Rudolf Rosa, David Mareček, Aleš Tamchyna
{rosa,marecek,tamchyna}@ufal.mff.cuni.cz

Deepfix:

Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ACL SRW, Sofia, 6th August 2013

Motivation

- Source text in English:

  EU criticizes not only the Greek government.

Motivation

- Source text in English:

  EU criticizes not only the Greek government

- Google Translate to Czech (6th Aug 2013):

  EU kritizuje nejen řecká vláda

Motivation

- Source text in English:

  EU criticizes not only the Greek government

- Google Translate to Czech (6th Aug 2013):

  *EU kritizuje nejen **řecká vláda***$_{\textbf{nominative (subject)}}$

  - *Not only **the Greek government** criticizes EU*

Motivation

- Source text in English:

  EU criticizes not only the Greek government

- Google Translate to Czech (6th Aug 2013):

  EU kritizuje nejen ***řecká vláda***$_{\textbf{nominative (subject)}}$

  - *Not only* ***the Greek government*** *criticizes EU*

- Post-editation by Deepfix:

  EU kritizuje nejen ***řeckou vládu***$_{\textbf{accusative (object)}}$

  - *EU criticizes not only* ***the Greek government***

Outline

1. Problem definition
2. Sentence analysis
3. Sentence post-editing
4. Results

Outline

1. Problem definition
   - Errors in valency in SMT outputs
2. Sentence analysis
3. Sentence post-editing
4. Results

Outline

1. Problem definition
   - ➔ Errors in valency in SMT outputs
2. Sentence analysis (DEEP)
   - ➔ Deep dependency parsing
3. Sentence post-editing
4. Results

Outline

1. Problem definition
   - Errors in valency in SMT outputs
2. Sentence analysis (DEEP)
   - Deep dependency parsing
3. Sentence post-editing (FIX)
   - Statistical model of valency
4. Results

Outline

1. Problem definition
   - Errors in valency in SMT outputs
2. Sentence analysis (DEEP)
   - Deep dependency parsing
3. Sentence post-editing (FIX)
   - Statistical model of valency
4. Results
   - Automatic & manual evaluation of Deepfix

Subject – object dichotomy

- English: **position** (left/right constituent)
  - Subject *criticize* Object
- Czech: **morphological case** (nominative/other); word order relatively free
  - Subject$_{\text{nominative}}$ *kritizovat* Object$_{\text{accusative}}$
  - Object$_{\text{accusative}}$ *kritizovat* Subject$_{\text{nominative}}$
  - Subject$_{\text{nominative}}$ Object$_{\text{accusative}}$ *kritizovat*
  - Object$_{\text{accusative}}$ Subject$_{\text{nominative}}$ *kritizovat*

Valency of *criticize* (*kritizovat*)

- example sentence
  - *EU*$_{\text{subject}}$ *criticizes not only the Greek government*$_{\text{object}}$
  - *EU*$_{\text{nominative}}$ *kritizuje nejen řeckou vládu*$_{\text{accusative}}$

Valency of *criticize* (*kritizovat*)

- example sentence
  - *EU*$_{\text{subject}}$ *criticizes not only the Greek government*$_{\text{object}}$
  - *EU*$_{\text{nominative}}$ *kritizuje nejen řeckou vládu*$_{\text{accusative}}$
- a valency frame of a verb

subject	*criticize*	object
nominative	*kritizovat*	accusative

Valency of *criticize* (*kritizovat*)

- example sentence
  - *EU*$_{\text{subject}}$ *criticizes not only the Greek government*$_{\text{object}}$
  - *EU*$_{\text{nominative}}$ *kritizuje nejen řeckou vládu*$_{\text{accusative}}$
- a valency frame of a verb

subject	*criticize*	object	(position)
nominative	*kritizovat*	accusative	(cases)

Valency of *criticize* (*kritizovat*)

- example sentence
  - *EU*$_{\text{subject}}$ *criticizes not only the Greek government*$_{\text{object}}$
  - *EU*$_{\text{nominative}}$ *kritizuje nejen řeckou vládu*$_{\text{accusative}}$
- a valency frame of a verb

subject	*criticize*	object	(position)
nominative	*kritizovat*	accusative	(cases)

- decomposition into head-argument pairs
  - (*to criticize, government*) ~ (*kritizovat, vládu*)

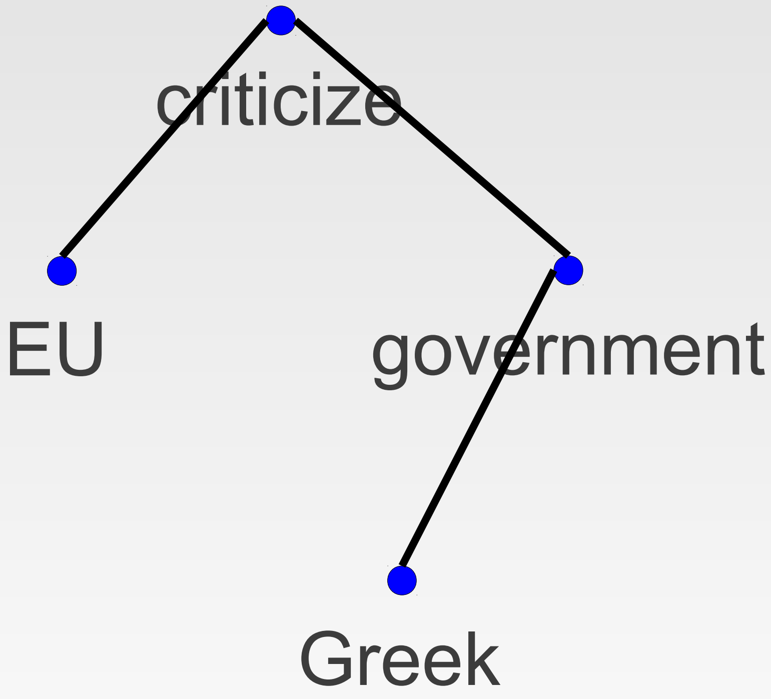
Sentence analysis (DEEP)

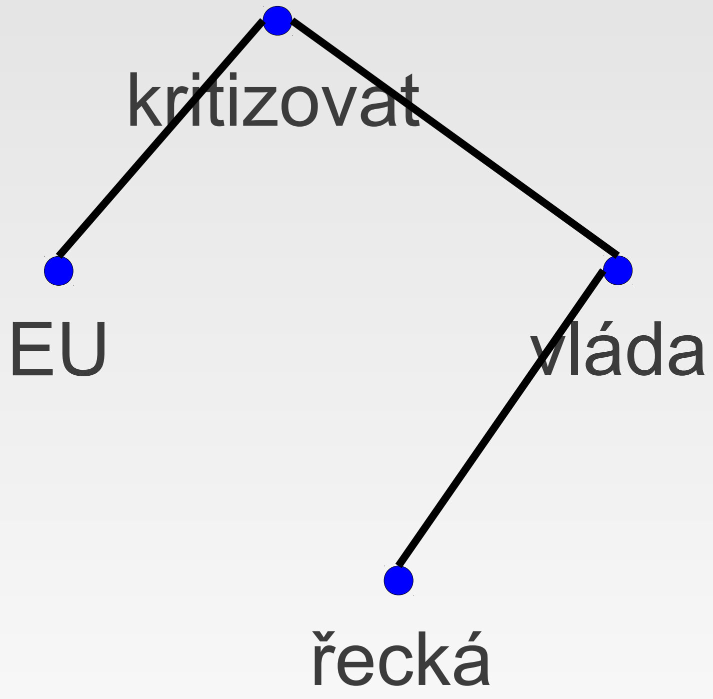
- tagging & lemmatization
    - combination of rule-based and statistical approach
- word-alignment
    - unsupervised methods (Giza++)
- dependency parsing
    - statistical, trained on manually created treebanks
    - parser adapted for parsing of SMT outputs
- induction of deep structure (tectogrammar)
    - rule-based

Deep syntactic dependency trees

EU criticizes
the Greek government

EU kritizuje
řecká vláda

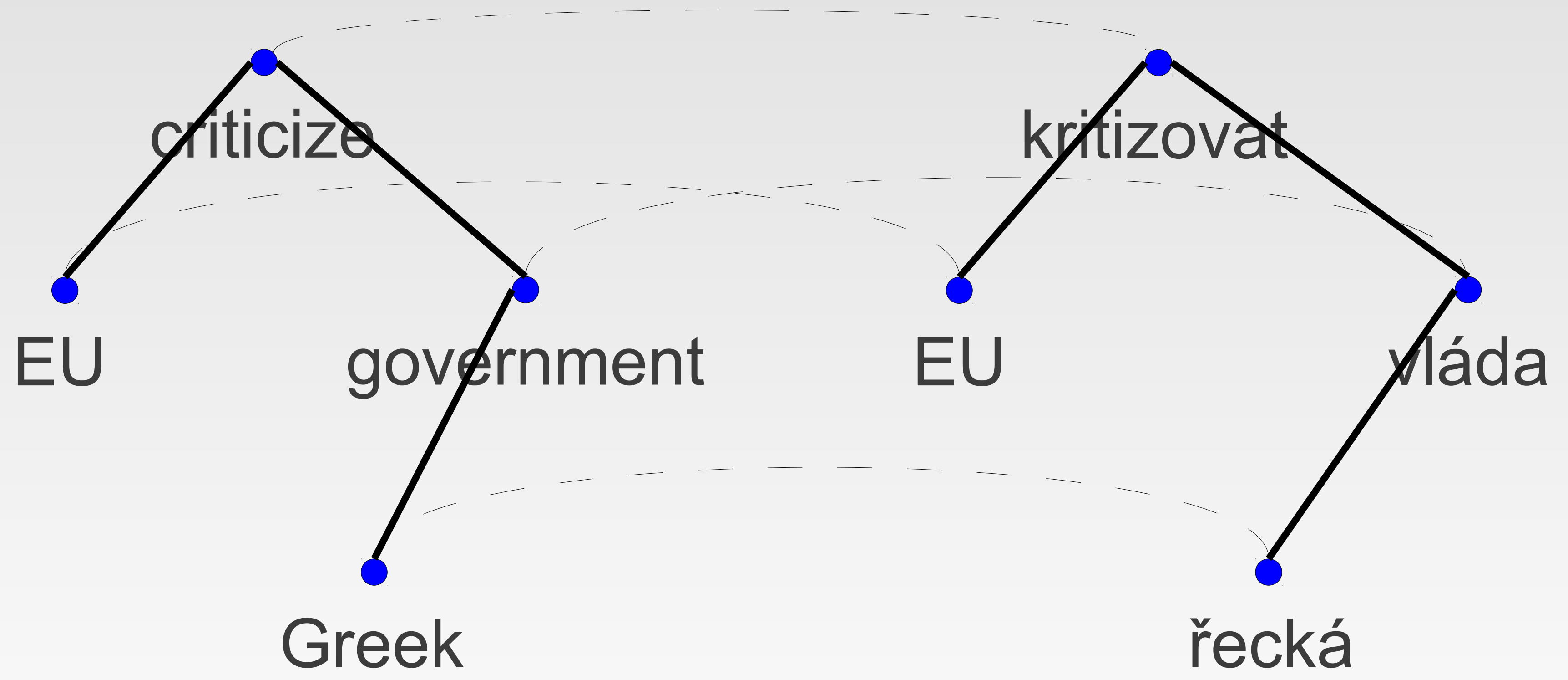




Deep syntactic dependency trees

EU criticizes the Greek government

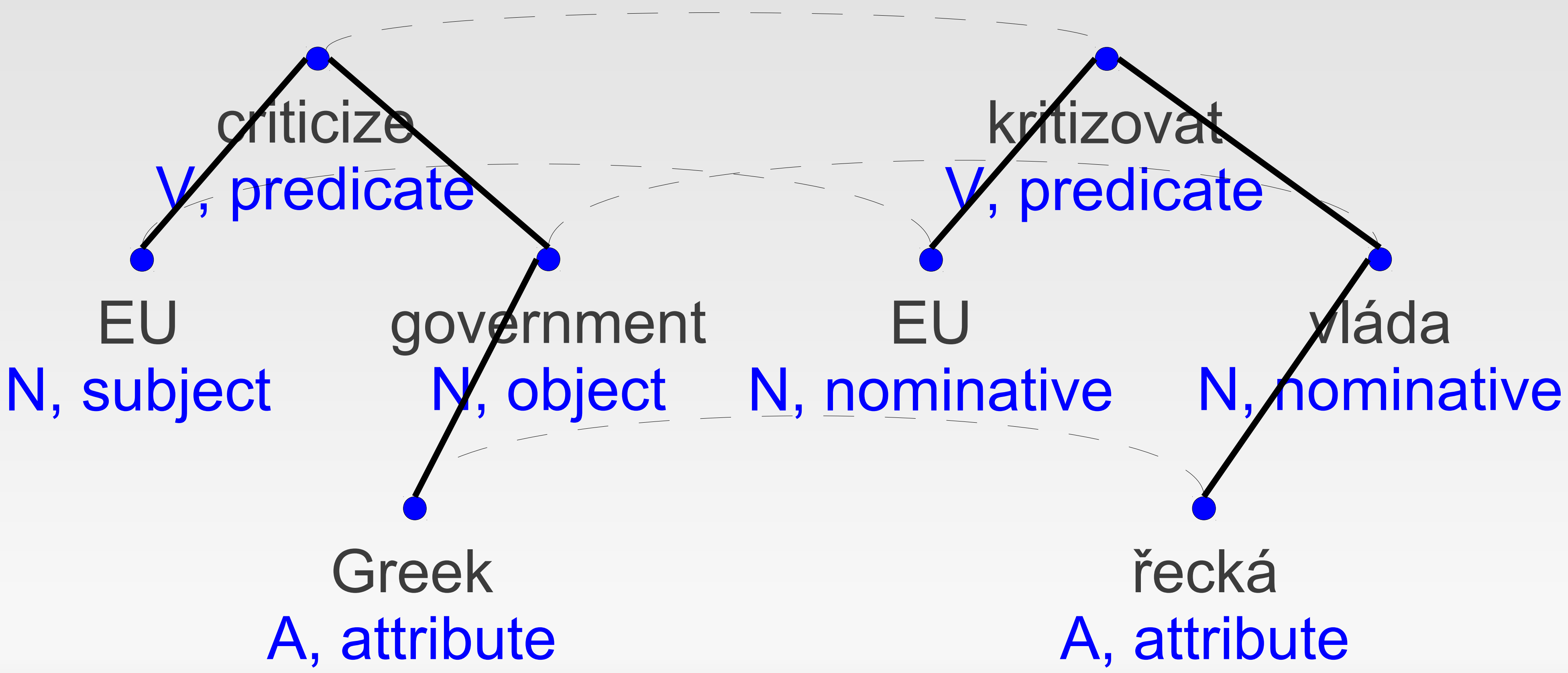
EU kritizuje řecká vláda



Deep syntactic dependency trees

EU criticizes
the Greek government

EU kritizuje
řecká vláda



criticize
V, predicate

EU
N, subject

government
N, object

Greek
A, attribute

kritizovat
V, predicate

EU
N, nominative

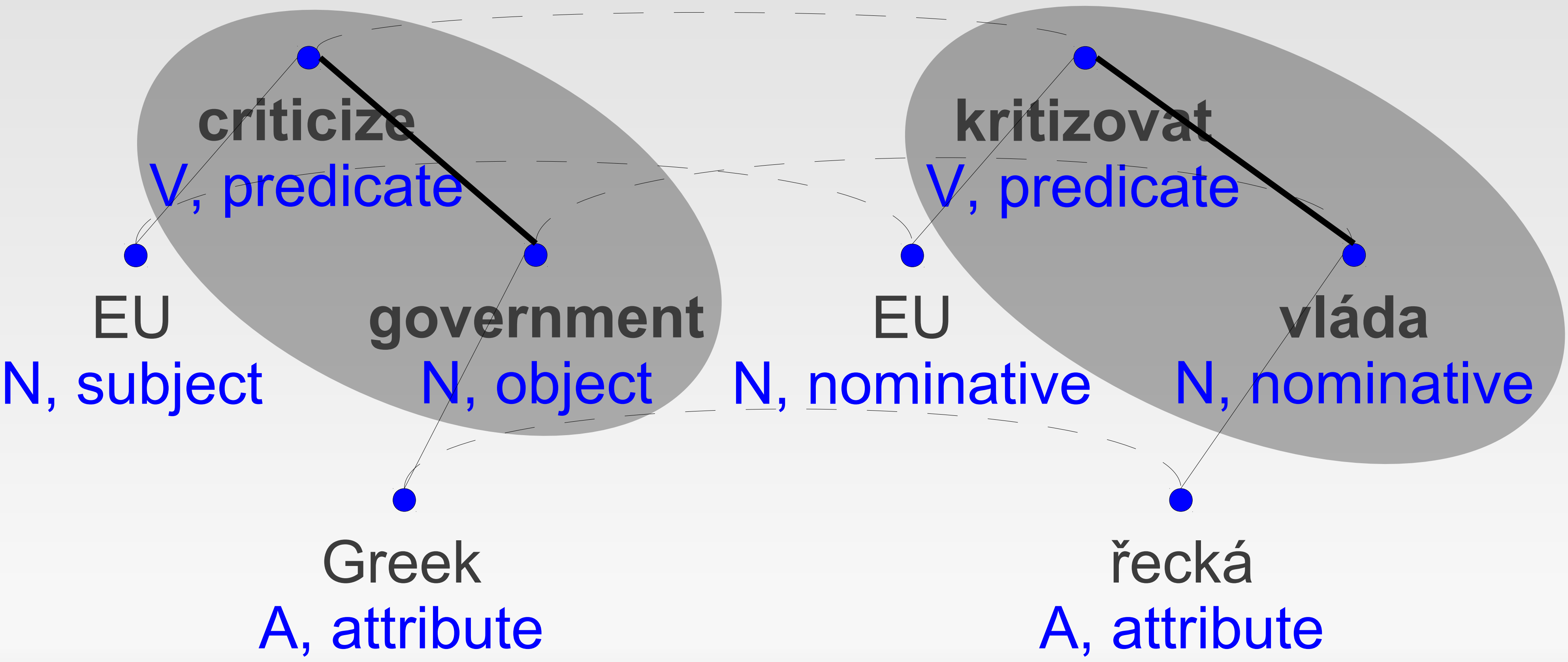
vláda
N, nominative

řecká
A, attribute

(head, arg) pair identification

*EU **criticizes***
*the Greek **government***

*EU **kritizuje***
*řecká **vláda***



criticize
V, predicate

EU
N, subject

government
N, object

Greek
A, attribute

kritizovat
V, predicate

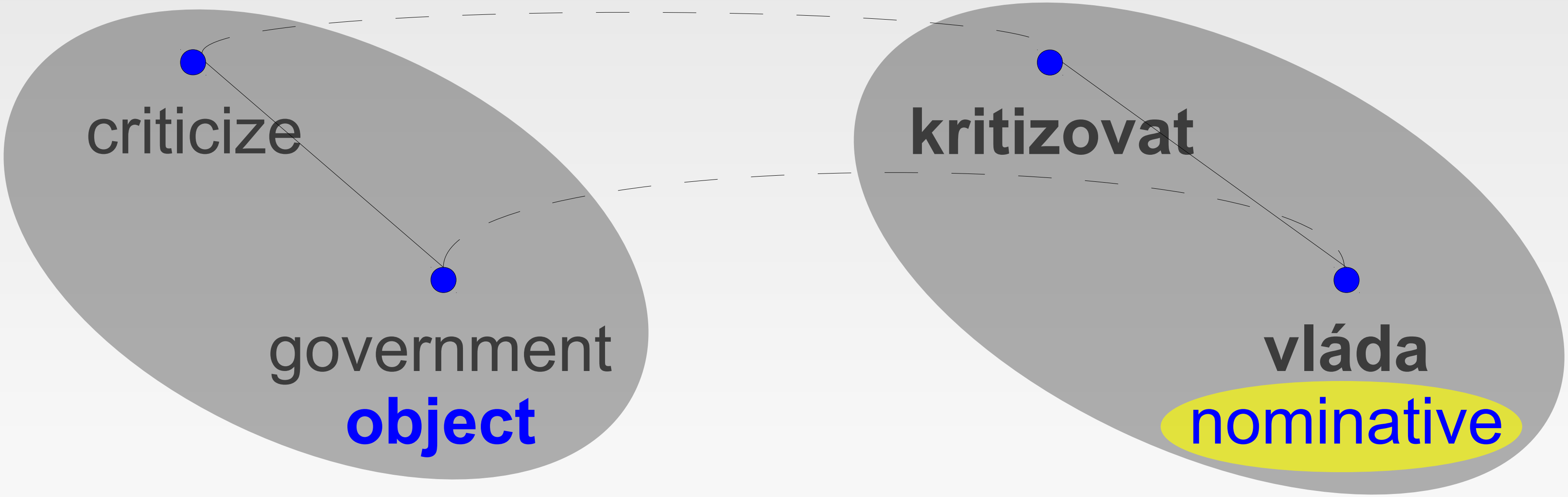
EU
N, nominative

vláda
N, nominative

řecká
A, attribute

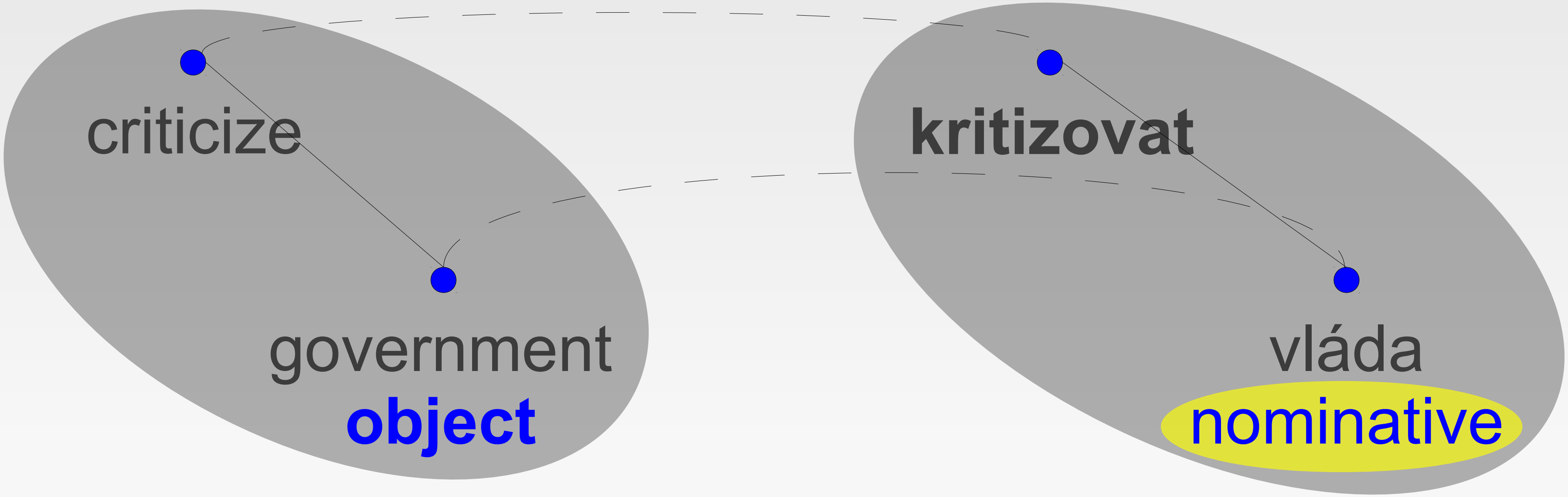
Valency models (FIX)

- $P(arg_{case} \mid head_{lemma}, English\_arg_{case})$
- $P(arg_{case} \mid head_{lemma}, English\_arg_{case}, arg_{lemma})$
- estimated from CzEng 1.0 (15M parallel stcs)



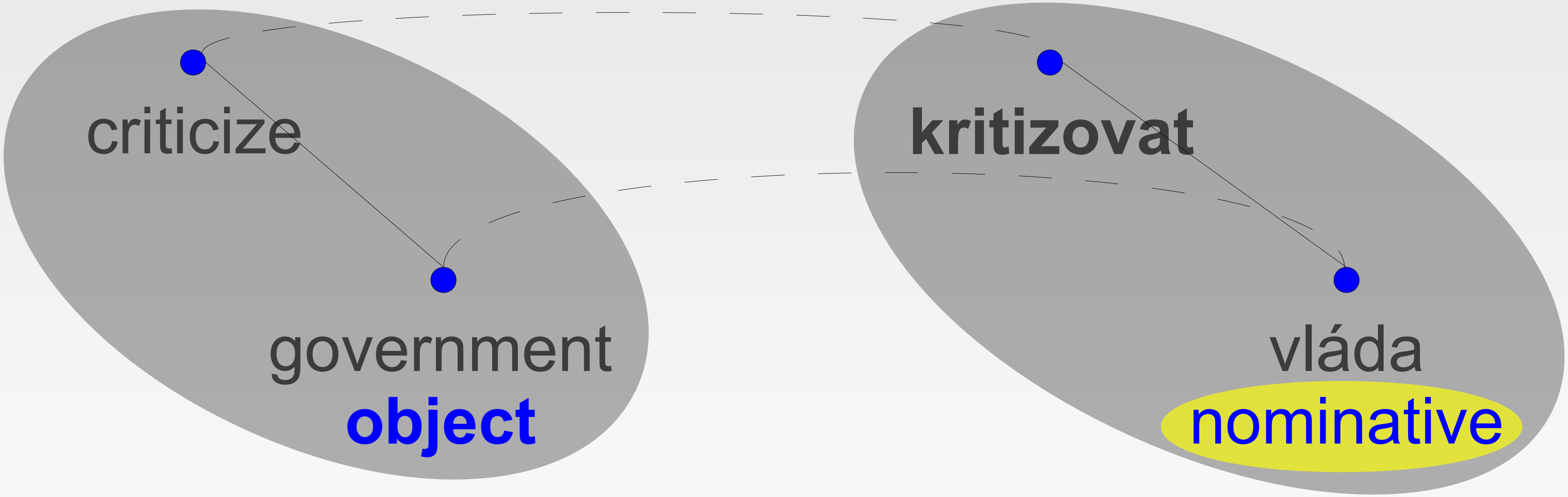
Argument case probabilities

- P(nominative | *kritizovat*, object) = 0.03
- P(accusative | *kritizovat*, object) = 0.80



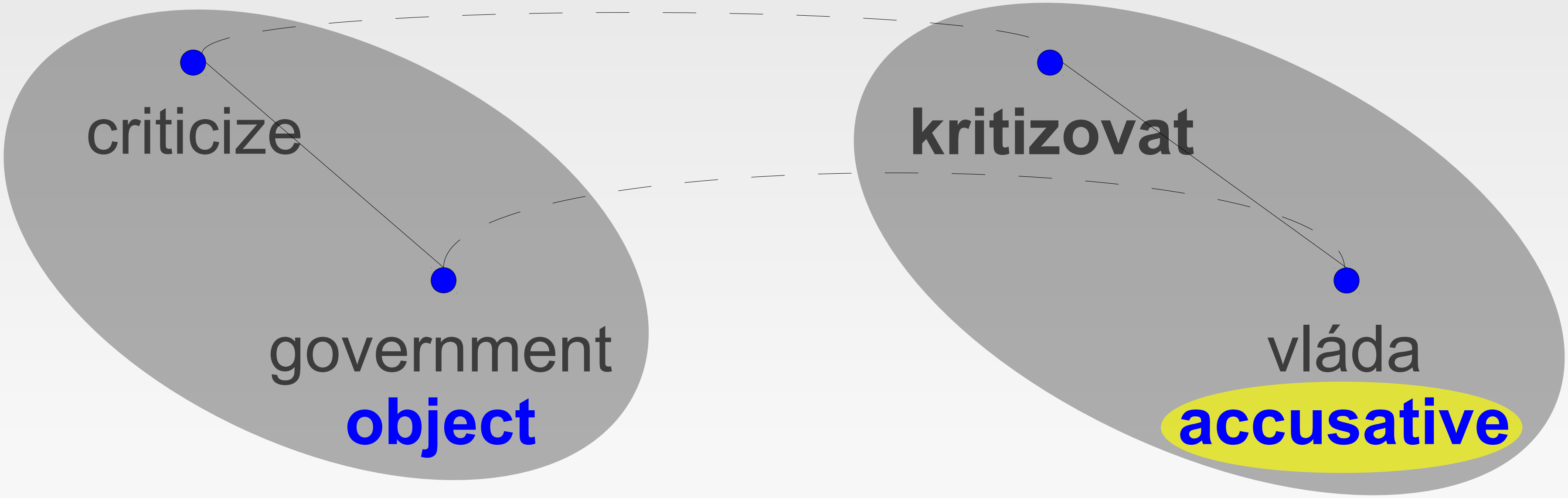
Argument case probabilities

- P(nominative | *kritizovat*, object) = 0.03
- P(accusative | *kritizovat*, object) = 0.80
- threshold: 0.55



Argument case correction

- P(nominative | *kritizovat*, object) = 0.03
- P(**accusative** | *kritizovat*, object) = **0.80**
- threshold: **0.55**



Sentence correction

- Statitical machine translation output:

  EU kritizuje nejen ***řecká***$_{\text{nominative}}$ ***vláda***$_{\text{nominative}}$

  - *Not only* ***the Greek government*** *criticizes EU*

Sentence correction

- Statitical machine translation output:

  *EU kritizuje nejen **řecká**$_{\textbf{nominative}}$ **vláda**$_{\textbf{nominative}}$*

  - *Not only **the Greek government** criticizes EU*

- Valency model correction:

  *EU kritizuje nejen **řecká**$_{\textbf{nominative}}$ **vládu**$_{\textbf{accusative}}$*

Sentence correction

- Statitical machine translation output:

  EU kritizuje nejen ***řecká***$_{\text{nominative}}$ ***vláda***$_{\text{nominative}}$

  - *Not only* ***the Greek government*** *criticizes EU*

- Valency model correction:

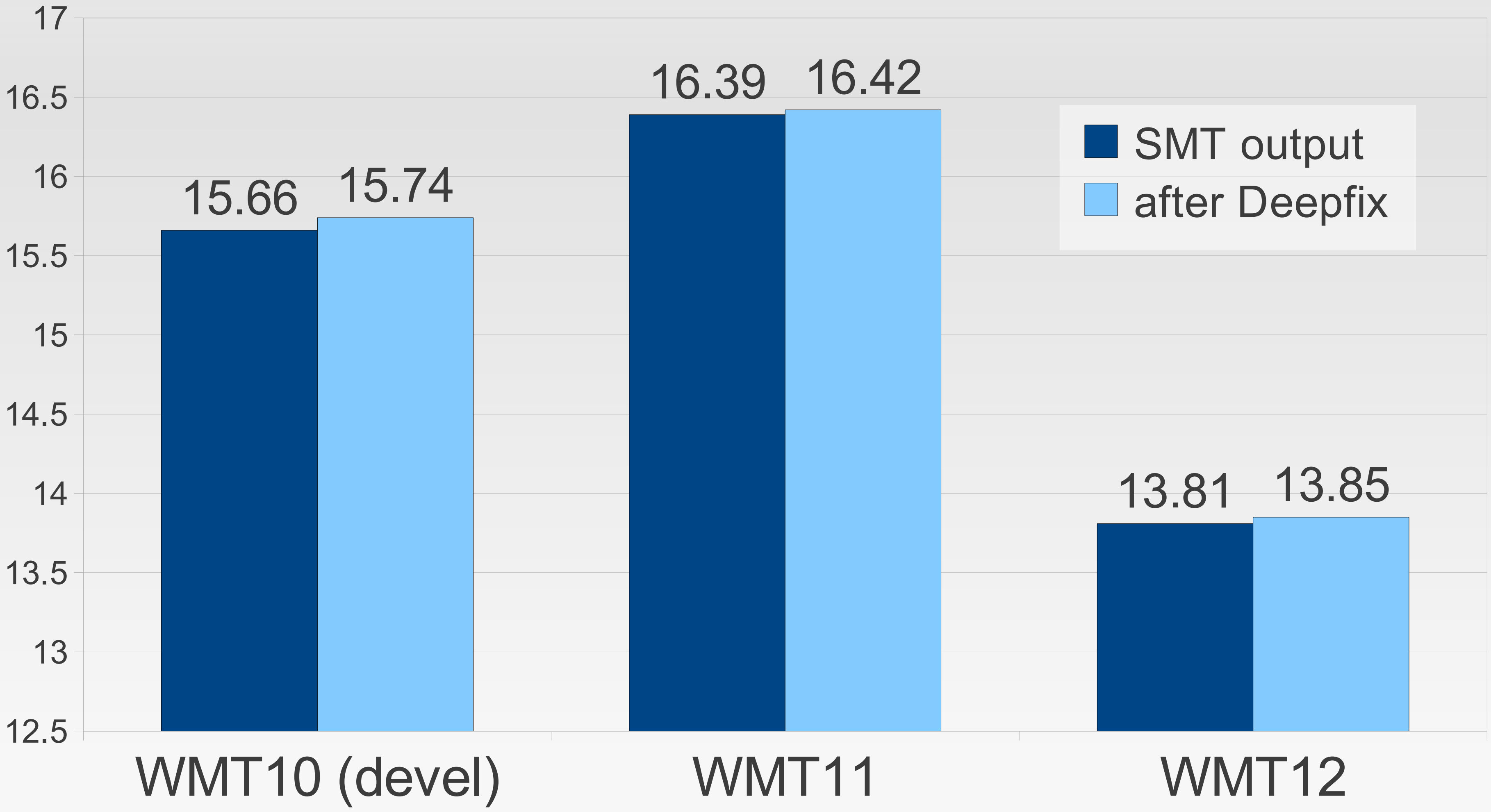
  EU kritizuje nejen ***řecká***$_{\text{nominative}}$ ***vládu***$_{\text{accusative}}$

- Agreement enforcement:

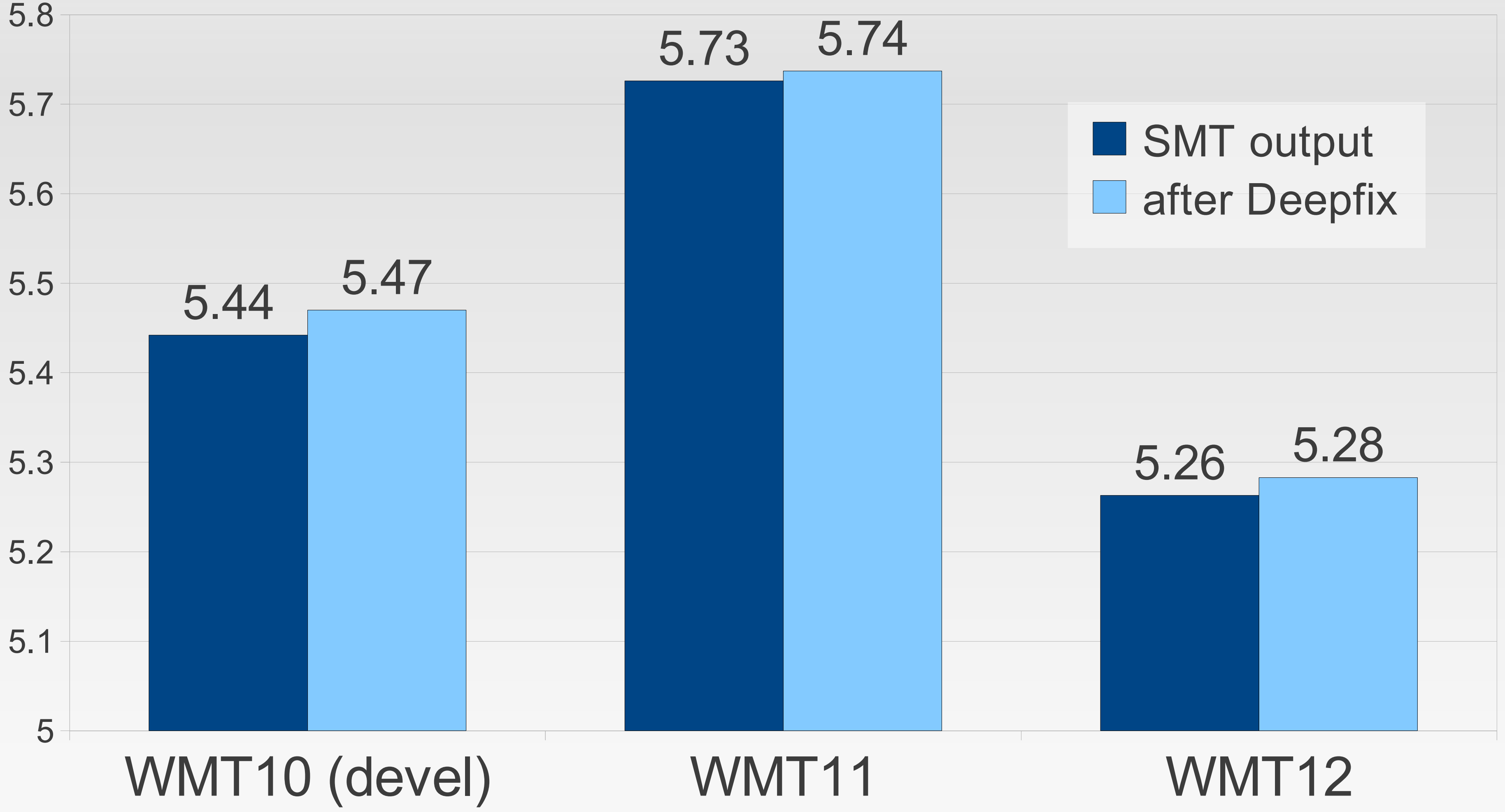
  EU kritizuje nejen ***řeckou***$_{\text{accusative}}$ ***vládu***$_{\text{accusative}}$

  - *EU criticizes not only* ***the Greek government***

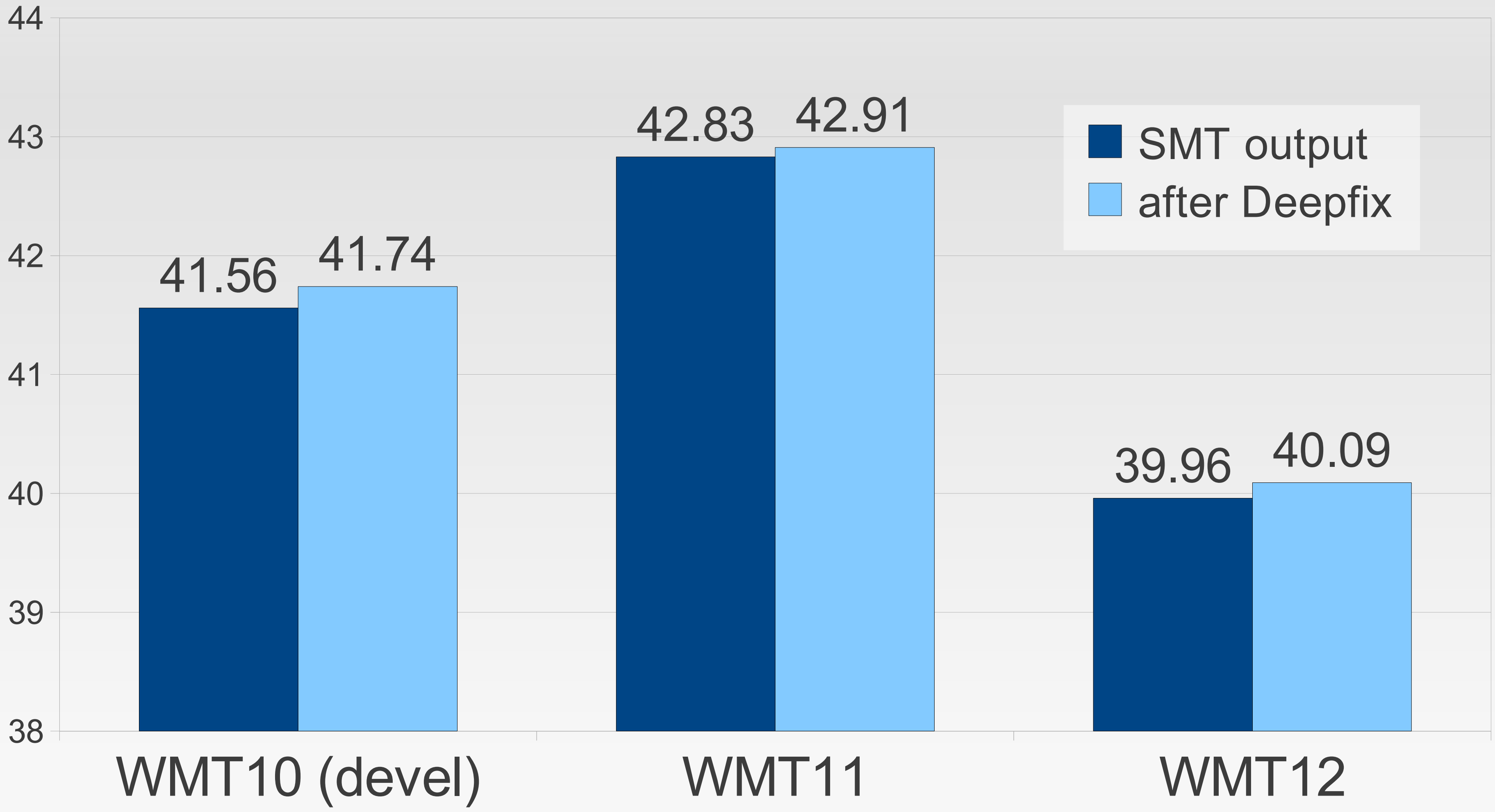
Automatic evaluation (BLEU)



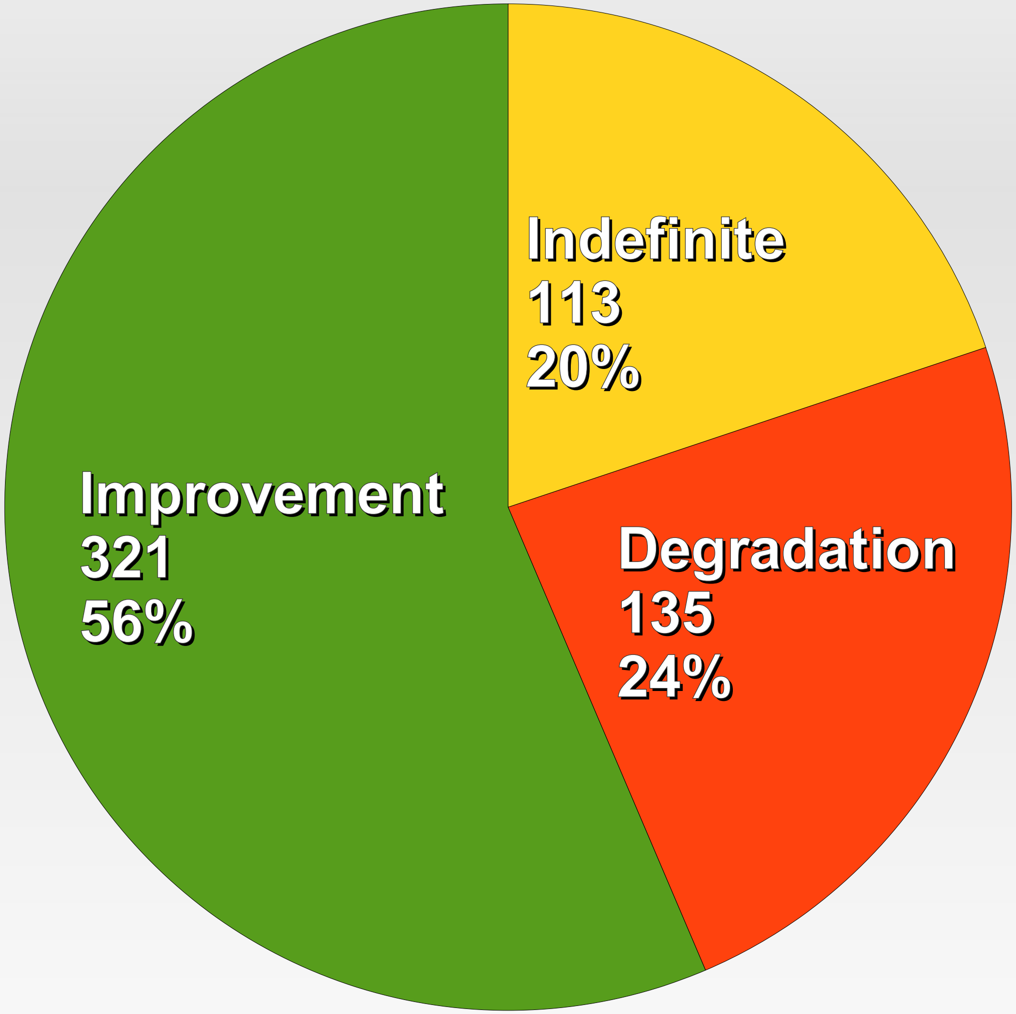
Automatic evaluation (NIST)



Automatic evaluation (1-PER)



Manual evaluation (changed stcs)



Conclusion

- address valency errors
  - statistical post-editing of SMT
- identify head-argument pairs (DEEP)
  - deep syntactic analysis
- find the best case for the arguments (FIX)
  - statistical valency model
- obtain slight improvement of translation quality
  - indicated by automatic evaluation
  - confirmed by manual evaluation

Future work

- explore existing valency lexicons
- more intricate modelling
  - combine more models
  - machine learning (now thresholds semi-manual, and overfitted to development data)
- further adapt underlying NLP tools (tagger)
- extend to other language pairs

Thank you for your attention

Rudolf Rosa, David Mareček, Aleš Tamchyna
{rosa,marecek,tamchyna}@ufal.mff.cuni.cz

**Deepfix:
Statistical Post-editing
of Statistical Machine Translation
Using Deep Syntactic Analysis**

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

For this presentation and other information, please visit:

http://ufal.mff.cuni.cz/~rosa/