Parsing under-resourced languages: Cross-lingual transfer strategies for Indian languages

Loganathan Ramasamy,
Advisor: Zdeněk Žabokrtský

Charles University in Prague

Sep 15, 2014

Outline

Current scenario in treebanking

- Around 30 treebanks are available in total.
- Many languages including some of the popular languages do not have treebanks.

#	Language	# Speakers	Size	Source
7	Bengali	193M	small	[HMAG10]
8	Russian	162M	large	[BGG+00]
9	Japanese	122M	medium	[KB00]
10	Javanese	84.3M	-	-
11	German	83.8M	large	[BDH+02]
12	Lahnda	82.7M	-	-
13	Telugu	74.0M	small	[HMAG10]
14	Marathi	71.8M	-	-
15	Tamil	68.8M	small	[Rv12]

Table: Treebanks (of all formalisms) availability for languages by number of speakers. $>$ 250000 tokens = *large*; 50000-250000 tokens =*medium*; $<$ 50000 tokens = *small*. Source: http://www.ethnologue.com/statistics/size as on June 7, 2013

Approaches to languages without treebanks

- Unsupervised methods
- Semi-supervised methods
- Cross-lingual transfer methods

Which approach is the best?

- Obviously supervised learning
- There is no single best approach when there's no treebank data
- Every language contains some amount of non-treebank resources (such as parallel corpus, POS tagger, translation system, etc.)
- Cross-lingual methods
  - Exploit resource-rich languages (key: concept of language universals)
  - Use resources such as parallel corpus and POS taggers

Thesis goals and contributions

- Goal
  - The main of goal of the thesis is to identify best cross-lingual transfer strategies for Indian languages (ILs).
- Contributions
  - Cross-lingual dependency transfer results for Indian languages (ILs)
  - New under-resourced scenarios such as using machine translated parallel texts instead of human translated texts
  - New treebank resource plus parallel corpus for Tamil language

Are Indian languages under-resourced?

- Basic Language Resource Kit (BLARK) as defined by [Kra03]

  "the minimal set of language resources that is necessary to do any precompetitive research and education at all"

- BLARK includes [Kra03]: spoken/written language corpora, grammars, modules (taggers morphological analyzers, parsers, speech recognizers, text-to-speech), annotation standards and tools, and so on.

Are Indian languages under-resourced? Simple survey

Lang.	Treebank	Parallel corpora	POS tagger	Web corpora	GT
Hindi	HyDT	JHU HindEnCorp EMILLE	✓	W2C EMILLE	✓
Bengali	HyDT	JHU EMILLE	✓	W2C EMILLE	✓
Telugu	HyDT	JHU	✓	W2C EMILLE	✓
Marathi	×		✓	W2C EMILLE	✓
Tamil	TamilTB	EnTam JHU	✓	W2C EMILLE	✓
Urdu	UDT	UMC005 JHU EMILLE	✓	W2C EMILLE	✓
Gujarati	×	EMILLE	✓	W2C EMILLE	✓
Kannada	×		✓	W2C EMILLE	✓
Malayalam	×	JHU	✓	W2C EMILLE	×
Oriya	×			EMILLE	×

Table: Resource availability for Indian languages.

Approaches that we use

- Bitext projection
- Delexicalized parsing
- Transfer with machine translated texts

Table of Contents

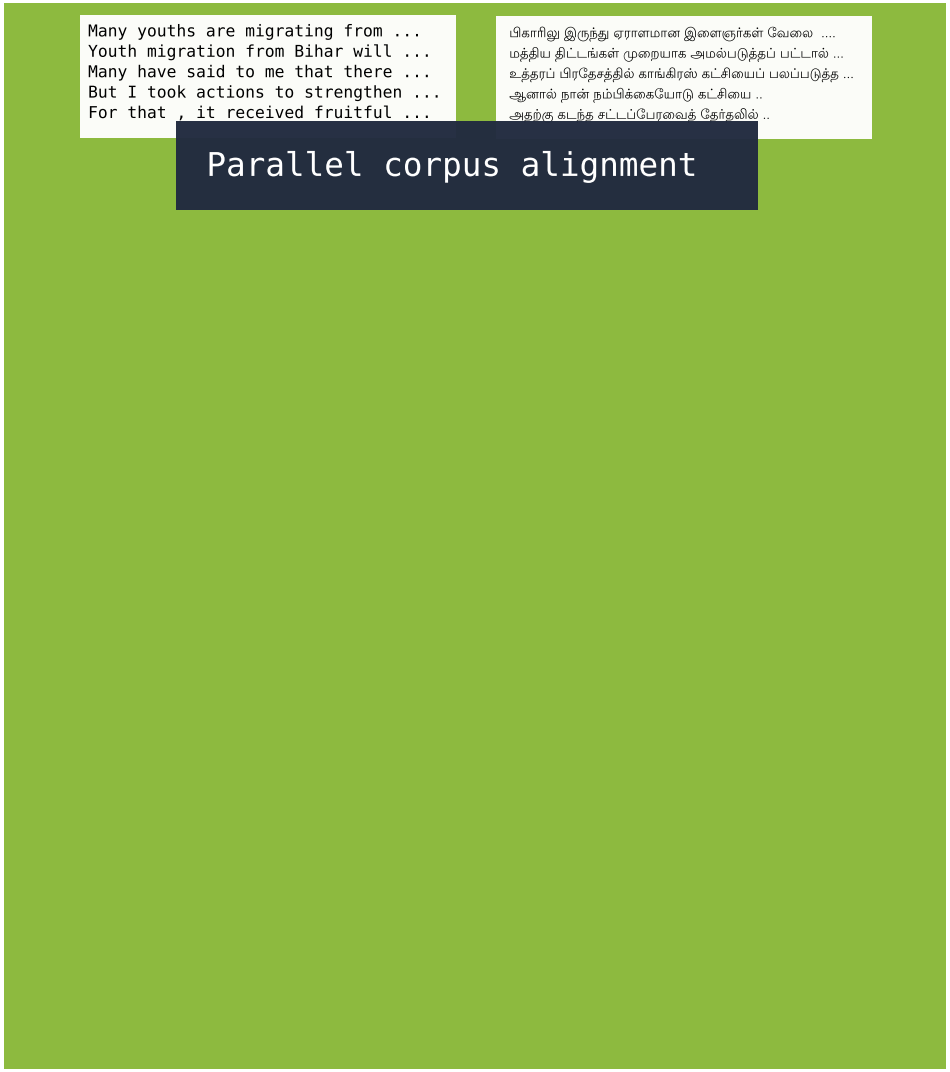
Overview

- [HRW02, HRW$^+$05] was the first to use projection based approach to transfer syntactic dependencies.
  - Direct Correspondence Assumption (DCA)
- Modified algorithm
  - Initialization of target tree
  - Chunk head identification

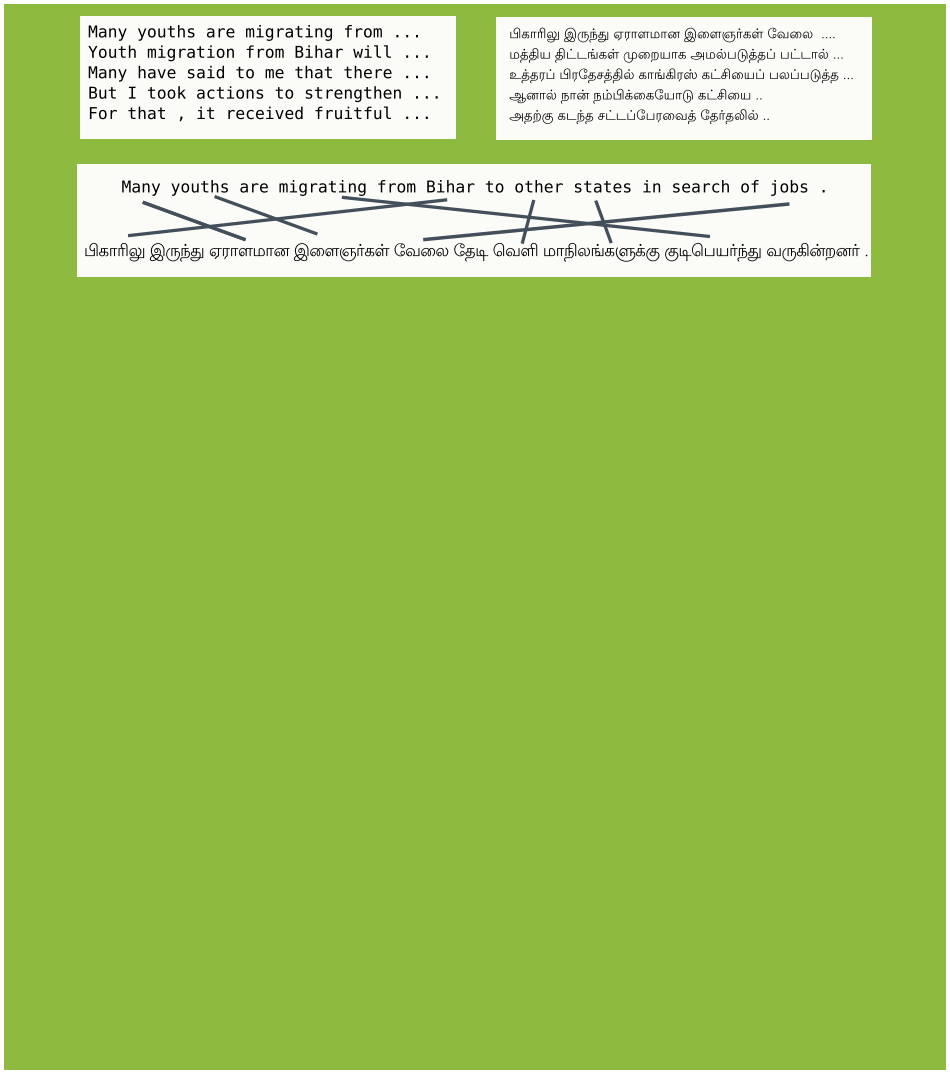
Overview


Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..

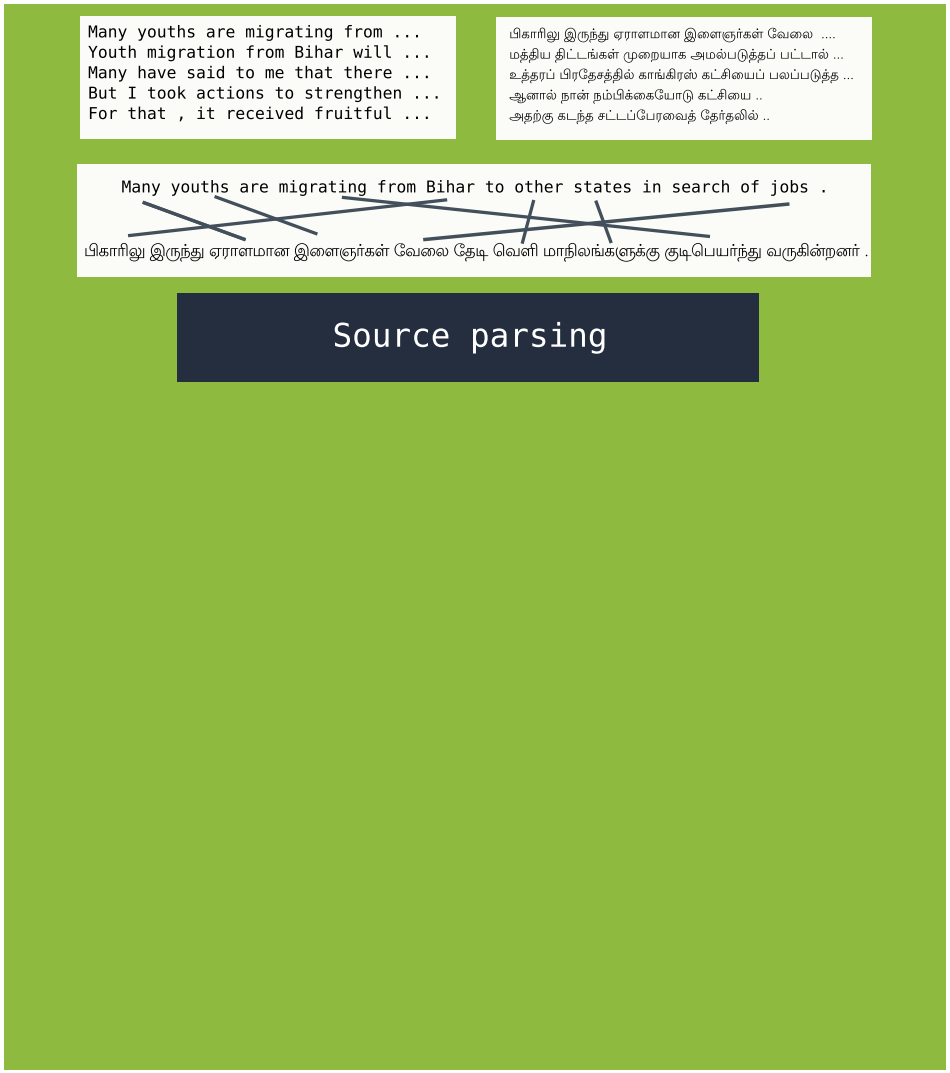
Overview


Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..
Parallel corpus alignment

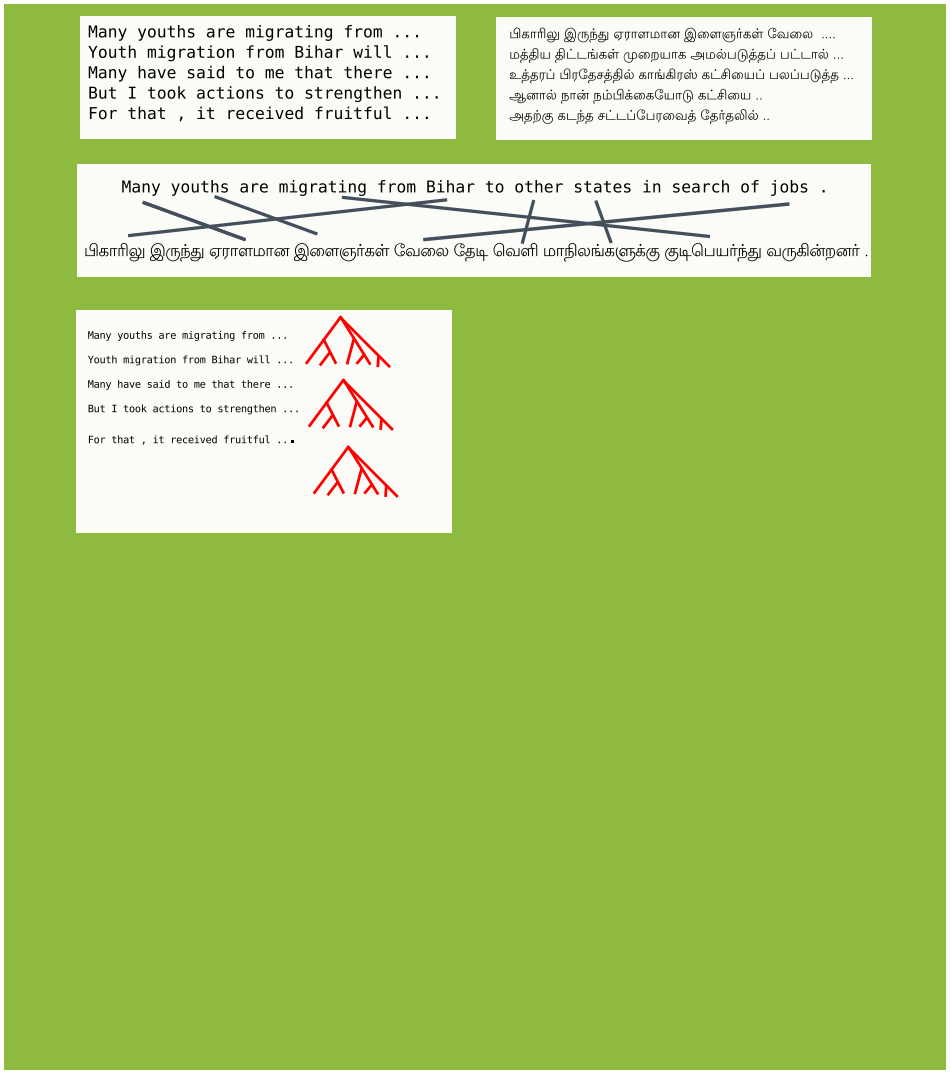
Overview


Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..
Many youths are migrating from Bihar to other states in search of jobs .
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை தேடி வெளி மாநிலங்களுக்கு குடிபெயர்ந்து வருகின்றனர் .

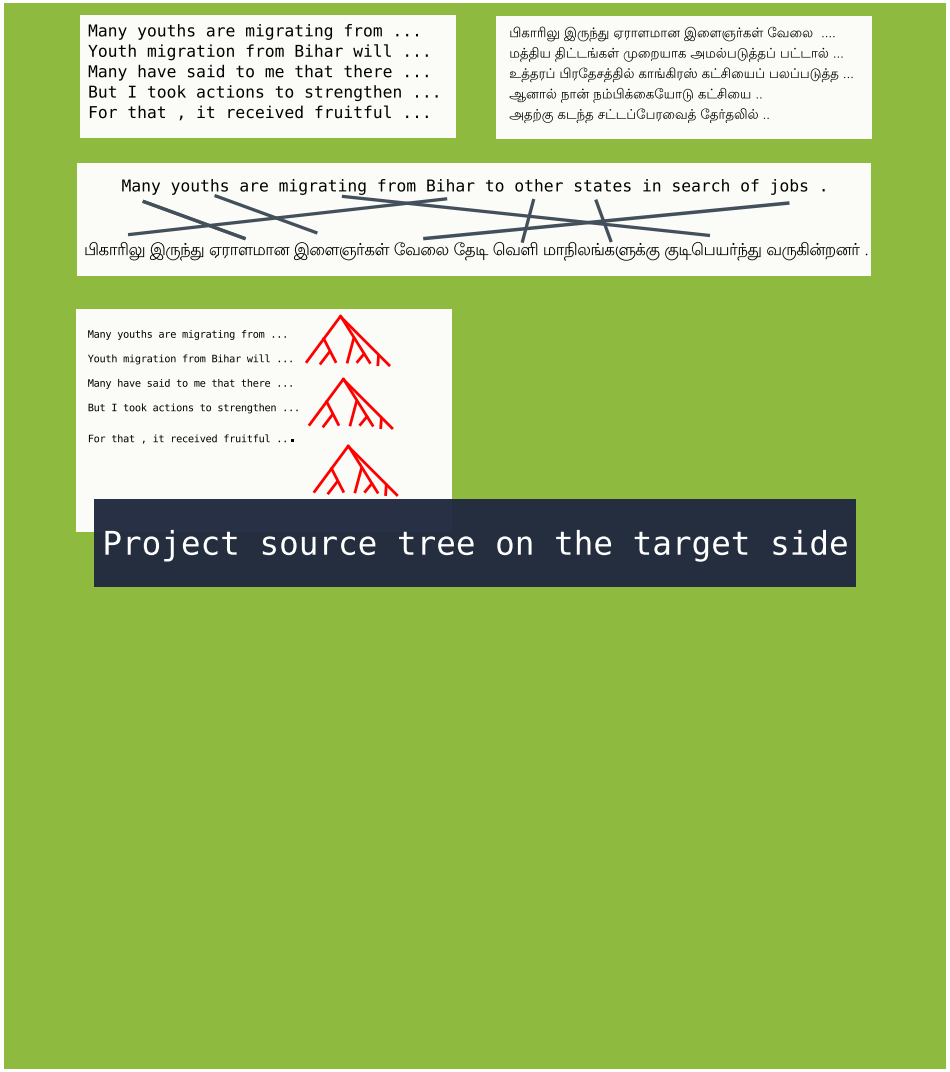
Overview


Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..
Many youths are migrating from Bihar to other states in search of jobs .
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை தேடி வெளி மாநிலங்களுக்கு குடிபெயர்ந்து வருகின்றனர் .
Source parsing

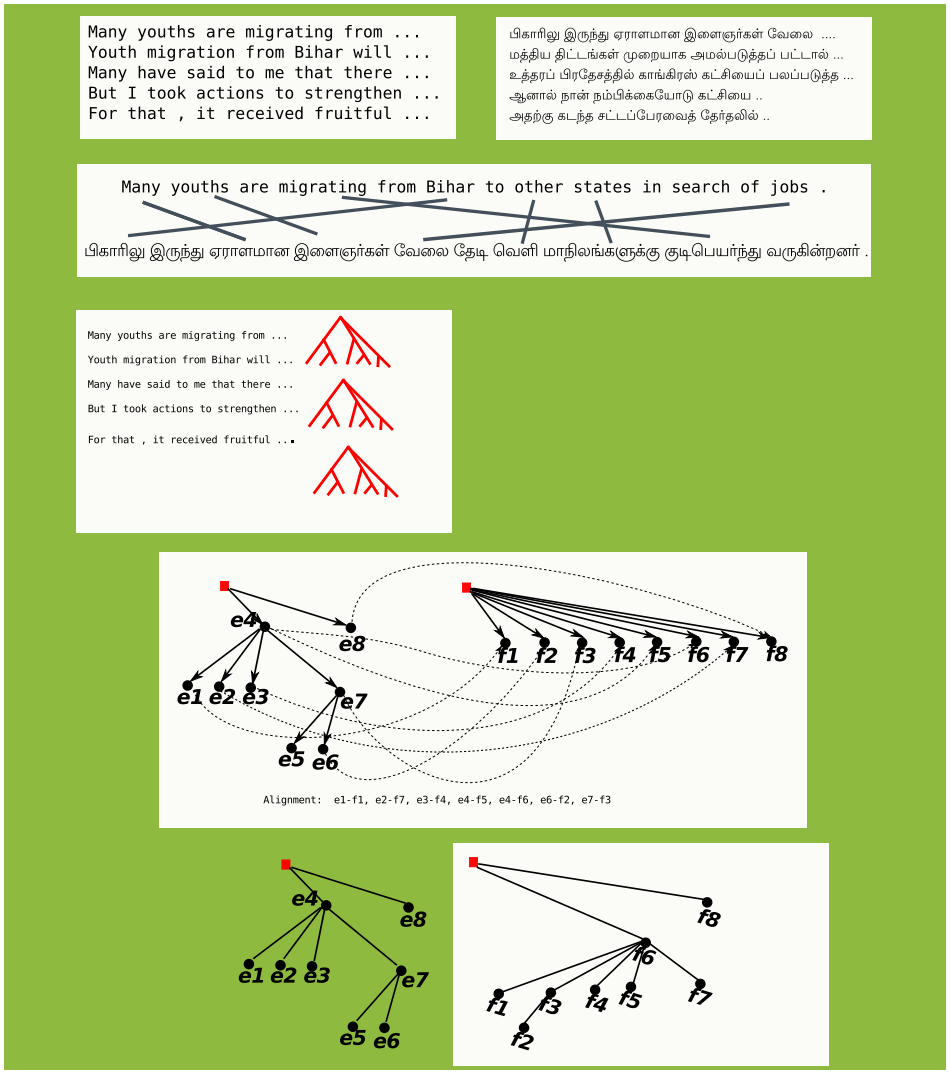
Overview


Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..
Many youths are migrating from Bihar to other states in search of jobs .
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை தேடி வெளி மாநிலங்களுக்கு குடிபெயர்ந்து வருகின்றனர் .
Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...

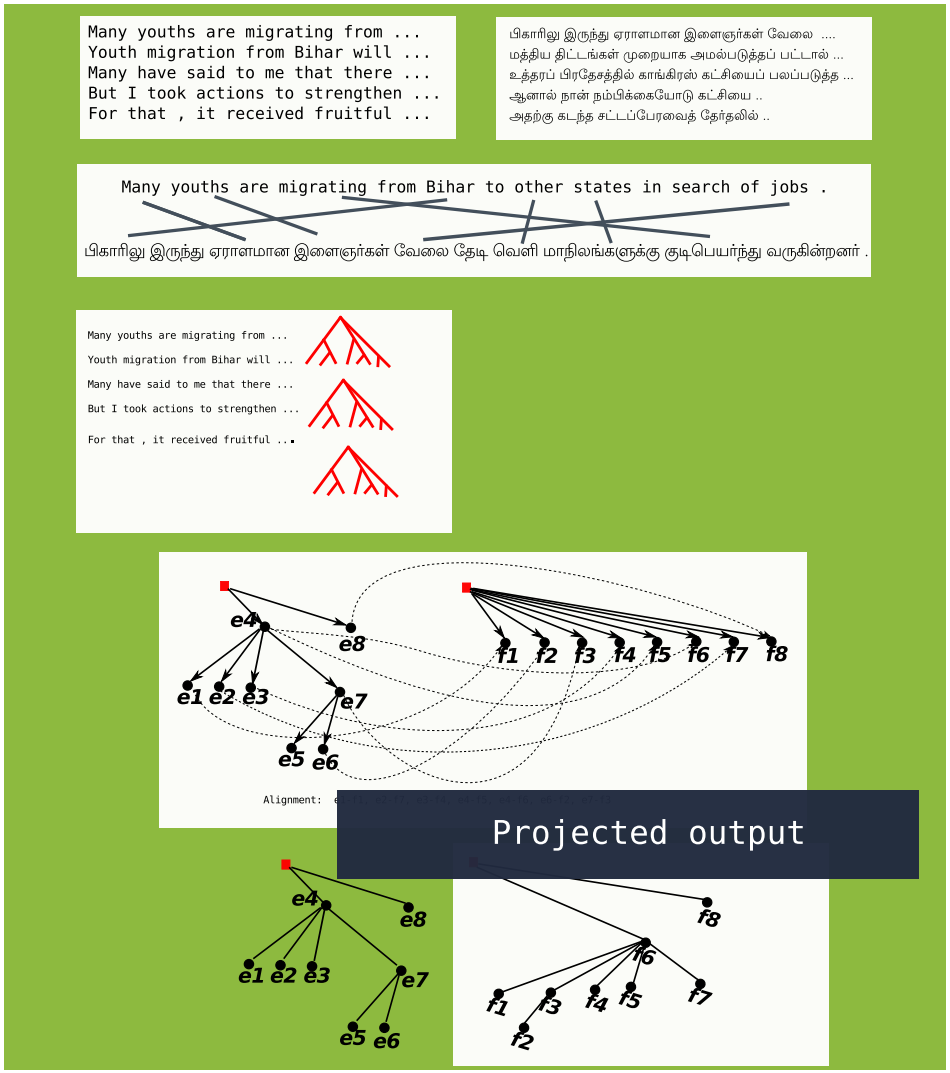
Overview


Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..
Many youths are migrating from Bihar to other states in search of jobs .
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை தேடி வெளி மாநிலங்களுக்கு குடிபெயர்ந்து வருகின்றனர் .
Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
Project source tree on the target side

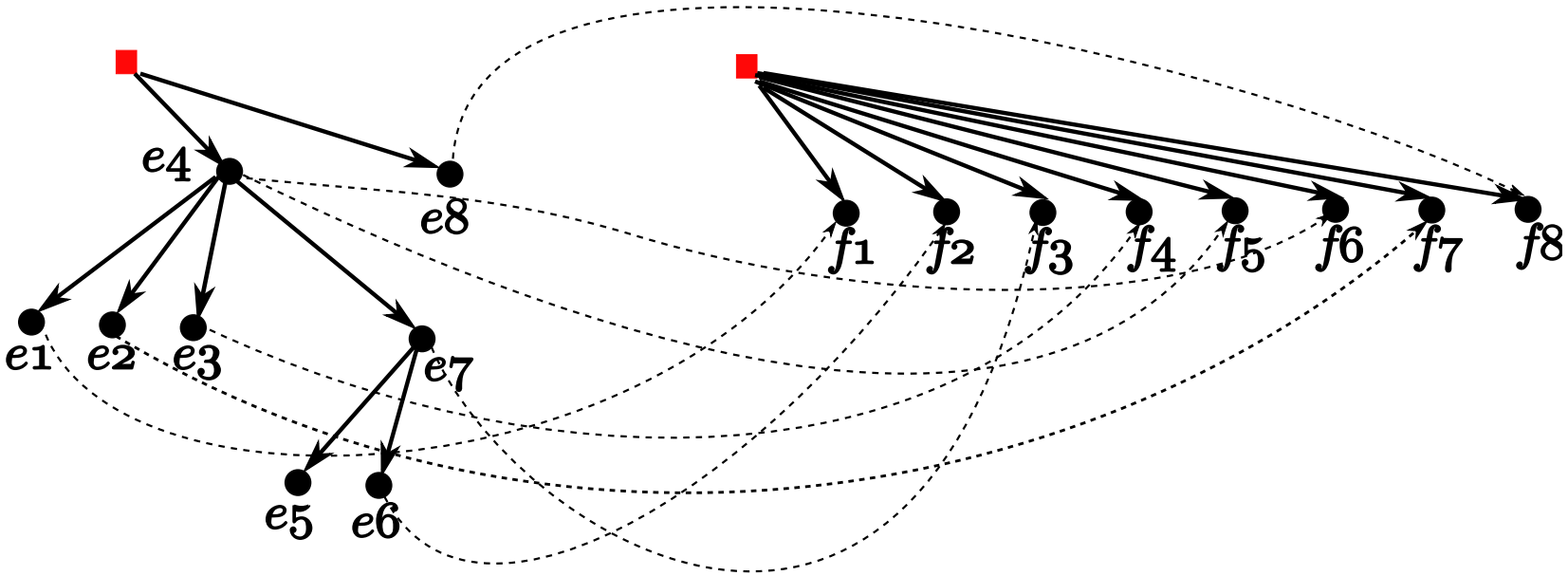
Overview


Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..
Many youths are migrating from Bihar to other states in search of jobs .
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை தேடி வெளி மாநிலங்களுக்கு குடிபெயர்ந்து வருகின்றனர் .
Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
e4
e8
e1 e2 e3
e7
e5 e6
f1 f2 f3 f4 f5 f6 f7 f8
Alignment: e1-f1, e2-f7, e3-f4, e4-f5, e4-f6, e6-f2, e7-f3
e4
e8
e1 e2 e3
e7
e5 e6
f8
f6
f1 f3 f4 f5 f7
f2

Overview


Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..
Many youths are migrating from Bihar to other states in search of jobs .
பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை தேடி வெளி மாநிலங்களுக்கு குடிபெயர்ந்து வருகின்றனர் .
Many youths are migrating from ...
Youth migration from Bihar will ...
Many have said to me that there ...
But I took actions to strengthen ...
For that , it received fruitful ...
e4
e8
e1 e2 e3
e7
e5 e6
f1 f2 f3 f4 f5 f6 f7 f8
Alignment:
Projected output
e4
e8
e1 e2 e3
e7
e5 e6
f8
f6
f1 f3 f4 f5 f7
f2

Algorithm




```
Alignment:  e1-f1, e2-f7, e3-f4, e4-f5, e4-f6, e6-f2, e7-f3
```

Figure: Aligned sentence pair

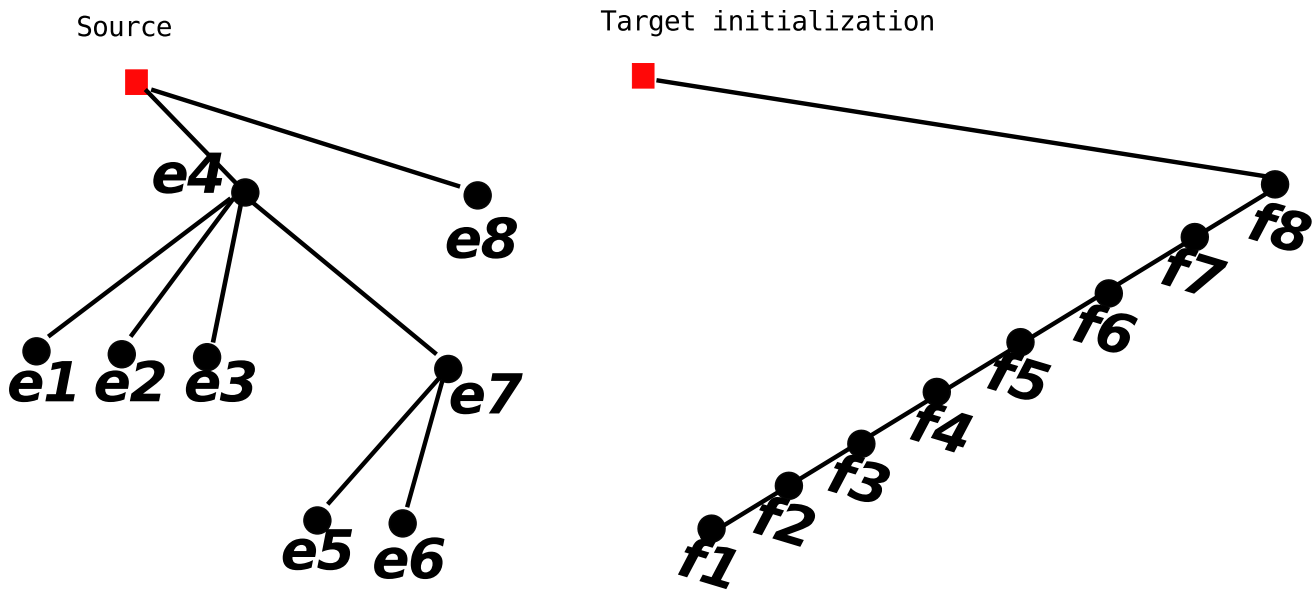


Figure: Projection: Step-by-step

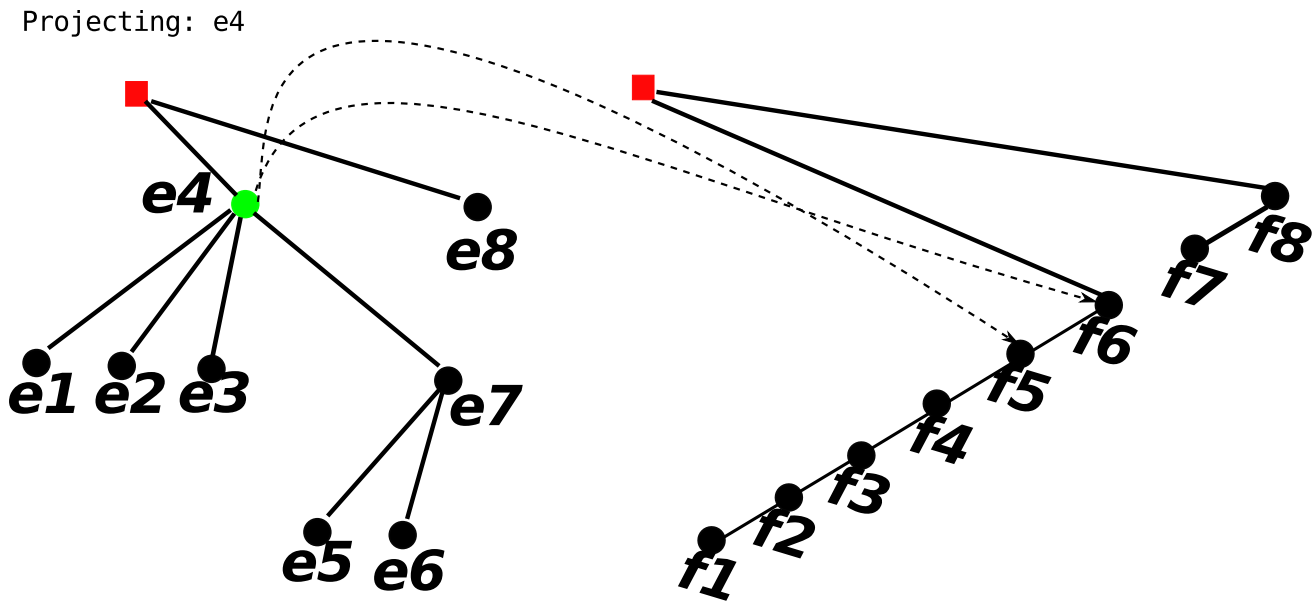


Figure: Projection: Step-by-step

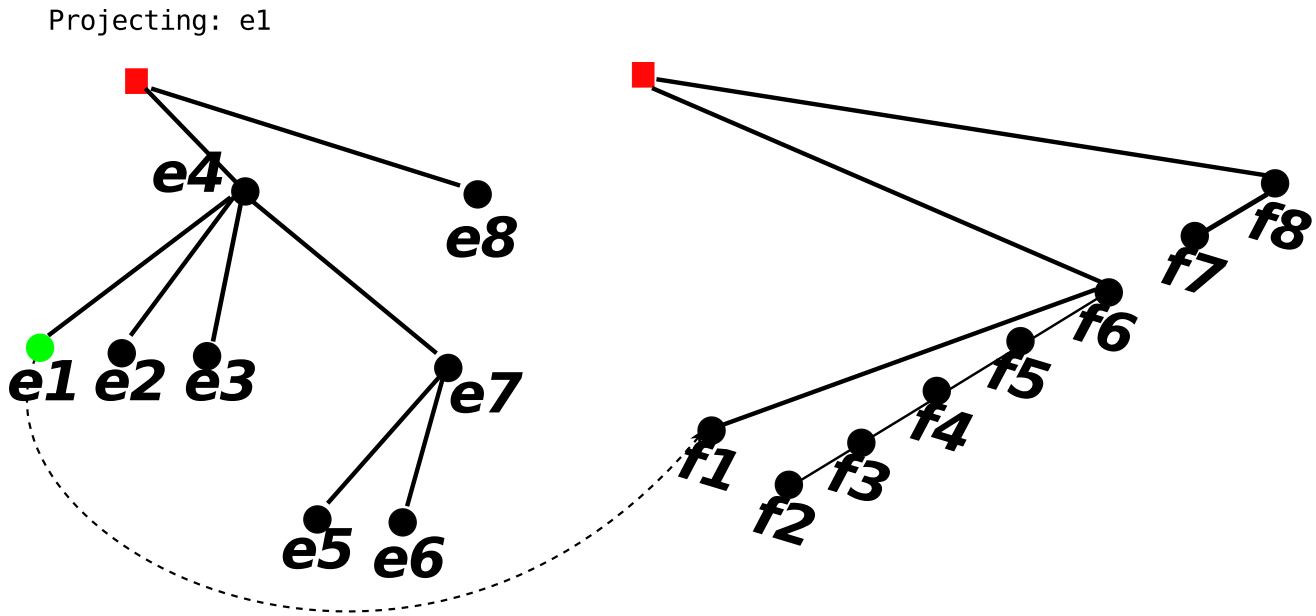


Figure: Projection: Step-by-step

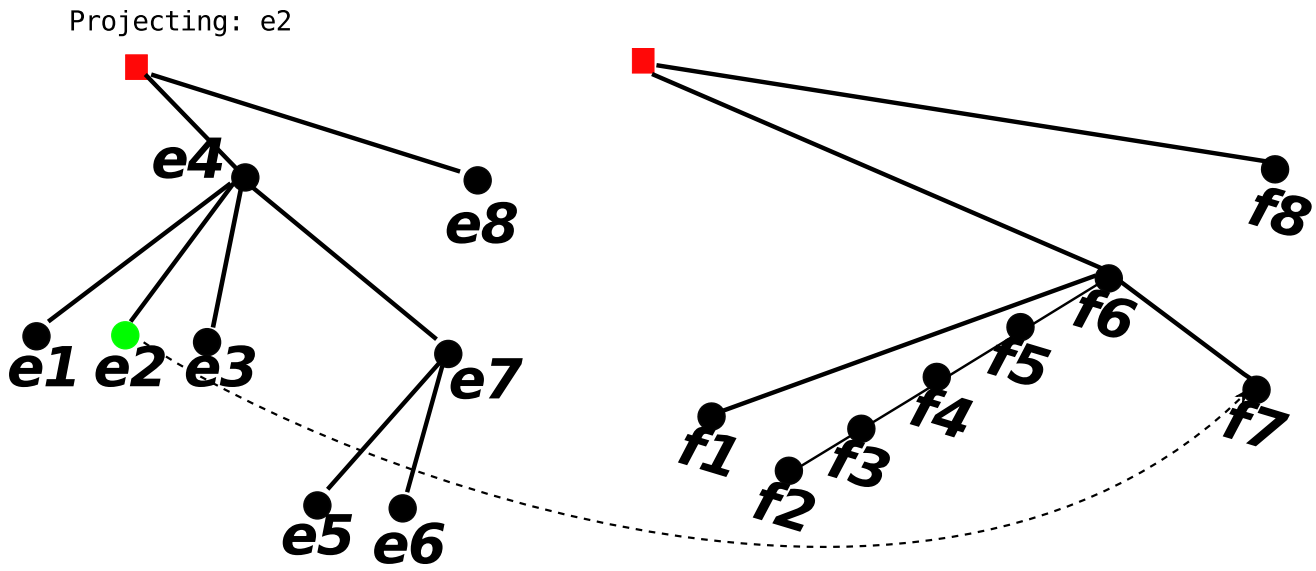


Figure: Projection: Step-by-step

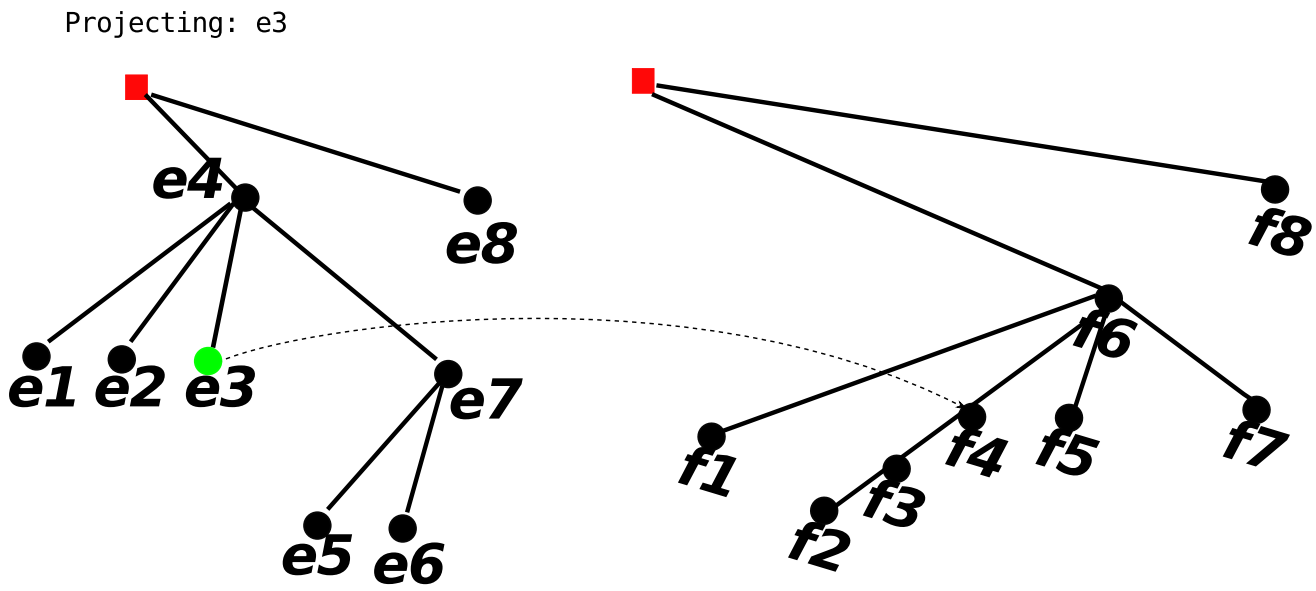


Figure: Projection: Step-by-step

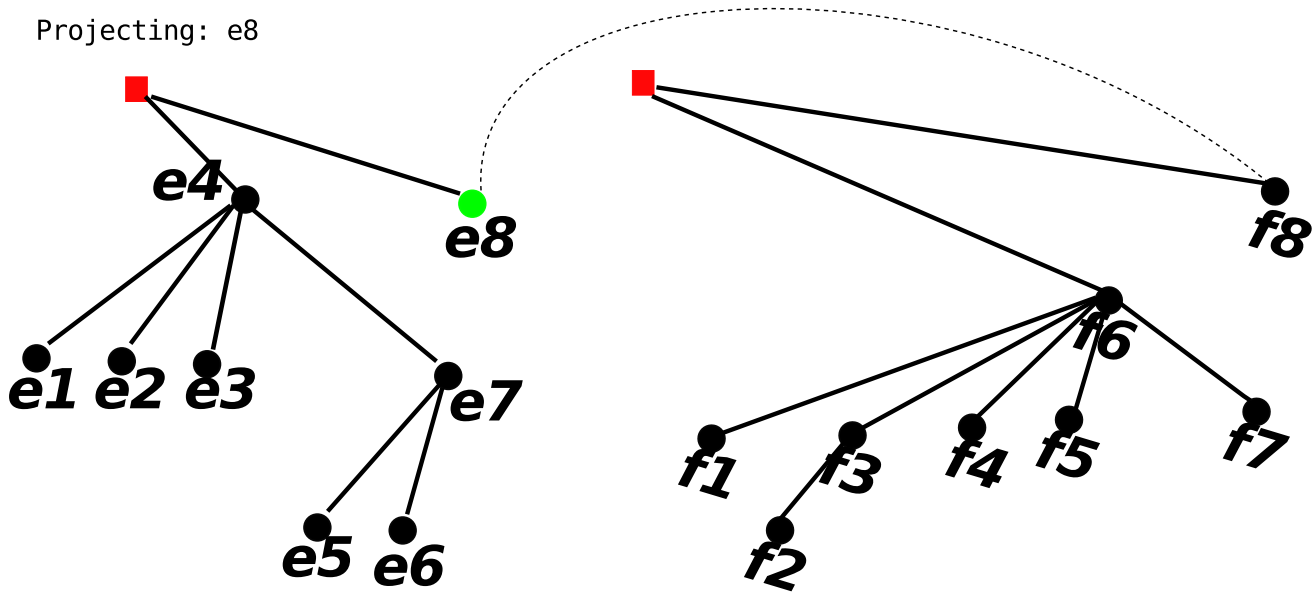


Figure: Projection: Step-by-step

Experimental setup

- **Source** - English
- **Target** - Indian languages
- Source parsing: MST parser (2^{nd} order, projective)
- Alignment
  - Berkeley aligner
  - HMM and HMM syntax
- No transformation rules
- Target tagger is required

Data - Parallel corpus

Tr. dir.	lang. pair	source	corpus size	#tokens		#avg. sen. length	
				src (en)	tgt (IL)	src (en)	tgt (IL)
EN→IL	English-Hindi (en-hi)	[BDR+14]	50.0K	636.4K	12.7	668.8K	13.4
	English-Tamil (en-ta)	[RBv12]	50.0K	1431.0K	28.6	1155.2K	23.1
	English-Urdu (en-ur)	[JZ11]	12.8K	267.6K	20.9	323.9K	25.2
IL→ EN	English-Bengali (en-bn)	[PCBO12]	20.8K	323.7K	15.6	264.9K	12.7
	English-Hindi (en-hi)	[PCBO12]	37.6K	622.9K	16.6	692.5K	18.4
	English-Tamil (en-ta)	[PCBO12]	35.0K	455.9K	13.0	417.9K	11.9
	English-Telugu (en-te)	[PCBO12]	43.0K	583.0K	13.6	480.4K	11.2
	English-Urdu (en-ur)	[PCBO12]	33.8K	530.9K	15.7	658.4K	19.5

Data - Treebanks

lang.	source	# sentences		avg. sen. len.		tagset size		
		train	test	train	test	dep.	CPOSTAG	POSTAG
Bengali (bn)	ICON2010 [HMAG10]	979	150	6.6	5.4	42	14	21
Hindi (hi)	MTPIL (COLING2012) [HMAG10]	12041	1233	22.3	21.4	119	36	20
Tamil (ta)	TamilTB 1.0 [Rv11]	480	120	15.1	15.8	25	12	465
Telugu (te)	ICON2010 [HMAG10]	1300	150	3.9	4.0	41	16	24
Urdu (ur)	UDT [BS12]	2808	313	13.1	12.4	71	15	29

Results - Baseline

Lang.	Left	Right	Supervised parser	
			pred. POS	gold POS
bn	53.6	04.6	70.5/72.2	76.2/78.3
hi	24.4	27.3	80.3/80.2	85.5/85.4
ta	50.9	09.5	59.8/60.7	76.4/76.8
te	65.8	02.4	83.1/83.1	87.5/87.0
ur	42.9	06.3	63.4/66.3	66.7/68.2

Results - Baseline

Lang.	Left	Right	Supervised parser	
			pred. POS	gold POS
bn	53.6	04.6	70.5/72.2	76.2/78.3
hi	24.4	27.3	80.3/80.2	85.5/85.4
ta	50.9	09.5	59.8/60.7	76.4/76.8
te	65.8	02.4	83.1/83.1	87.5/87.0
ur	42.9	06.3	63.4/66.3	66.7/68.2

- **IL treebanks are mostly left-branching**

Results - Projection

Corpus type	Lang. pair	HMM (syntax)		HMM	
		reparse	reparse_{10}	reparse	reparse_{10}
EN-IL	en-hi	25.3	28.4	26.9	32.2
	en-ta	**54.2**	52.9	**52.9**	52.0
	en-ur	42.9	49.0	42.9	48.4
IL-EN	en-bn	52.3	52.8	52.2	52.8
	en-hi	26.1	31.8	26.7	34.6
	en-ta	49.8	47.7	46.8	45.7
	en-te	**70.3**	71.6	**66.8**	68.0
	en-ur	**43.7**	49.2	**43.2**	48.4

Limitations of bitext projection

- Transformation rules for each source and target for **annotation compatibility**
- Availability of parallel corpus

Table of Contents

Overview

- Method that parses target language using a **source parser**
- Introduced by [ZR08]
- Main focus: Addressing annotation difference

How delexicalization works

Source

Many/JJ youths/NNS are/VBP migrating/VBG from/IN .../:

Youth/NNP migration/NNP from/IN Bihar/NNP will/MD .../:

Many/DT have/VBP said/VBN to/TO me/PRP that/RB there/RB .../:

But/CC I/PRP took/VBD actions/NNS to/TO strengthen/VB .../:

For/IN that/DT ,/, it/PRP received/VBD fruitful/JJ .../.

Target

பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..

How delexicalization works

Source

Many/JJ youths/NNS are/VBP migrating/VBG from/IN .../:
Youth/NNP migration/NNP from/IN Bihar/NNP will/MD .../:
Many/DT have/VBP said/VBN to/TO me/PRP that/RB there/RB .../:
But/CC I/PRP took/VBD actions/NNS to/TO strengthen/VB .../:

POS Harmonization (+ dependency)

Target

பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..

How delexicalization works

Source

Many/JJ youths/NNS are/VBP migrating/VBG from/IN .../:

Youth/NNP migration/NNP from/IN Bihar/NNP will/MD .../:

Many/DT have/VBP said/VBN to/TO me/PRP that/RB there/RB .../:

But/CC I/PRP took/VBD actions/NNS to/TO strengthen/VB .../:

POS Harmonization (+ dependency)

Many/A youths/N are/V migrating/V from/R .../:

Youth/N migration/N from/R Bihar/N will/V .../:

Many/A have/V said/V to/T me/P that/D there/D .../:

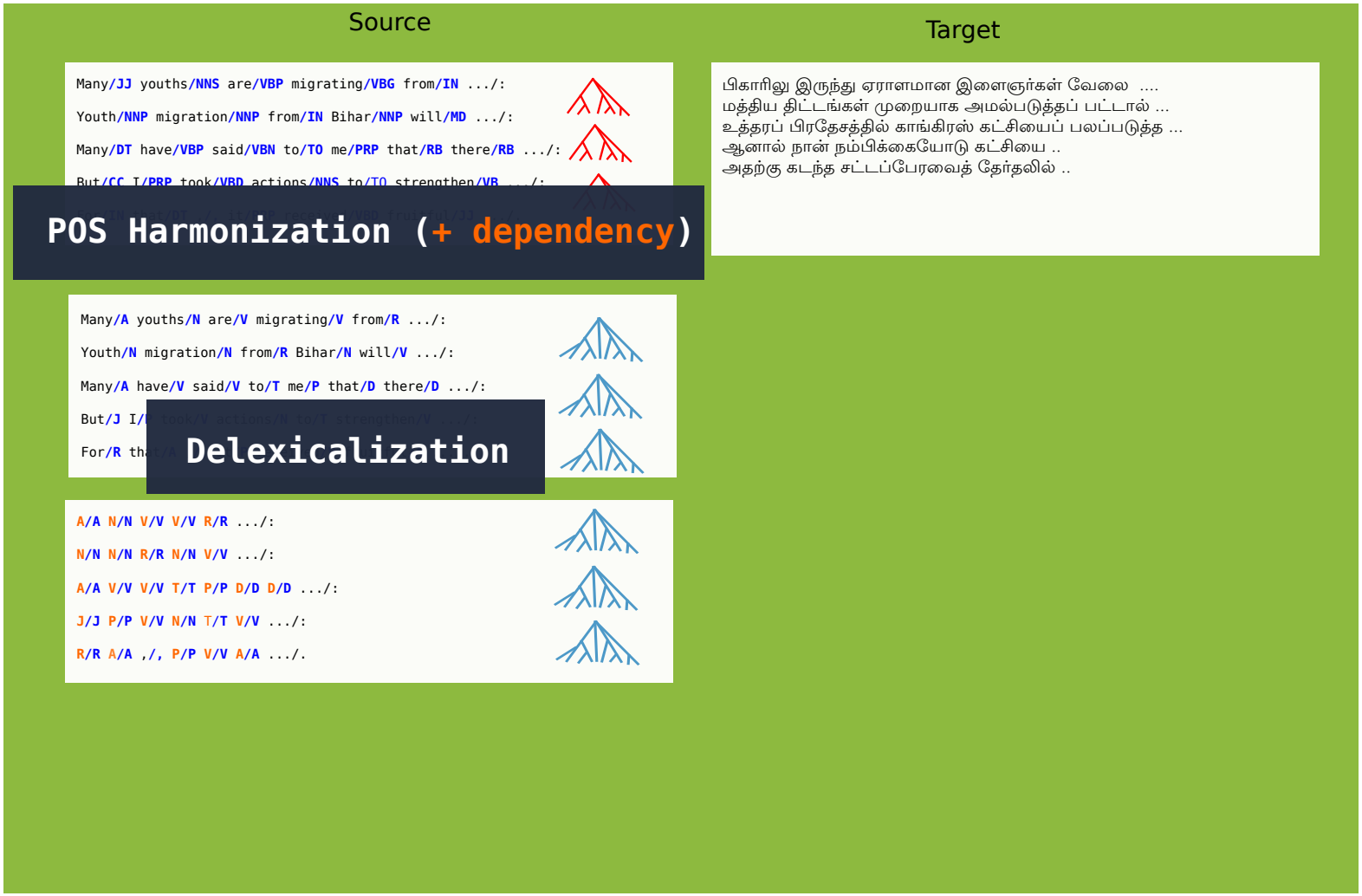
But/J I/P took/V actions/N to/T strengthen/V .../:

For/R that/A ,/, it/P received/V fruitful/A .../.

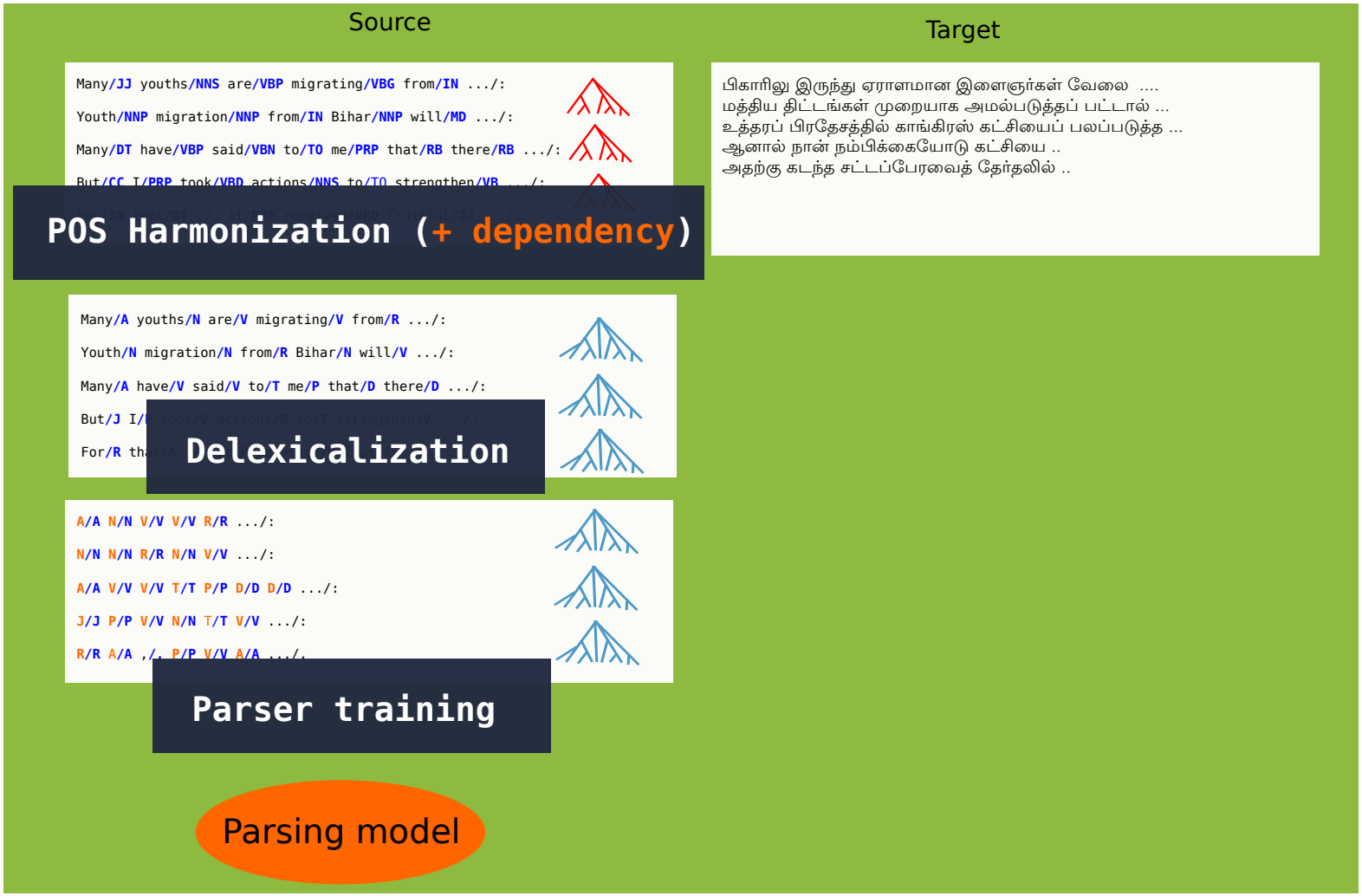
Target

பிகாரிலு இருந்து ஏராளமான இளைஞர்கள் வேலை ....
மத்திய திட்டங்கள் முறையாக அமல்படுத்தப் பட்டால் ...
உத்தரப் பிரதேசத்தில் காங்கிரஸ் கட்சியைப் பலப்படுத்த ...
ஆனால் நான் நம்பிக்கையோடு கட்சியை ..
அதற்கு கடந்த சட்டப்பேரவைத் தேர்தலில் ..

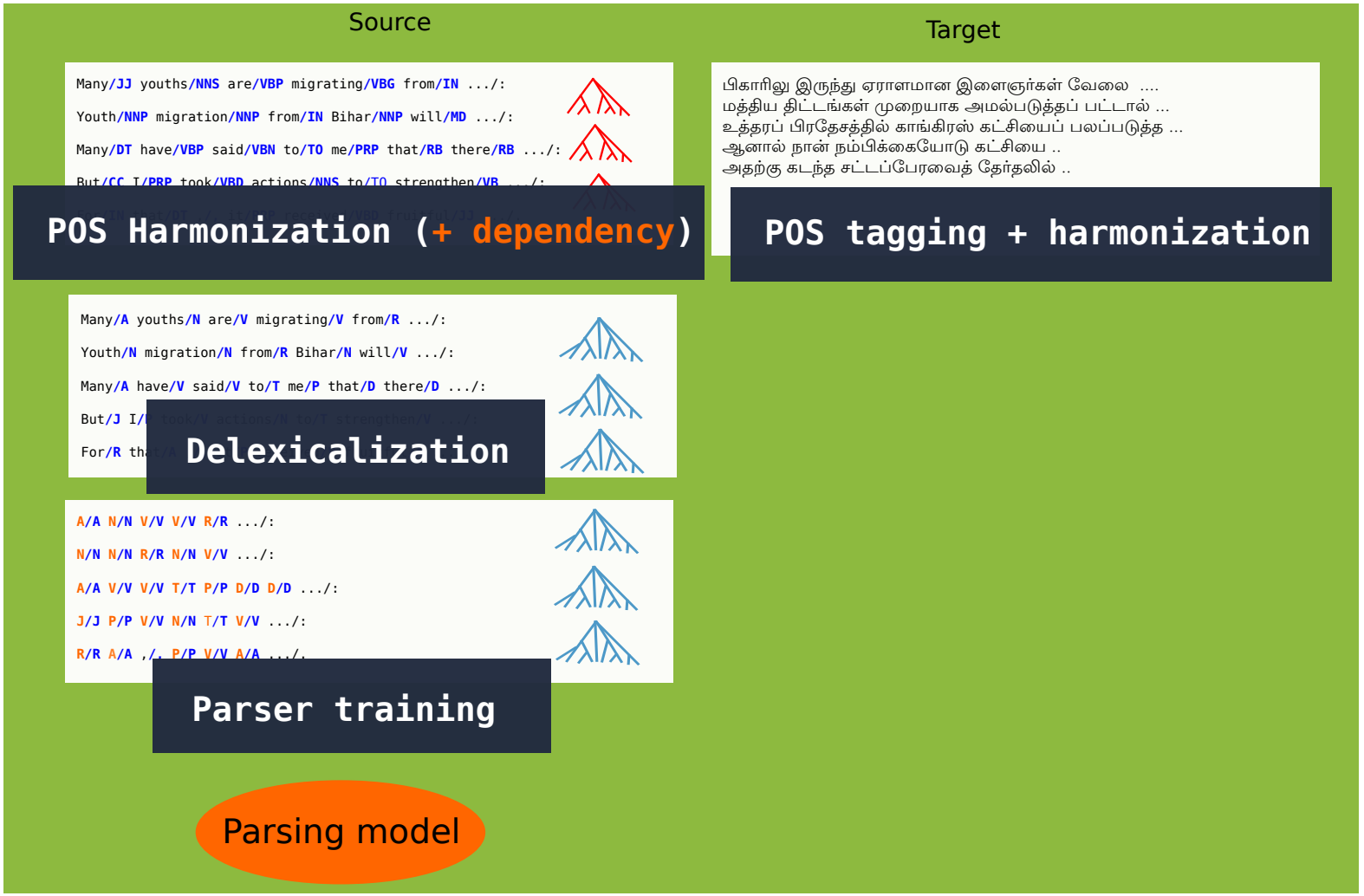
How delexicalization works



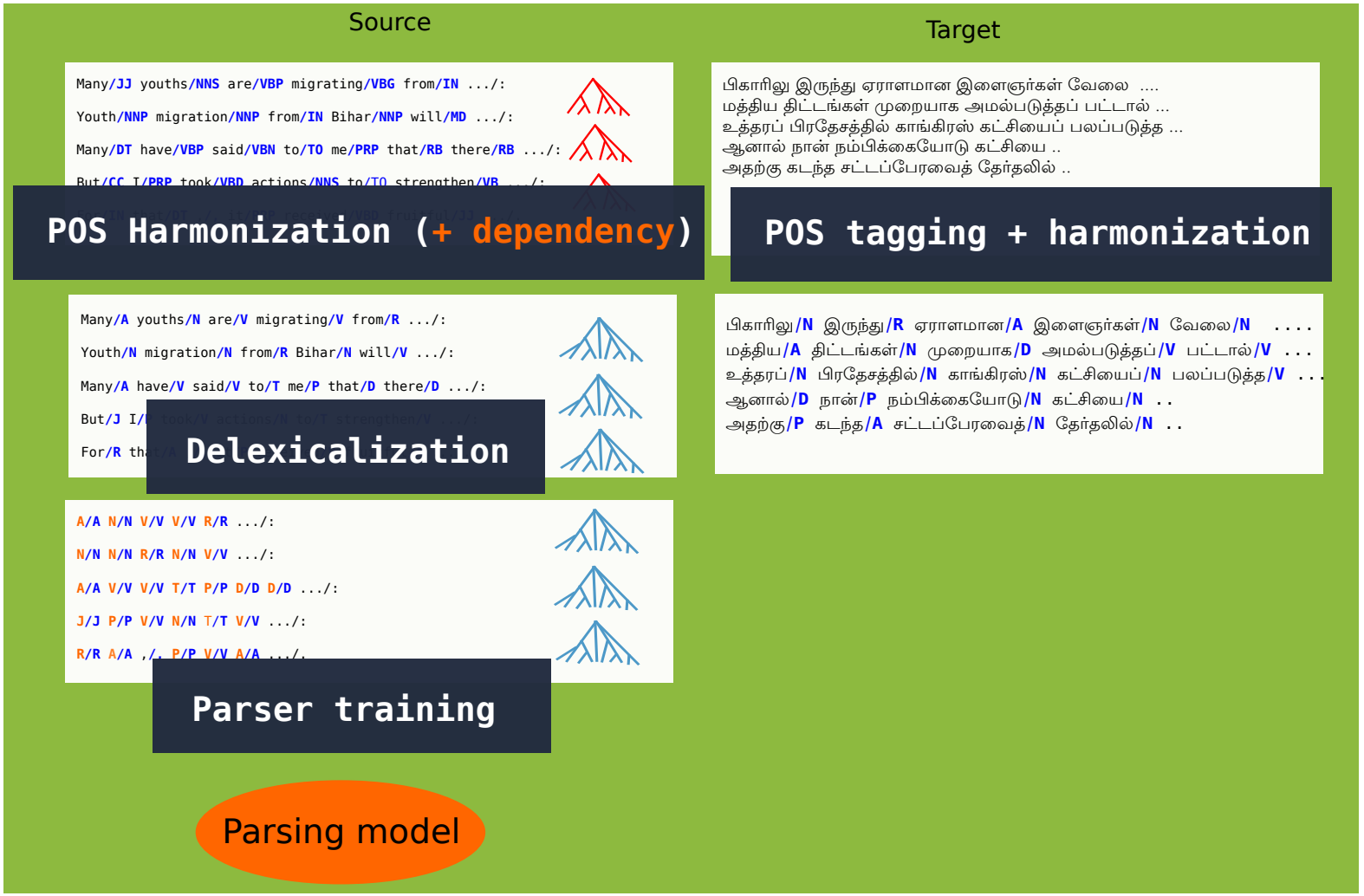
How delexicalization works



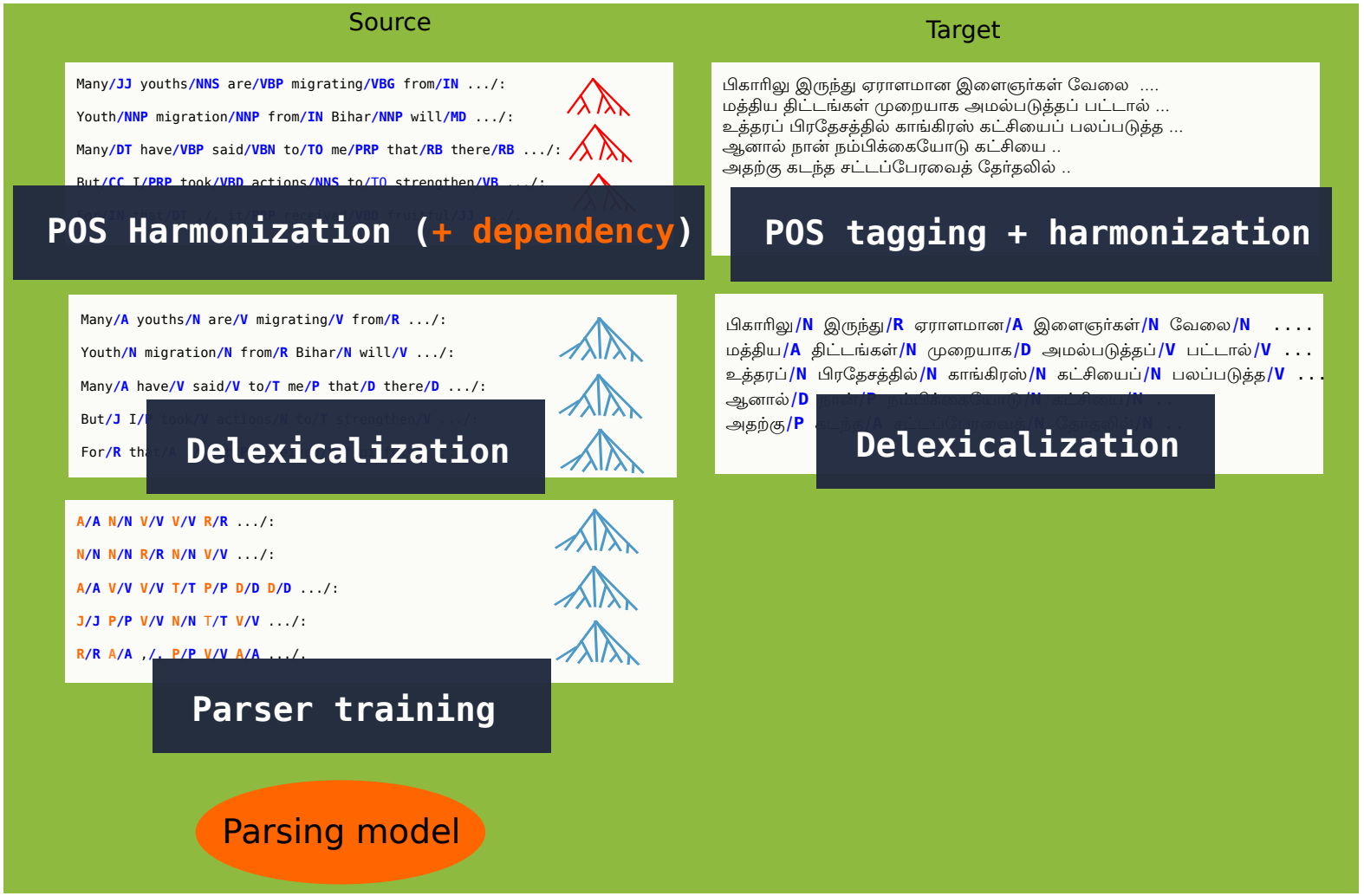
How delexicalization works



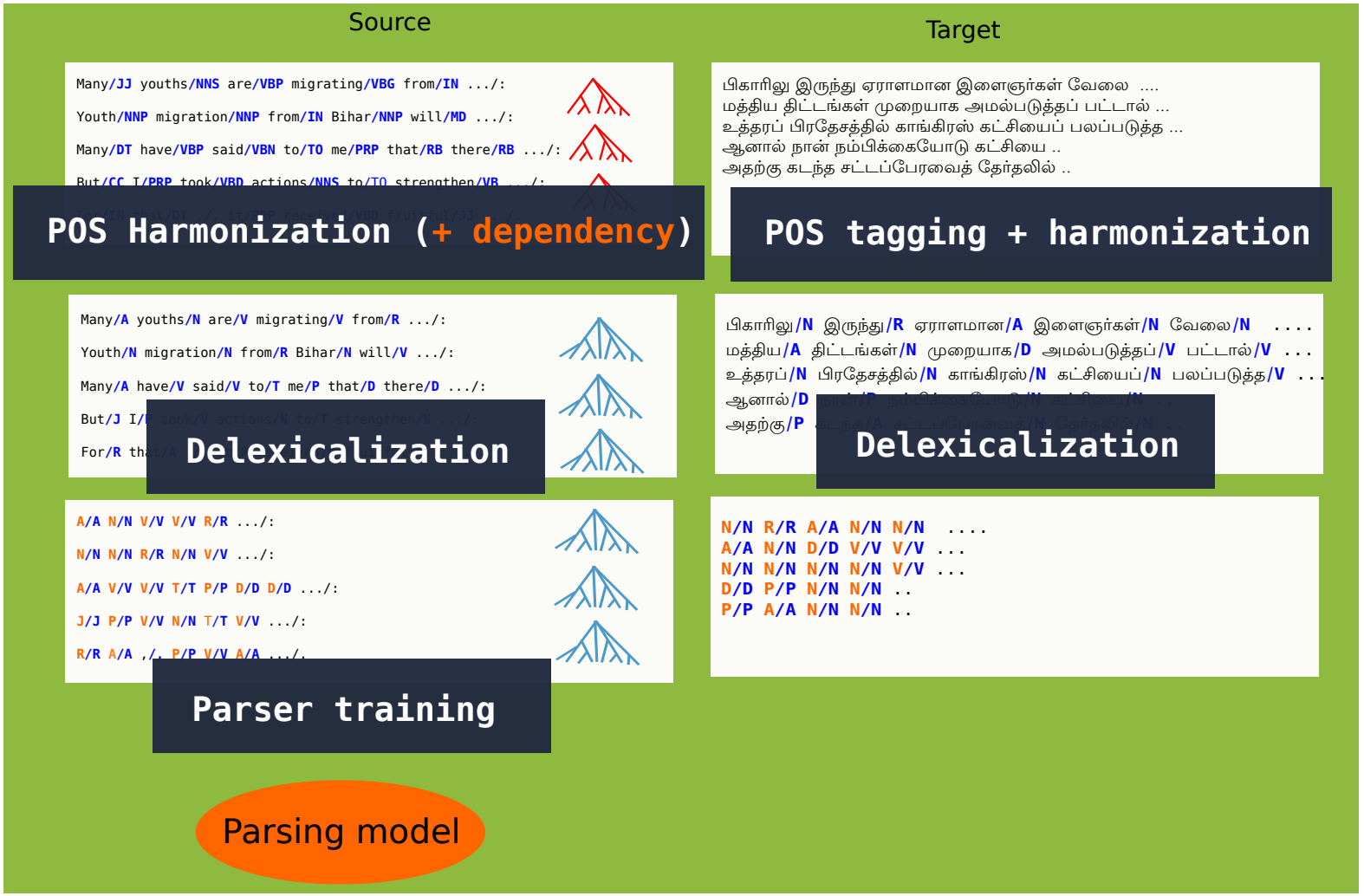
How delexicalization works



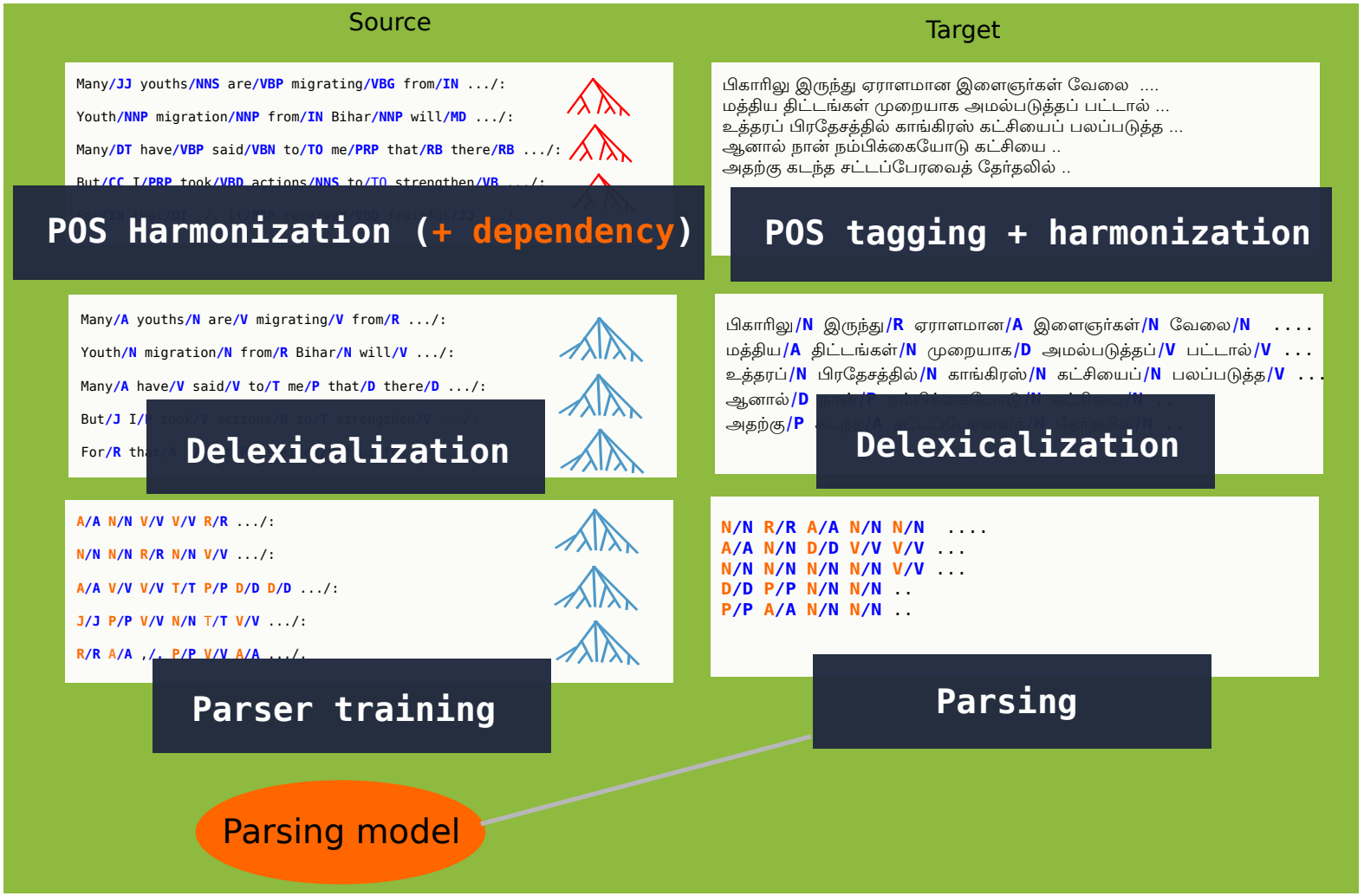
How delexicalization works



How delexicalization works



How delexicalization works



Annotation differences

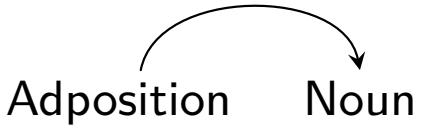
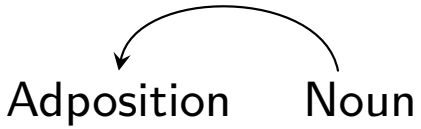
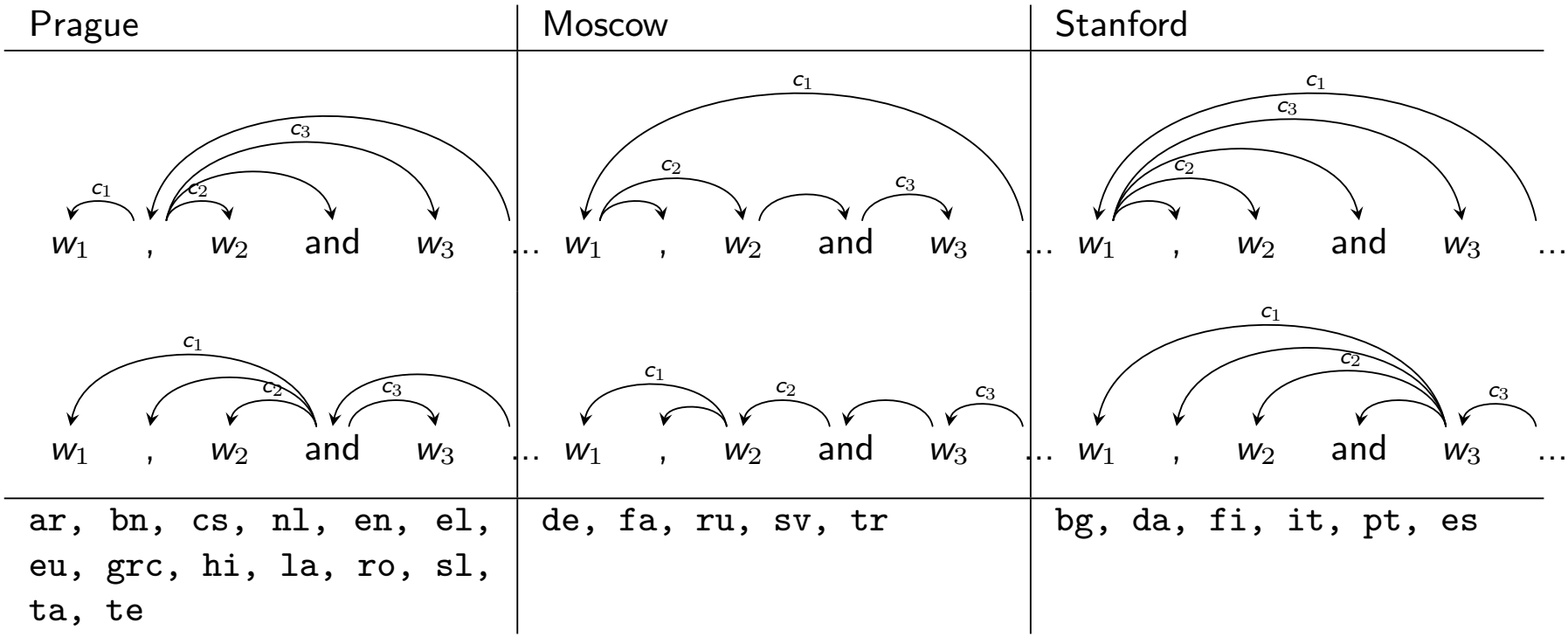
Pattern	Treebanks	Structure
adposition-noun	ar, bg, ca, cs, da, de, el, en, es, et, eu, fa, grc, he, hu, is, it, ja, la, nl, pl, pt, ru, sl, sk, sv, ta, tr, zh	
	fi, hi	

Table: Annotation differences in treebanks

Annotation differences



Prague	Moscow	Stanford
ar, bn, cs, nl, en, el, eu, grc, hi, la, ro, sl, ta, te	de, fa, ru, sv, tr	bg, da, fi, it, pt, es

Table: Annotation differences in treebanks (coordination structures - [PMv+13])

Annotation differences - Solutions

- We need multilingual approach
- The way forward : Common annotation style
  - MULTEXT-East [Erj10]: a multilingual standardized dataset for 16 central and eastern European languages
  - Interset [Zem08]and HamleDT [ZMP+12]
  - Universal POS tagset [PDM12] and Universal dependency treebank [MNQB+13]
- HamleDT
  - POS harmonization
  - Dependency harmonization
  - Harmonized version available for 30 treebanks
  - Includes harmonization for **4 Indian languages (`bn`, `hi`, `ta`, `te`)**

POS harmonization

Treebank	Orig. size	Interset fine	Interset coarse
bn	21	29	12
hi	36	344	12
ta	219	79	10
te	23	58	12
ur	29	10	10

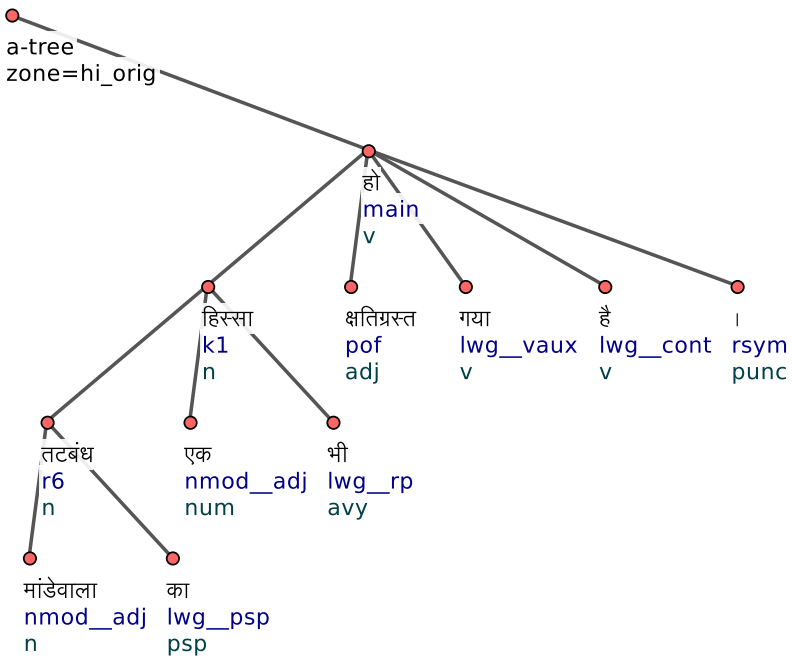
Table: Tagset size: original vs. harmonized

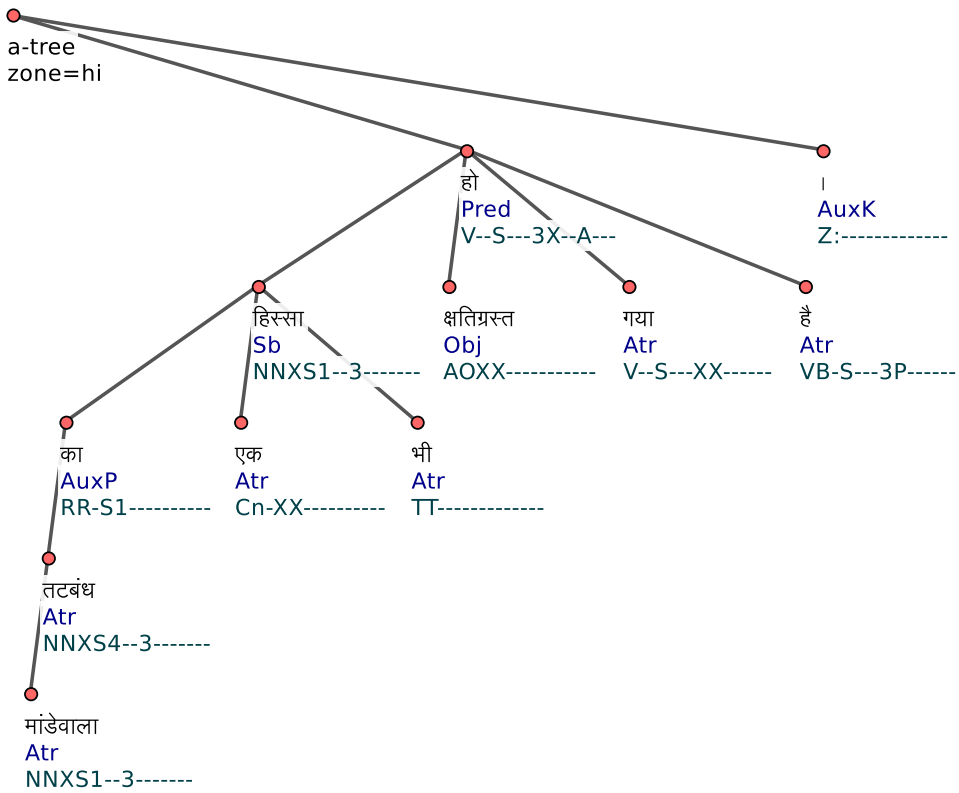
Dependency harmonization

Lang.	No punc	With punc
bn	99.9	99.1
hi	58.0	56.3
ta	100.0	99.8
te	100.0	99.4
ur	-	-

Table: UAS scores between original and harmonized treebanks

Harmonization example





[hi_orig] मांडेवाला तटबंध का एक हिस्सा भी क्षतिग्रस्त हो गया है ।
[hi] मांडेवाला तटबंध का एक हिस्सा भी क्षतिग्रस्त हो गया है ।

Figure: Hindi dependency tree: original vs. harmonized

Experiments - Setup

- **Source:** HamleDT trained delexicalized parsers
- **Target:** ILs
- POS harmonization
- Dependency harmonization

Data - HamleDT

- Key resource
- Delexicalized parsing experiments
  - HamleDT 2.0
  - POS harmonized
  - Dependency harmonized

Results - Baseline

Lang.	POS harmonized				POS+dep harmonized			
	left	right	pred.	gold	left	right	pred.	gold
bn	53.6	04.6	72.1	77.7	53.6	04.6	73.0	77.8
hi	24.4	27.3	76.0	78.5	53.3	07.7	75.8	78.4
ta	50.9	09.5	57.6	67.2	50.9	09.5	58.7	68.6
te	65.8	02.4	82.6	86.2	65.8	02.4	83.1	86.2
ur	42.9	06.3	-	-	-	-	-	-

Table: Supervised parser results: POS harmonized vs. POS+dep harmonized

Results - IL→IL delexicalized transfer

Lang.	← bn →		← hi →		← ta →		← te →		← ur →	
	D_P	D_{PD}	D_P	D_{PD}	D_P	D_{PD}	D_P	D_{PD}	D_P	D_{PD}
bn	70.8	71.6	27.8	33.1	34.8	36.1	78.6	78.2	58.6	-
hi	57.7	55.1	79.6	80.2	41.6	46.4	67.6	68.3	48.7	-
ta	57.3	55.9	34.2	61.7	58.4	58.5	69.1	69.6	42.6	-
te	63.9	62.4	21.9	24.1	22.5	26.1	82.6	82.6	53.9	-
avg	**59.6**	**57.8**	**28.0**	39.6	33	36.2	**71.8**	**72.0**	**51.0**	-

Table: Delexicalized parser results in the case of ILs as source. D_P - POS harmonization; D_{PD} - POS+dependency harmonization;

- crosses baseline in **4 out of 5** cases in POS harmonization
- crosses baseline in **2 out of 2** cases in POS+dependency harmonization

Table of Contents

Transfer with machine translated texts

- Can we perform projection **without a parallel corpus ?**
- Use translation systems to obtain parallel corpus

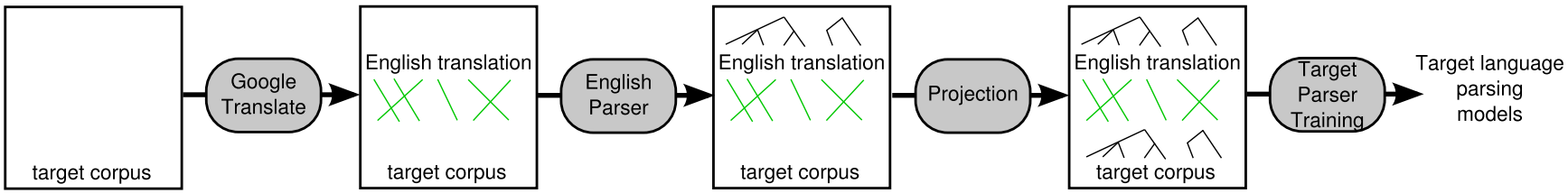


Figure: Schematic depiction of how the target parse trees are obtained (Credit: David Mareček)

- Joint work with David Marecek and Zdenek Zabokrtsky

Experimental setup

- Use artifically created parallel corpus for projection
- **Source:** English
- **Target:** Indo-European languages
- Comparison with Unsupervised dependency parsing
- Projection results for ILs

Data - HamleDT

- 18 treebanks translated to English
- Translations obtained over a period through Google Translate API
- Thanks to Rudolf

Results - Projection

Lang.	Baseline		UDP	Projection (+ reparsing)					sup
	left	right		unsup40	dir proj	univ	sup	gold	
ar	5.2	**58.8**	34.1	51.8	40.3	56.1	58.3	60.0	74.2
bg	17.9	38.8	56.7	46.4	41.4	**55.3**	53.6	56.3	81.5
ca	24.7	28.8	24.6	56.9	46.0	-	**59.0**	60.2	87.6
cs	24.1	28.9	55.0	45.4	51.4	54.7	**58.4**	62.3	74.9
da	13.2	**47.9**	42.0	41.6	36.9	41.6	42.1	43.1	79.8
de	24.2	18.9	43.5	46.4	43.0	47.8	**48.7**	50.1	84.2
el	32.0	18.5	30.9	49.2	52.0	59.2	**65.4**	65.2	78.5
es	24.7	29.0	36.3	56.9	47.0	-	**58.2**	59.0	88.1
et	34.1	17.4	**63.8**	54.8	58.0	-	58.3	66.0	72.9
fi	39.3	13.6	36.7	37.1	43.1	-	**39.5**	46.2	60.7
hi	24.4	27.3	15.6	24.9	24.6	-	**28.0**	28.3	75.1
hu	42.8	5.3	34.7	43.4	41.1	48.2	**51.2**	53.2	75.2
it	23.0	37.4	49.7	55.6	53.1	53.7	**59.3**	62.0	80.5
nl	27.9	24.7	27.6	**60.9**	53.2	-	59.3	61.4	76.7
pt	25.8	31.1	39.8	61.8	56.4	62.2	**62.9**	66.5	84.2
sl	24.4	26.6	45.9	44.4	50.4	45.7	**53.1**	56.6	76.7
sv	25.9	27.8	49.1	53.9	48.2	**56.8**	54.3	55.8	82.7
tr	**65.1**	2.0	42.3	46.0	51.4	-	57.2	57.0	75.5
avg	27.7	26.8	40.5	48.7	46.5	*52.8*	**53.7**	56.1	78.3

Results - Projection (ILs)

	Predicted POS		Gold POS	
Lang.	mt	human	mt	human
bn	47.8	51.8	49.2	50.8
hi	5.4	25.2	5.4	25.3
ta	42.9	45.9	41.4	42.9
te	58.2	65.9	60.5	68.1
ur	41.8	43.6	41.9	43.4
avg	39.2	46.5	39.7	46.1

Table: Projection results ILs: machine translated vs. human translated; Parallel corpus size: 2K.

Table of Contents

How to choose methodologies I

1. If there's a treebank available for a target language, then train a **supervised parser**.
2. If there's no treebank available for a target language but only a parallel corpus,
   - If the source language has a parser (implicitly POS tagger too), then try **projection** or **delexicalized parsing** or **both**.
3. If there's no treebank or parallel corpus available for a target language but only an MT system, then **create parallel corpus by translating target texts to a resource-rich source language** and apply step (2).
4. If there's no treebank, parallel corpus or MT system available for a target language but only a POS tagger is available, then try **delexicalized parsing**. Choosing source languages as closer to target as possible can enhance the accuracy.

How to choose methodologies II

5. If for any of the above steps the target POS tagger is not available, then **obtain target POS tags using unsupervised approach**.
6. If none of the above steps is viable, then use **unsupervised parsing** to obtain target structures.

Table of Contents

Summary

- Bitext projection
  - Outperforms left/right baseline in **3 out of 8** parallel datasets
  - Hardly any difference between professionally translated vs. translation through crowdsourcing
- Delexicalized parsing
  - Used HamleDT treebank parsers (30) as source and parsed IL texts
  - **Maximum benefit for IL-IL** delexicalized parsing
- Dependency transfer through machine translated bitexts
  - A new under-resourced scenario
  - Performs better than unsupervised dependency parsing in the case of Indo-European languages in HamleDT

Bibliography I


Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith.
The TIGER treebank.
In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 2002.


Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Aleš Tamchyna, and Dan Zeman.
Hindi-English and Hindi-only Corpus for Machine Translation.
In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC'14)*, Reykjavik, Iceland, May 2014. ELRA, European Language Resources Association.
in prep.

Bibliography II


Igor Boguslavsky, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Kreidlin, and Nadezhda Frid.
Dependency treebank for Russian: Concept, tools, types of information.
In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 987–991. Association for Computational Linguistics Morristown, NJ, USA, 2000.


Riyaz Ahmad Bhat and Dr. Dipti Misra Sharma.
Dependency treebank of urdu and its evaluation.
In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 157–165, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics.

Bibliography III

Tomaž Erjavec.
Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora.
In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

Bibliography IV


Samar Husain, Prashanth Mannem, Bharat Ambati, and Phani Gadde.
The ICON-2010 tools contest on Indian language dependency parsing.
In *Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing*, Kharagpur, India, 2010.


Rebecca Hwa, Philip Resnik, and Amy Weinberg.
Breaking the Resource Bottleneck for Multilingual Parsing.
Technical Report LAMP-TR-086,CS-TR-4355,UMIACS-TR-2002-35, University of Maryland, College Park, April 2002.

Bibliography V


Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak.
Bootstrapping parsers via syntactic projection across parallel texts.
Nat. Lang. Eng., 11(3):311–325, September 2005.


Bushra Jawaid and Daniel Zeman.
Word-order issues in English-to-Urdu statistical machine translation.
2011.


Yasuhiro Kawata and Julia Bartels.
Stylebook for the Japanese treebank in Verbmobil.
In *Report 240*, Tübingen, Germany, September 29 2000.

Bibliography VI


Steven Krauwer.
The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap.
In *Proceedings of the 2003 International Workshop Speech and Computer SPECOM-2003*, pages 8–15, 2003.


Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee.
Universal dependency annotation for multilingual parsing.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Bibliography VII


Matt Post, Chris Callison-Burch, and Miles Osborne.
Constructing parallel corpora for six indian languages via crowdsourcing.
In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada, June 2012. Association for Computational Linguistics.


Slav Petrov, Dipanjan Das, and Ryan McDonald.
A universal part-of-speech tagset.
In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

Bibliography VIII


Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zděněk Žabokrtský.
Coordination structures in dependency treebanks.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.


Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský.

Morphological processing
for english-tamil statistical machine translation.
In *Proceedings of the Workshop on MT and Parsing in Indian Languages*, December 2012.

Bibliography IX


Loganathan Ramasamy and Zdeněk Žabokrtský.
Tamil Dependency Parsing: Results Using Rule Based and Corpus Based Approaches.
In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, CICLing'11, pages 82–95, Berlin, Heidelberg, 2011. Springer-Verlag.


Loganathan Ramasamy and Zdeněk Žabokrtský.
Prague dependency style treebank for Tamil.
In *Proceedings of LREC 2012*, İstanbul, Turkey, 2012.

Bibliography X


Daniel Zeman.
Reusable tagset conversion using tagset drivers.
In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC 2008*, pages 28–30, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).


Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zděněk Žabokrtský, and Jan Hajič.
Hamledt: To parse or not to parse?
In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors,

Bibliography XI

Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

Daniel Zeman and Philip Resnik.
Cross-language parser adaptation between related languages.
In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India, 2008. Asian Federation of Natural Language Processing, International Institute of Information Technology.