Identifying Urdu Complex Predication via Bigram Extraction

Miriam Butt[1] Tina Bögel[1] Annette Hautli[1]

Sebastian Sulger[1] Tafseer Ahmed[2]

[1]University of Konstanz, Germany
[2]University of Karachi, Pakistan

COLING 2012 in Mumbai, India

The situation

- Spoken and written language in Urdu/Hindi: heavy usage of complex predicates (CPs)
- Different types of CPs (Butt 1995):
  - Aspectual V+V CPs: *gɪr paṛ-na* 'to fall suddenly (lit. fall fall)'
  - Permissive V+V CPs: *jane de-na* 'to let go (lit. go give)'
  - ADJ+V CPs: *saf kar-na* 'to clean (lit. clean do)'
  - N+V CPs: *yad kar-na* 'to remember (lit. memory do)'
- In other languages:
  - *take a bite out of X* (lit. to bite X)
  - *give X a stir* (lit. to stir X)
  - *außer Acht lassen* 'to ignore (lit. let out of sight)'
- General problem in shallow and deep parsing approaches to Urdu/Hindi: proper treatment of complex predicates

The challenges

- Automatic distinction of CPs from simplex verbs
- Extraction of subcategorization frames
- Semantic role labeling
- Drawing semantic inferences

The challenges

- Automatic distinction of CPs from simplex verbs
- Extraction of subcategorization frames
- Semantic role labeling
- Drawing semantic inferences

Research questions:

- Can we blindly apply common statistical methods to extract the relevant patterns?
- Can we confirm existing theoretical hypotheses of N+V CP classes?
- Can visualization help us with this task?

Outline

N+V CPs

- Combination of a *noun* which adds the main predicational content and a *light verb* which expresses subtle lexical semantic differences
- Highly productive constructions
- Proposal for different classes of N+V complex predicates based on a small case study (Ahmed and Butt 2011)

N+V CPs

- Combination of a *noun* which adds the main predicational content and a *light verb* which expresses subtle lexical semantic differences
- Highly productive constructions
- Proposal for different classes of N+V complex predicates based on a small case study (Ahmed and Butt 2011)

N+V Type	**Light Verb** kar 'do'	ho 'be'	hu 'become'	**Analyis**
CLASS A	+	+	+	psych-predications
CLASS B	+	−	−	only agentive
CLASS C	+	+	−	subject ≠ undergoer

N+V CPs

Class A: Psych predications (**Noun** + **Light verb**)

(1) √ laṛki=ne kahani **yad** **k-i**
girl.F.Sg=Erg story.F.Sg.Nom **memory.F.Sg.Nom** **do-Perf.F.Sg**
'The girl remembered a/the story.'
(lit.: 'The girl did memory of the story.')

N+V CPs

Class A: Psych predications (**Noun** + **Light verb**)

(2) √ laṛki=ne kahani **yad** **k-i**
girl.F.Sg=Erg story.F.Sg.Nom **memory.F.Sg.Nom** **do-Perf.F.Sg**
'The girl remembered a/the story.'
(lit.: 'The girl did memory of the story.')

√ laṛki=ko kahani **yad** **hɛ**
girl.F.Sg=Dat story.F.Sg.Nom **memory.F.Sg.Nom** **be.Pres.3P.Sg**
'The girl remembers/knows a/the story.'
(lit.: 'Memory of the story is at the girl.')

N+V CPs

Class A: Psych predications (**Noun** + **Light verb**)

(3) √ laṛki=ne kahani **yad** **k-i**
girl.F.Sg=Erg story.F.Sg.Nom **memory.F.Sg.Nom** **do-Perf.F.Sg**
'The girl remembered a/the story.'
(lit.: 'The girl did memory of the story.')

√ laṛki=ko kahani **yad** **hɛ**
girl.F.Sg=Dat story.F.Sg.Nom **memory.F.Sg.Nom** **be.Pres.3P.Sg**
'The girl remembers/knows a/the story.'
(lit.: 'Memory of the story is at the girl.')

√ laṛki=ko kahani **yad** **hu-i**
girl.F.Sg=Dat story.F.Sg.Nom **memory.F.Sg.Nom** **become-F.Sg**
'The girl came to remember a/the story.'
(lit.: 'Memory of the story became to be at the girl.')

N+V CPs

Class B: Agentive (transitive) CPs (**Noun** + **Light verb**)

(4) √ bılal=ne mɑkan **tɑmir** **ki-ya**
Bilal.M.Sg=Erg house.M.Sg.Nom **construction.F.Sg** **do-Perf.M.Sg**
'Bilal built a/the house.'

N+V CPs

Class B: Agentive (transitive) CPs (**Noun** + **Light verb**)

(5) √ bɪlal=ne mɑkan **tɑmir** **ki-ya**
Bilal.M.Sg=Erg house.M.Sg.Nom **construction.F.Sg** **do-Perf.M.Sg**
'Bilal built a/the house.'

— *bɪlal=ko mɑkan **tɑmir** **hɛ**
Bilal.M.Sg=Dat house.M.Sg.Nom **construction.F.Sg** **be.Pres.3.Sg**

N+V CPs

Class B: Agentive (transitive) CPs (**Noun** + **Light verb**)

(6) √ bɪlal=ne mɑkan **tɑmir** **ki-ya**
Bilal.M.Sg=Erg house.M.Sg.Nom **construction.F.Sg** **do-Perf.M.Sg**
'Bilal built a/the house.'

— *bɪlal=ko mɑkan **tɑmir** **hɛ**
Bilal.M.Sg=Dat house.M.Sg.Nom **construction.F.Sg** **be.Pres.3.Sg**

—
*bɪlal=ko mɑkan **tɑmir** **hu-a**
Bilal.M.Sg=Dat house.M.Sg.Nom **construction.F.Sg** **become-M.Sg**

N+V CPs

Class C: Subject no undergoer (**Noun** + **Light verb**)

(7) √ bɪlal=ne yɪh ʃɑrṭ **tɑslim** **ki**
Bilal.M.Sg=Erg this condition.F.Sg **acceptance.M.Sg** **do-Perf.F.Sg**
'Bilal accepted this condition.'

N+V CPs

Class C: Subject no undergoer (**Noun** + **Light verb**)

(8) √ bɪlal=ne yɪh ʃɑrṭ **tɑslim** **ki**
Bilal.M.Sg=Erg this condition.F.Sg **acceptance.M.Sg** **do-Perf.F.Sg**
'Bilal accepted this condition.'

√ bɪlal=ko yɪh ʃɑrṭ **tɑslim** **hɛ**
Bilal.M.Sg=Dat this condition.F.Sg **acceptance.M.Sg** **be-3.Sg**
'Bilal accepted this condition.'

N+V CPs

Class C: Subject no undergoer (**Noun** + **Light verb**)

(9) √ bɪlal=ne yɪh ʃarṭ **taslim** **ki**
Bilal.M.Sg=Erg this condition.F.Sg **acceptance.M.Sg** **do-Perf.F.Sg**
'Bilal accepted this condition.'

√ bɪlal=ko yɪh ʃarṭ **taslim** **hɛ**
Bilal.M.Sg=Dat this condition.F.Sg **acceptance.M.Sg** **be-3.Sg**
'Bilal accepted this condition.'

???
bɪlal=ko yɪh ʃarṭ **taslim** **hui**
Bilal.M.Sg=Dat this condition.F.Sg **acceptance.M.Sg** **become-F.Sg**

Our investigation

- Confirm the proposal by Ahmed and Butt (2011) with a larger empirical basis
- Extend the number of light verbs to four:
  1. *kar* 'do'
  2. *ho* 'be'
  3. *hʊ* 'become'
  4. *rak*^{h} 'put'
- Start "naively" with commonly used statistical measures
- See whether these measures work for our data

Outline

Extraction

Steps:

1. Use raw corpus of 7.9 million words harvested from the BBC Urdu website
2. Extract all bigrams which have one of the four light verbs as the right element
3. Data clean-up
4. Rank bigrams with the X^2 measure
5. Throw away bigrams with weak co-occurrence strength

Extraction

6. Combine bigram lists to show the relative frequency of each noun with each light verb

		Relative frequencies with light verbs			
ID	Noun	*kar*	*ho*	*hu*	*rakH*
1	*h2Asil* 'achievement'	0.771	0.222	0.007	0.000
2	**a2*lAn* 'announcement'	0.982	0.011	0.007	0.000
3	*bAt* 'talk'	0.853	0.147	0.000	0.000
4	*SurUa2* 'beginning'	0.530	0.384	0.086	0.000

Automatic transliteration as in Bögel (2012): unknown short vowels are represented as '*'

Hold-ups

- Spelling variation in Urdu words
- Inconsistent usage of "real" white space and zero-width non-joiner
- Homonymy
  - *ki* either feminine perfective form of *kar* 'do' or genitive marker
- Homography
  - *kyA* → 'that', *kıyA* → 'do.Perf.M.Sg'
- Nouns can be scrambled away from their light verbs
  → Bigram approach helpless
- Light verbs can also be main verbs and auxiliaries in Urdu
  → Much noise

Clustering

Automatic clustering of the data set

- Clusters based on the pattern of relative co-occurrence with the four light verbs
- Problem: How good are these clusters?

Clustering

Automatic clustering of the data set

- Clusters based on the pattern of relative co-occurrence with the four light verbs
- Problem: How good are these clusters?

→ **Visual analysis** of the data set

Outline

The concept

- Tight coupling of algorithms for automatic data analysis with visual components
- Eight visual variables: *position* (two variables x and y), *size*, *value*, *texture*, *color*, *orientation* and *shape*
- Exploit human perceptive abilities to support pattern detection

Purpose of visualization

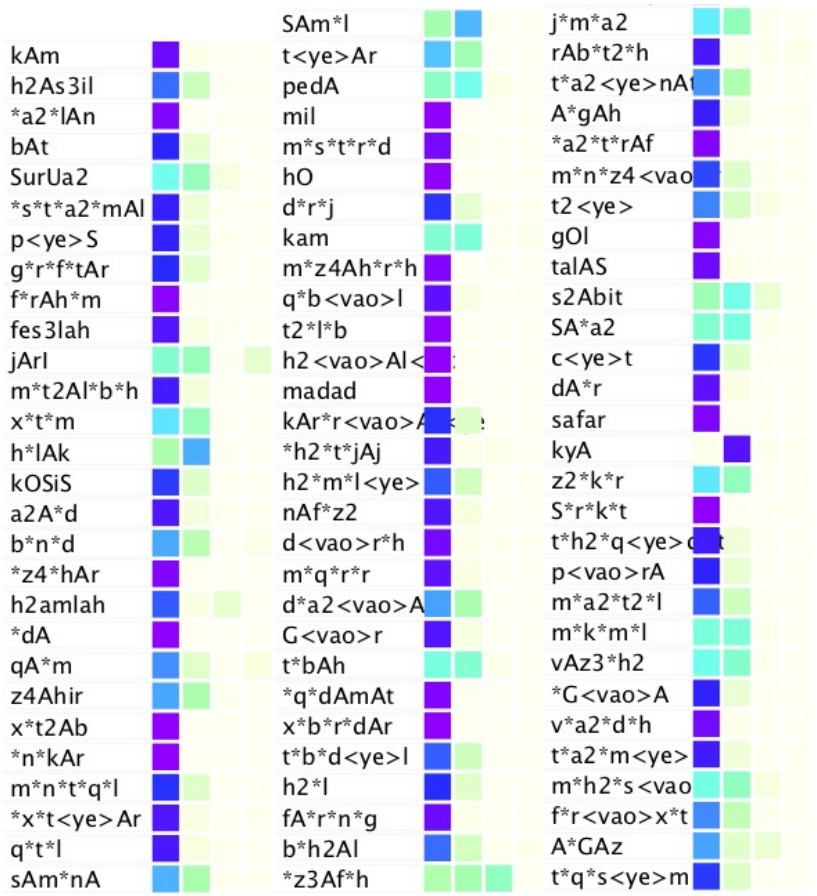
1. Overview of complex data sets
2. Starting point for an interactive exploration of data
3. Generation of new hypotheses, verification of existing hypotheses

Visualization – round 1

```
ID,kar,ho,hu,rakH
kAm,0.953,0.047,0.000,0.000
h2As3il,0.771,0.222,0.007,0.000
*a2*lAn,0.982,0.011,0.007,0.000
bAt,0.853,0.147,0.000,0.000
SurUa2,0.530,0.384,0.086,0.000
*s*t*a2*mAl,0.873,0.121,0.006,0.000
p<ye>S,0.864,0.131,0.005,0.000
g*r*f*tAr,0.841,0.159,0.000,0.000
f*rAh*m,0.995,0.005,0.000,0.000
fes3lah,0.920,0.065,0.015,0.000
jArI,0.461,0.379,0.005,0.155
m*t2Al*b*h,0.896,0.104,0.000,0.000
x*t*m,0.611,0.376,0.012,0.000
h*lAk,0.288,0.692,0.020,0.000
kOSiS,0.823,0.177,0.000,0.000
a2A*d,0.912,0.088,0.000,0.000
b*n*d,0.695,0.261,0.000,0.045
*z4*hAr,0.981,0.019,0.000,0.000
h2*m*l*h,0.790,0.064,0.146,0.000
*dA,1.000,0.000,0.000,0.000
qA*m,0.733,0.170,0.019,0.079
z4Ahir,0.699,0.289,0.012,0.000
x*t2Ab,1.000,0.000,0.000,0.000
*n*kAr,1.000,0.000,0.000,0.000
m*n*t*q*l,0.835,0.165,0.000,0.000
*x*t<ye>Ar,0.914,0.075,0.000,0.011
q*t*l,0.902,0.078,0.019,0.000
sAm*nA,0.686,0.301,0.013,0.000
SAm*l,0.322,0.678,0.000,0.000
t<ye>Ar,0.659,0.341,0.000,0.000
pedA,0.416,0.527,0.057,0.000
mil,1.000,0.000,0.000,0.000
m*s*t*r*d,0.970,0.030,0.000,0.000
h0,1.000,0.000,0.000,0.000
d*r*j,0.831,0.147,0.023,0.000
kam,0.466,0.485,0.018,0.031
m*z4Ah*r*h,0.982,0.000,0.018,0.000
q*b<vao>l,0.941,0.059,0.000,0.000
t2*l*b,1.000,0.000,0.000,0.000
h2<vao>Al<ye>,1.000,0.000,0.000,0.000
madad,1.000,0.000,0.000,0.000
kAr*r<vao>A*<ye>,0.835,0.165,0.000,0.000
*h2*t*jAj,0.899,0.047,0.055,0.000
h2*m*l<ye>,0.791,0.209,0.000,0.000
nAf*z2,0.917,0.083,0.000,0.000
d<vao>r*h,0.966,0.034,0.000,0.000
m*q*r*r,0.935,0.065,0.000,0.000
d*a2<vao>A,0.708,0.292,0.000,0.000
G<vao>r,0.914,0.086,0.000,0.000
t*bAh,0.503,0.457,0.040,0.000
*q*dAmAt,0.983,0.017,0.000,0.000
x*b*r*dAr,1.000,0.000,0.000,0.000
t*b*d<ye>l,0.786,0.214,0.000,0.000
h2*l,0.836,0.164,0.000,0.000
fA*r*n*g,0.956,0.044,0.000,0.000
b*h2Al,0.769,0.205,0.000,0.026
*z3Af*h,0.280,0.315,0.405,0.000
j*m*a2,0.594,0.395,0.011,0.000
rAb*t2*h,0.903,0.041,0.024,0.032
t*a2<ye>nAt,0.720,0.280,0.000,0.000
A*gAh,0.879,0.100,0.000,0.021
*a2*t*rAf,0.986,0.014,0.000,0.000
m*n*z4<vao>r,0.811,0.176,0.012,0.000
t2<ye>,0.742,0.199,0.059,0.000
g0l,0.985,0.015,0.000,0.000
talAS,0.963,0.037,0.000,0.000
s2Abit,0.349,0.521,0.130,0.000
SA*a2,0.452,0.505,0.043,0.000
c<ye>t,0.828,0.172,0.000,0.000
dA*r,0.938,0.062,0.000,0.000
safar,0.977,0.023,0.000,0.000
kyA,0.049,0.928,0.023,0.000
z2*k*r,0.607,0.393,0.000,0.000
S*r*k*t,1.000,0.000,0.000,0.000
t*h2*q<ye>qAt,0.893,0.107,0.000,0.000
p<vao>rA,0.860,0.140,0.000,0.000
m*a2*t2*l,0.784,0.216,0.000,0.000
m*k*m*l,0.492,0.481,0.027,0.000
vAz3*h2,0.537,0.450,0.013,0.000
*G<vao>A,0.863,0.137,0.000,0.000
v*a2*d*h,0.967,0.033,0.000,0.000
t*a2*m<ye>r,0.891,0.109,0.000,0.000
m*h2*s<vao>s,0.515,0.404,0.081,0.000
f*r<vao>x*t,0.737,0.242,0.021,0.000
A*GAz,0.702,0.167,0.131,0.000
t*q*s<ye>m,0.825,0.175,0.000,0.000
```

- Difficulty with detecting patterns among bare figures
- Requirement of a visual cue for the inspection of the clusters

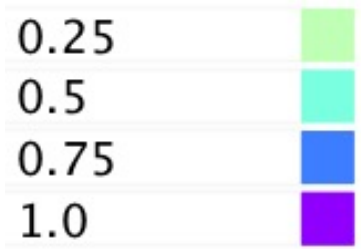
Visualization – round 1



- Difficulty of detecting patterns among bare figures
- Requirement of a visual cue for the inspection of the clusters

Visualization – round 1

- Mapping of relative frequencies to the visual variable *color*
- The higher the frequency, the darker the color
- Reference visualization of relative frequencies:



- Proportional mapping between relative frequency and color

Visualization – round 1

Raw data

Noun	*kar*	*ho*	*hu*	*rakH*
h2Asil	0.771	0.222	0.007	0.000
*a2*lAn	0.982	0.011	0.007	0.000
bAt	0.853	0.147	0.000	0.000
SurUa2	0.530	0.384	0.086	0.000

Visualization – round 1

Raw data

Noun	*kar*	*ho*	*hu*	*rakH*
h2Asil	0.771	0.222	0.007	0.000
*a2*lAn	0.982	0.011	0.007	0.000
bAt	0.853	0.147	0.000	0.000
SurUa2	0.530	0.384	0.086	0.000

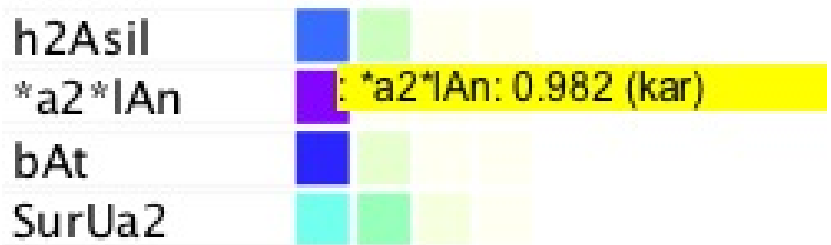
Visualized data

Visualization – round 1

Raw data

Noun	*kar*	*ho*	*hu*	*rakH*
h2Asil	0.771	0.222	0.007	0.000
*a2*lAn	0.982	0.011	0.007	0.000
bAt	0.853	0.147	0.000	0.000
SurUa2	0.530	0.384	0.086	0.000

Visualized data



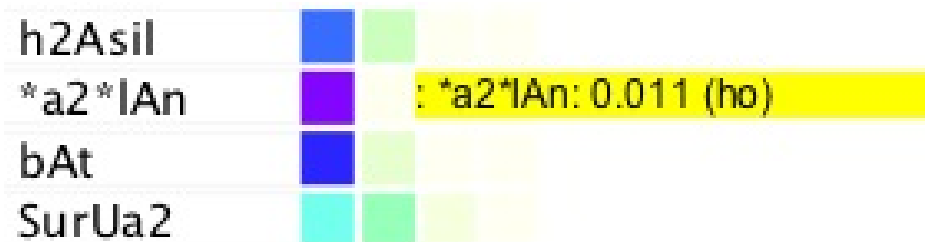
- Tool facilitates zooming and mousing over to see the underlying data set

Visualization – round 1

Raw data

Noun	*kar*	*ho*	*hu*	*rakH*
h2Asil	0.771	0.222	0.007	0.000
*a2*lAn	0.982	0.011	0.007	0.000
bAt	0.853	0.147	0.000	0.000
SurUa2	0.530	0.384	0.086	0.000

Visualized data



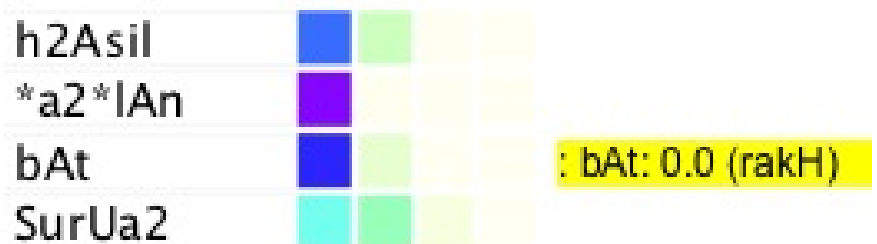
- Tool facilitates zooming and mousing over to see the underlying data set

Visualization – round 1

Raw data

Noun	*kar*	*ho*	*hu*	*rakH*
h2Asil	0.771	0.222	0.007	0.000
*a2*lAn	0.982	0.011	0.007	0.000
bAt	0.853	0.147	0.000	0.000
SurUa2	0.530	0.384	0.086	0.000

Visualized data



- Tool facilitates zooming and mousing over to see the underlying data set

Visualization – round 1

Raw data

Noun	*kar*	*ho*	*hu*	*rakH*
h2Asil	0.771	0.222	0.007	0.000
*a2*lAn	0.982	0.011	0.007	0.000
bAt	0.853	0.147	0.000	0.000
SurUa2	0.530	0.384	0.086	0.000

Visualized data

- Tool facilitates zooming and mousing over to see the underlying data set

Visualization – round 1

Benefits of visualizing the initial clustering result

- At-a-glance detection of outliers, e.g. behavior of the verb *uTHA* 'to lift'

- Quick detection of clusters within clusters
- Visual evaluation of the goodness of the clustering

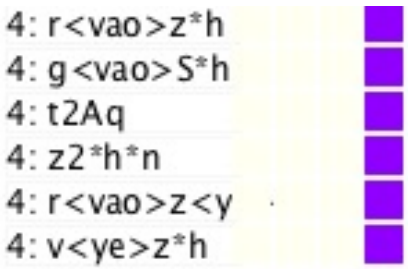
Visualization – round 1

Result:

- K-means clustering with k=5 best clustering algorithm according to the visualization
- Removal of clusters with consistently false hits (clusters 1, 3 and 4)
- Reduction of the list of bigrams from around 20.000 bigrams to 1.090
- Clusters 0 and 2 with many N+V and ADJ+V CPs are kept

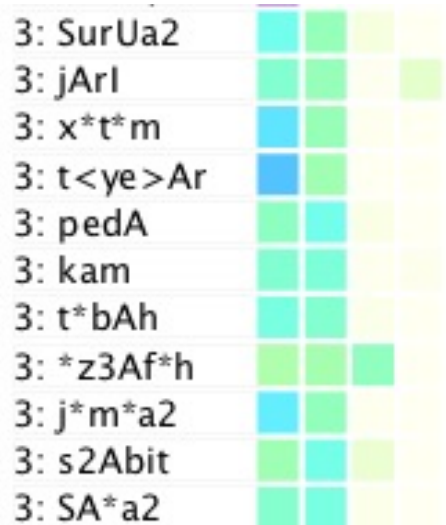
Next step: Reclustering and visualization of the reduced data set

Visualization – round 2



Cluster 4:

- Much co-occurence of item with *rak*^{h} ‘put’
- Mixed cluster without complex predicates

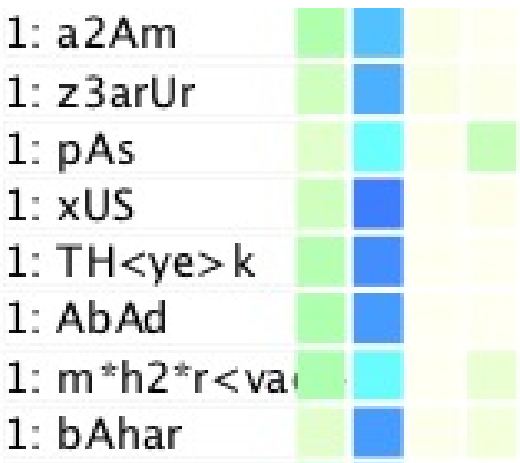


Cluster 3:

- Items occur equally often with *kar* ‘do’ and *ho* ‘be’
- Cluster contains mostly ADJ+V sequences but hardly any CPs

Visualization – round 2

Cluster 1:



- Occurs mostly with *ho* ‘be’ and *kar* ‘do’
- Cluster contains mostly ADJ+V sequences (also some valid N+V complex predicates)
- Interpreted as resultative constructions

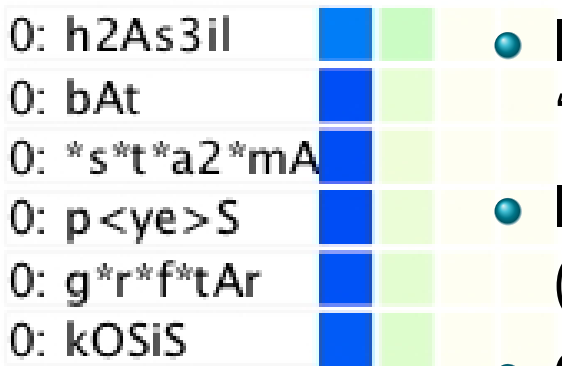
Visualization – round 2

Cluster 2:

- Largest cluster of all (around 600 members)
- Cluster 2 contains mostly N+V sequences, but not all are N+V CPs
- If N+V CP, then of class B in Ahmed and Butt (2011) (no dative subjects allowed)

Visualization – round 2

Cluster 0:



- Items occur mostly with *kar* ‘do’ and *ho* ‘be’
- Items also possible with *hu* ‘become’ (known from theoretical investigations)
- Contains valid N+V complex predicates that correspond to Ahmed & Butt’s class A (psych predications)

Visualization – round 2

Result:

N+V Type	Light Verb: kar 'do'	Light Verb: ho 'be'	Light Verb: hu 'become'	Analyis
CLASS A	+	+	+	psych-predications
CLASS B	+	−	−	only agentive
CLASS C	+	+	−	subject ≠ undergoer

- N+V CPs of class A and B can be extracted from corpora
- Class C is difficult to detect

Discussion

Data sparsity

- Known N+V combinations are not present in the corpus
- Problem of missing POS-tagged text for the language

Discussion

Data sparsity

- Known N+V combinations are not present in the corpus
- Problem of missing POS-tagged text for the language

BUT:

- Partial confirmation of the N+V CP classes established by Ahmed and Butt (2011)
- Detection of A+V CPs
- Facilitation of data cleanup using visual keys
- Evaluation of clusters using methods from visualization

Future work

- Exploration of N+V and ADJ+V CPs in POS-tagged corpora (Urooj et al. 2012)
- Exploit existing information to extract scrambled N+V CPs
- Further extension of the visualization component:
  - Increasing the interaction with the data
  - Development of different methods for cluster visualization

Summary

Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?

Summary

Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?

  No, linguistic knowledge is required.

Summary

Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?

  No, linguistic knowledge is required.

- Can we confirm existing theoretical hypotheses of N+V CP classes?

Summary

Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?

  No, linguistic knowledge is required.

- Can we confirm existing theoretical hypotheses of N+V CP classes?

  Yes, some clusters correspond to theoretically motivated CP classes.

Summary

Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?

  No, linguistic knowledge is required.

- Can we confirm existing theoretical hypotheses of N+V CP classes?

  Yes, some clusters correspond to theoretically motivated CP classes.

- Can visualization help us with this task?

Summary

Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?

  No, linguistic knowledge is required.

- Can we confirm existing theoretical hypotheses of N+V CP classes?

  Yes, some clusters correspond to theoretically motivated CP classes.

- Can visualization help us with this task?

  Definitely!

Thank you!

شکریہ