Towards Identifying Hindi/Urdu Noun Templates in Support of a Large-Scale LFG Grammar

Sebastian Sulger* & Ashwini Vaidya+

*Universität Konstanz
+University of Colorado at Boulder/Universität Konstanz

5th WSSANLP Workshop
COLING 2014, Dublin, Ireland
The situation

- Spoken and written Hindi/Urdu: heavy, *productive* use of complex predicates (CPs) across domains

- Different types of CPs:
  - Aspectual V+V CPs: *gir par* ‘suddenly fall’ (lit. ‘fall fall’)
  - Permissive V+V CPs: *jane de* ‘let go’ (lit. ‘go give’)
  - N+V CPs: *yad kar* ‘remember’ (lit. ‘memory do’)

- In other languages:
  - *take a bath* (≈ ‘bathe’)
  - *give a stir* (≈ ‘stir’)
  - *in Betracht ziehen* ‘consider’ (lit. ‘in look-at pull’)
  - Most of these are restricted in use and/or much less productive than South Asian CPs.
The challenges

- General problem in deep and shallow parsing methods for Hindi/Urdu (and other South Asian languages): proper treatment of complex predicates
  - Automatic distinction of CPs from simplex verbs
  - Extraction of subcategorization frames
  - Semantic role labeling
  - Drawing semantic inferences

Research questions:

What existing resources may be employed to explore CP usage?
Can we confirm/reject existing theoretical hypotheses of N+V CPs?
How far can clustering algorithms take us?
How “good” / “coherent” are the resulting classes?
How can our LFG grammar benefit from the results?
The construction

- Combination of noun and light verb to form a single predicational unit
- Noun contributes main predicational content (including argument(s)), light verb dictates case marking and expresses subtle lexical semantic differences
- Highly productive constructions
- [Ahmed and Butt, 2011]: proposal for different classes of N+V CPs based on a small case study of 45 nouns and 3 light verbs (kar ‘do’, ho ‘be’, hu- ‘become’)

<table>
<thead>
<tr>
<th>N+V type</th>
<th>light verb</th>
<th>analyis</th>
<th>example N</th>
</tr>
</thead>
<tbody>
<tr>
<td>class a</td>
<td>+</td>
<td>+</td>
<td>psych predications</td>
</tr>
<tr>
<td>class b</td>
<td>+</td>
<td>-</td>
<td>only agentive</td>
</tr>
<tr>
<td>class c</td>
<td>+</td>
<td>-</td>
<td>subject not an undergoer</td>
</tr>
</tbody>
</table>

**Table:** Classes of nouns identified by [Ahmed and Butt, 2011]
Goals of the investigation

- How do the proposals by [Ahmed and Butt, 2011] hold up towards a larger empirical basis (i.e., bigger corpora)?
- Extend the set of light verbs
- Apply different strategies of acquiring knowledge about CPs:
  - “Brute-force” statistical approach, based on bigram extraction, collocation analysis and clustering [Butt et al., 2012]
  - “Seed list” approach, using knowledge amassed from treebanks and clustering, and evaluate clusters (this paper)
- Come up with noun templates:
  - Nouns using one template will behave in a coherent way with respect to the light verbs they may occur with
  - Of great use for Hindi/Urdu grammar: extend noun lexicon/coverage
  - May inform further work on semantic classification of CPs
Contents

1 Hindi/Urdu Noun-Verb Complex Predicates

2 Corpus study

3 Evaluation

4 Analysis

5 Grammar integration
In a recent corpus study on Hindi\textsuperscript{1}, we used the approach below:

1. Collect corpus of 21 million words harvested from BBC Hindi website, Hindi wikipedia, and the Hindi-Urdu Treebank (HUTB) [Bhatt et al., 2009]
   - POS tagged, lemmatized using a state-of-the-art Hindi tagger [Reddy and Sharoff, 2011]

2. Look at a set of seven light verbs: \textit{kar} ‘do’, \textit{ho} ‘be’, \textit{de} ‘give’, \textit{le} ‘take’, \textit{rak\textsuperscript{h}} ‘put’, \textit{lag} ‘be attached’, \textit{a} ‘come’ (seven most frequently occurring light verbs)

\textsuperscript{1}The corresponding Urdu study is pending.
Methodology

3. Make use of the annotations in the HUTB [Bhatt et al., 2009]
   - Includes dependency annotation scheme
   - Employs label \( pof \) (for part of) to annotate complex predicates
   - Extract all items that are tagged as nouns and carry \( pof \) label
   → “Seed list” of nouns that we know take part in N-V CPs

4. Extract bigrams of pattern \( seed \ list \ noun \ item + light \ verb \ lemma \) from corpus
   - Assume that noun occurs right next to verb [Mohananan, 1994]
   - Cases where noun is removed from verb are rare (∼1% in HUTB)

5. Apply cutoff value \( c \) (noun occurrences across all light verbs)
   - Initial value \( c = 50 \): make statements about well-attested nouns
   - Also applied cutoff of \( c = 3 \) for comparison purposes
Methodology

Compute *relative frequencies* of nouns combined with light verbs
\( (c = 50: \, 522 \text{ nouns}; \, c = 3: \, 987 \text{ nouns}) \)

<table>
<thead>
<tr>
<th>ID</th>
<th>noun</th>
<th>kar 'do'</th>
<th>ho 'be'</th>
<th>de 'give'</th>
<th>le 'take'</th>
<th>rak'put'</th>
<th>lag 'attach'</th>
<th>a 'come'</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tanav 'tension'</td>
<td>0.115</td>
<td>0.562</td>
<td>0.058</td>
<td>0.058</td>
<td>0.000</td>
<td>0.000</td>
<td>0.207</td>
</tr>
<tr>
<td>2</td>
<td>b'hag 'part'</td>
<td>0.149</td>
<td>0.365</td>
<td>0.119</td>
<td>0.253</td>
<td>0.000</td>
<td>0.000</td>
<td>0.115</td>
</tr>
<tr>
<td>3</td>
<td>ag 'fire'</td>
<td>0.110</td>
<td>0.251</td>
<td>0.087</td>
<td>0.000</td>
<td>0.000</td>
<td>0.443</td>
<td>0.055</td>
</tr>
<tr>
<td>4</td>
<td>mazuri 'sanction'</td>
<td>0.000</td>
<td>0.000</td>
<td>0.757</td>
<td>0.243</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>d'hava 'attack'</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>kripa 'mercy'</td>
<td>0.409</td>
<td>0.486</td>
<td>0.000</td>
<td>0.000</td>
<td>0.105</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Table:** Relative frequencies of co-occurrence of nouns with light verbs
Methodology

7. Apply clustering algorithm to the data
   - Clustering the nouns based on their occurrence patterns with light verbs
   - $k$-means and GVM clustering

→ Problem: how to evaluate?
   - We already know that our combinations (seed list noun item + light verb lemma) form legitimate CPs.
   - What we don’t know is how semantically coherent the clusters are.
   - We also don’t know which $k$ and $c$ values give us the best (i.e. most expressive/semantically most coherent) clusters.

---

Contents

1 Hindustani/Urdu Noun-Verb Complex Predicates

2 Corpus study

3 Evaluation

4 Analysis

5 Grammar integration
Preliminary evaluation using WordNet

2. Follow the technique described by e.g. [Van de Cruys, 2006] for each $k = 2, ..., 10$ and for $c = \{3; 50\}$
   1. Extract synonyms, hypernyms and hyponyms for every noun in a cluster
   2. Choose cluster centroid: noun with most semantic relations with every other noun in cluster
   3. Extract co-hyponyms, i.e. the hyponyms of the hypernyms (sisters in the ontology tree), for each centroid from WordNet (along with their synonyms, hypernyms and hyponyms)
   4. Calculate coherence for each cluster: count number of nouns that overlap with nouns in centroid’s relations & divide by number of words in cluster
   5. Maximize coherence across $k$ and $c$
## Preliminary evaluation using WordNet

The table below shows the semantic coherence values for different sizes of $k$ and cluster numbers $c$ using the GVM and $k$-means algorithms.

<table>
<thead>
<tr>
<th>Size of $k$</th>
<th>$c = 3$</th>
<th>$c = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GVM</td>
<td>$k$-means</td>
</tr>
<tr>
<td>5</td>
<td>0.049</td>
<td>0.060</td>
</tr>
<tr>
<td>6</td>
<td>0.066</td>
<td>0.055</td>
</tr>
<tr>
<td>7</td>
<td>0.089</td>
<td>0.056</td>
</tr>
<tr>
<td>8</td>
<td>0.084</td>
<td>0.089</td>
</tr>
<tr>
<td>9</td>
<td>0.082</td>
<td>0.081</td>
</tr>
</tbody>
</table>

**Table:** Semantic coherence values for $k = 5 - 9$

→ Most coherent clusters according to evaluation with $k = 5$, $c = 50$ using $k$-means algorithm
Contents

1 Hindi/Urdu Noun-Verb Complex Predicates

2 Corpus study

3 Evaluation

4 Analysis

5 Grammar integration
Overview

Figure: Cluster visualization for \( k = 5, \ c = 50 \) (using tool by [Lamprecht et al., 2013])

Cluster description:

- Cl. 1 (light green): \( \text{kar} \) ‘do’
- Cl. 2 (dark blue): \( \text{ho} \) ‘be’
- Cl. 3 (pink): \( \text{de} \) ‘give’
- Cl. 4 (dark green): alternating between \( \text{rak}^h \) ‘keep’, \( \text{lag} \) ‘attach’, \( \text{a} \) ‘come’
- Cl. 5 (light blue): \( \text{le} \) ‘take’
Observations:

- Continuum between cl. 1/2: $kar/ho$ alternation, psych predication [Ahmed and Butt, 2011]
- Continuum between cl. 1/3: $kar/de$ alternation, “transfer”
- “Isolated” clusters 4/5: lexicalized incorporated idioms [Davison, 2005]

**Figure:** Cluster visualization for $k = 5$, $c = 50$ (using tool by [Lamprecht et al., 2013])
Productivity

<table>
<thead>
<tr>
<th>Light verb</th>
<th>Gloss</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>de</em></td>
<td>‘give’</td>
<td>0.75</td>
</tr>
<tr>
<td><em>kar</em></td>
<td>‘do’</td>
<td>0.08</td>
</tr>
<tr>
<td><em>le</em></td>
<td>‘take’</td>
<td>0.06</td>
</tr>
<tr>
<td><em>ho</em></td>
<td>‘be’</td>
<td>0.06</td>
</tr>
<tr>
<td><em>a</em></td>
<td>‘come’</td>
<td>0.02</td>
</tr>
<tr>
<td><em>rakh</em></td>
<td>‘keep’</td>
<td>0.02</td>
</tr>
<tr>
<td><em>lag</em></td>
<td>‘attach’</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Table:** LV properties of cluster 3, measured at centroid

- Frequencies: *likeness* of nouns in group to co-occur with LVs
- Productive patterns *more likely* to be valid CPs, less productive patterns *more likely* to be non-CP combinations [Butt et al., 2012]
Contents

1 Hindi/Urdu Noun-Verb Complex Predicates

2 Corpus study

3 Evaluation

4 Analysis

5 Grammar integration
Noun templates

<table>
<thead>
<tr>
<th>Light verb</th>
<th>Gloss</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>'give'</td>
<td>0.75</td>
</tr>
<tr>
<td>kar</td>
<td>'do'</td>
<td>0.08</td>
</tr>
<tr>
<td>le</td>
<td>'take'</td>
<td>0.06</td>
</tr>
<tr>
<td>ho</td>
<td>'be'</td>
<td>0.06</td>
</tr>
<tr>
<td>a</td>
<td>'come'</td>
<td>0.02</td>
</tr>
<tr>
<td>rak\textsuperscript{h}</td>
<td>'keep'</td>
<td>0.02</td>
</tr>
<tr>
<td>lag</td>
<td>'attach'</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Table:** LV properties of cluster 3, measured at centroid

- *Could* define grammar templates that model “absolute” choices:
  
  \[
  \text{NVGROUP3} = \{ \text{NV-CP-VERB} = \text{dE} \\
  \quad | \text{NV-CP-VERB} = \text{kar} \\
  \quad | \sim \text{NV-CP-VERB} \\
  \}. 
  \]

  iSArA NOUN @NVGROUP3.  

*signal (give/do)*
Even better: noun templates with optimality choices

<table>
<thead>
<tr>
<th>Light verb</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>0.75</td>
</tr>
<tr>
<td>kar</td>
<td>0.08</td>
</tr>
<tr>
<td>le</td>
<td>0.06</td>
</tr>
<tr>
<td>ho</td>
<td>0.06</td>
</tr>
<tr>
<td>a</td>
<td>0.02</td>
</tr>
<tr>
<td>rak\textsuperscript{h}</td>
<td>0.02</td>
</tr>
<tr>
<td>lag</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Table:** LV properties of cluster 3

**Advantages:**

- Do not make strict assertions about CP formation
- Rather, model statistical preferences over analyses
- Boost grammar coverage, robustness

**Or:** grammar templates that model *preferences*, via marks inspired by *Optimality Theory* (OT):

\[
\text{OPTIMALITYORDER cp-dispref non-cp-dispref} \\
\text{+cp-pref +non-cp-pref.}
\]

\[
\text{NVGROUP3} = \{ \text{NV-CP-VERB} = \text{dE} \text{ OT-MARK cp-pref} \text{ VERB = dE} \text{ OT-MARK non-cp-dispref} \text{ ...} \text{ NV-CP-VERB = lag} \text{ OT-MARK cp-dispref} \text{ VERB = lag} \text{ OT-MARK non-cp-pref} \}
\]

\[
i\text{SArA NOUN @NVGROUP3. signal (give/do)}
\]
Summary

- Some nouns heavily lexicalized towards a peculiar semantic configuration (i.e., compatible with a smaller subset of light verbs)
- Others may occur with a wider range of light verbs
- In dire need of further theoretical linguistic work to (possibly) link “noun templates” defined here with semantic classes

→ Use for grammar development?
  - Lexicon development
  - Can define templates, based on classification
  - Handle new coinages/borrowings, predict their usage

- Future work:
  - Apply method to Urdu data
  - Refine/narrow down clusters (using more data/more features/more light verbs)
Thank you all for your attention!
References

Discovering Semantic Classes for Urdu N-V Complex Predicates.


IndoWordNet.

Urdu and the Modular Architecture of ParGram.

Identifying Urdu Complex Predication via Bigram Extraction.

The Parallel Grammar Project.
References II

Urdu in a Parallel Grammar Development Environment.


A Visual Analytics System for Cluster Exploration.

*Argument Structure in Hindi*.
CSLI Publications.

Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources.
ParGramBank: The ParGram Parallel Treebank.

Semantic Clustering in Dutch.
Background — the Hindi/Urdu ParGram Grammar

- Computational LFG grammar in development in Konstanz
- Aim: large-scale LFG grammar for parsing Urdu/Hindi
- Overview publications are e.g. [Butt and King, 2007], [Bögel et al., 2009]
- Grammar is part of the ParGram project
  - Collaborative, world-wide research project
  - Development of parallel, linguistically well-motivated LFG grammars for a variety of languages
  - Features and analyses are kept parallel for easy transfer between languages
  - Languages involved: English, German, French, Indonesian, Japanese, Norwegian, Welsh, Georgian, Hungarian, Turkish, Chinese, Urdu/Hindi
- Overview publications are e.g. [Butt et al., 2002], [Sulger et al., 2013]
Background — the Hindi/Urdu ParGram Grammar

Example parse:
kīsan \( \text{ṭrektər=ko} \) bec-ta he
farmer.M.Sg.Nom tractor.M.Sg.Obl=Acc sell-Impf.M.Sg be.Pres.3.Sg
‘The farmer sells the tractor.’

More information on the Hindi/Urdu ParGram Grammar:
http://ling.uni-konstanz.de/pages/home/pargram_urdu/

More information on ParGram: http://pargram.b.uib.no/
Class A: psych predcitions

- Occur with all three light verbs examined by [Ahmed and Butt, 2011]

(1) a. larki=ne kahani 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 the story.’)

b. larki=ko kahani 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 the story is at the girl.’)

c. larki=ko kahani yad hu-i
   girl.F.Sg=Dat story.F.Sg memory.F.Sg be.Perf-F.Sg
   ‘The girl came to remember a/the story.’
   (lit. ‘Memory of the story became to be at the girl.’)
Class B: agentive CPs

- Require an agentive (ergative-marked) subject and light verb *kar* 'do'

\[(2)\]

a. bilal=ne makan tamir ki-ya  
   Bilal.M.Sg=Erg house.M.Sg construction.F.Sg do-Perf.M.Sg  
   'Bilal built a/the house.'  
   (lit. 'Bilal did construction of the house.')

b. * bilal=ko makan tamir he  
   Bilal.M.Sg=Dat house.M.Sg construction.F.Sg be.Pres.3.Sg

c. * bilal=ko makan tamir hu-a  
   Bilal.M.Sg=Dat house.M.Sg construction.F.Sg be.Perf-M.Sg
Grammar integration

Class C: subject not an undergoer

- Exclude the light verb *hu-* ‘become’

(3) a. bilal=ne yih ñarţtaslim k-i
Bilal.M.Sg=Erg this condition.F.Sg acceptance.M.Sg do-Perf.F.Sg
‘Bilal accepted this condition.’
(lit. ‘Bilal did acceptance of this condition.’)

b. bilal=ko yih ñarţtaslim he
Bilal.M.Sg=Dat this condition.F.Sg acceptance.M.Sg be.Pres.3.Sg
‘Bilal accepted this condition.’
(lit. ‘Acceptance of this condition was at Bilal.’)

c. * bilal=ko yih ñarţtaslim hu-a
Bilal.M.Sg=Dat this condition.F.Sg acceptance.M.Sg be.Perf-M.Sg
[Ahmed and Butt, 2011] looked at a set of three light verbs

Extending the set of light verbs brings up new questions

Nouns that occur with kar ‘do’ and de ‘give’ (but exclude other light verbs)

(4) a. nadya=ne laṛki=ko paramarʃ ː ki-ya
   Nadya.F.Sg=Erg girl.F.Sg=Acc advice.M.Sg do-Perf.M.Sg
   ‘Nadya advised the girl.’
   (lit. ‘Nadya did advice to the girl.’)

b. nadya=ne laṛki=ko paramarʃ ː di-ya
   Nadya.F.Sg=Erg girl.F.Sg=Acc advice.M.Sg give-Perf.M.Sg
   ‘Nadya advised the girl.’
   (lit. ‘Nadya gave advice to the girl.’)
Grammar integration

And beyond ...

- Nouns that occur with *kar ‘do’* only, not with *de ‘give’*

(5) a. bilal=ne makan tamir ki-ya
   Bilal.M.Sg=Erg house.M.Sg construction.F.Sg do-Perf.M.Sg
   ‘Bilal built a/the house.’
   (lit. ‘Bilal did construction of a/the house.’)
   [Ahmed and Butt, 2011, p. 3]

b. * bilal=ne makan tamir di-ya
   Bilal.M.Sg=Erg house.M.Sg construction.F.Sg give-Perf.M.Sg
Nouns that occur with *le* ‘take’ only, not with any other light verb

(6) a. nadya=ne  ləɾki=ko  god  lɨ-ya
   Nadya.F.Sg=Erg girl.F.Sg=Acc lap.F.Sg take-Perf.M.Sg
   ‘Nadya adopted the girl.’
   (lit. ‘Nadya took lap to the girl.’)

b. * nadya=ne  ləɾki=ko  god  kɨ-ya
   Nadya.F.Sg=Erg girl.F.Sg=Acc lap.F.Sg do-Perf.M.Sg

c. * nadya=ne  ləɾki=ko  god  dɨ-ya
   Nadya.F.Sg=Erg girl.F.Sg=Acc lap.F.Sg do-Perf.M.Sg