | Nominative | mera ‘my’ |
| | | | | | Oblique | mere ‘my’ |
| | | | | | Plural | mere ‘my’ |
| Feminine | Singular | Nominative | mera ‘my’ |
| | | Oblique | mere ‘my’ |
| | | Plural | mere ‘my’ |
| Masculine | Singular | Nominative | hamara ‘our’ |
| | | Oblique | hamare ‘our’ |
| | | Plural | hamare ‘our’ |
| Feminine | Singular | Nominative | hamari ‘our’ |
| | | Oblique | hamari ‘our’ |
| | | Plural | hamari ‘our’ |

Table 3.2: Possessive pronouns in Hindi/Urdu

7Hindi and Urdu differ in that Hindi uses a single word to render the third person possessive pronouns (e.g., uska ‘her/his/its’ etc.), while Urdu uses two separate tokens (us=ka ‘her/his/its’ etc.).

8For a more complete overview of possessive pronouns in the language, see e.g., Kachru (2006), p. 63.
3.2. THE HINDI/URDU GENITIVE CASE

First and second person possessive pronouns always have animate antecedents, while this may not always be the case with third person possessive pronouns, just as the third person pronoun voh ‘she/he/it’ (from which uska/us = ka is derived) does not necessarily refer to an animate entity (Schmidt, 1999, p. 19). Possessive pronouns in Hindi/Urdu distribute exactly like their kp counterparts; therefore, in my analysis, I treat possessive pronouns as syntactically equivalent to full genitive kps.

3.2.4 The Distribution of the Genitive

Regarding the distribution of the genitive case in Hindi/Urdu, Butt and King (2004b) state the following:

(12) The genitive may be roughly characterized as marking subjects of nonfinite clauses [...], subjects of finite copula clauses [...], and specifiers of nominals [...].

(Butt and King, 2004b, p. 11)

Mohanan (1994) further notes that two classes of N-V CPs occur with genitive subjects. One class is characterized by Mohanan (1994) as involving irreversible events, or events that result in a relatively permanent state. CPs in this class include janm ho ‘be born’ and vivah ho ‘get married’. An example using the CP janm ho ‘be born’ is given in (13).

(13) sri krɪʃn=ka matʰʊra=mẽ janm hu-a
    Sri Krishna.M.SG = GEN.M.SG Mathura = LOC.IN birth.M.SG be.PERF-M.SG
    ‘Sri Krishna was born in Mathura.’ (Mohanan, 1994, p. 180)

The second class includes N-V CPs such as vicar ho ‘think’, dava ho ‘claim’ and iradaho ‘intend’. Mohanan (1994) characterizes this class as describing deliberate mental activities. An example for this class is given in (14).

(14) ram=ka [gʰɜɾ lauᵗ-ne]NP = ka
    vicar tʰ-a
    thought.M.SG be.PAST-M.SG
    ‘Ram was thinking of returning home.’ (Mohanan, 1994, p. 180)

Note that Mohanan (1994) distinguishes these two sets of N-V CPs on the basis of their semantics, not in terms of their syntax. Regarding the syntax, it can be observed that they both involve a genitive subject as well as an N-V CP. Thus, both of the two classes obey to a single syntactic pattern. I therefore define the following working definition of the distribution of the genitive in Hindi/Urdu.

I survey the use of the genitive in NPs in §3.5. I will not attempt an analysis of the genitive subject N-V CP patterns identified by Mohanan (1994); here, I will have nothing to add to Mohanan’s take and acknowledge that the genitives in N-V CPs do behave as matrix subj.s. See Chapter §7 for a description of N-V CPs. I also exclude the genitive’s use on subjects of non-finite and finite copula clauses from my survey; I am convinced,
however, that the treatment of the genitive in NPs worked out in this chapter can easily be extended to include the genitive’s use in such clauses.

Before I turn to my discussion of the Hindi/Urdu genitive, the following two sections take excursions that will lay the ground for my analysis. §3.3 presents previous LFG accounts of genitive constructions in different languages — English, Hungarian, as well as Welsh. It turns out that some of these previous accounts draw close similarities between the subj function and the GF attributed to the possessor. Thus, §3.4 reviews tests for subjecthood that have been suggested for Hindi/Urdu. These tests will lend support to the subsequent discussion of the Hindi/Urdu data in §3.5.

3.3 RELATED WORK IN LFG THEORY

3.3.1 ENGLISH

The account of English possessors given by Bresnan (2001) identifies possessors with a unique GF: the POSS function. Bresnan (2001), p. 293, assumes that the lexical forms of nouns may be augmented by a lexical predication template to select the POSS function as an argument. An example of such an augmentation is given in (15b), when applied to (15a). The resulting f-structure for the NP Mary’s horse will look like in Figure 3.1 (slightly adapted from Bresnan (2001)).

(15) a. horse N (↑ PRED) = ‘horse’
   b. ‘horse’ ⇒ ‘horse-of<(↑ POSS)>’

Figure 3.1: POSS function for possessors

The POSS function is restricted to the nominal domain; Bresnan (2001) assumes that it is used for possessors as in Figure 3.1 as well as genitive subjects of nominal gerunds as in Roseanne’s singing of the song. The application of the POSS augmentation rule to a nominal gerund is shown in (16). The restriction of POSS to the nominal domain can be modeled using the inverse mapping of the $\phi$ function with reference to the label of the nominal category (NP/DP).

(16) a. singing N (↑ PRED) = ‘singing<(↑ OBL)}’ (Bresnan, 2001, p. 294)
   b. ‘singing<(↑ OBL)}’ ⇒ ‘singing-of<(↑ POSS) (↑ OBL)}’ (Bresnan, 2001, p. 294)

\[\lambda\] is the labelling function associating category labels with c-structure nodes. $\phi$ is the mapping function projecting f-structure from c-structure. Thus, (17) restricts POSS to appear only in f-structures associated with c-structures that belong to the nominal domain.
3.3. RELATED WORK IN LFG THEORY

Verbal gerunds, on the other hand, do not distribute like nominal gerunds and receive the category V under Bresnan’s account; since they nevertheless exclusively occur in the nominal domain, the constraint in (17) needs to be extended as in (18), which states that the \( \text{PRED} \) attribute of any gerundive verb must be projected off a category inside the nominal domain.

\[
\text{(17)} \quad n \in \lambda (\varphi^{-1}(\text{POSS} \uparrow)) \quad \text{(Bresnan, 2001, p. 294)}
\]

The verbal gerund \textit{visiting} as in \textit{Mary’s frequently visiting Fred}, being derived from the verb \textit{visit}, has the lexical entry in (19). Its subcategorization frame does not include a \text{POSS} function, which means that the resulting f-structure will be neither coherent (the \text{POSS} function contributed by the possessive \textit{dp Mary’s} is not selected by any predicate) nor complete (the \text{SUBJ} function designated by the gerundive is not provided). \text{(Bresnan, 2001)} proposes the solution in (20), where a lexical rule identifies the possessor of gerundive verbs with their subject; this results in \text{POSS} acting as a functional controller of the \text{SUBJ}, and the f-structure will be coherent and complete.

\[
\text{(18)} \quad \text{V (GERUNDIVE)} \Rightarrow n \in \lambda (\varphi^{-1}(\text{PRED} \uparrow)) \quad \text{(Bresnan, 2001, p. 294)}
\]

\[
\text{(19)} \quad \text{visiting} \quad \text{V (GERUNDIVE)} \quad (\uparrow \text{PRED}) = \text{‘visiting} <(\uparrow \text{SUBJ}) (\uparrow \text{OBJ})>’
\]

\[
\text{(20)} \quad \text{V (GERUNDIVE)} \Rightarrow (\uparrow \text{POSS}) = (\uparrow \text{SUBJ}) \quad \text{(Bresnan, 2001, p. 294)}
\]

The \text{POSS} function came to be widely used in the LFG literature for modeling possessors. Among the researchers that make use of \text{POSS} are \text{Sadler (2000)} (who takes \text{POSS} to be a subclass of the \text{SUBJ} \text{GF}), \text{Laczkó (2000)}, and \text{Falk (2001)}. Meanwhile, the origin of the \text{POSS} label as such is not made explicit anywhere, and may be connected to the intuition that possession (or perhaps a more abstract concept of possession, as discussed by \text{Barker and Dowty (1993)}) is the prototypical relation expressed by possessives, even though the \text{POSS} function in general is taken to be a function not semantically restricted to possession.

Taking a closer look at the lexical rule in (20), an alternative suggests itself, namely to regard the possessor as projecting a \text{SUBJ} \text{GF} itself inside the nominal domain. This is the solution advocated by \text{Dalrymple (2001)} as well as \text{Chisarik and Payne (2003)}. The augmentation template in (15) would then be restated as in (21), rendering the f-structure in Figure 3.2 for the \textit{np Mary’s horse}.

\[
\text{(21)} \quad \text{a. horse} \quad \text{N (\uparrow \text{PRED})} = \text{‘horse’}
\]

\[
\text{b. ‘horse’} \Rightarrow \text{‘horse-of} <(\uparrow \text{SUBJ})>’
\]

Aside of circumventing the problem regarding the English gerundives, effectively getting rid of the need to specify the lexical rule in (20), this approach has a couple of other advantages, as suggested by \text{Chisarik and Payne (2003)}. First, the \text{SUBJ} function is intrinsically taken to be semantically unrestricted; this would readily account for the
3.3. RELATED WORK IN LFG THEORY

wide variety of thematic roles that may be realized as “possessives” across languages. Second, the SUBJ function in LFG is generally regarded as a discourse-related function: in English clauses, the subject typically functions as a topic, just like the possessor does in English NPs. Third, anaphoric relations could also be explained naturally if possessors were assumed to be SUBJS. The behavior of the reflexive pronoun himself in the English sentence Jo criticized himself, for example, is assumed to be identical to the one in Jo’s criticism of himself. In both cases, the antecedent of the reflexive is a SUBJ in its binding domain (Chisarik and Payne, 2003, p. 184). They argue that ADNOM needs to be an unrestricted function, due to the wide semantic array that is available to this position; they present the sample data in (22) in support of their argument:

(22) a. the green eyes of the girl sitting opposite me BODY PART
    b. the sister of the man who had been arrested KINSHIP
    c. the debut of the young flautist from Abergavenny AGENT
    d. the new house of the Vice-Chancellor elect OWNER

The ADNOM function, contrary to SUBJ, is shown to be non-discourse-related. It can never serve as the antecedent of reflexive pronouns, and pronouns (which, in the unmarked case, function as topics in English) tend to occur in the possessor SUBJ function, and not in the ADNOM function; see the example in (23), taken from (Chisarik and Payne, 2003, p. 187). Thus, for the example in (24), Chisarik and Payne (2003) propose the f-structure in Figure 3.3.

Figure 3.2: SUBJ function for English possessors

\[
\begin{array}{c}
\text{PRED} \; \text{‘horse-of-<↑SUBJ>’} \\
\text{SUBJ} \; g: [\text{PRED} \; \text{‘Mary’}] \\
\end{array}
\]

Thus, the minimal complete nucleus in Jo criticized himself is the whole clause, while the minimal complete nucleus in Jo’s criticism of himself is the whole NP.

10The binding domain of the English reflexive himself is generally taken to be as in (i).

(i) The antecedent of the English reflexive pronoun himself must appear in the Minimal Complete Nucleus containing the pronoun. (Dalrymple, 2001, p. 280)

The Minimal Complete Nucleus itself is defined with respect to the presence of a SUBJ function as in (ii):

(ii) Minimal Complete Nucleus containing an f-structure f:
The smallest f-structure that contains f and a SUBJ function. (Dalrymple, 2001, p. 281)

11The terminology may be confusing. Bresnan (2001) only uses the term “possessors” in her discussion of gerundives and means the English prenominal s-genitive. The discussion by Chisarik and Payne (2003) revolves around the functional difference between what they refer to as the “genitive” and the “of-oblique”. I adapt the latter terminology for the discussion of the English data.
3.3. RELATED WORK IN LFG THEORY

(23)  
   a. her car  
   b. ??the car of her

(24)  
   the daughter of the king

\[
\begin{array}{c|c|c}
\text{ADNOM} & \text{PRED} & \text{king} \\
\hline
\text{PRED} & \text{daughter}<((\uparrow \text{ADNOM})>) \\
\end{array}
\]

Figure 3.3: ADNOM function for English possessors

As will be seen in §3.5, the Hindi/Urdu data differs in several respects from the English data. First, there is no constraint on the number of possessors in Hindi/Urdu. While this is only true of a certain type of Hindi/Urdu possessors, which I will refer to as attributive genitives, it nevertheless is a syntactic observation that needs to be dealt with. In English, on the other hand, there is a maximum of one pre-nominal and post-nominal genitive each. This motivates the use of grammatical functions instead of an adjunct. For Hindi/Urdu, however, the coherence requirement of LFG (Kaplan and Bresnan, 1982) would seem to rule out using governable GFS, at least for attributive genitives. Unless we can posit a (possibly very) large number of governable GFS and can motivate somehow that nominal predicates subcategorize for all of these, an adjunct treatment is clearly preferable.12

Second, Hindi/Urdu makes use of only a single genitive construction, a genitive KP headed by the clitic k- (see §3.2). English, on the other hand, uses two distinct constructions, the prenominal genitive as well as the postnominal of-oblique.

3.3.2 HUNGARIAN

Chisarik and Payne (2003) also discuss possessor constructions in Hungarian. In this language, there are two separate possessor constructions. Chisarik and Payne (2003) refer to the pattern in (25a) as the dative possessor construction, and to the pattern in (25b) as the genitive possessor construction.13

(25)  
   a. [a király-nak] a lány-a  
       ART king-DAT ART daughter-SUBJ.3  
       ‘the king’s daughter’  
       (Chisarik and Payne, 2003)
   b. [a király] lány-a  
       ART king daughter-ADNOM.3  
       ‘the king’s daughter’  
       (Chisarik and Payne, 2003)

12A brief note on the argument-adjunct distinction can also be found in §3.5.4 after the data discussion.
13Other researchers have referred to the pattern in (25b) as the nominative possessor construction (e.g., Laczkó, 2000; Szabócskí, 1994). The details of this terminological difference are purely morphological and not of importance for the present discussion.
Crucially, they argue that the two constructions realize distinct GFs at f-structure, in a way quite similar to how they analyze the English data. The Hungarian dative possessor construction is shown to occupy the SUBJ function, since it is discourse-related as well as semantically unrestricted. Thus, for (25a) they propose the f-structure in Figure 3.4. The genitive possessor construction shown in (25b), on the other hand, is treated like the postnominal of-oblique in English in that it resides in an ADNOM function (Figure 3.5).

![Figure 3.4: SUBJ function for Hungarian possessors](Chisarik and Payne, 2003)

![Figure 3.5: ADNOM function for Hungarian possessors](Chisarik and Payne, 2003)

What complicates the Hungarian account is the observation that any given nominal may only be modified by at most one of the above constructions. That is, the dative and the genitive possessor constructions exclude each other in a single NP. The LFG analysis by Laczkó (1995, 2000) captures this by assuming that both constructions are functionally equivalent and subsumed under the POSS GF, an assumption which nicely accounts for the mutual blocking of structures. Given that they propose two different GFs, Chisarik and Payne (2003) need to come up with a different approach. They settle on the generalization in (26), which states that the realization of two [−r] arguments is blocked in Hungarian NPs; this is referred to as the “Asymmetrical Possessor Parameter” (APP), and is an extension into the nominal domain of the “Asymmetrical Object Parameter”, originally proposed by Bresnan and Moshi (1990) for objects of clauses.

(26) Asymmetrical Possessor Parameter (Chisarik and Payne, 2003, p. 197)

\* \[ \theta \mid \theta \]
\[ \mid \mid \]
\[ [−r] \mid [−r] \]

The APP is argued to be active in Hungarian, but not in English, which does allow two unrestricted arguments (§3.3.1). This explains the distributional differences in Hungarian vs. English possessives.

There are several differences between the Hungarian data and the Hindi/Urdu data, as will be seen shortly. First, as stated above, Hindi/Urdu imposes no restriction on the number of genitive modifiers possible, contrary to what is seen in Hungarian. Second, Hungarian is similar to English in that it makes use of two distinct possessives; thus,
the difference in GF assignment is reflected by a difference in the morphosyntax of both languages. Hindi/Urdu only has a single genitive construction. What remains to be seen below is in how far the Hindi/Urdu construction displays variation in terms of its functional distribution.

3.3.3 **Welsh**

Sadler (2000) discusses possession in Welsh NPs. After discussing and rejecting the movement account by Rouveret (1994) (Rouveret, 1994 takes a late pre-minimalist Principles and Parameters approach (Chomsky and Lasnik, 1993) to generative grammar) Sadler proposes an LFG account. She argues that the possessor is structurally the specifier of NP, and that semantically, the possessor can fill a wide variety of thematic roles (e.g., themes and agents); further, virtually every head noun may take a possessor, which may correspond to either a semantic argument or a semantic non-argument (i.e., an adjunct). In Welsh, only a single possessor is allowed per head noun, which leads Sadler to propose a GF which she calls “subjective” and discourse-oriented, choosing the label POSS. The assumed discourse-orientedness of POSS results from its ability to bind other arguments in the NP domain. Sadler (2000) assumes that all nominal predicates may optionally subcategorize for the POSS GF which is structurally associated with the NP specifier position.

The lexical entries in (27) and (28) proposed by Sadler (2000) differ, however, from each other in a crucial point. In the entry in (27), the POSS function occurs outside the angle brackets, while in (28), it occurs inside the angle brackets. This reflects a difference in thematic vs. non-thematic selection of the POSS argument; another, maybe more common application of the notation in (27) are expletive arguments such as the English expletive it in It is raining. (Butt et al., 1999b, Dalrymple, 2001). What the notation means with respect to possessors is that the head noun llyfr ‘book’ does not intrinsically subcategorize for a poss, but is augmented to include it (probably via a lexical rule in the fashion of (15), but this is not further specified by Sadler (2000)). Compared to, e.g., (21b), Sadler’s approach in (27) is even more restrictive regarding the selectional status of the possessive. Sadler’s notation in terms of a non-thematic designator (i.e., the bracket-external notation) follows Bresnan (1982a): thematic functions are mapped to arguments of the semantic relation and thus must satisfy its selectional restrictions, while non-thematic functions are purely syntactic and may be filled, for example, by semantically empty expletives, as in It rains. Thus, the notation by Sadler (2000) entails that possessives are treated on a par with expletives in terms of being purely syntactic, hence not enforcing any selectional restrictions.

(27) ‘llyfr<> (↑ POSS)’ ‘book’
(Sadler, 2000, p. 97)

(28) ‘disgrifiad<<(↑ POSS)>>’ ‘description’
(Sadler, 2000, p. 97)
Note that, for Sadler (2000), the difference in the notation between (27) and (28) also alludes to the difference between intrinsic and extrinsic possession worked out by Barker (1995). Under his account, verb-derived event nominals such as description or kinship terms such as daughter intrinsically (i.e., lexically) select for a possessor, while common nouns like book or firetruck do not intrinsically do so, but nevertheless allow modification by extrinsic possessors. Both of these selection processes do not per se, as acknowledged by Barker (1995), put any constraints on the thematic role encoded by the possessor; however, in the case of intrinsic possession, the specific thematic relation depends on inherent qualities of the head noun.

Welsh poses an additional difficulty in that the definiteness of the whole NP is determined by the most deeply embedded possessor. For the example in (29), this means that the whole NP is definite, since the most deeply embedded possessor banc ‘bank’ is marked by the definite article y ‘the’. Sadler (2000) proposes the annotations in (30) on the possessives in specifier position; the annotations take care of the definiteness requirement and result in the f-structure in Figure 3.6.

(29) cath merch rheolwr y banc
cat daughter manager the bank
‘the bank manager’s daughter’s cat’
(Sadler, 2000, p. 100)

(30) NP → N’ NP
   ↑ = ↓ (↑ POSS) = ↓
   (↑ DEF) = (↓ DEF)

\[
\begin{align*}
PRED & \quad \text{‘cat<>(↑ POSS)’} \\
DEF & \quad + \\
PRED & \quad \text{‘daughter<((↑ POSS)>’} \\
DEF & \quad + \\
PRED & \quad \text{‘manager<>((↑ POSS)’} \\
DEF & \quad + \\
PRED & \quad \text{‘bank’} \\
DEF & \quad + \\
\end{align*}
\]

Figure 3.6: POSS function for Welsh possessors

While the treatment of Welsh NPs seems coherent in and of itself, it is not entirely clear what Sadler’s motivation is for assuming the POSS function, instead of the SUBJ function (which is also an unrestricted, discourse-oriented GF). Sadler herself calls POSS a subjective function, alluding to the apparent similarities between the two functions,

\[\text{14}\]

I will invoke the account given by Barker (1995) in my description of the functional realm of Hindi/Urdu possessors in §3.5.2.
and from her account alone it is not clear on what basis she rules out calling possessors subj. It is also questionable in how far Sadler’s account can be carried over to the Hindi/Urdu data. As mentioned above, Welsh only allows a single possessor per head noun; this is not the case in Hindi/Urdu (at least not for all types of possessors). Moreover, the treatment of augmented possessives by Sadler in terms of a non-thematic function is not without issues. By using a non-thematic function, Sadler (2000) effectively eliminates the possibility that there is any kind of selectional restriction enforced between a common noun head and its possessor, a prediction along the lines of “anything goes” that may turn out to be too strong across languages (Barker, 1995).

3.4 Testing for Subjecthood in Hindi/Urdu

There is considerable evidence from other languages, as was seen in the preceding section, that the GF realized by possessives inside the f-structure of the nominal head is closely related (Bresnan, 2001, Sadler, 2000), if not identical (Chisarik and Payne, 2003), (Dalrymple, 2001) to the subj function that is otherwise known only from verbal predication. Thus, I take the subj function to be a contender for the function filled by the possessor KP in Hindi/Urdu. Fortunately, there are several tests for subjecthood in Hindi/Urdu identified by Mohanan (1994), three of which can readily be applied to the possessors. I will establish these tests in this section and make use of them throughout the remainder of the chapter.

3.4.1 Reflexive Binding

The first subject test for Hindi/Urdu discussed by Mohanan (1994) is the apna test. The reflexive adjective apna ‘self’ can take as its antecedent only logical subjects or grammatical subjects, but not objects of any kind. This diagnostic is given in (31). In both of the examples in (32), apna ‘self’ binds the subj of the clause. In other words, the subj is the only possible antecedent of the reflexive; ravi in (32b) is not a possible antecedent, since it is neither the logical (semantic) nor the grammatical (syntactic) subject of the clause.

(31) The reflexive apna ‘self’ must be bound by an L-SUBJ or subj within its minimal finite clause. (Mohanan, 1994, p. 125)
3.4. TESTING FOR SUBJECTHOOD IN HINDI/URDU

(32) a. ravi apni saikāl = par bētʰ-a
‘Ravi sat on self’s bike.’ (Mohanan, 1994, p. 122)

b. vijay = ne ravi = ko apni saikāl = par
bētʰ-a-ya
sit.CAUS-PERF.M.SG
‘Vijay seated Ravi on self’s bike.’ (Mohanan, 1994, p. 122)

Since the antecedent of apna ‘self’ must occur inside the minimal finite clause containing the reflexive, the antecedent cannot be outside its minimal finite clause. Thus, in (33), raja cannot be an antecedent of apna ‘self’, as raja is the subj of a finite clause that is not the same finite clause as the clause containing apna ‘self’.

(33) raja = ne kah-a ki mantri apne
gār ga-ya
house.M.SG.ACC go-PERF.M.SG
‘The king said that the minister went to self’s house.’ (Mohanan, 1994, p. 125)

Note that the concept of minimal finite clause corresponds to the notion of minimal finite domain defined in LFG as in (34), according to Dalrymple (2001), who builds on Dalrymple (1993).

(34) The minimal finite domain corresponds to the smallest f-structure f where f contains an attribute TENSE. (adapted from Dalrymple, 2001, p. 283)

For my purposes, since I couch my analysis in the framework of LFG, the diagnostic in (31) can thus be restated as in (35).

(35) The reflexive apna ‘self’ must be bound by an L-SUBJ or SUBJ within its minimal finite domain (i.e., the smallest f-structure with an attribute TENSE).

3.4.2 PRONOMINAL COREFERENCE

A second subject test identified by Mohanan (1994) relates to pronominal coreference. Mohanan (1994) argues that the generalization in (36) is a reliable diagnostic for subjecthood in Hindi/Urdu.

(36) The antecedent of a pronoun cannot be a c-commanding SUBJ of the minimal finite clause that contains the pronoun. (Mohanan, 1994, p. 135)

The notion of c-command used in (36), originally formulated by Reinhart (1976) as a relation on syntactic trees, is analogous to the relation of f-command in LFG. F-command
was originally defined by Bresnan (1982a) as in (37). This is also the definition given by Dalrymple (2001). 16

(37) F-command: \( f \) f-commands \( g \) if and only if \( f \) does not contain \( g \), and all \( f \)-structures that contain \( f \) also contain \( g \).

(38) The antecedent of a pronoun cannot be an \( f \)-commanding \( \text{SUBJ} \) of the minimal finite clause that contains the pronoun.

As with the diagnostic regarding reflexive binding, the notion of minimal finite clause can be replaced by the LFG-compatible concept of minimal finite domain. Hence, the final version of the diagnostic is as in (39).

(39) The antecedent of a pronoun cannot be an \( f \)-commanding \( \text{SUBJ} \) of the minimal finite domain (i.e., the smallest \( f \)-structure with an attribute \text{TENSE}) that contains the pronoun.

16 Dalrymple (2001), p. 159, further provides an extension of the definition in (37), arguing that (37) may not make the correct predictions in cases where different attributes have identical values, for example, in functional control structures. For my purposes, though, the definition in (37) is sufficient. It is also important to note that while \( f \)-command may be analogous to \( c \)-command, as it was defined to state similar generalizations, the two notions are fundamentally different nonetheless. \( c \)-command is exclusively a relation between (parts of) syntactic trees. \( f \)-command, on the other hand, is in principle independent of constituent configurations as it is defined over \( f \)-structures. For that reason, \( f \)-command makes very different predictions, and has been argued to be the determining factor in formulating binding generalizations across languages (Bresnan, 1982a, 2001; Dalrymple, 1993, 2001; Zaenen, 1983). Just to recount a run-of-the-mill example from Bresnan (1982a) (p. 334): In Contradicting himself will discredit Mr. Jones, the NP Mr. Jones does not \( c \)-command, but \( f \)-command the subject position inside the complement clause Contradicting himself.
To see how the diagnostic in (39) works, consider the example in (40). In (40b), ravi as well as the pronoun us are occurring in the same minimal finite domain; in particular, ravi fcommands the pronoun. Since ravi can never be the antecedent of the pronoun, it can be the subj by (39). This prediction turns out to be correct. In (40c), both vijay and ravi f-command the pronoun; however, only ravi is a possible antecedent of the pronoun, which means that vijay is the subj of the clause.

(40)  
a. ravi saikal = par bɛṭʰ-a
   Ravi.M.SG.NOM bicycle.F.SG = LOC.ON sit-PERF.M.SG
   ‘Ravi sat on a/the bike.’

b. ravi us = ki saikal = par
   Ravi.M.SG.NOM PRON.3.SG.OBL = GEN.F.SG bicycle.F.SG = LOC.ON
   bɛṭʰ-a
   sit-PERF.M.SG
   ‘Ravi, sat on his bike.’ (Mohanan, 1994, p. 126)

   c. vijay = ne ravi = ko us = ki
       saikal = par bɛṭʰ-a-ya
       bicycle.F.SG = LOC.ON sit.CAUS-PERF.M.SG
       ‘Vijay seated Ravi on his bike.’ (Mohanan, 1994, p. 126)

3.4.3 Control

In addition to reflexive binding (the apna test) and pronominal coreference, Mohanan (1994) presents a third test for subjecthood. Control of obligatory pronominal control sites in participial adjunct clauses is only possible for nominals that are subjects. She gives the formal definition in (41).

(41)  
A nominal that can control a participial adjunct clause with an obligatory control site must be a subj. (Mohanan, 1994, p. 135)

The diagnostic is motivated by examples such as (42). Here, the participial adjunct clause includes an obligatory pronominal control site. The subject inside the participial adjunct clause may not be realized by an overt NP/KP; this is shown by (42b). Instead, it must be controlled by another GF in the matrix clause. (42c) then shows that the controlling GF must be the subj GF of the matrix clause.

(42)  
a. ravi = ne vijay = ko bɛṭʰ-a-ya
   ‘Ravi seated Vijay.’

b. * ravi = ne vijay = ko nina = ke
   muskura-te hu-e bɛṭʰ-a-ya
   smile-IMPF.M.SG.OBL be.PERF-M.SG.OBL sit.CAUS-PERF.M.SG
   ‘Ravi seated Vijay while Nina smiling.’ (Mohanan, 1994, p. 128)
c. ravi = ne vijay = ko muskura-te hu-e
Ravi.M.SG = ERG Vijay.M.SG = ACC smile-IMPF.M.SG.OBL be.PERF-M.SG.OBL
bi[t]a-ya
sit.CAUS-PERF.M.SG
‘Ravi, seated Vijay while __i/sj smiling.’ (Mohanan, 1994, p. 128)

Mohanan (1994) notes that while (41) is a reliable subjecthood indicator, its inverse is not necessarily also true. That is, a nominal that cannot control an obligatory pronominal control site may also turn out to be a subject. She gives the example in (43). (43a) includes a stative predicate, and while the *apna test for reflexive binding (43b) and the pronominal coreference test (43c) both indicate that mohan is the SUBJ of the clause, the result of the control test is negative (43d), even though mohan clearly is the subject of (43a).

(43) a. mohan us kamre = mē he
Mohanan.M.SG.NOM PRON.3.SG.OBL room.M.SG.OBL = LOC.IN be.PRES.3.SG
‘Mohan is in that room.’ (Mohanan, 1994, p. 129)

b. mohan apne kamre = mē he
‘Mohan is in self’s room.’ (Mohanan, 1994, p. 129)

c. mohan us = ke
Mohanan.M.SG.NOM PRON.3.SG.OBL = GEN.M.SG.OBL
kamre = mē he
room.M.SG.OBL = LOC.IN be.PRES.3.SG
‘Mohan is in his room.’ (Mohanan, 1994, p. 129)

d. * mohan muskura-te hu-e
Mohanan.M.SG.NOM smile-IMPF.M.SG.OBL be.PERF-M.SG.OBL
us kamre = mē he
PRON.3.SG.OBL room.M.SG.OBL = LOC.IN be.PRES.3.SG
* ‘Mohan is in that room __i/sj smiling.’ (Mohanan, 1994, p. 130)

I conjecture that the example in (43d) is ungrammatical for a different reason, namely the clash between the stative matrix predicate and the active embedded predicate. If one controls against differences in aspectual category between the matrix and the embedded clause, the resulting clauses are grammatical and the matrix SUBJ may serve as a controller of the obligatory pronominal control site in the embedded clause. See (44) for an example, where both the embedded as well as the matrix predicate are of stative nature.

(44) cuha bahot cʰota janwar ho-te
rat.M.SG.NOM very small.M.SG animal.M.SG.NOM be-IMPF.M.SG.OBL
hu-e us kamre = mē he
be.PERF-M.SG.OBL PRON.3.SG.OBL room.M.SG.OBL = LOC.IN be.PRES.3.SG
‘The rat is in that room __i being a very small animal.’
Given the grammaticality of (44), I assume for my purposes that the inverse of (41) is also true and reformulate (41) to give the diagnostic in (45).

(45) A nominal that can control a participial adjunct clause with an obligatory control site must be a **subj**. A nominal that cannot control a participial clause with an obligatory control site cannot be a **subj**, provided that other factors (e.g., aspectual differences) are controlled for.

With the tests for subjecthood in place, I now turn to a detailed survey of the possessive in Hindi/Urdu. The survey looks at the genitive phrases as they occur within the limits of the NP (§3.5), focusing on their constituent as well as functional properties.

3.5 Hindi/Urdu Genitive Arguments

This section surveys the use of the Hindi/Urdu genitive. I discuss its constituent properties as well as its functional properties. Regarding its constituent properties, I show that the linear order of the genitive phrase with respect to other modifiers in the NP is less strict than is usually acknowledged in the literature; for example, the genitive may occur non-initially in the NP. Genitives inside the NP may not, however, occur to the right of the NP head. In addition, the genitive phrase may also be scrambled outside of the NP they are licensed in. Concerning the functional properties of genitive modification, the section shows that Hindi/Urdu distinguishes four types of head nouns: common nouns, event nouns, relational nouns as well as proper nouns. The section further demonstrates that these noun types differ in the way they specify genitive arguments, and that three GFS can be distinguished for genitive KPS: **adjunct**, **subj** as well as **obj**.

3.5.1 Constituent Properties

This section surveys the constituent properties of the genitive in the NP. I discuss the genitive’s properties regarding coordination, linear order as well as stacking. Genitive scrambling to positions outside of the NP the genitive is licensed in is discussed in a separate section in §3.6.

3.5.1.1 Coordination

The NP marked by the genitive case clitic may be coordinated with further NPs. In this case, the clitic may attach to the whole coordinated NP. Furthermore, the whole genitive KP may also be coordinated with other KPs. This gives rise to ambiguous structures as can be seen in (46). The referents of (46a) unambiguously are cars that simultaneously belong to the boy as well as the girl. (46b) is ambiguous; the referents may be cars that simultaneously belong to the boy as well as the girl, or the referents may be cars that belong to the boy as well as (other) cars that belong to the girl.
   boy.M.SG.OBL and girl.F.SG = GEN.F.PL car.F.PL
   ‘the boy’s and the girl’s cars’

   ‘the boy’s and the girl’s cars’

The fact that the genitive case clitic may take scope over the whole (optionally coordinated) NP is typical for the case clitics found in Hindi/Urdu (Butt, 1995, Butt and King, 2004b, Mohanan, 1994); it is an essential part of the line of argumentation in favor of treating the case markers as clitics (see §2.3.2). For more examples regarding the genitive pattern in coordination, see Payne (1995).

3.5.1.2 LINEAR ORDER

Several elements may intervene between the possessor and possessum. While the general pattern is as in (47), other orders are possible. (47–51) are examples; (49) has been taken from the Hindi-Urdu Treebank (HUTB, Bhatt et al., 2009, Palmer et al., 2009), while (51) is among the results of a Google search.17 (47–49) show that an attributive adjective modifying the head noun may precede the possessor or intervene between the possessor and the head noun. Numerical adjectives may also precede the possessor (50–51).

(47) a. [ram=ka bɑṛa mɑkan]NP
   ‘Ram’s big house’

b. [bɑṛa ram=ka mɑkan]NP
   ‘Ram’s big house’

(48) a. [ram=ki nili gaṛi]NP
   Ram.M.SG = GEN.F.SG blue.F.SG car.F.SG
   ‘Ram’s blue car’

b. [nili ram=ki gaṛi]NP
   blue.F.SG Ram.M.SG = GEN.F.SG car.F.SG
   ‘Ram’s blue car’

(49) kɑmai=ka accʰa sadʰɑn
   earning.F.SG = GEN.M.SG good.M.SG measure.M.SG
   ‘the acceptable size of the earning’

(50) a. [nadya=ke do beṭe]NP
   Nadya.F.SG = GEN.M.PL two son.M.PL
   ‘Nadya’s two sons’

17I searched Google for the query दो उसे doʊs=ke ‘two of his’. In the search result (51) was extracted from, it was clear from context that (51) was a single NP.

18HUTB file fullnews_id_2489472-a_date_30_5_2004.dat, sentence ID 9
3.5. HINDI/URDU GENITIVE ARGUMENTS 99

b. [do nadya = ke bete]NP
two Nadya.F.SG = GEN.M.PL son.M.PL
‘Nadya’s two sons’

(51) do us = ke buce
two PRON.3.SG.OBL = GEN.M.SG child.M.PL
‘two of his/her children’

Examples where the adjective precedes the genitive phrase may be ambiguous; the adjective may modify the noun marked by the genitive, or it may modify the head noun of the overall NP, provided the adjective is semantically compatible with both, and its morphological agreement features permit both readings. An ambiguous example is given in (52); in such cases, it is found that local attachments (i.e., readings where the adjective modifies the genitive-marked noun) are preferred over non-local attachments. The preference for local attachment is well-known from cognitive science, and is discussed by Kimball (1973) as the “Right Association” principle; Gibson (1991) reformulated the principle as the “Recency Preference”.

(52) coti larki = ki kitab
‘the small girl’s book’
preferred over
‘the small book of the girl’

While the semantic compatibility between adjectives and nouns are impossible to test for in a computational LFG grammar, the possible syntactic attachments can be controlled for by checking for the morphological agreement features. If these features permit ambiguity, preferences such as the one in (52) can be implemented in XLE using OT marks (§2.5.3.10). See §3.6.2.1 for a further preference of local attachment concerning genitive scrambling.

In (1–3), it was seen that the canonical word order within NPs is possessor first, posses- sum last. NPs where the possessor occurs to the right of the possessum are unacceptable; the Hindi/Urdu NP conforms to the general head-final word order of the language (Butt, 1995, Kachru, 2006, Mohanan, 1994, Schmidt, 1999).

(53) a. [ram = ka bara makan]NP
‘Ram’s big house’

http://www.srijangatha.com/

I would like to suggest that the ambiguity created by linear orders as in (52) is one of the reasons why the canonical NP-internal order is genitive-initial, as shown in §3.2.2.1. The fact that possible ambiguities of adjective attachment are averted would readily explain the canonical order, although a precise theoretical assessment of these effects requires more research, in particular with respect to other NP-internal modifiers such as quantifiers etc.
b. [bərə ram = ka makan]_{NP}  
‘Ram’s big house’
c. * [bərə makan ram = ka]_{NP}  

(54) a. [ram = ki nili garî]_{NP}  
Ram.M.SG = GEN.F.SG blue.F.SG car.F.SG  
‘Ram’s blue car’
b. [nili ram = ki garî]_{NP}  
blue.F.SG Ram.M.SG = GEN.F.SG car.F.SG  
‘Ram’s blue car’
c. * [nili garî ram = ki]_{NP}  
blue.F.SG car.F.SG Ram.M.SG = GEN.F.SG

(55) a. [nadya = ke do beṭe]_{NP}  
Nadya.F.SG = GEN.M.PL two son.M.PL  
‘Nadya’s two sons’
b. [do nadya = ke beṭe]_{NP}  
two Nadya.F.SG = GEN.M.PL son.M.PL  
‘Nadya’s two sons’
c. * [do beṭe nadya = ke]_{NP}  
two son.M.PL Nadya.F.SG = GEN.M.PL

The assumption that the genitive has to precede its head inside the NP is corroborated by the fact that KPs with overt case clitics heading genitive-final NPs are ungrammatical, cf. the b. examples in (56–57). This is a clear indication that the genitive phrases cannot be right-adjointed to the NP head.

(56) a. [[nadya = ke do beṭe]_{NP} = ne]_{KP} garî = ko cɑla-yi
he be.PRES.3.SG  
‘Nadya’s two sons have driven the car.’
b. * [[do beṭe nadya = ke]_{NP} = ne]_{KP} garî = ko
two son.M.PL Nadya.F.SG = GEN.M.PL = ERG car.F.SG = ACC
cɑla-yi he drive-PERF.F.SG be.PRES.3.SG

(57) a. amin [[anjʊm = ki beṭi]_{NP} = se]_{KP}
Amin.M.SG.NOM Anjum.F.SG = GEN.F.SG daughter.F.SG = INST
zu = tak  
gɑy-a
zoo.M.SG = LOC.TOWARDS go-PERF.M.SG
‘Amin went to the zoo with Anjum’s daughter.’
3.5. HINDI/URDU GENITIVE ARGUMENTS

Finally, Hindi/Urdu determiners may co-occur with the possessor, either preceding or following it. This is shown in the examples in (58–59); the examples further show that the linear order among attributive adjectives, determiners and possessors is completely free, while the meaning is preserved.

(58)  

a. nina=ka yeh surx makan  
Nina.F.SG = GEN.M.SG this red house.M.SG  
‘this red house of Nina’

b. nina=ka surx yeh makan

c. yeh nina=ka surx makan

d. yeh surx nina=ka makan

e. surx yeh nina=ka makan

f. surx nina=ka yeh makan

(59)  

a. ʊstad=ka kʊch hoʃyar talɪb-ɪlm  
teacher.M.SG = GEN.M.SG some smart student.M.PL  
‘some smart students of the teacher’

b. ʊstad=ka hoʃyar kʊch talɪb-ɪlm

c. kʊch ʊstad=ka hoʃyar talɪb-ɪlm

d. kʊch hoʃyar ʊstad=ka talɪb-ɪlm

e. hoʃyar kʊch ʊstad=ka talɪb-ɪlm

f. hoʃyar ʊstad=ka kʊch talɪb-ɪlm

3.5.1.3 STACKING

It can further be observed that the possessors may be stacked on top of one another. That is, a possessor KP may itself be modified by a possessor KP as in the examples in (60). Recall that the genitive case clitic agrees in number, gender as well as morphological form with its head noun. In (60a), the possessor nina=ke ‘Nina’s’ modifies beṭe ‘son’; since beṭe is in the oblique form (itself being marked by the genitive), the genitive in nina=ke needs to be in the oblique as well. The complex phrase nina=ke beṭe=ki ‘Nina’s son’s’ in turn modifies gaṛi ‘car’, and here, both the case clitic and the head noun bear nominative morphology. A very similar example with different agreement morphology is given in (60b).

(60)  

a. nina=ke beṭe=ki gaṛi  
Nina.F.SG = GEN.M.SG.OBL son.M.SG.OBL = GEN.F.SG car.F.SG  
‘Nina’s son’s car’
b. ləɾke=ki  topi=ka  rang  
‘the boy’s hat’s color’  

Genitive stacking also gives rise to strings where multiple structures are possible (i.e., where the agreement morphology does not rule out either structure). For the NP in (61a) the readings indicated in (61b–c) are possible. In (61b) the possessor nina=ki ‘Nina’s’ modifies sone ‘gold’, and the complex phrase nina=ki sone=ki ‘Nina’s gold’s’ in turn modifies gʰɑṛi ‘watch’, rendering the meaning ‘the watch of Nina’s gold’ (i.e., a reading where the watch was fabricated using Nina’s gold). In contrast, in (61c), sone=ki modifies gʰɑṛi, and nina=ki modifies sone=ki gʰɑṛi, rendering ‘Nina’s golden watch’ (a watch belonging to Nina and made of gold, but not necessarily Nina’s gold).

(61)  
a. nina=ki sone=ki gʰɑṛi  
b. [[nina=ki  
sone=ki] = ki  
gʰɑṛi]  
‘the watch of Nina’s gold’  
c. [nina=ki  
sone=ki  
gʰɑṛi]]  
‘Nina’s golden watch’  

3.5.1.4 SUMMARY

I summarize the findings regarding the linear order properties of the genitive in NPs as follows. While the general order of constituents is as identified by Payne (1995), repeated in (52a), the variations in the ordering regarding the genitive shown in (62b–c) are also possible. Note that the actual Hindi/Urdu NP template is more complicated; for purposes of simplification, I omit further interactions with quantifiers, adverbs, pronouns, etc. For a more complete overview, including complex NPs with relative and adverbial clauses, see e.g. Kachru (2006) and Verma (1971). An implementation in terms of a computational grammar needs to account for these various possible word orderings of the genitive.

(62)  
a. Possessor — Determiner — (Numeral) Adjective — Noun  
(Payne, 1995, p. 294)  
b. Possessor — (Numeral) Adjective — Determiner — Noun  
c. Determiner — Possessor — (Numeral) Adjective — Noun  
d. Determiner — (Numeral) Adjective — Possessor — Noun  
e. (Numeral) Adjective — Determiner — Possessor — Noun  
f. (Numeral) Adjective — Possessor — Determiner — Noun
3.5.2 Functional Properties

What kind of f-structure should the genitive phrases receive in an LFG analysis? If one was to choose among the grammatical functions as they were described in §2.4.4, which would it be? Is the function subcategorized for by the nominal head, or must the genitive phrases be treated as adjuncts? Are genitive phrases restricted to realizing a certain thematic role, or must they be treated as an unrestricted function? Is the function discourse-oriented, e.g., can it bind an anaphora in the NP domain? In this section, I discuss the functional properties of the genitive phrases in NPs. After a general statement about semantic unrestrictedness, I turn to a discussion of the different kinds of Hindi/Urdu nouns (common, event, relational as well as proper nouns) and the ways that they may be modified by possessors. The discussion will directly inform the implementation of genitive phrase nominal arguments in XLE.

3.5.2.1 Semantic Unrestrictedness

As noted by many researchers across a wide variety of languages, possessors encompass many more semantic roles than just true possession. This is also true of Hindi/Urdu. The examples in (63) illustrate; see also the overview descriptions in Platts (1909), Raza (2010) and Bögel and Butt (2013). Note that the examples in (63) merely give an overview of the possible relations encoded by the genitive in Hindi/Urdu; I am not aware of a complete characterization of its semantics. Listing all usages of the genitive in a given language has proven to be notoriously difficult, if not impossible altogether (Williams, 1982). However, there have been some attempts at describing the semantic mechanisms that build up genitive relations (e.g. Barker, 1995). Note that in the examples in this and later sections, I occasionally use labels such as MATERIAL, KINSHIP, AGENT etc. to indicate the semantic role realized by the genitive phrase. The semantic role labels are not taken from a well-defined set of nominal thematic roles; in a fashion similar to Chisarik and Payne (2003), I use these labels only to indicate the variation in the semantic roles of genitive phrases.

(63) a. lɑkṛi =ki mez
   wood.F.SG = GEN.F.SG table.F.SG
   ‘the wooden table’ MATERIAL

b. uska beta
   PRON.3.SG.OBL = GEN.M.SG son.M.SG
   ‘her/his son’ KINSHIP

c. nina = ka makan
   Nina.F.SG = GEN.M.SG house.M.SG
   ‘Nina’s house’ POSSESSION

d. nina = ka nak
   Nina.F.SG = GEN.M.SG nose.M.SG
   ‘Nina’s nose’ (HAVING) BODYPART
I will not attempt a full semantic characterization of the possessive relation. My argument is that there are three natural subgroups of the possessive in Hindi/Urdu that cut across thematic roles. These three subgroups need to be distinguished in terms of their syntactic function, i.e., the GF that they encode inside the nominal f-structure that they are licensed in. The three GFs to be identified are ADJUNCT, SUBJ as well as OBJ. Prompted by recent work on possessives in LFG (see §3.3), I consider the possibility that possessives are selected as SUBJ by the nominal head; however, this option only turns out to be the correct analysis for a parts of the data; other nouns subcategorize for an OBJ, and all nouns may further be freely modified by ADJUNCT genitives.

3.5.2.2 Testing for Subjecthood inside NPs

Unfortunately, only one of the tests defined in §3.4 can be applied to the NP domain in Hindi/Urdu. Since *apna* ‘self’ is an adjective, it can be inserted into a noun phrase after the genitive KP to test whether it can be bound by the NP marked by the genitive. If this turns out to be possible, this constitutes strong evidence for the genitive KP realizing the SUBJ function inside the NP. Note, however, that if this is possible, it also requires that the subject diagnostic concerning *apna* ‘self’ in (35) be changed. The diagnostic, as it is formulated now, refers to the smallest f-structure with an attribute TENSE as the binding domain where *apna* ‘self’ is bound by a SUBJ. NPS, however, lack TENSE. Thus, if it is an NP-internal SUBJ that binds the reflexive, the reflexive binding diagnostic must be reformulated: *apna* ‘self’ is not bound within its minimal finite domain, but within the minimal domain containing a subject. If there is no NP-internal SUBJ, the reflexive is bound by a clausal SUBJ. In LFG’s binding theory, the relevant binding domain is the minimal complete nucleus. I reformulate the *apna* diagnostic in §3.8.
3.5.2.3 Common Nouns

Common count nouns in Hindi/Urdu, as in the other languages surveyed above, do not come with lexically specified genitive arguments; they are inherently non-relational in the sense of Barker (1995). They may still, however, be modified by genitive kps of two different sorts. (64) lists some examples of Hindi/Urdu common count nouns.

(64) * mez ‘table’, gari ‘car’, seb ‘apple’, makan ‘house’, ...

Common mass nouns turn out to behave exactly like common count nouns with respect to genitive modification; therefore, I discuss mass nouns, such as the example nouns in (65), in this section together with count nouns.

(65) pani ‘water’, rɔʃni ‘light’, lakṛi ‘wood’, ...

First, common nouns may be modified by genitive phrases that describe some attribute of the head noun. Raza (2010) has termed these genitive phrases “attributive genitives”; I adopt this term for the present discussion. Some examples are shown in (66).

(66) a. lakṛi=ki mez
    wood.F.SG = GEN.F.SG table.F.SG
    ‘the wooden table’ MATERIAL

b. sorx rɑng=ki gari
    red color.M.SG = GEN.F.SG car.F.SG
    ‘the car of red color’ COLOR

c. do sal=ki billi
    two year.M.PL = GEN.F.SG cat.F.SG
    ‘the two-year-old cat’ AGE

To test for the GF that is filled by the attributive genitive, one can apply the test for reflexive binding. As the examples in (67) show, apna cannot be bound by attributive genitives (more exactly, by the nominal that is marked by the genitive in attributive genitives).

(67) a. * lakṛi = ki apnimez
    wood.F.SG = GEN.F.SG self.F.SG table.F.SG

b. * sorx rɑng = ki apnigari
    red color.M.SG = GEN.F.SG self.F.SG car.F.SG

c. * do sal = ki apnibilli
    two year.M.PL = GEN.F.SG self.F.SG cat.F.SG

Common nouns may be modified by multiple attributive genitives at the same time (Bögel and Butt, 2013, Raza, 2010, 2011). In the examples in (68), both possessors are
co-modifying the head noun, a fact that can be determined by looking at the morphosyntax of the genitive case clitics; if the structures in (68) were hierarchical in nature, one would find oblique marking on the first clitic, given the agreement pattern. Given the syncretism of the oblique and the nominative in the feminine, the pattern can be seen most clearly with the masculine gender as in (68b), where the first genitive case clitic is not in the oblique form and can thus not be analyzed as modifying surx rang ‘red color’, which itself is in the oblique.

(68)  

a. lakṛi=ki surx rang=ki mez
    ‘the table of red color made of wood’

b. lakṛi=ka surx rang=ka makan
    ‘the house of red color made of wood’

c. sone=ki do sal=ki gʰɑṛi
    ‘the two year old watch made of gold’

Hindi/Urdu is different from other languages such as English or German in that it does not place any constraints on the number of genitive modifiers possible in NPs. The example in (69a) from Hindi/Urdu involves three distinct possessors modifying makan ‘house’; the corresponding examples from English and German are ungrammatical.

(69)  

a. lakṛi=ka surx rang=ka do
    wood.F.SG = GEN.M.SG red color.M.SG = GEN.M.SG two
    sal=ka makan
    year.M.PL = GEN.M.SG house.M.SG
    ‘the two year old house of red color made of wood’

b. * the house of wood of red color of two years

c. * das Haus roter Farbe
    the.N.SG.NOM house.N.SG.NOM red.F.SG.GEN color.F.SG.GEN
    festen Holzes
    firm.N.SG.GEN wood.N.SG.GEN

The negative result from the apna test shows that attributive genitives cannot be subjects. The fact that multiple attributive genitives may co-modify common nouns is evidence for assuming that these genitive phrases are not subcategorized as GFS by the common noun at all. If they were subcategorized GFS, the LFG constraint on coherence (Kaplan and Bresnan, 1982) would rule out grammatical structures like the ones in (68) and (69a). Thus, these genitive phrases must be treated as ADJUNCTS.

Aside of attributive genitives, common nouns may also be modified by other genitives that respond positively to the apna test. (70) shows examples. The thematic roles encoded by such genitives vary; possession, membership, bodypart, part, agent, creator, are just some of the possible thematic roles. Thus, common nouns occur with what
Barker (1995) has termed extrinsic possessors (see §2.2.1): even though they do not intrinsically select for a possessor, they may be modified by an extrinsic possessor, which can take on various semantic roles. Syntactically, the argument is realized as a subj, as indicated by the apna test.22

(70) a. mera apna makan  
PRON.POSS.M.SG self.M.SG house.M.SG  
‘my; own; house’ POSSESSION
b. nina=ka apna fariq  
‘Nina’s; own; team’ MEMBER
c. nina=ka apna nak  
‘Nina’s; own; nose’ (HAVING) BODYPART
d. makan=ka apna darvaza  
‘the house’s; own; door’ (HAVING) PART
e. nina=ki apni kitab  
‘Nina’s; own; book’ AGENT/CREATOR
f. ram=ka apna ag  
‘Ram’s; own; fire’ INITIATOR

Extrinsic possessors may combine with attributive genitives to co-modify the head noun; however, apna ‘self’ may only be bound by the extrinsic possessor, not by the attributive genitive.

(71) nina=ka lakṛi=ka apna makan  
‘Nina’s; own, of wood;’

If extrinsic possessors are SUBJS, then LFG’s principle of coherence (Kaplan and Bresnan, 1982) predicts that multiple extrinsic possessors do not co-occur in a single NP. This prediction is borne out by the data. Consider the examples in (72). In (72a), the extrinsic possessor realizes an agent/creator role; the extrinsic possessor in (72b) is a

---

22 In fact, the reflexive binding subject test as defined in (31) suggests an alternative to the conclusion that the genitive fills the subj function. This alternative would be to assume that apna ‘self’ is bound by the logical subj, and not by the syntactic subj, in the NP, and that the nominal marked by the genitive ends up as a different grammatical function. This would imply that there is an argument-changing operation interpolated between the predicate-argument structure and the grammatical-functional structure, e.g., in terms of a lexical rule such as passivization. Since, in LFG, such rules are generally assumed to apply exclusively to verbs (e.g., Dalrymple, 2001, p. 201), I assume it is the syntactic subj that is the antecedent of apna ‘self’.
true possessor. (72c) shows that it is not possible to realize both of these roles with two extrinsic possessors co-modifying a head noun; to render the desired meaning Nina’s house that was built by Ram, one would need to use a participial construction as in (72d).

(72) a. ram = ka makan
Ram.M.SG = GEN.M.SG house.M.SG
‘Ram’s house’ AGENT/CREATOR
b. nina = ka makan
Nina.F.SG = GEN.M.SG house.M.SG
‘Nina’s house’ POSSESSION

(73) a. makan = ka kʊṭṭa
house.M.SG = GEN.M.SG dog.M.SG
‘the house’s dog’
b. * makan = ka apnaıı kʊṭṭa

(73a) In (73a), an animate head noun is modified by an inanimate genitive kp. (73b) indicates that that kp cannot be a subj. Under the adjunct reading of (73a) (e.g., where the genitive expresses a non-possessive association of some sort), the overall np is nevertheless grammatical. A similar example is in (73a); again, the inanimate genitive gaṛi = ki can only be an adjunct to orat ‘woman’.

Animacy is well-known to influence possessor selection in a number of other languages. In English, for example, animacy is known for significantly influencing the choice between the of-oblique and the prenominal genitive; for studies on this, see,
for example, Chisarik and Payne (2003), Rosenbach (2012), Szmrecsanyi (2013), Taylor (1989) as well as the contributions to Börjars et al. (2013). Klamer and Kratochvíl (2006) describe Teiwa and Abui, two Papuan languages, where animacy predicts the properties of morphological marking on the possessor.

Moreover, possessor selection is not the only place in Hindi/Urdu grammar that is influenced by animacy. It is, for example, well-known that inanimate direct objects are accusative-marked only if definite, while animates should be marked by accusative/dative case even if indefinite (Allen, 1951, Butt and King, 2004b, Masica, 1991, Mohanan, 1994, among others). Examples are provided in (74); the animate object in (74) is only grammatical if carrying accusative case, while the inanimate object is grammatical with nominative (74b) or accusative (74c) case, receiving a definite interpretation in the latter case.

(74) a. ɪla=ne  baco =ko/*bacco  uṭʰa-ya
    Il.la.F.SG = ERG child.M.SG.OBL = ACC/*child.M.SG.NOM lift-PERF.M.SG
    ‘Ila lifted a/the child.’ (Mohanan, 1994, p. 80)

b. ɪla=ne  har  uṭʰa-ya
   Il.la.F.SG = ERG necklace.M.SG lift-PERF.M.SG
   ‘Ila lifted a/the necklace.’ (Mohanan, 1994, p. 80)

c. ɪla=ne  har =ko  uṭʰa-ya
    Il.la.F.SG = ERG necklace.M.SG = ACC lift-PERF.M.SG
    ‘Ila lifted the necklace.’ (Mohanan, 1994, p. 80)

It is thus clear that the implementation of genitives, but also the Urdu ParGram grammar in general can benefit from the annotation of animacy in nominals.

I conclude that common nouns may be modified by attributive genitives as well as extrinsic possessors. The former must be treated as ADJUNCTS, while the latter must be treated as SUBJ.s, as indicated by the combinatory possibilities as well as reflexive binding. Since count nouns do not come with a lexically specified SUBJ, they must be augmented in the lexicon to subcategorize for one. Animate count nouns can only be augmented to take animate SUBJ.s. §4.3.3 describes an analysis and implementation of common noun argument selection.

3.5.2.4 EVENT NOUNS

There is a class of nouns in Hindi/Urdu that is derived from verbal roots. The verbal roots these nominals are derived from either stem from Hindi/Urdu itself or, in the case of Urdu, may stem from other languages such as Arabic or Persian (Raza, 2010). These nouns intrinsically select a genitive-marked argument; semantically, they are different from common nouns in that they specify the number and type of argument genitives they license in their lexical entries. I refer to these nominals as “event nouns”. Event nouns take part in what Barker (1995) refers to as intrinsic possession in that they put
strict conditions on the thematic role that is realized by the argument. (75) lists some examples of event nouns in Hindi/Urdu.


Most event nouns may only take a single genitive argument. An example for an event noun derived from an intransitive verbal root and modified by a possessor is given in (76a); as can be seen in (76b), the possessor may bind the reflexive apna. Nouns derived from transitive verbal roots do not place tight restrictions on their thematic role realized by the genitive. For example, the noun tabahi ‘destruction’, derived from a transitive verb, may have its agent or its patient realized as a genitive. In such cases, apna may be bound by the genitive phrase, no matter the thematic relation encoded by it. In both cases, the genitive phrase is a SUBJ.

(76)  a.  ṭrain=ki ravangi
    train.M.SG = GEN.F.SG departure.F.SG
    ‘the train’s departure’  AGENT [Raza, 2010]

           b.  ṭrain, =ki apni, ravangi
    train.M.SG = GEN.F.SG self.F.SG,NOM departure.F.SG
    ‘the train’s own departure’  AGENT

(77)  a.  ᴛɪnsan=ki tabahi
    ᴛɪnsan.M.SG = GEN.F.SG destruction.F.SG
    ‘the destruction of man’  PATIENT [Raza, 2010]

           b.  ᴛɪnsan, =ki apni, tabahi
    ‘man’s own destruction’  PATIENT

(78)  a.  ᴛɪnsan=ki tabahi
    ᴛɪnsan.M.SG = GEN.F.SG destruction.F.SG
    ‘the destruction by man’  AGENT [Raza, 2010]

           b.  ᴛɪnsan, =ki apni, tabahi
    ‘man’s own destruction’  AGENT

If an intrinsic possessor SUBJ is realized with event nouns, it bears the thematic role of the verbal root and is interpreted accordingly; in many cases, this role will be an agent or patient, but in principle, it depends entirely on the lexical semantics of the nominal. This view is consistent with theoretical views on nominalization expressed by Barker (1995) and Rappaport (2006), who also note that the interpretation of the genitives depends on the lexical argument structure of the event nominal.

There is a smaller class of event nominals that may select two distinct thematic roles as genitive phrases. Raza (2010) mentions, for example, the noun gʰɛrao ‘circumvention’,
explaining that it may take both a subject and a second GF, which he states is OBJ. Note that *apna* ‘self’ is only bound by the subject genitive (79b) and cannot be bound by the other genitive. (80) is another example. One of the GFs found with such nominals must be regarded as an unrestricted and discourse-oriented SUBJ (Chisarik and Payne, 2003, Sadler, 2000).

(79) a. *naoja*vanô = ka \[t^h\text{ane} = ka\]
youngster.M.PL.OBL = GEN.M.SG police-station.M.SG.OBL = GEN.M.SG
gʰɛrao
circumvention.M.SG
‘the circumvention of the police station by the youngsters’

(80) a. *sadr* = ka \[ɪlɛkʃan\]

president.M.SG = GEN.M.SG election.M.SG.NOM
dr-ne = ka \[ɛlan\]
do.CAUS-INF.M.SG.OBL = GEN.M.SG announcement.M.SG
‘the announcement made by the president to conduct elections’

b. *sadr* = ka \[ɪlɛkʃan\]

president.M.SG = GEN.M.SG election.M.SG.NOM
kara-ne = ka \[apna_{i,sj} ɛlan\]
dr.CAUS-INF.M.SG.OBL = GEN.M.SG self.M.SG announcement.M.SG
‘the announcement made by the president himself to conduct elections’

c. *risale* = ka \[ɪlɛkʃan\]
pamphlet.M.SG.OBL = GEN.M.SG election.M.SG.NOM
kara-ne = ka \[apna_{i,sj} ɛlan\]
dr.CAUS-INF.M.SG.OBL = GEN.M.SG self.M.SG announcement.M.SG
‘the announcement from the pamphlet itself to conduct elections’

The status of the other genitive (which *Raza* (2010) states is an object) is worthy of discussion. First, note that event nouns have verbal counterparts which can reasonably

*Raza* (2010) glosses *gʰɛrao* as ‘circumvention’; the noun can also be fittingly translated to English as *encirclement, surrounding*. Note that *gʰɛrao* has found its way into the Concise Oxford English Dictionary in 2004 under the definition in (i):

(i) *Gherao* n (pl. *gheraos*). Indian; a protest in which workers prevent employers leaving a place of work until demands are met; Origin: From Hindi. (Oxford Dictionaries, 2004, p. 598)
be assumed to select for a subject and an object; an example for a verbal expression corresponding to (79a) is shown in (81).

(81) \[ \text{nãojavanō = ne} \]
\[ tʰane = ko \]
\[ \text{youngster. M.PL. OBL = ERG. M.SG} \]
\[ \text{police-station. M.SG. OBL = GEN. M.SG} \]
\[ gʰer-a \]
\[ \text{surround-perf. M.SG} \]

‘The youngsters surrounded the police station.’

If these same functions are assumed to project in an NP as they do in verbal predication, then we find a SUBJ and an OBJ with nouns such as \( gʰer\)-ao, both of which are unrestricted functions in LFG. This, however, would entail that the “Asymmetrical Possessor Parameter” formulated by Chisarik and Payne (2003) for Hungarian (APP; see (26) in § 3.3.2) is not active in Hindi/Urdu, at least not for event nominals. The APP is repeated below in (82); it states that two unrestricted functions are not permitted in nominals.

(82) Asymmetrical Possessor Parameter (Chisarik and Payne, 2003, p. 197)

\[
\begin{array}{c|c}
\theta & \theta \\
\hline
\text{[-r]} & \text{[-r]}
\end{array}
\]

The other option for the non-SUBJ genitive is to assume that they are thematically restricted obliques. However, this assumption runs into problems. If the phrases were obliques, one would expect that the head nominal itself would assign case to the oblique, in which case one would find semantically or idiosyncratically-marked phrases (Butt and King, 2004b, see also the discussion in Dalrymple (2001), p. 26–27); instead, the phrases carry canonical genitive case.

Obliques are a class of arguments which are difficult to define. In particular, the differences between obliques and objects can be quite subtle, in that arguments which

\[ 25 \text{I note in passing that some of the event nouns that occur with multiple genitive-marked arguments can also appear with complement clauses, which are usually analyzed as occupying the COMP GF in LFG; an example is the noun } \epsilon\text{lan 'announcement' in (i). In such cases, however, the complement clause is not genitive-marked and occurs in postnominal position; I therefore exclude this particular construction from my present description.} \]

(i) \[ \text{sadr = ka} \]
\[ \text{el\epsilon\text{lan} = ne} \]
\[ \text{k\text{h us = ne} \]
\[ \text{ilek\text{jan}} \]
\[ \text{president. M.SG = GEN. M.SG} \]
\[ \text{announcement. M.SG} \]
\[ \text{that PRON. 3.SG. OBL = ERG election. M.SG. NOM} \]
\[ \text{k\text{ara-e-ga} \]
\[ \text{do.CAUS-SBJV-FUT. M.SG} \]

‘the announcement made by the president that he will conduct elections’

The COMP GF has recently attracted interest in LFG; in particular, its status as an independent construct has been challenged, and it has been proposed to merge it with the OBJ function (see, e.g., Alsina et al., 2005, Dalrymple and Lødrup, 2000, for discussions of the issues). I will not review these issues here, but I would like to state that if these issues are resolved towards including COMP in OBJ, the generalization that these nominals select a SUBJ and an OBJ would be preserved. What exactly drives the choice in the realization of the argument as a clausal complement vs. a simple genitive must be left for future work.
3.5. HINDI/URDU GENITIVE ARGUMENTS

look like obliques on the surface may in fact be objects. Butt et al. (1999b) discuss this issue with respect to English verbs such as refer (83), where the object is preceded by a preposition; the fact that document is an object can be seen from passivization (83b).

(83) a. Our employees frequently refer to this document. (Butt et al., 1999b, p. 51)
    b. This document is frequently referred to by our employees. (Butt et al., 1999b, p. 51)

I would like to argue that phrases marked by the Hindi/Urdu genitive cannot be obliques. My analysis and implementation involves a purely structural account of genitive case marking, where the genitive is assigned as an “elsewhere” case in the absence of other (semantic or idiosyncratic) case-marking requirements. Thus, I side with Raza (2010) in that I assume that nouns such as gʰerəo ‘circumvention’ take a subj and an obj, which implies that the app is not active in Hindi/Urdu, like it is not active in English (see §3.3.2).

Event nouns may be further modified by attributive genitive ADJUNCTS. Two such examples are given in (84). In (84a), for example, accʰe sɑdʰan=ka ‘of good measure’ attributively modifies mɪntɪxab ‘selection’. As with common nouns, the result of the apna test is negative for these genitives (85).

(84) a. sɑdʳ=ka accʰe sɑdʰan=ka
    mɪntɪxab
    selection.M.SG
    ‘the president’s selection of good measure’
    b. sɑdʳ=ka ek sɑfɑɾ=ka
    ‘the president’s travel of one year’

(85) a. sɑdʳi=ka accʰe sɑdʰanɪ=ka
    apnaɪ,ṣj mɪntɪxab
    self.M.SG selection.M.SG
    ‘the president’s own selection of good measure’
    b. sɑdʳi=ka ek sɑfɑɾɪ=ka apnaɪ,ṣj
    travel.M.SG
    ‘the president’s own travel of one year’

Although event nouns intrinsically subcategorize for one or two gfs, nps headed by these nouns may be missing one or both of their gfs. This is expected given Barker’s
According to Barker, suppressing a nominal argument results in the argument being existentially bound; i.e., the nominal argument is not realized in the syntax, but nevertheless implied by the semantics of the nominal. In the example in (47), the event nominal’s agent argument is suppressed; nevertheless, it is implied that someone is doing the traveling.

(86) safar ek hafte=tak rah-a
    travel.M.SG.NOM one week.M.SG.OBL = LOC.TO remain-PERF.M.SG
    ‘The travel lasted for a week.’

Event nouns that can select two GFs may be missing one (87b–c) or all (87a) of their GFs. If only single genitive appears with such nouns, it is either realized as the subj or as the obj of the predicate. That is, there are cases where single genitives appearing with these nouns are bound by apna (87b), but there are other cases where this is not possible (87c).

(87) a. gʰerao ek hafte=tak rah-a
    circumvention.M.SG one week.M.SG.OBL = LOC.TO remain-PERF.M.SG
    ‘The circumvention lasted for a week.’

b. nɑojɑvanõ = ka apna ɡʰerao ek
    hafte=tak rah-a
    week.M.SG.OBL = LOC.TO remain-PERF.M.SG
    ‘The circumvention by the youngsters themselves lasted for a week.’

c. * tʰane = ka apna ɡʰerao ek
    hafte=tak rah-a
    week.M.SG.OBL = LOC.TO remain-PERF.M.SG

The fact that any or all of the nominal arguments may be missing is only exemplary of nominal predication in general, as acknowledged by the theoretical works on the topic (see §2.2.1); nominal arguments are syntactically entirely optional (Barker and Dowty, 1993; Dowty, 1989; Higginbotham, 1983; Rappaport, 2006).27

27From an LFG point of view, it may seem odd to assume that a predicator can select for a single non-subject GF, here, an OBJ. Some LFG works have argued for the “Subject Condition”, which states that every predicator must have a subject (Baker, 1983; Bresnan, 2001, p. 311). On the other hand, the influential work by Rappaport (2006) suggests that while nouns share the argument structure of the verbs they are derived from, the GF inventories and mechanisms for linking the arguments to GFs in the spirit of LFG’s Lexical Mapping Theory (Bresnan and Kanerva, 1989; Bresnan and Moshi, 1990; Bresnan and Zaenen, 1990) have to be different for the two categories. In addition, prepositions across many languages (English, German and French are just some) usually take a single GF that is not considered a subject (cf. the computational LFG account by Butt et al. (1999b)), again questioning the “Subject Condition”. Here, I remain agnostic regarding the mapping of arguments and the relation between verbal argument frames and nominal argument frames, and register the fact that the data indicate that a single OBJ can occur as a GF selected by a nominal.
I conclude that event nouns intrinsically select for one or two arguments, based on the valency of the verbs they are derived from. Some event nominals derived from transitive verbal roots may occur with two distinct arguments; in this case, they subcategorize for a \textsc{subj} as well as an \textsc{obj}. Most event nouns may only subcategorize for a single \textsc{subj}, though. Additionally, the nominal argument(s) of an event noun may be suppressed, in which case they do not turn up in the subcategorization frame of the noun. Nevertheless, the existence of the argument(s) are implied by the lexical semantics of the nominal. Regarding event nouns that can appear with two distinct genitive \textsc{gfs}, either of the arguments may be suppressed.

3.5.2.5 RELATIONAL NOUNS

Relational nouns lexically entail a relationship with another entity. Examples from Hindi/Urdu are in (88). Relational nouns are different from event nouns in that they are not derived from verbal roots. They also can never take more than one genitive argument. A subclass of relational nouns are body part nouns; these also take a single genitive argument. Some examples for body part nouns are in (89).

(88) \textit{beṭi} ‘daughter’, \textit{bap} ‘father’, \textit{sɪra} ‘top’, ...
(89) \textit{nak} ‘nose’, \textit{ṭang} ‘leg’, \textit{ankʰ} ‘eye’, ...

Relational nouns intrinsically select for a possessive argument in the sense of Barker (1995). If a particular entity is a daughter, that fact alone entails the existence of a different entity who is the parent of that daughter. Similarly, it is hard to imagine the noun \textit{sɪra} ‘top’ without it being the top of something. Such relations are a part of the lexical semantics of the individual relational noun. (90-92) are some examples of NPs with relational head nouns. The b. examples show that \textit{apna} ‘self’ is bound by the genitive KP. Relational nouns thus subcategorize a \textsc{subj} possessor argument.

(90) a. \textit{mez=ka sɪra}  
\textit{table.F.SG = GEN.M.SG top.M.SG} 
‘the table’s top’
b. \textit{mez\textsubscript{i}=ka apna\textsubscript{i} sɪra}  
‘the table’s own top’

(91) a. \textit{nina=ki beṭi}  
\textit{Nina.F.SG = GEN.F.SG daughter.F.SG} 
‘Nina’s daughter’
b. \textit{nina\textsubscript{i}=ki apni\textsubscript{i} beṭi}  
\textit{Nina.F.SG = GEN.F.SG self.F.SG daughter.F.SG} 
‘Nina’s own daughter’
3.5. HINDI/URDU GENITIVE ARGUMENTS

(92)  a. meri ṭang
PRON.POSS.F.SG leg.F.SG
‘my leg’
meri ṭang
PRON.POSS.F.SG self.M.SG leg.F.SG
‘my own leg’

In addition, just like event and common nouns, they may be modified by additional attributive genitives, as in (93).

(93)  a. mez = ka surx rāng = ka ṭapna_i,ṣ j sira
‘the table’s own red top’
b. nina = ki do sal_j = ki ṭapni_i,ṣ j beṭi
‘Nina’s own daughter of two years’

The possessor argument of relational nouns may be suppressed, just like the argument(s) of event nominals as seen in the previous section. In (94), the relational noun beṭi ‘daughter’ occurs without its possessor argument. Nevertheless, it is implied by the lexical semantics that beṭi ‘daughter’ is related to a parent argument.

(94)  beṭi = ne kuṭṭe = ko kʰila-ya
daughter.F.SG = ERG dog.M.SG.OBL = ACC eat.CAUS-PERF.M.SG
‘The daughter fed the dog.’

(95) illustrates the animacy constraint discussed in §3.5.2.3 with a relational noun.
Just as animate common nouns, animate relational nouns may only be modified by animate subj genitives.

(95)  a. gʰudsal = ka beṭa
stable.M.SG = GEN.M.SG son.M.SG
‘the stable’s son’
b. * gʰudsal_j = ka ṭapni_i beṭa

Body part nouns, however, do not require animate possessor subjs (96).

(96)  a. mez = ki ṭapni_i ṭang
table.F.SG = GEN.F.SG self.F.SG leg.F.SG
‘the table’s own leg’
b. sui = ki ṭapni_i ankʰ
needle.F.SG = GEN.F.SG self.F.SG eye.F.SG
‘the needle’s own eye’
c. bɔtɑl = ki ṭapni_i gardɑn
bottle.F.SG = GEN.F.SG self.F.SG neck.F.SG
‘the bottle’s own neck’
3.5.2.6 PROPER NOUNS & PRONOUNS

Proper nouns as well as pronouns are generally assumed to not be able to take arguments at all. This is acknowledged both by the theoretical (Barker, 1995; Grimshaw, 1990) as well as the NLP literature (Meyers et al., 2004a,b,c) on the subject. A look at the Hindi/Urdu data reveals the same generalization: proper nouns cannot subcategorize for a subj (or obj) gf, and neither can pronouns. The ungrammatical data in (97) illustrate; here, the proper noun ram is modified by different genitive kps, none of which can be bound by apna ‘self’. Corresponding examples with pronouns are in (98).

(97) a. * Nina = ka apni Ram
b. * mera apni Ram
   PRON.POSS.M.SG self.M.SG Ram.M.SG
c. * bæce = ka apni Ram
d. * mokan = ka apni Ram

(98) a. * Nina = ka apni mē
   Nina.F.SG = GEN.M.SG self.M.SG PRON.1.SG
b. * bæce = ka apni mē
   child.M.SG.OBL = GEN.M.SG self.M.SG PRON.1.SG

Proper nouns and pronouns differ with respect to modification by attributive genitives as in (100–99). Genitive kp modification in general is not felicitous with pronouns (99). With proper nouns, attributive genitives are exceptional, but not impossible. For example, the NPs in (100) are acceptable, insofar as my informants were able to come up with situations where such modification is thinkable. However, if an attachment of the attributive genitive kp to another, non-proper noun is possible (given the possibility of genitive scrambling, see §3.6), they highly prefer the alternative reading. For example, in (101), it is possible that surx rang=ki ‘of red color’ non-locally modifies gari ‘car’, a reading which is preferred instead of a modification of the proper noun nina. In general, the generalizations concerning pronouns and proper nouns are expected given the highly referential character of these nominal expressions.

(99) a. * do sal = ki mē
   two year.M.PL = GEN.F.SG PRON.1.SG
b. * surx rang = ki mē
   red color.M.SG = GEN.F.SG PRON.1.SG

(100) a. do sal = ki nina
   two year.M.PL = GEN.F.SG Nina.F.SG
   ‘two year old Nina’
3.5. HINDI/URDU GENITIVE ARGUMENTS

b. \( \text{surx} \ \text{rang} = \text{ki} \quad \text{nina} \)  
\( \text{red} \ \text{color.M.SG} = \text{GEN.F.SG Nina.F.SG} \)  
‘Nina of red color’ (e.g., Nina wearing red clothes that day)

\( \text{(101)} \) \( \text{surx} \ \text{rang} = \text{ki} \quad \text{nina} = \text{ne} \quad \text{gaři} \quad \text{dek^h-i} \)  
\( \text{red} \ \text{color.M.SG} = \text{GEN.F.SG Nina.F.SG car.F.SG see-PERF.F.SG} \)  
‘A car of red color, Nina saw.’

preferred over  
‘Nina of red color saw a car.’

3.5.3 PRO-DROP OR ARGUMENT SUPPRESSION?

Above, I have argued that event and relational nouns undergo argument suppression quite freely. That is, when a nominal argument is realized, it is filled by a genitive phrase and, if it is a subject, is bound by the reflexive \( \text{apna} \) ‘self’; when it is suppressed, it is not realized in the syntax, but existentially bound by the lexical semantics of the nominal as per the account given by Barker (1995).

An alternative view to argument suppression involves an analysis of the non-overt genitive arguments in terms of pro(nominal)-drop. Under this view, the argument(s) of the nominal would never be suppressed, but always realized, though they may be dropped under the right circumstances. Several accounts of pro-drop assume that what drives pro-drop is overt agreement morphology on the syntactic heads that subcategorize for the dropped constituent (e.g., Koeneman, 2000, Prasad, 2000, Rizzi, 1986, among others). Other accounts argue that agreement morphology is not a necessary condition for pro-drop, but that discourse/information structure need to be taken into account, e.g., Butt and King (2000) for Hindi/Urdu and Butt (2007) for Punjabi. See Neeleman and Szendrői (2007) for a cross-linguistic discussion of pro-drop, also arguing that agreement is not a necessary condition for pro-drop.

What is common among the above approaches is that pro-drop is usually discussed with respect to verbal predication; that is, in those languages that permit pro-drop, it is commonly assumed that it is verbal arguments that are dropped. Pro-drop in the nominal domain has not found as much interest as verbal pro-drop in theoretical linguistics research, it seems; exceptions are Salzmann (2011) as well as references therein. An LFG account involving nominal pro-drop is Strunk (2005) (working on Low Saxon), the main features of which are summarized below.

An example from Low Saxon is given in (102); Strunk (2005) calls this construction the possessive pronoun construction (PPC). Here, the possessor is a fully-fledged posses-
sive pronoun, occurring in the same syntactic position as articles, demonstratives, etc., and being in complementary distribution with them; Strunk (2005), along with Dipper (2003), analyzes the pronoun to be of category D and a co-head to NP. Agreement-wise, the pronoun fulfills two different functions: its stem provides agreement features for its

\[\text{28}\] See Chapter §5 for other types of case marking on nominal arguments.
antecedent (i.e., the possessor), while at the same time the pronoun has to agree with the possessum in number, gender and case (103). Strunk (2005) proposes the lexical entries in (104a–b) for the pronoun and a head noun, respectively; the resulting c- and f-structures for (102) are shown in Figure 3.8.

(102) ehr Gesicht
her.F.SG face.N.SG
‘her face’
(adapted from Strunk, 2005)

(103) a. he geiht sien-en Weg
he goes his-M.SG.ACC way.M.SG.ACC
‘He goes his way.’

b. * he geiht sien-e Weg
he goes his-F.SG/PL.ACC way.M.SG.ACC

(104) a. Lexical entry for Low Saxon possessive pronoun (Strunk, 2005):
ehr D (↑ POSS PRED) = ‘pro’
(↑ POSS PERS) = 3
(↑ POSS NUM) = sg
(↑ POSS GEND) = f
(↑ NUM) = sg
(↑ GEND) = n
(↑ CASE) = acc

b. Lexical entry for Low Saxon head noun Gesicht ‘face’ (Strunk, 2005):
Gesicht N (↑ PRED) = ‘face<(↑ POSS)⟩’
(↑ PERS) = 3
(↑ NUM) = sg
(↑ GEND) = n
(↑ CASE) = acc

Low Saxon has another possessive construction, shown in (105a), which Strunk (2005) calls the possessive linker construction (PLC). Here, the possessor is a full DP, and is followed by a pronoun, which is of the same form as the possessive pronoun in (102); the pronominal in (105a) agrees with the possessor DP in number and gender, just like the pronoun in (102) provides agreement features for its antecedent. Strings without the pronoun are ungrammatical (105b). The question then follows what exactly the function of the pronoun in (105a) is, which Strunk (2005) calls the “possessive linker” in a PLC.

(105) a. [[de’n Jung] sien Vadder]
‘the boy’s father’

b. * [[de’n Jung] Vadder]
the.M.SG.ACC boy.M.SG.ACC father.M.SG
‘the boy’s father’
Strunk (2005) dismisses analyses involving a presumptive pronoun for empirical reasons, but instead argues that this represents a case of nominal pro-drop: when the possessor DP is non-overt (i.e., in the possessive pronoun construction), the possessive pronoun provides a PRED feature to the POSS feature, thus satisfying the completeness principle. On the other hand, when the possessor DP is overt (in the possessive linker construction), the pronoun does not have PRED, i.e., is not interpreted anaphorically, but merely agrees with the possessor by projecting agreement features. The updated lexical entry for the pronoun thus looks as in (106), where the PRED feature supplement for POSS has been made optional. This follows the standard approach to verbal pro-drop in LFG (Bresnan, 2001); in verbal pro-drop, the verbal morphology is taken to function like agreement morphology in presence of an overt complement phrase; if no complement is overt to agree with, the morphology provides a pronominal interpretation for a missing complement (Bresnan, 2001, p. 150). To rule out (105b), Strunk (2005) proposes to annotate the pronoun to project a POSS MARKING feature (line 2 of (106)), which is then required by the c-structure rule in (107). The result is the analysis in Figure 3.9 for (105a).


---

Reasons for rejecting a DP-external resumption analysis are that the PLC always forms one DP constituent and that it can occur in the middle of a clause. Reasons for rejecting a DP-internal resumption analysis include the fact that the DP in a PLC may contain a wh-word or a negative pronoun, which should be pragmatically odd if the linker pronoun were indeed a resumptive.

---

Figure 3.8: C- and f-structure for Low Saxon possessive pronoun construction
3.5. HINDI/URDU GENITIVE ARGUMENTS

Updated lexical entry for Low Saxon possessive pronoun (Strunk, 2005):

ehr  D  ((↑ POSS PRED) = ‘pro’)

(↑ POSS MARKING) = +

(↑ POSS PERS) = 3

(↑ POSS NUM) = sg

(↑ POSS GEND) = f

(↑ NUM) = sg

(↑ GEND) = n

(↑ CASE) = acc

(107) DP  →  DP  D’

(↑ POSS) = ↓  ↑ = ↓

(↑ POSS MARKING) = c +

Since Strunk (2005) assumes only a single lexical entry for a given possessive pronoun, i.e., the one in (106) for ehr ‘her’, the f-structure in Figure 3.8 should in fact also include MARKING + for the POSS function.

There are several differences between the Hindi/Urdu and the Low Saxon data. First, Hindi/Urdu does not exhibit use of pronouns similar to that of Low Saxon. The reflexive ɑpna ‘self’ has to be bound by a possessor SUBJ and is thus similar to the use of e.g. ehr ‘her’ in (105a) [31], but it otherwise generally cannot function as a possessive pronoun indexing a possessor on its own. An example is in (108). (108b) is ungrammatical for most of my informants since ɑpni is not bound by any antecedent. For some of my informants, though, (108b) is acceptable with the reading My/our daughter saw a car., where ɑpni is indeed reanalyzed as a possessive pronoun, maybe by invoking “logophoric conversion” (Reuland, 2001). However, this is not possible for all of the informants, and

[31]I have not discussed the precise semantics of ɑpna ‘self’ when bound by a possessor; here, more research is needed to work out the exact reasons for realizing the reflexive.
for none of them, *apni can be bound by *nina from (108a). If one were to apply an analysis involving a pro-dropped possessor SUBJ (i.e., as in the intended reading of (108b)), *apni could be bound by the dropped pronoun. This is not possible, although there is no a priori reason to assume that a dropped pronoun cannot bind an anaphor; cf. e.g. Gill (1999) for Korean.

(108)

<table>
<thead>
<tr>
<th>a.</th>
<th>*nina</th>
<th>bazar=mẽ</th>
<th>tʰ-i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nina.F.SG.NOM</td>
<td>market.M.SG = LOC.IN</td>
<td>be.PAST-F.SG</td>
<td></td>
</tr>
<tr>
<td>'Nina was in the market.'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.</td>
<td>*apni</td>
<td>beṭi=ne</td>
<td>gaṛi</td>
</tr>
<tr>
<td>self.F.SG daughter.F.SG = ERG car.F.SG see-PERF.F.SG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intended: ‘(Her/Nina’s) own daughter saw a car.’</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Second, *apna does not agree in any morphosyntactic feature with the possessor it is bound to, but exclusively with the head noun, which is further evidence of its status as a reflexive (i.e., not a fully-fledged pronoun) in the absence of logophoric conversion. Given these observations, it is hard in general to argue for a pro-drop analysis in Hindi/Urdu nominals, which is why I assume that argument suppression is taking place, and the reflexive has to be bound by an overt genitive; see also §3.6.3 for a comparison to Turkish possessives.

3.5.4 ARGUMENTS OR ADJUNCTS?

In light of the above, it is appropriate to challenge the argument status of the genitive arguments. In particular, since they can be freely omitted, why not treat them as adjuncts? The distinction between arguments and adjuncts is a frequently discussed issue in the linguistics literature; much of the literature focuses on the distinction in the context of verbal predication (e.g., Carnie, 2006, Haegeman, 1994, Kroeger, 2004, Radford, 2004, Tallermann, 2005), while there is also some discussion distinguishing arguments from adjuncts within nominals (e.g., Grimshaw, 1990, Lebeaux, 2000, Needham and Toivonen, 2011). In general, the concepts “argument” and “adjunct” are of crucial importance to linguistic theories. In LFG, for example, the distinction is central throughout: lexical entries of predication items (whether that be nominals, adjectives or verbs) only list arguments, not adjuncts; at f-structure, coherence and completeness are checked only with respect to governable GFS (Kaplan and Bresnan, 1982). Beyond syntactic theory, there are studies indicating that the distinction is in fact psycholinguistic reality (Lebeaux, 2000, Tutunjian and Boland, 2008).

The distinction between arguments and adjuncts is also a prominent topic in computational linguistics. In lexicon development, one would like to define only core participants of events for efficiency reasons (although PropBank, for example, lists some adjuncts with verbs while at the same time maintaining the distinction (Bonial et al., 2015, Palmer et al., 2005)). In parsing, one is in need of the distinction for applications
down the pipeline, such as sentence condensation, where adjuncts are a prime candidate for removal (Riezler et al., 2003).

The problem with the argument-adjunct distinction is that there is no single valid and universally agreed-upon definition of either concept, in spite of their importance across theories; rather, there are more or less vague descriptions, relating to event structure and its participants, as stated by e.g., Needham and Toivonen (2011):

[...], the general intuition is that, roughly speaking, arguments are the central, necessary participants in the event, whereas adjuncts provide “extra” information about where, when and how the event occurred. (Needham and Toivonen, 2011)

Note, however, that this diagnostic again cannot be applied across the nominal domain, since not all nominals contain event structure; see the discussions by Barker (1995) and Grimshaw (1990). Relational nouns in particular form an interesting case; Barker (1995) convincingly argues that they need to have argument structure, even though lacking event structure altogether.

Many of the syntactic tests identified in the literature are only useful with respect to verbal predicates. Needham and Toivonen (2011) discuss a battery of ten tests connected to prepositional selection, optionality, iterativity, VP anaphora and other features, noting that most of the tests sport weaknesses and should be used with some care; with respect to event nominals, they note that their arguments are an unclear case where the optionality indicates their adjunct status, but the fact that they are core participants of an event hints at them being arguments. Above, I have relied quite heavily on the apna test on reflexive binding as well as on optionality/iterativity to test for the argument/adjunct status of genitive phrases. I would like to argue that if these tests are combined, the result is the pattern illustrated above in Table 3.3. A similar picture regarding the argument/adjunct distinction in nominals emerges from Lebeaux (2000). In Chapter §4, I show how the pattern as well as the restrictions concerning nominal type and animacy can be analyzed in LFG and implemented in an XLE grammar.

3.5.5 SUMMARY

To summarize, I have argued that in general, there is no limitation on the number of possessors used to modify the noun (unlike in English, German, Hungarian or Welsh). Instead, what is found is that there are four types of nouns that differ with respect to genitive modification. Common nouns may be augmented for an extrinsic possessor subj, as may proper nouns, if the possessor itself is a proper noun. Event nominals and relational nominals intrinsically select for a subj. Most event nominals derived from transitive verbs can realize either of their arguments as the subj, while some of them may realize both, in which case they map onto a subj and an obj function, the latter of which is instantiated to a specific. Additionally, I have argued that attributive
genitives are ADJUNCTS, and that all of the nominal types above may be modified by such
genitives quite freely. The properties of the different nominal types are summarized in
Table 3.3, together with the types of genitive modifiers they occur with; EP is short
for extrinsic possession, while IP stands for intrinsic possession. The Binds apna ‘self’?
column specifies the behavior of the respective genitive type with respect to binding the
reflexive apna ‘self’. The last column titled GF gives the proposed grammatical function
for each genitive type. Proper nouns and pronouns are missing from this table, since
they do not occur with genitive modifiers.

<table>
<thead>
<tr>
<th>Noun type</th>
<th>Genitive type</th>
<th>Amount</th>
<th>Binds apna ‘self’?</th>
<th>GF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common count nouns</td>
<td>Attributive</td>
<td>0–∞</td>
<td>X</td>
<td>ADJUNCT</td>
</tr>
<tr>
<td></td>
<td>EP</td>
<td>0–1</td>
<td>✓</td>
<td>SUBJ</td>
</tr>
<tr>
<td>Common mass nouns</td>
<td>Attributive</td>
<td>0–∞</td>
<td>X</td>
<td>ADJUNCT</td>
</tr>
<tr>
<td></td>
<td>EP</td>
<td>0–1</td>
<td>✓</td>
<td>SUBJ</td>
</tr>
<tr>
<td>Event nouns</td>
<td>Attributive</td>
<td>0–∞</td>
<td>X</td>
<td>ADJUNCT</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>0–2</td>
<td>✓/X</td>
<td>SUBJ/OBJ</td>
</tr>
<tr>
<td>Relational nouns</td>
<td>Attributive</td>
<td>0–∞</td>
<td>X</td>
<td>ADJUNCT</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>0–1</td>
<td>✓</td>
<td>SUBJ</td>
</tr>
</tbody>
</table>

Table 3.3: Genitive KP types in the Hindi/Urdu NP

3.6 GENITIVE SCRAMBLING

“Scrambling” refers to the process of altering the basic order of constituents of a clause
(or smaller parts of a clause) to a different, alternative order. The term was first coined
by Ross (1967) and has since been established beyond the Chomskyan tradition through-
out generative grammar to describe many non-canonical word order phenomena across
languages. Scrambling usually affects sisters in the constituent tree; in many free word-
order languages, however, words may also be scrambled out of their constituents into the
clause. Usually, this is made possible in such languages by rich agreement morphology
in terms of case, gender and/or number. Hindi/Urdu is one of these languages.

Generally, scrambling is characterized across languages as an operation that reflects
extra-syntactic preferences, e.g., pragmatic/information-structural configurations such
as emphasis, topicalization etc. (Choi, 1999b, Karimi, 2003, Kidwai, 2000, among oth-
ers). My investigation is focused on describing the syntactic constraints that hold for the
operation; these constraints will then directly inform my implementation in the Urdu

---

I have deliberately ignored smaller common noun subclasses, such as measure nouns (e.g., sal ‘year’,
kalo ‘kilo’) or partitive nouns (e.g., ard‘anf ‘half’). There are several reasons for ignoring such nouns. First,
such nouns often display special requirements in terms of case marking and thus have to be excluded from
the present discussion of possessor/genitive marking. Second, most of these nouns cannot take arguments,
but may behave in distinct ways in other respects that go beyond the scope of this chapter.
ParGram grammar.

3.6.1 General Description

In addition to the variable word order inside NPs, there are examples showing that the genitive modifiers can occur outside of the NPs they modify. I refer to this as “Genitive Scrambling”. In (109a), the genitive occurs in the canonical position inside the NP to the left of the head noun. In (109b), the genitive is scrambled outside of the subject NP to the end of the clause; still, it must be analyzed as a modifier of the head noun *dost* ‘friend’, since it cannot be argued to be an argument of the intransitive verb *a* ‘come’.

\[(109)\]
\[\begin{align*}
\text{a. } & \text{ram}=\text{ka} \quad \text{dost} \quad \text{ay-a} \\
\text{Ram.M.SG} &= \text{GEN.M.SG friend.M.SG.NOM come-PERF.M.SG} \\
\text{‘Ram’s friend came.’} \\
\text{(Butt and Zinsmeister, 2009)}
\end{align*}\]
\[\begin{align*}
\text{b. } & \text{dost} \quad \text{ay-a} \quad \text{ram}=\text{ka} \\
\text{friend.M.SG.NOM come-PERF.M.SG Ram.M.SG} &= \text{GEN.M.SG} \\
\text{‘Ram’s friend came.’} \\
\text{(Butt and Zinsmeister, 2009)}
\end{align*}\]

In (110a), the object *gari* ‘car’ is modified by the genitive *us=ki* ‘her/his/its’. The genitive can be scrambled out of the object to the beginning of the clause as in (110b). From the morphosyntax, it is clear that in (110b) the feminine-inflected *us=ki* ‘her/his/its’ modifies *gari* ‘car’, since that is the only feminine nominal in the sentence. A very similar example is in (111). (112b) shows an example of a scrambled possessive pronoun; the canonical word order is shown in (112a).

\[(110)\]
\[\begin{align*}
\text{a. } & \text{ram}=\text{ne} \quad \text{us}=\text{ki} \quad \text{gari} \\
\text{Ram.M.SG} &= \text{ERG PRON.3.SG.OBL = GEN.F.SG car.F.SG.NOM} \\
\text{bazar} &= \text{mè} \quad \text{dekʰ-i} \\
\text{market.M.SG} &= \text{LOC.IN see-PERF.F.SG} \\
\text{‘Ram saw her/his car in the market.’} \\
\text{(adapted from Bögel and Butt, 2013, p. 301)}
\end{align*}\]
\[\begin{align*}
\text{b. } & \text{us}=\text{ki} \quad \text{ram}=\text{ne} \quad \text{gari} \\
\text{PRON.3.SG.OBL} &= \text{GEN.F.SG Ram.M.SG} = \text{ERG car.F.SG.NOM} \\
\text{bazar} &= \text{mè} \quad \text{dekʰ-i} \\
\text{market.M.SG} &= \text{LOC.IN see-PERF.F.SG} \\
\text{‘His/her car, Ram saw in the market.’} \\
\text{(adapted from Bögel and Butt, 2013, p. 301)}
\end{align*}\]

\[(111)\]
\[\begin{align*}
\text{a. } & \text{tom}=\text{ne} \quad \text{kis}=\text{ki} \quad \text{kitab} \quad \text{xarid-i?} \\
\text{you} &= \text{ERG who.SG.OBL = GEN.F.SG book.F.SG.NOM buy-PERF.F.SG} \\
\text{‘Whose book did you buy?’} \\
\text{(adapted from Bögel and Butt, 2013, p. 301)}
\end{align*}\]
\[\begin{align*}
\text{b. } & \text{kis}=\text{ki} \quad \text{tom}=\text{ne} \quad \text{kitab} \quad \text{xarid-i?} \\
\text{who.SG.OBL} &= \text{GEN.F.SG you} = \text{ERG book.F.SG.NOM buy-PERF.F.SG} \\
\text{‘Whose book did you buy?’} \\
\text{(Bögel and Butt, 2013, p. 301)}
\end{align*}\]
3.6. GENITIVE SCRAMBLING

(112) a. \[\text{nina}=\text{ne} \quad \text{mera} \quad \text{makan} \quad \text{dek}^b-\text{a}\]
\[\text{Nina.F.SG} = \text{ERG PRON.POSS.M.SG house.M.SG.NOM see-PERF.M.SG}\]
‘Nina saw my house.’

b. \[\text{mera} \quad \text{nina}=\text{ne} \quad \text{makan} \quad \text{dek}^b-\text{a}\]
\[\text{PRON.POSS.M.SG Nina.F.SG} = \text{ERG house.M.SG.NOM see-PERF.M.SG}\]
‘My house, Nina saw.’

Genitives may also be scrambled to the right. In (113a), a permutation of (110a), the object is topicalized to the front of the clause. In (113b), the genitive phrase modifying the object is scrambled to the right and occurs after the subject. A similar example is given in (114), where \[\text{kis}=\text{ki} \] ‘whose’ modifies \[\text{kitab} \] ‘book’, but is not in the same constituent.

(113) a. \[\text{us}=\text{ki} \quad \text{ga}=\text{r}i \quad \text{ram}=\text{ne}\]
\[\text{PRON.3.SG.OBL} = \text{GEN.F.SG car.F.SG.NOM Ram.M.SG} = \text{ERG}\]
\[\text{bazar}=\text{m}ě \quad \text{dek}^b-\text{i}\]
\[\text{market.M.SG} = \text{LOC.IN see-PERF.F.SG}\]
‘His/her car, Ram saw in the market.’
\[(\text{adapted from Bögel and Butt, 2013, p. 301)}\]

b. \[\text{ga}=\text{r}i \quad \text{ram}=\text{ne} \quad \text{us}=\text{ki}\]
\[\text{car.F.SG.NOM Ram.M.SG} = \text{ERG PRON.3.SG.OBL} = \text{GEN.F.SG}\]
\[\text{bazar}=\text{m}ě \quad \text{dek}^b-\text{i}\]
\[\text{market.M.SG} = \text{LOC.IN see-PERF.F.SG}\]
‘His/her car, Ram saw in the market.’
\[(\text{adapted from Bögel and Butt, 2013, p. 301)}\]

(114) a. \[\text{kis}=\text{ki} \quad \text{kitab} \quad \text{tum}=\text{ne} \quad \text{xarid}-\text{i}^?\]
\[\text{who.SG.OBL} = \text{GEN.F.SG book.F.SG.NOM you} = \text{ERG buy-PERF.F.SG}\]
‘Whose book did you buy?’
\[(\text{adapted from Bögel and Butt, 2013, p. 301)}\]

b. \[\text{kitab} \quad \text{tum}=\text{ne} \quad \text{kis}=\text{ki} \quad \text{xarid}-\text{i}^?\]
\[\text{book.F.SG.NOM you} = \text{ERG who.SG.OBL} = \text{GEN.F.SG buy-PERF.F.SG}\]
‘Whose book did you buy?’
\[(\text{Bögel and Butt, 2013, p. 301)}\]

Recall that the order within NPs is genitive \[\text{KP} \] first, then possessum (i.e., head noun) (§3.5.1). As seen in (113–114), however, when genitives are scrambled outside of their NP, this canonical order is not necessarily preserved. Using the terminology of Fanselow and Féry (2006), I am going to refer to scrambled genitives that occur before their heads in the sentence as \textit{non-inverted} scrambled genitives, and to scrambled genitives that occur after their heads as \textit{inverted} scrambled genitives.\footnote{Fanselow and Féry (2006) is a cross-linguistic survey of the morphosyntactic and prosodic features of discontinuous NPs, which is an umbrella term for different types of displacements of otherwise \textit{NP}-internal elements.}

I contend that scrambling of the genitive modifier is possible since the genitive displays rich morphology which agrees with its head, which enables speakers to identify...
the nominal in the sentence modified by the genitive. Fanselow and Féry (2006) identify agreement inside NPs as a main factor influencing the availability of discontinuous NPs across languages, but there are also counter-examples against this generalization; Turkish, for example, has discontinuous NPs, in spite of the absence of agreement inside nominal projections (see §3.6.3 for a comparison of the Hindi/Urdu data to Turkish).

3.6.2 PREFERENCES AND CONSTRAINTS

The operation of genitive scrambling does not occur without constraints. This section sums up these constraints, which serve as the empirical background for my implementation of genitive scrambling in §4.5. The constraints relate to case, scrambling from complement clauses and adjuncts, as well as scrambling from more deeply embedded GFS. In addition, there is a clear preference regarding local attachments where multiple configurations are possible.

3.6.2.1 LOCAL ATTACHMENTS ARE PREFERRED

Consider (115a), which involves a topicalized object. The possessor of that object can be scrambled to the right as in (115b). In cases such as (115b), us = ki is either a scrambled genitive modifying gari ‘car’ or a canonical genitive locally attached to bag ‘park’; the agreement morphology does not rule out either. Where the agreement morphology permits both scrambled as well as locally attached genitives, local attachments turn out to be highly preferred. Here, informants judge us = ki ‘his/her’ as modifying bag ‘park’, but acknowledge that it may also modify gari ‘car’.

(115)  a. us = ki  
       gari  
       ram = ne  
       PRON.3.SG.OBL = GEN.F.SG car.F.SG.NOM Ram.M.SG = ERG 
       bag = mé  
       dekʰ-i  
       park.F.SG = LOC.IN see-PERF.F.SG  
       ‘Her/his car, Ram saw in the park.’  

       b. gari  
       ram = ne  
       us = ki  
       car.F.SG.NOM Ram.M.SG = ERG PRON.3.SG.OBL = GEN.F.SG 
       bag = mé  
       dekʰ-i  
       park.F.SG = LOC.IN see-PERF.F.SG  
       ‘The car, Ram saw in her park.’  
       preferred over  
       ‘His/her car, Ram saw in the park.’

The preference for local attachment is a principle well-known from cognitive science, first discussed by Kimball (1973) as the “Right Association” principle, and reformulated by Gibson (1991) as the “Recency Preference”.

The preference for local attachment is a principle well-known from cognitive science, first discussed by Kimball (1973) as the “Right Association” principle, and reformulated by Gibson (1991) as the “Recency Preference”.

The preference for local attachment is a principle well-known from cognitive science, first discussed by Kimball (1973) as the “Right Association” principle, and reformulated by Gibson (1991) as the “Recency Preference”.

The preference for local attachment is a principle well-known from cognitive science, first discussed by Kimball (1973) as the “Right Association” principle, and reformulated by Gibson (1991) as the “Recency Preference”.

The preference for local attachment is a principle well-known from cognitive science, first discussed by Kimball (1973) as the “Right Association” principle, and reformulated by Gibson (1991) as the “Recency Preference”.

The preference for local attachment is a principle well-known from cognitive science, first discussed by Kimball (1973) as the “Right Association” principle, and reformulated by Gibson (1991) as the “Recency Preference”.

The preference for local attachment is a principle well-known from cognitive science, first discussed by Kimball (1973) as the “Right Association” principle, and reformulated by Gibson (1991) as the “Recency Preference”.

The preference for local attachment is a principle well-known from cognitive science, first discussed by Kimball (1973) as the “Right Association” principle, and reformulated by Gibson (1991) as the “Recency Preference”. 
3.6. GENITIVE SCRAMBLING

3.6.2.2 SCRAMBLING AND CASE

The examples above involve genitives that are scrambled out of bare NPs. Genitives may also be scrambled out of NPs that are overtly case-marked; in this case, inverted scrambled genitives are ungrammatical, and the genitive has to precede its head in the clause. Examples are shown in (116). In both sentences, \textit{ram} = \textit{ke} ‘Ram’s’ modifies \textit{baccō} = \textit{ne} ‘children = ERG’, but since the latter is ergative-marked, the former has to precede it.

(116)

a. \text{ram} = \text{ke} \quad \text{kal} \quad \text{baccō} = \text{ne} \quad \text{yih}  \\
\text{Ram.M.SG} = \text{GEN.M.SG.OBL} \quad \text{yesterday} \quad \text{child.M.PL.OBL} = \text{ERG} \quad \text{this}  \\
\text{gana} \quad \text{ga-ya} \quad \text{tʰ-a}  \\
\text{song.M.SG.NOM} \quad \text{sing-PERF.M.SG} \quad \text{be.PAST-M.SG}  \\
‘Ram’s children sang this song yesterday.’

b. * \text{baccō} = \text{ne} \quad \text{kal} \quad \text{ram} = \text{ke} \quad \text{yih}  \\
\text{child.M.PL.OBL} = \text{ERG} \quad \text{yesterday} \quad \text{Ram.M.SG} = \text{GEN.M.SG.OBL} \quad \text{this}  \\
\text{gana} \quad \text{ga-ya} \quad \text{tʰ-a}  \\
\text{song.M.SG.NOM} \quad \text{sing-PERF.M.SG} \quad \text{be.PAST-M.SG}  \\
‘The children saw Ram’s dog yesterday.’

A similar example involving a genitive scrambled from an overtly marked object NP is given in (117): \text{ram} = \text{ke} ‘Ram’s’ needs to precede its head \text{kütte} = \text{ko} ‘dog = ACC’.

(117)

a. \text{baccō} = \text{ne} \quad \text{ram} = \text{ke} \quad \text{kal}  \\
\text{child.M.PL.OBL} = \text{ERG} \quad \text{Ram.M.SG} = \text{GEN.M.SG.OBL} \quad \text{yesterday}  \\
\text{kütte} = \text{ko} \quad \text{dekʰ-a}  \\
\text{dog.M.SG.OBL} = \text{ACC} \quad \text{see-PERF.M.SG}  \\
‘The children saw Ram’s dog yesterday.’

b. * \text{baccō} = \text{ne} \quad \text{kütte} = \text{ko} \quad \text{kal}  \\
\text{child.M.PL.OBL} = \text{ERG} \quad \text{dog.M.SG.OBL} = \text{ACC} \quad \text{yesterday}  \\
\text{ram} = \text{ke} \quad \text{dekʰ-a}  \\
\text{Ram.M.SG} = \text{GEN.M.SG.OBL} \quad \text{see-PERF.M.SG}  \\
Recall that genitive KPs modifying nominals in overtly case-marked KPs need to have oblique nominal morphology. One might assume, then, that examples such as (117b) are bad simply because there are several options for the genitive KP to modify a nominal, given the high amount of syncretism in genitive case marking for the oblique; e.g., in (117b) the genitive could modify both \text{baccō} and \text{kütte}. (118) shows that this cannot be the issue. Here, the genitive can modify both nominals, being in linear precedence to both of them; cf. also (116b), which is ungrammatical, even though the agreement morphology clearly rules out any other possibilities of modification aside of \text{baccō}.

(118)

\text{ram} = \text{ke} \quad \text{kal} \quad \text{baccō} = \text{ne}  \\
\text{Ram.M.SG} = \text{GEN.M.SG.OBL} \quad \text{yesterday} \quad \text{child.M.PL.OBL} = \text{ERG}  \\
\text{kütte} = \text{ko} \quad \text{dekʰ-a}  \\
\text{dog.M.SG.OBL} = \text{ACC} \quad \text{see-PERF.M.SG}  \\
‘The children saw Ram’s dog yesterday.’

or

‘Ram’s children saw the dog yesterday.’
I take the case constraint to be an idiosyncratic requirement of Hindi/Urdu nominal arguments; clearly, more research in the domain of discontinuous NPs/argument scrambling is needed in order to arrive at a clear explanation of this particular pattern. As will be seen in §5.4.2.2, locative and instrumental nominal arguments obey the same constraint.

3.6.2.3 SCRAMBLING FROM COMPLEMENT CLAUSES

Another constraint concerns complement clauses. None of my informants judge possessors scrambled out of finite complement clauses as grammatical; cf. the ungrammatical examples in (119). However, a majority of my informants indicate that it is grammatical to scramble genitive phrases from within non-finite complement clauses (e.g., the clause headed by the modal verb in (120)). This is in line with the findings by Mahajan (1990) and Kidwai (1999, 2000), who state that scrambling of arguments from within finite complement clauses is generally not accepted, whereas scrambling from infinite complement clauses is.

(119) * us=ki ram=ne kʰ-a kʰ
PRON.3.SG.OBL = GEN.F.SG Ram.M.SG = ERG say-PERF.M.SG that
[nina=ne gᵃɾⁱ dekʰ-i] Nina.F.SG = ERG car.F.SG.NOM see-PERF.F.SG

(120) us=ki ram gᵃɾⁱ dekʰ sak-a
PRON.3.SG.OBL = GEN.F.SG Ram.M.SG.NOM car.F.SG.NOM see can-PERF.M.SG
‘His/her car, Ram could see.’

3.6.2.4 NO SCRAMBLING OUT OF ADJUNCTS

The third constraint concerning genitive scrambling is that genitive KPs may not be scrambled from within adjuncts. In (121a), meri ‘my’ is an extrinsic possessor modifying bag ‘park’, which itself is locative case-marked and an adjunct to the overall clause. It is found that the possessor may not move from its constituent to any other position in the clause (121b–c).

(121) a. ram=ne meri bag=mᵉ haṭʰi
dekʰ-a see-PERF.F.SG
‘Ram saw an elephant in my park.’

b. * meri ram=ne bag=mᵉ
PRON.Poss.F.SG Ram.M.SG = ERG park.F.SG = LOC.IN
haṭʰi dekʰ-a
elephant.M.SG.NOM see-PERF.F.SG
3.6. GENITIVE SCRAMBLING

c. * ram = ne
   bag = mē
   hatʰi
   ram = ne
   bag = mē
   hatʰi
   Ram.M.SG = ERG
   park.F.SG = LOC.IN
   elephant.M.SG = NOM
   meri
dekʰ-a
   PRON.POSS.F.SG = see-PERF.F.SG

Island behavior, i.e., the unavailability of constituents for movement/scrambling, is symptomatic for clausal adjuncts and is well-known throughout the literature, first discussed by Ross (1967). It is also a well-known diagnostic for distinguishing arguments from adjuncts, as discussed by, e.g., Needham and Toivonen (2011) in an LFG setting.

3.6.2.5 NO SCRAMBLING FROM DEEP WITHIN

All the examples so far indicate that it is possible to scramble genitive phrases out of grammatical functions (except for adjuncts and finite complement clauses) throughout the clause. The last constraint to be discussed here indicates that it is not possible to scramble genitive phrases that are selected by nominals further down a path of grammatical functions. Consider the examples in (122). ḋơhɑr ‘husband’ is a relational noun taking a SUBJ orɑt = ke ‘the woman’s’. ḋơhɑr = ki, in turn, is an extrinsic possessor SUBJ modifying the overall object of the clause, ɡɑrɪ ‘car’. The structure is as indicated by the bracketing in (122b). In the similar example (123), sorx rɑng = ke ‘of red color’ is an ADJUNCT modifying mɑkan ‘house’.

(122) a. ram = ne
   orɑt = ke
   ḋơhɑr = ki
   Ram.M.SG = ERG
   woman.F.SG = GEN.M.SG.OBL
   husband.M.SG = GEN.F.SG
   car.F.SG = NOM
   see-PERF.F.SG
   ‘Ram saw the woman’s husband’s car.’

b. ram = ne
   [[orɑt = ke]SUBJ
   ḋơhɑr = ki]SUBJ
   ɡɑrɪ OBJ
   dekʰ-a

(123) a. nina = ne
   sorx rɑng = ke
   mɑkan = ka
   Nina.F.SG = ERG
   red
   color.M.SG = GEN.M.SG
   house.M.SG = GEN.M.SG
   door.M.SG = see-PERF.M.SG
   ‘Nina saw the red house’s door.’

b. nina = ne
   [[sorx rɑng = ke]ADJUNCT
   mɑkan = ka]SUBJ
   ɡɑrɪ OBJ
   dekʰ-a

Given such situations, consider the examples in (124–125). In (124a–b), for example, orɑt = ke ‘the woman’s’, the SUBJ genitive KP modifying ḋơhɑr ‘husband’, cannot appear outside of the NP it is embedded in, i.e., outside the NP headed by ɡɑrɪ ‘car’, since it is embedded to far down in that NP, its GF path being (starting from the main clause) (↑ OBJ
SUBJ SUBJ). (125a–b) shows that the same restriction holds for attributive genitives such as sorx rɑng = ke ‘of red color’, which has the path (↑ OBJ SUBJ ADJUNCT) here.
3.6. GENITIVE SCRAMBLING

(124) a. * orɑt = ke ram = ne ḥɔhɑr = ki
    woman.F.SG = GEN.M.SG.OBL Ram.M.SG = ERG husband.M.SG = GEN.F.SG
gari dekʰ-i car.F.SG.NOM see-PERF.F.SG

b. * ram = ne ḥɔhɑr = ki gari
    Ram.M.SG = ERG husband.M.SG = GEN.F.SG car.F.SG.NOM
orɑt = ke dekʰ-i
    woman.F.SG = GEN.M.SG.OBL see-PERF.F.SG

gɑṛi car. f.sg.ном
dekʰ-i see-perf. f.sg

(125) a. * surx rang = ke nina = ne makan = ka
dɑrvaza dekʰ-a
door.M.SG see-PERF.M.SG

b. * nina = ne makan = ka dɑrvaza surx
    rang = ke dekʰ-a
color.M.SG = GEN.M.SG see-PERF.M.SG

Since my informants were not able to come up with a reasonable example of multiple 
excomps stacked on top of each other (e.g., modal verbs stacked in a fashion similar 
to English examples such as I need you to want to go to the theatre with me.), I am not 
including this possibility in my account of Hindi/Urdu genitive scrambling. In fact, the 
translations my informants were able to come up with all involve combinations of non-
finite and finite complement clauses — but scrambling out of the latter is not licit to 
begin with.

3.6.3 A COMPARISON TO TURKISH

The Hindi/Urdu data from genitive scrambling is similar to the Turkish data discussed 
by e.g., Kornfilt (2003). I show some examples from Kornfilt (2003) in (126). (126a) is 
the canonical word order, while in (126b) the possessor has been scrambled to the right.
Kornfilt assumes that there is an empty pro category in the possessor in-situ position 
which is licensed by the head of the possessive DP. Her motivation for assuming an empty 
pro is that possessive phrases exhibit overt agreement on their heads with the possessor, 
and that pro shows the typical binding properties known from other pronouns in Turkish. 
Kornfilt (2003) reanalyzes pro as a resumptive pronoun bound by the possessor, and 
argues that what is found here is not an instance of scrambling proper, but rightward 
dislocation.

(126) a. [Ahmed-in karr-sin]-i tani-mi-yor-um
    Ahmed-GEN wife-3SG-ACC know-NEG-PRES.PROG-1SG
    ‘I don’t know Ahmed’s wife.’ (Kornfilt, 2003, p. 131)

b. [pro, karr-sin]-i tani-mi-yor-um Ahmed-in,
    PRO wife-3SG-ACC know-NEG-PRES.PROG-1SG Ahmed-GEN
    (Kornfilt, 2003, p. 131)
In Hindi/Urdu, the reflexive \textit{apna} ‘self’ can be bound by a genitive phrase scrambled out of its NP, as in the examples in (127), where (127a) is the canonical word order, and (127b) as well as (127c) involve a scrambled genitive. Given the rich morphological agreement in the genitive, it is clear that \textit{ram=ka} throughout (127) modifies \textit{makan}.

\begin{itemize}
  \item (127a) \textit{nina=ne} \textit{ram=ka \textit{apna} \textit{makan}} N
\textit{k\text{al x\text{arid-a}} yesterday buy-PERF.M.SG}
\text{‘Nina bought Ram’s own house yesterday.’}
  \item (127b) \textit{ram=ka} \textit{nina=ne \textit{apna} \textit{makan}} 
\textit{k\text{al x\text{arid-a}} yesterday buy-PERF.M.SG}
\text{‘Nina bought Ram’s own house yesterday.’}
  \item (127c) \textit{nina=ne} \textit{\textit{apna} \textit{makan} k\text{al}}
\textit{ram=ka x\text{arid-a yesterday buy-PERF.M.SG}}
\text{‘Nina bought Ram’s own house yesterday.’}
\end{itemize}

In §3.8, I will provide an explanation of the observed reflexive binding patterns in Hindi/Urdu nominals. What is crucial here is that LFG does not need to posit empty categories to explain binding patterns because LFG’s binding theory is largely formulated over f-structure, not c-structure constraints (Dalrymple, 1993, 2001). In fact, the principle of “economy of expression” (Bresnan, 2001, Ch. 6) forbids the positing of empty categories, if generalizations concerning binding, scrambling and other phenomena can be traced back to explicit functional information on categories, e.g., agreement features or case; using functional uncertainty in combination with agreement, the scrambled genitive can, at f-structure, reliably be embedded under its head noun. In particular, there is no need for an analysis in terms of rightward dislocation involving an in-situ resumptive pronoun, for which there is otherwise no discernible morphosyntactic evidence; in an LFG setting, the examples above can naturally be treated as instances of scrambling.

According to Kornfilt (2003), Turkish possessors further obey a constraint, called the “barrier constraint” by Chomsky (1986), which rules out subconstituents right-adjointed to arguments; thus, (128b) is ungrammatical in Turkish. (128a) shows the canonical word order in Turkish.

\footnote{Additional binding constraints which make recourse to f-precedence (which, in turn, refers to linear c-precedence) have been formulated for pronominal binding in languages such as Malayalam (Mohanan, 1982), Japanese (Kameyama, 1984) as well as Korean (Choi, 1999b); see the discussions in (Dalrymple, 2001, p. 288–289) and (Bresnan, 2001, p. 193–203).}
3.6. GENITIVE SCRAMBLING

The constraint does not exist in Hindi/Urdu. Here, a genitive kp may occur next to its head. In other words, a genitive phrase may be scrambled throughout the sentence, including directly to the right of its head. Examples are in (129–131); here, the a. examples show the canonical position of the genitives to the left of their heads, while the b. examples are grammatical cases of scrambled possessors to the right of their heads.

Recall from §3.5.1.2, examples (53–57), that the genitive kps cannot right-adjoin to their heads; this means that the genitives in (129–131) need to be analyzed as scrambled genitives that are sisters to their heads at a higher level, i.e., the clause.

3.6.4 SUMMARY

I conclude that genitive scrambling is possible in Hindi/Urdu. Genitive modifiers of main clause gfs may freely be scrambled throughout a sentence. However, local attachments of modifiers are generally preferred over non-local (scrambled) modification, in cases where the morphosyntax permits both readings. Genitive modifiers of gfs embedded in non-finite complement clauses may scramble across clause boundaries, while genitive
modifiers of GFs embedded in finite complement clauses may not do so. It was seen that genitive KPs scrambled to the right of their heads are acceptable, unlike what is known from Turkish, and that scrambling out of adjuncts as well as scrambling out of GFs embedded further down a GF path is not felicitous in Hindi/Urdu.

Possessors are not the only NP modifiers that may be scrambled in Hindi/Urdu. Although their main focus is on NP-internal discontinuities, Raza and Ahmed (2011a) also mention NPs that are discontinuous at the clause-level, i.e., NPs whose modifiers have been scrambled outside of the NP. Examples are given below in (132). (132b) is an example of quantifier float in Hindi/Urdu; here, the head noun am ‘mangos’ is scrambled to the beginning of the clause, while its quantifier stays in situ.\(^{35}\) (132a) shows the canonical word order. The phenomenon is well-known also from other languages including English ((133), see, e.g., Payne, 2010) and German ((134), see, e.g., Vater, 1980).

(132) a. ali=ne bahut am
‘Ali ate many mangos.’ (Raza and Ahmed, 2011a)

b. am ali=ne bahut kʰa-e
‘Ali ate many mangos.’ (Raza and Ahmed, 2011a)

(133) a. All the children have eaten mangos.
b. The children have all eaten mangos.

(134) a. Alle Kinder haben Mangos gegessen.
all.NOM child.N.PL.NOM have.PRES.3.PL mango.F.PL.ACC eat.PART.PERF
‘The children have all eaten mangos.’

b. Die Kinder haben alle Mangos
de.N.PL.NOM child.N.PL.NOM have.PRES.3.PL all mango.F.PL.ACC
gegessen.
eat.PART.PERF
‘The children have all eaten mangos.’

(135) is an example of relative clause extraction in Hindi/Urdu. Noted before by Dayal (1994), it is possible to scramble a relative clause outside of its NP. Again, the process of extraposing a relative clause is also well-known from English (e.g., (136), Culicover and Rochemann, 1990, Ross, 1967) as well as German (e.g., (137), Kathol, 2000, Keller, 1995).

(135) a. ek jaks jis = ki darʰi
one person.M.SG.NOM who.SG.OBL = GEN.F.SG beard.F.SG.NOM
tʰ-i ay-a
be.PAST-F.SG come-PERF.M.SG
‘One person, who had a beard, came.’ (Raza and Ahmed, 2011a)

\(^{35}\)See e.g. Fanselow and Féry (2006) for some more examples of quantifier float in Hindi/Urdu.
b. ek जान्स अय-अ जिस = कि 
one person.M.SG.NOM come-PERF.M.SG who.SG.OBL = GEN.F.SG 
dाऱि तः-ि 
beard.F.SG.NOM be.PAST-F.SG 
‘One person, who had a beard, came.’ (Raza and Ahmed, 2011a)

(136) a. The children, who were sad, cried. 
b. The children cried, who were sad.

(137) a. Die Kinder, die traurig waren, weinten. 
the.N.PL.NOM child.N.PL.NOM which sad be.PAST.3.PL cry.PAST.3.PL 
‘The children, who were sad, cried.’ 
b. Die Kinder weinten, die traurig waren. 
the.N.PL.NOM child.N.PL.NOM cry.PAST.3.PL which sad be.PAST.3.PL 
‘The children cried, who were sad.’

Thus, seeing how other NP modifiers such as quantifiers as well as relative clauses may 
be scrambled out of the NPs they modify, it is not surprising that possessors may also 
occur outside the NP.

Most analyses in Chomskyan frameworks assume scrambled constituents to be derived 
in terms of movement or copying; see Karimi (2003), Corver and van Riemsdijk (1994) 
among many others for such approaches. In LFG, scrambled constituents are usually 
dealt with using base generation of the phrases in scrambled position combined with 
functional uncertainty annotations to embed the phrases in the correct f-structure (Ka- 
plan and Zaenen, 1995). In §4.5, I will discuss an LFG/XLE solution that correctly models 
the generalizations for Hindi/Urdu.

3.7 THE GENITIVE IN POSSESSIVE CLAUSES

As Hindi/Urdu lacks a full verb to express possession (such as, for example, English 
*have*), the language uses the verb *ho* ‘be’ in combination with the genitive case, which 
marks the possessor, and a second nominal (the possessum). I refer to such sentences as 
*possessive clauses*. Examples are in (138).

(138) a. nina=का मकःन हें 
Nina.F.SG = GEN.M.SG house.M.SG.NOM be.PRES.3.SG 
‘Of Nina is a house.’ = ‘Nina has a house.’ 
b. ram=की चाँदिकी हें 
‘Of Ram is a watch of silver.’ = ‘Ram has a silver watch.’

---

36Two English translations are provided here, a literal one involving the verb *be*, as well as one which 
sounds more natural in English using *have*. I will only give the natural translation in the examples below.
3.7. THE GENITIVE IN POSSESSIVE CLAUSES

3.7.1 Constituent Properties

This section discusses the constituent properties of the genitive in possessive clauses. It is shown that the facts concerning coordination as well as linear order mirror the facts for the NP-internal genitives. Thus, constituent properties alone do not motivate a treatment where the genitive KP is seen as an independent node in the tree; in fact, these properties are in support of treating possessive clauses as intransitives with only a single complex NP.

3.7.1.1 Coordination

The facts concerning coordination in possessive clauses mirror the facts from coordination of the genitive in noun phrases (as discussed in §3.5.1.1). Reconsider the NP example in (139), repeated from (46). The very same coordination pattern may be applied to the possessive clauses as can be seen in (140), resulting in the same meaning: while the reading of (140a) is unambiguously one where the cars belong to the boy and the girl simultaneously, (140b) either means that a set of cars belong to both at the same time, or that the boy and girl have two (not necessarily identical) sets of cars.
3.7. THE GENITIVE IN POSSESSIVE CLAUSES

(139) a. [[[lɑrke]_{np} or [lɑrkì]_{np} = ki]_{kp} garì]_{np}  
   boy.M.SG.OBL and girl.F.SG = GEN.F.PL car.F.PL  
   ‘the boy’s and the girl’s cars’

b. [[[lɑrke]_{np} = ki]_{kp} or [lɑrkì]_{np} = ki]_{kp} garì]_{np}  
   ‘the boy’s and the girl’s cars’

(140) a. [[[lɑrke]_{np} or [lɑrkì]_{np} = ki]_{kp} garì]_{np}  
   ‘The boy and girl have cars.’

b. [[[lɑrke]_{np} = ki]_{kp} or [lɑrkì]_{np} = ki]_{kp} garì]_{np}  
   boy.M.SG.OBL = GEN.F.PL and girl.F.SG = GEN.F.PL car.F.PL.NOM  
   be.PRES.3.PL  
   ‘The boy and the girl have cars.’

Thus, judging from how coordination patterns apply in the possessive clauses, these are simply complex NPs containing genitive modifiers combined with the verb ho ‘be’.

3.7.1.2 LINEAR ORDER

In §3.5.1.2, it was shown that several elements may intervene between the possessor phrase and the possessum inside the NP. In possessive clauses, the facts are entirely the same. Notice that the examples in (141–145) parallel the NPs in (47–51), except for the verb ho ‘be’.

(141) a. [ram=ka bɑrə mɑkan]_{np} hɛ  
   ‘Ram has big house.’

b. [bɑrə ram=ka mɑkan]_{np} hɛ  
   ‘Ram has a big house.’

(142) a. [ram=ki nili garì]_{np} hɛ  
   Ram.M.SG = GEN.F.SG blue.F.SG car.F.SG be.PRES.3.SG  
   ‘Ram has a blue car.’

b. [nili ram=ki garì]_{np} hɛ  
   blue.F.SG Ram.M.SG = GEN.F.SG car.F.SG be.PRES.3.SG  
   ‘Ram has a blue car.’

(143) a. [nadya=ke do beṭe]_{np} hɛ  
   Nadya.F.SG = GEN.M.PL two son.M.PL be.PRES.3.SG  
   ‘Nadya has two sons.’

b. [do nadya=ke beṭe]_{np} hɛ  
   two Nadya.F.SG = GEN.M.PL son.M.PL be.PRES.3.SG  
   ‘Nadya has two sons.’
3.7. THE GENITIVE IN POSSESSIVE CLAUSES

(144) kamai = ka əcc^a əd^a ən hɛ
earning=F.SG = GEN.M.SG əd^a ən ease.M.SG measure.M.SG.NOM be.PRES.3.SG
‘The earning has an acceptable size.’

(145) do əs = ke əbɛ ən hɛ
two PRON.3.SG.OBL = GEN.M.SG əd^a ən,hɛ
‘He/she has two children.’

The range of linear orders of elements in the possessive clause thus parallels the range that was observed inside NPs. Again, this constitutes evidence that the constituency in these sentences involves complex NPs with possessors inside them.

Variation in the word order of possessive clauses is possible. While the general pattern is as in (138) (i.e., possessor — possessum — ho ‘be’), examples such as (146–147) can also be found, where the possessum occurs before the possessor.

(146) jivan əs = ka ən hɛ
vitality,M.SG.NOM PRON.3.SG.OBL = GEN.M.SG be.PRES.3.SG
‘Vitality is of him/her.’

(147) yeh makaŋ əs əstri = ka ən hɛ
this house,M.SG.NOM PRON.3.SG.OBL woman,F.SG = GEN.M.SG be.PRES.3.SG
‘This house is of that woman.’ = ‘This house is that woman’s.’

As was seen in §3.6, scrambling of the possessor is possible with verbs other than ho ‘be’; a relevant example, repeated from (109), is given below in (148).

(148) a. ram = ka əbɛ əy-a
Ram,M.SG = GEN.M.SG friend.M.SG.NOM come-PERF.M.SG
‘Ram’s friend came.’

b. əbɛ əy-a ram = ka
friend,M.SG.NOM come-PERF.M.SG Ram,M.SG = GEN.M.SG
‘Ram’s friend came.’

With respect to the linear order of constituents, it can thus be seen that the possessors in the possessive clauses behave like other possessives inside NPs in that they may be removed from their respective nominal heads. I will return to the differing word orders in §3.7.3.

3.7.2 FUNCTIONAL PROPERTIES

This section reviews the subject tests introduced in §3.4 with respect to the genitive in possessive clauses. It turns out that two of the subject tests do not clearly indicate the status of the genitive in these clauses; the control test, however, clearly shows that the genitive phrase cannot be a clausal SUBJ.

3.7. THE GENITIVE IN POSSESSIVE CLAUSES

3.7.2.1 Reflexive Binding

To test for the clausal subjecthood of the genitive KPs in possessive clauses, the *apna* test as seen in §3.4.1 is not sufficient. To see this, note the following considerations. My claim in §4 will be that possessive clauses are intransitives with a single SUBJ, where the genitive KP is embedded in the SUBJ, and an intransitive verb *ho* ‘be’. Now consider the sentence in (149a). Since the genitive KP binds the reflexive adjective *apna* ‘self’, one could take such examples as evidence that the genitive KP must be the SUBJ of the clause, while the other NP must receive some other GF. This analysis is depicted by the bracketing in (149b).

(149) a. ram =ka apna ek bʰi beṭa nahī tʰ-a
Ram.M.SG = GEN.M.SG self.M.SG one FOC son.M.SG.NOM not be.PAST-M.SG
‘Ram, didn’t have even one son of his own.’

b. [ram =ka]SUBJ [apna ek bʰi beṭa]GF nahī tʰ-a

However, if one considers the alternative bracketing as indicated in (150a), it becomes clear why the *apna* test cannot be used to show the SUBJ status of the relevant KP. *apna* ‘self’ is an adjective that occurs within NPs. Since the Hindi/Urdu NP constituent structure may be fairly complex (as seen in §3.5.1), the entire string bracketed in (150a) constitutes a valid NP; see the example in (150b).

(150) a. [ram =ka apna ek bʰi beṭa]SUBJ nahī tʰ-a

b. ram =ka apna ek bʰi beṭa
‘Ram, ’s only one own son’

Assuming that (150a) is a valid bracketing, *apna* ‘self’ may bind a local SUBJ inside the NP *apna* occurs in, exactly in the way that was shown in §3.4.1. Thus, the *apna* test alone as defined in (35) is not a sufficient indicator for the clausal subjecthood of the genitive KP.

3.7.2.2 Pronominal Coreference

The subject test concerning pronominal coreference is also problematic. Recall that pronouns cannot be bound by an f-commanding SUBJ of the minimal finite domain that contains the pronoun. The minimal finite domain in LFG is defined as the smallest f-structure with an attribute TENSE. Now, for the sake of argument, assume that the genitive KP *ram =ki* in (151a) occupies the clausal SUBJ function. That SUBJ would then f-command the f-structure of the other GF in a minimal finite domain. This analysis is sketched by the bracketing in (151b) as well as by the f-structure in Figure 3.10a. The alternative analysis, where the genitive KP is an embedded SUBJ inside a larger clausal SUBJ, is shown in the bracketing in (151c) as well as by the f-structure in Figure 3.10b.
3.7. THE GENITIVE IN POSSESSIVE CLAUSES

(151)  

a. ram = ki hasin ankʰ hẽ  
   ‘Ram has beautiful eyes.’

b. [ram = ki]SUBJ [hasin ankʰ]GF hẽ

c. [ram = ki hasin ankʰ]SUBJ hẽ

Figure 3.10: F-command relations in possessive clauses

When an additional genitive phrase with the pronoun us is added, as in (152a), the majority of my informants state that us cannot be an antecedent of ram. The only person who states otherwise also mentions that it would be more natural to insert apna ‘self’ for referring back to ram. The fact that ram cannot be an antecedent of us is again compatible with both analyses, shown in (152b) and (152c) and their f-structures in Figure 3.11 and Figure 3.12, respectively. This is because under both analyses, the antecedent of the pronoun is a SUBJ that f-commands the pronoun in the minimal finite domain containing the pronoun. Therefore, the test from pronominal coreference does not rule out either analysis.

(152)  

a. ram = ki us = ke  
   Ram.M.SG = GEN.F.PL PRON.3.SG.OBL = GEN.M.SG.OBL  
   makan = ke rang = ki hasin ankʰ  
   house.M.SG = GEN.M.SG.OBL color.F.PL = GEN.M.SG beautiful.F.PL  
   eye.F.PL.NOM be.PRES.3.PL  
   ‘Ram has beautiful eyes of the color of his house.’

b. [ram = ki]SUBJ [us = ke makan = ke rang = ki hasin ankʰ]GF hẽ

c. [ram = ki us = ke makan = ke rang = ki hasin ankʰ]SUBJ hẽ

3.7.2.3 CONTROL

Finally, the diagnostic regarding control of a participial adjunct clause is clear indication that the genitive KPs in possessive clauses cannot be clausal SUBJS. The sentence in (153a) is a possessive clause. In (153b), the genitive KP cannot serve as the controller of
the participial adjunct clause. By (45), the genitive KP can then not be the grammatical SUBJ of the clause.

(153)  
a. sri krïʃn=ka makan he  
      Sri Krishna.M.SG = GEN.M.SG house.M.SG.NOM be.PRES.3.SG  
      “Sri Krishna, has a house.’

   b. * sri krïʃn=ka makan amir ho-te  
      Sri Krishna.M.SG = GEN.M.SG house.M.SG.NOM rich be-IMPF.M.SG.OBL  
      hu-e he  
      be.PERF-M.SG.OBL be.PRES.3.SG  
      ‘Sri Krishna, has a house ___vi being rich.’
3.7. THE GENITIVE IN POSSESSIVE CLAUSES

I conclude that testing for the subjecthood of genitive KPs in possessive clauses is not straightforward. In particular, distinguishing a clausal SUBJ reading from a reading where the genitive is selected as a SUBJ inside a more complex NP is not easy, and it was seen that two of the three subjecthood tests identified before fail to clearly indicate either of the two possibilities. The test from control, however, has to be regarded as motivation for assuming a complex NP reading where the genitive stays in its nominal domain. In absence of further diagnostics that would account for an alternative analysis, I will assume this particular analysis.

3.7.3 INFORMATION STRUCTURE

So far, I have ignored an important aspect of the analysis of possessive clauses on purpose, namely the part of the analysis that extends into the dimension of information structure. I do not attempt an actual in-depth analysis of the information structure of genitives/possessive clauses in the LFG framework; I confine myself to describing the c-structural as well as f-structural properties of the patterns. I will, however, describe the relevant data and mention previous attempts at analyzing the information structure of Hindi/Urdu.

Examples of possessive clauses that are relevant to the discussion in this section are repeated below. As was seen in §3.7.1, the word order in possessive clauses may be different. Both of the word orders in (154) are possible; in (154a), the possessor genitive KP occurs sentence-initially, while in (154b), it occurs in immediately preverbal position. The different word orders were attributed to the syntactic operation of scrambling, i.e., the genitive KP in (154b) is scrambled out of its nominal domain exactly in the way that was already described in §3.6. What was not discussed yet is the reason why scrambling occurs in the first place.

(154)  
\begin{align*}
a. & \text{nina}=\text{ka} \qquad \text{makan} \quad \text{he} \\
& \text{Nina.F.SG} = \text{GEN.M.SG} \quad \text{house.M.SG.NOM be.PRES.3.SG} \\
& \text{‘Of Nina is a house.’} = \text{‘Nina has a house.’} \quad = (138a) \\
b. & \text{yeh makan} \quad \text{us} \quad \text{stri}=\text{ka} \quad \text{he} \\
& \text{this house.M.SG.NOM PRON.3.SG.OBL woman.F.SG} = \text{GEN.M.SG be.PRES.3.SG} \\
& \text{‘This house is of that woman.’} = \text{‘This house is that woman’s.’} \quad = (147) \\
& \quad \text{(Spencer, 2008, p. 16)}
\end{align*}

To see the different information-structural properties of the two orderings, one has to view them in discourse context. Examine the following dialogs. In (155a), a particular house is under discussion. In (155b), the house is again referred to, and the fact that it is Ram’s house is introduced. As the referent makan ‘house’ had already been introduced in the discourse, it is not felicitous to say (155c) instead of (155b), since in (155c), makan ‘dog’ occurs in the immediately preverbal position, which is cited as the position in Hindi/Urdu encoding focus (Butt and King, 1996, 2000; Kidwai, 2000; Mahajan, 1990).
Focused information has been characterized as new information (i.e., information not previously known to the hearer) by a wealth of literature on the subject of information structure (Butt and King, 1996, 2000, Choi, 1999b, Givón, 1979, King, 1995, Kiss, 1995, among others).

(155)  

a. yeh makan = mě cuha he  
this house.M.SG = LOC.IN rat.M.SG be.PRES.3.SG  
‘In this house, there is a rat.’

b. makan ram = ka he  
house.M.SG Ram.M.SG = GEN.M.SG be.PRES.3.SG  
‘The house is of Ram.’ = ‘The house is Ram’s.’

c. ?? ram = ka makan he  
Ram.M.SG = GEN.M.SG house.M.SG be.PRES.3.SG  
‘Ram has a house.’

As a second example, consider the dialog in (156). In (156a), a speaker asks a question about a discourse referent ram, i.e., about a referent already present in the discourse. To answer in a coherent way, the other speaker chooses (156b), where the information that is being asked for is in the preverbal focus position, signaling that this is new information. (156c) makes for a bad answer in this particular context since it has the new information in sentence-initial position, the position described as the topic position in Hindi/Urdu (Butt, 1995, Butt and King, 1996, 2000, Kidwai, 2000, Mahajan, 1990, Mohanan, 1994). The discourse function of topic, in turn, has been characterized cross-linguistically as encoding “old” information already known to the hearer (Büring, 1999, Butt and King, 1996, 2000, King, 1995, Kiss, 1995, among others).

(156)  

a. ram = ka kya he  
Ram.M.SG = GEN.M.SG what be.PRES.3.SG  
‘What is of Ram?’ = ‘What does Ram have?’

b. ram = ka makan he  
Ram.M.SG = GEN.M.SG house.M.SG be.PRES.3.SG  
‘Of Ram is a house.’ = ‘Ram has a house.’

c. ?? makan ram = ka he  
house.M.SG Ram.M.SG = GEN.M.SG be.PRES.3.SG  
‘The house is Ram’s.’

Thus, none of the alternatives in (155b–c) and (156b–c) can be considered ungrammatical; they are merely different ways of packaging up the same information and presenting it to the addressee. In particular, notice how the interpretation of the referents in e.g. (156) is changed pragmatically: even though there is no definite determiner in Hindi/Urdu, the sentence-initial topialized element receives a definite interpretation. This is in sharp contrast to the preverbal element, which is interpreted as an indefinite.
3.7. THE GENITIVE IN POSSESSIVE CLAUSES

The assumptions made here are in line with previous analyses regarding information-structural properties of Hindi/Urdu. Butt and King (1996) present an LFG analysis of the free word order languages Urdu and Turkish, proposing to analyze the free word order of both languages as base-generated possibilities (i.e., alternatives that are not derived by movement from a deeper representation) that directly reflect different information structures. The view by Butt and King (1996) is presented in §2.3.5. For the analysis presented below, I exclude the level of information structure and state that, while I am sympathetic to approaches that postulate additional projections for information and/or discourse structure, much research remains to be done to arrive at a more complete picture of how these levels interact with the purely syntactic description in terms of c- and f-structure.

3.7.4 A Restriction on Event Nouns

Looking at the distribution of possessive clauses, it turns out that common nouns as well as relational nouns feature in these clauses. Event nouns are generally bad in possessive clauses. The examples in (157) illustrate.

(157) a. *ṭrain=ki rɑvangi he
   train.M.SG = GEN.F.SG departure.F.SG.NOM be.PRES.3.SG

b. *ɪnsan=ki tɑbahi he
   man.M.SG = GEN.F.SG destruction.F.SG.NOM be.PRES.3.SG

(158) a. *ɪnsan=ki tɑbahi he
   man.M.SG = GEN.F.SG destruction.F.SG.NOM be.PRES.3.SG

b. *nɑojɑvanõ=ka tʰane=ka
   youngster.M.PL.OBL = GEN.M.SG police-station.M.SG.OBL = GEN.M.SG
gʰɛrao he
   circumvention.M.SG.NOM be.PRES.3.SG

If possessive clauses are treated as intransitives, serving merely to predicate the existence of the head nominal, it becomes clear why event nouns are not felicitous in this clause type: such clauses are stative and therefore lack any aspectual information, but event nominals generally require an aspectual context in order to be felicitous in intransitive clauses (Comrie, 1976, but see Sabbagh (2009) for an account of Tagalog “event” existentials). An example with an aspectual context is shown in (159); (159) is a modification of (157a).

(159) kal ṭrain=ki rɑvangi hu-i
   yesterday train.M.SG = GEN.F.SG departure.F.SG.NOM be.PERF-F.SG

   ‘Yesterday, the departure of the train happened.’ = ‘Yesterday, the train left.’

Thus, I assume that event nominal heads are generally not felicitous as subjects of possessive clauses. To implement this constraint in an LFG grammar, lexical information on the nominal head is needed that sets apart event nouns from other types of nouns.
3.8 REFLEXIVE BINDING EXPLAINED

The analysis I propose poses a problem for the original version of the subject diagnostic concerning reflexive binding, as formulated by Mohanan (1994). To see why, reconsider the diagnostic in (160), repeated from (160).8

(160) The reflexive *apna* ‘self’ must be bound by an L-SUBJ or SUBJ within its minimal finite domain (i.e., the smallest f-structure with an attribute TENSE).

As is, the diagnostic does not explain why the reflexive *apna* ‘self’ may be bound by the NP marked by the genitive; there is no TENSE feature in the NP f-structures, and the NP f-structures do not represent minimal finite domains. Therefore, I propose a change to the diagnostic. The binding domain of *apna* ‘self’ cannot be the minimal finite domain; instead, *apna* ‘self’ must be bound in what Dalrymple (1993) and Dalrymple (2001) have defined as the minimal complete nucleus, see (161).

(161) The minimal complete nucleus corresponds to the smallest f-structure $f$ where $f$ contains some predicate and a function SUBJ.

(adapted from Dalrymple, 2001, p. 283)

Thus, the revised reflexive binding diagnostic is as in (162).

(162) The reflexive *apna* ‘self’ must be bound by an L-SUBJ or SUBJ within its minimal complete nucleus (i.e., the smallest f-structure containing some predicate and a function SUBJ).

Under this version of the diagnostic, all of the binding facts can be resolved. In the nominal domain, the minimal complete nucleus is the head noun’s f-structure; here, a possessor SUBJ must bind *apna* ‘self’. At the same time, (162) accounts for the binding facts outside the nominal domain. In (163), there is no SUBJ function within the reflexive’s immediate (i.e., nominal) f-structure, and it must thus be bound by the next available SUBJ, i.e., *ravi*, which is part of the smallest available f-structure containing some predicate and a function SUBJ. Here, this is the clausal f-structure; thus, *apna* ‘self’ must be bound as shown in the f-structure in Figure 3.13.

(163) *ravi, apni, saikal = par bɛṭʰ-a*


‘Ravi sat on self’s bike.’

(Mohanann, 1994, p. 122)

---

8[160] is my own reformulation of the diagnostic that is compatible with LFG’s binding theory, and equivalent with the original diagnostic by Mohanan (1994). See §3.4.1.
The analysis as presented so far results in a significant deal of ambiguity. To see why this is the case, consider the \textit{np} in (164). Given the above account of genitive functional assignment in Hindi/Urdu, this \textit{np} yields a number of analysis. The head nominal \textit{mez} ‘table’ is a common noun that may optionally augmented to subcategorize for a \textit{subj}, realizing extrinsic possession. Other than that, the nominal may be modified by attributive genitives that are \textit{adjunct}s. Since the order of the individual genitive types is free, the above \textit{np} receives a total of four different analyses, as indicated in (164b–e)\textsuperscript{39}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.13.png}
\caption{Full clause f-structure, reflexive binding}
\end{figure}

(164) a. makan = ki laķr̥i = ki suṛx rān̥g = ki
   mez table.F.SG
   ‘the house’s table of red color made of wood’

b. [makan = ki]_{ADJUNCT} [laķr̥i = ki]_{ADJUNCT} [suṛx rān̥g = ki]_{ADJUNCT} mez

\textsuperscript{39}Here, I assume that all three genitive \textit{kp}s in the example are attached in a flat manner to the head noun, thus ignoring other possibilities of attachment, e.g., one where \textit{makan = ki} ‘the house’s’ attaches under \textit{laķr̥i} ‘wood’; this (semantically less plausible) analysis would yield a meaning where the table of red color is made of some house’s wood.
Disambiguating between (164b–e) is not a matter of syntax, but belongs in the domain of (lexical) semantics. Given the identical morphosyntactic realization of \textit{subj} and \textit{adjunct} genitives, syntax cannot know whether a house is more likely than wood to be an extrinsic possessor of a table. Likewise, it cannot know whether a house can be an extrinsic possessor of a table at all, and it cannot know all the attributes that a table might have. Knowledge like this is part of large-scale lexical ontologies such as WordNet \cite{Bhattacharyya2010, Fellbaum1998}. Nevertheless, I have shown that there are certain constraints on possessive modification that one has to factor in, and which help in disambiguating the input.

3.10 CONCLUSION

This concludes the chapter on the Hindi/Urdu genitive. It was shown that the genitive case must be treated as a case clitic in the sense of \textit{Butt and King} \cite{Butt2004b}, and that it is the only case clitic that inflects in agreement with a head noun. After discussing related work on possessives in LFG, the chapter has proceeded to a discussion of subjecthood diagnostics in Hindi/Urdu, prompted by the common treatment of possessives as \textit{subj}s licensed by nominals. §3.5 has identified three different classes of nouns, common nouns, event nouns, relational nouns as well as proper nouns and pronouns, which differ in the way they license genitives. The \textit{apna} ‘self’ test on reflexive binding proved to be appropriate for distinguishing three types of genitives: \textit{subj}s, \textit{obj}s and \textit{adjunct}s.

Concerning the linear order within the Hindi/Urdu NP, data was introduced showing that the order of NP internal constituents is free, with the restriction that the order needs to be head-final; in particular, genitive phrases modifying the head noun do not occur after the head NP-internally.

Overall, the treatment results in ambiguity where the functional interpretation of the genitive KP or the possessive pronoun cannot uniquely be determined. To that end, I have further discussed different selectional restrictions in terms of animacy and nominal type. Regarding animacy, animate nouns were shown to subcategorize only animate \textit{subj}s as possessors. Regarding nominal type, pronouns were shown to resist all types of genitive modification, while proper nouns allow only \textit{adjunct} genitives.

I have also provided a description of genitive scrambling, which was shown to affect all major GFs in a clause, but not adjuncts. The GF the genitive is scrambled from may be embedded in an XCOMP. In addition, the selectional restrictions concerning animacy and proper nouns/pronominals apply in scrambled configurations as well. It was also found
that generally, local attachments are preferred over scrambled configurations. The next chapter discusses an LFG analysis and accompanying XLE implementation of the genitive patterns.
4 LFG Analysis & XLE Implementation of the Genitive

4.1 INTRODUCTION

In this chapter I develop a novel analysis of the Hindi/Urdu genitive. I see the Hindi/Urdu genitive as a strictly nominal case; the genitive can only be assigned inside the nominal domain by nominal heads. All nominals in the language may be modified by genitives; the genitives are not tied to a particular semantic configuration or thematic role. As done by others (Butt and King, 2004b, Mohanan, 1994), I thus characterize the Hindi/Urdu genitive as an instance of structural or default case marking. The nature of the modification must be strictly lexical, though; while common count nouns are augmented with the proper subcategorization frame via an augmentation template for lexical predication (as suggested by Bresnan (2001)), relational nouns and event nouns come with fully specified subcategorization frames. The former operation realizes what Barker (1995) refers to as extrinsic possession, while the latter possibility is what Barker (1995) calls intrinsic possession. Proper nouns do not predicate and thus may only be modified by genitive modifiers that are adjuncts.

4.2 GENITIVE CASE MARKING

The Hindi/Urdu genitive is a strictly nominal case. That is, its distribution is confined to nominal f-structures only. In possessive clauses, nominals marked by the genitive case

1At first sight, the postpositions mentioned in §2.3.3 and §3.2.2.1 seem to form an exception to the generalization that the genitive is only assigned by nominals: they also require genitive complements.
do not behave as subjects; furthermore, there is not a single verb in Hindi/Urdu that requires the genitive case on any of its GFS. Thus, the genitive case is never assigned outside of nominal domain. Given the discussion of case in Hindi/Urdu in §2.3.2, the genitive is characterized as a structural (or “default”) case assigned by nominals. As will be seen in Chapter §3, there are nominals that lexically assign other types of case, but in the absence of additional requirements, the genitive is assigned for nominal arguments.

The lexical entries of the genitive case markers are given in (1). Recall the agreement pattern of the genitive case marker in Table 3.1: in XLE, constraining equations (see §2.5.3.1) can account for the requirements concerning gender, number as well as morphological form. In (1), the constraints are in the form of inside-out constraining equations (§2.5.3.4). To see how these constraints work, reconsider my proposal for the GF assignment of the genitive KP in Table 3.3. The genitive KP may either be embedded in a SUBJ, ADJUNCT or in an OBJ f-structure inside the head noun’s f-structure. The equation in the second line of (1a), for example, states that the GEN attribute of the f-structure embedding either the SUBJ, ADJUNCT or OBJ f-structure needs to bear the value fem. The fact that genitive phrases are confined to the nominal domain (i.e., either selected as arguments by a nominal, or realized as adjuncts) is implemented by defining an inside-out existential constraint (see §2.5.3.2) of the form (({SUBJ|OBJ|ADJUNCT} ~) NTYPE), which states that in the f-structure that the genitive phrase is embedded in, there needs to be an NTYPE feature. The constraint is in the template NTYPE-REQ (2), which is called from the entries in (1). The constraint (~ CASE) =c gen requires the KP headed by the clitic to be in the genitive case; the case is not assigned here, but structurally in the NP rule (see below). Note that, for readability and efficiency reasons, in (1b–c) the morphological constraints on the case marker are handled using a local variable (see §2.5.3.8), named GEN-IO. The same mechanism is used in other annotations below. The scope of local variables is local (i.e., they are limited to the annotation in which they are defined), which is why the variable definition is repeated in each annotation.

\[
\begin{align*}
(1) \quad & a. \quad kI \quad K * (\sim CASE) =c gen \\
& \quad \quad \quad \quad ((\{SUBJ\mid OBJ\mid ADJUNCT \}$ \sim) \text{GEND}) =c \text{fem} \\
& \quad \quad \quad \quad \quad \text{@NTYPE-REQ}. \\
& b. \quad kA \quad K * (\sim CASE) =c gen \\
& \quad \quad \quad \quad ((\{SUBJ\mid OBJ\mid ADJUNCT \}$ \sim) = %\text{GEN-IO} \\
& \quad \quad \quad \quad (%\text{GEN-IO NUM}) =c \text{sg} \\
& \quad \quad \quad \quad (%\text{GEN-IO GEND}) =c \text{masc} \\
& \quad \quad \quad \quad (%\text{GEN-IO CHECK _NMORPH}) =c \text{nom} \\
& \quad \quad \quad \quad \quad \text{@NTYPE-REQ}. 
\end{align*}
\]

However, the generalization can still be argued to hold for two reasons. First, as discussed in §2.3.3, some of the postpositions allow for nominal inflection, which hints at their nominal status, if only in past stages of Hindi/Urdu (cf. Masica (1991, p. 234–235)). Second, recall that native speakers do not associate the (mostly invariant) case form ke with the genitive case to begin with, but rather take it to be part of a complex postposition, which suggests that reanalysis has taken place or is taking place.
4.2. GENITIVE CASE MARKING

C. kE K * (^ CASE) =c gen
   ({SUBJ|OBJ|ADJUNCT $} ~) = %GEN-IO
   {(%GEN-IO NUM) =c sg
   (%GEN-IO CHECK _NMORPH) =c obl
   |(%GEN-IO NUM) =c pl}
   (%GEN-IO GEND) =c masc
   @NTYPE-REQ.

(2) NTYPE-REQ = ((({SUBJ|OBJ|ADJUNCT $} ~) NTYPE).

The lexical entries for the possessive pronouns mera/mer/meri ‘my’ and uska/uske/uski ‘her/his/its’ are shown in (3–4). Masculine forms of possessive pronouns call up the template in (3a), feminine forms use (3b). Recall that the feminine inflection pattern is underspecified for number and morphological form of the modified nominal, which is why the templates in (3) differ in the number of arguments they take: POSS-PRONOUN-FEM does not enforce additional agreement values for the NUM and CHECK _NMORPH features of the modified nominal. The first and second person pronouns are also annotated for animacy by calling the ParGram common template ANIM with the value “+”, while third person pronouns are underspecified for animacy, hence are not annotated for the feature. In addition, the third person pronoun forms are marked for deixis, in terms of a distinction between proximal and distal forms; this is achieved via a call to the template DEIXIS. uska/uske/uski ‘her/his/its’ are distal forms. The third person pronouns are underspecified for animacy; this was implemented by making the template call to ANIM optional.\footnote{See the XLE documentation on encoding optional f-annotation at http://www2.parc.com/isl/groups/nltt/xle/doc/notations.html#N.9.3.4}

(3) a. mErA PRON * @(POSS-PRONOUN-MASC meN 1 sg masc sg nom)
   @(ANIM +).

b. mErE PRON * { @(POSS-PRONOUN-MASC meN 1 sg masc sg obl)
   | @(POSS-PRONOUN-MASC meN 1 sg masc pl nom)
   | @(POSS-PRONOUN-MASC meN 1 sg masc pl obl)
   } @(ANIM +).

c. mErI PRON * @(POSS-PRONOUN-FEM meN 1 sg fem)
   @(ANIM +).

(4) a. uskA PRON * @(POSS-PRONOUN-MASC vuh 3 sg masc sg nom)
   @(DEIXIS distal)
   { @(ANIM +)}.\footnote{See the XLE documentation on encoding optional f-annotation at http://www2.parc.com/isl/groups/nltt/xle/doc/notations.html#N.9.3.4}
4.2. GENITIVE CASE MARKING

b. uskE PRON * { @(POSS-PRONOUN-MASC vuh 3 sg masc sg obl) |
                @(POSS-PRONOUN-MASC vuh 3 sg masc pl nom) |
                @(POSS-PRONOUN-MASC vuh 3 sg masc pl obl) } 
   @(DEIXIS distal) 
   { @(ANIM +) } .
c. uskI PRON * @(POSS-PRONOUN-FEM vuh 3 sg fem) 
   @(DEIXIS distal) 
   { @(ANIM +) } .

(5)  
a. POSS-PRONOUN-MASC(_F _P _N _CG _CN _CM) =  
   @(PRED _F)  
   @(PERS _P)  
   @(NUM _N)  
   @(NSYN pronoun)  
   @(PRON-TYPE poss)  
   ( ({SUBJ|OBJ|ADJUNCT $} ^ ) = %GEN-IO  
     (%GEN-IO GEND) =c _CG  
     (%GEN-IO NUM) =c _CN  
     (%GEN-IO CHECK _NMORPH) =c _CM  
   @NTYPE-REQ.  
b. POSS-PRONOUN-FEM(_F _P _N _CG) =  
   @(PRED _F)  
   @(PERS _P)  
   @(NUM _N)  
   @(NSYN pronoun)  
   @(PRON-TYPE poss)  
   ( ({SUBJ|OBJ|ADJUNCT $} ^ ) GEND) =c _CG  
   @NTYPE-REQ.  

The rule in (6), repeated from §2.7.3.3, constructs the Hindi/Urdu KP. The rule follows the analysis of Hindi/Urdu case clitics put forward by Butt and King (2004b) in that the case clitic is a functional head that takes an NP as its complement. The third line in (6) constrains the morphological form of the nominal marked by the case clitic to be in the oblique. Bear nominatives are handled by making the k head optional.  

(6)  
   KP --> NP  
   (K: ( ^ CHECK _NMORPH) =c obl).  

---

As is, the rule in (6) does not account for coordinated NPs within genitive KPs as described in §3.5.1.1. A complete discussion of KP and NP coordination and its implementation in LFG would lead to far afield here. However, I am confident that the Hindi/Urdu facts can be explained in a straightforward fashion using the approach to NP coordination in Dalrymple (2001), p. 380ff.
4.2. GENITIVE CASE MARKING

The rule in (7a) constructs the Hindi/Urdu NP constituent. NP nodes may be either simple personal pronouns, or nouns (which are possibly modified). Thus, the right side of the rule is a disjunction between the two options. Nadj is a metacategory (see §2.5.3.7); its definition is shown in (7b). Nadj is modified from what was seen in §2.7.3.3 to include the reanalyzed genitive phrases. The first line in Nadj attaches a KP which is assigned to bear genitive case; this models the fact that the genitive is a structural case. The KP is functionally annotated as either a SUBJ, an OBJ or as part of an ADJUNCT set. The Kleene star (*) on the KP node label results in a possibly infinite number of KPs attached to the NP (see §2.5.2). Naturally, infinite attachment is only applicable to the genitives annotated as parts of the ADJUNCT set; multiple SUBJ or OBJ functions will be ruled out as inconsistent in LFG (Kaplan and Bresnan, 1982). The meta-category further makes use of the shuffle operator “,” introduced in §2.5.3.6, to model the free word order inside the modified NP. Possessive pronouns are annotated in the same way as genitive KPs, since the two types of modifiers do not display functional differences; the lexical entries and templates for the possessive pronouns are in (3–4) and (5), respectively.

(7)  a. NP --> {PRON

| Nadj}.

b. Nadj = KP*: (! CASE) = gen

! <h ^

{@SUBJ|@OBJ|@ADJUNCT} ,

PRON*: (! PRON-TYPE) = poss

! <h ^

{@SUBJ|@OBJ|@ADJUNCT} ,

AP*: @ADJUNCT

{f::RS* = ^

| f::RS* ~= ^

@ (OT-MARK attach)} ,

N.

Recall from §3.5.1.2 that genitive KPs do not occur after the head in the Hindi/Urdu NP. In XLE, this can be modeled using the head precedence annotation “! <h ^”; this constraint is true only if the f-structures “!” (the genitive KP/possessive pronoun

---

Nadj is simplified in order not to clutter up the description, e.g., it does not include quantifiers. The description here excludes locative nominal arguments, which are the subject of Chapter 5.

A detailed study of the possible word order variations in the Hindi/Urdu NP remains to be carried out. However, several different orders were identified in §3.5.1, which point to a generally free order warranting the use of the shuffle operator.

The genitive KP may, however, be scrambled out of the NP to various positions, including to the right of the NP they are licensed in; the implementation of genitive scrambling is described in §4.5.
4.3. SELECTING THE GENITIVE

f-structure) and “¬” (the nominal head f-structure) have heads and the head of “!” precedes the head of “¬” in the c-structure (see §2.5.3.3). In addition, it was seen in §3.5.1.2 that if adjectives can have multiple attachments based on their morphological agreement features, local attachments are preferred as in the example in (52) (repeated from (52)):

(8)  
coti  larķi = ki  kita
‘the small girl’s book’
preferred over
‘the small book of the girl’

The preference is modeled in (7b) by attaching a constraint on the c-structure metavariable RS* (see §2.5.3.9 for a description of c-structure metavariables) to the functional annotation of the adjective phrase (AP). The constraint states that either the f-structure of the immediately right c-structure sister node is equal to the f-structure of the adjective’s head’s f-structure (ˆ), in which case there are no additional constraints, or that it is not equal to the f-structure of the adjective’s head’s f-structure, in which case the annotation points to the ParGram common template OT-MARK which assigns an ot mark with the name attach to the analysis. The attach mark is specified as a dispreference mark: an analysis with that mark will be dispreferred over a competing analysis that does not have the mark.

4.3 SELECTING THE GENITIVE

In this section, I provide an overview of how the different nominal types discussed in this chapter (common, event, relational, proper nouns and pronouns) are annotated for possessor argument selection in the Urdu ParGram grammar. Specifically, expressing the correct generalizations concerning different noun types turns out to be straightforward, given the ParGram feature system for nominals (see §2.6.3).

4.3.1 DISTINGUISHING NOMINALS

As discussed in §2.6.3.2 NSYN specifies the syntactic type of the noun. Specifically, it distinguishes common nouns, pronouns and proper nouns. Here, the notion of common nouns must be seen as a broad syntactic notion that does not help in distinguishing common nouns that are semantically different; in particular, it does not set apart count nouns from nouns that express events or relations (§3.5.2.3–§3.5.2.5). This is accomplished by NSEM, and specifically, NSEM COMMON (§2.6.3.3) which annotates different semantic types.

7In the generation direction, one would ultimately want to state preferences as well; for example, if the generator receives as input an f-structure where the adjective modifies the head noun, one would want the grammar to only generate genitive-initial strings. A more complete generation model of preferences in NP-internal orders would have to be the result of an extensive corpus study, for which there is no space here.
4.3. SELECTING THE GENITIVE

<table>
<thead>
<tr>
<th>Noun type</th>
<th>NSYN</th>
<th>NSEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common count nouns</td>
<td>common</td>
<td>COMMON count</td>
</tr>
<tr>
<td>Common mass nouns</td>
<td>common</td>
<td>COMMON mass</td>
</tr>
<tr>
<td>Event nouns</td>
<td>common</td>
<td>COMMON event</td>
</tr>
<tr>
<td>Relational nouns</td>
<td>common</td>
<td>COMMON count</td>
</tr>
<tr>
<td>Proper nouns</td>
<td>proper</td>
<td>PROPER ...</td>
</tr>
<tr>
<td>Pronouns</td>
<td>pronoun</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 4.1: Hindi/Urdu noun type annotation

of common nouns (Butt et al., 1999b). Recall from §2.6.3.3 that apart of COMMON, other sub-features of NSEM are used for nouns that do not take arguments and thus are not of importance here; proper nouns, for example, receive NSEM PROPER (and further sub-features for distinguishing different types of proper nouns/names), while pronouns do not have an NSEM feature.

For event nouns (§3.5.2.4), I propose an annotation in terms of NSEM COMMON event; this is an addition to the ParGram standard NSEM COMMON feature space, which currently only offers gerund as a value for gerund nouns. In §3.5.2.4, I have not made a distinction between different types of event nominals in terms of their morphosyntax; for example, I have not distinguished nominals derived from verbal roots from non-derived event nominals, since such nouns behave alike in terms of predication. It remains to be seen whether different types of event nominals behave differently syntactically, to a level where one would want to annotate such differences at f-structure. That being the case, it would not seem plausible to enforce an annotation using a gerund value, which entails that all such nouns are gerunds. The final feature space for the nominals discussed in this chapter is as in Table 4.1.

As can be seen from Table 4.1, the feature space proposal does not set apart common count nouns from relational nouns. To a level that is relevant for functional information encoded at f-structure, this underspecification is intentional. Recall from §3.5.2.3 and §3.5.2.5 that both nominal types may or may not occur with a SUBJ, and both occur with ADJUNCT genitives. They also behave alike with respect to other syntactic properties, e.g., countability or modification using adjectives, quantifiers, etc. As far as I am aware, none of the ParGram grammars employ different values for these two types of nouns. The fact that relational nouns intrinsically take a SUBJ (which may be suppressed), and common count nouns extrinsically may be augmented to take a SUBJ, is a purely semantic observation, ultimately resulting in identical behavior at f-structure.
4.3. SELECTING THE GENITIVE

4.3.2 INTERFACING MORPHOLOGY & SYNTAX FOR NOUNS

In the Urdu ParGram grammar, common (count and mass) nouns, relational nouns, event nouns and proper nouns are part of the finite-state morphological analyzer (FSMA) and are fed to the grammar via the morphology-syntax interface (see §2.7.3.1). Pronouns, on the other hand, are dealt with exclusively in the lexicon of the grammar. This design decision was made due to the fact that pronouns constitute a minor word class and display idiosyncratic behavior that warrants a treatment in the lexicon; here, the implementation follows the suggestions made by Kaplan et al. (2004).

(9) shows the morphological analyses for a common count noun; (9b), (9c), (9d) and (9e) show analyses for a common mass, a relational, an event and a proper noun, respectively. Relational nouns receive the same morphological analysis as common count nouns.

(9)

a. Common count noun: gari ‘car’
   gARI ↔ gARI+Noun+Count+Fem+Sg
b. Common mass noun: roʃni ‘light’
   roSnI ↔ roSnI+Noun+Mass+Fem+Sg
c. Relational noun: beṭi ‘daughter’
   bETI ↔ bETI+Noun+Count+Anim+Fem+Sg
d. Event noun: tabahi ‘destruction’
   tabAhI ↔ tabAhI+Noun+Event+Fem+Sg
e. Proper noun: ram ‘Ram’
   rAm ↔ rAm+Noun+Name+Masc

The sublexical rule in (10) parses the N category, i.e., all nominals in the FSMA fed to the syntax component. The major disjunction in the rule distinguishes common from non-common (i.e., proper) nouns in terms of the NSYN feature. The first disjunct is used for common nouns; here, the N-T node is annotated calling the common template NSYN with the value common. Common nouns are then divided into count, mass and event nouns in a smaller disjunction. The second major disjunct is used for proper nouns, and N-T is annotated for NSYN proper. This disjunct includes the category NAME, which may be filled by the tag +Name used for person names.8

8 Other NAME tags in the Urdu ParGram grammar include e.g. +City, +Country.
4.3. SELECTING THE GENITIVE

(10) \( N \rightarrow \text{NOUN-S} \)
\[
\{N-T: \emptyset(\text{NSYN common}) \\
\{\text{COUNT} \\
\{\text{ANIM} \\
\{\text{MASS} \\
\{\text{EVENT}} \\
\text{GEND} \\
\text{NUM} \\
(N-MORPH) \\
|N-T: \emptyset(\text{NSYN proper}) \\
\text{NAME} \\
(GEND) \\
\text{(NUM)}\}.
\]

The guesser in (11), repeated here for convenience from §2.7.3.1, then ensures that any item in the morphology may be analyzed as a noun stem, provided the tag sequence attached to the item matches the sublexical noun rule definition in (10). The guesser sends all items successfully analyzed as NOUN-S to the NOUN template in (12) (also repeated from §2.7.3.1). Note that this applies to all noun stems, including proper nouns. Recall that the special -unknown lexical entry thus matches stems analyzed by the FSMA which are not given explicit entries elsewhere in the lexicon. This aspect will become relevant in the discussion on event nominals (§4.3.5).

(11) -unknown NOUN-S XLE \(\emptyset(\text{NOUN }\%\text{stem})\).

(12) NOUN(_P) = \(\emptyset(\text{PRED }_P)\).

An overview of the sublexical entries used in (10) is given in Table 4.2. Using this system, all nouns provided by the FSMA are systematically annotated for various morphosyntactic and lexical-semantic features, which can be used in constraining nominal argument selection and genitive modification in general.

4.3.3 COMMON & RELATIONAL NOUNS

Common nouns do not, by themselves, subcategorize any possessor arguments; their lexical argument structure does not include any such arguments. As was seen in §3.5.2.3, they may still be modified by extrinsic possessors, which map onto a SUBJ function. To account for this, I follow the proposals by Bresnan (2001) and Sadler (2000) in positing a lexical predication template that provides a subcategorization frame with a SUBJ function for common nouns. However, contrary to Bresnan and Sadler, but following Chisarik and Payne (2003), under my analysis common nouns are augmented to include a single SUBJ function instead of a POSS function, signaling the discourse-sensitivity as well as the thematic unrestrictedness of possessors. For a common noun such as darvaza ‘door’
### 4.3. SELECTING THE GENITIVE

<table>
<thead>
<tr>
<th>Subl. category</th>
<th>Usage</th>
<th>Example entry</th>
<th>Example annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN-S</td>
<td>stem</td>
<td>gARI</td>
<td>@(NOUN %stem)</td>
</tr>
<tr>
<td>N-T</td>
<td>POS tag</td>
<td>+Noun</td>
<td>—</td>
</tr>
<tr>
<td>GEND</td>
<td>gender</td>
<td>+Masc</td>
<td>@(GEND masc)</td>
</tr>
<tr>
<td>NUM</td>
<td>numer</td>
<td>+Sg</td>
<td>@(NUM sg)</td>
</tr>
<tr>
<td>N-MORPH</td>
<td>morph. form</td>
<td>+Ob1</td>
<td>@(NMORPH obl)</td>
</tr>
<tr>
<td>COUNT</td>
<td>count nouns</td>
<td>+Count</td>
<td>@(COMMON count)</td>
</tr>
<tr>
<td>MASS</td>
<td>mass nouns</td>
<td>+Mass</td>
<td>@(COMMON mass)</td>
</tr>
<tr>
<td>EVENT</td>
<td>event nouns</td>
<td>+Event</td>
<td>@(COMMON event)</td>
</tr>
<tr>
<td>NAME</td>
<td>proper nouns (“names”)</td>
<td>+Name</td>
<td>@(PROPER-TYPE name)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>@(ANIM +)</td>
</tr>
<tr>
<td>ANIM</td>
<td>animate count nouns</td>
<td>+Anim</td>
<td>@(ANIM +)</td>
</tr>
</tbody>
</table>

Table 4.2: Hindi/Urdu sublexical categories for nouns

with the lexical entry in (13a), the lexical predication template for SUBJ augmentation looks as in (13b).

(13) a.  
         darvaza  N  (↑ PRED) = ‘darvaza’

        b. ‘darvaza’ ⇒ ‘darvaza<(↑ SUBJ)>’

Conversely, relational nouns may have their SUBJ suppressed. In LFG, this can be represented by assuming a suppression template as in (14) for beṭi ‘daughter’, which is formally the inverse of (13).

(14) a.  
         beṭi  N  (↑ PRED) = ‘beṭi<(↑ SUBJ)>’

        b. ‘beṭi<(↑ SUBJ)>’  ⇒  ‘beṭi’

In the Urdu grammar, both of these processes are implemented by altering the NOUN template from (12) as in (15). The first disjunct in the template is applied to proper nouns by constraining the NSYN feature to have the value proper; thus, all proper nouns from the FSMA pass through the first disjunct (having been annotated as NSYN proper by (10)). Notably, the first disjunct does not include a subcategorization frame, modeling the fact that proper nouns do not take arguments. The second disjunct is applied to common, non-proper nouns. The template is formally equivalent to the augmentation template stated in (13) as well as the suppression template in (14), since nouns (relational, common count or common mass) calling up the template may or may not receive a subcategorization frame selecting for a SUBJ.
4.3. SELECTING THE GENITIVE

(15) \text{NOUN(_P)} = \{(\wedge \text{NTYPE NSYN}) = \text{c proper} \\
\quad \circ (\text{PRED _P}) \\
\quad | (\wedge \text{NTYPE NSYN}) = \text{c common} \\
\quad \{\circ (\text{PRED _P}) \\
\quad | \circ (\text{SUBJ_core _P})\}\}.

An example of an NP with a common head noun darvaza ‘door’ is given in (16). The NP by itself is ambiguous between two readings, one where the genitive KP maps to an adjunct, and one where it’s a subject. That is, it may or may not be an extrinsic possessor. As discussed in §3.5.2, there are specific constraints in possessor selection; this case, however, is entirely ambiguous between the two readings.\footnote{To see that the genitive KP in (16) may be used attributively as an adjunct, consider, for example, a situation where the door is painted with a house-like shape.} The resulting c-structure output (identical for both analyses) is shown in Figure 4.1; the two f-structures are in Figure 4.2. Figures 4.3 and 4.4 are c- and f-structures for a relational noun head.

(16) makan=ka darvaza
    house.M.SG = GEN.M.SG door.M.SG
    ‘the house’s door’

(17) nina=ki beṭi
    Nina.F.SG = GEN.F.SG daughter.F.SG
    ‘Nina’s daughter’

\begin{figure}[h]
\centering
\begin{tikzpicture}
\node (NP) at (0,0) {NP};
\node (KP) at (-1.5,-1) {KP};
\node (N) at (1.5,-1) {N};
\node (NP-kA) at (-3,-2) {NP kA};
\node (N-kA) at (-1.5,-2) {N kA};
\node (makAn) at (-3,-3) {makAn};
\node (darvAzA) at (-1.5,-3) {darvAzA};

\draw[->] (NP) -- (KP);
\draw[->] (NP) -- (N);
\draw[->] (NP-kA) -- (N-kA);
\draw[->] (N-kA) -- (makAn);
\draw[->] (N-kA) -- (darvAzA);
\end{tikzpicture}
\caption{Hindi/Urdu NP c-structure, common noun head}
\end{figure}

4.3.4 PROPER NOUNS & PRONOUNS

Given the template in (15), proper nouns are excluded from being augmented to select a SUBJ; they are annotated as (\wedge \text{NTYPE NSYN proper}) and thus can only enter the first disjunct. Pronouns are not part of the morphology of the grammar and hence never

\footnote{CHECK features are shown here once, but are suppressed in the remaining f-structures in this chapter.}
call up (15); the functional annotation pertaining to pronouns was already described in §2.7.3.2 for the pronoun  \( mɛ̃ \) ‘I’ and is not repeated here.

For modeling the fact that proper nouns, but not pronouns, may be modified by attributive (adjunct) genitives, I propose an addition to (7b). (18) states that either (\^\text{NTYPE NSYN}) is set to common, or it is set to proper. In the former case, no further constraints are enforced; in the latter case, an OT mark called proper-mod is specified. The mark is specified in the optimality marks ranking of the grammar as a dispreference
mark; that is, an analysis with that mark will be dispreferred over a competing analysis that does not have the mark, e.g., a configuration where the genitive is analyzed as a scrambled genitive of a different (non-proper) nominal in the clause (see §4.5). Adding (18) to the ! $ (ADJUNCT) option in the KP and PRON entries in (7b) results in pronouns being excluded from and proper nouns being dispreferred for attributive genitive modification.11

(18) Constraint on attributive genitives and proper nouns/pronouns:

\[
\{(\sim \text{NTYPE NSYN}) =c \text{ common } \\
| (\sim \text{NTYPE NSYN}) =c \text{ proper } \\
\oplus (\text{OT-MARK proper-mod})\}
\]

### 4.3.5 Event Nouns

Event nouns come with fully specified subcategorization frames, and thus do not partake in the lexical augmentation rule shown in (13b). They do, however, undergo argument suppression as schematized in (14). The amount of GFS licensed is lexically fixed. First, as was shown in §3.5.2.4, a single, thematically unrestricted genitive phrase may be

11The implementation of OT marks in XLE is explained in §2.5.3.10.
selected as a SUBJ. An example lexical entry for tabahi ‘destruction’ is given in (19). This particular noun, although being derived from a transitive verb, mandates only a single genitive argument; this genitive may realize either the agent or the patient argument as a SUBJ. All event nouns receive the tag +Event in the morphological analyzer; the tag calls up the template in (20) as described in §4.3.2.

(19) tabAhI NOUN-S XLE { @(SUBJ_core tabAhI) | @(PRED tabAhI) }.

(20) EVENT = @(COMMON event)
 | @(NSYN common).

Event nouns are also part of the FSMA of the Urdu grammar; the lexical category for these nouns is NOUN-S, which matches the sublexical rule in (10). By providing lexical entries for event nouns as in (19), an unwanted analysis in terms of the -unknown entry is effectively blocked. This also means that event nouns do not call up the NOUN template, so that their subcategorization information is specified exclusively in their lexical entries.

(21) is an example involving the noun in (19). The c- and f-structures in Figures 4.5 and 4.6 are identical for the two readings of (21), where gari=ki ‘the car’s’ is either an agent or patient. The disambiguation of (21) is left to the semantics, which is adequate given that the syntax does not rule out either reading.

(21) gari=ki tabahi
car.F.SG = GEN.F.SG destruction.F.SG
‘the car’s destruction’

Second, there is a small class of event nouns that select two distinct genitive arguments: an unrestricted SUBJ function and a restricted OBJ function. The noun gʰɛrao

---

---
4.4. ENCODING ANIMACY

Throughout §3.5.2, it was seen that Hindi/Urdu places a constraint on genitive modification of animate head nouns; such nouns uniformly require their \textsc{subj} genitives to be

\begin{verbatim}
Figure 4.6: Hindi/Urdu NP f-structures, event noun head l
\end{verbatim}

‘circumvention’ is a member of this class; the lexical entry for \texttt{gʰɛrao} is given in (22). In addition, one or all of the \texttt{gfs} may be missing.

\begin{verbatim}
(22) gHerAO NOUN-S XLE {@(SUBJ-OBJ_core gHerAO) |
| @(SUBJ_core gHerAO) |
| @(OBJ_core gHerAO) |
| @(PRED gHerAO)}. 
\end{verbatim}

(23) is an example, repeated from (79), with the noun in (22). C- and f-structures for (23) are given in Figures 4.7 and 4.8. Given the annotation of the lexical entry in (22), other readings are produced by the grammar (e.g., an analysis where both arguments are suppressed and the genitive \texttt{kps} are interpreted as adjuncts), but have to be ruled out outside of the grammar.

\begin{verbatim}
(23) noojavanō = ka tʰane = ka 
youngster.M.PL.OBL = GEN.M.SG police-station.M.SG.OBL = GEN.M.SG gʰɛrao 
circumvention.M.SG
‘the circumvention of the police station by the youngsters’ (Raza, 2010) 
\end{verbatim}
animate as well. At the same time, animate nouns may freely be modified by inanimate adjunct genitives. The constraint is independent of the noun type; relational nouns as well as common nouns, if animate, only take animate genitive SUBJS.

To account for this, the Hindi/Urdu FSMA used by the grammar was extended to encode animacy. Person names as well as animate count nouns are annotated as animates via a call to a common template: @(ANIM +). For person names, the call is part of the tag entry of +Name; for animate count nouns, the FSMA provides an analysis in terms of a +Anim tag, the annotation of which includes the call (see (9) as well as Table 4.2). The template call results in a feature ANIM, valued +. Note that animacy is currently encoded in terms of underspecification, that is, only animate nouns receive the ANIM feature. This treatment is parallel to the way the English ParGram grammar encodes human/non-human nominals using HUMAN +.

In the grammar, the animacy restriction can be enforced by adding the constraint in (24) to the annotation of genitive KP s and possessive pronouns in the Nadj metacategory. The constraint states that either (∼ ANIM) (the head noun’s animacy feature) is set to +, in which case the feature needs to agree with the genitive’s feature, or (∼ ANIM) is not
set to +. The annotation thus ensures that the features only agree when the head noun is animate. If the head noun is inanimate (and hence does not have (\(^ \neg \ ANIM \)) in the current implementation), the lower disjunct is used as it only constrains (\(^ \neg \ ANIM \)), but does not exclude its absence. The notation is also compatible with future changes to the grammar, which may involve a specification of (\(^ \neg \ ANIM \)) in terms of a - value for inanimates (e.g., in terms of the incorporation of additional lexical resources).

(24) Animacy constraint annotation:

\[
\{(\neg \ ANIM) = c + \\
(\neg \ ANIM) = c (! \ ANIM) \\
(\neg \ ANIM) = c +
\} \\
\]

(24) is added to those disjuncts in (7b) that annotate the KP/possessive pronoun as a SUBJ. The final Nadj template with all necessary constraints concerning nominal type and animacy is shown in (25).\(^{13}\) Given the animacy constraint, the non-animate genitive kp modifying an animate head noun in (26) is analyzed exclusively as an adjunct (Figure 4.9).

(25) Nadj = KP*: (! CASE) = gen

! <h ^

\{\@SUBJ

\{(\neg \ ANIM) = c + \\
(\neg \ ANIM) = c (! \ ANIM) \\
(\neg \ ANIM) = c +
\} \\
\@OBJ \\
\@ADJUNCT

\{(\neg \ NTYPE NSYN) = c common \\
(\neg \ NTYPE NSYN) = c proper \\
(\neg \ NTYPE NSYN) = c proper \}

\}, \\
\@ADJUNCT \\
\}, \\
N.

(26) garĩ=ki orat
car.F.SG = GEN.F.SG woman.F.SG

‘the car’s woman’

Aside from possessor selection, it was seen in §3.5.2.3 that there are other possible applications of the ANIM feature; for example, animate objects have to be marked by the

\(^{13}\)The disjunct for possessive pronouns is not shown here, but is identical to the KP disjunct, except for the fact that it assigns (\(^ \neg \ PRON-TYPE \)) = poss instead of genitive case.
"gARI kI orat"

**Figure 4.9:** Hindi/Urdu NP f-structure, animate noun head

accusative case, no matter if definite or indefinite. Given the present implementation of animacy, this can easily be realized by notating the constraint in (27) on main clause KP nodes that are functionally annotated as clausal OBJ.

\[(27)\] \[\text{KP: } \@\text{OBJ}\]
\[
\{(! \text{ANIM}) = c + \\
!(\text{CASE}) = c \text{ acc} \\
|!-(! \text{ANIM})\}
\]

The constraint rules out ungrammatical sentences such as (28a) with the nominative object, but accepts grammatical sentences such as (28b) with the accusative object, (28c) as well as (74c). (28) is repeated from (74).

\[(28)\]
\[\text{a. } ila=ne \text{ bacce=ko/*bacca } ut^b\text{a-ya} \]
\[\text{Ila.F.SG = ERG child.M.SG.OBL = ACC/*child.M.SG.NOM lift-PERF.M.SG} \]
\[\text{‘Ila lifted a/the child.’} \quad (\text{Mohanan, 1994, p. 80})\]
\[\text{b. } ila=ne \text{ har } ut^b\text{a-ya} \]
\[\text{Ila.F.SG = ERG necklace.M.SG lift-PERF.M.SG} \]
\[\text{‘Ila lifted a/the necklace.’} \quad (\text{Mohanan, 1994, p. 80})\]
\[\text{c. } ila=ne \text{ har = ko } ut^b\text{a-ya} \]
\[\text{Ila.F.SG = ERG necklace.M.SG = ACC lift-PERF.M.SG} \]
\[\text{‘Ila lifted the necklace.’} \quad (\text{Mohanan, 1994, p. 80})\]

There are currently 45 count nouns analyzed as animate in the Urdu grammar; moreover, the grammar includes 31 person names, which are also annotated as animate. Adding in additional lexical items, for example by tapping into existing online or offline resources, can further improve the above approach.

### 4.5 Genitive Scrambling

As was seen in §3.6, there are also cases where the genitive KP is scrambled outside of the NP it modifies. An example of genitive scrambling is repeated from (109) in (29).
4.5. GENITIVE SCRAMBLING

(29) a. ram = ka
dost ay-a
Ram.M.SG = GEN.M.SG friend.M.SG.NOM come-PERF.M.SG
‘Ram’s friend came.’

b. dost ay-a ram = ka
friend.M.SG.NOM come-PERF.M.SG Ram.M.SG = GEN.M.SG
‘Ram’s friend came.’

To account for genitive scrambling, genitive KPs must be allowed to distribute across the Hindi/Urdu clause. The S rule is repeated from §2.7.3.3 in (30a); since the example in (29) involves the unaccusative verb a ‘come’, I also provide the S_unacc rule for constructing clauses with unaccusative verbs in (30b). S_unacc, like the other meta-categories in (30a), has been extended by a call to the meta-category KP-SCRAMBLE.

(30) a. S --> {S_unerg|S_unacc|S_trans|S_cop}.

b. S_unacc = KP: @SUBJ,
   @CL-ADJUNCTS,
   @KP-SCRAMBLE,
   VCmain: (^ LEX-SEM VERB-CLASS) =c unacc.

KP-SCRAMBLE, the definition of which is shown in (31a), adds an arbitrary number of KP nodes, which are required to bear genitive case. The second line in (31a) instantiates an f-structure from KP-SCRAMBLE-PATH and saves it as a local variable called %PATH (see §2.5.3.8). KP-SCRAMBLE-PATH is a regular predicate abbreviation making use of a functional uncertainty expression (§2.5.3.5); the abbreviation expands as in (31b). As seen in §3.6, it is possible to scramble genitives out of all core GFS as well as long-distance out of a non-finite complement clause; in the grammar, such clauses map to the XCOMP function, which is the GF traditionally used in LFG for non-finite complement clauses. Thus, KP-SCRAMBLE-PATH expands to all core GFS, possibly embedded in an XCOMP; notice the optionality brackets around XCOMP. KP-SCRAMBLE-PATH excludes ADJUNCT, as genitives may not be scrambled out of adjuncts.

15The grammar also includes a corresponding version of KP-SCRAMBLE for scrambled possessive pronouns.
4.5. GENITIVE SCRAMBLING

(31) a. KP-SCRAMBLE = KP*: (! CASE) = gen
   (~ KP-SCRAMBLE-PATH) = %PATH
   {(%PATH CASE) = c nom
   | (%PATH CASE) ~= nom
   ! <h %PATH}
   {(%PATH SUBJ) = !
   {(%PATH ANIM) = c (! ANIM)
   | ~ (%PATH ANIM)}
   | (%PATH OBJ) = !
   | ! $( %PATH ADJUNCT)
   {(%PATH NTYPE NSYN) = c common
   | (%PATH NTYPE NSYN) = c proper
   @(OT-MARK proper-mod)}
   @(OT-MARK attach).

   b. KP-SCRAMBLE-PATH = (XCOMP) {SUBJ|OBJ|OBL|OBJ-GO|OBJ-TH}.

Thus, the current approach includes genitive KPs that are scrambled outside of the core main clause GFs as well as GFs inside an XCOMP; the approach is warranted given the data that was established in §3.6. The implementation does not, however, allow for genitive KPs that are scrambled from within COMP f-structures (used for finite complement clauses in ParGram grammars/LFG; see Butt et al. (1999b), p. 30) (§3.6.2.3), from adjuncts (§3.6.2.4), or from other GFs more deeply embedded in some path (§3.6.2.5).

Lines 3–5 realize the case constraint discussed in §3.6.2.2. %PATH’s CASE feature is either required to be nominative (i.e., not overtly case-marked), in which case no additional constraints are enforced, or it’s required to be non-nominative, in which case head precedence (§2.5.3.3) is used to model the fact that the genitive phrase needs to precede its head in the clause. This way, examples such as the one in (32b) are ruled out, while in (32a), ram = ke ‘Ram’s’ is analyzed as a genitive phrase modifying kūṭṭe = ko ‘dog = ACC’.

(32) a. baccō = ne ram = ke kal
   child.M.PL.OBL = ERG Ram.M.SG = GEN.M.SG.OBL yesterday
   kūṭṭe = ko dekʰ-a
do̱g.M.SG.OBL = ACC see-PERF.M.SG
   ‘The children saw Ram’s dog yesterday.’

   b. * baccō = ne kūṭṭe = ko kal
      child.M.PL.OBL = ERG dog.M.SG.OBL = ACC yesterday
      ram = ke dekʰ-a
      Ram.M.SG = GEN.M.SG.OBL see-PERF.M.SG

To factor in the animacy constraint on the genitive SUBJ, the %PATH variable is checked for its animacy feature (lines 7–8). Concerning genitive adjunct modification of pronouns and proper nouns, %PATH is checked for the value of its NTYPE NSYN feature. If the feature
is valued common, no further constraints are applied. If the value is proper, the OT mark proper-mod is introduced as discussed above. Finally, the last line in KP-SCRAMBLE points to a ParGram common template which assigns an OT mark with the name attach to scrambled genitives. The attach mark is specified as a dispreference mark, and will thus disprefer the analysis over a competing analysis without the mark (as discussed above in §4.2). This models the fact (discussed in §3.6.2.1) that a genitive which is scrambled away from its head is dispreferred against an alternative reading where the genitive is analyzed as a local modifier (provided the morphosyntax allows for both readings). The resulting XLE parse for (29b) is shown in Figures 4.10 and 4.11.

Figure 4.10: Hindi/Urdu c-structure, genitive scrambling

"dOst AyA rAm kA"

Figure 4.11: Hindi/Urdu f-structure, genitive scrambling
4.6 Possessive Clauses

With the implementation of Hindi/Urdu NPs in place, I now turn to possessive clauses. Under my analysis, these are treated as intransitives. That is, they differ from the NPs reviewed above by including an intransitive verb, ho ‘be’, which embeds the NP in a spatio-temporal stative context. The genitive KP is not treated as a clausal subj, as proposed by e.g. Raza (2011) as well as Mohanan (1994), but as a GF embedded in a complex NP, i.e., the NP whose existence is predicated.

The KP and NP rules are already in place; see (6) and (7)/(25), respectively. Intransitive ho is an unaccusative verb; that is, it differs from other, unergative intransitive verbs in that it does not appear with ergative subjects (§2.3.2). The V-SUBJ_unacc template, shown in (33b), annotates intransitive ho accordingly. The template calls in (33b) are self-explanatory. Here, I only mention the template NON-AGENTIVE (33c), which annotates unaccusative verbs as non-agentive; the LEX-SEM AGENTIVE feature (together with other features not relevant here) controls for the availability of ergative case marking, by being checked against in the lexical entry of the ergative case marker. The S_unacc rule for clauses with unaccusatives was shown in (30b).

\[(33)\]

a. ho V-S XLE @(V-SUBJ_unacc ho);

b. V-SUBJ_unacc(_P) = @(SUBJ_core _P)
   @NON-AGENTIVE
   @(VERB-CLASS unacc)
   @NOPASSIVE
   @(SUBCAT-FRAME V-SUBJ).

c. NON-AGENTIVE = (^ LEX-SEM AGENTIVE) = -.

Together with the rules previously defined, this yields the c- and f-structure in Figures 4.14 and 4.13 for the possessive clause in (34).

\[(34)\]

ram=ka makan he
Ram.M.SG = GEN.M.SG house.M.SG.NOM be.PRES.3.SG
‘Of Ram is a house.’ = ‘Ram has a house.’

§3.7.2.3 has shown that the genitive KP clearly does not behave like a clausal subj with respect to the control diagnostic in (45). That is, a genitive KP cannot control an obligatory control site in a participial adjunct clause; e.g., in (35b), repeated from (153b), sri krijn=ka cannot control into the participial adjunct clause. The present analysis accounts for this in a straightforward fashion. The genitive KP is embedded in the clausal subj and thus cannot control into the participial clause.

\[\text{\textsuperscript{16}}\text{As will be seen in Chapters §5 and §7, there are other uses of ho, which are not displayed in (33a).}\]

\[\text{\textsuperscript{17}}\text{The lexical entry for the ergative case marker was given in \(\text{§87}\) in \(\text{§2.7.3.2}\). It is the VOLITION template in (87) that constrains the appearance of the ergative.}\]


Figure 4.12: Hindi/Urdu c-structure, possessive clause

"rAm kA makAn he"

Figure 4.13: Hindi/Urdu f-structure, possessive clause

(35) a. sri krñn = ka makan he
    Sri Krishna.M.SG = GEN.M.SG house.M.SG.NOM be.PRES.3.SG
    “Sri Krishna, has a house.’

b. * sri krñn = ka makan amir ho-te
    Sri Krishna.M.SG = GEN.M.SG house.M.SG.NOM rich be-IMPF.M.SG.OBL
    hu-e he
    be.PERF-M.SG.OBL be.PRES.3.SG
    ‘Sri Krishna, has a house _ni being rich.’
4.6. POSSESSIVE CLAUSES

Note that under the present analysis, the two sentences in (36) receive the same f-structures, which are given in Figure 4.15 and 4.16. This reflects the assertions in §3.7.3 that the two word orders convey the same functional information. Thus, any pragmatic and information-structural differences between the two utterances must be dealt with in the appropriate part of the grammar; for example, the difference in word order registered at the level of c-structure (Figure 4.14) can feed into the information-structural component, which can then assign TOPIC and FOCUS roles accordingly, as per the proposals by Butt and King (1996) and Butt and King (2000). The S_unacc rule defined in (30b) effectively treats the genitive KP in 36 as a rightward-scrambled genitive. The genitive KP is attached directly to the S node in the c-structure, but annotated as a SUBJ of the matrix SUBJ in the f-structure.

(36)  a. tale = ki cabi nāhī he
      lock.M.SG.OBL = GEN.F.SG key.F.SG.NOM not be.PRES.3.SG
     ‘Of the lock is not a key.’ = ‘The lock has no key.’
  b. cabi tale = ki nāhī he
      key.F.SG.NOM lock.M.SG.OBL = GEN.F.SG not be.PRES.3.SG
     ‘The key is not of a lock.’

Finally, it was seen that event nouns are not felicitous in possessive clauses (§3.7.4). An ungrammatical example is repeated from (157) in (37).

(37)   *ṭrain = ki rāvangi he
       train.M.SG = GEN.F.SG departure.F.SG.NOM be.PRES.3.SG
To implement this restriction, it is crucial for the grammar to rely on a differentiated treatment of nominal types. In the present implementation, the $S_{\text{unacc}}$ rule in (30b) was adapted as in (38), where a constraint is attached to the subject $KP$; the constraint states that the subject's value for $\text{NTYPE NSEM COMMON}$ is either $\text{event}$ (i.e., that the subject is an event noun), in which case the $TNS-ASP$ complex feature needs to have an $\text{ASPECT}$ feature as per the existential constraint in line 3 of (38), or that the subject is not an event noun.
The constraint could also be annotated on the EVENT template (20); there, however, the annotation would have to make use of local variables and/or inside-out function application. It was therefore decided to annotate the constraint on $S_{\text{unacc}}$ in favor of a clearer grammar. In sum, the constraint rules out unaccusatives such as (37) and allows unaccusatives with aspectual information and event noun subjects.

4.7 A COMPARISON TO THE PARGRAM STATE OF THE ART

The detailed feature system of ParGram is shown to be helpful in constraining parses and ruling out certain infelicitous configurations; in particular, the $\text{NSYN}$ and $\text{NSEM}$ features can be used and extended to model the syntactic constraints.

I have presented the state-of-the-art approach to nominal predication and the representation of nominal arguments in the ParGram project in §2.6. Recall that ParGram currently uses the $\text{SPEC}$ feature for encoding possessors; specifically, possessors end up in a POSS f-structure inside the $\text{SPEC}$ feature complex. An f-structure produced by the German ParGram grammar for (39) is repeated below.

(39) Karls Hund
  Karl.M.SG.GEN dog.M.SG.NOM
  ‘Karl’s dog’

"Karls Hund"

Figure 4.17: ParGram f-structure for German possessive
The novel analysis of genitive phrases proposed here departs from the standard ParGram analysis in suggesting a more detailed approach to genitive configurations. An underspecified treatment of possessors cannot pay tribute to the specific patterns of genitives/possessors in all languages; the situation in Hindi/Urdu clearly requires a more refined approach. It results in some amount of ambiguity, not all of which can be resolved at the level of syntax (§3.9); but this cost is necessary in order to be able to explain the patterns in Hindi/Urdu. Let us assume for a moment an implementation of the Hindi/Urdu genitive by means of SPEC POSS; at least the following issues would immediately arise:

- The reflexive binding of *apna* ‘self’, the process of which is identical for a SUBJ genitive inside a nominal and a clausal SUBJ, would not be explained; instead, one would have to assume that the reflexive may be bound by a possessive specifier. However, all genitive phrases (including attributive and OBJ genitives) would end up under SPEC POSS, which would predict that these, too, can bind *apna*, which is not the case.

- The animacy restriction in genitive SUBJS could not be enforced. Recall that animate head nouns may only be modified by animate genitive SUBJS; the constraint is not present if the genitive is an adjunct. If genitive SUBJS and ADJUNCTS were treated alike, there would be no way of stating the constraint and deriving the correct facts in the grammar.

Thus, being faithful to the facts of nominal predication in Hindi/Urdu entails abandoning the SPEC POSS analysis for genitives. The issue will be discussed in the light of parallelism within ParGram in the overall conclusion of the thesis in Chapter §8.

4.8  CONCLUSION

This concludes the chapter on the implementation of the Hindi/Urdu genitive. The analysis presented here argues for a differentiated functional treatment of the genitive that is anchored in the lexical specification of the head nouns, contra a less refined approach in terms of a unified SPEC as is usually assumed in ParGram grammars. In Chapter §8, I will discuss implications for the ParGram project that follow from my proposal.

Overall, the treatment results in ambiguity where the functional interpretation of the genitive KP or the possessive pronoun cannot uniquely be determined. To that end, I have further discussed different selectional restrictions in terms of animacy and nominal type. Regarding animacy, animate nouns were shown to subcategorize only animate SUBJS as possessors. Regarding nominal type, pronouns were shown to resist all types of genitive modification, while proper nouns allow only ADJUNCT genitives. Given the rich feature system of ParGram, these restrictions can be enforced in straightforward ways.

I have also described a novel implementation of genitive scrambling, which was shown in §3.6 to affect all major GFS in a clause, but not adjuncts. The GF the genitive is
scrambled from may be embedded in an XCOMP. In addition, the selectional restrictions concerning animacy and proper nouns/pronominals apply in scrambled configurations as well. It was also found that generally, local attachments are preferred over scrambled configurations. The implementation makes use of different LFG/XLE tools that accurately model the data.
5.1 Introduction

The last chapter has looked at genitive nominal arguments. This chapter goes beyond genitives and looks at other nominal arguments that occur with locative as well as instrumental case. Formally, the case markers involved turn out to be typical Hindi/Urdu case markers, in that they show the typical behavior of case clitics in the language and head case phrases (kps). Contrary to the genitive case, they do not display morphological agreement. Regarding the selection of locative and instrumental kps inside NPS, what will emerge from this chapter is that there are two kinds of nominal heads in Hindi/Urdu: those that intrinsically license them and those that don’t. Those nouns that intrinsically license them come from two camps: event nouns and abstract relational nouns. Some common nouns may nevertheless be augmented to select a locative KP; I will call such nouns picture nouns. Functionally, I present evidence that the locative kps must be treated as arguments, not as adjuncts. Since they uniformly realize a theme argument, they must be analyzed as thematically restricted obliques in LFG. Semantically, the locative/instrumental case markers do not mark locations/instruments in a narrow sense, but are extended into non-spatial fields (Mohanan, 1994).

In locative clauses, the second major topic of the chapter, locative kps realize a location. Locative kps realizing a location may never appear on NPS, a feature that sets Hindi/Urdu apart from languages such as English, French or German. Locations always need a verbal element to be licensed in Hindi/Urdu. Locative clauses are argued to involve a copular verb ho ‘be’. In locative clauses, the word order may differ. What I will refer to as inverted locative clauses have the locative KP in sentence-initial position; here,
the locative \( kp \) behaves like a subject. In canonical locative clauses, the located nominal \( np \) occurs in sentence-initial position, and it behaves like a subject. Thus, in locative clauses in Hindi/Urdu, linear position is very much tied to functional behavior, which is curious given the general free-word order nature of Hindi/Urdu.

The chapter is structured as follows. §5.2 introduces the Hindi/Urdu locative and instrumental case clitics and gives a general overview of their distribution. §5.3 moves on to discuss in detail the usage of the locative and instrumental on elements within \( np \)s, reviewing their constituent as well as functional properties. §5.4 focuses on the scrambling possibilities of locative and instrumental arguments. §5.5 examines locative clauses, again regarding constituency as well as functional behavior. In §5.6, the facts concerning the word order in locative clauses are shown to require an analysis in terms of information structure. §5.7 concludes the chapter.

5.2 THE HINDI/URDU LOCATIVE & INSTRUMENTAL CASE

5.2.1 GENERAL DESCRIPTION

Hindi/Urdu features a variety of locative case markers as well as the instrumental marker \( se \). The locative case markers are clitics that feature the prototypical properties of the Hindi/Urdu case clitics as described in §2.3.2 (Ahmed, 2009; Butt and King, 2004b). An overview of the Hindi/Urdu locative case clitics is given in Table 5.1. The list given here is taken from Butt and King (2004b). An extensive discussion of the locative case clitics and other spatial markers across South Asian languages is also given by Ahmed (2009). Two of the locative case markers in Table 5.1 feature in other types of case marking as well: \( se \) is also used to mark the instrumental case, while \( ko \) also features as a dative/accusative case marker (Butt and King, 2004b).

<table>
<thead>
<tr>
<th>Clitic</th>
<th>Case</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>( mẽ )</td>
<td>Locative</td>
<td>‘in’</td>
</tr>
<tr>
<td>( par )</td>
<td>Locative</td>
<td>‘on’</td>
</tr>
<tr>
<td>( tak )</td>
<td>Locative</td>
<td>‘towards’</td>
</tr>
<tr>
<td>( \emptyset/ko )</td>
<td>Locative</td>
<td>‘to’</td>
</tr>
<tr>
<td>( se )</td>
<td>Locative</td>
<td>‘from’</td>
</tr>
<tr>
<td>( se )</td>
<td>Instrumental</td>
<td>‘with, by’</td>
</tr>
</tbody>
</table>

Table 5.1: Overview of Hindi/Urdu locative and instrumental case clitics

---

\(^1\)I exclude the postpositions discussed briefly in §2.3.3, only some of which carry a locative meaning. Recall that Butt and King (2004b) provide evidence that postpositions display differences in form as well as distribution when compared to case clitics.
5.2. THE HINDI/URDU LOCATIVE & INSTRUMENTAL CASE

5.2.2 THE DISTRIBUTION OF THE LOCATIVE

Regarding the distribution of the locative case in Hindi/Urdu, Butt and King (2004b) mention that it may mark an oblique grammatical function (i.e., non-object-like phrases that are subcategorized for by a verb) or an adjunct. (1–2) shows some examples of locative obliques. The examples in (1b) and (2b), where the locative phrases are missing, demonstrate that these are in fact obliques and are subcategorized for by the verbs in the respective sentences: it is ungrammatical to utter such sentences out of context.

Further evidence for the locative phrases’ status as obliques is the fact that in (1c) and (2c), the two locative phrases cannot bear the same status, i.e. they cannot both define the direction of the putting or climbing: one of them must be an adjunct defining the location of the event.

(1)

a. ram=ne kɪtab=ko mez=pɑr rɑkʰ-a
   ‘Ram put the book on the table.’

b. * ram=ne kɪtab=ko rɑkʰ-a

c. * ram=ne kɪtab=ko mez=pɑr
dabbe=mẽ rɑkʰ-a
   box.M.SG,OBL = LOC.IN put-PERF.M.SG

(2)

a. bɪlli dabbe=mẽ gʊs-i
cat.F.SG,NOM box.M.SG,OBL = LOC.IN enter-PERF.F.SG
   ‘The cat climbed in(to) the box.’

b. * bɪlli gʊs-i
cat.F.SG,NOM climb-PERF.F.SG

c. * bɪlli dabbe=mẽ mez=pɑr
cat.F.SG,NOM box.M.SG,OBL = LOC.IN table.F.SG = LOC.ON
gʊs-i climb-PERF.F.SG

The sentences in (3–4) are examples of locative adjuncts. The sentences are perfectly grammatical whether or not the locative phrase is expressed, even without further discourse context, which is clear evidence that the locative phrase is not subcategorized for by the verb (or by any other element in the clause, for that matter). Additional locative phrases may be added to further specify the location of the event/predicate.

(3)

a. nina=ne ram=ko (mɑkan=mẽ) dekʰ-a
   Nina.F.SG = ERG Ram.M.SG = ACC (house.M.SG = LOC.IN) see-PERF.M.SG
   ‘Nina saw Ram (in the house).’

2If the appropriate discourse context is given, one could imagine the obliques to be subject to pro-drop in (1b) and (2b); cf. Butt (2007), Butt and King (2000).
5.2. THE HINDI/URDU LOCATIVE & INSTRUMENTAL CASE

b. nina = ne ram = ko (makan = mê)
Nina.F.SG = ERG Ram.M.SG = ACC (house.M.SG = LOC.IN)
(gusulxan = mê)  
dekʰ-a
(bathroom.M.SG = LOC.ON) see-PERF.M.SG
‘Nina saw Ram (in the house) (in the bathroom).’

(4) a. nina (pahari = par) hans-i
Nina.F.SG.NOM (hill.F.SG = LOC.ON) laugh-PERF.F.SG
‘Nina laughed (on (top of) the hill).’

b. nina (pahari = par) (apni gaṛi = mê)
Nina.F.SG.NOM (hill.F.SG = LOC.ON) (self.F.SG car.F.SG = LOC.IN)
hans-i
laugh-PERF.F.SG
‘Nina laughed (on (top of) the hill) (in her car).’

Moreover, Mohanan (1994) discusses the use of the locative case on matrix subjects. This usage is of particular importance for this chapter. According to Mohanan, there are L-SUBJ which are associated with locative semantic configurations (such as in X, towards X, at X, near X), but extended to abstract semantic fields, such as mental states. Mohanan proposes that only in this case, a locative L-SUBJ gets realized as a grammatical SUBJ.

Examples from Mohanan (1994) for what she refers to as locative subjects are given below in (5). Note that the locative-marked KP is able to bind the reflexive apna ‘self’, which, together with other diagnostics, is motivation for Mohanan to treat the locative phrase as a SUBJ. (5a) describes the abstract containment of an inherent characteristic within Nina, while (5b) describes the abstract location of a burden or responsibility. I provide simplified versions of the sentences in (5) in (6). The sentences all involve a locative KP, which is a SUBJ, a theme NP, which bears some other grammatical function (GF), and the verb ho ‘be’, which acts as a copula verb.

(5) a. nina = mê apni mausi = ke liye baṛi
mãmta  he
affection.F.SG.NOM be.PRES.3.SG
‘Nina has a lot of affection for self’s aunt.’ (Mohanan, 1994, p. 176)

3 Recall that Mohanan (1994) distinguishes logical subjects (L-SUBJ), which correspond to the most prominent semantic argument of an individual argument structure (the deep-structure subject of Chomsky (1965), or the highest role of a theta-structure in the sense of Bresnan and Kanerva (1989) and Bresnan and Zaenen (1990), from grammatical subjects (SUBJ), which correspond to the SUBJ in LFG terms and ranks highest on a hierarchy of grammatical functions (Bresnan, 2001, Dalrymple, 2001). While the former is a semantic category relating to predicate-argument structure (e.g., the L-SUBJ may not always be identical with the SUBJ, for example in passivized structures), the latter is a syntactic entity that is relevant to the modeling of Hindi/Urdu nominal predication in terms of c- and f-structure that I propose in this chapter. Thus, I will only be concerned with generalizations that can be stated with respect to grammatical-functional concepts (e.g., Mohanan’s SUBJ).
5.2. THE HINDI/URDU LOCATIVE & INSTRUMENTAL CASE

b. ravi = par  
apne  
parivar = ki  
sari
jimmevari  
he
responsibility.F.SG.NOM be.PRES.3.SG
‘Ravi is solely responsible for (the care of) self’s family.’

(Mohan, 1994, p. 176)

(6) a. nina = mē  
mamta  
he
Nina.F.SG = LOC.IN affection.F.SG.NOM be.PRES.3.SG
‘In Nina is affection.’ = ‘There is affection in Nina.’

b. ravi = par  
jimmevari  
he
Ravi.M.SG = LOC.ON responsibility.F.SG.NOM be.PRES.3.SG
‘On Ravi is responsibility.’ = ‘There is responsibility on Ravi.’

Mohan (1994) effectively restricts the distribution of locative subjects to abstract semantic fields (which can be traced back to concrete spatial uses). That is, under Mohan’s account, locative kps only behave as subjs when they do not bear purely locative semantics, but are tied to abstract semantics, as is the case for the locatives in (5)/(6); in both of these sentences, the location must be seen as non-concrete and abstract, in the sense that e.g., mamta ‘affection’ is not located inside of Nina.

Other, semantically concrete locations, such as those in (7), are dismissed by Mohan (1994) as possible subjects, since they cannot plausibly be extended to non-spatial fields. The bulk of this chapter goes to show that this claim is too restrictive, and that concrete locations also behave as subjs in Hindi/Urdu. Thus, the analysis worked out here treats sentences such as the ones in (6) and (7) on a par — I will refer to such finite clauses as locative clauses. I examine these clauses in detail in §5.5.

(7) a. makan = mē  
cuhā  
he
house.M.SG = LOC.IN rat.M.SG.NOM be.PRES.3.SG
‘In the house is a rat.’ = ‘There is a rat in the house.’

b. mez = par  
kitab  
he
table.F.SG = LOC.ON book.F.SG.NOM be.PRES.3.SG
‘On the table is a book.’ = ‘There is a book on the table.’

In contrast to the pattern displayed by the genitive, locative-marked phrases do not occur inside nps as adjuncts in Hindi/Urdu. That is, nps such as the ones in (8) are ungrammatical. This is in contrast to other languages such as English and German (examples in (9) and (10), respectively), where the corresponding np-internal locative adjuncts are grammatical. The examples in (11) further illustrate; here, the nps are ungrammatical unless the locative kps are licensed by the verb in an adverbial phrase.

(8) a. * [makan = mē  
cuhā]np
house.M.SG = LOC.IN rat.M.SG
‘the rat in the house’
5.2. THE HINDI/URDU LOCATIVE & INSTRUMENTAL CASE

b. *[mez = par kɪtab]NP
table.F.SG = LOC.ON book.F.SG
‘the book on the table’

(9) a. the rat in the house
b. the book on the table

(10) a. die Ratte im Haus
the.F.SG.NOM rat.F.SG.NOM in.the.N.SG.DAT house.N.SG.DAT
‘the rat in the house’
b. das Buch auf dem Tisch
‘the book on the table’

(11) a. [mez = par *(par-i hu-i) kɪtab]NP
table.F.SG = LOC.ON *(lie-PERF.F.SG be.PERF-F.SG) book.F.SG
‘the book lying on the table’ (Raza and Ahmed, 2011b)
b. [sɑkul = mē *(beṭʰ-e hu-e) bɑcce]NP
school.M.SG = LOC.IN *(sit-PERF.M.PL be.PERF-M.PL) child.M.PL
‘the children sitting in the school’

The situation in Hindi/Urdu is more complicated, however. Raza and Ahmed (2011a) as well as Raza (2011) give an overview of nouns that occur with locative arguments. These nouns seem to involve the locative case marker par ‘on’ (Raza and Ahmed, 2010, 2011a,b). Examples of such nouns are given in (12–14). Semantically, these nouns do not seem to form a uniform class; abstract relational (12), event (13) as well as common nouns (14) all may occur with locative argument kps. Uttering nouns with multiple locative phrases (as in the b. examples) renders the NPs ungrammatical.

(12) a. [ram = par bʰɑrɔsa]NP
Ram.M.SG = LOC.ON trust.M.SG
‘trust in Ram’
b. * [ram = par nina = par bʰɑrɔsa]NP
Ram.M.SG = LOC.ON Nina.F.SG = LOC.ON trust.M.SG

(13) a. [sɑlamti = par barifɪng]NP
security.F.SG = LOC.ON briefing.F.SG
‘the briefing on security’ (Raza and Ahmed, 2011a)
b. * [sɑlamti = par saman = par barifɪng]NP
security.F.SG = LOC.ON equipment.M.SG = LOC.ON briefing.F.SG

(14) a. [mɑntɪq = par kɪtab]NP
logic.M.SG = LOC.ON book.F.SG
‘the book about logic’ (Raza, 2011)
b. * [mɑntɪq = par ʃɪkʃa = par kɪtab]NP
logic.M.SG = LOC.ON teaching.F.SG = LOC.ON book.F.SG
I will review NP-internal locative arguments in detail in §5.3, where it will become clear that they must be OBL-TH: thematically restricted obliques (OBL), instantiated to the thematic role of a theme.

5.2.3 The Distribution of the Instrumental

The instrumental case marker se is very versatile. It may occur on instrumental adjuncts (15a), on phrases expressing sources both locative (15b) and material (15c), as well as comitatives (15d). For more uses of the instrumental, refer to Butt and King (2004b).

(15) a. nadya = ne darvaza cabi = se kʰol-a
    Nadya.F.SG = ERG door.M.SG.NOM key.F.SG = INST open-PERF.M.SG
    ‘Nadya opened the door with a key.’
    (Butt and King, 2004b, p. 10)

b. nadya aj lahɔr= se a-yi
    Nadya.F.SG = ERG today Lahore = INST come-PERF.F.SG
    ‘Nadya came from Lahore today.’
    (adapted from Butt and King, 2004b, p. 10)

c. sonar = ne sone = se har
bana-ya
make-PERF.M.SG
‘The goldsmith made a necklace out of the gold.’
    (Butt and King, 2004b, p. 10)

d. nadya saddaf= se bat kar ruh-i
    Nadya.F.SG.NOM Saddaf.F.SG = INST talk.F.SG.NOM do stay-PERF.F.SG
he
be.PRES.3.SG
‘Nadya is talking to Saddaf.’
    (Butt and King, 2004b, p. 10)

Important for the present purposes are two things: first, the instrumental, like the locative, does not occur on adjuncts of nominals; cf. (16).

(16) a. * lahɔr= se nadya
    Lahore = INST Nadya.F.SG
    ‘Nadya from Lahore’

b. * sone = se har
gold.M.SG.OBL = INST necklace.M.SG.NOM
‘the necklace made out of gold’

Second, there are several event as well as abstract relational nouns that occur with instrumental nominal arguments. Examples are in (17).

(17) a. hatʰi= se dɑr
    elephant.M.SG = INST fear.M.SG
    ‘fear of an elephant’
5.3. THE LOCATIVE & INSTRUMENTAL IN NPS

This section surveys the NP-internal use of the locative and instrumental. I discuss constituent as well as functional properties of the relevant KPs. Regarding their constituent properties, I show that their distribution parallels the distribution of genitive KPs; e.g., they never succeed their head nominals, but may otherwise occur in various positions in the NP. Concerning the selectional properties of locatives and instrumentals, they turn out to be selected by nominals across noun classes: event nouns, relational nouns as well as a class of common nouns referred to as picture nouns. The section demonstrates that these noun types differ in the way they specify genitive arguments: while event nouns and abstract relational nouns intrinsically select for locative KPs as arguments, picture nouns do so by invoking a lexical rule that enables them to take locative KPs as arguments. Functionally, I show that the locative/instrumental (LI) KP must uniformly be analyzed as OBLs across all noun types, since they are thematically restricted to realizing a theme argument.

5.3.1 CONSTITUENT PROPERTIES

In this section, I discuss the constituent properties of NP-internal locative and instrumental arguments with respect to coordination, linear order as well as stacking.

5.3.1.1 COORDINATION

In NPs headed by nouns that license LI arguments, the NP marked by the locative or instrumental may be coordinated with other NPs. Here, the case clitic is attached to the whole coordinated NP, as in (18). Also, the whole LI KP can be coordinated with other LI KPs. Just like with the genitive (see §3.5.1.1), this may give rise to ambiguous structures. While (18a) unambiguously refers to trust in both Ram and Nina (at the same time), (18b) is ambiguous: the NP may refer to trust in both Ram and Nina, or it may refer to trust in Ram and a (possibly different) trust in Nina. Further examples are in (19) and (20).
5.3. THE LOCATIVE & INSTRUMENTAL IN NPs

(18) a. [[[ram]_{NP} or [nina]_{NP} = par]_{KP} bʰɑɾəsa]_{NP}
    Ram.M.SG and Nina.F.SG = LOC.ON trust.M.SG
    ‘trust in Ram and Nina’

b. [[[ram]_{NP} = par]_{KP} or [[nina]_{NP} = par]_{KP} bʰɑɾəsa]_{NP}
    Ram.M.SG = LOC.ON and Nina.F.SG = LOC.ON trust.M.SG
    ‘trust in Ram and in Nina’

(19) a. [[[salamti]_{NP} or [saman]_{NP} = par]_{KP} barifing]_{NP}
    security.F.SG and equipment.M.SG = LOC.ON briefing.F.SG
    ‘the briefing on security and equipment’

b. [[[salamti]_{NP} = par]_{KP} or [[[saman]_{NP} = par]_{KP} barifing]_{NP}
    security.F.SG = LOC.ON and equipment.M.SG = LOC.ON briefing.F.SG
    ‘the briefing on security and on equipment’

(20) a. [[[mantɪq]_{NP} or [[ʃɪkʃa]_{NP} = par]_{KP} kɪtab]_{NP}
    logic.M.SG and teaching.F.SG = LOC.ON book.F.SG
    ‘the book on logic and teaching’

b. [[[mantɪq]_{NP} = par]_{KP} or [[[ʃɪkʃa]_{NP} = par]_{KP} kɪtab]_{NP}
    logic.M.SG = LOC.ON and teaching.F.SG = LOC.ON book.F.SG
    ‘the book on logic and on teaching’

(21) a. [[[haṭʰi]_{NP} or [cuhe]_{NP} = se]_{KP} ḏɑr]_{NP}
    ‘fear of an elephant and mouse’

b. [[[haṭʰi]_{NP} = se]_{KP} or [[[cuhe]_{NP} = se]_{KP} ḏɑr]_{NP}
    ‘fear of an elephant and of a mouse’

Again, as was seen with the genitive, this behavior is expected and typical for the Hindi/Urdu case clitics (Butt, 1995, Butt and King, 2004b, Mohanan, 1994). The locative behaves just like the other case clitics in this respect (§2.3.2).

5.3.1.2 LINEAR ORDER

Several elements may intervene between the head nominal and the LI argument in the NP. While the general pattern is as in (12), i.e., the locative argument occurs right to the left of its head, other orderings are also possible. (22–27) are examples. It can be seen that an attributive adjective modifying the head noun may precede the LI argument, but the adjective may also intervene between the argument and the head noun.

(22) a. [salamti = par tɑfsili barifing]_{NP}
    security.F.SG = LOC.ON detailed.F.SG briefing.F.SG
    ‘detailed briefing on security’

b. [tɑfsili salamti = par barifing]_{NP}
    detailed.F.SG security.F.SG = LOC.ON briefing.F.SG
    ‘detailed briefing on security’
5.3. THE LOCATIVE & INSTRUMENTAL IN NP

(23)  
a. [nina = par cauda bʰɑrɔsa]_{NP}  
Nina.F.SG = LOC.ON broad.M.SG trust.M.SG  
‘broad trust in Nina’

b. [cauda nina = par bʰɑrɔsa]_{NP}  
broad.M.SG Nina.F.SG = LOC.ON trust.M.SG  
‘broad trust in Nina’

(24)  
a. [mɑntɪq = par aeʰi kɪtab]_{NP}  
logic.M.SG = LOC.ON good.F.SG book.F.SG  
‘the good book on logic’

b. [aeʰi mɑntɪq = par kɪtab]_{NP}  
good.F.SG logic.M.SG = LOC.ON book.F.SG  
‘the good book on logic’

(25)  
a. [muqaddamat = se cauda ɪstɪsna]_{NP}  
‘broad immunity from court cases’

b. [cauda muqaddamat = se ɪstɪsna]_{NP}  
‘broad immunity from court cases’

Examples where the adjective precedes the argument may be ambiguous; the adjective may modify the argument, or it may modify the head noun, provided the adjective is semantically compatible with both, and its morphological agreement features permit both readings. An ambiguous example is shown in (26). In such cases, it is again the local attachment that is preferred (cf. §3.5.1.2).

(26)  
hoʃyar lɑṛke = par bʰɑrɔsa  
intelligent boy.M.SG.OBL = LOC.ON trust.M.SG  
‘trust in the intelligent boy.’

preferred over  
‘intelligent trust in the boy’

One does not find examples where the head nominal precedes the locative argument inside the NP. For example, in the training section of the Hindi-Urdu Treebank (HUTB, Bhatt et al., 2009, Palmer et al., 2009), there are 7 occurrences of the nominal bʰɑrɔsa ‘trust’ together with a locative (i.e., par ‘on’) marked argument directly preceding it, but there is not a single instance of the locative occurring to the right of the head nominal inside an NP. This is also what native speaker judgments indicate: strings such as the ones in (27) are not felicitous NPs.

(27)  
a. * [cauda bʰɑrɔsa nina = par]_{NP}  
broad.M.SG trust.M.SG Nina.F.SG = LOC.ON

b. * [tɑfsili barīfing sɑlamti = par]_{NP}  
detailed.F.SG briefing.F.SG security.F.SG = LOC.ON
5.3. THE LOCATIVE & INSTRUMENTAL IN NPS 187

c. * [acʰi kitab mɑntiq = par]_NP
   good.F.SG book.F.SG logic.M.SG = LOC.ON

Another example illustrating the same syntactic generalization is in (28). Here, the NP with the locative argument is inside an ergative-marked KP. It is not possible according to my informants to reverse the order within said NP such that the argument succeeds the head.

(28) a. [[ram = par caude bʰɑrɔsə]_NP = ne]_KP nadya = ko
taqat
di-ya
strength.F.SG.NOM give-PERF.F.SG

   ‘Broad trust on Ram gave Nadya strength.’

b. * [[caude bʰɑrɔsə ram = par]_NP = ne]_KP
    broad.M.SG.OBL Ram.M.SG = LOC.ON trust.M.SG = ERG
    nadya = ko taqat di-ya
    Nadya.F.SG = ACC strength.F.SG.NOM give-PERF.F.SG

As with the genitive KPs, the order with respect to determiners/adjectives is quite unconstrained, as long as the argument precedes its head. Consider the examples in (29a–f) and (30a–f), showing that the linear order among attributive adjectives, determiners and locative arguments within the NP is completely free (cf. also (22)–(27)). The examples parallel the ones in examples (58) and (59) of §3.5.1.2.

(29) a. nina = par yeh cauda bʰɑrɔsə
    Nina.F.SG = LOC.ON this.SG broad.M.SG

   ‘this broad trust in Nina’

b. nina = par cauda yeh bʰɑrɔsə
c. yeh nina = par cauda bʰɑrɔsə
d. yeh cauda nina = par bʰɑrɔsə
e. cauda yeh nina = par bʰɑrɔsə
f. cauda nina = par yeh bʰɑrɔsə

(30) a. armi-cif = se kʊch tɑfsile mʊtalbɑh

   ‘some detailed demands to the army chief’

b. armi-cif = se tɑfsile kʊch mʊtalbɑh
c. kʊch armi-cif = se tɑfsile mʊtalbɑh
d. kʊch tɑfsile armi-cif = se mʊtalbɑh
e. tɑfsile kʊch armi-cif = se mʊtalbɑh
f. tɑfsile armi-cif = se kʊch mʊtalbɑh

Note that abstract nouns such as bʰɑrɔsə ‘trust’ do not pluralize in Hindi/Urdu (and thus behave like their English and German counterparts), which is why they are not compatible with some determiners such as the indefinite quantifier kʊch ‘some’. (30) uses the event head noun mʊtalbɑh ‘demand’. 
5.3. THE LOCATIVE & INSTRUMENTAL IN NPS

5.3.1.3 STAKING

Stacking NPs with LI arguments on one another is also possible in Hindi/Urdu. That is, the LI KP may itself be modified by another LI argument. Consider the example in (31). Here, the locative KP xuda=par ‘in god’ is an argument selected by yaqin ‘belief’, while the complex phrase xuda=par yaqin=par ‘on the belief in god’ is an argument of the overall head of the NP, b‘arosa ‘trust’.

(31) a. xuda=par yaqin=par b‘arosa
   god.M.SG = LOC.ON belief.M.SG = LOC.ON trust.M.SG
   ‘trust in the belief in god’
   b. xuda=par yaqin=par kitab
   ‘the book about the belief in god’
   c. moqaddamat=se istsne=par kitab
     ‘the book about immunity from court cases’

5.3.1.4 SUMMARY

The summary of the findings regarding the word order properties of LI arguments in NPs is as follows. The general word order obeys the head-final ordering of Hindi/Urdu, but is otherwise free. Determiners and (numeral) adjectives may intervene between the argument and its head. A complete c-structural account in XLE has to cover all the variations enumerated in (32). I omit other possible interactions with quantifiers, adverbs as well as pronouns and refer to descriptions of the Hindi/Urdu NP in e.g. Kachru (2006) and Verma (1971).

(32) a. LI argument — Determiner — (Numeral) Adjective — Noun
   b. LI argument — (Numeral) Adjective — Determiner — Noun
   c. Determiner — (Numeral) Adjective — LI argument — Noun
   d. Determiner — LI argument — (Numeral) Adjective — Noun
   e. (Numeral) Adjective — Determiner — LI argument — Noun
   f. (Numeral) Adjective — LI argument — Determiner — Noun

5.3.2 FUNCTIONAL PROPERTIES

What kind of f-structure should the LI KP receive in an LFG analysis? If one was to choose among the grammatical functions (GF) as they were described in §2.4.4, which would it be? Is it a function selected by the nominal head, or must the KPs be treated as adjuncts? Are the LI phrases restricted to realizing a certain thematic role, or do they represent an unrestricted GF? Is the function discourse-oriented, e.g., can it bind an anaphora in the
5.3. THE LOCATIVE & INSTRUMENTAL IN NP

NP domain? In this section, I discuss the functional properties of the LI phrases in NPs. After a general statement about semantic restrictedness, I turn to a discussion of the different kinds of Hindi/Urdu nouns that select for LI arguments, common nouns, event nominals and abstract relational nominals.

5.3.2.1 SEMANTIC RESTRICTEDNESS

Looking at the semantics of LI nominal arguments, it turns out that the head nominal places tight restrictions on the thematic role of the argument. Examples are in (33–34). The nominals in (33) occur with a theme argument, which is realized by a locative KP, while the nominals in (34) occur with source arguments, which are realized by instrumental KP.

(33) a. nina = par bʰɑrɔsa
Nina.F.SG = LOC.ON trust.M.SG
‘trust in Nina’ THEME
b. xʊda = par yɑqin
god.M.SG = LOC.ON belief.M.SG
‘belief in god’ THEME

c. sɑlamti = par bɑrifɪng
security.F.SG = LOC.ON briefing.F.SG
‘the briefing on security’ THEME (Raza and Ahmed, 2011a)
d. mɑntɪq = par kɪtab
logic.M.SG = LOC.ON book.F.SG
‘the book about logic’ THEME (Raza, 2011)
e. gɑndʰi = par fɪlm
Gandhi.M.SG = LOC.ON film.F.SG
‘the film about Gandhi’ THEME

(34) a. moqaddɑmat = se ɪstɪsna
court.case.M.PL = INST immunity.M.SG
‘immunity from court cases’ SOURCE (Raza and Ahmed, 2011a)
b. tsɛlab = se tɑbahi
flood.M.SG = INST destruction.F.SG
‘destruction due to the flood’ SOURCE (adapted from Raza, 2011, p. 168)
c. haṭʰi = se ɗɑr
elephant.M.SG = INST fear.M.SG
‘fear of an elephant’ SOURCE (Raza and Ahmed, 2011a)

The important generalizations that will emerge in this section about the data from Hindi/Urdu locatives assigned in NPs are the following. First, regarding its syntactic function, LI KPs uniformly behave as thematically restricted obliques. That is, they are semantically restricted to encode a specific thematic role, lexically specified by the head
noun. I reject the possibility that the locative KP’s are SUBJS/ADJUNCTS, based on the *apna ‘self’ test on reflexive binding defined in (162) as well as evidence from the combinatory possibilities of the LI arguments.

The second generalization concerns the way in which the LI is selected. Recall that not all event nouns and relational nouns can take LI arguments. I will argue that for those nouns that can take such arguments, the LI argument is always lexically selected, in a fashion similar to what Barker (1995) refers to as intrinsic possession regarding the lexical specification of genitive arguments. The choice between locative and instrumental case on the nominal argument is lexically driven: if the argument is a theme, it will carry locative case; otherwise, if the argument is a source, it will receive instrumental case. LI case marking must hence be regarded as semantic case marking (Butt and King, 2004b); the occurrence of locative or instrumental case on a nominal’s arguments is predictable from knowing about the particular nominal’s argument structure.

5.3.2.2 COMMON NOUNS

Above, common nouns with locative KP’s as in (35) (repeated from (8)) were described as ungrammatical. Still, the examples in (36) are grammatical. Here, the locative KP does not encode a location, but a theme argument. I am not aware of any common nouns that occur with an instrumental argument.

(35) a. * makan = mē cuha
   house.M.SG = LOC.IN rat.M.SG
   ‘the rat in the house’

   b. * mez = par kitab
      table.F.SG = LOC.ON book.F.SG
      ‘the book on the table’

(36) a. mantiq = par kitab
      logic.M.SG = LOC.ON book.F.SG
      ‘the book about logic’

      b. pyar = par gana
         love.M.SG = LOC.ON song.M.SG
         ‘the song about love’

      c. gandʰi = par film
         Gandhi.M.SG = LOC.ON film.F.SG
         ‘the film about Gandhi’

The ability to take theme locative KP’s is a lexical property of the particular common noun; it entirely depends on whether the noun’s meaning can be extended to be “descriptive” of a certain concept. For example, while it may be conceivable that a book may describe a certain table (e.g., one that is custom-built, very expensive, etc.), and thus, (35b) may be acceptable in this particular context, the phrases in (37) cannot plausibly be conceptualized.
5.3. THE LOCATIVE & INSTRUMENTAL IN NPS

(37) a. *mantiq = par gari
    logic.M.SG = LOC.ON car.F.SG
b. *pyar = par qalam
    love.M.SG = LOC.ON pen.M.SG
c. *gandʰi = par cuha
    Gandhi.M.SG = LOC.ON rat.M.SG

The subclass of common nouns that can take locative-marked arguments in Hindi/Urdu correspond to what has been labeled “picture nouns” in the literature, first described by Harris (1976). Examples for picture nouns in English taken from Soames and Perlmutter (1979), p. 293, are in (38).

(38) picture, photograph, sketch, portrait, story, anecdote, joke, poem, novel, book, movie, play, rumor, lie

When the apna ‘self’ test for reflexive binding (§3.3) is applied, it is found that the locative argument of picture nouns cannot bind the reflexive; judging by this test alone, the locative kps when modifying picture nouns do not behave like subjs.

(39) a. *mantiqi = par apni kitab
b. *pyarī = par apnai gana
c. *gandʰii = par apni film
    Gandhi.M.SG = LOC.ON self.F.SG film.F.SG

The kps do not, however, behave like adjuncts, either. The examples in (40) show that multiple locative kps modifying a picture noun render the resulting np ungrammatical, representing strong evidence that the locative kp is in fact an argument selected by the picture noun; there can be at most one locative argument expressed with picture nouns.

(40) a. *mantiq = par ʃiʃa = par kitab
    logic.M.SG = LOC.ON teaching.F.SG = LOC.ON book.F.SG
b. *pyar = par nafrat = par gana
    love.M.SG = LOC.ON hate.F.SG = LOC.ON song.M.SG
c. *gandʰi = par hɪndʊstan = par film
    Gandhi.M.SG = LOC.ON India.M.SG = LOC.ON film.F.SG

When picture nouns are augmented by an extrinsic possessor (see §3.5.2.3), it is the extrinsic possessor subj that binds apna ‘self’ (41).

(41) meri, apni, gandʰi = par film
    PRON.POSS.F.SG self.F.SG Gandhi.M.SG = LOC.ON film.F.SG
    ‘my own film on Gandhi’
Picture nouns enforce case restrictions on the locative KP: I was not able to find a picture noun that may occur with a non-genitive argument marked by a case clitic other than locative par ‘on’. This is further motivation for assuming a semantically narrow OBLθ GF for the locative KP. The examples in (42) illustrate.

(42)  a. mantiq = par /* = mē */ = tak /* = se */ = ∅ logic.M.SG = LOC.ON /* = LOC.IN */ = LOC.TOWARDS */ = INST */ = LOC.TO */ = ko kītab
*/ = LOC.TO book.F.SG
‘the book about logic’

b. gandr = par /* = mē */ = tak /* = se Gandhi.M.SG = LOC.ON /* = LOC.IN */ = LOC.TOWARDS */ = INST */ = ko film
*/ = LOC.TO */ = LOC.TO film.F.SG
‘the film about Gandhi’

It comes as no surprise that picture nouns also occur without the locative argument (43). Seeing how the head noun puts tight constraints on the case as well as the thematic role of its argument, it can safely be assumed that the respective arguments are part of the lexical argument structure of picture nouns, but may be suppressed; this goes against the alternative view which would assume that picture nouns are augmented to predicate the OBLθ GF, in a fashion similar to the one discussed in §3.5.2.3. For similar accounts see, e.g., Zagona (2002) for Spanish or Dobrovie-Sorin (2003) for Hebrew (among others).

(43)  a. nina = ne kal accʰi film dekʰ-i
Nina.F.SG = ERG yesterday good.F.SG film.F.SG.NOM see-PERF.F.SG
‘Nina saw a good film yesterday.’

b. ram = ne dɪljɑsp kītab xɑrid-i
‘Ram bought an interesting book.’

Since “picture” common nouns universally instantiate a theme argument in their OBLθ GF, I will call the final GF to be used in my analysis and implementation OBL-TH.

5.3.2.3 Event Nouns

Event nouns that are at the same time picture nouns occur with locative arguments; examples are in (44a–c). There are, however, other event nouns that occur with instrumental arguments; examples are shown in (44d–e). Again, to test for the GF status of the locative or instrumental KP, one can insert the apna ‘self’ reflexive; as the examples in (45) show, the reflexive cannot be bound by the nominal marked with the locative case. Multiple arguments carrying the same case marker are ungrammatical here, as well.

5It is important not to confuse the OBL-TH function with the “meta-function” OBLθ GF, the former being a specific semantic instantiation of the latter.
5.3. THE LOCATIVE & INSTRUMENTAL IN NPS

(44)  a. salamti = par (*barda(t = par) barifing
security.F.SG = LOC.ON tolerance.M.SG = LOC.ON briefing.F.SG
‘briefing on security’ THEME Raza and Ahmed, 2011a
b. ʃetan = par (*xuda = par) lanat
satan.M.SG = LOC.ON god.M.SG = LOC.ON curse.M.SG
‘curse on satan’ THEME

c. ʃetan = par (*xuda = par) lanat
satan.M.SG = LOC.ON god.M.SG = LOC.ON curse.M.SG
‘curse on satan’ THEME

d. ʃetan = par (*xuda = par) lanat
satan.M.SG = LOC.ON god.M.SG = LOC.ON curse.M.SG
‘curse on satan’ THEME

e. film = par (*kɪtab = par) bahas
film.F.SG = LOC.ON book.F.SG = LOC.ON discussion.F.SG
‘discussion about the film’ THEME

Of course, event nouns may be further modified by genitive phrases, which do not map
to OBLθ GF, but are SUBJ, OBJ or ADJUNCT as described in Chapter §3. An example with
the noun mutalbah ‘demand’ is in (46).

(45)  a. * salamti, = par apni, barifing
security.F.SG = LOC.ON self.F.SG briefing.F.SG
b. * ʃetan, = par apni, lanat
satan.M.SG = LOC.ON self.F.SG curse.M.SG
c. * film, = par apni, bahas
film.F.SG = LOC.ON self.F.SG discussion.F.SG
d. * armi cif, = se apna, mutalbah
army.F.SG = INST self.F.SG demand.M.SG

LI arguments of event nouns may be suppressed, i.e., not realized in the syntax (47).

(46)  mera armi cif = se apna mutalbah
‘my own demand to the army chief(,”)

(47)  a. barifing bahot tafsili tʰ-i
briefing.F.SG.NOM very detailed.F.SG be.PAST-F.SG
‘The briefing was very detailed.’

b. sadar = ne mutalbah kal son-a lekin
president.M.SG = ERG demand.M.SG.NOM yesterday hear-PERF.M.SG but
nahĩ man-a
not accept-PERF.M.SG
‘The president heard the demand yesterday but did not accept (it).’
5.3. THE LOCATIVE & INSTRUMENTAL IN NPS

Event nouns, just like common nouns, enforce case restrictions on the locative kp; a specific locative case marker, depending on the semantic role that is realized, is needed for the np to be felicitous. The examples in (48) illustrate. The generalization here is that theme arguments of event nouns that are also “picture nouns” receive locative case (i.e., the case marker par), while arguments realizing a source, recipient or stimulus will be realized with instrumental case. The precise thematic role realized must be specified in the lexical semantics of the head noun; for an LFG/XLE analysis, this entails that the lexical entries of the nominals must be stated together with information about case requirements on their arguments.

(48) a. salamti = par  /∗ = mē  /∗ = tak  /∗ = se
   security.F.SG = LOC.ON  /∗ = LOC.IN  /∗ = LOC.TOWARDS  /∗ = INST
   /∗ = ∅  /∗ = ko  barifeng
   /∗ = LOC.TO /∗ = LOC.TO briefing.F.SG
   ‘briefing on security’

   b. film = par  /∗ = mē  /∗ = tak  /∗ = se  /∗ = ∅
   film.F.SG = LOC.ON  /∗ = LOC.IN  /∗ = LOC.TOWARDS  /∗ = INST  /∗ = LOC.TO
   /∗ = ko  bahas
   /∗ = LOC.TO discussion.F.SG
   ‘discussion about the film’

Thus, event nouns that are picture nouns take an OBL-TH GF, instantiated from OBLθ GF; picture nouns that occur with source-like arguments take an OBL-SRC GF, instantiated from OBLθ GF. Both of these GFs are thematically restricted for realizing a specific thematic role, depending on the noun; the thematic relation is reflected in the case marking.

5.3.2.4 RELATIONAL NOUNS

The third class of nouns that occurs with li arguments is a class of relational nouns; examples are provided in (49). These are relational nouns that involve an abstract meaning; again, the reflexive binding diagnostic as defined in (162) involving apna ‘self’ is negative, which shows that the locative kp cannot be a SUBJ. That is, if apna is included in these NPS, the string is ungrammatical since apna is not bound.

(49) a. nina = par    (*apna)  bʰar̩̃sa
   Nina.F.SG = LOC.ON (self.M.SG) trust.M.SG
   ‘trust in Nina’          THEME

   b. ram = par    (*apna)  yaqin
   ‘belief in Ram’          THEME

   c. tivitɑr = par    (*apna)  rɑvɑya
   twitter.M.SG = LOC.ON (self.M.SG) attitude.M.SG
   ‘attitude on Twitter’     THEME
5.3. THE LOCATIVE & INSTRUMENTAL IN NPS

As expected, multiple se or par marked arguments are not grammatical:

(50) a. nina=par (*ram=par) bʰɑrɔsa
   Nina.F.SG = LOC.ON Ram.M.SG = LOC.ON trust.M.SG
   ‘trust in Nina’
   b. nadya=se (*ravi=se) pyar
   ‘love for Nadya’

Regarding case, it is again found that relational nouns enforce case form restrictions on their LI arguments (51). The choice between the two available case markers par and se depends on the lexical semantics of the head noun; namely, whether the head noun specifies a theme (obl-th) or a source (obl-src) argument. Naturally, arguments of relational nouns may be suppressed (52).

(51) a. nina=par /*=mẽ /*=tak /*=se /*=∅
   Nina.F.SG = LOC.ON /*=LOC.IN /*=LOC.TOWARDS /*=INST /*=LOC.TO
   /*=ko bʰɑrɔsa
   /*=LOC.TO briefing.F.SG
   ‘trust in Nina’
   b. moqaddamat=se /*=par /*=mẽ /*=tak
court.case.M.PL = INST /*=LOC.ON /*=LOC.IN /*=LOC.TOWARDS
   /*=∅ /*=ko istisna
   /*=LOC.TO /*=LOC.TO immunity.M.SG
   ‘immunity from court cases’

(52) istisna zɑrura he
   immunity.M.SG.NOM necessary.M.SG be.PRES.3.SG
   ‘Immunity is important.’

Relational nouns may also be modified by genitive phrases as described in Chapter §. An example with the noun bʰɑrɔsa ‘trust’ is in (53).

(53) ram=ka armi cif=par apna bʰɑrɔsa
   ‘Ram’s own trust in the army chief’
Abstract relational nouns, just like their counterparts in other languages, do not have plural forms. Thus, they must not be treated as count nouns.

5.3.3 Pro-Drop or Argument Suppression?

The same reasons that hold against a pro-drop analysis of genitive arguments (see §3.5.3) also hold against a pro-drop analysis for li arguments; here, the premise is even harder to confirm since li arguments never respond to reflexive binding. Moreover, as already mentioned elsewhere (§3.5.3), agreement morphology is never found on the nominal head, neither in presence nor in absence of the argument. Therefore, I propose that when the li arguments are not realized, they are existentially bound by the lexical semantics of the noun, in a way very similar to what is proposed by Barker (1995); after all, trust and belief are concepts that invariably involve some entity they are directed towards, a book or a film cannot be thought of without depicting something, and a demand is always directed towards someone.

5.3.4 Arguments or Adjuncts?

The reasons for an analysis in terms of arguments instead of adjuncts, given for genitives in §3.5.4, also pertain to the li kps; here, though, the evidence for the argument status of the phrases is even stronger, given the case marking constraints discussed above. Needham and Toivonen (2011) acknowledge that fixed case or prepositional constraints are useful for identifying arguments.

In the LFG tradition, the OBLθ GF is generally used for non-objective, thematically-restricted arguments (Bresnan, 2001, Bresnan and Kanerva, 1989, Dalrymple, 2001, see also the discussion in §2.4.4). OBLθs occur with prepositional marking in English, but may also be case-marked in other languages; languages of the latter type are, for example, the Australian languages discussed by Nordlinger (1998), as well as Hindi/Urdu (Butt and King, 2004b). Within OBLθ, the theory further distinguishes semantically-marked ones as well as idiosyncratically-marked ones. For semantically-marked OBLθ, the preposition or case marker used is meaningful, as it indicates the semantic role of the oblique. A standard example given in the literature is the to-phrase in English sentences such as (54a), where the preposition to indicates that the oblique is a goal (OBL-GOAL). Similarly, the preposition from in (54b) suggests that the oblique here is a source (OBL-SRC).

(54)  a. Hans gave the book to Steffi.
      b. Steffi took the book from Hans.

For idiosyncratically-marked obliques, the phrase is required to bear a specific case marker or preposition, which depends on the predicate that governs it. An example is the English verb rely (55a); here, the requirement that the oblique phrase must occur with the preposition on must be part of the lexical entry of rely. Needham and Toivonen (2011)
further give the example in (55b), where the nominal predicate trust constrains the form of the preposition on the complement PP. Bresnan (1982a) suggests that in such cases, a special FORM feature, specified by the predicate, constrains the form of the preposition on the OBL∅. Butt et al. (1999b) provide some more discussion of idiosyncratically-marked obliques.

(55)  
a. Hans relied on/*to/*about/*from/*with Luise.

b. trust in/*on/*about/*from/*with Carl

(adapted from Needham and Toivonen, 2011)

The Hindi/Urdu li nominal arguments reviewed here must be treated as semantically-marked obliques. It is found that the case marker on the argument is meaningful: it indicates the semantic role of the oblique (source or theme). For an analysis in terms of idiosyncratic/quirky case marking, one would have to show that different predicates require different (lexically-fixed) forms of case markers on li arguments; this is not the case: I have not found a noun that marks a theme argument not using par ‘on’, and I am not aware of nouns that mark sources not with se, but with some other case marker.

More recently, the status of semantically-marked restricted obliques as arguments has been challenged, though: Zaenen and Crouch (2009) argue that those OBL∅s that are semantically marked should be merged with the ADJUNCT GF. Arguing from a computational perspective in terms of the English ParGram grammar, they state that the classic tests for argumenthood do not work insofar as to clearly distinguish adjuncts from OBL∅, and that the uniqueness criterion as discussed by Bresnan (1982b) is problematic.

Zaenen and Crouch (2009) provide the examples in (56–58). In (56), on you must be analyzed as an argument, which implies that on your kindness must be seen as a type of parenthetical; here, the preposition assignment is mediated by the verbal predicate, hence (56) must be analyzed as an instance of an idiosyncratic OBL∅. On the other hand, one can analyze (57) as involving two separate locative adjuncts. Such an approach runs into issues with respect to sentences such as (58), where the PPs could be analyzed as three distinct directional adjuncts, as three independent semantically-marked OBL∅s (i.e., along the lines of OBL-SRC, OBL-PATH and OBL-GOAL), or as any mixture of the above, thus yielding multiple analyses.

(56)  I count on you, on your kindness. (Zaenen and Crouch, 2009, p. 646)

(57) He lives in France, in a small village. (Zaenen and Crouch, 2009, p. 646)

(58) He drove from Paris to Venice via Milan. (Zaenen and Crouch, 2009, p. 646)

In the English ParGram grammar, where a verb such as drive also features simple transitive and intransitive subcategorization frames to cover sentences such as (59–60), it was chosen to limit possible ambiguities by assuming only a single OBL function, which naturally is subject to uniqueness. The English grammar further makes use of OT marks in XLE (§2.5.3.10) in order to punish the double adjunct analysis. However, Zaenen
and Crouch (2009) note that the possibilities are still multiplied if local ADJUNCT (NADJUNCT) attachments to nominals are taken into account; e.g., for (61a), the PP to the school could also be attached to the noun house, so that one is in fact left with the three analyses in (61b) for (61a).

(59) John drove.  
(Zaenen and Crouch, 2009, p. 648)

(60) John drove the car.

(61) a. John drove from the house to the school.  
(Zaenen and Crouch, 2009, p. 648)

b. OBL, ADJUNCT; ADJUNCT, OBL; OBL, NADJUNCT

Even more local attachments are possible in (62). The system of OT marks currently implemented in the English grammar prefers the OBL analysis over the ADJUNCT analysis for each PP in (62), but there are no OT marks preferring/punishing the local ADJUNCT attachment of the PPs to preceding nouns. The result is an ambiguous parse as shown in (63); the wanted analyses for the sentence in (62) corresponding to the ones in (61b) for (61a) have been dispreferred in favor of the local attachment of the PPs to nouns.

(62) John drove the car from the house to the school.  
(Zaenen and Crouch, 2009, p. 648)

(63) OBL, NADJUNCT; NADJUNCT, OBL; NADJUNCT, NADJUNCT

The solution proposed by Zaenen and Crouch (2009) is to treat semantically-marked OBL$_8$s as ADJUNCTs, since the distinction between the two cannot be drawn determinately, and abolish the notion of semantically-marked OBL$_8$s for LFG parsing altogether. They argue that this would result in a reduction of OBJ/ADJUNCT ambiguity as well as a gain in parsing performance. The OT marks controlling the ambiguity could be simplified, and lexical entry maintenance would be eased, since ADJUNCTs are never listed there.

The issue described by Zaenen and Crouch (2009) is not immediately relevant to the data discussed here. The focus of Zaenen and Crouch (2009) is on the distinction between verbal obliques vs. nominal adjuncts; in Hindi/Urdu, however, the locative and instrumental phrases never occur as nominal adjuncts. A possibly similar issue could be raised with respect to the distinction between nominal (LI) arguments and verbal adjuncts in Hindi/Urdu; consider the example in (64), where the locative phrase could either be interpreted as a nominal argument of kitab ‘book’ or as a clausal adjunct.

(64) nina = ne dara$x$ = par kitab par$^b$.i
‘Nina read a book on the tree.’

However, in such cases, the adjunct reading can easily be ruled out by dispreferring adjuncts in general, as is done in the English as well as the Urdu ParGram grammars;
since there is no possibility for the kp *daraxt=par* to be a nominal adjunct, one is left with a single optimal analysis in terms of a nominal oblique.

Finally, following the suggestions made by Zaenen and Crouch (2009) would mean to give up the intuitive appeal of an analysis of semantically-marked OBLₜₘs as arguments, given the uniqueness criterion that they are subject to. Also, as it stands, the issues discussed in Zaenen and Crouch (2009) with respect to English obliques seem to derive mainly from issues revolving around the correct setting of OT marks in the English ParGram grammar, which I am not convinced is sufficient motivation for changing the status of OBLₜₘs, neither in theory nor in grammar development practice.

### 5.3.5 Summary

To summarize then, there are three distinct sets of Hindi/Urdu nouns that may take LI arguments: a set of common (“picture”) nouns, a set of event nouns as well as a set of (abstract) relational nouns. With respect to the reflexive binding diagnostic, it can be stated that the arguments do not behave like SUBJS; this, together with the fact that the LI arguments uniformly realize a specific thematic role depending on the noun (locatives realize a theme, instrumentals realize a source or a related role), leads to the conclusion that they must receive an analysis in terms of OBLₜₘ. I will present my LFG/XLE analysis of LI arguments of nouns in §6.5.2.

The properties of the arguments are summarized in Table 5.2, in relation to the types of nouns that they occur with. The *Selection?* column specifies whether the head noun intrinsically selects the LI as an argument. The *Binds apna ‘self’* column specifies the behavior of the LI arguments with respect to binding the reflexive *apna ‘self’*, while the last column titled *GF* indicates the proposed grammatical function.

<table>
<thead>
<tr>
<th>Noun type</th>
<th>LI type</th>
<th>Amount</th>
<th>Selection?</th>
<th>Binds <em>apna ‘self’</em></th>
<th>GF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common count nouns (“picture nouns”)</td>
<td>Theme</td>
<td>0–1</td>
<td>✓</td>
<td>x</td>
<td>OBL-TH</td>
</tr>
<tr>
<td>Event (“picture nouns”)</td>
<td>Theme</td>
<td>0–1</td>
<td>✓</td>
<td>x</td>
<td>OBL-TH</td>
</tr>
<tr>
<td>Event (non-“picture nouns”)</td>
<td>Source</td>
<td>0–1</td>
<td>✓</td>
<td>x</td>
<td>OBL-SRC</td>
</tr>
<tr>
<td>Relational</td>
<td>Theme/Source</td>
<td>0–1</td>
<td>✓</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Locative/instrumental KP types in the Hindi/Urdu NP

### 5.4 LI SCRAMBLING

In §3.6, I have discussed the properties of genitive scrambling: the displacement of genitives out of the NP they are licensed in. Raza and Ahmed (2011a) present several
examples showing that LI arguments may be scrambled, as well; I refer to this process as LI scrambling. Raza and Ahmed (2011a) confine their description to arguments that are scrambled within nominal domains. In this section, I briefly discuss the data discussed in Raza and Ahmed (2011a); moreover, I show that LI arguments, like genitives, may be scrambled throughout clauses, further away from their heads, subject to the same syntactic constraints as genitive arguments.

5.4.1 General Description

I will briefly discuss the examples given by Raza and Ahmed (2011a) for NP-internal scrambling, then move on to scrambling out of NPs. In the example in (65a), the following structure is found. The adjective hasl ‘possessed’ takes a dative-marked argument sadar = ko ‘by the president’, while the overall head of the complex NP istisna ‘immunity’ takes an instrumental-marked source argument, muqaddamat = se ‘from court cases’. In the example in (65a), the instrumental argument occurs in its canonical position to the left of its head nominal. In (65b), on the other hand, the instrumental occurs at the beginning of the NP, separated from its head. In (65c), it occurs in the second position, preceded only by the argument of the adjective. Raza and Ahmed (2011a) state that the order in (65a) is the canonical order, where the arguments occur next to their heads; but the orders in (65b–c) are also found frequently. The meaning is preserved in each order.

(65) a. [sadar = ko hasl muqaddamat = se aini
   president.M.SG = DAT possessed court-case.M.PL = INST constitutional
   istisna]NP
   immunity.M.SG
   ‘constitutional immunity from court cases possessed by the president’

   Raza and Ahmed (2011a)

b. muqaddamat = se sadar = ko hasl aini istisna
   Raza and Ahmed (2011a)

c. sadar = ko muqaddamat = se hasl aini istisna
   Raza and Ahmed (2011a)

The example in (65) shows that locative arguments may be scrambled further to the left, away from their heads. The example is a complex NP; both the locative argument and its head are within the boundaries of a single NP. The examples in (66–68), on the other hand, indicate that the locative argument may also occur outside of the NP that contains its head. Here, in the a. examples, the canonical word order is given, with the argument to the immediate left of the head noun (or separated from it only by other NP-internal material, such as adjectives (67a) or quantifiers (68a)). In the b. and c. examples, the LI arguments are separated from their heads by material that is clearly

---

\(^6I\) have not discussed argument-taking adjectives; for a discussion of such adjectives in Urdu, see Raza (2011).

\(^7I\)n fact, Raza and Ahmed (2011a) note that the non-canonical ordering in (65c) is the one most frequently found in news corpora.
at clausal level (the clausal adjunct do sal=ke pele ‘two years ago’ in the b. examples, the predicate adjective zaruri ‘necessary’ in the c. examples). Note also that while it was seen in §5.3.1.2 that LI arguments need to precede their heads in NPs, this order need not be preserved under scrambling; in the c. examples of (66–68), the heads precede their arguments. I will refer to those LI arguments that precede their heads in a clause as non-inverted scrambled LI arguments, and those that follow their heads as inverted scrambled LI arguments (in agreement with the terminology used by Fanselow and Féry (2006)).

(66)  
\begin{align*}
a. & \text{ do sal=ke pele ram=par bʰarɔsa} \\
   & \text{two year.M.PL = GEN.M.SG.OBL before Ram.M.SG = LOC.ON trust.M.SG.NOM} \\
   & \text{zaruri tʰ-a} \\
   & \text{necessary.M.SG be.PAST-M.SG} \\
   & \text{‘Two years ago, trust in Ram was necessary.’} \\

b. & \text{ ram=par do sal=ke pele bʰarɔsa zaruri tʰ-a} \\

c. & \text{ do sal=ke pele bʰarɔsa zaruri ram=par tʰ-a} \\

(67)  
\begin{align*}
a. & \text{ do sal=ke pele salamti=par tafsili} \\
   & \text{two year.M.PL = GEN.M.SG.OBL before security.F.SG = LOC.ON detailed.F.PL} \\
   & \text{barifing zaruri tʰ-i} \\
   & \text{briefing.F.PL.NOM necessary.F.PL be.PAST-F.PL} \\
   & \text{‘Two years ago, detailed briefings on security were necessary.’} \\

b. & \text{ salamti=par do sal=ke pele tafsili barifing zaruri tʰ-i} \\

c. & \text{ do sal=ke pele tafsili barifing zaruri salamti=par tʰ-i} \\

(68)  
\begin{align*}
a. & \text{ do sal=ke pele muqaddamat=se} \\
   & \text{two year.M.PL = GEN.M.SG.OBL before court-case.M.PL = INST} \\
   & \text{aini ɪstɪsna zaruri tʰ-i} \\
   & \text{constitutional immunity.M.SG.NOM necessary.F.PL be.PAST-F.PL} \\
   & \text{‘Two years before, constitutional immunity from court cases was necessary.’} \\

b. & \text{ muqaddamat=se do sal=ke pele aini ɪstɪsna zaruri tʰ-i} \\

c. & \text{ do sal=ke pele aini ɪstɪsna zaruri muqaddamat=se tʰ-i} \\

In (66–68), the LI arguments are scrambled out of subjects. (69–71) show that it is also possible to scramble them from other GFS. (69b) and (70b) involve locative KPs scrambled from objects; in (71b), the locative KP is extracted from a locative oblique. This goes to show that locative KPs may be extracted from different GFS in Hindi/Urdu.

(69)  
\begin{align*}
a. & \text{ nɪda=ne mɑntɪq=par ek kitab xɑrid-i} \\
   & \text{Nida.F.SG = ERG logic.M.SG = LOC.ON one book.F.SG.NOM buy-PERF.F.SG} \\
   & \text{he be.PRES.3.SG} \\

b. & \text{ mɑntɪq=par nɪda=ne ek kitab xɑrid-i he} \\

c. & \text{ nɪda=ne ek kitab xɑrid-i mɑntɪq=par he}
5.4. LI SCRAMBLING

(70) a. do sal=ke pele nina=ne ram=par
two year.M.PL=GEN.M.SG.OBL before Nina.F.SG = ERG Ram.M.SG = LOC.ON
bʰɑrɔse=ko kʰo-ya
trust.M.SG.OBL = ACC lose-PERF.M.SG
‘Two years ago, Nina lost the trust in Ram.’

b. ram=par do sal=ke pele nina=ne bʰɑrɔse=ko kʰo-ya

(71) a. nɪda=ne pani mantɪq = par kitab = par
ḍal-a
pour-PERF.M.SG
‘Nina poured water on the book on logic.’

b. mantɪq = par nɪda = ne pani kitab = par ḍal-a

Note that the predicate-argument structure of the b. examples of (66–71) is clear without ambiguity: The locative arguments can only be analyzed as arguments of their respective heads (bʰɑrɔse ‘trust’, bɑrifɪng ‘briefing’ and kɪtab ‘book’). Given the above account of semantic case marking on the nominal arguments, the scrambling possibilities are expected: the predicate-argument structure can be reconstructed precisely because of the case requirements on the nominal arguments. This is in line with typological accounts of discontinuous noun phrases (Fanselow and Féry, 2006).

5.4.2 PREFERENCES AND CONSTRAINTS

The goal of this section is to show that li scrambling is subject to the same constraints like genitive scrambling concerning local attachment preference, case, complement clauses, adjuncts as well as scrambling from deeper within.

5.4.2.1 LOCAL ATTACHMENTS are Preferred

In general, if the case marker on the argument permits multiple embeddings in a clause, a local attachment is preferred. The corresponding preference with respect to genitive phrases is discussed in §3.6.2.1. Consider the example in (72). Given the possibilities on scrambling as discussed above, the locative argument here could modify both kitab ‘book’ locally as well as rɪsala ‘magazine’ as a scrambled argument. The local attachment is in fact preferred over the non-local one.

(72) ram = ne mantɪq = par kitab = ko
rɪsala = par rɑkʰ-a
magazine.M.SG = LOC.ON put-PERF.M.SG
‘Ram put the book on logic on the magazine.’
preferred over
‘Ram put the book on the magazine on logic.’
5.4. SCRAMBLING AND CASE

As with scrambled genitives, inverted scrambled li arguments are ungrammatical if they are scrambled out of an overtly case-marked kp. The examples in (73) as well as (74) (repeated from (28)) illustrate.

(73) a. do sal = ke pele nina = ne ram = par
two year.M.PL = GEN.M.SG.OBL before Nina.F.SG = ERG Ram.M.SG = LOC.ON
b^harse = ko k^h-o-ya
trust.M.SG.OBL = ACC lose-PERF.M.SG
Two years ago, Nina lost the trust in Ram.’

b. * do sal = ke pele nina = ne
   two year.M.PL = GEN.M.SG.OBL before Nina.F.SG = ERG
   b^harse = ko ram = par k^h-o-ya
   Ram.M.SG = LOC.ON trust.M.SG.OBL = ACC lose-PERF.M.SG

(74) a. ram = par caude b^harse = ne nadya = ko
taqat di-ya
   strength.F.SG.NOM give-PERF.F.SG
   ‘Broad trust on Ram gave Nadya strength.’

b. * caude b^harse = ne nadya = ko taqat
   ram = par di-ya
   Ram.M.SG = LOC.ON give-PERF.F.SG

5.4.2.3 SCRAMBLING FROM COMPLEMENT CLAUSES

LI scrambling out of finite clauses is not possible; see the example in (75). The arguments can, however, be scrambled out of non-finite clauses, e.g., as in (76).

(75) * ram = par ila = ne soc-a kih [nina = ne
   Ram.M.SG = LOC.ON ila.F.SG = ERG think-PERF.M.SG that Nina.F.SG = ERG
   b^harse = ko k^h-o-ya]
   trust.M.SG.OBL = ACC see-PERF.F.SG lose-PERF.M.SG

(76) mantiq = par ram ktab xarid sak-a
   logic.M.SG = LOC.ON Ram.M.SG.NOM book.F.SG.NOM buy can-PERF.M.SG
   ‘On logic, Ram could buy a book.’

5.4.2.4 NO SCRAMBLING OUT OF ADJUNCTS

In §3.6.2.4, it was seen that it is not possible to scramble genitive modifiers out of adjuncts is not possible. The examples in (77–78) show that the same generalization holds for locative and instrumental arguments, as well. The locative argument in (77b) occurs inside an adjunct; the argument cannot be scrambled out of the adjunct (77b). (78) is a corresponding example involving an instrumental argument.
5.4. LI SCRAMBLING

Locative and instrumental arguments that are embedded in some path deeper inside a GF may not be scrambled out of their path. Recall from §5.3.1.3 that nouns selecting LI arguments may be stacked on top of one another, which gives rise to complex NPs such as the one in (79a) (repeated from (31c)). When such complex NPs occur in sentences as in (79b), the deeply embedded LI argument (here, the argument is an OBL-SRC embedded in the path (↑ OBJ OBL-TH) cannot be scrambled out of its NP, as seen in (79c).

(79) a. mʊqɑddɑmat=se ɪstɪsne=par diljasp ktab
‘the interesting book about immunity from court cases’

b. mē = ne mʊqɑddɑmat=se ɪstɪsne=par
PRON.1.SG.OBL = ERG court-case.M.PL = INST immunity.M.SG.OBL = LOC.ON
diljasp ktab kal xarid-i
interesting book.F.SG.NOM yesterday buy-PERF.F.SG
‘I bought an interesting book about immunity from court cases yesterday.’

c. * mē = ne ɪstɪsne=par diljasp
PRON.1.SG.OBL = ERG immunity.M.SG.OBL = LOC.ON interesting
ktab kal mʊqɑddɑmat=se xarid-i

Another example is given in (80). (80a) is a complex NP where tabahi ‘destruction’ takes an instrumental argument, and natija ‘result’ is the overall head of the NP, taking
5.5. THE LOCATIVE IN LOCATIVE CLAUSES

a genitive that is a subject. In (80b), the complex NP from (80a) is an object marked by the accusative case. (80c) is an example of a non-inverted scrambled LI argument. The example is not ruled out by the case constraint discussed above in §5.4.2.2 (which only affects inverted scrambled arguments); nevertheless, it is ungrammatical. The reason is that it is embedded not in a main clause GF, but inside a deeper path (↑ OBJ SUBJ).

(80) a. tsɛlab = se tabahi = ka natija
    flood.M.SG = INST destruction.F.SG = GEN.M.SG result.M.SG
    ‘the result of the destruction due to the flood’

    b. nina = ne tsɛlab = se tabahi = ke
    Nina.F.SG = ERG flood.M.SG = INST destruction.F.SG = GEN.M.SG.OBL
    natije = ko dekʰ-a
    result.M.SG.OBL = ACC see-PERF.M.SG
    ‘Nina saw the result of the destruction due to the flood.’

    c. * tsɛlab = se nina = ne tabahi = ke
    flood.M.SG = INST Nina.F.SG = ERG destruction.F.SG = GEN.M.SG.OBL
    natije = ko dekʰ-a
    result.M.SG.OBL = ACC see-PERF.M.SG

5.4.3 SUMMARY

The preferences and constraints in LI argument scrambling are the same as in genitive scrambling. In general, local attachments are preferred over scrambled configurations if multiple readings are possible. Scrambling from complement clauses is only possible if the complement clause is non-finite. If the argument is scrambled from a case-marked nominal, only non-inverted scrambling is possible.

5.5 THE LOCATIVE IN LOCATIVE CLAUSES

To locate a nominal in space, Hindi/Urdu uses the copula verb ho ‘be’ in combination with a nominal marked with the locative case, which marks the location, and a second nominal, which is the nominal to be located (Kachru 2006, Schmidt 1999). I refer to such sentences as locative clauses. Some examples are in (81). The specific locative case marker is chosen to reflect the intended semantics of the locative relation. In addition, the nature of the location may be concrete (81a–d) or abstract (81e–g). Without the verbal predicate (i.e., bare NPs with a location), the strings are ungrammatical.

(81) a. mez = par cuha *(he)
    table.F.SG = LOC.ON rat.M.SG.NOM be.PRES.3.SG
    ‘On the table is a rat.’ = ‘There is a rat on the table.’

---

8 natija ‘result’ is a relational noun of the type discussed in §3.5.2.5.
5.5. THE LOCATIVE IN LOCATIVE CLAUSES

(b) mez = par kitab *(he)
table.F.SG = LOC.ON book.F.SG.NOM be.PRES.3.SG
‘On the table is a book.’ = ‘There is a book on the table.’

c. makan = mē kutta *(he)
house.M.SG = LOC.IN dog.M.SG.NOM be.PRES.3.SG
‘In the house is a dog.’ = ‘There is a dog in the house.’

d. makan = mē admi *(he)
house.M.SG = LOC.IN person.M.SG.NOM be.PRES.3.SG
‘In the house is a person.’ = ‘There is a person in the house.’

e. nina = mē pyar *(he)
Nina.F.SG = LOC.IN love.M.SG.NOM be.PRES.3.SG
‘In Nina is love.’ = ‘There is love in Nina.’

f. ram = mē mamta *(he)
Nina.F.SG = LOC.IN affection.F.SG.NOM be.PRES.3.SG
‘In Ram is affection.’ = ‘There is affection in Ram.’

g. yasin = mē bʰai *(he)
Nina.F.SG = LOC.IN fear.M.SG.NOM be.PRES.3.SG
‘In Yassin is fear.’ = ‘There is fear in Yassin.’

The ungrammaticality of NPs without the copula verb in (81) is strong evidence that these nominals by themselves cannot license a location, but that it is the copula ho ‘be’ that enables the expression of a location. I take ho ‘be’ in (81) to not be an existential verb when used in locative clauses; instead, it must be seen as a copula verb with two distinct grammatical functions in its subcategorization frame. More evidence for this claim will be presented throughout this chapter. A crucial difference between the L1 arguments in NPs discussed in §5.3 and the locatives in locative clauses relates to their thematic role.

Locative clauses are purely stative predicates: the copula verb ho ‘be’ is essentially used to “link” a nominal to its location (Kachru, 2006; Sulger, 2012). As a direct consequence, not all of the locative case clitics identified in Table 5.1 occur in locative clauses, since some of them require the predicate to be non-stative. The locative case markers ًَ ‘to’, which is used to mark directions of movement, as well as tak ‘towards’ and se ‘from’ imply an active predicate and are never used in stative predicates. Thus, they never occur in locative clauses. The examples in (82) illustrate.

(82) a. * tol abeb = ko admi he
Tel Aviv = LOC.TO person.M.SG.NOM be.PRES.3.SG

b. * mez = tak kutta he
table.F.SG = LOC.TOWARDS dog.M.SG.NOM be.PRES.3.SG

c. * mez = se kitab he
table.F.SG = LOC.ON book.F.SG.NOM be.PRES.3.SG
What complicates matters is that the complex NPs reviewed in §5.3 above also occur in verbal predication, as in (83). There has been some confusion about such sentences; in particular, they were mistaken for locative clauses involving a copula verb (Abbas and Raza, 2014; Raza, 2011). I assume this is partially due to the fact that on the surface, there is nothing in the morphosyntax that hints at a predicational difference. I will refer to such clauses as “pseudo”-locative clauses.

(83) a. mɑntɪq=pɑr
logic.M.SG = LOC.ON kitab.F.SG.NOM
book be.PRES.3.SG
‘There is a book on logic.’

b. nadya=se
Nadya.F.SG = INST pyar
love.M.SG.NOM
be.PRES.3.SG
‘There is love for Nadya.’

Below, I will review the constituent as well as functional properties of locative clauses. In particular, I present clear evidence that sets apart the locative clauses in (81) from the predications in (83), the latter of which have to be treated as purely intransitive (i.e., like the possessive clauses surveyed in §3.7).

5.5.1 CONSTITUENT PROPERTIES

This section discusses the constituent properties of locative clauses. As mentioned above, locative or instrumental KPs never occur inside NPs in Hindi/Urdu unless licensed by their nominal heads. Thus, locative or instrumental phrases never attach to noun phrases in the c-structure. Instead, as evidenced by the survey in the below sections, tests in §5.5.1.2, a flat c-structure for locative clauses must be assumed.

5.5.1.1 COORDINATION

One may freely coordinate either the locative KP or the located nominal NP in locative clauses. The examples in (84–85) are all grammatical. In addition to the coordination pattern of the locative KP in (85), one may decide to coordinate the NP that is embedded in the locative KP as in (86). Thus, from the point of view of coordination, the locative KP and the located nominal NP behave like standard KPs/NPs in the language.

(84) a. [makan=mẽ]kp
house.M.SG = LOC.IN
[[cuha]NP
rat.M.SG.NOM
or
[kutta]NP
dog.M.SG.NOM
be.PRES.3.SG
‘In the house is a rat or a dog.’ = ‘There is a rat or a dog in the house.’

b. [ram=mẽ]kp
Ram.M.SG = LOC.IN
[[mamta]NP
affection.F.SG.NOM
or
[bʰai]NP
fear.M.SG.NOM
be.PRES.3.PL
‘In Ram are affection and fear.’ = ‘There are affection and fear in Ram.’
5.5. THE LOCATIVE IN LOCATIVE CLAUSES

(85) a. \[
\text{[mez} = \text{par}]_{\text{KP}} \quad \text{or} \quad \text{[almari} = \text{mê}]_{\text{KP}} \quad \text{[kitabê]_{NP}} \\
\text{table.F.SG} = \text{LOC.ON and cupboard.F.SG} = \text{LOC.IN book.F.PL.NOM} \\
\text{hê} \\
\text{be.PRES.3.PL}
\]

‘On the table and in the cupboard are books.’ = ‘There are books on the table and in the cupboard.’

b. \[
\text{[ram} = \text{mê}]_{\text{KP}} \quad \text{or} \quad \text{[nina} = \text{mê}]_{\text{KP}} \quad \text{[mumta]_{NP}} \\
\text{Ram.M.SG} = \text{LOC.IN and Nina.F.SG} = \text{LOC.IN affection.F.SG.NOM} \\
\text{hê} \\
\text{be.PRES.3.SG}
\]

‘In Ram and in Nina is affection.’ = ‘There is affection in Ram and in Nina.’

Coordination of the KPs and NPs is also possible in pseudo-locative clauses. (87-90) are examples corresponding to the coordination patterns in (84-86). Interestingly, a sentence such as (87a) can be bracketed in either of two ways, with different results in meaning. The bracketing in (87b) involves a complex NP including a locative argument; the NP is coordinated with a simple NP. The bracketing in (87c), on the other hand, involves two NPs coordinated in an NP, which then takes a locative argument modifying both nouns. The former type of bracketing is not available for (84). (88) is an example similar to (87).

(87) a. \[
\text{mantiq} = \text{par} \quad \text{kitab} \quad \text{ya} \quad \text{risala} \quad \text{hê}
\]

b. \[
\text{[mantiq} = \text{par}]_{\text{KP}} \quad \text{[kitab]_{NP}} \quad \text{ya} \quad \text{[risala]_{NP}} \quad \text{hê} \\
\text{logic.M.SG} = \text{LOC.ON book.F.SG.NOM or magazine.M.SG.NOM be.PRES.3.SG} \\
\text{*‘On logic is a book or a magazine.’ = ‘There is a book or a magazine on logic.’} \\
\text{‘There is a book on logic or a magazine.’ (i.e., the magazine is not necessarily on logic)}
\]

c. \[
\text{[mantiq} = \text{par}]_{\text{KP}} \quad \text{[[kitab]_{NP}} \quad \text{ya} \quad \text{[risala]_{NP}}_{\text{NP}} \quad \text{hê} \\
\text{logic.M.SG} = \text{LOC.ON book.F.SG.NOM or magazine.M.SG.NOM be.PRES.3.SG} \\
\text{*‘There is a book on logic or a magazine.’} \\
\text{‘On logic is a book or a magazine.’ = ‘There is a book or a magazine on logic.’ (i.e., both are on logic)}
\]

(88) a. \[
\text{nadya} = \text{se} \quad \text{pyar} \quad \text{ya} \quad bʰai \quad \text{hê}
\]
5.5. THE LOCATIVE IN LOCATIVE CLAUSES

b. \[[nadya=se]\text{k}p \text{pyar}]_{\text{NP}} \text{ ya [bʰai]}_{\text{NP}} \text{ hé}
\text{Nadya.F.SG = INST love.M.SG.NOM or fear.F.SG.NOM be.PRES.3.SG}
*‘For Nadya is love or fear. = There is love for or fear of Nadya.’
‘There is love for Nadya or fear.’ (i.e., the fear is not necessarily of Nadya)

c. \[[nadya=se]\text{k}p \text{ [pyar]}_{\text{NP}} \text{ ya [bʰai]}_{\text{NP}} \text{ hé}
\text{Nadya.F.SG = INST love.M.SG.NOM or fear.F.SG.NOM be.PRES.3.SG}
*‘There is love for Nadya or fear.’
‘For Nadya is love or fear. = There is love for or fear of Nadya.’ (i.e., both
are related to Nadya)

\((89-90)\) parallel \((85-86)\) above, which is expected given the coordination patterns of kps
\((\text{Butt and King}, 2004b, \text{Kachru}, 2006, \text{Schmidt}, 1999)\).

\((89)\) a. \[[mɑntɪq=par]\text{k}p or [zɑban=par]\text{k}p \text{[kitabẽ]}_{\text{NP}}
\text{logic.M.SG = LOC.ON and language.F.SG = LOC.ON book.F.PL.NOM}
\text{ hé be.PRES.3.PL}
‘On logic and language are books.’ = ‘There are books on logic and lan-
guage.’

b. \[[[nadya=se]\text{k}p or [ram=se]\text{k}p \text{[pyar]}_{\text{NP}} \text{ hé}
\text{Nadya.F.SG = INST and Ram.M.SG = INST love.M.SG.NOM be.PRES.3.SG}
‘For Nadya and for Ram is love.’ = ‘There is love for Nadya and for Ram.’

\((90)\) a. \[[mɑntɪq\text{[}NP \text{ ya [zɑban=par]}_{\text{k}p \text{[kitab]}_{\text{NP}}
\text{logic.M.SG = LOC.ON and language.F.SG = LOC.ON book.F.SG.NOM}
\text{ hé be.PRES.3.SG}
‘On logic or language is a book.’ = ‘There is a book on logic or language.’

b. \[[[nadya]\text{NP or [ram]}_{\text{NP} = se}\text{k}p \text{[pyar]}_{\text{NP}} \text{ hé}
\text{Nadya.F.SG = INST and Ram.M.SG = INST love.M.SG.NOM be.PRES.3.SG}
‘For Nadya and Ram is love.’ = ‘There is love for Nadya and Ram.’

The coordination patterns, and the availability of bracketings/readings such as the ones
in \((87b)\) and \((88b)\) are in clear support of an analysis in terms of “pseudo”-locative
clauses as purely intransitive, on a par with possessive clauses.

5.5.1.2 LINEAR ORDER

The elements of the locative clause may be freely shuffled. The sentences in \((91b)\)
through \((91d)\) show permutations of the example in \((91a)\). The located nominal may
appear either sentence initially or not, and the same holds for the location. Further,
the copula verb ho ‘be’ may intervene between the two. All of the sentences in \((91)\)
are grammatical. Varying the word order, as in the case of possessive clauses \((3.7)\),
is driven by information-structural properties of the clauses; I briefly discuss these ef-
facts below in \(5.6\). The two most common word orders are the ones in \((91a)\) as well as
5.5. THE LOCATIVE IN LOCATIVE CLAUSES

(91) a. cuha mez = par he
cuha rat.M.SG.NOM mez = par
table.F.SG = LOC.ON be.PRES.3.SG
‘The rat is on the table.’

b. mez = par cuha he
cuha table.F.SG = LOC.ON
rat.M.SG.NOM = be.PRES.3.SG
‘On the table is a rat.’ = ‘There is a rat on the table.’

c. mez = par he cuha
cuha table.F.SG = LOC.ON
rat.M.SG.NOM be.PRES.3.SG
‘On the table is a rat.’ = ‘There is a rat on the table.’

d. cuha he mez = par
cuha rat.M.SG.NOM be.PRES.3.SG
table.F.SG = LOC.ON
‘The rat is on the table.’

In contrast to the genitive, there are restrictions on adjectival modification. Recall
that the order of genitive and adjectival/quantifying modifiers is free with respect to
one another in possessive clauses, as discussed in §3.7.1.2; the relevant examples are
(92–93), two of which are repeated in (92–93). In §3.7.1.2, I take this as motivation
for assuming a complex NP c-structure as indicated by the bracketing in the examples.

(92) [nili ram = ki gaṛi]NP he
blue.F.SG Ram.M.SG = GEN.F.SG
car.F.SG.NOM = be.PRES.3.SG
‘Ram has a blue car.’

(93) [do nadya = ke beṭe]NP he
two Nadya.F.SG = GEN.M.PL
son.M.PL.NOM = be.PRES.3.PL
‘Nadya has two sons.’

The same analysis cannot be maintained for locative clauses. The example in (94a),
corresponding to the adjectival modification in (92), is unacceptable in the reading There
are blue books on the table. That is, the adjective can only modify mez ‘table’, even
though, given the syncretism in feminine inflection, the morphology on the adjective
does not rule out either possibility. For the adjective to be able to modify kitabē ‘books’,

9The reason for these specific labels do not matter at this point; they will become clear in my analysis
of these clauses in Chapter 6.

10An analysis of (94a) involving adjectival scrambling is thus not acceptable for my informants; one
of them (a Hindi speaker) states that for nili ‘blue’ to be able to modify kitabē ‘books’, he would have to
use rather dramatic prosody and prefer the paraphrase in (94c). Adjectival scrambling is otherwise not
described in accounts of Hindi/Urdu scrambling or extraposition; cf. e.g. Kidwai (2000), Raza and Ahmed
(2011a).
one could utter (94b) in a discourse context that supports the prosody, involving the topicalization marker to.11

(94) a. nili mez = par kitabê hê
   blue.F.SG/PL = LOC.ON book.F.PL.NOM be.PRES.3.PL
   ‘On the table are blue books.’ = ‘There are blue books on the table.’
   ‘On the blue table are books.’ = ‘There are books on the blue table.’

b. nili = to mez = par tin kitabê th-i
   blue.F.SG/PL = TOP table.F.SG = LOC.ON three book.F.PL.NOM be.PAST-F.PL
   ‘As for blue (things), there were three books on the table.’

(95a), corresponding to the word order of (93), is ungrammatical in the reading where do ‘two’ modifies kitabê ‘books’. Here, it is even harder to come up with a prosodic configuration that is licit, since numerals such as do ‘two’ and other quantifiers do not lend themselves to the topicalization marker to.

(95) a. * do mezõ = par kitabê hê
   two table.F.PL.OBL = LOC.ON book.F.PL.NOM be.PRES.3.PL
   ‘On the tables are two books.’ = ‘There are two books on the tables.’

b. do mezõ = par kitabê hê
   two table.F.PL.OBL = LOC.ON book.F.PL.NOM be.PRES.3.PL
   ‘On the two tables are books.’ = ‘There are books on the two tables.’

All in all, together with the fact that locative kps may not occur inside nps (unless selected by the head nominal in the fashion discussed in §5.3), this is a strong indication that one cannot maintain a complex np analysis for locative clauses, but that one must assume clausal structures with two independent nominal nodes. On the other hand, “pseudo”-locative clauses do not exhibit the same behavior. Consider the sentence in (96), repeated from (26). The reading where the adjective modifies the locative argument is preferred (since it is the more local attachment), but the important observation is that hoʃyar ‘intelligent’ can also modify the overall head noun, bʰɑrɔsa ‘trust’, again hinting at a crucial structural difference between sentences like (96) and (94a).

(96) hoʃyar larke = par bʰɑrɔsa hê
   intelligent boy.M.SG.OBL = LOC.ON trust.M.SG.NOM be.PRES.3.SG
   ‘There is trust in the intelligent boy.’
   preferred over
   ‘There is intelligent trust in the boy.’

Various elements may intervene between the located nominal and the locative kp. Two examples in (97–98) illustrate. In (97), the sentence-level temporal adjunct kal ‘yesterday’ is added at various positions within the sentence. In (98), the adjunct bac-cô = ke par³-ne = ke liye ‘for the children to read’ is added, again modifying the clause as

11I thank Rajesh Bhatt for providing the examples illustrating the use of the topicalization marker to.
a whole. Such modifications are expected to be possible under the assumption that the locative and the located nominal are not embedded inside a single c-structural configuration.

(97) a. kal cuha mez = par tʰ-a yesterday rat.m.sg.nom table.f.sg = loc.on be.past-m.sg
‘Yesterday, the rat was on a/the table.’

b. cuha kal mez = par tʰ-a rat.m.sg.nom yesterday table.f.sg = loc.on be.past-m.sg
‘Yesterday, the rat was on a/the table.’

c. cuha mez = par kal tʰ-a rat.m.sg.nom table.f.sg = loc.on yesterday be.past-m.sg
‘Yesterday, the rat was on a/the table.’

d. mez = par kal cuha tʰ-a table.f.sg = loc.on yesterday rat.m.sg.nom be.past-m.sg
‘Yesterday, on the table was a rat.’ = ‘Yesterday, there was a rat on the table.’

(98) a. baccō = ke parʰ-ne = ke liye child.m.pl.obl = gen.m.sg.obl read-inf.m.sg.obl = gen.m.sg.obl for kitabē mez = par hē book.f.pl.nom table.f.sg = loc.on be.pres.3.pl
‘The books are on the table for reading by the children.’ = ‘The books are on the table for the children to read.’

b. kitabē baccō = ke book.f.pl.nom child.m.pl.obl = gen.m.sg.obl parʰ-ne = ke liye mez = par hē read-inf.m.sg.obl = gen.m.sg.obl for table.f.sg = loc.on be.pres.3.pl
‘The books are on the table for reading by the children.’ = ‘The books are on the table for the children to read.’

c. kitabē mez = par baccō = ke book.f.pl.nom table.f.sg = loc.on child.m.pl.obl = gen.m.sg.obl parʰ-ne = ke liye hē read-inf.m.sg.obl = gen.m.sg.obl for be.pres.3.pl
‘The books are on the table for reading by the children.’ = ‘The books are on the table for the children to read.’

d. mez = par baccō = ke table.f.sg = loc.on child.m.pl.obl = gen.m.sg.obl parʰ-ne = ke liye kitabē hē read-inf.m.sg.obl = gen.m.sg.obl for book.f.pl.nom be.pres.3.pl
‘On the table are books for reading by the children.’ = ‘There are books on the table for the children to read.’
5.5. THE LOCATIVE IN LOCATIVE CLAUSES

5.5.1.3  SUMMARY

To summarize, it was seen that the evidence from coordination as well as linear order support an analysis of a flat c-structure involving two separate nominal phrases for locative clauses. For “pseudo”-locative clauses, the data clearly point in a different direction: here, facts from linear order, adjectival modification as well as coordination suggest that what is found here is a complex NP, whose existence is predicated in an intransitive clause, much like the possessive clauses reviewed in §3.7.

5.5.2  FUNCTIONAL PROPERTIES

This section takes a detailed look at the functional properties of locative clauses, reviewing the subject tests identified by Mohanan (1994) which were introduced in §3.4. In particular, Mohanan’s claim that only those l-SUBJ that extend to non-spatial, abstract semantic fields may be realized as grammatical subjects is examined more closely. In the next subsections, the relevant tests are applied to both concrete, purely spatial uses of the locative and abstract, non-spatial uses. Furthermore, I examine both types of the locative clauses in turn — the canonical locative clause (CLC) as well as the inverted locative clause (ILC) — regarding the subjecthood of the nominals involved, dealing with the question of whether the subject stays the same when varying the word order. It turns out that the GF assignment switches with the word order variation: in the CLC (located nominal first, locative KP second), it is the located nominal that behaves as the SUBJ, while in the ILC (locative KP first, located nominal second), it is consistently the locative KP that behaves as the SUBJ.

5.5.2.1  REFLEXIVE BINDING

The apna ‘self’ diagnostic as defined in §3.8 is repeated in (99) — the only eligible antecedent of the reflexive apna ‘self’ is an L-SUBJ or a SUBJ in the reflexive’s minimal complete nucleus. Mohanan (1994) applies the diagnostic to test for the SUBJ in sentences such as (100a); in (100b), the locative-marked nominal is the only eligible antecedent of the reflexive, which means either that the locative KP is the grammatical subject (in the absence of any other binding possibilities), or that the sentence has no grammatical subject. To assume that the locative was merely an L-SUBJ (and not a SUBJ) would entail the possibility that the reflexive could be bound by the other nominal in the clause, manta ‘affection’. This is, however, not possible, which means that the locative is the SUBJ.

(99)  The reflexive apna ‘self’ must be bound by an L-SUBJ or SUBJ within its minimal complete nucleus (i.e., the smallest f-structure containing some predicate and a function SUBJ).
5.5. THE LOCATIVE IN LOCATIVE CLAUSES 214

(100) a. nina=mẽ bɑṛi mɑmta hɛ
    Nina.F.SG = LOC.IN much.F.SG affection.F.SG.NOM be.PRES.3.SG
    ‘In Nina is a lot of affection.’ = ‘There is a lot of affection in Nina.’ = ‘Nina has a lot of affection.’
    (Mohanan, 1994, p. 176)

b. nina=mẽ apni mɑsĩ=ke liye bɑṛi
    mɑmta hɛ
    affection.F.SG.NOM be.PRES.3.SG
    ‘Nina has a lot of affection for self’s aunt.’
    (Mohanan, 1994, p. 176)

The locative in (100b) is used in an abstract sense; the reflexive binding is not applied to concrete uses of the locative by Mohanan (1994). The examples in (101) make use of concrete spatial locations. Both examples in (102) are grammatical, and the reflexive is bound by the locative. The apna test is thus positive for both abstract and concrete spatial uses of the locative.

(101) a. mɑkan=mẽ mez hɛ
    house.M.SG = LOC.IN table.F.SG.NOM be.PRES.3.SG
    ‘In the house is a table.’ = ‘There is a table in the house.’

b. mez=pɑr ɑʊzar hɛ
    table.F.SG = LOC.ON tool.M.PL.NOM be.PRES.3.PL
    ‘On the table are tools.’ = ‘There are tools on the table.’

(102) a. mɑkan=mẽ apne bɑre kamre=mẽ
    mez hɛ
    table.F.SG.NOM be.PRES.3.SG
    ‘In the house, in self’s big room, is a table.’ = ‘There is a table in the house, in self’s big room.’

b. mez=pɑr apni mɑraṁmat=ke liye
    table.F.SG = LOC.ON self.F.PL repair.F.SG = GEN.M.SG.OBL for
    ɑʊzar hɛ
    tool.M.PL.NOM be.PRES.3.PL
    ‘On the table, for self’s repair are tools.’ = ‘There are tools on the table, for self’s repair.’

The examples in (100–102) are lccb — what if their word order is changed to render cllcs? (103) is a cllc with an abstract location; even without the reflexive binding test applied, the sentence is not acceptable without further specification in context. This is due to the information-structural properties of the clause; observe that the abstract nominal mɑmta ‘affection’ is in the sentence-initial topic position (Butt, 1995, Butt and King, 1996, 2000, Kidwai, 2000, Mahajan, 1990, Mohanan, 1994, see also §3.7.3 in (103), but abstract nominals are generally less adequate topics than other types of nouns (Langacker, 1999, Schmid, 2007). The English counterpart of (103) is also awkward,
possibly due to the same reason. The information-structural properties of locative clauses are discussed in some detail in §5.6.12

(103)  ? mamta nina = mē he affection.F.SG NOM Nina.F.SG = LOC.IN be.PRES.3.SG
    ? ‘The affection is in Nina.’

Inverting the order of the locative clauses with concrete spatial expressions in (101) yields the examples in (104). Applying the reflexive binding test as in (105), it can be seen that apna ‘self’ may in fact only be bound by the located NP, not by the locative KP. In absence of other binding possibilities, this suggests that in the CLCS, the located NP is the SUBJ (unless the sentences have no SUBJ), and the locative KP is realized as some other GF.

(104) a.  mez makan = mē he table.F.SG.NOM house.M.SG = LOC.IN be.PRES.3.SG
    ‘The table is in a/the house.’

   b.  ɑʊzar mez = par he tool.M.PL.NOM table.F.SG = LOC.ON be.PRES.3.PL
    ‘The tools are on a/the table.’

    ‘The table is in the house for self’s repair.’

    ‘The tools are on the table for self’s thorough examination.’

5.5.2.2 PRONOMINAL COREFERENCE

I repeat the definition of the subject diagnostic concerning pronominal coreference in (106). Regarding locatives, Mohanan (1994) gives the example in (107); she argues that because the pronoun us = ki ‘her’ cannot be coreferent with the locative L-SUBJ nina = mē (which is in its minimal finite clause), the locative L-SUBJ must be the SUBJ.

(106) The antecedent of a pronoun cannot be an f-commanding SUBJ of the minimal finite domain (i.e., the smallest f-structure with an attributive TENSE) that contains the pronoun.

12(103) may be acceptable in a certain discourse context where a specific kind of affection is under discussion, and then several people are described with respect to whether they have that kind of affection in them. Thus, such a sentence is certainly not ungrammatical, but quite restricted in its use.
Again, the locative in (107) is used in an abstract sense. To show that the usage of locative subjects extends to purely spatial, non-abstract senses, one needs to show that (106) is also true for those purely spatial uses of the locative. (108) is such an example. As in (107), it is found that the locative-marked ground makan=mẽ ‘in the house’ cannot be the antecedent of the pronoun us=ka ‘its’. Judging by the diagnostic in (106), this is evidence that the locative is the SUBJ, even though the locative is being used in a concrete spatial sense.

The examples in (107) and (108) constitute I LCS — i.e., the locative occurs sentence-initially. Turning to CLCS, consider the sentence in (109), in which again a purely spatial, concrete use of the locative is found.

It turns out that the same restriction concerning the coreferent of us holds for CLCS; ram, judging by this test, behaves like a SUBJ when occurring sentence-initially, while in (108) the locative makan=mẽ ‘in the house’ behaves like a SUBJ when occurring sentence-initially.

13I omit the pronominal coreference test for the non-inverted, canonical version of (107), where the locative carries an abstract meaning; in addition to the issue mentioned above regarding the inadequacy of abstract nominals such as mɑmta ‘affection’ for being topics, it is difficult to relate a pronoun such as us to such a nominal in any semantically plausible way.

14I must note that the judgments of my informants regarding (108) and (109) were not uniform. While a majority stated that the pronoun may refer neither to the locative in (108) nor to the located NP in (109) in compliance with Mohanan’s test, thus evidencing their SUBJ status, one informant’s intuition was that the pronoun may corefer either with some other nominal or with the locative in (108) the located nominal in (109). The same informant did not share Mohanan’s original judgment of (107).
5.5. THE LOCATIVE IN LOCATIVE CLAUSES

5.5.2.3 Control

Finally, the control test is a clear indication that concrete, spatial uses of the locative behave like the abstract uses discussed by Mohanan (1994) in that they may be realized as clausal SUBJS. Recall that Mohanan (1994) uses control of participial adjunct clauses as a subject diagnostic. I have defined the diagnostic in §3.4.3; it is repeated in (110). To show that locatives may behave as SUBJS with respect to control, Mohanan gives the example in (111). The locative-marked nominal controls an obligatory pronominal control site inside the participial adjunct clause, which Mohanan argues is a property of grammatical SUBJS only.

(110) A nominal that can control a participial adjunct clause with an obligatory control site must be a SUBJ. A nominal that cannot control a participial clause with an obligatory control site cannot be a SUBJ, provided that other factors (e.g., aspeclional differences) are controlled for.

(111) nina=mẽ [bacce=ki avasta dekʰ kar] Nina.F.SG = LOC.IN child.M.SG.OBL = GEN.F.SG condition.F.SG.NOM see do mamta puda hu-i affection.F.SG.NOM born be.PERF-FEM.SG

‘Nina, i seeing the child’s condition, began to feel affection (for the child).’

(Mohanan, 1994, p. 176)

Sentences such as (111) are best analyzed as resultatives (Ahmed et al., 2012): the result of an event (e.g., the bracketed portion in (111)) is that some property is predicated of some entity; here, the property of puda ‘born’ is predicated of nina=mẽ ‘in Nina’. Resultatives are known as instances of secondary predication (Nedjalkov, 1988, Simpson, 2006, and references therein). The locative clauses under examination in this chapter are different from resultatives in two ways. First, resultatives occur with a selection of different verbs in Hindi/Urdu, such as ho ‘be’ and kar ‘do’, while locative clauses invariably use the copula verb ho ‘be’. Second, resultatives belong to a different Aktionsart; they always involve the achievement of some result state by finishing an event (e.g., Herweg, 2013, Simpson, 2006). Thus, they are semantically very different from locative clauses, which are purely stative in nature. Another example of a resultative is given in (112).

(112) yasin=ne mez saf k-i Yassin.M.SG = ERG table.F.SG.NOM clean do-PERF.F.SG

‘Yassin made the table (be) clean.’ = ‘Yassin cleaned the table.’

(Ahmed et al., 2012)

Nevertheless, the question may arise whether locative-marked phrases behave the same way in locative clauses as in the resultative example provided by Mohanan (1994). The example in (113) cannot be analyzed as a resultative, as it lacks aspeclional information...
and is purely stative in nature. Thus, it must be analyzed as a locative clause. What is found here is that the locative-marked nominal ram = mē controls the control site of the participial adjunct clause (ek) acʰa larka ho-ne ‘being a good boy’; by (110), ram = mē must hence be a subj.

(113) ram = mē [(ek) acʰa larka
Ram.M.SG = LOC.IN (one) good.M.SG boy.M.SG.NOM
he be.PRES.3.SG
‘Ram i, __, being a good boy, has a lot of affection.’

(113) shows that at least non-spatial, abstract uses of the locative cause nominals to behave like subjects with respect to participial adjunct clause control. What about concrete uses of the locative? In (114), the reading of the locative cannot be plausibly extended to any non-spatial field (such as abstract containment as in (111) or (113)). Still, the locative-marked nominal controls an obligatory control site within the participial adjunct clause. Thus, judging by (110), the nominal is makaŋ=mē behaves like a subj.

(114) is makaŋ=mē [cʰota ho-ne = ki
this.OBL house.M.SG = LOC.IN small.M.SG be-Inf.M.SG.OBL = GEN.F.SG vaja = se] ek hi cuha he
cause.F.SG = INST] one FOC rat.M.SG.NOM be.PRES.3.SG
‘This house i, __, being a small house, has only one rat (in it).’

As all the examples above are ilcs where the locative occurs in sentence-initial position, one can again ask how this subject test pans out in clcs, as in (115) or (116).

(115) janwarō = ki pyar [(ek) acʰa
animal.M.PL.OBL = GEN.F.SG love.M.SG.NOM (one) good.M.SG
asas ho-ni] nina = mē he
feeling.M.SG.NOM be-INF.F.SG] Nina.F.SG = LOC.IN be.PRES.3.SG
‘The love for animals i, __, being a good feeling, is in Nina.’

(116) cuha [bɑhʊt cʰota janwar ho-na]
rat.M.SG.NOM very small.M.SG animal.M.SG.NOM be-INF.M.SG]
makaŋ=mē he
house.M.SG = LOC.IN be.PRES.3.SG
‘The rat i, __, being a small animal, is in the house.’ (i.e., it fits inside the house, thus it does not need to be outside)

It is clear that, if the located nominal is topicalized, it is the one which behaves like a subj with respect to control, and not the locative. janwarō = ki pyar ‘love for animals’
5.6. INFORMATION STRUCTURE

in (115) and cuha ‘rat’ in (116) are able to control the obligatory control site inside the adjunct clause, which shows that they act as the SUBJS of their respective sentences.\footnote{Again, sentences such as (i) are awkward due to the topicalized abstract nominal pyar ‘love’:

(i) pyar nina = mē he
love.m.sg.nom Nina.F.sg = loc.in be.pres.3.sg
‘The love is in Nina.’}

5.5.2.4 SUMMAR Y

Summing up the functional properties of locative clauses, I state that these clauses do not perform in a uniform fashion with respect to the grammatical functions involved, depending on the word order; all of the subject tests performed in this section indicate that when the locative occurs sentence-initially, it is the one that displays all of the properties of a subject, whereas if the located nominal occurs sentence-initially, it is the one with the subject properties, according to the subject tests put forward by Mohanan\footnote{(1994)}. Moreover, the discussion in this section has shown that it is not only abstract locations that take part in this pattern, but that the pattern extends to purely spatial and concrete uses of the locative, contrary to the claims by Mohanan\footnote{(1994)}. In my analysis in Chapter §6, I am going to explain the pattern by assuming the operation of locative inversion for the copula verb ho ‘be’, and show how the alternation is implemented in XLE.

5.6 INFORMATION STRUCTURE

It was seen earlier in this chapter that, not unlike the genitive (see §3.7.3), the word order is generally free in locative copula clauses. This section goes into more detail considering the semantic variation from the angle of discourse. Again, as with the genitive, I will not try to provide a complete analysis of the information-structural properties of locative clauses, and I do not implement the generalizations made in this section in the Urdu ParGram Grammar. However, it is useful to look at the data from the angle of information structure to understand why Hindi/Urdu speakers make the word-order choices in the way they do.

As was seen in §5.5.1.2, the word order in locative clauses may be different. Both of the word orders in the a. and b. examples of (117–118) are grammatical; in the a. examples, the locative-marked kp occurs sentence-initial, while the other nominal occurs in the medial, preverbal position. In the b. examples, the word order is inverted. Based on this data, in §5.5.1.2, I have argued that the c-structure of locative clauses must be a flat one, with two independent nominals. What was not yet discussed, however, is the reason why speakers produce these different word orders in the first place. In what follows, (115), where the abstract nominal is modified by a genitive phrase, is nevertheless fine. A reasonable assumption for explaining this effect is that the difference in the grammaticality is due to the modification by the genitive phrase. §5.6 elaborates on these issues.
I suggest that the appropriate choice of the ordering depends on the discourse context, since different orderings represent different information-structural configurations.

\[(117)\]
\[\text{a. } \text{mez} = \text{par} \quad \text{cuha} \quad \text{he} \quad \text{table.} \text{F.SG} = \text{LOC.ON rat.} \text{M.SG.NOM be.} \text{PRES.3.SG} \]
\[\text{‘On the table is a rat.’} = \text{‘There is a rat on the table.’} \]
\[\text{b. } \text{cuha} \quad \text{mez} = \text{par} \quad \text{he} \quad \text{rat.} \text{M.SG.NOM table.} \text{F.SG} = \text{LOC.ON be.} \text{PRES.3.SG} \]
\[\text{‘The rat is on the table.’} \]

\[(118)\]
\[\text{a. } \text{kamre} = \text{mē} \quad \text{admi} \quad \text{he} \quad \text{room.} \text{M.SG.OBL} = \text{LOC.IN man.} \text{M.SG.NOM be.} \text{PRES.3.SG} \]
\[\text{‘In the room is a man.’} = \text{‘There is a man in the room.’} \]
\[\text{b. } \text{admi} \quad \text{kamre} = \text{mē} \quad \text{he} \quad \text{man.} \text{M.SG.NOM room.} \text{M.SG.OBL} = \text{LOC.IN be.} \text{PRES.3.SG} \]
\[\text{‘In the room is a man.’} = \text{‘There is a man in the room.’} \]

Again, embedding the examples in a discourse context helps seeing their different information-structural properties. Consider the following dialogs. In \[(119)\]a, a particular dog is introduced as a discourse referent. In the locative clause in \[(119)\]b, the dog is again referred to, and the fact that the dog is inside some room is introduced. As the referent \text{kutta} ‘dog’ had already been introduced in the discourse, it is not felicitous to say \[(119)\]c instead of \[(119)\]b, since in \[(119)\]c, \text{kutta} ‘dog’ occurs in the immediately preverbal position, which is cited as the position in Hindi/Urdu encoding focus (Butt and King, 1996, 2000, Kidwai, 2000, Mahajan, 1990). Focused information has been characterized as new information (i.e., information not previously known to the hearer) by a wealth of literature on the subject of information structure (Butt and King, 1996, 2000, Choi, 1999b, Givón, 1979, King, 1995, Kiss, 1995, among others).

\[(119)\]
\[\text{a. } \text{ram} = \text{ka} \quad \text{kutta} \quad \text{he} \quad \text{Ram.} \text{M.SG} = \text{GEN.M.SG dog.} \text{M.SG.NOM be.} \text{PRES.3.SG} \]
\[\text{‘Of Ram is a dog.’} = \text{‘Ram has a dog.’} \]
\[\text{b. } \text{kutta} \quad \text{kamre} = \text{mē} \quad \text{he} \quad \text{dog.} \text{M.SG.NOM room.} \text{M.SG.OBL} = \text{LOC.IN be.} \text{PRES.3.SG} \]
\[\text{‘The dog is in the room.’} \]
\[\text{c. } \text{?} \text{ kamre} = \text{mē} \quad \text{kutta} \quad \text{he} \quad \text{room.} \text{M.SG.OBL} = \text{LOC.IN dog.} \text{M.SG.NOM be.} \text{PRES.3.SG} \]
\[\text{‘There is a dog in the room.’} \]

As a second example, consider the dialog in \[(120)\]. In \[(120)\]a, a speaker asks a question about the contents of a discourse referent \text{kamra} ‘room’. This referent occurs in sentence-initial position, and thus is treated as a topic (i.e., a referent already present in the discourse). The interrogative pronoun \text{kya} ‘what’ occupies the preverbal focus position. To answer in a coherent way, the other speaker chooses \[(120)\]b, where the
information that is being asked for is in the preverbal focus position, signaling that this
is new information. (120c) makes for a bad answer in this particular context since it
has the new information in the sentence-initial topic (Butt, 1995, Butt and King, 1996,
of topic, in turn, has been characterized cross-linguistically as encoding information al-
ready known to the hearer (Büring, 1999, Butt and King, 1996, 2000, King, 1995, Kiss,
1995, among others). In particular, notice how the interpretation of the referents in
(120–121) is changed pragmatically: even though there is no definite determiner in
Hindi/Urdu, the sentence-initial topicalized element receives a definite interpretation.
This is in sharp contrast to the preverbal element, which is interpreted as an indefinite.

(120)  
a.  kəmre=mẽ  kya  hɛ  
   room.M.SG.OBL = LOC.IN what be.PRES.3.SG
   ‘What is in the room?’

b.  kəmre=mẽ  kʊtta  hɛ  
   room.M.SG.OBL = LOC.IN dog.M.SG.NOM be.PRES.3.SG
   ‘In the room is a dog.’

c.  ??  kʊtta  kəmre=mẽ  hɛ  
    dog.M.SG.NOM room.M.SG.OBL = LOC.IN be.PRES.3.SG
   ‘The dog is in the room.’

(121)  
a.  kʊtta  kahã  hɛ  
    dog.M.SG.NOM where be.PRES.3.SG
   ‘Where is the dog?’

b.  kʊtta  kəmre=mẽ  hɛ  
    dog.M.SG.NOM room.M.SG.OBL = LOC.IN be.PRES.3.SG
   ‘The dog is in the room.’

c.  ??  kəmre=mẽ  kʊtta  hɛ  
    room.M.SG.OBL = LOC.IN dog.M.SG.NOM be.PRES.3.SG
   ‘There is a dog in the room.’

Thus, it is found that, in a particular context, one alternative word order may be more
fitting than the other. The different word orders represent different ways of conveying
the same information and presenting it to the addressee.

As with the possessive clauses, there are certain locative clauses that are not felicitous
due to information-structural constraints. Consider the word-order contrast in (122);
here, (122b) is frowned upon by native speakers of Hindi/Urdu, because it contains an
abstract bare nominal pyar ‘love’ in the topic position. Without any further modification,
such abstract nominals are bad topics: It is indeed hard to come up with a situation where
the abstract entity pyar ‘love’ is a discourse referent salient enough to be a topic. On the
other hand, (122c) indeed sounds less bad to my informants, even though it features
an NP with an abstract noun head as the topic. This is presumably due to the fact that
the abstract noun is modified by a genitive kp, and hence it is easier to construct it as a
topic in a discourse; such considerations are, however, outside the scope of the present discussion and present an interesting direction for future work.\[16\]

\[(122)\] (a) \(\text{nina} = \text{mē} \quad \text{pyar} \quad \text{he}\) 
\[\text{Nina.F.SG} = \text{LOC.IN} \quad \text{love.M.SG.NOM} \quad \text{be.PRES.3.SG}\]

‘In Nina is love.’ = ‘There is love in Nina.’

(b) ?? \(\text{pyar} \quad \text{nina} = \text{mē} \quad \text{he}\) 
\[\text{love.M.SG.NOM} \quad \text{Nina.F.SG} = \text{LOC.IN} \quad \text{be.PRES.3.SG}\]

‘The love is in Nina.’

(c) ?? \(\text{janwarō} = \text{ki} \quad \text{pyar} \quad \text{nina} = \text{mē}\) 
\[\text{animal.M.PL.OBL} = \text{GEN.SG} \quad \text{love.M.SG.NOM} \quad \text{Nina.F.SG} = \text{LOC.IN}\]
\[\text{he} \quad \text{be.PRES.3.SG}\]

‘The love for animals is in Nina.’

I exclude the level of information structure from my analysis of locative clauses. I am confident that approaches making use of the correspondence architecture in LFG (e.g., Butt and King, 2000, King, 1997, King and Zaenen, 2004, Laczkó and Rákosi, 2010, O’Connor, 2004, Sulger, 2009b) can handle the information-structural properties of locative clauses, which parallel the generalizations about possessive clauses drawn in §3.7.3.

5.7 CONCLUSION

This concludes the chapter on Hindi/Urdu locative and instrumental (li) nominal arguments. I have discussed the constituent as well as functional properties of these nominal arguments and argued for an analysis in terms of thematically-restricted oblique functions; locative nominal arguments are OBL-TH, realizing a theme argument, while instrumental nominal arguments are OBL-SRC. The possibilities of LI scrambling turned out to be identical to those of genitive scrambling, subject to the same preferences and constraints.

I have then turned to a discussion of locative clauses; these were shown to not behave like the possessive clauses surveyed in §3.7, but to belong to a different predicational pattern: they involve two separate nominals that are linked by a copula verb. It was further shown that locative clauses partake in a specific alternation that bears on the grammatical-functional structure of the clauses: in what I have called canonical locative clauses, the located nominal behaves like a subject, while in inverted locative clauses, the locative KP is the subject. In Chapter §6, I make sense of the construction by invoking

\[\text{\[16\]}\] Note also that abstract entities are on the low end of the ontological salience scale given by Langacker (1999). The scale, provided in (i), captures salience properties (i.e., properties that lend themselves to an entity being constructed as a discourse topic) that are inherent in the entities themselves.

(i) speaker > hearer > human > animal > physical object > abstract entity
locative inversion, a syntactic alternation on the argument structure of the copula, and present an complete LFG/XLE analysis of the construction.
6

LFG Analysis & XLE Implementation of the Locative & Instrumental

6.1 INTRODUCTION

The analysis of the Hindi/Urdu locative and instrumental (li) arguments makes a strict distinction between usages inside NPs and usages in locative clauses. On the one hand, when used in NPs, the li KP realizes a theme (locatives) or source (instrumentals) argument and is uniformly mapped to an OBL-TH or OBL-SRC GF. On the other hand, when used in locative clauses, it is either the locative KP which is realized as the SUBJ, or the located nominal. Instrumental KPs never feature in locative clauses. The syntactic mechanism responsible for the “inversion” of GFs in locative clauses is the operation of locative inversion (Bresnan and Kanerva, 1989). Before moving on to the actual spell-out of the analysis and the implementation, I provide descriptions of both locative inversion as discussed in the LFG setting (§6.2) as well as the different styles of copula analyses available in LFG (§6.3).

6.2 LOCATIVE INVERSION

In this section, I discuss the theoretical concept of “locative inversion”, explain how it accounts for the alternation found in Hindi/Urdu locative clauses, and discuss a possible implementation in a computational grammar. The account given here lays the ground for the implementation of the pattern in the Urdu ParGram grammar.
6.2.1 THEORETICAL EXPLANATION

I make the following proposal: Hindi/Urdu makes use of locative inversion (Bresnan and Kanerva, 1989). Locative inversion applies to verbs that specify a theme and a location argument (Bresnan and Kanerva, 1989, Kibort, 2007). This is precisely the function of the Hindi/Urdu copula verb ho ‘be’ in locative clauses.

I provide the argument frame of the copula ho ‘be’ in (1). Here, the analysis relies on standard Lexical Mapping Theory (LMT) assumptions, defined as part of LFG theory for mapping from thematic argument structure to grammatical-functional structure (Bresnan and Kanerva, 1989, Bresnan and Zaenen, 1990). ho ‘be’ is a two-place predicate with a theme and a location. Usually, in such a configuration, the argument-GF mapping in Figure 6.1 applies. Here, the theme argument is intrinsically classified as a $[-r]$ argument, while the locative argument is intrinsically classified as a $[-o]$ argument (see §2.4.4 for a discussion of the use of the $[\pm o]$ and $[\pm r]$ features in LFG). By default, the locative receives a $[+r]$ value (by virtue of being a semantically restricted role); the theme ends up as the SUBJ of the predication, while the locative ends up as a different GF (naturally, not the SUBJ). For (2), this will result in the f-structure shown in Figure 6.2.

\[(1) \quad \text{ho} < \text{th loc} >\]

\[
\begin{array}{c|c|c|c|c|c}
\text{ho} & < & \text{th} & \text{loc} & > \\
& | & | & | \\
\text{intrinsic} & [-r] & [-o] & & \\
\text{defaults} & & & [+r] & \\
\hline
\text{OBJ/SUBJ} & \text{GF2 (\neg SUBJ)} \\
\text{well-formedness} & \text{SUBJ} & \text{GF2 (\neg SUBJ)} \\
\end{array}
\]

\[\text{Figure 6.1: Linking analysis for predicative locatives}\]

\[(2) \quad \text{kutta} \quad \text{kamre=m\=e} \quad \text{he} \\
\text{dog.M.SG.NOM} \quad \text{room.M.SG.OBL} = \text{LOC.IN be.PRES.3.SG} \\
\text{‘The dog is in the room.’}\]

In inverted locative clauses, on the other hand, the assignment of grammatical functions is reversed, as was seen in §5.5.2; the locative behaves like a SUBJ, and the theme does not. In the original formulation of locative inversion within the framework of LFG, Bresnan and Kanerva (1989), working on the Bantu language Chichewa, motivate this... \[\text{1For the non-SUBJ GF, I use GF2 as a placeholder throughout this section; the status of this particular non-SUBJ GF is at issue in the discussion in §6.3.}\]
alternate assignment in terms of discourse functions. Taking non-inverted locatives such as the one in (2) as the default case, they argue that inverted locatives are the marked case, having a presentational function whereby the theme is focussed and the locative is a topic. This is precisely what is found in Hindi/Urdu locative clauses (see my argumentation in §5.6); in non-inverted locative clauses, the theme argument is in the topic position, while in the inverted clauses, the locative argument is.

Kibort (2007), taking up the markedness explanation, argues that in inverted locatives, the theme must receive a [+o] value, leaving the locative to become the subject by well-formedness conditions. I adopt this approach to locative inversion, shown in Figures 6.3 and 6.4. The resulting f-structure for the inverted version of (2), (3), is given in Figure 6.5.

(3) kamre = mē kutta he room.M.SG.OBL = LOC.IN dog.M.SG.NOM be.PRES.3.SG
    ‘In the room is a dog.’ = ‘There is a dog in the room.’
Locative inversion is generally thought of as a process applying to a verb that predicates over a theme and a locative argument (Bresnan and Kanerva, 1989, Freeze, 1992, Kibort, 2007). Here, I confine myself to describing the copula ho ‘be’ in its usage in locative clauses, which carries exactly that function: tying a nominal to some location. Of course, there are other predicates in Hindi/Urdu that realize the very same function; for example, the verb ja ‘go’ comes to mind. See the examples in (4).

(4) a. kʊtta zu=ko ga-ya
dog.M.SG.NOM zoo.M.SG = ACC go-PERF.M.SG
‘The dog went to the zoo.’

b. zu=ko kʊtta ga-ya
zoo.M.SG = ACC dog.M.SG.NOM go-PERF.M.SG
‘The dog went to the zoo.’

For such verbs, one would certainly expect effects similar to the effects found with the copula. It would be surprising if locative inversion were found to apply exclusively to the copula verb ho ‘be’, but not to other verbs that involve the same argument configuration (Bresnan and Kanerva, 1989). Further investigations into the domain of locative inversion and its application to predicates in Hindi/Urdu are needed and outreach the frame of copula predication set in this chapter.

6.2.2 COMPUTATIONAL IMPLEMENTATION

In xle, the alternation resulting from locative inversion can be adequately modeled for Hindi/Urdu via a template such as the one in (5); the template states that it is either the subject which is locative-marked, in which case the other GF has to be nominal case-marked, or it is the other way round: the subject is nominative case marked, and the other GF is a locative. In either case, the subject has to precede the second GF.

(5) LOC-INV = {¬SUBJ CASE =c loc

(¬GF2 CASE =c nom

| ¬SUBJ CASE =c nom

(¬GF2 CASE =c loc

(¬SUBJ) ≪ h (¬GF2).}
thus models the generalization from locative inversion with respect to the linear order constraints found in locative clauses: it is uniformly the sentence-initial nominal that exhibits subject behavior. The template can be called from a lexical entry of the verb that takes part in the alternation, resulting in the correct GF assignment.

An approach to locative predication in terms of two independent thematic roles and, indeed, grammatical functions, makes the correct predictions given the data. Locative inversion is the process that accounts for the GF alternation in locative clauses: GF assignment is tied very much to c-structural positions in these clauses, rendering two syntactically fixed patterns in a language that is otherwise a free word order language.

Comparing the locative pattern with the genitive pattern makes for a strong case against views that aim at unifying accounts of locatives and possessives (e.g., Clark, 1978, Freeze, 1992). For Hindi/Urdu, such a view turns out to be not feasible. The possessive construction is an intransitive predicate, and the genitive is a strictly nominal case. Locative clauses, on the other hand, predicate differently in that they make use of a two-place copula verb. The next section concerns the status of the second GF as part of the copula analysis as a whole.

6.3 THE COPULA ANALYSIS

In LFG, there have been several attempts at analyzing copula constructions. Since linguists working within LFG have tried to come up with analyses that can be applied to related phenomena cross-linguistically, and since there is a great deal of variation in the realization of the copula construction across languages, no definitive LFG approach for the copula has been settled on (Attia, 2008, Butt et al., 1999b, Dalrymple et al., 2004a, Laczkó, 2012, Sulger, 2009a). An overview of the different types of copula analyses in LFG is given in Figure 6.6, which is adapted from Attia (2008).

![Figure 6.6: Copula construction analyses in LFG](image)

Within the ParGram project, most grammars use a double-tier analysis for copula predication. An approach advocating a uniform treatment of copulas cross-linguistically was advocated in the early years of ParGram (Butt et al., 1999b); however, a uniform analysis could not do justice to the typological variation found with copulas (as will be seen in the discussion in this section). As it stands, ParGram reflects the typological difference.
by employing all of the analysis in Figure 6.6, with each language making a language-specific choice (Sulger et al., 2013). In this section, I briefly review each of these analysis for their appropriateness for modeling Hindi/Urdu locative clauses.

### 6.3.1 The Single-Tier Analysis

In the “single-tier analysis”, the predicate of the copula construction on its own functions as the sentential head; it itself selects for a subject. Dalrymple et al. (2004a) claim that this is the appropriate analysis for copula constructions where the copula can be omitted under certain conditions; in those cases, it is argued that the predicate can subcategorize the subject. Two examples from Japanese are given in (6); here, the adjective is argued to be the head of the clause, subcategorizing for a subj. Dalrymple et al. (2004a) propose that a complement should be treated as selecting for a subj if the complement can be used in copula clauses with the copula missing. This entails that adjectives such as aka ‘red’ in Japanese will be treated as subcategorizers in the language, even in contexts outside of copula clauses. The resulting f-structure for (6a) is given in Figure 6.7.

(6)  

| (a) | hon wa akai.  
| --- | book TOP red  
| ‘The book is red.’ | Japanese (Dalrymple et al., 2004a, p. 190)  
| b. | sono hon wa akai desu.  
| this book TOP red is  
| ‘This book is red.’ | Japanese (Dalrymple et al., 2004a, p. 191)  

![Figure 6.7: Single-tier analysis for Japanese (Dalrymple et al., 2004a, p. 191)](image)

In Japanese, the situation is slightly more complicated in that nominal predicates, in contrast to adjectival predicates, do not allow the copula to be omitted. Hence, adapting the analysis by Dalrymple et al. (2004a) entails that one would have to assume that nouns in Japanese are not able to subcategorize for a subject, but adjectives are, and the copula may therefore be omitted. Indeed, the conclusion drawn by Dalrymple et al. (2004a) is that, within Japanese, one has to assume different analyses for copula clauses that look very similar on the surface, but nevertheless behave differently as far as copula occurrence is concerned, depending on the category of the predicative.

---

2In this section, I use the terms “copula predicate” and “copula complement” for referring to the same thing: the material in the copula clause that maps to the grammatical function that is predicated of the subject.
Copula occurrence also seems to be governed by tense (Dalrymple et al., 2004a). Cross-linguistically, in languages where a null copula is allowed, it seems that the copula’s tendency to be null is stronger in the present tense, while it generally must be overt in past and future tenses. Languages that employ this pattern are, for example, Russian and Arabic (Dalrymple et al., 2004a, and references therein). For such languages, Dalrymple et al. (2004a) argue that a unified analysis is desirable, as there does not seem to be any difference in the syntax and semantics of such constructions. They discuss both a single-tier analysis and a closed complement double-tier (“predlink”) analysis (see §6.3.2.2) for such languages, claiming, however, that a unified analysis might not be possible for all languages where copula occurrence is partially controlled by tense.

Attia (2008) argues against positing divergent analyses for languages such as Japanese, where Dalrymple et al. (2004a) motivate the differentiation in analysis with the presence vs. absence of the copula. Attia (2008) claims that one should only use different syntactic analyses for the same function (e.g., copula predication) when there is motivation for doing so. The overt realization of the copula in different contexts is not enough motivation to assume diverging analyses, whatever the deciding factor may be: category of the predicate constituent or tense. Attia (2008) further provides evidence from Arabic, where it is clear that the predicate cannot be the head of the clause because it does not assign case to the subject. However, it remains questionable in how far the statement about Arabic case-assigning predicates holds valid as cross-linguistic evidence that copula predicates can never be the head of a sentence. In essence, however, the criticism by Attia (2008) is valid and points strongly into one direction: whether the copula is in the sentence or not does not matter; the syntactic function of the copula predicate is the same in both contexts: linking the subject and the predicate, thereby ensuring predication. Posing different analyses for copula constructions, either within a language or cross-linguistically, does not capture the generalization of predication across copula sentences.

Another piece of work advocating the use of the single-tier analysis for copula predication is Nordlinger and Sadler (2006). They discuss examples from Abkhaz, a Caucasian language spoken in Georgia and Turkey. Here, the predicate carries morphological features which are normally present on verbs, such as tense, aspect and mood. In (7), a whole sentence is formed containing a single word, namely the inflected predicate.

(7)   Də-psə́-w-p’.
3SG.SBJ-dead-PRES-DECL
‘He is dead.’ Abkhaz (Nordlinger and Sadler, 2006, p. 146)

This is a case where it is indeed questionable whether one can do without the single-tier analysis. Attia (2008) acknowledges this, and states that cases such as (7) are compelling motivation for a single-tier analysis as in Figure 6.7. Thus, in cases such as (7), i.e., in a specific language and under specific conditions, the predicate is realized as a subcategorizing head, argues Attia (2008). Supposedly, by the specific conditions, cases
Currently, the only grammar among the ParGram grammars that makes use of the single-tier analysis for copula predication is the Indonesian grammar (Arka, 2012).

6.3.2 The Double-Tier Analyses

Under both of the double-tier analyses, the copula selects for two separate GFS, a SUBJ and another GF, where the intuition is that the latter is predicated of the former. The predicate GF can be annotated on a number of different c-structure categories; e.g., there are predicative NPS, APS, PPS. The same is true for Hindi/Urdu; although I have only discussed predicative (locative) KPs in Chapter 8, the predicate function may in fact be filled by different categories. For some examples, see (8-10).

(8) Predicative NP:
   a. Sam is a teacher. English
   c. Il est enseignant. PRON.M.SG.NOM be.PRES.3.SG teacher.M.SG.NOM ‘He is a teacher.’ French
   d. ram aḍhupak he Ram.M.SG.NOM teacher.M.SG.NOM be.PRES.3.SG ‘Ram is a teacher.’ Hindi/Urdu

(9) Predicative AP:
   a. The house is red. English
   b. Das Haus ist rot. the.N.SG.NOM house.N.SG.NOM be.PRES.3.SG red ‘The house is red.’ German
   c. La maison est rouge. the.F.SG house.F.SG.NOM be.PRES.3.SG red ‘The house is red.’ French
   d. makan surx he house.M.SG.NOM red be.PRES.3.SG ‘The house is red.’ Hindi/Urdu

(10) Predicative PP/KP:
   a. The dog is on the roof. English

---

3In such cases, however, the boundaries between categories are blurred, so that the question can be asked whether includes an incorporated verbal or copular head. I will not go into further detail here.
There are two types of double-tier analyses that differ with respect to the status of the predicate function. On the one hand, the XCOMP analysis employs an open complement function XCOMP for the predicate (and has therefore been dubbed “the open complement double-tier analysis”); on the other hand, there is the PREDLINK analysis, which makes use of a closed complement PREDLINK in what is also called “the closed complement double-tier analysis”. Both types of analyses are briefly reviewed below.

6.3.2.1 THE XCOMP ANALYSIS

Under one analysis, the predicative phrase selects for a subject, which is functionally controlled by the main subject of the sentence. In LF, functional control is realized by an XCOMP (“predicate complement”) attribute; the subject of XCOMP is unified with the main subject of the sentence. XCOMP is an “open” function, meaning that control equations are defined between the subject of the sentence and the XCOMP’s subject. The partial XCOMP f-structure alone does not contain a subject value, i.e. is not complete; it receives the subject’s value by functional control through the linking verb (Bresnan, 1982a, 2001, Butt et al., 1999b, Lødrup, 2011). Due to this “open” nature of the copula complement, this analysis is also referred to as the “open complement double-tier analysis” (Attia, 2008, Dalrymple et al., 2004a).

Dalrymple et al. (2004a) state that this is the preferable analysis for cases in which the predicate shows agreement with its subject. Agreement is seen as a strong indication for a control relation between the subject and the predicate. Dalrymple et al. (2004a) give the example of French adjectives in predicate position, agreeing with the subject of the main clause much like verbs agree with subjects in French. See (11a) and (11b) and the resulting f-structure in Figure 6.8, where the index 1 indicates that the XCOMP’s SUBJ is functionally controlled by the subject of the main clause. The examples are taken from Dalrymple et al. (2004a), who also state that if one assumes an XCOMP type of analysis, one can write basic lexical entries for the copula as in (12a), and for the predicate as in (12b).

(11) a. Il est petit.
he.M.SG be.PRES.3.SG small.M.SG
‘He is small.’
French (Dalrymple et al., 2004a, p. 195)
6.3. THE COPULA ANALYSIS

b. Elle est petite.
    she.F.SG be.PRES.3.SG small.F.SG
    ‘She is small.’
    French (Dalrymple et al., 2004a, p. 195)

(12) a. est V (↑ PRED) = ‘be<(↑ SUBJ), (↑ XCOMP)>’
    (↑ SUBJ) = (↑ XCOMP SUBJ)

b. petite A (↑ PRED) = ‘small<(↑ SUBJ)>’
    (↑ SUBJ NUM) = c sg
    (↑ SUBJ GEND) = c fem

Figure 6.8: Open double-tier analysis of French copula

Thus, the XCOMP analysis allows a simple annotation, and naturally accounts for the subject-predicate agreement in terms of functional control. Crucially, the predicate is taken to select for its own SUBJ, which is why agreement between predicate and subject can be enforced exactly as in between verb and subject. There are, however, also disadvantages to the XCOMP analysis in terms of agreement. Attia (2008) makes the valid point that in French, the agreement of the copula predicate with the subject is not the same type of agreement verbs show with their subjects: while verbs always agree with subjects in person, copula predicates do not necessarily do so; e.g., adjectives do not agree in person with the subject. This would entail that one would need to define two separate sets of agreement relations for copula-predicate vs. subject-verb agreement.

In addition, Attia (2008) mentions that agreement cannot be governed by the syntax alone, but that the semantics must be involved. Given the data in (13–14), it is in fact reasonable to assume that agreement is licensed via the semantics rather than the syntax; otherwise, it would be hard to explain why (13b) is ungrammatical while (14a-b) are grammatical. By consequence, agreement relations cannot be stated solely on the basis of syntactic rules; vice versa, deciding on a syntactic analysis on the basis of agreement remains questionable.

4The fact that the semantics must be involved in agreement is acknowledged, at least for predicative adjectives, by Dalrymple et al. (2004a), stating the following:
6.3. THE COPULA ANALYSIS

(13)  a. They are doctors.  
      b. *They are a doctor.  
      (Attia, 2008, p. 103)

(14)  a. They are the cause of our trouble.  
      b. They are a big problem.  
      (Attia, 2008, p. 103)

Another argument contra the XCOMP open complement analysis comes from a predicational perspective. The analysis in Figure 6.8 is not suitable for copula constructions according to Attia (2008) because this is exactly the way normal subject raising verbs (such as seem, appear) are analyzed in LFG. Assuming an f-structure like Figure 6.8 would therefore mean that there is no difference between copula constructions and subject raising verbs, such as the one in (15) (with the corresponding f-structure in Figure 6.9).

(15)  Elle semble contente.  
      ‘She seems happy.’

     PRED  ‘seem<(↑ XCOMP)> (↑ SUBJ)’
     SUBJ  ‘she’
           NUM  sg  1
           GEND fem
     XCOMP  ‘happy<(↑ SUBJ)>’
          SUBJ [ ] 1
          ATYPE predicative

Figure 6.9: Open double-tier analysis of French raising verb

Note that there is some tradition in generative grammar in treating copula constructions as raising constructions (Adger and Ramchand, 2003; Bresnan, 2001; Carnie, 1997; Moro, 1997). Attia (2008) rejects this view. He claims that one has to differentiate between raising constructions with verbal complements, and raising constructions without verbal complements; sentences like (16a) and (16b) are argued to have different underlying structures, because the predication is different.

(i) In other languages, however, some considerations may weaken the status of agreement as an argument for assuming an XCOMP analysis. In languages like Norwegian, for example, there is no subject-verb agreement, so that subject-adjective agreement must be treated differently from subject-verb agreement in any case. Another issue is that predicative adjective agreement may be governed by semantic rather than syntactic features.  
      (Dalrymple et al., 2004a, p. 196)
(16)  
(a) He seems to go.  
(Attia, 2008, p. 104)  
(b) He seems happy.  
(Attia, 2008, p. 104)

The verbal complement in (16a) clearly selects for a subject. go occurs regularly elsewhere with subjects, so unless one is to assume several different lexical entries for go, it is plausible to assume a raising construction. Attia (2008) claims, however, that it does not seem plausible, and also hard to prove, that the adjective phrase in the complement position in (16b) (and also adverbial phrases, noun phrases, prepositional phrases) subcategorizes for a subject. The analysis proposed by Attia (2008) treats the two verbs seem in (16a) and (16b) differently; seem in (16a) is analyzed as an actual raising verb with functional control, while seem in (16b) is treated as a "quasi-copula", linking subject and predicate. A relevant quote is given in (17):

(17) The difference between *he seems to go* and *he seems happy* is the same as the difference between *he goes* and *he is happy* which are completely different syntactic structures. The first is a verbal construction while the second is a predicational construction.  
(Attia, 2008, p. 16)

Dalrymple et al. (2004a) themselves provide the most compelling evidence against an XCOMP analysis, further emphasizing their argumentation for a pluralist approach towards copula constructions. In cases where the post-copular complement already has a subject which is different from the subject of the main clause, the closed complement PREDLINK analysis is the preferred analysis. See the examples in (18a) and (18b).

(18)  
(a) The good thing is that he did not throw the snowball.  
(b) The main goal is (for the student) to succeed in the exam.

If an XCOMP analysis is assumed for sentences like these, the result is a clash of PRED values: because of the control equations, the XCOMP f-structure would contain two subjects which are not unifiable. See the illformed f-structure in Figure 6.10 for sentence (18a).

```
Figure 6.10: Clashing open double-tier analysis with divergent subjects
```
6.3. THE COPULA ANALYSIS

6.3.2.2 THE PREDLINK ANALYSIS

Another commonly used copula analyses, both in LFG theory (Attia, 2008, Laczkó, 2012, Sulger, 2009a) as well as in LFG grammar writing within the ParGram project (Butt et al., 1999b, Sulger et al., 2013), is the PREDLINK analysis. Here, the copula verb is thought of as a functional element that links the subject to its predicate; the latter is projected to a PREDLINK grammatical function. Thus, the PREDLINK analysis models copula predication in a regular way: the main PRED of the f-structure expresses that a specific property is predicated of the subject. Exactly this is captured by the PREDLINK function (Butt et al., 1999b). An example f-structure from Butt et al. (1999b) for the English sentence in (19) is given in Figure 6.11.

(19) The tractor is red.

```
Figure 6.11: (Simplified) f-structure for (19) (Butt et al., 1999b, p. 70)
```

One main advantage of the PREDLINK approach is that since it does not rely on control equations, it does not have any issues with sentences such as (18a) and (18b). On the other hand, when there is subject-predicate agreement, the relevant constraints can also be formulated using the PREDLINK analysis (Attia, 2008, Butt et al., 1999b). Re-consider the example in (11); to model the same data under a PREDLINK analysis, one uses a lexical entry for the copula as in (20a) as well as inside-out constraints (Dalrymple, 2001, p. 143) as in (20b). Agreement in this case occurs outside the adjective’s partial f-structure (i.e., the agreement is not governed locally by the adjective, since the subject is not in the subcategorization frame of the adjective). The result is the f-structure in Figure 6.12.

(20) a. est v (↑ PRED) = ‘be<(↑ SUBJ), (↑ PREDLINK)>’
    b. petite a (↑ PRED) = ‘small’
       ((PREDLINK ↑) SUBJ NUM) = c sg
       ((PREDLINK ↑) SUBJ GEND) = c fem
       (↑ ATYPE) = predicative

According to Dalrymple et al. (2004a), the notation in (20b) is cumbersome. Since the XCOMP analysis shown in Figure 6.8 and (12) allows a simpler annotation, and since it is more similar to the analysis of other cases of agreement, such as subject-verb-agreement, the XCOMP analysis should be preferred over the closed complement analysis. However,
the non-standard equations shown in (20b) can be rewritten using standard notation. Here, Attia (2008) states that one should not prefer one syntactic analysis over the other because of notational preferences. Moreover, Attia (2008) points out that the feature constraints could be annotated in the syntax of the grammar (i.e. in the phrase-structure rules), and not in the lexical entries of the adjectives, since it seems implausible to generally assume that all adjectives subcategorize over a subject and agree with it.

There are several other advantages to the double-tier closed complement PREDLINK analysis. First, it does not matter what kind of constituent the copula complement is. The PREDLINK analysis seems to be the only one that succeeds in providing valid representations for all constituent types, which can take different semantic roles (see Attia (2008) for an overview). Other approaches seem to have problems with a unified analysis. Bresnan (2001), for example, uses the XCOMP analysis, and thus assumes that e.g. adjectives can subcategorize for subjects; she also assumes that nouns and prepositional phrases can do so. To account for this, she proposes to manipulate the PRED of the noun or preposition by means of lexical rules; see the sentences in (21) and the corresponding rules in (22a) and (22b), cited by Lødrup (2011).

(21)  a. The pills made him a monster.  (Lødrup, 2011, p. 22)
    b. She seems in a bad mood.  (Lødrup, 2011, p. 22)

(22)  a. ‘monster’  ⇒  ‘be-a-monster<↑SUBJ>’  (Lødrup, 2011, p. 22)
    b. ‘in<↑OBJ>’  ⇒  ‘be-in-a-state-of<↑SUBJ>, (↑OBJ)>’  (Lødrup, 2011, p. 22)

Both Attia (2008) and Lødrup (2011) find this approach problematic, since it not only results in artificial and complex annotation, but also presupposes that any PP or NP in a given language can in principle subcategorize for a subject. Dalrymple et al. (2004a) and also Rosén (1996) in an earlier paper maintain that this type of analysis is certainly not desirable. Within the closed complement analysis, these problems vanish, since there is no XCOMP f-structure, and hence there is no need for a subject that is functionally
controlled. I give the f-structure for (21b) in Figure 6.13, assuming a closed complement double-tier analysis.

\[
\begin{array}{c}
\text{PRED} & \text{‘seem<($\uparrow$ SUBJ), ($\uparrow$ PREDLINK)>’} \\
\text{SUBJ} & \begin{array}{c}
\text{PRED} \text{ ‘she’} \\
\text{PRED} \text{ ‘in<($\uparrow$ OBJ)>’}
\end{array} \\
\text{PREDLINK} & \begin{array}{c}
\text{OBJ} \begin{array}{c}
\text{PRED} \text{ ‘mood’} \\
\text{ADJUNCT} \{\text{PRED ‘bad’}\}
\end{array}
\end{array}
\end{array}
\]

Figure 6.13: Well-formed closed double-tier analysis of She seems in a bad mood.

In languages where the copula may be non-overt, the PREDLINK analysis has great intuitive appeal, since, at f-structure level, it mirrors the juxtaposition of constituents when the copula is missing (Attia, 2008). Since there are many languages that contain be-less sentences (Carnie, 1995), the analysis further gains cross-linguistic appeal. Attia (2008) further argues that the presence vs. absence of the copula itself is a parameter of variation. Since the copula is generally considered as semantically empty, there is no functional distinction to be made between sentences containing the copula and sentences without the copula. The predication in the absence of the copula can be modeled using a null-be predicator in LF rule notation. See the sentence in (23) and the rule in (24); the resulting f-structure is shown in Figure 6.14.

(23) hwa ṭālibun
he student
‘He is a student.’ Arabic (Attia, 2008, p. 19)

(24) \[ S \rightarrow \begin{array}{c}
\text{NP} \\
\Downarrow \text{SUBJ} = \downarrow \\
\Downarrow \text{TENSE} = \text{pres}
\end{array}\]

The rule in (24) looks complicated, and some clarification is necessary. The rule states that a sentence can be formed out of an NP, annotated for the subject GF; an empty element with the predicate null-be, selecting for a subject and a PREDLINK and

\[ \{\text{NP} | \text{AP}\} \]

\[ \Downarrow \text{PREDLINK} = \downarrow \]

\[ \Downarrow \text{GEND} = \Downarrow \text{SUBJ GEND} \]

\[ \Downarrow \text{NUM} = \Downarrow \text{SUBJ NUM} \]

Note that the \( \varepsilon \) in (24) is not an empty category, but an empty element; the difference is crucial in that no additional nodes will appear in the c-structure resulting from (24). \( \varepsilon \) only affects f-structure: the
providing the sentence with present tense; and a noun phrase or an adjective phrase as the head of the \textsc{predlink} f-structure, agreeing in gender and number with the subject of the main clause. Note that the tense feature is provided by the empty element $\epsilon$, which captures the insight that the copula in Arabic can only be omitted in present tense. If a copula was present in the sentence, then the tense feature would be provided by the copula itself.

6.3.3 \textsc{Choosing an Analysis for Hindi/Urdu Locative Clauses}

The data from Hindi/Urdu locative clauses are best analyzed using the \textsc{predlink} analysis. The reasons I have opted for this particular analysis are listed below.

6.3.3.1 \textsc{Agreement}

Agreement between \textsc{subj} and the copula predicate in Hindi/Urdu only occurs in the case of \textsc{aps} and \textsc{np}s, as in examples (25)-(26) below. The predicate may also be a \textsc{pp} (postpositional phrase, §2.3.3) as in (27); agreement does not occur in this case.

(25) Predicative \textsc{ap}:
\begin{itemize}
  \item a. makan bara he
      house.M.SG.NOM big.M.SG be.PRES.3.SG
      ‘The house is big.’
  \item b. * makan bari he
      house.M.SG.NOM big.F.SG be.PRES.3.SG
\end{itemize}

(26) Predicative \textsc{np}:
\begin{itemize}
  \item a. ram ustad he
      Ram.M.SG.NOM teacher.M.SG.NOM be.PRES.3.SG
      ‘Ram is a/the teacher.’
  \item b. * ram ustani he
      Ram.M.SG.NOM female-teacher.F.SG.NOM be.PRES.3.SG
\end{itemize}

\textit{associated annotations will be introduced as if they were associated with the other categories in (24). See also \url{http://ling.uni-konstanz.de/pages/xle/doc/notations.html#NƦ} for discussion of the empty element and its implementation in XLE.}
As discussed above, agreement by itself is not a sufficient criterion for choosing one analysis over the other. This ends up being a notational difference only, and not one of linguistic adequateness, since the agreement constraints can be modeled under the predlink analysis as well as under the xcomp analysis.

Further, there are grammatical examples of predicative NPs in Hindi/Urdu that lack subject-copula agreement (i.e., examples corresponding to e.g., English They are a big problem). Two such examples are given in (28). This does not necessarily argue in favor of either syntactic analysis. Such data rather suggests that one should not posit agreement as the deciding factor at any rate, but that the semantics plays a major role, a route which I will not further investigate here.

   ‘They are a big problem.’

b. vʊh ek lɑmbi kɑhani hẽ PRON.3.PL.NOM one long.F.SG story.F.SG.NOM be.PRES.3.PL
   ‘They are a long story.’

6.3.3.2 Predicate Categories

In general, there are various c-structure nodes that can be predicates in copula clauses in Hindi/Urdu: APS, KPS, NPS as well as PPS may all be used predicatively, although predicate types other than the KP are not discussed in this chapter at length. Under the xcomp analysis, one would have to assume that all these nodes may select for a subj (which is then functionally controlled by the main clause subj). There is no a priori reason for assuming this. Moreover, a lexical rule augmenting the relevant lexical entries would result in a doubling of the relevant lexical entries in the grammar.

6.3.3.3 Cross-linguistic Appeal

The predlink analysis overall has more cross-linguistic appeal. Using its various extensions (inside-out functional constraints for agreement, the null-pred predicate for dropped copulas), it can cover more data than the xcomp analysis and does not run into problems where the copula predicate already includes a subj, as in e.g., (18) discussed above.

\[^6\] I was not able to find such examples with predicative APS.
6.3.4 Summary

To sum up the discussion about the different approaches towards copula constructions in LFG, I stress that each one of the possible analyses has its advantages and disadvantages. While one has to assume a subcategorization frame for predicate elements in the open-complement types of analysis which might be less appropriate for some languages than for others, the PREDLINK approach is more neutral in this respect. However, the annotation methods required in connection with the PREDLINK might be a little more complicated than the intuitive functional control annotations of the XCOMP analysis. I think in this discussion it is important to see LFG in the context of parallel (i.e. cross-linguistic) grammar designing and engineering. As long as there are no serious reasons to object the PREDLINK analysis, I maintain that a universally applicable analysis should be favored to increase the cross-linguistic appeal of LFG. The main features of the XCOMP and PREDLINK types of copula analyses in LFG are summarized in Table 6.1.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Predicate function</th>
<th>SUBJ control</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-tier main PRED</td>
<td></td>
<td>X</td>
<td>Non-overt copula</td>
</tr>
<tr>
<td>Open complement</td>
<td>XCOMP</td>
<td>✓</td>
<td>Agreement, simple notation</td>
</tr>
<tr>
<td>double-tier</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closed complement</td>
<td>PREDLINK</td>
<td>X</td>
<td>Regular predication, no doubled lexical entries</td>
</tr>
<tr>
<td>double-tier</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: Copula analysis types in LFG

6.4 Locative & Instrumental Case Marking

Generally, Hindi/Urdu locative and instrumental case is only assigned in verbal environments. Exceptions from this generalization are:

- a set of common nouns known as “picture nouns”, which occur with a single locative-marked OBL-TH;
- a set of event nouns, which may occur with either a locative OBL-TH or an instrumental OBL-SRC;
- a set of relational nouns, which may occur with either a locative OBL-TH or an instrumental OBL-SRC.

To avoid confusion, it is important to note that here, OBL-TH does not stand for a generic thematically-restricted oblique (OBLθ), but for an oblique instantiated for the thematic role of a theme.

7A case in point here is the account for Abkhaz predication presented by Nordlinger and Sadler (2006) and discussed in §6.3.1.
The lexical entries of the locative case markers are given in (29). Unlike the genitive case marker, but like the other case markers of Hindi/Urdu, none of the case markers in (29) agree with a head noun. The first line in each of the lexical entries assigns the locative case to the KP; compare this to the analysis/implementation of the genitive case, which is assigned structurally (§4.2). The second line in each entry defines a lexical-semantic concept of location for each case marker by calling up the template in (30) that assigns the value accordingly. In the Urdu ParGram grammar, the SEM-PROP feature complex defines lexical-semantic properties of GFSs that are driven by semantic case marking (cf. §2.3.2) as well as inherent semantic properties of nominals. The values used for the LOCATION feature as part of SEM-PROP are in line with the description of locative case marking in Hindi/Urdu by Butt and King (2004b) (cf. Table 5.1). The third line in each entry is an inside-out function application that models the distribution for each of the case clitics: it is the clitics themselves that indicate their distribution in terms of the GFSs they occur on. The implementation is reminiscent of what has been termed “constructive case” by Nordlinger (1998) for Australian languages where the case morphology indicates the GF.

(29) a. mEN K * @(CASE loc)
    @ (SEM-LOC in)
    (@ {ADJUNCT $|OBL|PREDLINK|SUBJ} ^).

b. par K * @(CASE loc)
    @ (SEM-LOC on)
    (@ {ADJUNCT $|OBL|OBL-TH|PREDLINK|SUBJ} ^).

c. tak K * @(CASE loc)
    @ (SEM-LOC towards)
    (@ {ADJUNCT $|OBL} ^).

d. ko K * @(CASE loc)
    @ (SEM-LOC towards)
    (@ {ADJUNCT $|OBL} ^).

(30) SEM-LOC(_P) = (~ SEM-PROP LOCATION) = _P.

Thus, the locative case markers not only contribute to the GF identification in the clause, but also provide detailed semantic information as to the exact denotation of the location. This is what Butt and King (2004b) refer to as semantic case. It can be seen from (29), however, that the case markers do not distribute evenly with respect to GF assignment. It was seen in Chapter §5 that the clitics mē ‘in’ and par ‘on’ occur in the locative clauses surveyed, since they are the ones compatible with a stative predicate. Thus, they feature on SUBJ as well as PREDLINK functions. In addition, par occurs on OBL-TH (as a theme argument of a “picture noun”). tak ‘towards’ and ko ‘to’ do not occur in locative clauses, as indicated by their GF assignment; they only feature on obliques selected by verbal predicates and on adjuncts.
The implementation of verbal obliques in the Urdu ParGram grammar currently makes use of an underspecified $\text{OBL}$ function; a more differentiated analysis of different types of obliques selected by distinct verb classes requires additional work that is outside the scope of this thesis. There are, however, efforts towards the construction of an Urdu VerbNet that can further inform the grammar development process in this direction (Hautli-Janisz, 2014).

The lexical entry for the instrumental case marker $\textit{se}$ is given in an abridged version in (31). Recall that $\textit{se}$ is a versatile case marker ($§5.2.3$); in particular, it marks directional locations as well as more abstract sources.

\[(31) \quad \textit{se} \quad \textit{K} * \{\textit{OBL} \}
\]
\[
\quad \text{K} * \{\textit{SEM-LOC directional}\}
\]
\[
\quad \{\textit{ADJUNCT} \textit{OBL} \}
\]
\[
\quad \text{K} * \{\textit{ADJUNCT} \}
\]
\[
\quad \{\textit{OBL-SRC} \}
\].

The KP rule for case phrases is not repeated here; it was already seen in (5) in $§4.2$. What needs to be adapted is the Nadj template given in (25) in $§4.4$; its updated form is shown in (32). The KP entry in (32) has been significantly changed; it now features a tripartite distinction between the genitive (SUBJ/OBJ/ADJUNCT), locative (OBL-TH) and instrumental assignment (OBL-SRC). The head precedence constraint is stated over each of the KPs (i.e., outside of the disjunct).

\[(32) \quad \text{Nadj} = \text{KP}*: !<h
\]
\[
\quad \{(! \text{CASE}) = \text{gen}
\]
\[
\quad \{\text{SUBJ}|\text{OBJ}|\text{ADJUNCT}\}
\]
\[
\quad |\text{OBL-TH}
\]
\[
\quad |\text{OBL-SRC}
\]
\[
\quad ,\text{PRON}*: (! \text{PRON-TYPE}) = \text{poss}
\]
\[
\quad !<h
\]
\[
\quad \{\text{SUBJ}|\text{OBJ}|\text{ADJUNCT}\}
\]
\[
\quad ,\text{AP}*: \text{ADJUNCT}
\]
\[
\quad \{f::RS* = ~
\]
\[
\quad |f::RS* -= ~
\]
\[
\quad |\{\text{DT-MARK attach}\}
\]
\[
\quad ,\text{N}.
\]

---

8For example, the instrumental case is also used on oblique agents in passive constructions, a possibility not accounted for by (31), but included in the grammar.

9The templates for OBL-TH and OBL-SRC assignment, not shown here, were not part of the ParGram common templates and were defined separately.
6.5 SELECTING THE LOCATIVE

The feature system for nominals that is used in the Urdu ParGram grammar was already described in §4.3.1; the mechanism of interfacing morphology and syntax for nominals was laid out in §4.3.2. Given this setup, this section describes how different types of nouns select for locative/instrumental arguments under the present analysis.

6.5.1 COMMON NOUNS

Most common nouns in Hindi/Urdu never select locative/instrumental arguments (see §5.3). The NP rule in (32) only allows locative/instrumental KPs that are governable GFs; it does not allow such KPs to be adjuncts, contrary to the genitive pattern. Ungrammatical NPs such as the ones in (33) will naturally be ruled out since the head nominal cuha ‘rat’ is not specified with the proper subcategorization frame in terms of an OBL-TH or OBL-SRC, which means that there is a governable GF that is not governed by any predicate, resulting in incoherency.

(33) a. * mɑkan=mẽ cuha
    house.M.SG = LOC.IN rat.M.SG

b. * mɑkan=par cuha
    house.M.SG = LOC.ON rat.M.SG

Picture nouns, however, do appear with locative KP arguments. Three examples of such nouns are repeated from (36) in (34).

(34) a. mantiq=par kitab
    logic.M.SG = LOC.ON book.F.SG

    ‘book about logic’

b. pyar=par gana
    love.M.SG = LOC.ON song.M.SG

    ‘song about love’

c. gɑndʰi=par film
    Gandhi.M.SG = LOC.ON film.F.SG

    ‘film about Gandhi’

The lexical entries for kitab ‘book’, gana ‘song’ and film ‘film’ are shown in (35). The entries are noun stems inserted into the sublexical rule for nouns as described in §4.3.2 and further point to a template N-SUBJ-OBLth-med, shown in (36a); the template is a “mediator template” in that it redirects to four further templates that assign subcategorization frames to the noun. The first, N-SUBJ-OBLth, is used if the picture noun occurs with a locative OBL-TH as well a genitive SUBJ; the second and third is used if the noun

10 "mẽ ‘in’, the locative case marker in (33a), cannot be part of an OBL-TH function to begin with, as specified in (29a).
occurs with either one of these arguments suppressed; and the last simple PRED template is the default template if all arguments are suppressed.

\[(35)\]

\begin{enumerate}
\item a. kitAb \(\text{NOUN-S XLE } @(N\text{-SUBJ-OBLth-med kitAb}).\)
\item b. gAnA \(\text{NOUN-S XLE } @(N\text{-SUBJ-OBLth-med gAnA}).\)
\item c. film \(\text{NOUN-S XLE } @(N\text{-SUBJ-OBLth-med film}).\)
\end{enumerate}

\[(36)\]

\begin{enumerate}
\item a. \(N\text{-SUBJ-OBLth-med}(\_P) = \{ @(N\text{-SUBJ-OBLth } \_P) \)
\item \(\text{\textit{\_P}}\text{-}(^\_\text{SUBJ})\text{-}(^\_\text{OBL-TH})\}\). \(\text{\textit{\_P}}\text{-}(^\_\text{OBL-TH})\text{-}\text{SEM-PROP}\text{-}\text{LOCATION}) = \textit{on}.
\item \(N\text{-OBLth}(\_P) = (\^ \text{PRED}) = '\_P<(' \text{SUBJ})(' \text{OBL-TH})>'\) 
\item \(\text{\textit{\_P}}\text{-}(^\_\text{OBL-TH})\text{-}\text{SEM-PROP}\text{-}\text{LOCATION}) = \textit{on}.
\item \(\text{SUBJ_core}(\_pred) = (\^ \text{PRED}) = '_\text{pred}<(' \text{SUBJ})>'\). 
\item \(\text{PRED}(\_pred) = (\^ \text{PRED}) = '_\text{pred}'.\)
\end{enumerate}

The locative case marking on arguments of picture nouns has been analyzed as an instance of idiosyncratic/quirky case: the theme argument of such nouns is lexically required to bear the locative case marker \(\textit{par} \text{ 'on'}.\) This is implemented by using a constraining equation on those nominal subcategorization templates that include the OBL-TH function. The constraint states that the value of the feature (\(^\text{\_SEM-PROP}\text{-\_LOCATION})\) of the OBL-TH f-structure must be equal to \textit{on}; this value is only assigned by the case marker \(\textit{par} (29b)\), and its occurrence is thereby enforced. The resulting c- and f-structures for (34a) are shown in Figure 6.15.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.15.png}
\caption{Hindi/Urdu NP c- and f-structure, picture noun}
\end{figure}

There are currently 10 picture nouns in the Urdu grammar that are annotated in the same fashion as the examples in (35). Given the descriptions by \textbf{Harris (1976)} and \textbf{Soames and Perlmutter (1979)}, I expect that there are many more picture nouns in

\cite{11The templates in (36d) and (36e) are part of the common templates file of ParGram.}
Hindi/Urdu; including more of these nouns is a matter of further native speaker consultation/translation. Here, one could also definitely experiment with semi-automatic methods; for example, one could try to make use of the NomBank resource for English (Meyers et al., 2004a,b) to extract picture nouns annotated as such in NomBank and translate them to Hindi/Urdu, double-checking the results with native speakers.

### 6.5.2 Event & Relational Nouns

Event and relational nouns have to be defined in the lexicon with subcategorization frames that model their predicational properties. It was seen that event nouns and relational nouns may either select a source argument or a theme argument. Event nouns that are at the same time picture nouns can occur with a theme argument that is locative marked; other event nouns select for a source argument, realized with instrumental case. Relational nouns also occur with both types of case marking, depending on what kind of argument they occur with. An overview of the relevant patterns with example nouns for each pattern is repeated from Table 5.2, Chapter 5, below in Table 6.2. Here, the last column $n$ indicates the number of lexical entries currently included in the Urdu ParGram grammar of the respective noun type.

<table>
<thead>
<tr>
<th>Noun type</th>
<th>LI type</th>
<th>Amount</th>
<th>GF</th>
<th>Example noun</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event (&quot;picture nouns&quot;)</td>
<td>Theme</td>
<td>0–1</td>
<td>OBL-TH</td>
<td>barifing ‘briefing’ bahas ‘discussion’</td>
<td>4</td>
</tr>
<tr>
<td>Event (non-&quot;picture nouns&quot;)</td>
<td>Source</td>
<td>0–1</td>
<td>OBL-SRC</td>
<td>tabahi ‘destruction’ ravangi ‘departure’</td>
<td>12</td>
</tr>
<tr>
<td>Relational</td>
<td>Theme</td>
<td>0–1</td>
<td>OBL-TH</td>
<td>b'arosa ‘trust’ faxar ‘pride’ istisna ‘immunity’</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Source</td>
<td>0–1</td>
<td>OBL-SRC</td>
<td>pyar ‘love’</td>
<td>10</td>
</tr>
</tbody>
</table>

**Table 6.2:** Locative/instrumental KP overview with example head nominals

Lexical entries for each example noun in Table 6.2 are given below in (37); (37a&c) point to the template shown above in (36a), while (37b&d) use the templates in (38). Again, (38a) is a mediator template that redirects to (38b–c) as well as (36d–e). It can be seen that the templates constructed to model nominal predication can be reused heavily, nevertheless yielding the correct subcategorization behavior while maintaining a lean grammar in terms of the number of templates.

(37)   a. barIfing NOUN-S XLE @(N-SUBJ-OBLth-med briefing).
       b. tabAhI NOUN-S XLE @(N-SUBJ-OBLsrc-med tabAhI).
       c. bHarOsA NOUN-S XLE @(N-SUBJ-OBLth-med bHarOsA).
       d. pyAr NOUN-S XLE @(N-SUBJ-OBLsrc-med pyAr).
(38) a. $\text{N-SUBJ-OBLsrc-med}(_P) = \{ @(\text{N-SUBJ-OBLsrc} _P) \\
| @(_P) \\
| @(_P) \\
| @(_P) \}$.  

b. $\text{N-SUBJ-OBLsrc}(_P) = (^\text{PRED}) = ^_P<(^\text{SUBJ})(^\text{OBL-SRC})>$.  
c. $\text{N-OBLsrc}(_P) = (^\text{PRED}) = ^_P<(^\text{OBL-SRC})>$. 

Figure 6.16 and 6.17 show an example parse for the NP in (39). Here, the head noun $b^\text{bara}$ 'trust' is modified by a locative theme argument (OBL-TH) as well as a genitive SUBJ.

(39) ram$\text{=}k\text{a}$ armi cif$\text{=}p\text{ar}$ b$^\text{bara}$osa  
Ram.$\text{M.SG=}\text{GEN.M.SG}$ army chief.$\text{M.SG=}\text{LOC.ON}$ trust.$\text{M.SG}$  
‘Ram’s trust in the army chief’

Figure 6.17: Hindi/Urdu NP f-structure, relational noun
Recall that abstract relational nouns such as *bʰɑroṣa* ‘trust’ have to be treated as non-count nouns, since they do not pluralize. Such nouns receive an analysis in the morphological component of the Urdu grammar in terms of a +Abstract tag, as in (40a) below; the tag is added to the morphology-syntax interface as described in §4.3.2. Attached to the tag is a call to the template shown in (40b), which annotates the noun as seen in Figure 6.17. The treatment thus distinguishes such nouns from count nouns by ways of a distinct analysis in the COMMON feature.\(^{12}\)

\[(40)\]

a. bHarOsa+Noun+Abstract+Masc+Sg+Nom
b. ABSTRACT = @(COMMON abstract)
   @(NSYN common).

6.6 LI SCRAMBLING

There are cases of locative/instrumental KP scrambling (§5.4). One such case is shown in (41). In (41b), the locative KP argument *ram=par* ‘in Ram’ of *bʰarɔsa* ‘trust’ in (41a) is scrambled to a c-structural position which is clearly outside of the NP in which the KP is licensed.

\[(41)\]

a. kɑl ram=par bʰɑroṣa zaruri
   yesterday Ram.M.SG = LOC.ON trust.M.SG NOM necessary.M.SG
tʰ-a
   be.PAST-M.SG
   ‘Yesterday, trust in Ram was necessary.’
b. ram = par kɑl bʰarɔsa zaruri tʰ-a
c. kɑl bʰarɔsa zaruri ram = par tʰ-a

It was seen that LI scrambling is subject to the same constraints as genitive scrambling; all that is needed to implement scrambled LI arguments is a separate disjunct for these phrases in the meta-category KP–SCRAMBLE (§4.5, example (31)). The updated meta-category is shown in (42a); here, the OT mark dispreferring non-local attachments is annotated for all KP types (line 1); the same applies to the case constraint, stating that scrambled KPs may only occur to the left of their heads if the heads are overtly case-marked (lines 3–5). All KP types identified use the same regular predicate abbreviation KP–SCRAMBLE–PATH, which identifies the functional uncertainty path the KPs may be embedded in (42b). The uniform use of the abbreviation for all KPs is warranted since it has been shown to apply to all KPs in a uniform way.

\footnote{Event nouns, annotated as COMMON event, do pluralize, however, which entails that event nouns can be treated as a special type of count noun.}
6.7. LOCATIVE CLAUSES

When locative case appears in locative clauses, it may either be in a SUBJ or a PREDLINK GF. The case marker heads a SUBJ GF in inverted locative clauses and a PREDLINK GF in canonical locative clauses. The functional association of the located nominal (a bare KP) is the inverse: it is the SUBJ in the canonical locative clause and the PREDLINK in the inverted locative clause. Thus, locative clauses in Hindi/Urdu display a very fixed structure-function association that is otherwise not known within the language (being a free word-order language): the SUBJ appears sentence-initially, while the PREDLINK (the copula clause predicate) appears in immediately preverbal position.

Recall that the copula ho ‘be’ is homonymous with the unaccusative verb ho ‘be’, occurring in possessive clauses and “pseudo”-locative clauses. I provide the lexical entry for ho ‘be’, updated from the one used in (33) in §4.6, below in (43). The verb’s reading as an unaccusative was already discussed in §4.6; here, I provide the template for the copula reading in (43b). The template states that the copula subcategorizes for a SUBJ and a PREDLINK function. PREDLINK may either be headed by a (predicative) adjective, a nominal or a postposition (i.e., the postpositions mentioned in §2.3.3); these possibilities are controlled for by whether PREDLINK has a feature indicating the type of adjective (ATYPE), noun (NTYPE), or postposition (PTYPE). If PREDLINK is headed by a noun, the
clause may either partake in locative inversion, or it may not do so if neither the subject nor the predicate function are locative-marked. The template for locative inversion is shown in (43c); it is repeated from (5), but the PREDLINK function has been substituted for GF2 given the analysis for copula predication chosen above. Also note that copula verbs receive an analysis in terms of purely predicational verbs; this is realized by a template call to VERB-CLASS, which annotates ⟨LEX-SEM VERB-CLASS⟩ with the value pred.

(43)  

a. \( h_0 \) \( V-S \) XLE \( ⟨V-SUBJ\text{\_unacc} \ h_0⟩ \);
\( V\text{cop}\)\( -S \) XLE \( ⟨V-SUBJ\text{-PREDLINK} \ h_0⟩ \);

... 

b. \( V\text{-SUBJ\_PREDLINK}(_P) = (^\text{\_PRED})='^\text{\_P}<(^\text{\_SUBJ})('^\text{\_PREDLINK})>' \)
\( @\text{\_NON\text{-AGENTIVE} }\)
\( @\text{(VERB-CLASS pred) }\)
\( @\text{\_NOPASSIVE }\)
\( @\text{(VTYPE copular) }\)
\{\( ^\text{\_PREDLINK ATYPE} \)
\|\( ^\text{\_PREDLINK NTYPE} \)
\{\( @\text{\_LOC\text{-INV} }\)
\|\( ^\text{\_SUBJ CASE} = c \text{nom} \)
\( ^\text{\_PREDLINK CASE} = c \text{nom} \}\)
\|\( ^\text{\_PREDLINK PTYPE} = c \text{sem} \}\).

c. \( \text{LOC\text{-INV}} = \{(^\text{\_SUBJ CASE}) = c \text{loc} \)
\( ^\text{\_PREDLINK CASE} = c \text{nom} \)
\|\( ^\text{\_SUBJ CASE} = c \text{nom} \)
\( ^\text{\_PREDLINK CASE} = c \text{loc} \}\)
\( ^\text{\_SUBJ} < h (^\text{\_PREDLINK}). \)

Finally, the clausal rule for copula clauses is given in (44b). Defined as a meta-category (like \( S\text\_unacc \) for unaccusatives) and called up in the main clausal rule \( S \), it consists of a KP that is a subject, a KP/AP/NP/PP that is a PREDLINK, clausal adjuncts, scrambled KPs as well as the verbal complex rule \( V\text{main} \) (which includes the copula verb head).

(44)  

a. \( S \) \( \rightarrow \{S\text{-unerg}|S\text{-unacc}|S\text{-trans}|S\text{-cop}|S\text{-cp}\}. \)
b. $S_{\text{cop}} = KP: \emptyset \text{SUBJ}$

\[
\{KP: \emptyset \text{PREDLINK} \\
|AP: \emptyset \text{PREDLINK} \\
|NP: \emptyset \text{PREDLINK} \\
|PP: \emptyset \text{PREDLINK}\}
\]

,  
@CL-ADJUNCTS  
,  
@KP-SCRAMBLE  
,  
VCmain: ($^\sim$ LEX-SEM VERB-CLASS) =c pred.

Given these rules and templates, the Urdu ParGram grammar produces the c- and f-structures in Figure 6.18 and 6.19 for the canonical locative clause in (45a); and it parses (45b) into the structures in Figure 6.20 and 6.21.

(45) a. do baccE makan=par ñe
    two child.M.PL.NOM house.M.SG = LOC.ON be.PRES.3.PL
    ‘The two children are on the house.’

b. makan=par do baccE ñe
    house.M.SG = LOC.ON two child.M.PL.NOM be.PRES.3.PL
    ‘There are two children on the house.’

Figure 6.18: Canonical locative clause c-structure
6.8. CONCLUSION 253

"dO baccE makAn par hEN"

Figure 6.19: Canonical locative clause f-structure

Figure 6.20: Inverted locative clause c-structure

6.8 CONCLUSION

In this chapter, I have discussed how the theoretical concept of locative inversion can adequately model the alternation found in locative clauses. Once it is assumed that the Hindi/Urdu copula is a two-place predicate linking a theme and a location, locative inversion is a regular process driving an alternation on the [±o] and [±r] features in argument-GF linking. The result is an inverse assignment of GFs in canonical vs. inverted locative clauses, implying that in such clauses, the GF assignment is very much linked to structural position, a surprising finding given the otherwise free-word order properties of Hindi/Urdu.
I have further discussed competing copula analysis in LFG and settled on the PREDLINK analysis, which involves a closed complement. Reasons for choosing the PREDLINK analysis come from considerations related to agreement, the possible predicate categories, as well as cross-linguistic appeal.

LI arguments are assigned the OBL-TH and OBL-SRC function, respectively. A system of subcategorization templates was shown to model the predicational patterns accordingly. The implementation of LI scrambling reuses the existing implementation for genitive scrambling and requires only minor streamlining of the relevant grammar parts. Finally, locative clauses make use of a specific template in the grammar that puts tight constraints on case as well as linear order by using a combination of constraining equations and head precedence.

The upcoming chapter looks beyond the boundaries of noun phrases and the verb ho ‘be’. There, I am concerned with the predicational properties of nominals in noun-verb complex predicates (N-V CPs); in particular, what will be at the center of the discussion is the combinatory possibilities between nouns and verbs (so-called “light verbs” in CP constructions) in N-V CPs.
Hindi/Urdu Noun-Verb Complex Predicates

7.1 INTRODUCTION

This chapter presents two corpus studies on noun-verb complex predicates (N-V CPs). Building on previous work, I give a definition of noun-verb complex predicates, discuss light verb alternations in Hindi/Urdu and present the problems the construction poses for natural language processing. After this, I turn to a description of two separate corpus studies, conducted by myself and colleagues, which are shown to tackle the challenges posed by the N-V CP construction. Finally, the chapter describes an implementation of N-V CPs in the Urdu ParGram grammar that accurately models the empirical results of the corpus studies.

This chapter has a goal that is very different from the approach to genitive (Chapter 3) and locative/instrumental nominal arguments (Chapter 5). There, I argue that the specific predicational properties of nominals have not been well-understood; the detailed approach in distinguishing how nominals subcategorize for different grammatical functions (GF) predicts the nominals’ behavior with respect to binding as well as their occurrence in larger predicational contexts (unaccusatives as well as copula clauses). This chapter takes a different route in that the predicational properties of N-V CPs in Hindi/Urdu are in fact quite well-understood, as will be discussed in §7.2. What is less understood are the combinatorial possibilities of nouns and light verbs in N-V CPs. To understand these properties better and to start incorporating them in a computational lexicon, looking at corpora manually and relying solely on native speaker intuition will not be sufficient, as will become clear shortly (§7.2.5). Instead, automatic and semi-automatic, data-oriented methods are applied that can come up with empirical general-
izations that in turn can inform and indeed speed up the grammar development process. In turn, the results are useful as input for theoretical reasoning concerning N-V CPs, since they include empirical findings that might otherwise be overlooked.

The chapter has the following structure. §7.2 broadly defines the construction the thesis is interested in: §7.2.1 and §7.2.2 in turn discuss the elementary units that take part in the construction, nominal hosts and light verbs, while §7.2.3 illustrates the combinatory possibilities of N-V CPs that give rise to different groups of N-V CPs. §7.2.4 discusses issues related to nominal arguments in N-V CPs. §7.2.5 describes the challenges that arise from N-V CPs from the viewpoint of grammar development. §7.3 briefly summarizes related work that has been done on CP identification and processing. Sections §7.4 and §7.5 in turn describe the two corpus studies. In §7.6, I discuss how the findings can be implemented in the Urdu ParGram grammar. §7.7 concludes the chapter.

7.2 NOUN-VERB COMPLEX PREDICATES

In N-V CPs, a noun is combined with a verb to form a single predicational unit. The verb, which is usually referred to as the light verb, dictates the case marking on the subject, determines agreement patterns, carries information about tense and aspect, and adds information about the degree of agentivity (Butt, 2003, 2010). The noun, on the other hand, is the main predicational element, featuring full lexical semantics and argument structure (Ahmed and Butt, 2011, Butt et al., 2012, Mohanan, 1994). In the examples from Hindi/Urdu below, it can be seen that nouns (yad ‘memory’ in (1), pyar ‘love’ in (2)) may be combined with different light verbs.

(1) a. lɑṛki=ne kɑhani yad k-i girl.F.SG = ERG story.F.SG.NOM memory.F.SG do-PERF.F.SG
   ‘The girl remembered a/the story.’ (lit. ‘The girl did memory of a/the story.’) (Butt et al., 2012, p. 412)

   b. lɑṛki=ko kɑhani yad hɛ girl.F.SG = DAT story.F.SG.NOM memory.F.SG be.PRES.3.SG
   ‘The girl remembers/knows a/the story.’ (lit. ‘Memory of a/the story is at the girl.’) (Butt et al., 2012, p. 412)

   c. lɑṛki=ko kɑhani yad hu-i girl.F.SG = DAT story.F.SG.NOM memory.F.SG be.PART-PERF.F.SG
   ‘The girl came to remember a/the story.’ (lit. ‘Memory of a/the story became to be at the girl.’) (Butt et al., 2012, p. 412)

(2) a. nadya=ne yasin=se pyar ki-ya Nadya.F.SG = ERG Yassin.M.SG = INST love.M.SG.NOM do-PERF.M.SG
   ‘Nadya loved Yassin.’ (lit. ‘Nadya did love for Yassin.’)

   b. nadya=ko yasin=se pyar hɛ Nadya.F.SG = DAT Yassin.M.SG = INST love.M.SG.NOM be.PRES.3.SG
   ‘Nadya is in love with Yassin.’ (lit. ‘Love for Yassin is at Nadya.’)
c. nadya = ko yasin = se pyar hu-a
Nadya.F.SG = DAT Yassin.M.SG = INST love.M.SG.NOM be.PART-PERF.M.SG
‘Nadya fell in love with Yassin.’ (lit. ‘Love for Yassin became to be at Nadya.’)

In both of the examples above, the noun and the verb form a single predicational
element. In (1), kahani ‘story’ is thematically licensed by the noun yad ‘memory’, but
it is not realized as a genitive, as would be typical for arguments of nouns (see §3).
Rather, kahani ‘story’ functions as an object of the overall joint predication. In (2a) the
noun yad ‘memory’ is combined with the light verb kar ‘do’. In this case the subject
must be ergative and the overall reading is one of an agentive, deliberate remembering.
The difference between (2b) and (2c) is one of eventive vs. stative, so that in (2b), larki ‘the girl’ is already taken to be in the state of remembering the story (and not actively
entering a state of remembering it). In (2c) the light verb is the participial form of ho ‘be’ and essentially means ‘become’. The same generalizations apply to (2).

At the same time, the grammatical-functional structure of the sentences in (1) is simple,
which means that there is only a single subject, and there are no embeddings, i.e., the
structure is monoclusal in nature. This is a defining characteristic of complex predicates
in general: A complex argument structure resulting in a simple grammatical-functional
structure (Butt, 1995, Mohanan, 1994). This property sets apart complex predicates from
embedding structures such as modal verbs, sentential or infinitival complement clauses
etc. I provide Butt’s (1995) classic definition for complex predicates in (3). This is the
definition I adhere to throughout the chapter.

(3) CP definition (Butt, 1995):

a. The argument structure is complex (two or more semantic heads contribute
arguments).

b. The grammatical functional structure is that of a simple predicate. It is flat:
there is only a single predicate (a nuclear PRED) and a single subject.

c. The phrase structure may be either simple or complex. It does not necessarily
determine the status of the complex predicate.

In the case of N-V CPSs, the two semantic heads that contribute to the argument structure are a noun and a light verb, jointly forming the complex predication in a process
called predicate composition (Alsina, 1996), argument merger (Mohanan, 1994) or argument fusion (Butt, 1995); this process is taken to be a result of the syntactic composition
of two separate predicational elements, thus not applying at a lexical level, but in the
syntax (Alsina, 1996, Butt, 1995). In the literature, the noun forming part of an N-V CP is
often referred to as the nominal host. The next two sections discuss some characteristics

1There are other categories that are used as host predicates in Hindi/Urdu CPSs: verbs, adjectives, adverbs,
nouns may all feature as hosts in Hindi/Urdu CPSs (Butt, 1995, Mohanan, 1994). The focus of this chapter
is on N-V CPSs.
of nominal hosts and light verbs in turn.

7.2.1 NOMINAL HOSTS

In N-V CPS, the nominal host carries the main predication content. Together with the light verb, it jointly determines the argument structure of the resulting CP. In (1) above, the nominal host yad ‘memory’ selects the argument kahani ‘story’, which is realized as an object in the resulting clause. Although this object is thematically licensed by yad ‘memory’, it is not realized as a genitive, as would be typical for arguments of nouns. Rather, it carries bare case marking, i.e., nominative case in the system of Butt and King (2004b); see §2.3.2.

Mohanan (1994) discusses some characteristics of N-V CPS related to their constituent structure. By presenting arguments from scrambling, Mohanan derives the structure in Figure 7.1, showing that the nominal host is in fact not a daughter of the clausal S node, but forms part of the verbal constituent.

Figure 7.1: Constituency of N-V CPS (Mohanan, 1994, p. 203)

Note that even though the light verb can topicalize in an N-V CP (Mohanan, 1994, p. 204–205), the nominal host cannot. See (4a) for an example of a light verb topicalized to the beginning of the clause, and (4b) for an ungrammatical example of a topicalized nominal host.

(4)  a. ki-ya ram = ne mohan = par bʰɑrɔsa do-perf.m.sg Ram.m.sg = erg Mohan.m.sg = loc.on reliance.m.sg.nom
    ‘Ram relied on Mohan.’ (lit. ‘Ram did reliance on Mohan.’) (Mohanan, 1994, p. 205)

    b. * bʰɑrɔsa ram = ne mohan = par reliance.m.sg.nom Ram.m.sg = erg Mohan.m.sg = loc.on
    ki-ya do-perf.m.sg
    ‘Ram relied on Mohan.’ (lit. ‘Ram did reliance on Mohan.’) (Mohanan, 1994, p. 206)

Further, it is generally not possible to modify the nominal host using adjectives, coordinate the nominal host with another noun, replace the nominal host using a wh word in a

Mohanan (1994) translates bʰɑrɔsa into English ‘reliance’, while I gloss it as ‘trust’ elsewhere in this thesis.
wh-question, or attach a relative clause to the nominal host (Mohanan, 1994). Aligning this further evidence with the contrast in (4b), Mohanan (1994) argues that the nominal host cannot be a maximal projection (i.e., not a fully-fledged NP) and thus should be analyzed as a lexical category, unlike the light verb, which may topicalize (4a) and thus must be a maximal projection. The argument is that in Hindi, only maximal projections may topicalize and take part in the other structures mentioned. This in turn derives the tree in Figure 7.2, which is an augmented version of the one in Figure 7.1. Here, maximal projections are indicated by bars over the category names.

![Figure 7.2: Maximal projections in N-V CPs (Mohanan, 1994, p. 211)](image)

### 7.2.2 Light Verbs

Butt (2003, p. 1) mentions that Jespersen (1965) originally coined the term *light verb* to allude to the fact that even though these items follow the general schema of other verbs, they cannot be said to predicate fully. Butt (2003, p. 1) gives examples from English that illustrate the intermediate status of light verbs; the examples are repeated in (5), where the light verbs are highlighted in italics.

(5) a. *have* a rest, a read, a cry, a think  
    b. *take* a sneak, a drive, a walk, a plunge, a shot  
    c. *give* a sigh, a shout, a shiver, a pull, a ring, an example, a break

The light verbs’ intermediate status arises from two apparently contradictory properties. On the one hand, they seem semantically somewhat deficient when compared to other verbs. Compare, for example, the denotation of *giving someone a key*, in the sense of *physically handing over a key*, with *giving a break*. It is clear that the former involves the full predicational power of the verb *give*, while the latter is semantically weak.

On the other hand, they are not semantically empty either. Butt (2003) notes that there is a clear difference between, e.g., *take a bath* and *give a bath*. One can thus argue...
that light verbs represent a mixed category in the verbal domain, in between auxiliaries and fully-fledged verbs, and that there are clear characteristics that set light verbs apart from the other members of the domain. While it seems established that their semantics is not empty, it is rather difficult to characterize exactly what their semantic contribution is.

In the N-V CP examples seen so far, the light verbs did not act as mere verbalizers of the nouns, but made substantial contributions to the event structure of the resulting complex predicate. Reconsider the examples in (6), repeated from (1).

(6)  

a. lɑṛki=ne kahani yad k-i  
girl.F.SG = ERG story.F.SG.NOM memory.F.SG do-PERF.F.SG  
‘The girl remembered a/the story.’ (lit. ‘The girl did memory of the story.’)  
(Butt et al., 2012, p. 412)

b. lɑṛki=ko kahani yad he  
girl.F.SG = DAT story.F.SG.NOM memory.F.SG be.PRES.3.SG  
‘The girl remembers/knows a/the story.’ (lit. ‘Memory of a/the story is at the girl.’)  
(Butt et al., 2012, p. 412)

c. lɑṛki=ko kahani yad hu-i  
girl.F.SG = DAT story.F.SG.NOM memory.F.SG be.PART-PERF.F.SG  
‘The girl came to remember a/the story.’ (lit. ‘Memory of a/the story became to be at the girl.’)  
(Butt et al., 2012, p. 412)

In all of the examples in (6), the light verbs do not predicate to their full extent, but are semantically weak. All of them are form-identical, though, to a full verb in Hindi/Urdu: the light verb kɑr ‘do’, for example, has a form-identical sibling in the main verb kɑr ‘do’. In all the CP examples above, the predication would be rendered semantically infelicitous or simply ungrammatical without the predicate contribution from the nominal host. At the same time, the sentences would be ungrammatical without the light verb, since there would be no verbal item in the clause to license the arguments. Further, the choice of the specific light verb is determined by event-structural considerations: (6a) features a stative psych predicate with a dative experiencer subject, (6b) an agentive predicate with an ergative subject, and (6c) a dynamic psych predicate, again with a dative experiencer subject.

A list of Hindi/Urdu light verbs given by Kachru (2006) is shown in (7).

(7)  

Kachru (2006) distinguishes light verbs from conjunct verbs, explaining that while light verbs combine with main verbs to form compound verbs (V-V CPSs in Butt’s (2003) terminology), conjunct verbs are the ones that combine with a noun in a conjunct verb construction (i.e., in N-V CPSs), so that different sets of light verbs combine with nouns
vs. verbs. While there is some overlap between these sets, the distribution is in principle different. Thus, not all of the light verbs in (7) are relevant for the discussion in this chapter, since not all of them combine with nouns. Mohanan (1994) provides a list of light verbs known to occur with nouns; see (8).


The dependence of the light verb on the full predicate has led researchers such as Mohanan (1994), Butt (1995), and Alsina (1996) to propose a theory of complex predication where light verbs contribute to the predication in a systematic way; this is realized in argument structure composition theories put forward by the authors. The general intuition is that the light verb is syntactically as well as semantically dependent on the main predicate; in the case of N-V CPS, the main predicate is contributed by the nominal host. At the same time, the nominal host itself is selected as one of the arguments in the complex predication, as argued by Mohanan (1994) (see §7.2.1). Thus, the light verb is the main syntactic head, selecting the nominal host as an argument, and through argument composition forming a complex predicate with the nominal host, selecting a single common set of GFS.

7.2.3 CLASSES OF N-V CPS

Mohanan (1994) already identified two subclasses of N-V CPS, following previous observations by e.g., Bailey (1956). In one of Mohanan’s classes, the light verb agrees with the nominal host and, confusingly, the noun is both a syntactic object and a part of the verbal predication. For this first subclass, Mohanan (1994) has shown that one can passivize the nominal host. This is evidence that the nominal host is a GF of the overall clause, namely its object. The nominal host is thus a predicator and an argument of the overall complex predicate at the same time. An example of this type of N-V CP is given in (9); the active version is shown in (9a), the passive in (9b). Another example is in (10), where the agreement relation is clearer due to the feminine nominal host.

(9) a. ila=ne mohan=ka apman ki-ya
   ‘Ila insulted Mohan.’ (lit. ‘Ila did insult of Mohan.’)
   (Mohanan, 1994, p. 228)

b. ila=se mohan=ka apman ki-ya
gu-ya
go-PERF.M.SG
   ‘Mohan was insulted by Ila.’ (lit. ‘Insult of Mohan was done by Ila.’)
   (Mohanan, 1994, p. 228)

The passive in (10b) was adapted from the example in the source only to include the optional agent, so that it is parallel to (9b).
(10) a. atãkvadi = ne das admiyō = ki hatya
k-i do-PERF.F.SG
‘The terrorist killed ten people.’ (lit. ‘The terrorist did murder of ten people.’) (Montaut, 2016, p. 149)

b. atãkvadi = se das admiyō = ki hatya
k-i ga-i do-PERF.F.SG go-PERF.F.SG
‘Ten people were killed by the terrorist.’ (lit. ‘Murder of ten people was done by the terrorist.’) (adapted from Montaut, 2016, p. 149)

In the other subclass identified by Mohanan (1994), the light verb does not agree with the nominal host. This is an indication that the nominal host is in fact not a syntactic object of the overall clause, but is part of the verbal domain. In these CPs, the nominal host does not passivize, which shows that it is not an object of the clause. An example of such an N-V CP was already given in (6); there, the agreement pattern cannot clearly be observed, since both the nominal host as well as the object introduced by it are feminine and nominative, which is why both of them are eligible for the verb to agree with. (11a) is another example of this type of N-V CP. Here, the light verb does not agree with the nominal host, even though the host is the only nominative argument in the clause and thereby should trigger agreement. Instead, the light verb carries default masculine morphology. Passivization of the example in (11a) yields (11b); in the passivized clause, however, it is not the nominal host which ends up as the subject, but the argument provided by the host (contrary to the pattern observed in (9)). According to Mohanan (1994, p. 229–232), this second subclass is very small compared to the first.

(11) a. īla = ne mohān = ko pāsand ki-ya
‘Īla liked/approved of Mohan.’ (lit. ‘Īla did liking to Mohan.’) (Mohanan, 1994, p. 229)

b. īla = se mohān pāsand ki-ya ga-ya
‘Mohan was liked/approved of by Īla.’ (lit. ‘Mohan was done liking by Īla.’) (Mohanan, 1994, p. 231)

Ahmed and Butt (2011), who conduct a small-scale corpus study involving 45 nouns as well as three light verbs, come up with three further classes that cut across Mohanan’s (1994) morphosyntactic distinction. Their classification is based on the observation that Mohanan (1994) cites the subject tests from reflexive binding (§3.4.1) as well as pronominal coreference (§3.4.2) to show that mohān is the subject in (11a): mohān can be bound by the reflexive apna ‘self’, and it cannot be the antecedent of the pronoun. These facts are expected if mohān is the subject.
not all nominal hosts are necessarily compatible with all light verbs; instead, there are subclasses of N-V CPs based on co-occurrence patterns of nominal hosts with light verbs. Table 7.1 (taken from Butt et al. 2012) summarizes the main results presented in Ahmed and Butt 2011.

<table>
<thead>
<tr>
<th>Light verb</th>
<th>ho ‘be’</th>
<th>kar ‘do’</th>
<th>hu- ‘become’</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class A</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>psych-predications</td>
</tr>
<tr>
<td>Class B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>only agentive</td>
</tr>
<tr>
<td>Class C</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>do not allow subject to be an undergoer</td>
</tr>
</tbody>
</table>

Table 7.1: Classes of nouns identified by Ahmed and Butt (2011)

Ahmed and Butt (2011) limit their investigation to a set of three light verbs, kar ‘do’, ho ‘be’ and hu- ‘become’. Explaining this restriction, they mention that these are the most frequent and basic light verbs, and that other light verbs such as de ‘give’ or a ‘come’ have a more complex distribution and are less frequent.

In Table 7.1, Class A refers to examples such as in (1)/(6). This type encompasses what is known as psych-predications, i.e., actions of remembering, thinking, feeling, etc. However, not all nouns may be used with as many light verbs as yad ‘memory’. Class B in Table 7.1 does not allow the subject to be non-agentive. That is, it does not allow combinations with those light verbs that require a dative subject. An example for this type of noun is given in (12). This particular class was identified as by far the largest class in Ahmed and Butt (2011).

(12) a. bilal=ne makan tamir ki-ya
   ‘Bilal built a/the house.’ (lit. ‘Bilal did construction of a/the house.’) (Ahmed and Butt, 2011)

b. * bilal=ko makan tamir he
   ‘Bilal built a/the house.’ (lit. ‘Construction of a/the house is at Bilal.’) (Ahmed and Butt, 2011)

c. * bilal=ko makan tamir hu-a
   ‘Bilal built a/the house.’ (lit. ‘Construction of a/the house became to be at Bilal.’) (Ahmed and Butt, 2011)

The third class identified by Ahmed and Butt (2011), Class C, allows dative subjects in principle, but not when they are the undergoer of the action. Thus, these nouns systematically refuse to combine with hu-, the participial form of ho ‘be’. An example with the noun mtsar ‘wait’ is given in (13). Other nouns that pattern similarly are taslim ‘acceptance’ and bardaf ‘tolerance’. This was again identified as a fairly small class in Ahmed and Butt (2011).
As already mentioned, while the classes identified by Ahmed and Butt (2011) seem promising, the corpus work was done manually and was limited to a total of 45 nouns. The goal of the experiments described in this chapter is to expand the search space by using automatized methods, to thus extract information about a significant number of nouns and to be able to confirm, expand or revise Ahmed and Butt’s (2011) proposal.

### 7.2.4 Nominal Arguments

In this section, some properties pertaining to the arguments introduced by the nominal hosts are discussed. The section establishes a clear relation between the case marking properties of nouns within the NP and the way nominal arguments are case-marked in N-V CPS.

#### 7.2.4.1 Absence of Nominal Arguments

In the examples seen so far, the nominal host introduces an additional argument that is realized as an argument of the resulting complex predicate. For example, in (13a), the nominal host yad ‘memory’ introduces the argument kahani ‘story’. This is exemplary of the complex argument structure that calls for a theory of argument structure composition as put forward e.g. by Butt (1995). An additional argument supplied by the nominal host, however, does not feature in all N-V CPS; that is, the complex argument structure of the CP is not always reflected in an additional argument supplied by the noun, but may find exponence in case marking.

Take (14a), for example, which does not feature any additional argument introduced by the noun kʰãsi ‘cough’. The example nevertheless holds up to other tests of complex predicatehood. For example, topicalizing kʰãsi ‘cough’ renders the example ungrammatical (14b); also, it is not possible to coordinate kʰãsi with another noun (14c).
7.2. NOUN-VERB COMPLEX PREDICATES

(14) a. nina=ko bɑhʊt kʰāsi ɛh Nina.F.SG = DAT much cough.F.SG be.PRES.3.SG
    ‘Nina has a severe cough.’ (lit. ‘Much cough is at Nina.’)

    b. * kʰāsi nina=ko bɑhʊt ɛh cough.F.SG Nina.F.SG = DAT much be.PRES.3.SG
    ‘Nina has a severe cough.’ (lit. ‘Much cough is at Nina.’)

    ‘Nina has a severe cough and fever.’ (lit. ‘Much cough and fever is at Nina.’)

A detailed analysis of such psych predicates is developed in Sulger (2012). The dative case on the subject is licensed by the light verb ho ‘be’, and the noun kʰāsi ‘cough’ contributes the main predicate to the clause. It is thus reasonable to assume that in (14a), kʰāsi ‘cough’ is a nominal host in an N-V CP with ho ‘be’. The host here does not, however, contribute any arguments to the overall predication.

7.2.4.2 CASE MARKING

Several different types of case marking may appear on the argument provided by the nominal host: locative (4a), instrumental (2), genitive (9), as well as nominative (bare) (1) and accusative case (11). The choice between nominative and accusative case is again influenced by animacy and specificity: animate objects and specific inanimates occur in the accusative, while non-specific inanimates occur in the nominative (see §2.3.2). Mohanan (1994) argues that the difference in the two agreement patterns discussed above is reflected in the case marking: on the one hand, when the light verb agrees with the nominal host, the host’s argument bears instrumental, locative or genitive case; on the other hand, when the light verb does not agree with the nominal host, the host’s argument carries either nominative or accusative case, as per the animacy/specificity contrast. This correlation between agreement and case marking on the nominal host argument is particularly clear in cases where both patterns are possible, as in (15).

(15) a. us=ne mohan=ko yad PRON.3.SG.OBL = ERG Mohan.M.SG = ACC memory.F.SG.NOM do-PERF.M.SG
    ‘She/He remembered Mohan.’ (Mohanan, 1994, p. 231)

    b. us=ne mohan=ki yad PRON.3.SG.OBL = ERG Mohan.M.SG = GEN.F.SG memory.F.SG.NOM
    k-i
do-PERF.F.SG
    ‘She/He remembered Mohan.’ (Mohanan, 1994, p. 231)

The light verb *kar* ‘do’ is dyadic and features a subject and an object. In (15a), *mohan* is the overall object (as shown by passivization), while *yad* ‘memory’ does not map to a GF of the main clause at all. In (15b), on the other hand, the host maps to the overall object of the clause and triggers agreement on the verb. Mohanan (1994) ultimately traces the different behavior in terms of agreement and case down to a lexical property of the nominal host: some hosts feature only in combinations without host-verb agreement, while others occur only in combinations where such agreement is enforced; still other nouns may occur in both patterns. *yad* ‘memory’ thus falls in the latter category, as evidenced by (15); *bʰarọsa* ‘trust’, on the other hand, is an example of a noun that only appears in the agreeing type, cf. (16a) vs. (16b), while *pɪnc* ‘pinch’ only partakes in the pattern lacking agreement between the host and the light verb, cf. (17a) vs. (17b).

I would like to argue, with Mohanan (1994), that the case on the nominal argument is very much tied to the lexical entry of the nominal host. In particular, I suggest that if the nominal host subcategorizes for an instrumental obl-src or locative obl-th argument, as per the account given in §5, it realizes that argument in the N-V CP, as well, with the accompanying case marking. The nominal host itself is then realized as the overall object, and the verb agrees with it, while the obl-src/obl-th GFs are realized as part of the main clause. Consider the examples in (18); all of the nominal hosts here can appear with a locative or an instrumental argument, and in all of the sentences the nominal host is the overall object. The nominal argument cannot appear with other types of case marking.
b. \( \text{ram} = \text{ne} \quad \text{yasin} = \text{par} \quad */ = \text{se} \quad */ = \text{ka} \quad */ = \text{ko} \)
\[ \text{Ram.M.SG} = \text{ERG} \quad \text{Yassin.M.SG} = \text{LOC.ON} \quad */ = \text{INST} \quad */ = \text{GEN.M.SG} \quad */ = \text{ACC} \]
\[ */ = \emptyset \quad \text{bhorosa} \quad \text{ki-ya} \]
\[ */ = \text{NOM} \quad \text{trust.M.SG.NOM} \quad \text{do-PERF.M.SG} \]
‘Ram trusted Yassin.’ (lit. ‘Ram did trust in Yassin.’)

c. \( \text{bilal} = \text{ko} \quad \text{baccō} = \text{se} \quad */ = \text{par} \quad */ = \text{ka} \)
\[ \text{Bilal.M.SG} = \text{DAT} \quad \text{child.M.PL.OBL} = \text{INST} \quad */ = \text{LOC.ON} \quad */ = \text{GEN.M.SG} \]
\[ */ = \text{ACC} \quad */ = \text{NOM} \quad \text{hate.F.SG.NOM} \quad \text{be.PART-PERF.F.SG} \]
‘Bilal came to hate (the) children.’ (lit. ‘Hate for (the) children became to be at Bilal.’)

On the other hand, if the nominal host does not subcategorize for an OBL-SRC or OBL-TH, the case on the nominal argument defaults to nominative/accusative or genitive case, as in (19a) (= 12a) or (19b) (= 13b), respectively; exactly what drives the choice between these two case assignments will have to be resolved in future work, but again, the assumption is that the choice between the case marking (and hence, the resulting overall syntactic pattern) is lexically fixed.

(19)

a. \( \text{bilal} = \text{ne} \quad \text{makan} \quad \text{tamir} \quad \text{ki-ya} \)
\[ \text{Bilal.M.SG} = \text{ERG} \quad \text{house.M.SG.NOM} \quad \text{construction.F.SG} \quad \text{do-PERF.M.SG} \]
‘Bilal built a/the house.’ (lit. ‘Bilal did construction of a/the house.’)
\[ \text{(Ahmed and Butt, 2011)} \]

b. \( \text{bilal} = \text{ne} \quad \text{nadya} = \text{ka} \quad \text{intizar} \quad \text{ki-ya} \)
\[ \text{Bilal.M.SG} = \text{ERG} \quad \text{Nadya.F.SG} = \text{GEN.M.SG} \quad \text{wait.M.SG.NOM} \quad \text{do-PERF.M.SG} \]
‘Bilal waited for Nadya.’ (lit. ‘Bilal did waiting of Nadya.’)
\[ \text{(Ahmed and Butt, 2011)} \]

### 7.2.4.3 Absent Nominal Arguments

The argument supplied by the nominal host may not be realized in a clause. That is, it is perfectly acceptable to utter sentences such as the ones in (20b–22b), which feature absent nominal arguments compared to the a. examples (which are repeated from above).

(20)

a. \( \text{nina} = \text{ne} \quad \text{us} = \text{par} \quad \text{bhorosa} \quad \text{ki-ya} \)
\[ \text{Nina.F.SG} = \text{ERG} \quad \text{PRON.3.SG.OBL} = \text{LOC.ON} \quad \text{trust.M.SG.NOM} \quad \text{do-PERF.M.SG} \]
‘Nina trusted her/him.’ (lit. ‘Nina did trust in her/him.’)

b. \( \text{nina} = \text{ne} \quad \text{bhorosa} \quad \text{ki-ya} \)
\[ \text{Nina.F.SG} = \text{ERG} \quad \text{trust.M.SG.NOM} \quad \text{do-PERF.M.SG} \]
‘Nina trusted.’ (lit. ‘Nina did trust.’)

(21)

a. \( \text{nadya} = \text{ko} \quad \text{yasin} = \text{se} \quad \text{pyar} \quad \text{he} \)
\[ \text{Nadya.F.SG} = \text{DAT} \quad \text{Yassin.M.SG} = \text{INST} \quad \text{love.M.SG.NOM} \quad \text{be.PRES.3.SG} \]
‘Nadya is in love with Yassin.’ (lit. ‘Love for Yassin is at Nadya.’)
b. nadya = ko pyar he
Nadya.F.SG = DAT love.M.SG.NOM be.PRES.3.SG
‘Nadya is in love.’ (lit. ‘Love is at Nadya.’)

(22) a. lɑṛki = ko kahani yad hu-i
girl.F.SG = DAT story.F.SG.NOM memory.F.SG be.PART-PERF.F.SG
‘The girl came to remember a/the story.’ (lit. ‘Memory of a/the story became
to be at the girl.’) (Butt et al., 2012, p. 412)
b. lɑṛki = ko yad hu-a
girl.F.SG = DAT memory.F.SG be.PART-PERF.M.SG
‘The girl came to remember.’ (lit. ‘Memory became to be at the girl.’)

However, the sentences in (20b–22b) must still be analyzed as N-V CPS, since they display
the usual properties of the construction. For all of these sentences, it is bad to topicalize
or coordinate the host (23–25); also, the case marking on the subjects could not possibly
be assigned by a main verb (i.e., non-light verb) ho ‘be’, but must be regarded as the
result of the complex predication as described above.

(23) a. * bʰɑrɔsa nina = ne kɪ-ya
trust.M.SG.NOM Nina.F.SG = ERG do-PERF.M.SG
‘Nina trusted.’ (lit. ‘Nina did trust.’)
b. * nina = ne bʰɑrɔsa ɔr pyar kɪ-ya
‘Nina trusted and loved.’ (lit. ‘Nina did trust and love.’)

(24) a. * pyar nadya = ko he
love.M.SG.NOM Nadya.F.SG = DAT be.PRES.3.SG
‘Nadya is in love.’ (lit. ‘Love is at Nadya.’)
b. * nadya = ko pyar ɔr bʰɑrɔsa he
Nadya.F.SG = DAT love.M.SG.NOM and trust.M.SG.NOM be.PRES.3.SG
‘Nadya is in love and trusts.’ (lit. ‘Love and trust are at Nadya.’)

(25) a. * yad lɑṛki = ko hu-a
memory.F.SG girl.F.SG = DAT be.PART-PERF.M.SG
‘The girl came to remember.’ (lit. ‘Memory became to be at the girl.’)
b. * lɑṛki = ko yad ɔr pyar hu-a
girl.F.SG = DAT memory.F.SG and love.M.SG be.PART-PERF.M.SG
‘The girl came to remember and to love.’ (lit. ‘Memory and love became
to be at the girl.’)

There are two possible explanations for the absence of the nominal arguments. Under
the first explanation, the nominal arguments undergo argument suppression. Given the
account that was established in the preceding chapters, this would be expected: nominal
arguments routinely undergo argument suppression, which existentially binds the
argument (i.e., its presence is implicitly understood), but leads to it not being overtly
realized (Barker, 1995).
On the other hand, we are in the verbal domain here, where it is well-known that Hindi/Urdu makes rampant use of pro(nominal)-drop. Pro-drop and some of the conditions that drive pro-drop were already discussed in §3.5.3. In a nutshell, any and all of the verbal arguments may be dropped in Hindi/Urdu under the right circumstances; discourse/information structure plays an important role in that it is predominantly given (i.e., topic) or old information that may be dropped (Butt, 2007; Butt and King, 2000). The absent arguments in (20b–22b), even though licensed by nominals, end up as arguments of the joint predication, and thus their absence could also be explained by pro-drop.

I am not going to resolve this theoretical conundrum here. I do, however, state that the subcategorization templates defined in Chapters §4 and §6 already handle argument suppression, so that if one is to assume an analysis of the patterns in terms of suppression, this particular aspect of N-V CPS is not problematic. The current implementation of N-V CPS realizes the suppression analysis. Data such as in (20) do, however, pose a serious problem with respect to grammar development in terms of distinguishing CPS from non-CPS instances (see §7.2.5.2), for reasons that are independent from the question of whether the data stem from argument suppression or pro-drop.

7.2.5 CHALLENGES FOR GRAMMAR DEVELOPMENT

In Hindi/Urdu, there are about 700 simple verbs (Humayoun, 2006), while the remaining verbal inventory consists of CPS. Thus, Hindi/Urdu augments its verbal predication by using CPS, and CPS are encountered frequently in general language use, as well as in newspaper corpora. For example, in the Hindi/Urdu treebank (HUTB, Bhatt et al., 2009, Palmer et al., 2009), almost 37% of the predicates have been annotated as light verb constructions (Vaidya et al., 2014). Thus, any NLP research effort dealing with Hindi/Urdu, whether shallow or deep, whether its goal be parsing, generation, question-answering or the construction of lexical resources like WordNet (Bhattacharyya, 2010) or PropBank (Vaidya et al., 2011), encounters complex predication sooner rather than later.

7.2.5.1 GENERAL CHALLENGES OF THE CONSTRUCTION

One challenge with respect to grammar development is to model the predicational properties of N-V CPS in an appropriate fashion. At the outset of LFG and, indeed, at the outset of computational grammar development in the ParGram project, the need for detailed predicate composition in terms of adding and combining different predicates and their subcategorization frames was not anticipated (Butt et al., 2003). For alternations over lexical items, lexical rules can be formulated and implemented in a straightforward fashion in XLE. One example is passivization, where GFS can either be renamed or suppressed: the OBJ is renamed SUBJ, the SUBJ is either renamed OBL-AG (for oblique agent)

---

7See, for example, the XLE documentation for lexical rules at http://ling.uni-konstanz.de/pages/xle/doc/notations.html#NƩ.Ʀ.
or suppressed, in case of the short passive (Bresnan, 1982d; Kaplan and Bresnan, 1982).

In N-V CPS (as in other types of complex predications), on the other hand, lexical rules do not suffice to encode the predicational properties. First, as was seen above, the nominal host may contribute to the overall argument structure in terms of supplying an additional GF; lexical rules as motivated in standard LFG and implemented in XLE do not allow for the addition of GFs. Second, the nominal host itself may either end up as the main clause object, or it may end up as a part of the main predicate; the latter case is not feasible using lexical rules, because lexical rules operate on lexical entries (verbs, typically), not on combinations of predicates.

To overcome this shortcoming in the LFG arsenal, Kaplan and Wedekind (1993) propose the restriction operator ‘\’, which allows for the restricting out of some attribute (including its value) from some F-structure. A simple example of the notation is shown in (26); here, the TENSE feature is restricted out from the F-structure f. If f is the F-structure shown in Figure 7.3a, the result of (26) is the F-structure shown in Figure 7.3b, which is equal to Figure 7.3a except that it lacks the feature TENSE.

(26) \TENSE

\[
\begin{align*}
  f : & \quad \text{SUBJ} [ \text{PRED ‘Chris’} ] \\
  & \quad \text{TENSE past}
\end{align*}
\]

\[
\begin{align*}
  f - \TENSE : & \quad \text{SUBJ} [ \text{PRED ‘Chris’} ]
\end{align*}
\]

(a) Complete F-structure prior to restriction

(b) F-structure after restriction

Figure 7.3: Simple restriction example, adapted from Dalrymple (2001, p. 162)

In its original formulation, restriction was taken to apply to lexical F-descriptions and not in the syntactic component; this made it possible to account for structural head mismatches well-known from machine translation, e.g., the adjunct-main verb mismatch between English *The baby just fell.* vs. French *Le bébé vient de tomber.* (Kaplan and Wedekind, 1993). However, this early implementation did not allow for the phrasal combination of CPS; in particular, it was not possible to productively combine several instances of restriction in a recursive manner, as required by instances of CP stacking discussed by Butt (1994). Restriction therefore was reimplemented in XLE; since then, the restriction operator can be used productively within the syntactic component, e.g., in functional annotations on phrasal nodes. §7.6.1 describes the basics of the current implementation of N-V CPS in the Urdu ParGram grammar.

---

8Strictly speaking, recursive composition could have been realized under the earlier implementation of restriction by stating the anticipated possible and impossible combinations in the lexical entries of the light verbs; this would essentially mean that one would have to encode (dynamic) phrasal combinatory possibilities in (static) lexical entries, which is intuitively not desirable.
7.2.5.2 Distinguishing cps from Non-cps

The data discussed in §7.2.4.3 pose a problem when the main verb counterparts of light verbs are taken into consideration. To see why this is the case, recall that several light verbs, just as their main verb relatives, require their subject to be agentive. This leads to contrasts such as the one in (27) ((27a) is repeated for convenience from (20a)). The sentences look similar on the surface; however, (27a) must be analyzed as an N-V CP, while (27b) is not a CP, but involves the main verb kar ‘do’ with a simple argument structure.

(27)

a. nina=ne bʰarɔsa ki-ya
   Nina.F.SG = ERG trust.M.SG.NOM do-PERF.M.SG
   ‘Nina trusted.’ (lit. ‘Nina did trust.’)

b. nina=ne hɔmwɑrk ki-ya
   Nina.F.SG = ERG homework.M.SG.NOM do-PERF.M.SG
   ‘Nina did a/the homework.’

bʰarɔsa ‘trust’ and kar ‘do’ form a legitimate N-V CP, but hɔmwɑrk ‘homework’ and kar ‘do’ do not; the information about the ability to function as a nominal host in a given context, must be part of the lexical entry of a given nominal. This also applies to the Urdu ParGram grammar, where the lexical entries of nouns can be augmented to include such information.

7.2.5.3 Productivity

CP formation is a very productive process in Hindi/Urdu. For example, nouns borrowed into Hindi/Urdu can instantly partake in N-V CP formation. With respect to NLP applications, the frequency and productivity of complex predication means that it is not possible to construct a static list of N-V combinations. Rather, there must be a way in which one can dynamically predict which kinds of combinations are possible and which should be impossible. In particular, understanding the constraints on N-V CP formation in more detail would be welcome for further NLP applications. To implement these possibilities and to extend grammar coverage in this area, one can pursue two different approaches. One can either investigate in detail the syntactic and lexical-semantic principles that are behind N-V CP formation, to start building hypotheses about the constraints. Once these hypotheses have been validated, one can then start to think about how to implement them in the lexicon or grammar. The corpus study reported in §7.4 is a contribution in this direction.

Another possibility, which is at the focus of the second corpus study reported in §7.5, is more shallow, but at the same time faster and easier to implement. Here, the main goal is not to derive the constraints on N-V CP formation via the lexical-semantics of their single parts, but to derive light verb usage from actual corpus evidence and group together individual nouns accordingly. The behavior of the nouns found in the corpus
can be used as input for the construction of grammar templates that model the nouns’ behavior with respect to light verb co-occurrence.

7.2.6 SUMMARY

In total, these facts about the syntax of Hindi/Urdu N-V CPS enable us to delimit the frame of N-V CPS to the following generalizations:

(28) a. An N-V CP in Hindi/Urdu consists of a nominal host and a light verb.

b. Generally, N-V CPS occur in clause-final position, and their sequence is nominal host + light verb. (Mohanan, 1994)

c. The list of light verbs available to N-V CP formation is limited and they form a separate syntactic class. (Butt, 2003)

d. All light verbs have direct, form-identical counterparts among the class of main verbs. (Butt, 2003)

e. Nominal hosts are nouns that do not project maximally, but are in general an open lexical class. (Mohanan, 1994)

f. Nominal hosts may or may not contribute additional arguments to the complex predication; at the very least, they contribute by licensing specific case marking on the arguments of the CP. (Butt, 2003)

When planning a corpus study, it is found that the generalizations in (28a–c) are rather beneficial for defining the search space. It is known where to look for the combinations, and it is also known what to look for. Less beneficial are the generalizations in (28d–f). The nominal hosts available to N-V CP formation form no coherent and well-defined semantic class; also, light verbs may easily be confounded with their main verb counterparts, which makes it difficult to distinguish instances of actual complex predication from combinations involving main verbs. The corpus studies reported on in Sections §7.4 and §7.5 aim at overcoming these obstacles.

7.3 RELATED WORK

Before I describe the two corpus studies that were designed to shed light on N-V CPS and their combinatory possibilities, I briefly present related work on the topic of identifying and classifying CPSs for NLP purposes in Hindi/Urdu as well as Persian, another language that makes use of CPSs.

Ahmed (2010) looked at V-V CPSs in Urdu. In V-V CPSs, the main predicational element is a main verb; the following light verb is of the same type as the light verbs discussed in §7.2.2. Ahmed (2010) manually examines the acceptability of verb classes with the light verbs ja ‘go’, le ‘take’ and de ‘give’. The results show that the light verb de ‘give’

---

9The canonical word order in Hindi/Urdu is head-final; see §2.3.1.
occurs preferentially with verbs of bodily expression, e.g. *hans* ‘laugh’ or *cix* ‘scream’, and verbs of sending away, e.g. *pʰenk* ‘throw’. On the other hand *ja* ‘go’ occurs more preferentially with change of state verbs such as *gr* ‘fall’. These results are promising as they indicate that similar behavior may be observable for nouns as well.

Bhattacharyya et al. (2006) describe the construction of a lexical knowledge base, i.e., a WordNet, for Indian languages. Their special focus is on Hindi. N-V as well as V-V complex predicates are identified linguistically and stored computationally in the lexical resource. Empirical tests are constructed to decide if a combination of two words, the second of which is a verb, is a complex predicate or not. Bhattacharyya et al. (2006) argue that such tests provide a principled way of deciding the status of complex predicates in Indian languages. Their tests include evidence from topicalization, coordination, adjectival modification as well as wh-questions (i.e., most of the tests already identified by Mohanan (1994)). They choose to encode CPs as separate entries in their lexical database.

A recent study on identifying Persian N-V CPs, Taslimipoor et al. (2012), uses distributional vector-space methods and shows that verb vectors are indeed a very useful indicator of CP status. Their base assumption is that, if a given noun N in a low-frequency (or novel) candidate N-V combination is semantically similar to a noun that tends to form (high-frequency, known) N-V CPs with the verb V, then it is more likely that the candidate N-V is an acceptable CP as well. Semantic similarity is measured by taking into account a vector-space model that is based on frequent words in a 10-word window around the target noun. They compare a model that it based on noun similarity only vs. another model that incorporates verb vectors, and find that the results are significantly better using the light verb dimension; Taslimipoor et al. (2012) state that this affirms their original intuition that a verb-based vector space model can better capture similarities across CPs. This finding is in agreement with our intuition that features based on light verbs best capture generalizations about N-V CPs.

### 7.4 Corpus Study 1: The “Brute Force” Approach

Here, the idea was to take advantage of statistical methods and proceed per standard methods currently embraced in the field. That is, use an available corpus that should in principle be large enough to adequately reflect language use and to extract bigrams from this corpus in order to identify patterns in N-V combinations and to use our knowledge of the extracted patterns in further NLP applications.

In pursuing this experiment, we were indeed able to adduce new information about combinatory possibilities in CPs (§7.4.3). However, our experiment also provides a cautionary tale with respect to diving into a corpus “blindly”, i.e., assuming that mere statis-
tical analysis will provide good enough results and any noise due to language particular considerations will simply wash out if the corpus is large enough. Some of the difficulties we encountered had to do with the non-standardized nature of Urdu orthography, some with the structure of Urdu (§7.4.1) and some with the complex nature of the data (§7.4.2).

With respect to the complex interrelationships between our data, we were able to achieve significant improvement by using methods coming from the field of visual analytics (Card et al., 1999, Keim et al., 2010, 2008, Thomas and Cook, 2005). Instead of trying to make sense of bare numbers, we used a visualization tool that maps figures to colors and therefore makes the statistical analysis immediately accessible via easy to process visual means. The visualization allowed us to assess our complex data “at-a-glance”, to take corrective measures and to generate new hypotheses as to CP formation and the validation of existing hypotheses.

7.4.1 Methodology

In this section, I describe the methodology used for harvesting CP candidates from a raw Urdu text corpus and for identifying classes among the candidate CPs via a simple bigram analysis in conjunction with techniques from visual analytics. I also discuss some problems we discovered to be associated with working with an unannotated “raw” corpus for a language like Urdu. The overall results make a strong case for the prioritization of (mainly manual) high quality resource building — it seems that significant progress with entirely shallow methods cannot be achieved unless high quality, linguistically informed resources can be drawn upon.

7.4.1.1 Corpus

A prerequisite for our experiment is access to a large corpus for Urdu. No large corpus for Urdu (annotated or not) is publicly available to date; the resource described by Urooj et al. (2012) was under development at the time we were conducting this study. We therefore decided to use a 7.9 million word corpus we had been harvesting from the bbc Urdu website for a number of months. The corpus consists of news articles on various different topics, e.g., entertainment, multimedia, science, sports. The articles were parsed into raw text using the Perl HTML module. Inspection of the corpus showed that the BBC Urdu script encoding is particularly clean and systematic in comparison to other Urdu newspaper sites. We therefore did not clean or preprocess the corpus with respect to punctuation, orthography or other normalization issues. In Figure 7.4 below, I provide the first 4 lines of the resulting raw text corpus file.

11 http://www.bbc.co.uk/urdu/
12 http://search.cpan.org/dist/HTML-Parser/Parser.pm
7.4. CORPUS STUDY 1: THE “BRUTE FORCE” APPROACH

As a first pass, we went through the corpus and extracted all the bigrams containing the four verbs *kar* ‘do’, *ho* ‘be’, *hu-* ‘become’ and *rakʰ* ‘put’. We decided to extend our set of light verbs beyond the ones used in Ahmed and Butt (2011) by including *rakʰ* ‘put’ in the hopes of arriving at a finer-grained picture of the distribution of the nouns in N-V CPs. We thus looked for all instances of these verbs (in all of their conjugated forms) and extracted them plus any word immediately preceding them. These bigrams were stored and their frequency was recorded. Note that by extracting bigrams, we adhered to the definition of the N-V CP pattern as given in 7.2, which states that the nominal host and the light verb are part of a fairly fixed phrasal configuration (i.e., they are both part of a single maximal verbal projection).

This procedure yielded four initial lists of bigrams, one for each of the four verbs *kar* ‘do’, *ho* ‘be’, *hu-* ‘become’ and *rakʰ* ‘put’. A manual inspection of these lists revealed that while we were finding N-V CPs of the type we were looking for, most of the highly frequent bigrams were either junk or not the kinds of combinations we were trying to find. An example of the former is given in (29), while (30) includes examples of the latter; the right-justified numbers indicate $n$ (overall frequency of the bigram). Recall that our corpus consisted of raw, untokenized text, which is why (29) includes a punctuation mark (i.e., a sentence boundary), thus not representing a valid bigram. We also had an issue with low frequency in that many bigrams were recorded just once or twice.

(29)

\[
\text{be.pres.3.sg period do} \quad n = 81
\]

(30)

\[
\begin{align*}
\text{a. ka bo} & \quad n = 393 \\
= \text{ka ho} & = \text{gen.m.sg be} \\
\text{b. vo ho} & = \text{pron.3.sg be} 
\end{align*}
\]

I designed the Perl scripts used for the bigram extraction, producing the initial four bigram lists.
Since we were working with a raw (i.e., untagged or otherwise preprocessed) corpus, it was clear to us from the outset that we would not end up with particularly clean data. However, we were not prepared for the amount of false positives we did get. Closer inspection of the bigram lists revealed that many of the false positives were due to certain case markers, conjunctions and pronouns, which all occur frequently before our set of verbs. This is actually to be expected, given the structure of the language (see §7.4.1.4 for further discussion). Since our set of verbs is very versatile in that they can not only act as light verbs in N-V CPs, but also function as main verbs and as auxiliaries, many of our top bigrams turned out to be verb-verb combinations of one type or another. Although these are all valid bigrams, they are unwanted noise in the context of our investigation. Further errors were introduced by punctuation and tokenization or white-space issues (see §7.4.1.4).

As a consequence, the initial bigram lists, which consisted of 16033 possible combinations, were pruned. For one, all bigrams which appeared less than six times in the corpus were removed. We assumed that most of them were tokenization errors and other machine processing lapses. The value of six turned out to be useful for pruning after manual inspection using several different values. We also removed all those bigrams which had a negative $\chi^2$ value (see below). This reduction left us with only around 4500 bigrams. In a second step, we constructed a list of stop words from various sources. For one, we removed any bigrams containing problematic closed class items such as case markers, conjunctions and pronouns. For another, we used a verb list containing all the conjugated forms of about 700 simple verbs in Urdu (15285 verb forms in total, from the verb conjugator in Raza 2011) in order to remove all those bigrams which we could identify as v-v sequences via this verb list.

After applying these steps, we were left with just 2154 candidate bigrams from the original 16033 bigram possibilities drawn from the raw corpus. These candidate bigrams still contained problematic items (see §7.4.2), but identifying and removing them at this stage would have involved intense manual inspection and labor. We therefore decided to work with this list of bigrams for further analysis by different methods.

As a first analysis step, the association strength between the bigram members was computed using the Chi-Square ($\chi^2$) measure. This ensures that the more often one of our light verbs occurs with a certain word compared to all other words, the stronger the association is and the higher the bigram is ranked among the group of bigrams. The statistics were computed by means of the ucs toolkit. We decided to use the $\chi^2$ association measure to determine the positive or negative association between the words

---

14 The verb-verb combinations were not, however, v-v complex predicates of the type discussed by Butt (1995), since these occur with a set of light verbs different from ours; see also the discussion in §7.2.2.

in the bigram for two reasons. First, papers using comparatively sized corpora have reported encouraging results for similar experiments (Hautli and Sulger, 2011, Kizito et al., 2009, Ramisch et al., 2008). Second, initial manual comparison between bigram lists ranked according to all measures implemented in the UCS toolkit revealed the most convincing results for the $\chi^2$ test. Based on the ranking, we reduced the list of bigrams and discarded all bigram instances with either a negative $\chi^2$ or a frequency below 6 (see above). A negative $\chi^2$ value indicates a negative word association, i.e. the bigram members do not occur together very often in comparison to their frequency in other bigrams.

At this point, we still had our four separate lists of bigrams for the verbs *kar* ‘do’, *ho* ‘be’, *hu* ‘become’ and *rakʰ* ‘put’, but the lists were now of ranked bigrams. Manual inspection revealed that the top items in these lists now did contain *N-V* combinations of the type that Ahmed and Butt (2011) were looking at, among them also some of the nouns that were discussed in that paper. Our various steps of filtering and ranking thus did allow us to extract a list of strongly associated lexical items of the right type. The next step was to proceed to an analysis of our extracted data. Since our interest lay in the determination of possible classes of *N-V* CPS, we decided to run a clustering algorithm on our data.

### 7.4.1.3 Automatic Clustering

Based on the filtered and ranked lists of bigrams from §7.4.1.2, we investigated the occurrences of words with light verbs across the different types of bigram combinations, i.e., we combined the information of the four lists into one list, recording the light verb behavior of every single word. That is, we had different classes of words: words occurring with all four light verbs or words with only *kar* ‘do’ and *ho* ‘be’, *hu* ‘become’ and *rakʰ* ‘put’ or words occurring with various combinations of these verbs. Table 7.2 shows a sample matrix of four nouns (*hasɪl* ‘achievement’, *alan* ‘announcement’, *bat* ‘talk’ and *ʃuru* ‘beginning’ in that order) and their relative frequency of co-occurrence with the four light verbs, i.e. out of all occurrences of noun 1 (*hasɪl* ‘achievement’) with one of the four light verbs, the relative frequency of it occurring with *kar* ‘do’ is 0.771.

<table>
<thead>
<tr>
<th>ID</th>
<th>Noun Transliteration/gloss</th>
<th><em>ho</em> ‘be’</th>
<th><em>kar</em> ‘do’</th>
<th><em>hu</em> ‘become’</th>
<th><em>rakʰ</em> ‘put’</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>حاصل <em>hasɪl</em> ‘achievement’</td>
<td>0.222</td>
<td>0.771</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>اعلان <em>alan</em> ‘announcement’</td>
<td>0.011</td>
<td>0.982</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>بات <em>bat</em> ‘talk’</td>
<td>0.147</td>
<td>0.853</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>شروع <em>ʃuru</em> ‘beginning’</td>
<td>0.384</td>
<td>0.530</td>
<td>0.086</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 7.2: Relative frequencies of co-occurrence of nouns with light verbs

---

16 UCS implements the one-sided version of the Chi-Square test for word association, where positive scores indicate positive association, and negative scores indicate negative association. See the UCS documentation at [http://www.collocations.de/UCS/UCS-Perl-html/UCS/AM.html](http://www.collocations.de/UCS/UCS-Perl-html/UCS/AM.html).
Note that although the motivation for our experiment stemmed from an interest in N-V
CPS, our bigrams in fact contain all kinds of POS in combination with our set of four verbs
(minus the bigrams that were removed during the pruning steps described in §7.4.1.2).
This is a direct and unavoidable consequence of using an untagged “raw” corpus.

Based on the pattern of relative co-occurrence with the four light verbs, the results
were clustered automatically. This was done by myself using a data mining platform de-
veloped at the University of Konstanz: KNIME. We used a standard k-means clustering
algorithm (MacQueen, 1967) in order to assign each noun to a cluster within KNIME.
k-means is used to group similar objects (here: the nouns) into k clusters based on their
behavior with respect to the given features (here: the light verb relative frequencies).
Note that the number k of clusters the objects are assigned to needs to be specified in
advance. Here, we experimented with different values between two and seven. How-
ever, just like the raw numbers as illustrated in Table 7.2, the quality of the resulting
clusters was difficult to measure in the absence of proper evaluation techniques. For
this corpus study, we therefore decided to experiment with techniques involving visual
analytics (e.g., Collins, 2010, Dörk et al., 2008, Mayer et al., 2010). The assumption was
that the quality of the clusters can be sufficiently estimated by a human by manually
inspecting the clusters when aided by visualizations of the data. Before moving to a
discussion of these techniques in §7.4.2, we take a step back and consider the language
particular knowledge we had to rely on so far.

7.4.1.4 DISCUSSION: LANGUAGE PARTICULAR ISSUES

As mentioned in §7.4.1.2, our first pass at bigram extraction resulted in a large number
of false positives. However, this was to be expected. Urdu is a language which is not
particularly morphologically complex, but it does use a significant amount of morphol-
ogy. One unfortunate feature from the perspective of NLP is that the same material is
used for several different purposes. For example, -a, -i and -e are morphemes used to
mark gender and number on nouns, adjectives, verbs, participles as well as the genitive
case. Additionally, there is significant homonymy with respect to frequently used words.
For example, one that had a significant impact is the perfect masculine form of kar ‘do’,
kiya, which is generally written the same way as the interrogative pronoun kya ‘what’.
Another example is the perfect feminine form ki of kar ‘do’, which is written the same
way as the complementizer ki ‘that’ as well as the feminine singular form of the genitive
case marker.

The genitive case marker in general posed a problem. It is structurally a clitic (Butt
and King, 2004b) and is written as a separate word in Urdu. Given that it is a genitive
case marker, it is naturally found adjacent to nouns. For us this meant that we extracted
many bigrams which turned out to be collocations of the feminine singular genitive

17http://www.knime.org; KNIME is freely available for research purposes.
18The i in kya, the perfect masculine form of kar ‘do’, is realized as a diacritic. Since diacritics rarely
occur in standard written Urdu, the form looks identical to the interrogative pronoun kya ‘what’.
marker and a noun. We therefore decided to remove all instances of bigrams with *ki*. This meant that we probably lost many “good” bigram candidates, but we did not see a way of filtering out the “bad” instances of the genitive while keeping the good instances of the perfect feminine singular of *kar ‘do’*.19

As already mentioned, our set of four light verbs *kar ‘do’, ho ‘be’, hu- ‘become’ and rakʰ ‘put’* can be employed as simple verbs in Urdu as well; moreover, *ho ‘be’, and hu- ‘become’* can also function as auxiliaries. This meant that our initial bigram extraction netted many v-v sequences in which an item of our set of four verbs occurred as an auxiliary after a main verb. We dealt with this by employing a list of verbs along with all of their inflections that was constructed as part of the work done by Raza (2011). Since this encompassed a total of 15285 verb forms, having access to this already existing resource was invaluable.

With respect to verb conjugation, we also naturally normalized over the bigrams we looked for. That is, we looked for a total of 238 different forms of our set of four verbs, but normalized the different inflected versions to just the stem form for purposes of bigram storage. This is necessary for the task at hand as the inflectional variability would cause a futile co-occurrence analysis.

Finally, our initial list of bigrams contained instances of words with punctuation attached to them. We naturally removed these; however, there are other problems arising with respect to the Urdu script that are not dealt with as easily. For one, there are several different ways of spelling certain words in Urdu. One preprocessing step that we could have done is to run a normalization module across the corpus. However, this also requires specialized knowledge about the language/orthography and this source of errors was not large enough for us to take this step. Similarly, our bigram counts contain instances of words which have not been spaced correctly. The Urdu script is such that each letter has joined and nonJoined versions. The conditions governing when to use a joined vs. non-joined version of a letter are fairly complex.

For example, take *پاک پاتا* *bat ‘talk’* from Table 7.2. The first “letter” is a combination of a ‘b’ and an ‘a’ (the joined forms), the second letter is the non-joined version of ‘t’. In order to differentiate between spaces within words and spaces between two words, two different types of spaces have been defined. One is a normal space, the other is zero-width non-joiner.20 However, authors are not always consistent in their use, thus giving rise to errors in the corpus, which we again cannot deal with without adding time-consuming manual inspection coupled with deep language-particular knowledge to the process. These errors thus remain in the bigram list we use for the analysis detailed in the next section.

---

19 One could argue that we could have kept those instances of *ki* where it occurred sentence-finally, since in those cases it is always a verb. While there is punctuation at sentence boundaries in our corpus (e.g., in Figure 7.4, is used as a final stop), it is not clear that keeping such instances is a reliable method, given processes such as genitive scrambling (§3.6).

20 The HTML code for the zero-width non-joiner character: \&zwj;
Visual Analytics is based on the tight coupling of algorithms for automatic data analysis and interactive visual components (Keim et al., 2010, Thomas and Cook, 2005). The idea is to exploit human perceptive abilities to support pattern detection (Card et al., 1999). This involves the mapping of data dimensions to eight visual variables (Bertin, 1983), namely position (two variables x and y), size, value, texture, color, orientation and shape. While some numerical data dimensions can be mapped directly to one visual variable, other data features may require complex layout algorithms that project a combination of multiple data dimensions to a combination of visual variables, e.g., to the combination of the two powerful positional variables x and y. Finally, a data analyst should be able to manipulate the visual display interactively for different perspectives on the data, following Shneiderman’s Visual Information-Seeking mantra “Overview first, zoom and filter, then details on demand” (Shneiderman, 1996).

The purposes of visualizing data are manifold. On the one hand, visualizations can be used to achieve an overview of complex datasets. On the other hand, the visualization approach can serve as a starting point for interactively exploring data, ideally detecting hidden patterns. In addition, new hypotheses can be generated and existing hypotheses verified.

The visualization employed here uses the visual variable color and encodes the relative frequency of occurrences with different light verbs. This relative frequency is mapped onto a linear saturation scale, i.e. the higher the relative frequency, the more saturated the color. Figure 7.5 shows a “Reference” visualization on the top row, where the saturation is exemplified with the relative frequencies of 0.25, 0.5, 0.75 and 1.0 in the columns from left to right. Below the reference visualization, the relative frequencies of the data in Table 7.2 are encoded visually. The left-most color column shows the relative frequencies with kar ‘do’, the next columns show the relative frequencies with the light verbs ho ‘be’, hu- ‘become’ and rakʰ ‘put’, respectively. The lexical item on the left hand side is the transliterated version of the Urdu input in Table 7.2, using the transliterator of Bögel (2012), following the transliteration scheme by Malik et al. (2010).

By using color-coding to represent the relative frequencies, the visualization eases the interpretation of the underlying data; e.g., it is straightforward to tell from the picture that the item bAt ‘talk’ occurs much more often with kar ‘do’, since the first square in the bAt row is the most saturated one; however, it also occurs with ho, given that the second square also features (lighter) saturation. It is also easy to see how additional hypothesis can be formed based on the visualization; for example, one can hypothesize

---

21Since short vowels are generally not encoded in the Urdu script, the transliterator puts a default “*” in places where a short vowel is expected. The ambiguous characters vao and ye (consonant or vowel) are represented with <vao> and <ye>, respectively. In case of entries with ‘??’, the automatic transliteration could not find an adequate transliteration — this can be due to typing errors or unusual vowel/consonant combinations of English loan words. We needed to use a transliterated version of our bigram lists as the visualization tool we used could not deal with UTF-8 input.
7.4. CORPUS STUDY 1: THE “BRUTE FORCE” APPROACH

Figure 7.5: Visualization of the relative frequencies in Table 7.2

that bAt ‘talk’ and h2As3i1 ‘achievement’ pattern more alike than any of the other items seen here. Such observations are harder to make based on raw frequency counts as in Table 7.2.

The visualization tool allows for an interactive exploration of the data in that the human investigator can scroll through the list of visually encoded N-V combinations, with the possibility of zooming in to get a detailed view on a restricted set of nouns, as well as zooming out to see a greater number of combinations and their overall behavior. Mousing over a colored box reveals the relative frequency of the N-V CP.

7.4.2.1 Visual Analysis: Round 1

After applying the visualization to the k-means clustering as discussed above and manually inspecting the visualization, it was found that best result for the clustering method described in section 7.4.1.3 was obtained with a specification of $k = 5$, i.e., five clusters. The visual analysis is described in this section.

The five clusters are numbered 0–4. Cluster 4 is a large cluster of about 1100 words. This cluster contained just those sets of words which occurred only or almost only with ho ‘be’. A manual inspection of this cluster showed that all of these words were either errors, false positives containing inflected verbs or words like ‘what’ and ‘how many’ or items like ‘small’, ‘teacher’, ‘market’, ‘tomorrow’ (the vast majority). The last category are items that occur in run-of-the-mill copula contexts like ‘He is a teacher.’ This class is of no interest to us as there are no useful combinatory constraints to be discovered. We therefore decided to discard this cluster from our list in its entirety.

Cluster 3 contained a small set of words that occurred with hu- ‘become’ 100% of the time. All of these were false positives and were discarded. Similarly, Cluster 1 was a comparatively small cluster that consisted of words occurring mainly with rak’h ‘put’. Manual inspection showed that all the words occurring with rak’h ‘put’ 100% of the time were false positives: the words were all objects of the main verb ‘put’ and not CPs of any kind. Again, we culled our bigram list to remove this set.

The other two clusters are more of a mixed bag. Cluster 2 has many items that occur only with kar ‘do’. Figure 7.6 shows the top part of Cluster 2 on the left. This contains nouns already identified as belonging to the ones combining in the type of N-V CP we are interested in. However, this class contains desired results as well as false positives which cannot be separated from one another on a visual (or purely numerical) basis. Pruning the cluster would involve intense manual labor, which we decided to forego.
Similarly, Cluster 0 contains desired results as well as false positives. Again, these are difficult to discern visually, though there are subclusters as this cluster contains those words which occur with all the four light verbs (at different levels of frequency). In Figure 7.6, hasɪl ‘achievement’, šuru ‘beginning’, and koʃɪʃ ‘struggle’ are the kinds of nouns we are interested in. On the other hand, xɑtɑm ‘finished’ and jari ‘continued’ are adjectives.

As part of our quick visual inspection, we also noted several types of unusual patterns found in the clusters, such as the ones illustrated in Figure 7.7 for the verb ʊṭʰa ‘raise’. These often turned out to be false positives. The visualization thus allowed for a quick and easy identification of such false positives, and all unusual patterns that were identified as false positives based on this method of analysis were also removed from our bigram list. Detecting these types of errors would have been neigh impossible without an easily accessible visual representation.

All in all, after the manual selection process, we were left with only 1090 instances of bigrams. This list still contains false positives, but removing these would take intensive manual labor. We therefore decided to rerun our automatic clustering algorithm on the (partially) cleaned data.

### 7.4.2.2 Visual Analysis: Round 2

The new clustering again yielded the best results with a specification of $k = 5$, i.e., five clusters. The visualization shows that the cleaning and culling described in the previous section has had a positive effect, but that even the remaining 1090 bigrams still contain a good amount of false positives which would need to be culled on the basis of intense manual inspection. Nevertheless, the overall results are encouraging as different types of nouns are indeed being distributed over the different clusters. The description of the visual analysis follows.

As shown in Figure 7.8, Cluster 0 still contains hasɪl ‘achievement’ and koʃɪʃ ‘struggle’.
Items from the previous Cluster 2 also now appear here: *bat* ‘talk’, *istemal* ‘use’ and *peʃ* ‘presentation/happening’. On the other hand, *fur[u* ‘beginning’, *xatam* ‘finished’ and *jari* ‘continued’ are now in Cluster 3, which mainly consists of adjectives, but also a few of the nouns we are interested in. While ADJ-V CPs were not the target of this investigation, this cluster contains potentially useful information with respect to ADJ-V CPs.

Cluster 4 contains words which occur almost exclusively with *rakʰ* ‘put’; by looking at these nouns, it turns out that this set still does not contain much that is of interest within the scope of this investigation; i.e., most items in this cluster do not partake in N-V CPs. On the other hand, Cluster 2 contains many nouns which occur in N-V CPs and indeed, contains many of the nouns that Ahmed and Butt (2011) identified as belonging to Class B. A comparison of our results with that of Ahmed and Butt (2011) shows that while we did not find all of the nouns used in Ahmed and Butt, of the nouns that we did find most are Class B nouns (no dative subjects allowed). These are distributed over Clusters 0 and 2, while the Class A nouns (full range of light verb use) are found in Clusters 0, 1 and 3. No Class C nouns were found in this particular study; possible reasons for this outcome are explored below. Table 7.3 summarizes the results of the study.

<table>
<thead>
<tr>
<th>Cluster</th>
<th><em>ho</em> ‘be’</th>
<th><em>kar</em> ‘do’</th>
<th><em>hu</em> ‘become’</th>
<th><em>rakʰ</em> ‘put’</th>
<th>Ahmed and Butt (2011)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>✓/✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Class A/B</td>
</tr>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Class A</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Class B</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Class A</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 7.3: Overview of the results — Corpus Study 1

7.4.3 Results and Discussion

In this study, we set out to see if we could find fine-grained information about the distributional properties of nouns in N-V CPs by extracting bigrams from a large “raw” corpus

---

22But note that Cluster 0 also contains adjectives, for instance *gṛifter* ‘arrested’ in Figure 7.8. Adjectives that are not part of CPs had already been removed in the manual filtering step in §7.4.2.1.
of Urdu. In addition to finding target instances of N-V combinations, our work also resulted in lists of possible ADJ-V combinations.

Unfortunately, our method has suffered from problems related to data sparsity. Manual inspection of some of the CPs we found showed that not all possible combinations of N/ADJ were in fact attested in the corpus. That is, our bigram lists will need to be complemented by manual native speaker work (traditional lexicography). Further, it is hard to give an indication of how many valid CPs we might have missed. Since we have no idea how many CPs were indeed contained in our corpus, we cannot say with certainty. And as N-V CPs are combinatorily dynamic, there are potentially infinitely many of these. All we can say is how many of the potentially infinitely many combinations we did find.

We also find it remarkable that out of a corpus of 7.9 million words, we are at this point left with only 1090 bigrams (and are aware that a portion of these bigrams would still need to be culled as they represent errors or false positives). This problem of data sparsity could perhaps be ameliorated by using a different type of corpus, but the suspicion is that for our type of enterprise, data sparsity might be a problem regardless of how large or different a corpus is chosen — one cannot guarantee that all possible combinations will indeed be attested in any given corpus.

The study also makes contributions with respect to two methodological points. In order to be analyze our data more perspicuously, we experimented with new methods from visual analytics. This experimentation was successful as it allows for quick visual analysis of our datasets, which in turn also enables a non-labor-intensive way of further cleaning and culling our data. In particular, the visual analysis allowed us to be able to quickly assimilate information about complex interrelationships in our data: which types of verbs does the word in question occur with with what level of frequency and how does this compare to other words in the list?

The visualization component proved to be an essential part of our data analysis. Another essential component turned out to be our language-particular knowledge and access to language-particular resources. The idea that we could work more or less “blindly” with a large corpus did not pan out. Rather, at every step we needed to be aware of language particular issues with respect to orthography, morphology and syntax. We could not have done our work without the use of a list of all the conjugated forms of 700 simple verbs (Raza, 2011) or a transliterator (Bögel, 2012) (needed to massage the input for the visualization component). We thus conclude that while we were partly successful in achieving what we set out to do, our work would have been able to be yield more precise results if we had access to standardized language-particular resources. In §7.5, I discuss a second corpus study that makes use of one such resource, the Hindi/Urdu Treebank (HUTB, Bhatt et al., 2009, Palmer et al., 2009).
7.4.4 Remaining Problems

Some problems remain after conducting this first corpus story. First, note that the corpus used was not POS tagged, which results in the need to include multiple pruning steps. Even after these steps, without additional manual filtering and pruning, the resulting lists are far from being usable as input for a computational lexicon, e.g., in the Urdu ParGram grammar. This may also in part be the reason why the clusters of the study are not entirely congruent with the classes identified by Ahmed and Butt (2011); e.g., no nouns from Class C were found. Another reason may be data sparsity; recall that the lists of Ahmed and Butt (2011) were supplemented by native speaker judgments.

Another issue with this particular study is that there was no objective, empirical evaluation. The evaluation was done manually, by examining the output of the visualization and deleting items that were identified as non-cp nouns. Applying a POS tagging step could have remedied part of this issue, since there would have been less junk items in the input to the clustering step. Note also that manually inspecting the visualization and deleting false positives is a tedious task and is error-prone.

Another drawback is that a short list of only 4 light verbs is used. As seen in §7.2.2, there are more than 4 light verbs that may be used in N-V CPs, and extending the list of light verbs can have two advantages. First, it is highly likely that the k-means clustering algorithm will turn out to be more effective if more features (i.e., more light verb relative frequencies) are used.

7.5 Corpus Study 2: The “Seed List” Approach

In the second corpus study, again the aim was to identify patterns of Hindi/Urdu N-V CPs. In contrast to the study discussed in §7.4, the methodology described here combines information extracted from a treebank with a clustering approach and visualization methods. We also use a larger corpus to overcome at least some of the issues with respect to data sparsity.23

Here, we aimed at overcoming several problems that came up during the first study. First, we were able to make use of new, manually-created resources. For example, there is a freely available part-of-speech (POS) tagger for Hindi (Reddy and Sharoff, 2011), as well as a manually-created treebank covering both Hindi and (on a smaller scale) Urdu: the Hindi/Urdu Treebank (HUTB, Bhatt et al., 2009, Palmer et al., 2009). Thus, we were trying to come up with ways to use these emerging resources to help us in the task at hand, namely identifying templates of N-V CP formation. Second, by including more light verbs in our examination than included in the first study, we increased the search space to be able to come up with a bigger set of generalizations about N-V CP formation. The intuition here was that by using manually-created resources, one should end up with

---

23The research described in this section was jointly carried out with my colleague Ashwini Vaidya. Throughout the section, I indicate my own contributions to the work. The study was previously published in Sulger and Vaidya (2014).
generalizations about the data that are in principle implementable in a lexicon as part of the Urdu ParGram grammar.

7.5.1 Methodology

In this section, I describe the methodology used in the second corpus study for harvesting N-V CP candidates and for identifying classes among the candidate CPS using clustering techniques. For reasons to become clear below, techniques from visual analytics played only a minor role in this particular study. The approach described here makes use of several, in part manually-annotated linguistic resources for Hindi/Urdu, and the reported results overcome some of the problems described in §7.4.4 (the ones mentioned in §7.4.4, in particular). The study thus again makes the point mentioned in §7.4.1, namely that once high-quality linguistic resources can be drawn upon, the outcome of such clustering methods as well as the resulting identification of semantic classes of CPS can be improved significantly.

7.5.1.1 Corpus

We were not aware of any freely available, reasonably-sized Hindi corpus, which is why we collected our own for this study. As done for Urdu in the approach described in §7.4, the BBC news site was scraped for Hindi news text. This gave us a corpus of about ~6.5 million running tokens of Hindi. We then downloaded the December 1st, 2013 dump of the Hindi wikipedia and parsed it into raw text using the Wikipedia Extractor toolset. This gave us an additional ~10.8 million running tokens of Hindi; by including the Wikipedia, the resulting corpus was extended beyond the newspaper domain. Finally, we also included the raw version of the Hindi portion of the HUTB (~211,000 running tokens). The final corpus contained ~17.5 million tokens. Below is an overview table of the resulting Hindi corpus.

<table>
<thead>
<tr>
<th>Source</th>
<th>Domain</th>
<th>Size (tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi Wikipedia dump</td>
<td>Wikipedia articles</td>
<td>10,808,459</td>
</tr>
<tr>
<td>BBC Hindi</td>
<td>news articles</td>
<td>6,529,304</td>
</tr>
<tr>
<td>HUTB</td>
<td>news articles</td>
<td>211,472</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>17,549,235</td>
</tr>
</tbody>
</table>

Table 7.4: Size distribution in Hindi corpus used

---

24 Recall from §2.7 that the Urdu ParGram grammar has been designed from the outset to parse both Hindi as well as Urdu text. The corpus study discussed here was carried out on a Hindi corpus, which the grammar can directly profit from. See also §7.5.3 for discussion.
25 http://www.bbc.co.uk/hindi
26 http://dumps.wikimedia.org/hiwiki/20131201
27 http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
Next, the corpus was tagged using the state-of-the-art Hindi tagger described in Reddy and Sharoff (2011). The tagger was trained on a corpus of ~30.4 million tokens and features a reported accuracy for Hindi text of 91.31%. The output of the tagger is formatted as in Table 7.5, where each word appears on a separate line together with its POS analysis as well as the lemma.

<table>
<thead>
<tr>
<th>Corpus token</th>
<th>Transliteration/gloss</th>
<th>POS tags</th>
<th>Lemmatized token</th>
</tr>
</thead>
<tbody>
<tr>
<td>कुछ</td>
<td>kʊch ‘some’</td>
<td>QF.adj.any.any..d</td>
<td>कुछ</td>
</tr>
<tr>
<td>अच्छी</td>
<td>acchɪ ‘good’</td>
<td>JJ.adj.f.any..any</td>
<td>अच्छी</td>
</tr>
<tr>
<td>फिल्मों</td>
<td>filmo ‘films’</td>
<td>NN.n.f.pl.3.o</td>
<td>फिल्मों</td>
</tr>
<tr>
<td>का</td>
<td>ka ‘of’</td>
<td>PSP.psp.m.sg..d</td>
<td>का</td>
</tr>
<tr>
<td>निर्माण</td>
<td>nirman ‘organization’</td>
<td>NN.n.m.sg.3.d</td>
<td>निर्माण</td>
</tr>
</tbody>
</table>

Table 7.5: POS tagged Hindi corpus excerpt

7.5.1.2 SEED NOUNS

Next, we built a seed list of nouns. In the HUTB, the POF label is the dependency label by which CPS can consistently be identified. The diagnostic criteria used for identifying CPS in the treebank is based on native speaker intuition; however, the final decision about the CP status of a given noun-verb sequence is up to the senior linguists heading the HUTB project (Bharati et al., 2012). We were thus confident that the presence of a POF label would be a reliable indication of a CP, and initial random samples reassured our confidence.

The POF label is used in the HUTB for adjectives and adverbs as well as nouns. We were, however, only interested in nouns carrying the POF label. Thus, we filtered based on part of speech; in the HUTB, nouns carry the POS label NNP. We were aware of the fact that filtering the items dependency-labeled as POF and POS-labeled as NNP excluded any nouns that were wrongly tagged as something other than NNP in the treebank, and introduced adjectives, adverbs etc. to our noun seed list if these were incorrectly tagged as NNP in the treebank. Since the treebank is hand-crafted, however, we considered this portion to be of small size.

This gave us an initial list of candidate nouns that were further filtered for spelling variations using Perl scripts. We were left with a seed list with 1207 nouns which we knew take part in CPS. In the following table, an excerpt of the noun seed list is shown; I also provide frequencies of occurrences in the HUTB. To summarize, the list is extracted

---

Note that not all tagger errors will have an impact on our methodology, since we are only interested in bigrams consisting of items tagged as nouns and items tagged as verbs; e.g., if the tagger mistakenly identified an adjective as an adverb, this will have no impact given our approach.

POF stands for “part of” (Bharati et al., 2012, Bhat and Sharma, 2012).

The HUTB labels proper nouns distinctly as NNP; since proper nouns never form N-V CPSs, nouns occurring with the tag NNP were excluded.
automatically from the Hindi portion of the Hindi/Urdu Treebank (HUTB, Bhatt et al., 2009, Palmer et al., 2009) and consists of all the items that carry the dependency label POF (which stands for “part of”) and are POS tagged as nouns, indicating that the item at issue takes part in an N-V CP.

<table>
<thead>
<tr>
<th>ID</th>
<th>noun</th>
<th>Transliteration/gloss</th>
<th>HUTB frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>प्रतिनिधित्व</td>
<td>pratnuditva ‘representation’</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>असहमति</td>
<td>asahamati ‘discrepancy’</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>अदा</td>
<td>ada ‘grace’</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>प्रस्ताव</td>
<td>prastav ‘motion, proposal’</td>
<td>56</td>
</tr>
</tbody>
</table>

Table 7.6: Hindi seed noun list excerpt

7.5.1.3 Bigram Collocation Extraction

Our aim was to establish co-occurrence patterns of the nouns in the seed list. At the same time, we wanted to only extract co-occurrence bigrams that were well-attested. Obviously, we could have turned exclusively to the actual source of our seed list nouns, the HUTB itself, to extract co-occurrence bigrams and run clustering algorithms based on this. For this task, however, the HUTB corpus (~211,000 words) by itself would not have been sufficient. For instance, if we applied a cutoff of 50 to the bigram occurrences in the treebank, we would be left with only 20 nouns, which would hardly have given us any meaningful groups in the clustering step. In addition, the method would have suffered severely from data sparsity due to the small corpus size.

Thus, we turned to our ~17.5 million token Hindi corpus, where we extracted our bigrams for further analysis. We were interested in bigrams adhering to the following pattern: seed list noun item + light verb. The list of light verbs we examined is given in (31). This list was determined after consulting the HUTB; these are the seven most frequent light verbs in the treebank, as indicated by counting their occurrences to the right of dependencies marked by the POF label (which differentiates the light verbs from their full verb counterparts). We decided to include a larger range of seven light verbs than the study reported in §7.4 because we wanted to be able to make more detailed assumptions about the different syntactic patterns of Hindi N-V CP formation. At the same time, however, our list of light verbs is smaller than the one given in Mohanan (1994), cf. the list in (8). This is due to the fact that the frequencies of the two light verbs missing from (31) (i.e., mar ‘hit’, kʰa ‘eat’) drop significantly in the HUTB; we were not sure we could make decent assumptions about those light verbs and the nominal hosts that co-occur with them.


We did not consider hu- ‘become’ as a separate light verb, for two reasons. First, the tagger we were using did not distinguish between different aspectual inflections of the
verb *ho* ‘be’; i.e., all inflections of *ho* were collapsed into a single lemma. Matching the surface word form with a specific word list consisting of inflected word forms of *hu-* would have been cumbersome. Second, it was not clear to us that we would gain much better results from including *hu-* Recall that Ahmed and Butt (2011) had identified the absence of *hu-* as the feature distinguishing Class C nouns from the other classes; Class C was characterized as a smaller class compared to the other ones. The previous study presented above, however, was not able to identify any Class C nouns, which suggests that the distinction is hard to identify for a clustering algorithm.

Each seed list noun item had to be tagged as **NN** in the corpus. That is, a match would only occur if one of the light verb stems occurred directly to the right of the noun. Thus, this approach, like the one described in §7.4, adheres to the definition of the **N-V CP** pattern as given in §7.2, which states that the nominal host and the light verb are part of a fairly fixed phrasal configuration (i.e., they are both part of a single maximal verbal projection) and thus represent bigrams.

Our initial lookup strategy was to look only at nouns that occurred more than 50 times in the corpus (i.e., 50 noun occurrences summed up across all light verbs), to make assumptions only about well attested nouns. This resulted in a total of 523 nouns extracted from the corpus, together with their light verb co-occurrences. We converted the light verb co-occurrences to relative frequencies; the resulting list formed the basis for our clustering experiments. An excerpt from the resulting list is given in Table 7.7. Here, the seed list noun items of Table 7.6 are given together with the absolute frequencies of occurrence with each of the light verbs. The same list, but with the converted relative frequencies of light verbs, is shown in Table 7.8.

To be able to make comparisons regarding data quality and cluster coherence, we chose to create a separate co-occurrence dataset, this time extracting all nouns that occurred more than 3 times (summarized over all light verbs) in the corpus. This returned 987 nouns (of the 1207 nouns that were originally part of the noun seed list). Here, we were curious whether the clusters would distribute differently if more data points (i.e., nouns) were taken into account.

### Automatic Clustering

In the next step, a clustering algorithm was applied to the data. This was done using the clustering tool described in Lamprecht et al. (2013). At the moment, the tool features two clustering algorithms: the *k*-means algorithm (MacQueen, 1967), already used in the corpus study described in §7.4, as well as the Greedy Variance Minimization (GVM) algorithm. Each data item *n* (each noun) was assigned to a cluster based on 7 features corresponding to the relative frequencies of light verb co-occurrences (see Table 7.8). For example, the nouns *prastav* ‘proposal’ and *vyay* ‘cost’ have feature values as shown in Table 7.9. Using the clustering tool, we created 18 different datasets: a set of nine
### Table 7.7: Hindi bigram matches: absolute frequencies

<table>
<thead>
<tr>
<th>ID</th>
<th>Noun</th>
<th>Transliteration/gloss</th>
<th><em>ho</em> 'be'</th>
<th><em>kar</em> 'do'</th>
<th><em>de</em> 'give'</th>
<th><em>le</em> 'take'</th>
<th><em>rakʰ</em> 'put'</th>
<th><em>lag</em> 'be attached'</th>
<th><em>a</em> 'come'</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>pratīna</em> 'representation'</td>
<td><em>pratīna</em> 'representation'</td>
<td>11</td>
<td>704</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>732</td>
</tr>
<tr>
<td>2</td>
<td><em>ada</em> 'grace'</td>
<td><em>ada</em> 'grace'</td>
<td>1</td>
<td>380</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>381</td>
</tr>
<tr>
<td>3</td>
<td><em>prastav</em> 'motion, proposal'</td>
<td><em>prastav</em> 'motion, proposal'</td>
<td>72</td>
<td>92</td>
<td>162</td>
<td>10</td>
<td>58</td>
<td>0</td>
<td>29</td>
<td>423</td>
</tr>
<tr>
<td>4</td>
<td><em>upeka</em> 'neglect, disregard'</td>
<td><em>upeka</em> 'neglect, disregard'</td>
<td>11</td>
<td>81</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td><em>vyay</em> 'cost'</td>
<td><em>vyay</em> 'cost'</td>
<td>28</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>67</td>
</tr>
</tbody>
</table>

### Table 7.8: Hindi bigram matches: relative frequencies

<table>
<thead>
<tr>
<th>ID</th>
<th>Noun</th>
<th>Transliteration/gloss</th>
<th><em>ho</em> 'be'</th>
<th><em>kar</em> 'do'</th>
<th><em>de</em> 'give'</th>
<th><em>le</em> 'take'</th>
<th><em>rakʰ</em> 'put'</th>
<th><em>lag</em> 'be attached'</th>
<th><em>a</em> 'come'</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>pratīna</em> 'representation'</td>
<td><em>pratīna</em> 'representation'</td>
<td>0.015</td>
<td>0.962</td>
<td>0.022</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td><em>ada</em> 'grace'</td>
<td><em>ada</em> 'grace'</td>
<td>0.003</td>
<td>0.997</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td><em>prastav</em> 'motion, proposal'</td>
<td><em>prastav</em> 'motion, proposal'</td>
<td>0.17</td>
<td>0.217</td>
<td>0.383</td>
<td>0.024</td>
<td>0.137</td>
<td>0</td>
<td>0.069</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td><em>upeka</em> 'neglect, disregard'</td>
<td><em>upeka</em> 'neglect, disregard'</td>
<td>0.12</td>
<td>0.88</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td><em>vyay</em> 'cost'</td>
<td><em>vyay</em> 'cost'</td>
<td>0.42</td>
<td>0.58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7.7: Hindi bigram matches: absolute frequencies

Table 7.8: Hindi bigram matches: relative frequencies
datasets for \( k = 2 \ldots 10 \), for both of the two cutoff points examined (\( c = 50 \), \( c = 3 \)). These 18 datasets were input to our evaluation method.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Transliteration/gloss</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
</tr>
</thead>
<tbody>
<tr>
<td>प्रस्ताव</td>
<td>prastav 'motion, proposal'</td>
<td>0.17</td>
<td>0.217</td>
<td>0.383</td>
<td>0.024</td>
<td>0.137</td>
<td>0</td>
<td>0.069</td>
</tr>
<tr>
<td>व्यय</td>
<td>vyay 'cost'</td>
<td>0.42</td>
<td>0.58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.9: Feature value examples for clustering

7.5.1.5 Cluster Evaluation

We made use of an automatic method using Hindi WordNet (Bhattacharyya, 2010) to choose the best \( k \) partition value. We followed the technique described in Van de Cruys (2006), which uses WordNet relations to arrive at the most semantically coherent clusters. We define semantic coherence as the similarity among items in a cluster, based on an overlap between their WordNet relations. Specifically, for each \( k = 2 \ldots 10 \), we iterated through the automatically generated clusters and performed the following steps:

1. Using WordNet, we extracted synonyms, hypernyms and hyponyms for every word in a cluster.

2. A word that had the most semantic relations with every word in the cluster was chosen as its centroid (see Table 7.11 below for the resulting “semantic centroid” words).

3. The co-hyponyms i.e., the hyponyms of the hypernyms for this centroid were extracted from WordNet (along with its synonyms, hypernyms and hyponyms).

4. In order to calculate precision for each cluster, we counted the number of words in that cluster that overlapped with the words in the centroid’s relations (the words extracted in step 3).

Precision is evaluated as an average across all clusters computed. Thus, in order to achieve perfect precision (1.0 or 100%), a model would need to group the data into clusters that are made up exclusively of nouns that display perfect semantic coherence with the centroid, so that all of the nouns would need to match one of the centroid’s relations. A precision value of 0.5 (or 50%) would entail that an average 50% of nouns across all clusters in a given model matches one of the centroid’s relations.

We averaged the precision across all clusters for every \( k \) value. We found that precision for each cluster gradually improved for \( k = 2 \ldots 4 \) until we got the most semantically coherent partitions for \( k = 5 \) using \( k \)-means, for \( n = 523 \) nouns and \( c = 50 \). Table 7.10 shows the evaluation figures from our WordNet evaluation method, for \( k = 5 \ldots 9 \).

In Figure 7.9, we plot the semantic coherence values against the number of clusters to show the best results. The \( k \)-means algorithm performed only slightly better than GVM,
7.5. CORPUS STUDY 2: THE “SEED LIST” APPROACH

<table>
<thead>
<tr>
<th>Size of $k$</th>
<th>$c = 3$</th>
<th>$c = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 5$</td>
<td>0.049</td>
<td>0.107</td>
</tr>
<tr>
<td>$k = 6$</td>
<td>0.066</td>
<td>0.121</td>
</tr>
<tr>
<td>$k = 7$</td>
<td>0.089</td>
<td>0.104</td>
</tr>
<tr>
<td>$k = 8$</td>
<td>0.084</td>
<td>0.108</td>
</tr>
<tr>
<td>$k = 9$</td>
<td>0.082</td>
<td>0.095</td>
</tr>
</tbody>
</table>

Table 7.10: Semantic coherence values for $k = 5\ldots 9$ for clustering algorithms GVM and k-means

and after $k = 5$, the semantic coherence of the clusters declined again. As a point of comparison, we also plotted k-means and GVM results for $c = 3$. In this configuration, the best results are achieved for a higher $k$ value (i.e., for $k = 7$ using k-means and $k = 8$ for GVM), but we rejected this on the basis of a better semantic coherence value for k-means with a cutoff of 50.

7.5.2 ANALYSIS

The tool developed by Lamprecht et al. (2013) is useful for visual cluster inspection; for instance, the tool created the visual clustering in Figure 7.10 using the k-means algorithm with $k = 5$. The visualization enables the user to inspect the data points and derive initial generalizations. Figure 7.10 shows a visualization of the clusters from the tool; here, each circle represents a data item. The colors encode membership within a cluster. The larger circles represent cluster centroids. Cluster centroids are not actual, but fictive
data items that possess feature values averaged across all items for a given cluster; in other words, the centroids indicate the typical noun/light verb distribution of a cluster (Lamprecht et al., 2013).

Recall that in the visualization used in §7.4, relative frequencies of noun items to co-occur with light verbs were correlated with color saturation; the more saturated the color to the right of the noun, the higher the relative frequency of the light verb to occur with it; the cluster number as well as the actual noun were shown to the left (Figures 7.5–7.8). The tool did not supply information on the coherence of the clusters or the cluster centroids, nor does the visualization show prototypical cluster patterns at-a-glance (Lamprecht et al., 2013).

The visualization used here, on the other hand, enables us to see three most frequently occurring light verbs, viz. kar ‘do’, ho ‘be’ and de ‘give’, represented by light green, dark blue and pink respectively. Many nouns alternate with kar ‘do’ and ho ‘be’, hence there is a visible continuum between the light green and dark blue data points. The two clusters in the centre show a dark green cluster, consisting of only a handful of nouns that alternate with the light verbs rakʰ ‘keep’, lag ‘be attached’ and a ‘come’. The light blue cluster on the other hand is larger and is dominated by the light verb le ‘take’.

In order to further interpret the results of our study, we also referred to a secondary result from our WordNet evaluation. Recall that while extracting the extent of overlap of the semantic relations, we also extracted the “semantic centroids” (sc), i.e., those words that had the most semantic relations with every other word in the cluster (see §7.5.1.4). For our best result of \( k = 5 \), we show the scs in Table 7.11.

The scs also revealed semantic similarities in the five clusters that we found. For in-
stance, dynamic events that are inanimate and abstract and take an agentive argument will lend themselves to combinations with kar ‘do’; these nouns end up in Cluster 2 and include nouns of Ahmed and Butt’s Class B. Similarly, events that include the semantic property of ‘transfer’ occur with de ‘give’ and are part of Cluster 1, although there is ostensibly an overlap here, as these events invariably also require agentive arguments. This particular class of nouns has previously gone unnoticed in previous studies, but the figures indicate a clear difference in behavior between nouns of Cluster 1 and nouns of Cluster 2. More nouns that end up in Cluster 1 are vacan ‘promise’ or manyata ‘recognition’.

Cluster 4 includes nouns that occur often with the light verb ho ‘be’. These are nouns that denote mental states, resulting in an experiencer subject. The nouns in this group also occur frequently with kar ‘do’. These are the psych-predications previously identified as Class A in Ahmed and Butt’s classification.

Less frequently occurring light verbs, especially rak’h ‘put’, lag ‘be attached’ and a ‘come’ show fewer alternations, as they do not occur in combination with all nouns and are grouped together in this result (Clusters 3 and 5). These light verbs often form N-V CPS with a more idiosyncratic meaning. Davison (2005) has argued that some of these light verbs may form “incorporation idioms” (rather than true N-V CPS). An example for a highly idiosyncratic combination is ag lag ‘burn’, featuring below in (32). The noun ag ‘fire’ occurs with the light verb lag ‘be attached’ 96% of the time in our dataset. A further indication for the idiomatic status of the nouns in Clusters 3 and 5 is the fact that these two clusters are the smallest ones in our dataset. Here, we do not go so far as to exclude these nouns from our investigation as well as from the implementation (§7.6); it is not clear what exactly the theoretical status of “incorporation idioms” is, nor how they can be implemented in LFG/XLE, which is why they are treated as N-V CPS here.

(32) gʰɑr=ko ag lag-i
    house.M.SG = DAT fire.F.SG.NOM be-attached-PERF.F.SG

    ‘The house was burning.’ (lit. ‘Fire was attached to the house.’)

---

32 A look at a random sample of the remaining 4% (i.e., those instances of ag ‘fire’ occurring with light verbs other than lag) has revealed that these combinations are either not valid CPS, or not part of complete sentences.
Table 7.12 shows the number of noun items in each cluster for $k = 5$; it also lists the average relative frequencies for each cluster and each light verb, averaged across noun items. Comparing the average figures in Table 7.12 with the feature values of the scs in Table 7.11 further confirms the status of the scs as prototypical members of each cluster.

<table>
<thead>
<tr>
<th>Cl.</th>
<th>n</th>
<th>ho ‘be’</th>
<th>kar ‘do’</th>
<th>de ‘give’</th>
<th>le ‘take’</th>
<th>rak ‘put’</th>
<th>lag ‘be attached’</th>
<th>a ‘come’</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>93</td>
<td>0.06</td>
<td>0.08</td>
<td>0.76</td>
<td>0.06</td>
<td>0.017</td>
<td>0.002</td>
<td>0.021</td>
</tr>
<tr>
<td>2</td>
<td>289</td>
<td>0.108</td>
<td>0.849</td>
<td>0.02</td>
<td>0.006</td>
<td>0.009</td>
<td>0.001</td>
<td>0.008</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0.14</td>
<td>0.057</td>
<td>0.124</td>
<td>0.014</td>
<td>0.002</td>
<td>0.592</td>
<td>0.072</td>
</tr>
<tr>
<td>4</td>
<td>97</td>
<td>0.701</td>
<td>0.192</td>
<td>0.035</td>
<td>0.017</td>
<td>0.018</td>
<td>0.006</td>
<td>0.03</td>
</tr>
<tr>
<td>5</td>
<td>33</td>
<td>0.092</td>
<td>0.051</td>
<td>0.098</td>
<td>0.458</td>
<td>0.167</td>
<td>0.008</td>
<td>0.126</td>
</tr>
</tbody>
</table>

Table 7.12: Relative frequencies for all clusters for $k = 5$

The average figures of N-V CP co-occurrences for a certain cluster inform us about the likelihood of a certain group of nouns to co-occur with a certain light verb. For instance, nouns that are part of Cluster 1 show a high likelihood of occurrence with de ‘give’ (0.76 or 76%), but they are less likely to occur with other light verbs. The distribution reflects a difference in the syntactic behavior of the nouns: While the productive patterns indicate CP formation, the less productive patterns do not represent CPs at all; that is, the more frequent a given combination is, the more likely it will be that it is a CP. This is a finding in line with Butt et al. (2012), who ended up deleting many low-frequency patterns which turned out to be non-CP combinations. Similar tendencies can be derived for the five groups of nouns derived from our clustering experiment above.

7.5.3 RESULTS AND DISCUSSION

Above, I have discussed a second corpus study of Hindi/Urdu N-V CPS that makes use of a novel methodology in terms of a noun seed list and an evaluation based on WordNet. It was found that the k-means algorithm with $k = 5$ and a frequency cutoff of 50 provided the best results in terms of semantic coherence of the resulting clusters. The resulting noun groups can be used in different NLP settings.

The study reported in this section was carried out on Hindi data. Since much of the differences in vocabulary between Hindi and Urdu arises in the nominal domain, one would have to run the same experiment on Urdu data separately to collect a lexical resource of Urdu nouns taking part in N-V CP formation. However, there is an obstacle that needs to be overcome to apply the above methodology to Urdu. The Urdu portion of the HUTB is still a lot smaller than the Hindi portion, which would result in coverage loss (Bhat and Sharma, 2012). Consequently, the noun seed list for Urdu would be smaller than the Hindi seed list collected here. It is to be expected, though, that many of the Hindi seed list items also feature as nominal hosts in Urdu, and thus translations of the Hindi nouns could provide an initial seed list for Urdu; this is certainly a possible direction for future work.
7.6 NOUN GROUPS IN HINDI/URDU GRAMMAR DEVELOPMENT

The experiments discussed here show that nouns can be grouped according to their distribution across light verbs; often, they occur with one dominant light verb, but nouns also feature the possibility of occurring with one or several other light verbs. The clusters do not represent absolute certainties about N-V CPS, but report tendencies of occurrences. In this section, I discuss the integration of the Hindi noun groupings into the Urdu ParGram grammar via the construction of templates (see §2.5.4) that can be augmented to model the relevant linguistic generalizations in terms of constraints inspired by optimality theory (OT, Prince and Smolensky (2004); see §2.5.3.10). A serious evaluation of the effect on the grammar of adding in this lexical resource is part of future work.

7.6.1 IMPLEMENTING N-V CPS

Before I go into further detail concerning the implementation of noun groups, I will briefly explain the basic implementation of N-V CPS in the Urdu ParGram grammar. For illustration purposes, I provide the example in (33) as a running example for the description. Here, it is found that the nominal host ifara ‘signal’ is combined with the light verb kar ‘do’. Further, ifara ‘signal’ provides an additional argument bilal to the predication, which is accusative case-marked. In the passivized version of (33a), which is shown in (33b), it is the host’s argument bilal which is the subject; it is not possible to passivize the nominal host, which is an indication that the CP must be of the non-agreeing type discussed by Mohanan (1994) (§7.2.3). The direct object of (33a) is bilal.

(33)  

\[ \begin{align*}
   &a. \quad nadya=ne \quad bilal=ko \quad ifara \quad ki-ya \\
   &\quad Nadya.F.SG = \text{erg} \quad Bilal.M.SG = \text{acc} \quad signal.M.SG \quad do\text{-perf}.M.SG \\
   &\quad ‘Nadya signaled Bilal.’ (lit. ‘Nadya did a signal towards Bilal.’) \\
   &b. \quad nadya=se \quad bilal \quad ifara \quad ki-ya \quad go-ya \\
   &\quad Nadya.F.SG = \text{inst} \quad Bilal.M.SG.NOM \quad signal.M.SG \quad do\text{-perf}.M.SG \quad go\text{-perf}.M.SG \\
   &\quad ‘Bilal was signaled by Nadya.’ (lit. ‘Bilal was done a signal by Nadya.’)
\end{align*} \]

The clausal rule for complex predicates is shown below in (34b), defined as a meta-category and called from the main S rule in (34a). In (34b), the additional disjuncts for OBL-TH as well as OBL-SRC are needed for the clausal attachment of locative and instrumental arguments provided by the nominal host, to be described further below. The existential constraint on the VTYPE feature in the VCmain entry ensures that only complex predicates are handled by this rule. The verbal complex VCmain includes an entry Vcp that is shown in (34b). Vcp supplies a value to (\text{\`VTYPE COMPLEX-PRED}) via a template CP-TYPE, based on the syntactic type of the CP; Vcp further constrains the nominal host to be singular in number, which is also in line with theoretical descriptions of NV-CPS.

(34)  

\[ \begin{align*}
   &a. \quad S \rightarrow \{S\_unergr|S\_unacc|S\_trans|S\_cop|S\_cp\}.
\end{align*} \]
7.6. NOUN GROUPS IN HINDI/URDU GRAMMAR DEVELOPMENT

b. \( S_{cp} = KP: @SUBJ \)
   \( KP*: \{ @OBJ-GO \\
   \quad | @OBL \\
   \quad | @OBL-TH \\
   \quad | @OBL-SRC \\
   \quad | @OBJ \\
   \quad \{(! ANIM) =c + \\
   \quad (! CASE) =c acc \\
   \quad |-(! ANIM)}\} \),
   @CL-ADJUNCTS,
   @KP-SCRAMBLE,
   @POSS-PRON-SCRAMBLE,
   VC\(main: (^ VTYPE COMPLEX-PRED)\).

c. \( V_{cp} \rightarrow \{ N: @NV-ANNOTATION \\
   \quad (! NUM) =c sg \\
   \quad @(CP-TYPE nv) \\
   \quad | A: @AV-ANNOTATION \\
   \quad @(CP-TYPE av)\} \)
   Vlight-nav.

(35) is the (simplified) lexical entry of the verb kar ‘do’. It features two readings, a
main verb one (V-S) as well as a light verb one (Vlight-nav-S). The main verb read-
ing is not particularly interesting; in the light verb reading, the entry points to the
templates shown in (35b–d). Crucially, the subcategorization frame of the light verb
(i.e., the template in (35b)) provides a SUBJ as well as a %PRED2 slot; the latter will be
filled by the nominal host predicate in predicate composition. The feature (\( ^VTYPE \)
COMPLEX-PRED-FORM\) registers the light verb used to form the particular cp. Finally, the
AGENTIVE template encodes the fact that kar ‘do’ encodes an agentive act and thus needs
to have an agentive subject with ergative case.

(35) a. kar \( V-S \ XLE @(V-SUBJ-OBJ %stem)\);
   Vlight-nav-S \( XLE @(PRED-LIGHT-VERB %stem)\)
   \( @(CP-FORM kar)\)
   \( @AGENTIVE\} \).

b. \( PRED-LIGHT-VERB(_P) = (^ PRED) = '_P<(' SUBJ %PRED2)'\).

c. \( CP-FORM(_P) = (^ VTYPE COMPLEX-PRED-FORM) = _P\).

d. \( AGENTIVE = (^ LEX-SEM AGENTIVE) = +\).

The lexical entry for the nominal host of the running example, ifara ‘signal’, is shown
in (36). The noun is of the same type as gʰɛrao ‘circumvention’ in that it may occur with
a subject and an object that are both structurally genitive-marked when occurring inside
an NP; ifara does not select a locative or an instrumental argument. Important for the
present purposes is the third line of (36), which states that \textit{ifara} 'signal' may select for an object. It does so in the case at hand; its object argument, \textit{bilal}, is realized as the overall object of the clause in (33a).

(36) \texttt{iSArA NOUN-S XLE \{ @(SUBJ-OBJ\_core iSArA) \}
| @(SUBJ\_core iSArA)
| @(OBJ\_core iSArA)
| @(PRED iSArA))}.

Finally, the template \texttt{NV-ANNOTATION}, called by \texttt{Vcp} in (34c), handles CP formation. The template annotates the host’s f-structure (!) as restricted using a check feature. Then, it supplies !’s predicate as the second argument to the light verb’s predicate, thereby realizing predicate composition. In addition, for both types of N-V CPs, all features pertaining to the nominal host need to be restricted out from its f-structure, so that they do not end up in the topmost f-structure (by means of the host being a part of the main PRED). These two mechanisms are common to all N-V CPs.

Further, the template features two disjuncts. The first disjunct handles the non-agreeing N-V CP type; here, the host does not feature as a GF in the clause. The host’s object, however, is annotated as the main clause object. The second disjunct handles the agreeing CP type. Recall that here, the nominal host is both a part of the overall predicate as well as an object of the overall clause, requiring that the object’s f-structure be assigned; this is realized by the template \texttt{ASSIGN-NV-OBJ} in (37b). Finally, the lower disjunction in (37a) ensures that locative and instrumental arguments are mapped to the top f-structure, if present.

(37) a. \texttt{NV-ANNOTATION} = (! CHECK _RESTRICTED) = +
\texttt{^ PRED ARG2} = (! PRED)
\texttt{! ^ PRED CHECK NTYPE GEND NUM PERS} =
\texttt{^ OBJ} = (! OBJ)
\texttt{| @ASSIGN-NV-OBJ}
\texttt{^ OBL-TH} = (! OBL-TH)
\texttt{^ OBL-SRC} = (! OBL-SRC)}.

b. \texttt{ASSIGN-NV-OBJ} = (\texttt{^ OBJ PRED}) = '(! PRED FN)'
\texttt{^ OBJ CASE} = nom
\texttt{^ OBJ PERS} = (! PERS)
\texttt{^ OBJ CHECK} = (! CHECK)
\texttt{^ OBJ NTYPE} = (! NTYPE).
7.6.2 An Example Parse

Given the system described here, the sentence in (33) receives the c- and f-structures in Figures 7.11 and 7.12. The structures are correct in that they model a complex predicate-argument structure (cf. the complex top f-structure PRED) as well as a flat grammatical-functional structure.

![Diagram of C-structure for N-V CP](image1)

![Diagram of F-structure for N-V CP](image2)

Next, however, consider what happens when the example in (33) is slightly changed, and the verb kar ‘do’ is substituted for the verb de ‘give’. (38a) is the result; this sentence, however, is not a CP at all, but features a main verb de ‘give’. The noun ɪʃara ‘signal’ may be topicalized and coordinated, as seen in (38b) and (38c), respectively.
7.6. NOUN GROUPS IN HINDI/URDU GRAMMAR DEVELOPMENT

(38)  a. nadya=ne  bilal=ko  ḳारa  ḍi-yā
    ‘Nadya gave Bilal a signal.’

   b. ḳār a  nadya=ne  bilal=ko  ḍi-yā
    ‘A signal, Nadya gave Bilal.’

   c. nadya=ne  bilal=ko  ḳār a  ḍr  pārmarāj
give-PERF.M.SG
    ‘Nadya gave Bilal a signal and some advice.’

The main problem, however, is that de ‘give’ also features as a light verb, sporting the exact same annotation as kar ‘do’ in (38a). Thus, without further information, the grammar produces the (incorrect) c- and f-structure in Figure 7.13 and Figure 7.14 for (38a) (aside of the correct structures not shown here, where de ‘give’ is a main verb selecting three independent GFS). This is where the Urdu ParGram grammar can profit from the information garnered in the corpus studies described above.

Figure 7.13: Incorrect c-structure for (38a)

7.6.3 PREFERENCE TEMPLATES FOR N-V CP FORMATION

To model the preferences of nouns for CP formation when combining with a given light verb, the results in Table 7.12 were translated into grammar templates (§2.5.4) in the Urdu ParGram grammar. As stated above, the templates make use of ot marks that model the statistical preferences of a given noun to co-occur as a nominal host with a light verb. The templates thus reflect the clustering result shown in Table 7.12.

The noun template for Cluster 2 is shown below in (39). During the translation of the results in Table 7.12 to grammar templates, the following approach was taken. The verb that features most prominently in a given cluster is assigned an ot mark that prefers a
Figure 7.14: Incorrect f-structure for (38a)

CP analysis, in case there is a competing analysis that is not a CP. All other (non-most prominent) verbs of the cluster are given dispreference marks, in case they are used as light verbs; this does not rule out a CP analysis for those cases, but will punish it if another non-CP analysis is available. Thus, the template turns out to be a disjunct over three choices for the parser:

- either there is a light verb, and it is not kar ‘do’, then disprefer the analysis;
- or there is a light verb, and it is kar ‘do’, then prefer the analysis;
- or there is no light verb, then do nothing (i.e., do not prefer/disprefer the analysis).

(39) NV-GROUP2 = {(^ VTYPE COMPLEX-PRED-FORM)
   (~ VTYPE COMPLEX-PRED-FORM) ~ = kar
   @(OT-MARK cp-dispref)
   | (~ VTYPE COMPLEX-PRED-FORM) = c kar
   @(OT-MARK cp-pref)
   | (~(~ VTYPE COMPLEX-PRED-FORM)).

Corresponding templates have been implemented for the rest of the clusters in Table 7.12, but are not shown here. The lexical entry of ɪʃara ‘signal’ can now be augmented with a template call to (39), as seen in (40). Consequently, the incorrect parse corresponding to Figures 7.13 and 7.14 is dispreferred and not shown in XLE output anymore.
Note that OT marks can only prefer/disprefer any given analysis over another one if the two are in direct competition with one another. For example, OT marks will not be able to prefer an analysis with the light verb ho ‘be’ over an analysis with the light verb kar ‘do’ simply because the two of them will never be in direct competition with one another; the former requires a dative subject, while the latter occurs with ergative subjects.

A possible extension of this work would be to make use of OT marks for generation, and, in particular, for machine translation. Consider, for example, a rule-based machine translation system translating from English to Hindi. Given that English often does not distinguish different translations for different N-V CPs, one could use OT marks to derive a preference hierarchy of translation candidates, based on the absolute counts of N-V combinations (cf. Table 7.7). An example is the English verb comfort. This may be translated into Hindi using a variety of N-V CPS, among which are sahayata kar ‘assistance do’ as well as afvasan de ‘comfort give’. Assuming the translation system has rules that map to all of these translations, it will be hard to decide which one(s) to generate. A set of OT marks could easily be computed that prefers certain translation candidates over others, given the list of absolute N-V counts; the OT marks could then be appended to the rewrite rule mapping from the English main verb to the Hindi N-V CPS. The angle of this approach is a little different from the one described above, since preferences would have to be computed across nominal hosts instead of over light verb co-occurrence.

7.7 Conclusion

This concludes the chapter on N-V CPS. I have discussed the linguistic background of the construction, challenges for grammar development as well as related work on identifying different groups of N-V CPS. I have then proceeded to describe in detail two different corpus studies that were undertaken to further understand the constraints on the productive combination of nouns and light verbs in N-V CPS. The Urdu ParGram grammar was shown to be able to profit directly from the results of the second corpus study, in terms of grammar templates that model the statistical preferences of noun-light verb combinations adequately. The complete noun list from the corpus study reported on in §7.5, their template assignment, as well as the preference templates themselves are distributed as part of the Urdu ParGram grammar and are available either from the CD-ROM attached to this document or from the online INESS platform for treebanking and LFG grammar.
With respect to explaining the lexical-semantic constraints behind $n$-$v$ complex predication, a lot remains to be done. While both of the corpus studies have confirmed the findings by Ahmed and Butt (2011) regarding psych-predications and combinations that only lend themselves to agentive subjects, nouns belonging to Ahmed and Butt’s Class C could not be found, for various reasons. In addition, it has not become clear why certain nouns display a clear preference for a single light verb, and what exactly in their lexical semantics accounts for this behavior; here, further investigations involving the concept of “incorporation idioms” by Davison (2005) are expected to provide more insight.

\[\text{Iness homepage is located at } \text{http://iness.uib.no}\]
8 Discussion and Conclusion

8.1 Nominal Predication in Hindi/Urdu

This thesis proposes a radically new way of treating nominal predication in computational LFG grammar development. In this chapter, I summarize the main findings of the thesis and discuss some implications for the ParGram project as well as directions for future work.

In Chapters §3 and §5, I take a detailed look at different types of case marking in Hindi/Urdu NPs; it turns out that Hindi/Urdu makes use of predicating nominals and nominal arguments in quite different ways. A unified treatment simply cannot be maintained here, and looking closely at the predicational patterns, the functional behavior, as well as the constituent properties of nominal arguments is necessary in order to arrive at the correct generalizations.

Genitive phrases behave as subjects, objects or adjuncts of nominals; a differentiated functional treatment is indispensable if the patterns are to be modeled correctly. At the same time, the genitive is argued to be a strictly nominal, structural case, which is not licensed outside of the nominal domain. Most nominals can only be modified by a single genitive argument at a time, while a few may occur with two distinct subject and object genitives. In addition, all nominals may be modified by zero or more attributive (adjunct) genitives. Possessive clauses are reanalyzed as intransitives, involving complex noun phrases with embedded possessors. This novel analysis better accounts for the facts from binding as well as control.

As part of the analysis of the genitive, I reevaluate the Hindi/Urdu subjecthood test for reflexive binding by *apna* ‘self’, originally given by Mohanan (1994); I present evidence
that the reflexive may be bound by a genitive subject licensed inside the nominal domain, and redefine the diagnostic accordingly, in accordance with LFG’s binding theory.

Locative and instrumental arguments, on the other hand, are shown to behave as obliques. In particular, they involve case marking that semantically identifies the encoded thematic roles as themes and sources, and are selected by nominals that subcategorize for these roles. Locations are never licensed in Hindi/Urdu nominals, but are restricted to the realm of verbal predication. Locative clauses, used for locating entities in space, are also discussed and argued to involve locative inversion (Bresnan and Kanerva, 1989), explaining the grammatical-functional patterns in these clauses.

Nominal arguments may regularly be omitted; argument suppression, as discussed by e.g. Barker (1995), can account for non-overt nominal arguments. An analysis in terms of pro(nominal) drop was considered, but rejected, on morphosyntactic grounds as well as evidence from reflexive binding.

Finally, Chapter §7 presents issues involving noun-verb complex predicates (N-V CPSs). Here, nominal predication extends beyond the strictly nominal domain, in that nouns and light verbs form a single predicate with a single common set of grammatical functions (Mohanan, 1994); the noun may contribute additional arguments to the predication. In particular, the combinatory possibilities of nouns with different light verbs were examined. To that end, the chapter presents two distinct corpus studies that approach the challenge of identifying these possibilities from different angles. Both studies show that nouns exhibit statistical preferences to occur with specific light verbs, depending on the lexical semantics of the nouns.

From the point of view of implementing the generalizations using the XLE grammar development platform (Crouch et al., 2015), a number of different constructs, mechanisms and notations are necessary, including functional uncertainty (for scrambled nominal arguments), optimality-theory marks (for several preferences across constructions), shuffling of elements (for free word order), head precedence (for word order constraints), among others. The resulting Urdu grammar is included on the CD-ROM attached to this document and can be tested using the online INESS platform for treebanking and LFG grammar testing.\footnote{The INESS homepage is located at \url{http://iness.uib.no}}

As it stands, the thesis presents a complete analysis and implementation of nominal argument identification and classification for Hindi/Urdu, taking into account the different case-marking strategies (structural vs. semantic), agreement, scrambling phenomena, animacy requirements, word-order preferences and related facts, while at the same time being truthful to LFG theory.

8.2 IMPLICATIONS FOR PARGRAM

I present the state-of-the-art approach to nominal predication and the representation of nominal arguments in the ParGram project in §2.6. Recall that ParGram currently uses
8.3. THE FUTURE

the SPEC feature for encoding possessors; however, the novel analysis of genitive phrases proposed here departs from the standard ParGram analysis in suggesting a more detailed approach to genitive configurations. An underspecified treatment of possessors cannot pay tribute to the specific patterns of genitives/possessors in all languages; the situation in Hindi/Urdu clearly requires a more refined account (§4.7).

How does the account presented in this thesis relate to the issue of parallelism, which is central to ParGram? Recall that ParGram grammars are parallel in the sense that they need to adhere to commonly defined analyses and features that are guided by common linguistic principles (Butt et al., 1999a, 2002, 1999b, King et al., 2005, Sulger et al., 2013). ParGram aims at testing the LFG formalism for its universality and coverage limitations to see how far parallelism can be maintained across a larger array of languages. However, parallelism should not be maintained if it was at the cost of misrepresenting the linguistic facts; that is, dissimilar structures or features to encode the same meaning are only allowed if maintaining equal structures would be at the cost of misrepresenting a language.

With respect to nominal arguments, parallelism is thus faced with a conundrum. One can choose to either abstract away from the peculiarities of nominal arguments, and give a more shallow analysis, e.g., in terms of a SPEC POSS feature; or one can try to model the predicational properties more closely. The former strategy is at the cost of predicting the wrong facts and lacking explanatory appeal. The latter strategy as followed by myself results in a certain amount of ambiguity, not all of which can be resolved at the level of syntax (§3.9); but this cost is necessary in order to be able to explain the patterns in Hindi/Urdu.

8.3 THE FUTURE

There is more work to be done with respect to nominal predication in Hindi/Urdu and the Urdu ParGram grammar as well as generally throughout the ParGram project. With respect to Hindi/Urdu, the lexical semantics that are at play in N-V CPs are still not fully understood, and a full formalization of the issues is required. Here, the results from the corpus studies reported in Chapter 7 can inform further investigations. The Urdu grammar itself can benefit from improvements to the lexicon. The implementation of case marking as well as the general rule system is now in place; however, the lexicon is still small in comparison to other large-scale ParGram grammars. The expectation here is that the coverage can be greatly increased if (semi-) automatic methods of lexicon extension are used. Finally, a large-scale evaluation of the grammar has yet to be carried out.

A possible future developments with respect to grammar engineering within ParGram in general includes a re-examination of the SPEC POSS feature across grammars. The analysis described in this thesis can serve as a model analysis of nominal arguments/nominal

2I limit myself to the discussion of genitive/possessors here.
predication for those grammars that find that the SPEC POSS approach is insufficient for deriving the correct generalizations. Orientating oneself towards theoretical LFG accounts of possessives/genitives can provide further input. Overall, I remain optimistic that ParGram can instigate a treatment of nominal arguments, genitives, possessives, and nominal predication in general that is more faithful to linguistic theory as well as computationally efficient.
Bibliography


Bresnan, Joan (ed.). 1982c. The Mental Representation of Grammatical Relations. MIT Press.


Butt, Miriam and Zinsmeister, Heike. 2009. ESSLLI 2009 Course on Case, Scrambling and Default Word Order, course material.


BIBLIOGRAPHY


Raza, Ghulam and Ahmed, Tafseer. 2010. Order and Discontinuity within Urdu NPs, presentation given at the ILCW’10 (First International Linguistic Conference in Warsaw), Poland.


Spencer, Andrew. 2008. The possessum-agreement construction or ‘Does Albanian have a genitive case?’, manuscript, University of Essex.


Tutunjian, Damon and Boland, Julie E. 2008. Do we need a distinction between arguments and adjuncts? Evidence from psycholinguistic studies of comprehension. Language and Linguistics Compass 2, 631–646.


Colophon

This thesis was typeset using \TeX, originally developed by Leslie Lamport and based on Donald Knuth’s \TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (x11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu.