Lexical Resources for South Asian Languages

Miriam Butt

University of Konstanz

Chennai, 18.12.2011
State of the Art

NLP of South Asian languages has shown considerable progress over the last decade in terms of producing resusable resources.

- Corpora
NLP of South Asian languages has shown considerable progress over the last decade in terms of producing resusable resources.

- Corpora
- POS Taggers
State of the Art

NLP of South Asian languages has shown considerable progress over the last decade in terms of producing resusable resources.

- Corpora
- POS Taggers
- Treebanks (e.g., the Hindi Treebank)
NLP of South Asian languages has shown considerable progress over the last decade in terms of producing resusable resources.

- Corpora
- POS Taggers
- Treebanks (e.g., the Hindi Treebank)
- First Lexical Resources (e.g., Hindi WordNet; Bhattacharyya 2010)
State of the Art

This talk — focus on Lexical Resources
- Why they are important
This talk — focus on Lexical Resources

1. Why they are important
2. What there is for English (and some other languages)
This talk — focus on Lexical Resources

1. Why they are important
2. What there is for English (and some other languages)
3. Why we can’t just copy existing solutions/architectures
State of the Art

This talk — focus on Lexical Resources

1. Why they are important
2. What there is for English (and some other languages)
3. Why we can’t just copy existing solutions/architectures
4. Report on current work done in Konstanz as part of UrduGram
Deep Processing and Lexical Resources

- Any type of deeper NLP requires knowledge about the lexical structure of a language.
Any type of deeper NLP requires knowledge about the lexical structure of a language.

For this reason, much effort has been put into the construction of ontologies as well as verbal resources.
Deep Processing and Lexical Resources

- Any type of deeper NLP requires knowledge about the lexical structure of a language.

- For this reason, much effort has been put into the construction of ontologies as well as verbal resources.

  - **WordNet** — thesaurus containing information about the individual contentful words of a language (mainly nouns, verbs and adjectives)
Deep Processing and Lexical Resources

- Any type of deeper NLP requires knowledge about the lexical structure of a language.
- For this reason, much effort has been put into the construction of ontologies as well as verbal resources.
  - **WordNet** — thesaurus containing information about the individual contentful words of a language (mainly nouns, verbs and adjectives)
  - **VerbNet** — detailed classification of verbs in terms of their arguments and syntactic properties (based on Beth Levin’s work on English)
Deep Processing and Lexical Resources

- Any type of deeper NLP requires knowledge about the lexical structure of a language.
- For this reason, much effort has been put into the construction of ontologies as well as verbal resources.
  - **WordNet** — thesaurus containing information about the individual contentful words of a language (mainly nouns, verbs and adjectives)
  - **VerbNet** — detailed classification of verbs in terms of their arguments and syntactic properties (based on Beth Levin’s work on English)
  - **FrameNet** — verbs encoded with information about the type of event and all of the participants involved, not just the arguments (based on Fillmore’s work on English)
Deep Processing and Lexical Resources

- Any type of deeper NLP requires knowledge about the lexical structure of a language.
- For this reason, much effort has been put into the construction of ontologies as well as verbal resources.
  - WordNet — thesaurus containing information about the individual contentful words of a language (mainly nouns, verbs and adjectives)
  - VerbNet — detailed classification of verbs in terms of their arguments and syntactic properties (based on Beth Levin’s work on English)
  - FrameNet — verbs encoded with information about the type of event and all of the participants involved, not just the arguments (based on Fillmore’s work on English)
  - PropBank — provides argument role labels for verbal propositions in terms of frame sets (project initiated by Martha Palmer)
Types of Lexical Resources

- VerbNet, FrameNet and PropBank have similar aims, but differ in the details.
Types of Lexical Resources

- VerbNet, FrameNet and PropBank have similar aims, but differ in the details.
  - Arguments vs. Adjuncts
Types of Lexical Resources

- VerbNet, FrameNet and PropBank have similar aims, but differ in the details.
  - Arguments vs. Adjuncts
  - Hierarchical Classification into classes and classes (e.g., VerbNet)
Types of Lexical Resources

- VerbNet, FrameNet and PropBank have similar aims, but differ in the details.
  - Arguments vs. Adjuncts
  - Hierarchical Classification into classes and classes (e.g., VerbNet)
  - Precise type of information encoded
Types of Lexical Resources

- VerbNet, FrameNet and PropBank have similar aims, but differ in the details.
  - Arguments vs. Adjuncts
  - Hierarchical Classification into classes and classes (e.g., VerbNet)
  - Precise type of information encoded

- Why so many versions? — No consensus yet about how to best represent lexical semantic information.
Types of Lexical Resources

- VerbNet, FrameNet and PropBank have similar aims, but differ in the details.
  - Arguments vs. Adjuncts
  - Hierarchical Classification into classes and classes (e.g., VerbNet)
  - Precise type of information encoded

- Why so many versions? — No consensus yet about how to best represent lexical semantic information.

- Most of the work to date has been on English — crosslinguistic comparison should yield a better understanding on what to represent and how.
Types of Lexical Resources

- VerbNet, FrameNet and PropBank have similar aims, but differ in the details.
  - Arguments vs. Adjuncts
  - Hierarchical Classification into classes and classes (e.g., VerbNet)
  - Precise type of information encoded
- Why so many versions? — No consensus yet about how to best represent lexical semantic information.
- Most of the work to date has been on English — crosslinguistic comparison should yield a better understanding on what to represent and how.
- Some crosslinguistic work already underway, more is needed.
Types of Lexical Resources

- VerbNet, FrameNet and PropBank have similar aims, but differ in the details.
  - Arguments vs. Adjuncts
  - Hierarchical Classification into classes and classes (e.g., VerbNet)
  - Precise type of information encoded
- Why so many versions? — No consensus yet about how to best represent lexical semantic information.
- Most of the work to date has been on English — crosslinguistic comparison should yield a better understanding on what to represent and how.
- Some crosslinguistic work already underway, more is needed.
- In particular, very little work has been done on understanding and representing the lexical semantics of South Asian languages.
Deep Processing with Lexical Information

Before moving on to South Asian issues, a brief demo of an English question-answer system.

- Created at PARC (Palo Alto Research Center; Bobrow et al. 2007)
Deep Processing with Lexical Information

Before moving on to South Asian issues, a brief demo of an English question-answer system.

- Created at PARC (Palo Alto Research Center; Bobrow et al. 2007)
- Uses an LFG grammar for deep parsing
Deep Processing with Lexical Information

Before moving on to South Asian issues, a brief **demo** of an English question-answer system.

- Created at PARC (Palo Alto Research Center; Bobrow et al. 2007)
- Uses an LFG grammar for deep parsing
- Combined information from VerbNet, WordNet and Cyc (an Ontology) for lexical semantic resources (Unified Lexicon; Crouch and King 2005)
Deep Processing with Lexical Information

Before moving on to South Asian issues, a brief demo of an English question-answer system.

- Created at PARC (Palo Alto Research Center; Bobrow et al. 2007)
- Uses an LFG grammar for deep parsing
- Combined information from VerbNet, WordNet and Cyc (an Ontology) for lexical semantic resources (Unified Lexicon; Crouch and King 2005)
- System was scaled up and used by Bing for some time
Deep Processing with Lexical Information

Demo Summary:

- Detailed information about verbs (e.g., factive vs. non-factive verbs).
- Information about what arguments a noun derived from a verb can have.
- Information about active/passive relations.
- Information about synonyms, hypernyms, etc. (coming from WordNet)

We need the same (and more) for South Asian languages.

Concrete Example: The Urdu ParGram grammar being built at the University of Konstanz.
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
ParGram (Parallel Grammars)

- ParGram is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- Aim: Produce wide coverage grammars for a variety of languages.
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- Aim: Produce **wide coverage grammars** for a variety of languages.
- Theoretical Framework: **LFG** (Lexical-Functional Grammar)
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- **Aim**: Produce **wide coverage grammars** for a variety of languages.
- **Theoretical Framework**: **LFG** (Lexical-Functional Grammar)
- **Grammar Development Platform**: **XLE**
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- Aim: Produce **wide coverage grammars** for a variety of languages.
- Theoretical Framework: **LFG** (Lexical-Functional Grammar)
- Grammar Development Platform: **XLE**
  - XLE was developed and maintained at PARC (Palo Alto Research Center)
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- Aim: Produce **wide coverage grammars** for a variety of languages.
- Theoretical Framework: **LFG** (Lexical-Functional Grammar)
- Grammar Development Platform: **XLE**
  - XLE was developed and maintained at PARC (Palo Alto Research Center)
  - Implemented in C (C++), used with emacs, tcl/tk
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- Aim: Produce **wide coverage grammars** for a variety of languages.
- Theoretical Framework: **LFG** (Lexical-Functional Grammar)
- Grammar Development Platform: **XLE**
  - XLE was developed and maintained at PARC (Palo Alto Research Center)
  - Implemented in C (C++), used with emacs, tcl/tk
  - Includes a parser, generator and transfer component.
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- Aim: Produce *wide coverage grammars* for a variety of languages.
- Theoretical Framework: **LFG** (Lexical-Functional Grammar)
- Grammar Development Platform: **XLE**
  - XLE was developed and maintained at PARC (Palo Alto Research Center)
  - Implemented in C (C++), used with emacs, tcl/tk
  - Includes a parser, generator and transfer component.
- Sample Industrial Applications:
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- Aim: Produce **wide coverage grammars** for a variety of languages.
- Theoretical Framework: **LFG** (Lexical-Functional Grammar)
- Grammar Development Platform: **XLE**
  - XLE was developed and maintained at PARC (Palo Alto Research Center)
  - Implemented in C (C++), used with emacs, tcl/tk
  - Includes a parser, generator and transfer component.
- Sample Industrial Applications:
  - Question-Answer and Information Retrieval in Powerset (bing.com, Microsoft)
ParGram (Parallel Grammars)

- **ParGram** is a collaborative effort by industrial and academic institutions around the world (Butt et al. 1999).
- Aim: Produce **wide coverage grammars** for a variety of languages.
- Theoretical Framework: **LFG** (Lexical-Functional Grammar)
- Grammar Development Platform: **XLE**
  - XLE was developed and maintained at PARC (Palo Alto Research Center)
  - Implemented in C (C++), used with emacs, tcl/tk
  - Includes a parser, generator and transfer component.

- Sample Industrial Applications:
  - Question-Answer and Information Retrieval in Powerset (bing.com, Microsoft)
  - Fuji Xerox: Information Extraction from Medical Records/Texts
ParGram Languages so far

- Chinese (PARC)
- German (IMS, Stuttgart)
- English (PARC, Powerset)
- French (Xerox Grenoble, PARC)
- Georgian (Bergen)
- Hungarian (Debrecen)
- Indonesian (ANU, Canberra)
- Japanese (Fuji Xerox)
- Malagasy (Oxford)
- Norwegian (Bergen)
- Spanish (Powerset)
- Tigrinya (Bergen)
- Turkish (Istanbul)
- Urdu (Konstanz)
- Welsh (Essex)
Parallel Representations

The ParGram philosophy is that analyses and representations should be as parallel as possible across languages:

- c(onstituent)-structure is allowed to differ (surface realization)
Parallel Representations

The ParGram philosophy is that analyses and representations should be as parallel as possible across languages:

- c(onstituent)-structure is allowed to differ (surface realization)
- f(unctional)-structure should be as similar as possible (deep structure)
Parallel Representations

The ParGram philosophy is that analyses and representations should be as parallel as possible across languages:

- c(onstituent)-structure is allowed to differ (surface realization)
- f(unctional)-structure should be as similar as possible (deep structure)
- Advantages: easier Machine Translation and retrieval of semantically relevant information
Parallel Representations — Example

Non-Parallel C-structure: English vs. Urdu future tense (Butt et al. 2004)

CS 1: *TOP*
    | ROOT
    | Sadj[fin] PERIOD
    | S[fin] .
    | NP
    | VPall[fin]
    | NPadj
    | VP[fut,fin]
    | NPzero AUX[fut,fin] VPv[base]
    | N will V[base] NP
    | Mary
    | VPv[base]
    | N see NPadj
    | NPzero
    | N
    | Ram

CS 1: ROOT
    | S
    | KP
    | NP NP K V
    | NPzero
    | N
    | anjum
    | rAm
Parallel Representations — Example

Mostly Parallel F-structure: English vs. Urdu future tense

"Mary will see Ram."

```
PRED 'see<[1:Mary], [146:Ram]>'
PRED 'Mary'
CHECK [ _LEX-SOURCE morphology, _PROPER known-name ]
SUBJ NTYPE [ NSEM [ PROPER [ NAME-TYPE first_name, PROPER-TYPE name ] ]
           [ NSYN proper ] ]
   CASE nom, GEND-SEM female, HUMAN +, NUM sg, PERS 3
PRED 'Ram'
CHECK [ _LEX-SOURCE morphology, _PROPER known-name ]
OBJ NTYPE [ NSEM [ PROPER [ NAME-TYPE first_name, PROPER-TYPE name ] ]
           [ NSYN proper ] ]
   CASE obl, GEND-SEM male, HUMAN +, NUM sg, PERS 3
CHECK [ _SUBCAT-FRAME V-SUBJ-OBJ ]
TNS-ASP [ MOOD indicative, PERF --, PROG --, TENSE fut ]
CLAUSE-TYPE decl, PASSIVE --, VTYPE main

"anjum rAm kO dEkHEgI"

```
PRED 'dEkH<[1:anjum], [17:rAm]>'
PRED 'anjum'
SUBJ NTYPE [ NSEM [ PROPER [ PROPER-TYPE name ] ]
            [ NSYN proper ] ]
   CASE nom, GEND fem, NUM sg, PERS 3
PRED 'rAm'
CHECK [ _NMORPH obl ]
OBJ NTYPE [ NSEM [ PROPER [ PROPER-TYPE name ] ]
           [ NSYN proper ] ]
   CASE acc, GEND masc, NUM sg, PERS 3
CHECK [ _VMORPH [ _MTYPE infl ]
       [ _REstricted --, _VFORM fut ] ]
LEX-SEM [ AGENTIVE + ]
TNS-ASP [ MOOD indicative, TENSE fut ]
CLAUSE-TYPE decl, PASSIVE --, VTYPE main
```
ParGram Architecture

- Semantic Representation
- XFR/Transfer Rules
- C-structure Rules
- F-structure annotations
- Possibly other annotations
- Lexicon:
  1) irregular or special items
  2) Subcategorization frames
- Morphological
  - Finite-State Morpho-phonology
  - Stem Lexicon
  - F-structure annotations
- Tokenization (& Transliteration)
  - Identifying Words, Punctuation

Input String: Generated String
The goal for the Urdu ParGram grammar is to be large-scale and robust. We therefore work on all parts of the architecture (Butt and King 1997, Bögel et al. 2009, Bögel et al. 2007).

Also:

- Transliterator to allow for processing of both Urdu and Hindi script (Malik et al. 2010)

\[\begin{align*}
\text{Hindi} & \quad \text{Urdu} \\
\text{घर बना} & \quad \text{گھर بنا} \\
gHar \text{ banA}
\end{align*}\]
The goal for the Urdu ParGram grammar is to be large-scale and robust. We therefore work on all parts of the architecture (Butt and King 1997, Bögel et al. 2009, Bögel et al. 2007).

**Also:**

- Transliterator to allow for processing of both Urdu and Hindi script (Malik et al. 2010)

\[
\text{Hindi} \quad \text{Urdu} \\
\text{घर बना} \quad \text{گھر بنا} \\
g\text{Har banA}
\]

- Semi-automatic acquisition and integration of lexical resources
Lexical Resources in UrduGram

Work mainly done by: Tafseer Ahmed, Annette Hautli and Ghulam Raza.

- Semi-automatic acquisition of subcategorization frames (Raza)
Lexical Resources in UrduGram

Work mainly done by: Tafseer Ahmed, Annette Hautli and Ghulam Raza.

- Semi-automatic acquisition of subcategorization frames (Raza)
- Integration of Hindi WordNet (Ahmed and Hautli)
Lexical Resources in UrduGram

Work mainly done by: Tafseer Ahmed, Annette Hautli and Ghulam Raza.

- Semi-automatic acquisition of subcategorization frames (Raza)
- Integration of Hindi WordNet (Ahmed and Hautli)
- Work on an Urdu VerbNet (Ahmed and Hautli)
Lexical Resources in UrduGram

Work mainly done by: Tafseer Ahmed, Annette Hautli and Ghulam Raza.

- Semi-automatic acquisition of subcategorization frames (Raza)
- Integration of Hindi WordNet (Ahmed and Hautli)
- Work on an Urdu VerbNet (Ahmed and Hautli)
- Classification of various types of complex predicates (Ahmed and Butt)
Lexical Resources in UrduGram

Work mainly done by: Tafseer Ahmed, Annette Hautli and Ghulam Raza.

- Semi-automatic acquisition of subcategorization frames (Raza)
- Integration of Hindi WordNet (Ahmed and Hautli)
- Work on an Urdu VerbNet (Ahmed and Hautli)
- Classification of various types of complex predicates (Ahmed and Butt)
- Building of further resources in cooperation with Lahore (Sarmad Hussain) as part of a DAAD funded project.
South Asian Challenges

Challenges: Variety of Complex Predicates

- South Asian languages tend to have a small verbal inventory: Urdu/Hindi only has about 500–800 verbs.
Challenges: Variety of Complex Predicates

- South Asian languages tend to have a small verbal inventory: Urdu/Hindi only has about 500–800 verbs.
- Instead, complex predicates are a major part of most South Asian languages — how to treat these?
South Asian languages tend to have a small verbal inventory: Urdu/Hindi only has about 500–800 verbs.

Instead, complex predicates are a major part of most South Asian languages — how to treat these?

- No good ready-made solutions available.
South Asian languages tend to have a small verbal inventory: Urdu/Hindi only has about 500–800 verbs.

Instead, complex predicates are a major part of most South Asian languages — how to treat these?

- No good ready-made solutions available.
- Can maybe list the most frequently occurring ones as single items corresponding to a verb (e.g., Hindi WordNet)
South Asian languages tend to have a small verbal inventory: Urdu/Hindi only has about 500–800 verbs.

Instead, complex predicates are a major part of most South Asian languages — how to treat these?

- No good ready-made solutions available.
- Can maybe list the most frequently occurring ones as single items corresponding to a verb (e.g., Hindi WordNet)
- But they are very productive and there are many different types of classes, not all of them easy to identify or analyze (cf. Workshop on Complex Predicates yesterday).
The verbal inventory of South Asian languages also does not line up straightforwardly with that of languages like English or German.

- For example, there is no *have* — instead the verb for ‘be’ takes up a variety of roles.
Different Alignment of Verb Classes

The verbal inventory of South Asian languages also does not line up straightforwardly with that of languages like English or German.

- For example, there is no *have* — instead the verb for ‘be’ takes up a variety of roles.
- The verb classes established by Beth Levin for English do not necessarily reflect the verb classes of South Asian languages.
South Asian languages make systematic use of case and case alternations to express semantic differences.
South Asian Challenges

Semantically Motivated Case Alternations

- South Asian languages make systematic use of case and case alternations to express semantic differences.
- This systematic use has only recently begun to be explored seriously (e.g., Ahmed 2011, Butt and Ahmed 2012)
Semantically Motivated Case Alternations

- South Asian languages make systematic use of case and case alternations to express semantic differences.

- This systematic use has only recently begun to be explored seriously (e.g., Ahmed 2011, Butt and Ahmed 2012)
  - Ahmed (2011) — survey of 8 South Asian languages in terms of systematic employment of different types of (object) case to express semantic differences.
Semantically Motivated Case Alternations

- South Asian languages make systematic use of **case** and **case alternations** to express semantic differences.
- This systematic use has only recently begun to be explored seriously (e.g., Ahmed 2011, Butt and Ahmed 2012)
  - Ahmed (2011) — survey of 8 South Asian languages in terms of systematic employment of different types of (object) case to express semantic differences.
  - Butt and Ahmed (2011) — show that this feature was a systematic part of the language as far back as Sanskrit (but no work done on Dravidian).
Example from Nepali:

(1) a. 

\[ \text{hasan=} \text{le} \quad \text{gaar}i \quad \text{calāũ}-\text{c}^{\text{h}}\text{a} \]
Hassan=Erg car.Nom drive-NonPast.3.Sg
‘Hassan drives cars (that’s what he does).’

b. 

\[ \text{hasan} \quad \text{gaar}i \quad \text{calāũ}-\text{c}^{\text{h}}\text{a} \]
Hassan.Nom car.Nom drive-NonPast.3.Sg
‘Hassan is driving a car/cars.’
Automatic Identification of Arguments and Adjuncts

- In languages like English and German the identification of a predicate’s arguments is mostly straightforward.
Automatic Identification of Arguments and Adjuncts

- In languages like English and German the identification of a predicate’s arguments is mostly straightforward.
  - In English, position is a fairly good indicator.
Automatic Identification of Arguments and Adjuncts

- In languages like English and German the identification of a predicate’s arguments is mostly straightforward.
  - In English, position is a fairly good indicator.
  - In German, case marking is a fairly good indicator.
Automatic Identification of Arguments and Adjuncts

- In languages like English and German the identification of a predicate’s arguments is mostly straightforward.
  - In English, position is a fairly good indicator.
  - In German, case marking is a fairly good indicator.

- Semi-automatic acquisition of subcategorization information and automatic verb classification has therefore worked fairly well for these languages (e.g., Schulte im Walde 2006, Kuhn, Eckle and Rohrer 1998 for German).
Automatic Identification of Arguments and Adjuncts

- In languages like English and German the identification of a predicate’s arguments is mostly straightforward.
  - In English, position is a fairly good indicator.
  - In German, case marking is a fairly good indicator.

- Semi-automatic acquisition of subcategorization information and automatic verb classification has therefore worked fairly well for these languages (e.g., Schulte im Walde 2006, Kuhn, Eckle and Rohrer 1998 for German).

- In contrast, in a language like Urdu/Hindi, neither position nor case provide straightforward clues about the subcategorization frame of a verb (or a noun or an adjective).
Automatic Identification of Arguments and Adjuncts


- 10 million word corpus of newspaper texts (BBC Urdu, Jang, etc.)
Automatic Identification of Arguments and Adjuncts


- 10 million word corpus of newspaper texts (BBC Urdu, Jang, etc.)
- Corpus is unannotated

- 10 million word corpus of newspaper texts (BBC Urdu, Jang, etc.)
- Corpus is unannotated
  - no annotated corpus available when he began his work

- 10 million word corpus of newspaper texts (BBC Urdu, Jang, etc.)
- Corpus is unannotated
  - no annotated corpus available when he began his work
  - existing POS-taggers for Urdu (Hardie 2003, Sajjad 2007) are too coarse-grained to be helpful

- 10 million word corpus of newspaper texts (BBC Urdu, Jang, etc.)
- Corpus is unannotated
  - no annotated corpus available when he began his work
  - existing POS-taggers for Urdu (Hardie 2003, Sajjad 2007) are too coarse-grained to be helpful

**Goal:** figure out valency and type of arguments a verb takes, e.g., *de* ‘give’ takes three arguments

- 10 million word corpus of newspaper texts (BBC Urdu, Jang, etc.)
- Corpus is unannotated
  - no annotated corpus available when he began his work
  - existing POS-taggers for Urdu (Hardie 2003, Sajjad 2007) are too coarse-grained to be helpful

**Goal:** figure out valency and type of arguments a verb takes, e.g., *de* ‘give’ takes three arguments

- argument that is marked ergative in certain situations (with perfect morphology) (agent)

- 10 million word corpus of newspaper texts (BBC Urdu, Jang, etc.)
- Corpus is unannotated
  - no annotated corpus available when he began his work
  - existing POS-taggers for Urdu (Hardie 2003, Sajjad 2007) are too coarse-grained to be helpful

**Goal:** figure out valency and type of arguments a verb takes, e.g., *de* ‘give’ takes three arguments

1. argument that is marked ergative in certain situations (with perfect morphology) (agent)
2. argument that is marked with dative/accusative *ko* (goal)

- 10 million word corpus of newspaper texts (BBC Urdu, Jang, etc.)
- Corpus is unannotated
  - no annotated corpus available when he began his work
  - existing POS-taggers for Urdu (Hardie 2003, Sajjad 2007) are too coarse-grained to be helpful
- **Goal:** figure out valency and type of arguments a verb takes, e.g., *de* ‘give’ takes three arguments
  - argument that is marked ergative in certain situations (with perfect morphology) (agent)
  - argument that is marked with dative/accusative *ko* (goal)
  - argument that is unmarked (theme)
Automatic Identification of Arguments and Adjuncts

Results:

- in the absence of reliable annotation, much of the data is not reliable and must be filtered out
- for example, complex predicates (of which there are many) have an effect on the number and type of arguments

(2) a. 

\[
\text{\texttt{ali=ne} dosa k}^h\text{a-ya} \\
\text{Ali.M=Erg Dosa.M.Sg. eat-Perf.M.Sg} \\
\text{‘Ali ate a dosa.’ (simple verb, ergative subject)}
\]

b. 

\[
\text{\texttt{ali} dosa k}^h\text{a par-\texttt{a}} \\
\text{Ali.M Dosa.M.Sg. eat fall-Perf.M.Sg} \\
\text{‘Ali fell to eating a dosa.’ (complex predicate, unmarked subject)}
\]
Results:

- Raza’s system SASU does fairly well for the portion of the data that is reliable.
Results:

- Raza’s system SASU does fairly well for the portion of the data that is reliable.
- However, it turns out there are more challenges to be overcome (see Raza 2011).
Automatic Identification of Arguments and Adjuncts

Results:

- Raza’s system SASU does fairly well for the portion of the data that is reliable.
- However, it turns out there are more challenges to be overcome (see Raza 2011).
  - Unanticipated patterns (not noted anywhere before)
Results:

- Raza’s system SASU does fairly well for the portion of the data that is reliable.
- However, it turns out there are more challenges to be overcome (see Raza 2011).
  1. Unanticipated patterns (not noted anywhere before)
  2. Ambiguous patterns/multifunctionality of case markers
Results:

- Raza’s system SASU does fairly well for the portion of the data that is reliable.
- However, it turns out there are more challenges to be overcome (see Raza 2011).
  1. Unanticipated patterns (not noted anywhere before)
  2. Ambiguous patterns/multifunctionality of case markers
  3. Case marked arguments of nouns and adjectives
Results:

- Raza’s system SASU does fairly well for the portion of the data that is reliable.
- However, it turns out there are more challenges to be overcome (see Raza 2011).
  
  1. Unanticipated patterns (not noted anywhere before)
  2. Ambiguous patterns/multifunctionality of case markers
  3. Case marked arguments of nouns and adjectives
  4. Non-contiguous dependencies within the NP
Multifunctionality of Case Markers

(3) a. 
\[ \text{ali=ne nida=ko bula-ya} \]
Ali.M=Erg Nida.F=Acc call-Perf.M.3Sg
‘Ali called Nida.’ (Accusative Argument)

b. 
\[ \text{ali=ne nida=ko xat lik^h-a} \]
‘Ali wrote a letter to Nida.’ (Dative Argument)

c. 
\[ \text{ali rat=ko a-ya} \]
Ali.M. night.F=Temp come-Perf.M.3Sg
‘Ali came at night.’ (Temporal Adjunct)

d. 
\[ \text{ali g^h ar=ko ga-ya} \]
Ali.M home.M=Loc go-Perf.M.3Sg
‘Ali went home.’ (Locative Argument)
Multifunctionality of Case Markers

(4) a.  
\underline{\text{اَلْيِّن}} \underline{\text{کِبَیْسِ}} \underline{\text{کِوُل}} \underline{\text{کِ}}  
\text{Ali.M=Erg key.F.Sg-Inst door.M.Sg open-Perf.M.3Sg}  
‘Ali opened the door with a key.’ (\textit{Instrumental Adjunct})

b.  
\underline{\text{اَلْيِّن}} \underline{\text{نِدَا}} \underline{\text{بَت}} \underline{\text{کِ}}  
\text{Ali.M=Erg Nida.F=Com talk.M.Sg do-Perf.F.3Sg}  
‘Ali talked to Nida.’ (\textit{Comitative Argument})

c.  
\underline{\text{اَلْيِّن}} \underline{\text{تَزِ}} \underline{\text{دَر}} \underline{\text{کِ}}  
\text{Ali.M fastness.F=Inst run-Perf.M.3Sg}  
‘Ali ran quickly.’ (\textit{Adverbial Phrase})

d.  
\underline{\text{اَلْيِّن}} \underline{\text{غَر}} \underline{\text{کِ}} \underline{\text{کِ}}  
\text{Ali.M home.M=Abl come-Perf.M.3Sg}  
‘Ali came from home.’ (\textit{Locative Adjunct})

e.  
further uses: “made of”, “instrumental agent”, comparison with
Multifunctionality of Case Markers

Ahmed (2011): this type of multifunctionality is not confined to Urdu/Hindi, but is typical of South Asian languages
Multifunctionality of Case Markers

Ahmed (2011): this type of multifunctionality is not confined to Urdu/Hindi, but is typical of South Asian languages.

However, different languages break up the semantic space differently.

**Functions of the Urdu/Hindi case**
- Instruments
- Agents of passives
- Expressions of (dis)ability ('Nadya cannot walk')
- Non-affected and indirect causees
- Comitative/Sociative (e.g., 'speak with')
- Lexically required with certain verbs ('love', 'see')
- Temporal and spatial expressions with the meaning of source (ablative)
- Made of Material ('made of steel')
- Comparison
- Manner
Multifunctionality of Case Markers

Punjabi

- *nal* ‘with’
  - Instruments
  - Comitative/Sociative (e.g., ‘speak with’)
  - Manner
  - Made of Material (‘made of steel’)

- *tō* ‘from’
  - Agents of passives
  - Expressions of (dis)ability
  - Non-affected and indirect causees
  - Temporal and spatial expressions with source meaning (ablative)
  - Comparison
Multifunctionality of Case Markers

Conclusion:

- Need a very clear idea of the semantic range of case markers on a language by language basis.
Multifunctionality of Case Markers

Conclusion:

- Need a very clear idea of the semantic range of case markers on a language by language basis.
- Need a methodology to differentiate between the different uses of the same case form — not clear how this can be done (semi-)automatically.
Multifunctionality of Case Markers

Conclusion:

- Need a very clear idea of the semantic range of case markers on a language by language basis.
- Need a methodology to differentiate between the different uses of the same case form — not clear how this can be done (semi-)automatically.
- Definitely need to encode which verbs require/allow for which kind of argument in a lexical resource.
Conclusion:

- Need a very clear idea of the semantic range of case markers on a language by language basis.
- Need a methodology to differentiate between the different uses of the same case form — not clear how this can be done (semi-)automatically.
- Definitely need to encode which verbs require/allow for which kind of argument in a lexical resource.
- Then need to take a further step and identify verb classes that behave similarly.
Multifunctionality of Case Markers

Furthermore: get structural ambiguity in addition to lexical ambiguity (Raza 2011).

(5) a. 
\[
\text{nīḍa=ne } [\text{zūkam=se } \text{bācāo}=\text{ki} \text{ dāvai}] \text{ xārid-i} \\
\text{Nīḍa=Erg } [\text{flu=Abl protection}=\text{Gen.F medicine.F}] \text{ buy-Perf.F} \\
'Nīḍa purchased medicine for protection from flu.'
\]

b. 
\[
\text{nīḍa=ne } \text{bāzār=se } [\text{zūkam=ki} \text{ dāvai}] \text{ xārīd-i} \\
\text{Nīḍa=Erg bāzār=Abl } [\text{flu=Gen.M medicine.F.Sg}] \text{ buy-Perf.F} \\
'Nīḍa purchased medicine for flu from the market.'
\]
Yet a further complication (Raza 2011):

- Urdu nouns and adjectives can also take case-marked arguments (pattern is unlike English and German)
Yet a further complication (Raza 2011):

- Urdu nouns and adjectives can also take case-marked arguments (pattern is unlike English and German)
- Which nouns and adjectives take what kinds arguments depends on whether they are originally drawn from Arabic, Persian, or are native.
Arguments of Nouns and Adjectives

Yet a further complication (Raza 2011):

- Urdu nouns and adjectives can also take case-marked arguments (pattern is unlike English and German)
- Which nouns and adjectives take what kinds arguments depends on whether they are originally drawn from Arabic, Persian, or are native.
- The arguments of nouns and adjectives distribute within the NP in terms of non-local dependencies.
## Examples of Argument-taking adjectives in Urdu

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Type of Argument</th>
<th>Example of Adjective Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>Dative Marked</td>
<td>sadār=ko hasīl president=Dat possessed ‘possessed by the president’</td>
</tr>
<tr>
<td>(ii)</td>
<td>Ablative Marked</td>
<td>ḍāliyāh=se xaīf courts=Abl afraid ‘afraid of courts’</td>
</tr>
<tr>
<td>(iii)</td>
<td>Locative Marked</td>
<td>būxār=mē mūbtāla fever=Loc.in suffered ‘suffered with fever’</td>
</tr>
<tr>
<td>(iv)</td>
<td>Adpositional</td>
<td>sīhāt=ke liye mūzīr health=Gen for harmful ‘harmful for health’</td>
</tr>
</tbody>
</table>
### Examples of Argument-taking nouns in Urdu

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Type of Argument</th>
<th>Example of Noun Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>Ablative</td>
<td>무qaddamat=se یستسنا court-case.M.Pl=Abl immunity.M ‘immunity from court-cases’</td>
</tr>
<tr>
<td>(ii)</td>
<td>Locative</td>
<td>아lamti=par یبارفیق security.F=Loc briefing.F ‘Briefing on security’</td>
</tr>
</tbody>
</table>
Interaction between Arguments and Modifiers of Nouns

In NPs

- the noun head is to the very right (unless you have ezafe)
- simple adjective modifiers come just before the head
- arguments of the noun are separated from their noun head

(6) a. 
\[ \text{mũqaddamat}=\text{se} \quad \text{istı̇sna} \]
\text{court-case.M.Pl}=\text{Abl immunity.M} \\
‘Immunity from court-cases’

b. 
\[ \text{mũqaddamat}=\text{se} \quad \text{aini} \quad \text{istı̇sna} \]
\text{court-case.M.Pl}=\text{Abl constitutional immunity.M} \\
‘Constitutional immunity from court-cases’
More Complex Interaction

- All heads are stacked to the right, all modifiers to the left.
- The most natural version (and one most found in the corpus) is (7a).
- This makes it extremely difficult to determine which argument belongs to which head, especially when you have several arguments marked by the same form (e.g., genitives, se or par).

(7) a. \( \text{sador}=\text{ko}_1 \text{ muqaddamat}=\text{se}_2 \text{ hasil}_1 \)
   president=Dat  court-cases=Abl  possessed
   aini  istısta\text{na}_2
   constitutional immunity.M
   ‘Constitutional immunity from court-cases possessed by the president’

b. \( \text{muqaddamat}=\text{se}_2 \text{ sador}=\text{ko}_1 \text{ hasil}_1 \text{ aini istısta\text{na}_2} \)

c. *\( \text{hasil}_1 \text{ muqaddamat}=\text{se}_2 \text{ sador}=\text{ko}_1 \text{ aini istısta\text{na}_2} \)
Arguments of Nouns and Verbs

Conclusion

- Badly need a lexical resource that lists
Conclusion

- Badly need a lexical resource that lists
  - which verbs take which arguments
Conclusion

- Badly need a lexical resource that lists
  - which verbs take which arguments
  - which nouns and adjectives can take which types of arguments.
Arguments of Nouns and Verbs

Conclusion

- Badly need a lexical resource that lists
  - which verbs take which arguments
  - which nouns and adjectives can take which types of arguments.
- Building the noun/adjective resource (semi)-automatically will be a challenge.
Conclusion

- Badly need a lexical resource that lists
  - which verbs take which arguments
  - which nouns and adjectives can take which types of arguments.
- Building the noun/adjective resource (semi)-automatically will be a challenge.
- But if one can build on the following kind of information, the perhaps there is a chance:
Arguments of Nouns and Verbs

Conclusion

- Badly need a lexical resource that lists
  - which verbs take which arguments
  - which nouns and adjectives can take which types of arguments.
- Building the noun/adjective resource (semi)-automatically will be a challenge.
- But if one can build on the following kind of information, the perhaps there is a chance:
  - knowledge about the make-up the NP
Arguments of Nouns and Verbs

Conclusion

- Badly need a lexical resource that lists
  - which verbs take which arguments
  - which nouns and adjectives can take which types of arguments.

- Building the noun/adjective resource (semi)-automatically will be a challenge.

- But if one can build on the following kind of information, the perhaps there is a chance:
  - knowledge about the make-up the NP
  - some initial list of seed words
Arguments of Nouns and Verbs

Conclusion

- Badly need a lexical resource that lists
  - which verbs take which arguments
  - which nouns and adjectives can take which types of arguments.

- Building the noun/adjective resource (semi)-automatically will be a challenge.

- But if one can build on the following kind of information, the perhaps there is a chance:
  - knowledge about the make-up the NP
  - some initial list of seed words
  - an annotated corpus like the Hindi/Urdu treebank (Palmer et al. 2007, Bhatt et al. 2009)
Complex Predicates

A further complication is introduced by N-V and Adj-V complex predicates.

These also require arguments, but in a different manner than what we saw NP internally (syntax and semantics differs considerably).

(8) a. 
\[ \text{ali=}ne \quad \text{kahani} \quad [\text{yad} \quad \text{k-i}] \]
Ali.M=Erg story.F.Sg memory.M.Sg do-Perf.F.Sg
‘Ali remembered Nida.’

b. 
\[ \text{ali=}ne \quad \text{kamre=}\text{ko} \quad [\text{saf} \quad \text{ki-ya}] \]
Ali.M=Erg room.M.Obl=Acc clean do-Perf.M.Sg
‘Ali cleaned the room.’
Complex Predicates

At the moment, in the Urdu grammar light verbs may combine with nouns and adjectives quite freely — no semantic restrictions are implemented.
Complex Predicates

At the moment, in the Urdu grammar light verbs may combine with nouns and adjectives quite freely — no semantic restrictions are implemented.

Demo
At the moment, in the Urdu grammar light verbs may combine with nouns and adjectives quite freely — no semantic restrictions are implemented.

However, there appear to be lexical semantic restrictions...
Complex Predicates

- At the moment, in the Urdu grammar light verbs may combine with nouns and adjectives quite freely — no semantic restrictions are implemented.

- **Demo**

- However, there appear to be lexical semantic restrictions

- Ahmed and Butt (2011): corpus study showing that semantic factors such as eventivity vs. stativity of a noun and agentivity vs. experience of an action play a role in the combinatory possibilities.
At the moment, in the Urdu grammar light verbs may combine with nouns and adjectives quite freely — no semantic restrictions are implemented.

**Demo**

However, there appear to be lexical semantic restrictions

Ahmed and Butt (2011): corpus study showing that semantic factors such as eventivity vs. stativity of a noun and agentivity vs. experience of an action play a role in the combinatory possibilities.

**Conclusion:** Need very detailed information about the lexical semantics of verbs, adjectives and nouns.
VerbNet

At Konstanz, we have concretely begun work on an Urdu VerbNet.
VerbNet

- At Konstanz, we have concretely begun work on an Urdu VerbNet.
- Methodology is a combination of
VerbNet

- At Konstanz, we have concretely begun work on an Urdu VerbNet.
- Methodology is a combination of
  - original language-driven work
VerbNet

- At Konstanz, we have concretely begun work on an Urdu VerbNet.
- Methodology is a combination of
  1. original language-driven work
  2. jump-starting a system based on Levin’s classification for English (see example for ‘put’ verbs done by Hautli and Ahmed)
VerbNet

- At Konstanz, we have concretely begun work on an Urdu VerbNet.
- Methodology is a combination of
  1. original language-driven work
  2. jump-starting a system based on Levin’s classification for English (see example for ‘put’ verbs done by Hautli and Ahmed)

Problems
At Konstanz, we have concretely begun work on an Urdu VerbNet. Methodology is a combination of

1. original language-driven work
2. jump-starting a system based on Levin’s classification for English (see example for ‘put’ verbs done by Hautli and Ahmed)

Problems

The English verb class organization differs from what is found in Urdu (and South Asian languages in general)
VerbNet

- At Konstanz, we have concretely begun work on an Urdu VerbNet.
- Methodology is a combination of
  1. original language-driven work
  2. jump-starting a system based on Levin’s classification for English (see example for ‘put’ verbs done by Hautli and Ahmed)
- Problems
  - The English verb class organization differs from what is found in Urdu (and South Asian languages in general)
  - Complex predicates pose a problem
At Konstanz, we have concretely begun work on an Urdu VerbNet. Methodology is a combination of

1. original language-driven work
2. jump-starting a system based on Levin’s classification for English (see example for ‘put’ verbs done by Hautli and Ahmed)

Problems

The English verb class organization differs from what is found in Urdu (and South Asian languages in general)

Complex predicates pose a problem

Should they be treated on a par with simple verbs? (does not seem right)
At Konstanz, we have concretely begun work on an Urdu VerbNet. Methodology is a combination of

1. original language-driven work
2. jump-starting a system based on Levin’s classification for English (see example for ‘put’ verbs done by Hautli and Ahmed)

Problems

The English verb class organization differs from what is found in Urdu (and South Asian languages in general)

Complex predicates pose a problem

- Should they be treated on a par with simple verbs? (does not seem right)
- How can one tell complex predicates from a simple verb with modifiers like *barbecue* = ‘cook on an open fire in the outdoors’. 
At Konstanz, we have concretely begun work on an Urdu VerbNet. Methodology is a combination of

1. original language-driven work
2. jump-starting a system based on Levin’s classification for English (see example for ‘put’ verbs done by Hautli and Ahmed)

Problems

The English verb class organization differs from what is found in Urdu (and South Asian languages in general)

Complex predicates pose a problem

- Should they be treated on a par with simple verbs? (does not seem right)
- How can one tell complex predicates from a simple verb with modifiers like *barbecue* = ‘cook on an open fire in the outdoors’. 
- More precisely: how can one tell whether one has a verbal equivalent to an English verb in Urdu and when not?
Verb Classes

Very little work has been done on identifying verb classes for South Asian languages.

- Patterns of causativization differ across verbs.
Verb Classes

Very little work has been done on identifying verb classes for South Asian languages.

- Patterns of causativization differ across verbs.
- There seems to be a class of *ingestive* verbs which pattern alike (e.g., 'drink, eat, read, learn').
Verb Classes

Very little work has been done on identifying verb classes for South Asian languages

- Pattrns of causativization differ across verbs.
- There seems to be a class of *ingestive* verbs which pattern alike (e.g., 'drink, eat, read, learn').
- Some work on Unaccusatives vs. Unergatives (Bhatt, Ahmed 2010, Richa 2009)
Verb Classes

Based on Levin’s methods, Ahmed (2011) identifies several different verb classes across South Asian languages.

<table>
<thead>
<tr>
<th>Class</th>
<th>Subject Marking</th>
<th>2nd Arg. Marking</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>NOM/ERG, DAT</td>
<td>ABL</td>
<td>fear</td>
</tr>
<tr>
<td>II</td>
<td>NOM/ERG</td>
<td>ABL</td>
<td>resign</td>
</tr>
<tr>
<td>III</td>
<td>NOM/ERG</td>
<td>LOC-on/ DAT</td>
<td>attack, bless</td>
</tr>
<tr>
<td>IV</td>
<td>NOM/ERG, DAT</td>
<td>LOC-on/ DAT</td>
<td>trust, doubt</td>
</tr>
<tr>
<td>V</td>
<td>NOM/ERG</td>
<td>COM/DAT</td>
<td>meet, marry</td>
</tr>
<tr>
<td>VI</td>
<td>NOM/ERG, DAT</td>
<td>COM</td>
<td>love, hate</td>
</tr>
</tbody>
</table>
Summary and Conclusion

- Much more linguistic work needs to be done on the lexical semantics of South Asian languages.
Summary and Conclusion

Much more linguistic work needs to be done on the lexical semantics of South Asian languages.
  
  Investigation of case distribution and function
Summary and Conclusion

- Much more linguistic work needs to be done on the lexical semantics of South Asian languages.
  - Investigation of case distribution and function
  - Identification of verb classes based on Levin’s methods
Summary and Conclusion

- Much more linguistic work needs to be done on the lexical semantics of South Asian languages.
  - Investigation of case distribution and function
  - Identification of verb classes based on Levin’s methods
- This should be coupled with concerted computational efforts
Summary and Conclusion

- Much more linguistic work needs to be done on the lexical semantics of South Asian languages.
  - Investigation of case distribution and function
  - Identification of verb classes based on Levin’s methods
- This should be coupled with concerted computational efforts
  - Corpus studies to detect patterns hitherto unnoticed phenomena and patterns of distribution
Summary and Conclusion

- Much more linguistic work needs to be done on the lexical semantics of South Asian languages.
  - Investigation of case distribution and function
  - Identification of verb classes based on Levin’s methods
- This should be coupled with concerted computational efforts
  - Corpus studies to detect patterns hitherto unnoticed phenomena and patterns of distribution
  - Implementation and gradual improvement of automatic subcategorization acquisition algorithms.
Summary and Conclusion

- Much more linguistic work needs to be done on the lexical semantics of South Asian languages.
  - Investigation of case distribution and function
  - Identification of verb classes based on Levin’s methods

- This should be coupled with concerted computational efforts
  - Corpus studies to detect patterns hitherto unnoticed phenomena and patterns of distribution
  - Implementation and gradual improvement of automatic subcategorization acquisition algorithms.
  - Experiment with existing semantic clustering methods — not sure if these would work well . . .
It seems that much of the needed resources will have to be compiled manually and very slowly, as they were originally done for English.

But perhaps the process can be speeded/assisted if we do get better and better POS-taggers as well as more and more annotated corpora.

In either case: great and interesting challenges still lied ahead!


References II


