MURI Year 2 Meeting

Noah Smith
CMU
nasmith@cs.cmu.edu
November 2, 2012

Syntactic Annotation Without Pain?

Noah Smith (CMU)
with Chris Dyer (CMU)
& Jason Baldridge (UT)

The Penn Treebank Approach

- $millions, 5 years
- One specific genre/dialect: 1980s *WSJ*!
- Methodologically hard to replicate
- Annotate *everything* …
  - Even though some parts are highly predictable.
  - And “everything” ain’t what it used to be!
  - Today, many people are “rebanking” …
- Licensing (data owned by Murdoch)

Syntax Annotation Attempt #1

- (Jason already told you about this)
- Phrase structure analysis, with heads:
  - Kinyarwanda (KGMC): 2,900 tokens
  - Malagasy (news): 2,300 tokens
- Great for exploring phenomena and *evaluating* text analyzers

… not *building* analyzers.

New Solution

- Allow *partial* annotation of the most important phenomena.
- Represent:
  - chunks (contiguous and non-contiguous)
  - dependencies among tokens, chunks
  - empty nodes
  - explicit "can't be bothered" annotations
- Intuitive, flexible text-based interface and interpreter for annotation
  - Less reliance on expertise

Twice-as-large dataset annotated in 3 weeks (vs. 3-4 months)

Graph Fragment Language

Chris: what it looks like

Jason: what we've done with it

Building Blocks

words

↕

morphology &
language models

↕

shallow syntax
(e.g., parts of speech)

↕

full syntax
(e.g., dependencies)

↕

semantics
(e.g., semantic graphs)

4. Marrying knowledge and learning for analyzing and predicting words (Jason Mielens, UT; Chris Dyer, CMU)

2. Type annotation for efficient prior POS knowledge use (Dan Garrette, UT)

3. Multilingual guidance for parser transfer (Noah Smith, CMU; Zach Hynes, MIT)

1. Synchronous hyperedge replacement grammars for efficient semantic analysis & generation (David Chiang, USC/ISI)

Text Analysis Transfer with Non-Parallel Multilingual Guidance

Shay Cohen CMU → Columbia University
Dipanjan Das CMU → Google Research
Noah Smith Carnegie Mellon University

Multilingual Transfer

using parallel data

no parallel data **(hard)**

joint learning for multiple languages

Snyder et al. (2009)

Naseem et al. (2010)

supervision in helper language(s)

Yarowsky and Ngai (2001)

Xi and Hwa (2005)

Zeman and Resnik (2008)

Smith and Eisner (2009)

Das and Petrov (2011)

McDonald et al. (2011)

joint learning for multiple languages

Cohen and Smith (2009)

Berg-Kirkpatrick and Klein (2010)

supervision in helper language(s)

this work:

Cohen et al. (2011)

Naseem et al. (2012)

Mixture Approach

Annotated data in helper languages + **Unlabeled** data in Portuguese = **Portuguese parameters**

German, English, Italian, Czech

Mixture Approach

Annotated data in helper languages + **Unlabeled** data in Portuguese = **Portuguese parameters**

German, English, Italian, Czech

Coarse, *universal* parameters

Coarse, *universal* parameters

1

Technique in a Nutshell

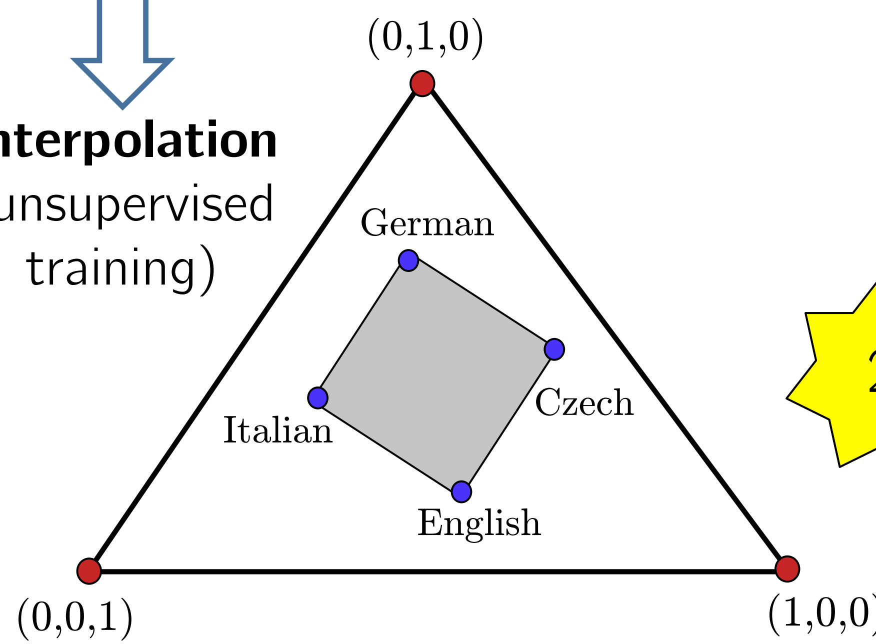
Annotated data in helper languages + **Unlabeled** data in Portuguese = **Portuguese parameters**

German, English, Italian, Czech

Coarse, *universal* parameters

Coarse, *universal* parameters

Interpolation (unsupervised training)



2

Mixture Approach

Annotated data in helper languages + **Unlabeled** data in Portuguese = **Portuguese parameters**

German, English, Italian, Czech

Coarse, *universal* parameters

Coarse, *universal* parameters

Interpolation (unsupervised training)

coarse parameters of Portuguese

2

Mixture Approach

Annotated data in helper languages + **Unlabeled** data in Portuguese = **Portuguese parameters**

German, English, Italian, Czech

Monolingual unsupervised training in Portuguese

Coarse-to-fine expansion and initialization

3

coarse parameters of Portuguese

Mixture Approach

Annotated data + **Unlabeled** data in Portuguese = **Portuguese parameters**

German, English, Italian, Czech

Monolingual unsupervised training in Portuguese

3

Results: Part-of-Speech Tagging

		our results	
	Direct Gradient (DG)	Uniform + DG	Mixture + DG
Number of Languages with Best Results	2 (Portuguese, Danish)	2 (Turkish, Bulgarian)	**6**
Average Accuracy	40.6	41.0	**43.3**

(without tag dictionary)

Results: Dependency Parsing

	EM	PR	PGI	Uniform	Mixture	Uniform + EM	Mixture + EM
Number of Languages with Best Results	0	2 (Turkish, Slovene)	0	3 (Bulgarian, Swedish, Dutch)	1 (Danish)	1 (Greek)	3 (Portuguese, Japanese, Spanish)
Average Accuracy	41.4	50.2*	53.6*	61.6	**62.2**	61.5	62.1

our results

Building Blocks

words

4. Marrying knowledge and learning for analyzing and predicting words (Jason Mielens, UT; Chris Dyer, CMU)

morphology & language models

2. Type annotation for efficient prior POS knowledge use (Dan Garrette, UT)

shallow syntax (e.g., parts of speech)

3. Multilingual guidance for parser transfer (Noah Smith, CMU; Zach Hynes, MIT)

full syntax (e.g., dependencies)

1. Synchronous hyperedge replacement grammars for efficient semantic analysis & generation (David Chiang, USC/ISI)

semantics (e.g., semantic graphs)

Tree-to-Tree Translation with Quasi-Synchronous Phrase Dependency Grammars

Kevin Gimpel

CMU $\rightarrow$ Toyota Technological Institute at Chicago

Advisor/presenter: Noah Smith (CMU)

Committee: Jaime Carbonell (CMU), David Chiang (USC/ISI)

Phrase Dependency

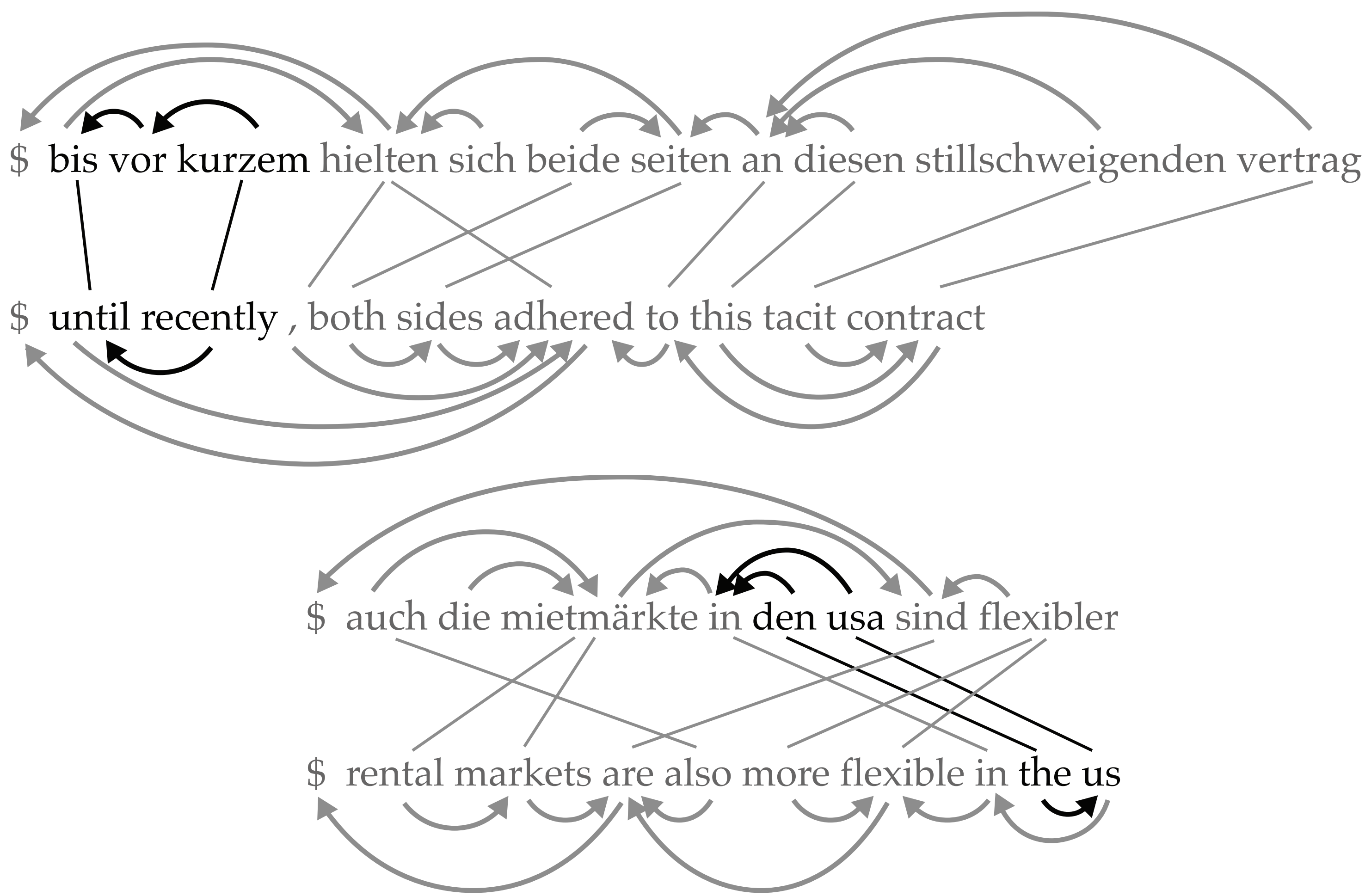
Kevin Gimpel

CMU $\rightarrow$ Toyota Technological Institute at Chicago

Advisor/presenter: Noah Smith (CMU)

Committee: Jaime Carbonell (CMU),
David Chiang (USC/ISI)

Syntactic Divergence in Real Data

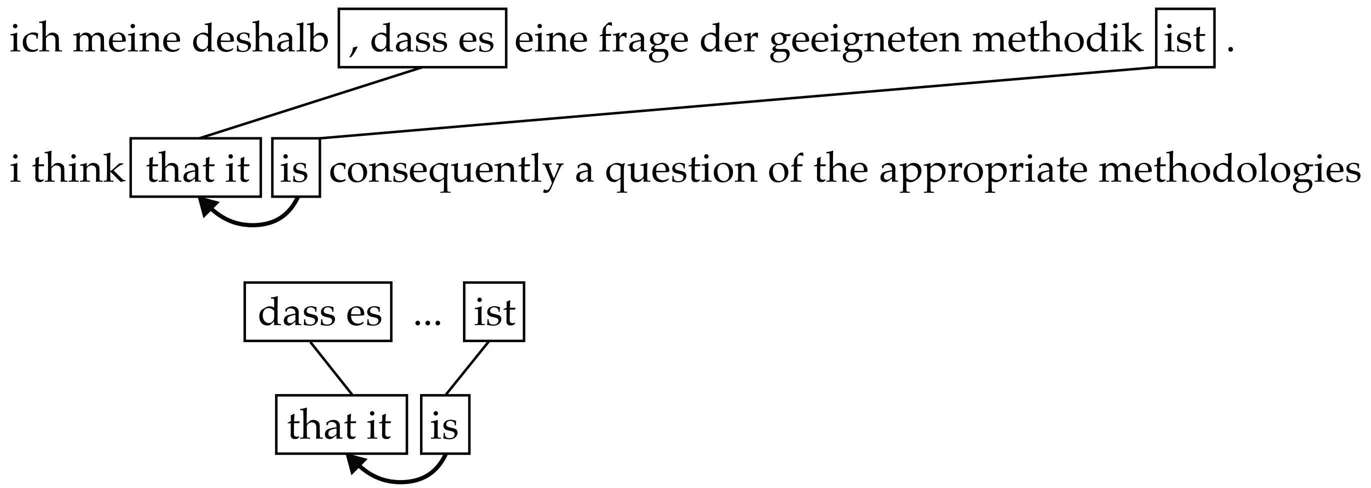


Problems & Solutions

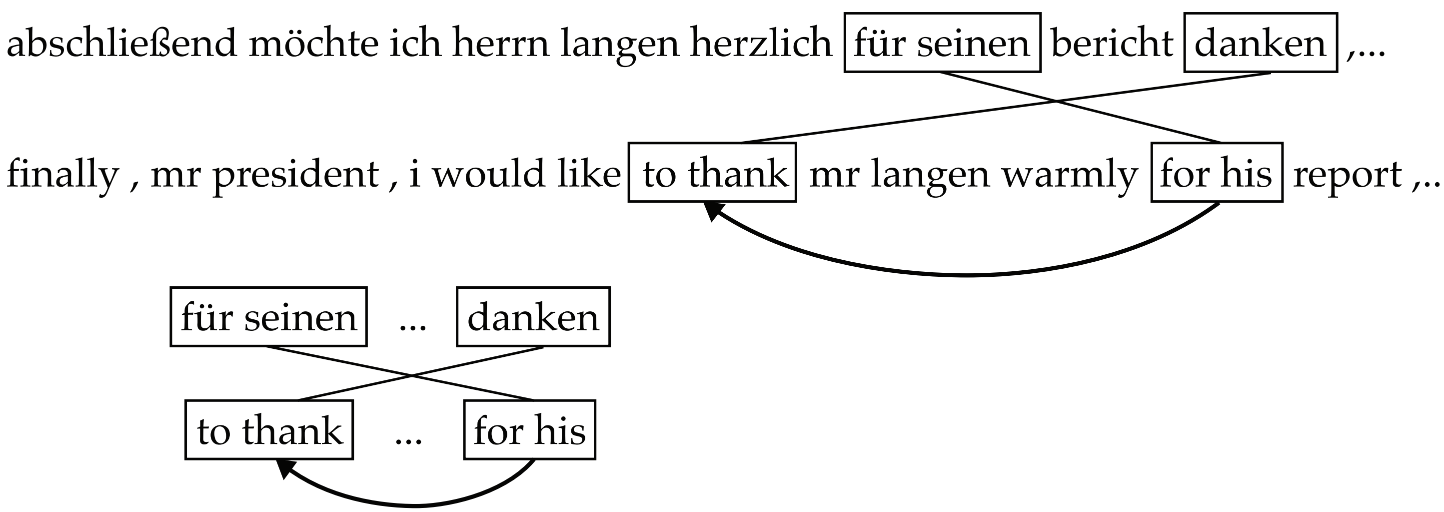
- Lexical dependencies align in many complex ways.
  - "Bigger" dependency rules on *phrases*.
- Across languages, dependency parsing ranges from pretty okay to nonexistent.
  - Source-side syntax is *optional*.
  - We explored *unsupervised parsing* as a stand-in.
- Flexible formalism $\longrightarrow$ expensive decoding!
  - Coarse-to-fine strategy (key advance over 2011).

Example Extracted Rules

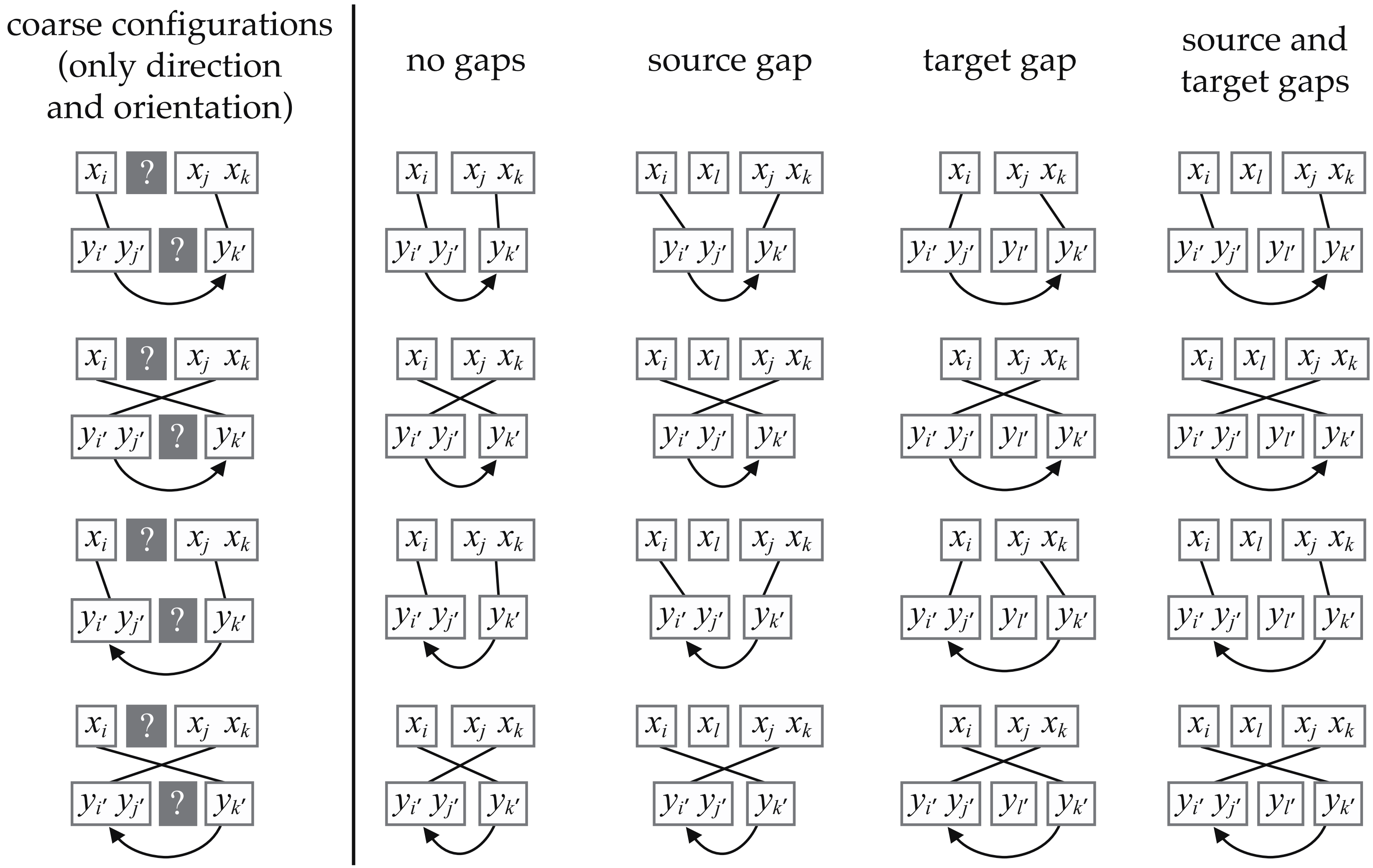
(a)



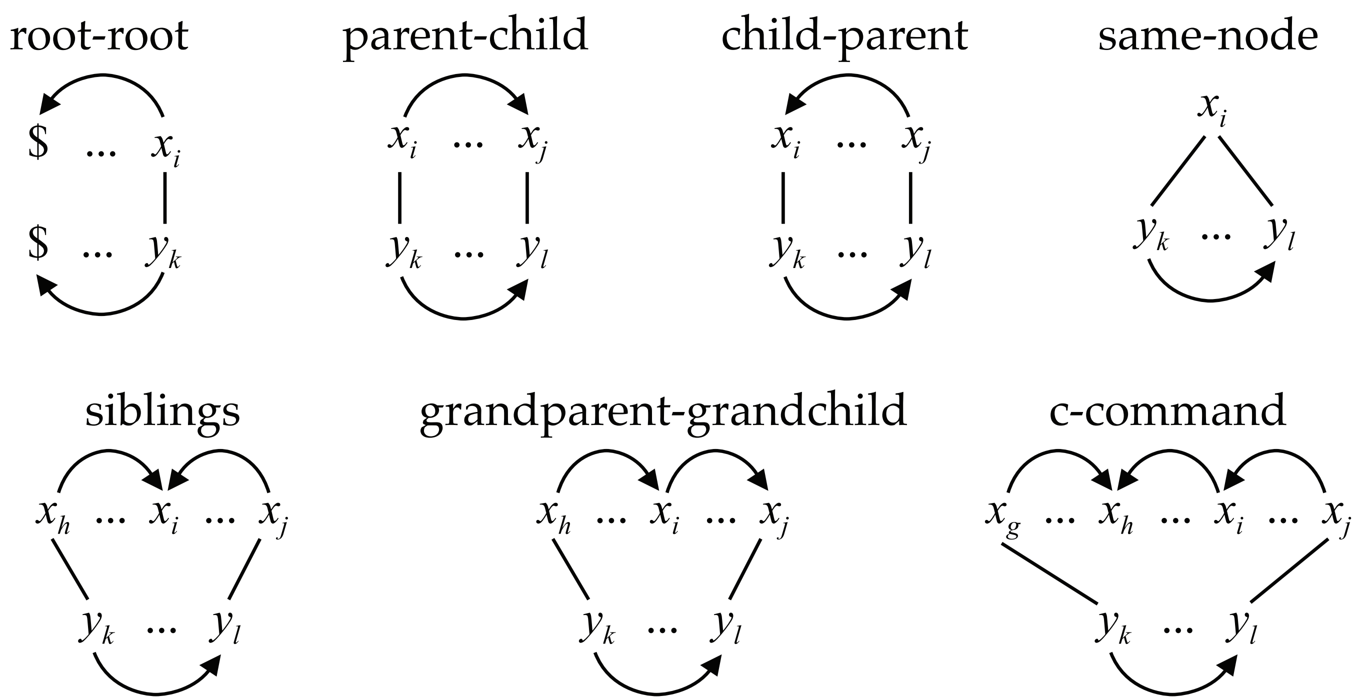
(b)



String-to-Tree Configurations

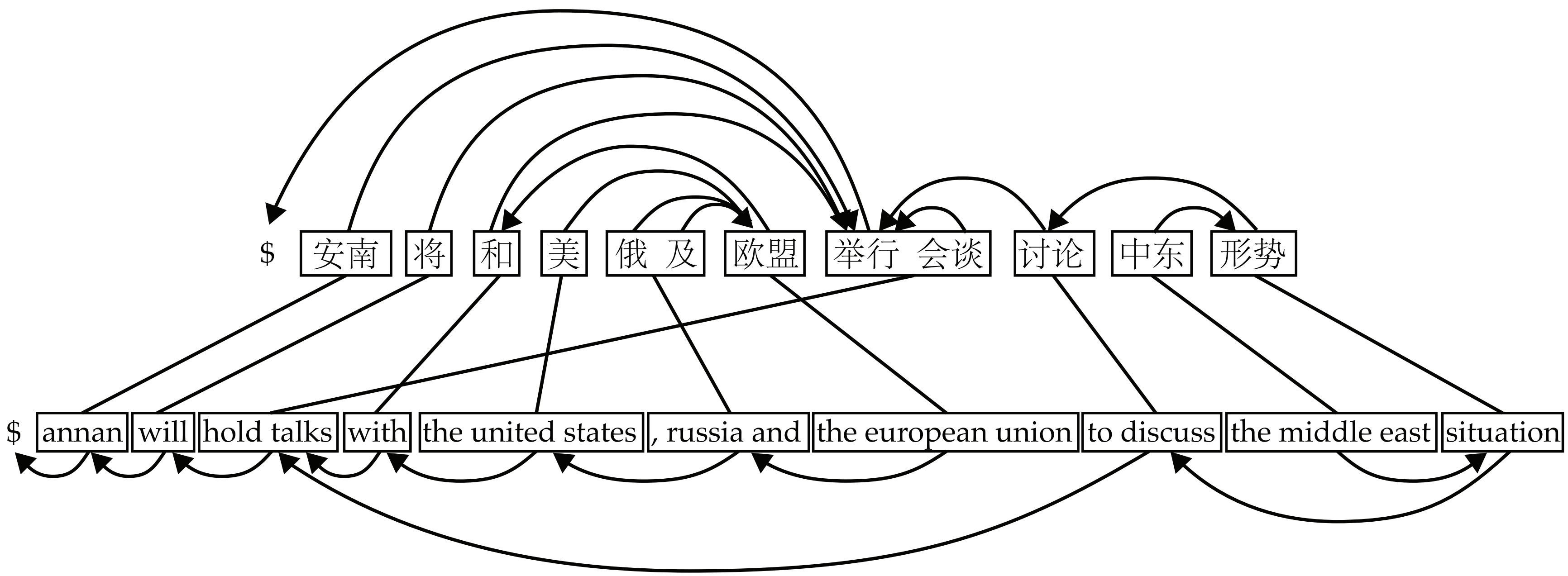


Tree-to-Tree Configurations



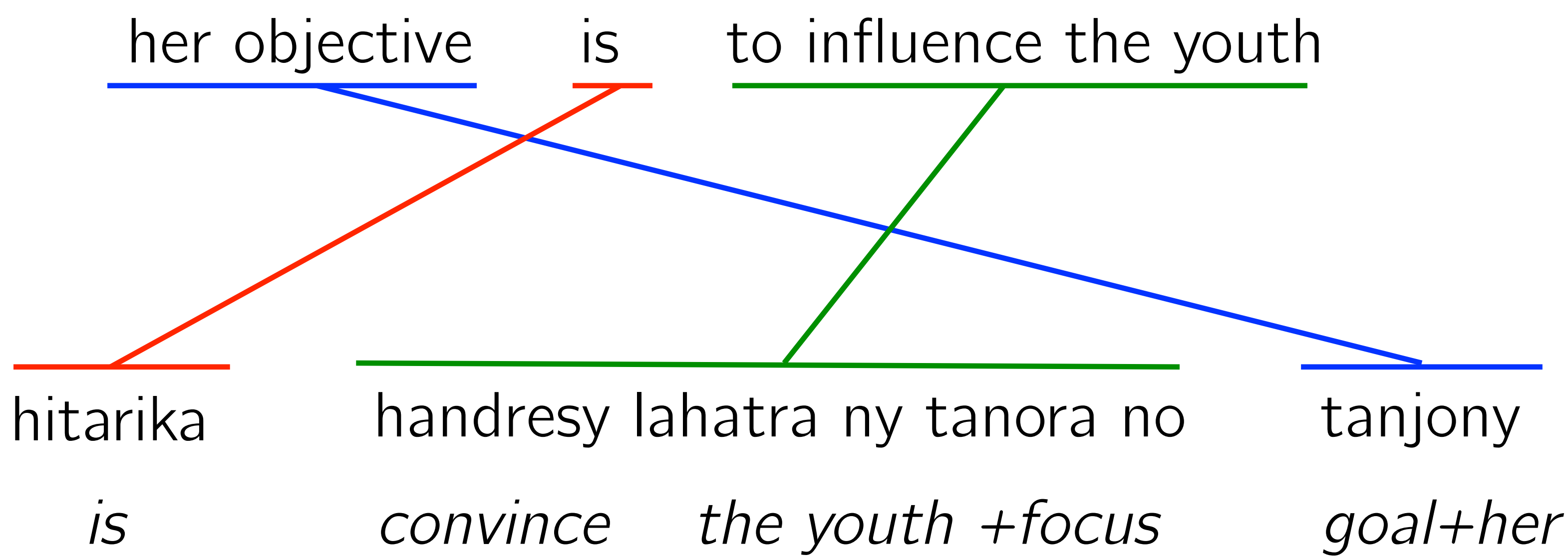
Credit: D. Smith and Eisner (2006)

Example Derivation Structure



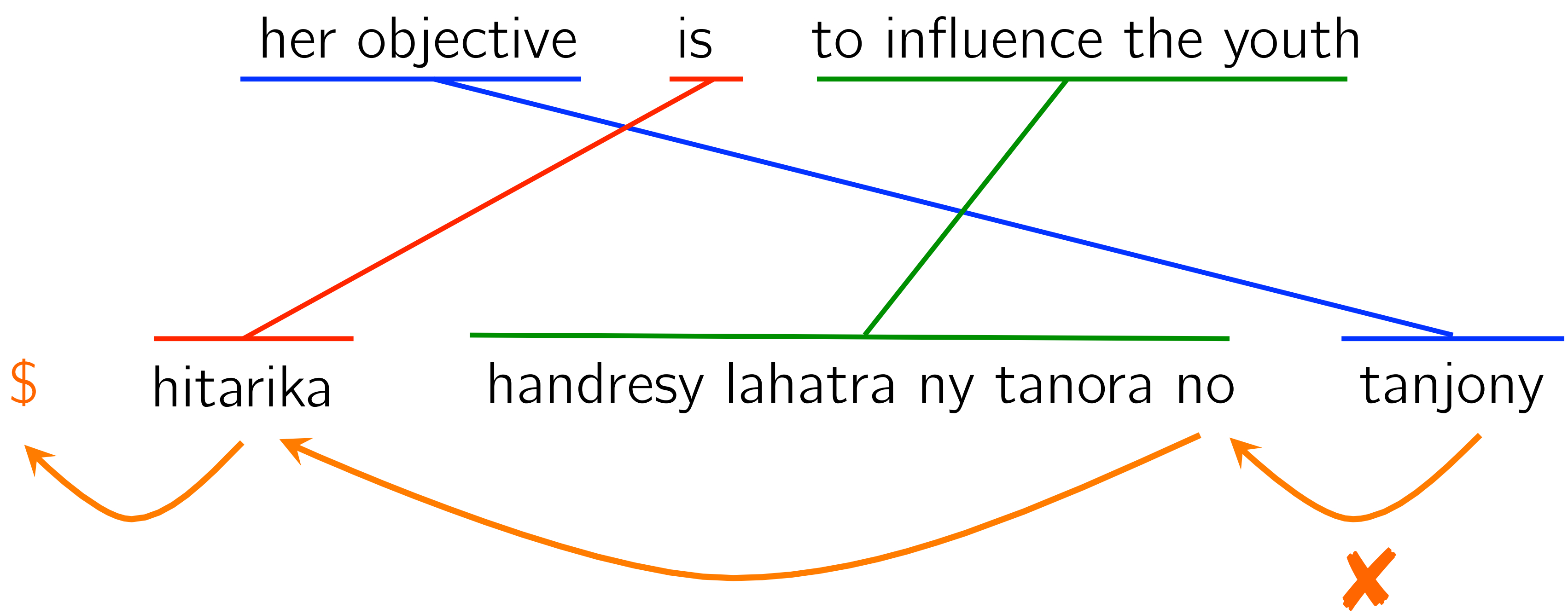
references: annan to hold talks with us , russia and eu over situation in middle east
annan will discuss middle east situation with u.s. , russia and european union
annan to discuss mideast situation with us , russia and eu
annan to meeting the us , russia and eu to discuss middle east crisis

An English-Malagasy Example



$$\text{S O V} \rightarrow \text{O V S}$$

Unsupervised Dependencies



Results: English→Malagasy

System	%BLEU (test)
Moses (stack size = 200)	15.05
Moses (stack size = 500)	15.08
QPDG	**15.54**

Lessons & Future

- ✓ QPDG: syntax + phrases can work together for better MT, even in syntax-hostile scenarios

- ❑ Use improved POS tagger (UT) and dependency parser (MIT) developed in year 2!

- ❑ Model morphology in the language model (Jason Mielens and Chris Dyer's talk this morning) and translation model (Waleed Ammar's talk later)