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

ÚFAL Seminar, Příchovice, 19th September 2013
(originally presented at 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***

DEEPFIX Outline

DEEPFIX Outline

- Problem definition
  - Errors in valency in SMT outputs

DEEPFIX Outline

- Problem definition
  - Errors in valency in SMT outputs
- Related work

DEEPFIX Outline

- Problem definition
  - Errors in valency in SMT outputs
- Related work
- Step 1 (DEEP): Sentence analysis
  - Deep dependency parsing (t-trees)

DEEPFIX Outline

- Problem definition
  - Errors in valency in SMT outputs
- Related work
- Step 1 (DEEP): Sentence analysis
  - Deep dependency parsing (t-trees)
- Step 2 (FIX): Sentence post-editing
  - Statistical model of valency

DEEPFIX Outline

- Problem definition
  - Errors in valency in SMT outputs
- Related work
- Step 1 (DEEP): Sentence analysis
  - Deep dependency parsing (t-trees)
- Step 2 (FIX): Sentence post-editing
  - Statistical model of valency
- Results

Subject – Object dichotomy

- English: **position** (left/right constituent)
  - *EU*$_{\textit{left}\text{ (Subject)}}$ criticizes *the government*$_{\text{right (Object)}}$
- Czech: **morphological case** (nominative/other)
  - *EU*$_{\textit{nominative}\text{ (Subject)}}$ kritizuje *vládu*$_{\textit{accusative}\text{ (Object)}}$
  - *vládu*$_{\textit{accusative}\text{ (Object)}}$ kritizuje *EU*$_{\textit{nominative}\text{ (Subject)}}$
  - *EU*$_{\textit{nominative}\text{ (Subject)}}$ *vládu*$_{\textit{accusative}\text{ (Object)}}$ kritizuje
  - *vládu*$_{\textit{accusative}\text{ (Object)}}$ *EU*$_{\textit{nominative}\text{ (Subject)}}$ kritizuje
  - (position may mark topic-focus articulation, stress...)

Valency of *criticize* (*kritizovat*)

- *EU*$_{\text{subject}}$ *criticizes the Greek government*$_{\text{object}}$
- *EU*$_{\text{nominative}}$ *kritizuje řeckou vládu*$_{\text{accusative}}$

Valency of *criticize* (*kritizovat*)

- *EU*$_{\text{subject}}$ *criticizes the Greek government*$_{\text{object}}$
- *EU*$_{\text{nominative}}$ *kritizuje řeckou vládu*$_{\text{accusative}}$
- a valency frame of a verb

subject	*criticize*	object
nominative	*kritizovat*	accusative

Valency of *criticize* (*kritizovat*)

- *EU*$_{\text{subject}}$ *criticizes the Greek government*$_{\text{object}}$
- *EU*$_{\text{nominative}}$ *kritizuje řeckou vládu*$_{\text{accusative}}$
- a valency frame of a verb
  - subject *criticize* object (position)
  - nominative *kritizovat* accusative (cases)

Valency of *criticize* (*kritizovat*)

- *EU*$_{\text{subject}}$ *criticizes the Greek government*$_{\text{object}}$
- *EU*$_{\text{nominative}}$ *kritizuje 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*)
  - (*to criticize,* Object) ~ (*kritizovat,* accusative)

Correction approach: rule-based?

- rule-based post-editing successful for many types of errors in English-to-Czech translation
  - morphological agreement, verb tenses, possessive constructions, passive constructions, negation...
  - in → SMT → Depfix → out (watch Monday seminar)
  - easy to do using Czech positional tagset & analysis of Czech and English to a-trees/t-trees in Treex

Correction approach: rule-based?

- rule-based post-editing successful for many types of errors in English-to-Czech translation
  - morphological agreement, verb tenses, possessive constructions, passive constructions, negation...
  - in → SMT → Depfix → out (watch Monday seminar)
  - easy to do using Czech positional tagset & analysis of Czech and English to a-trees/t-trees in Treex
- hard to fully cover valency by a set of rules
  - we need it parallelly for English and Czech
  - possible future work: use existing valency lexicons

Correction approach: statistical?

- statistical machine translation (SMT) works well
- statistical post-editing of rule-based machine translation (RBMT) outputs works well
    - in → RBMT → SMT → out (Simard et. al., 2007)
- statistical post-editing of SMT outputs
    - in → SMT → SMT → out
    - works for English-to-Turkish (Oflazer et. al., 2007)
    - works for French-to-English (Béchara et. al., 2011)

Correction approach: statistical?

- statistical machine translation (SMT) works well
- statistical post-editing of rule-based machine translation (RBMT) outputs works well
  - in → RBMT → SMT → out (Simard et. al., 2007)
- statistical post-editing of SMT outputs
  - in → SMT → SMT → out
  - works for English-to-Turkish (Oflazer et. al., 2007)
  - works for French-to-English (Béchara et. al., 2011)
  - does not work for English-to-Czech (Tamchyna)

Correction approach: combine!

- capturing valency by rules – not good
  - ➔ let's use statistics for that!
- simple statistical post-editing of SMT – not good
  - ➔ let's get some inspiration from the linguistically motivated rule-based approaches!
- Step 1: analyze the sentences in Treex
  - linguistically motivated, combines rules and statistics
- Step 2: correct valency with a statistical model
  - simple statistics, but operating on the t-layer

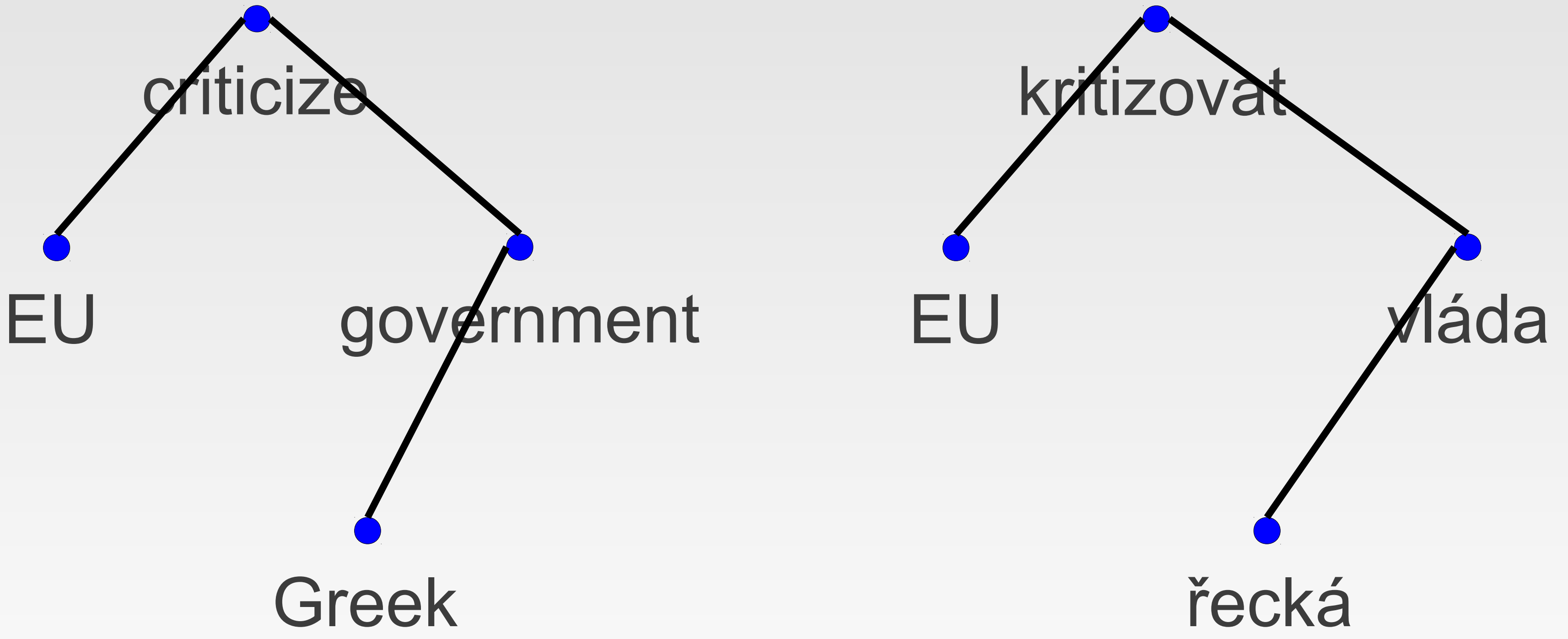
Sentence analysis (DEEP)

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

Deep syntactic dependency trees

EU criticizes
the Greek government

EU kritizuje
řecká vláda



criticize

EU

government

Greek

kritizovat

EU

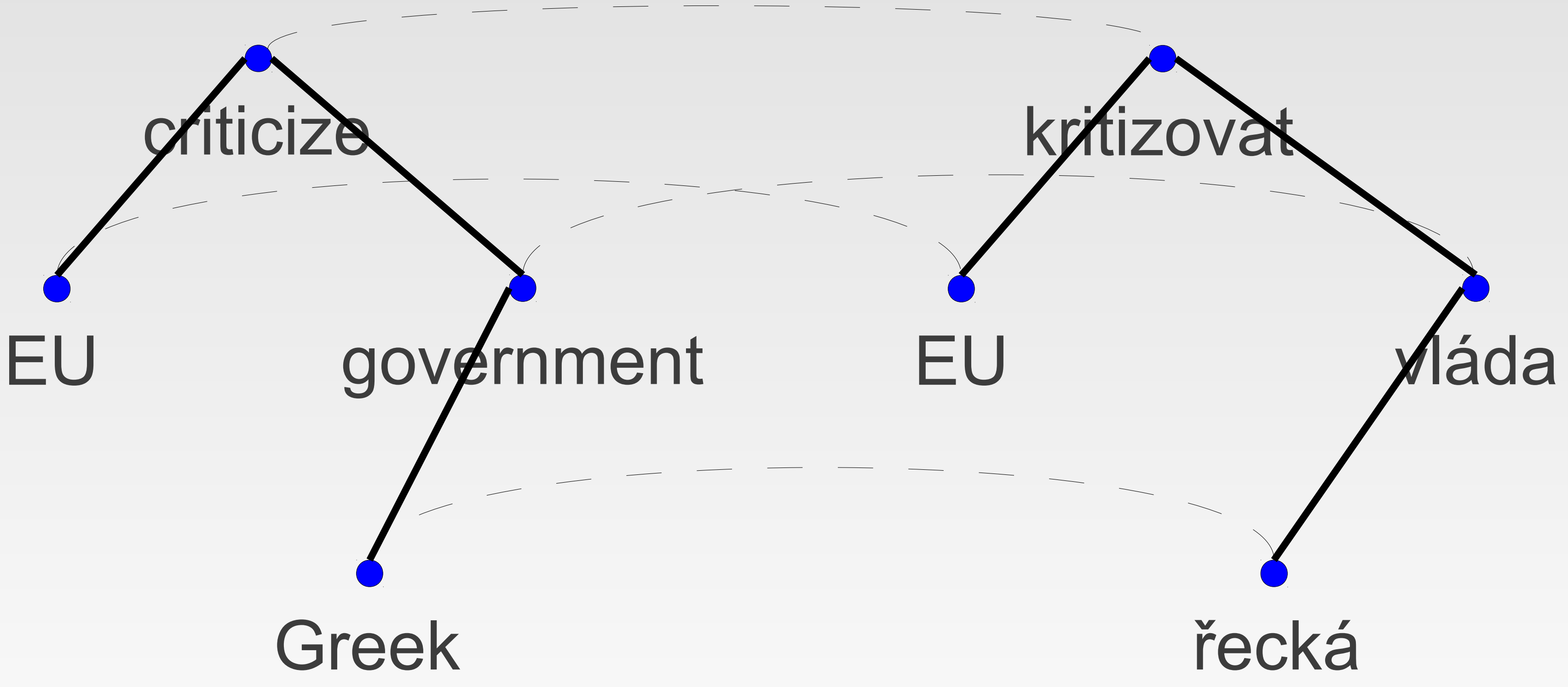
vláda

řecká

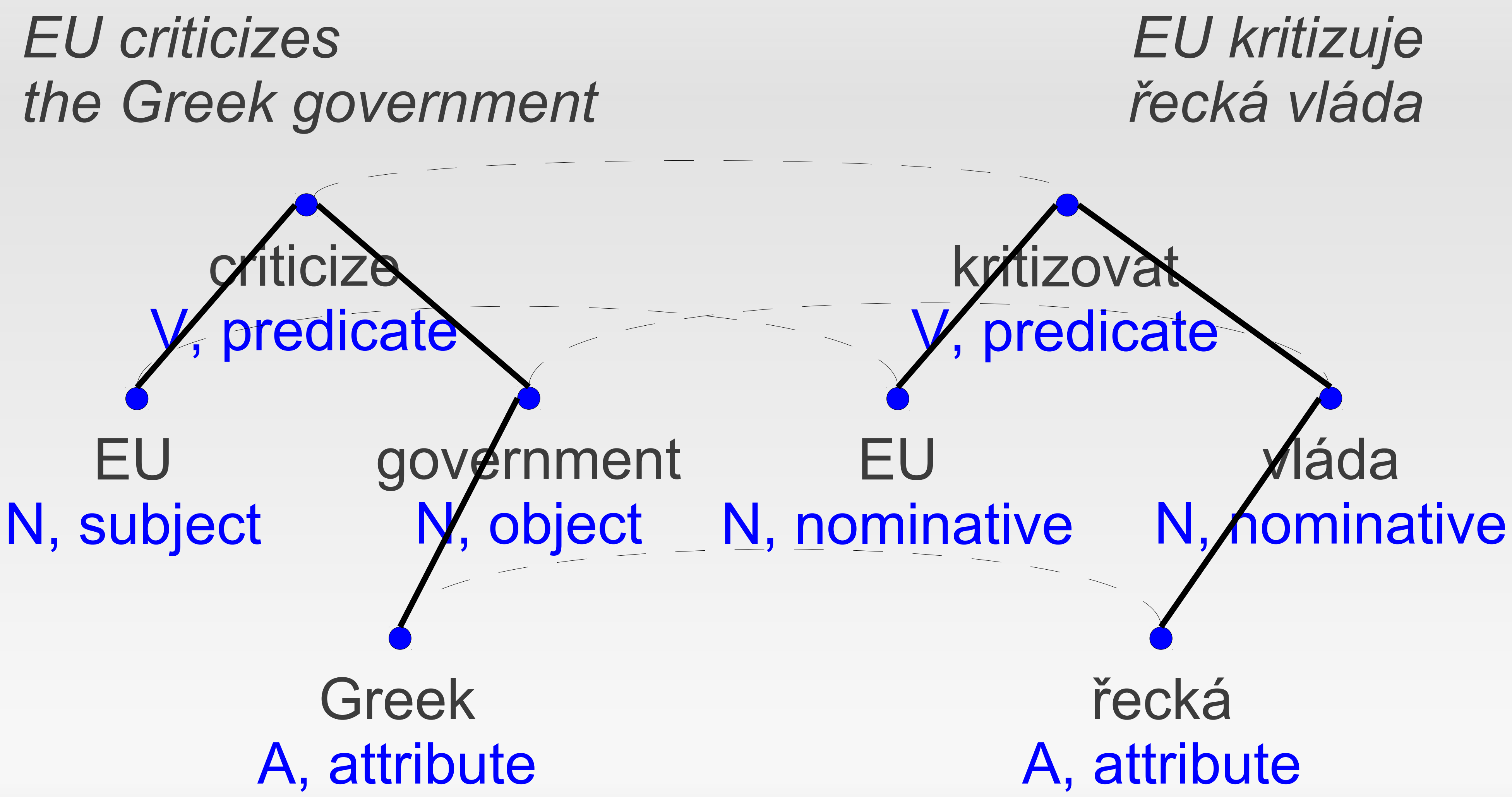
Deep syntactic dependency trees

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

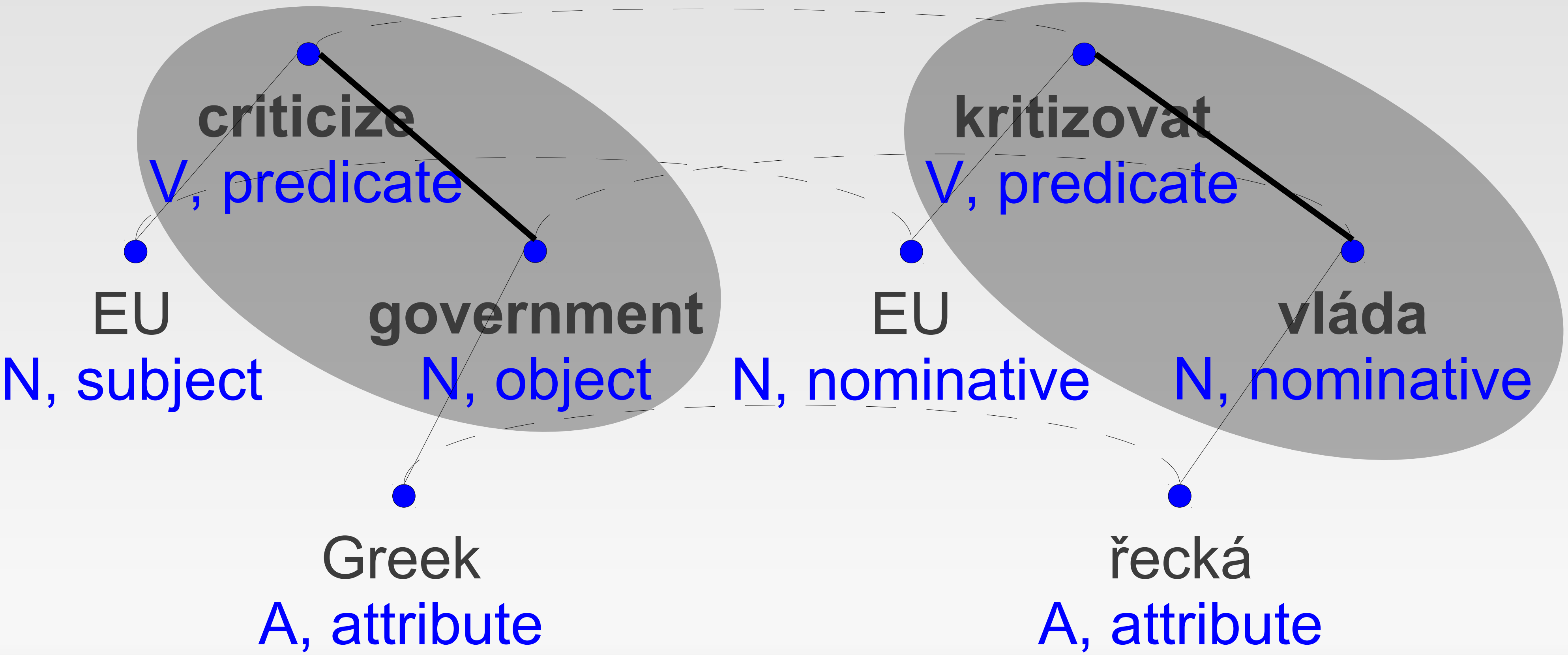
vláda
N, nominative

řecká
A, attribute

(head, arg) pair identification

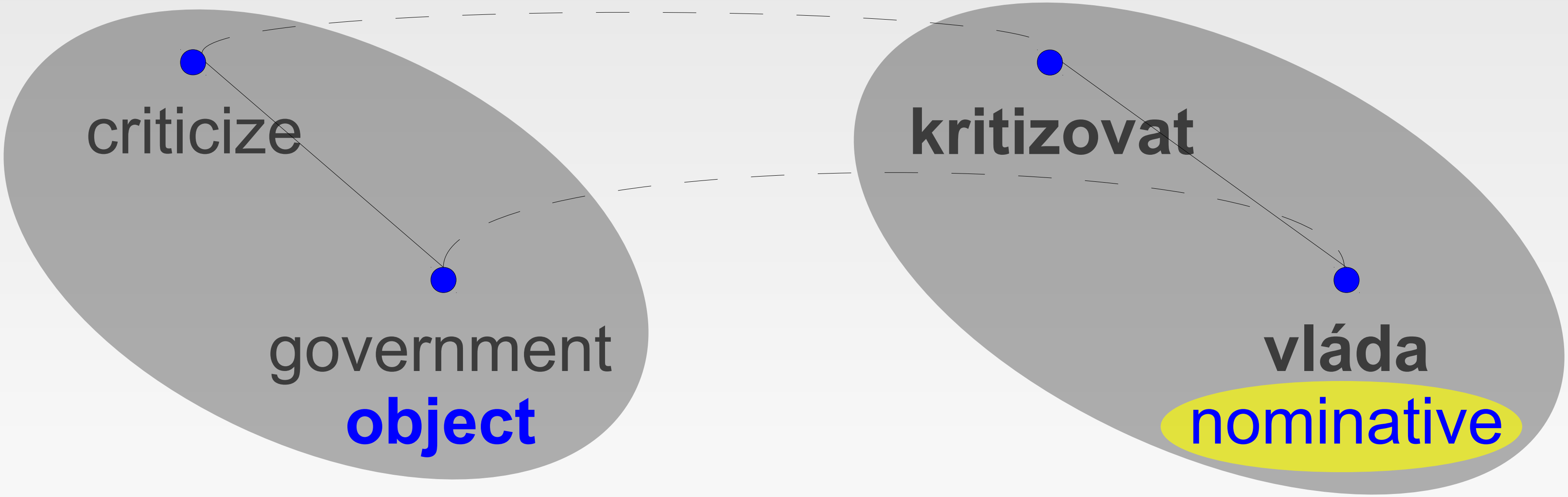
*EU **criticizes** the Greek **government***

*EU **kritizuje** řecká **vláda***



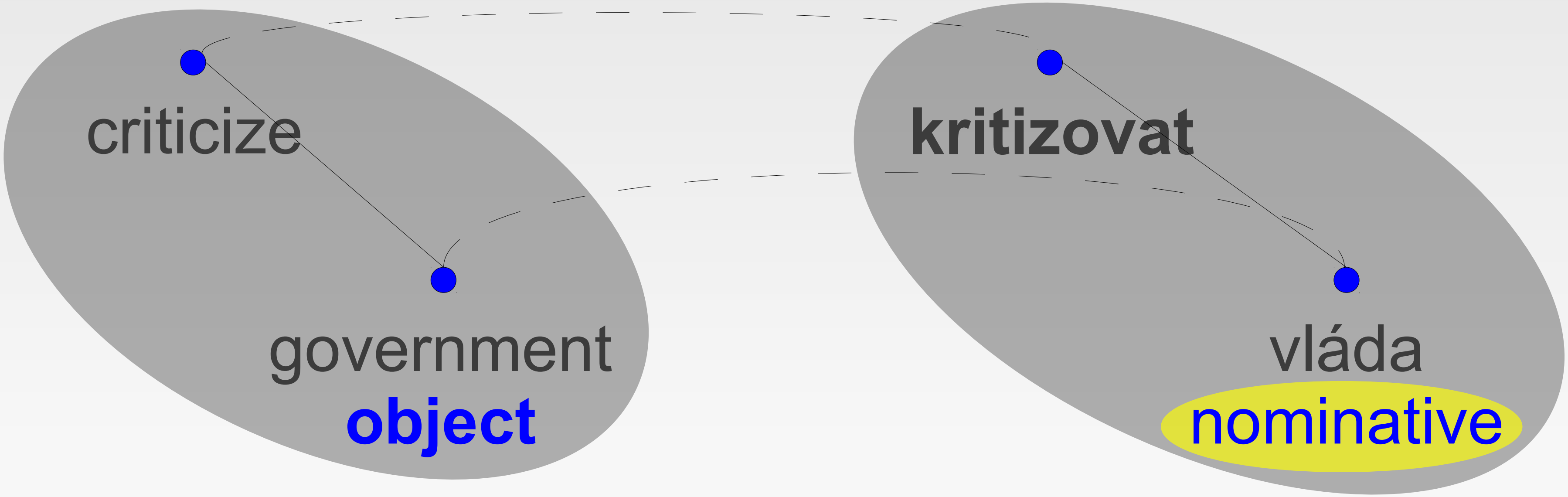
Valency models (FIX)

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



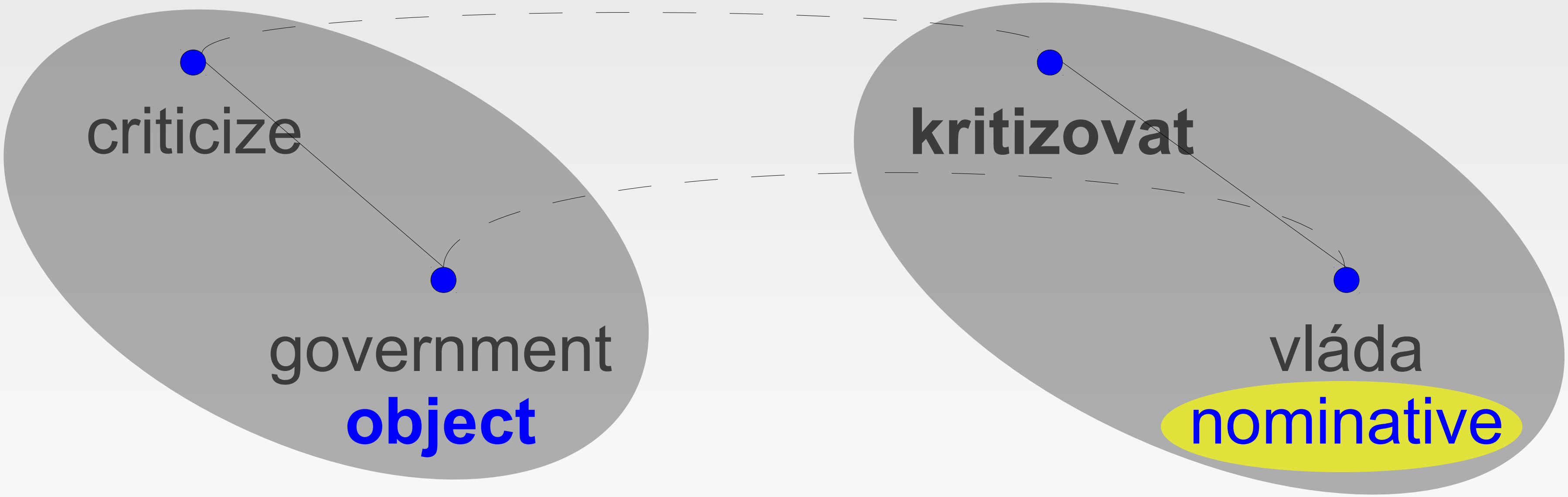
Argument case probabilities

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



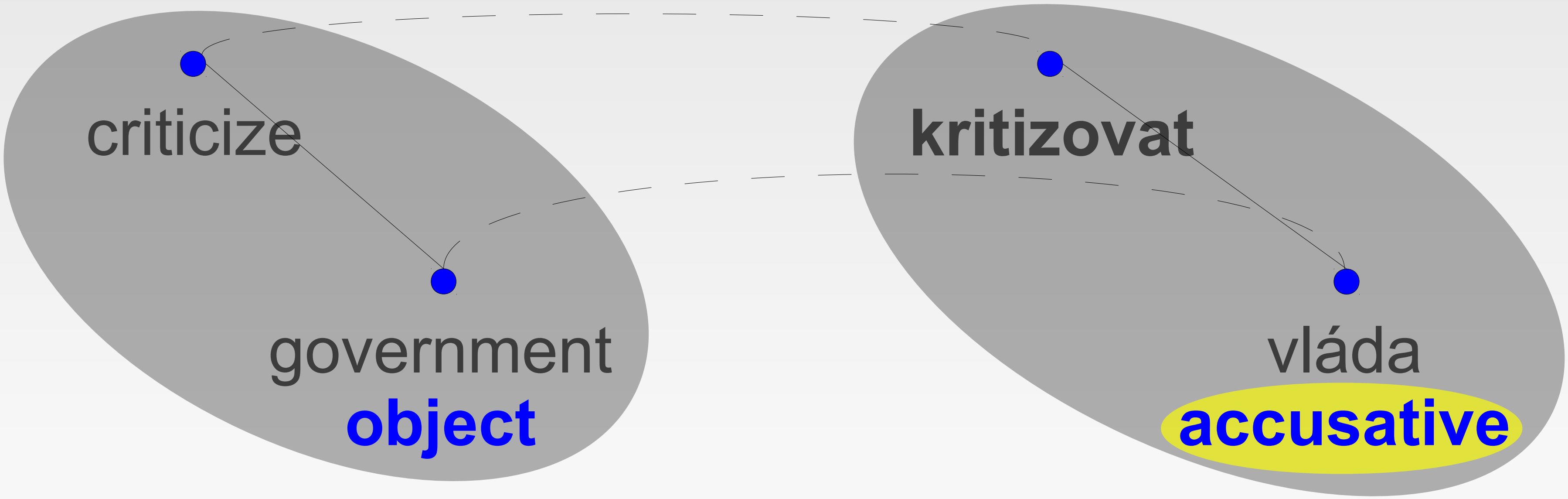
Argument case probabilities

- 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***$_{\textbf{nominative}}$ ***vláda***$_{\textbf{nominative}}$

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

Sentence correction

- Statitical machine translation output:

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

  - *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***$_{\textbf{nominative}}$ ***vláda***$_{\textbf{nominative}}$

  - *Not only* ***the Greek government*** *criticizes EU*
- Valency model correction:

  EU kritizuje nejen ***řecká***$_{\textbf{nominative}}$ ***vládu***$_{\textbf{accusative}}$
- Agreement enforcement:

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

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

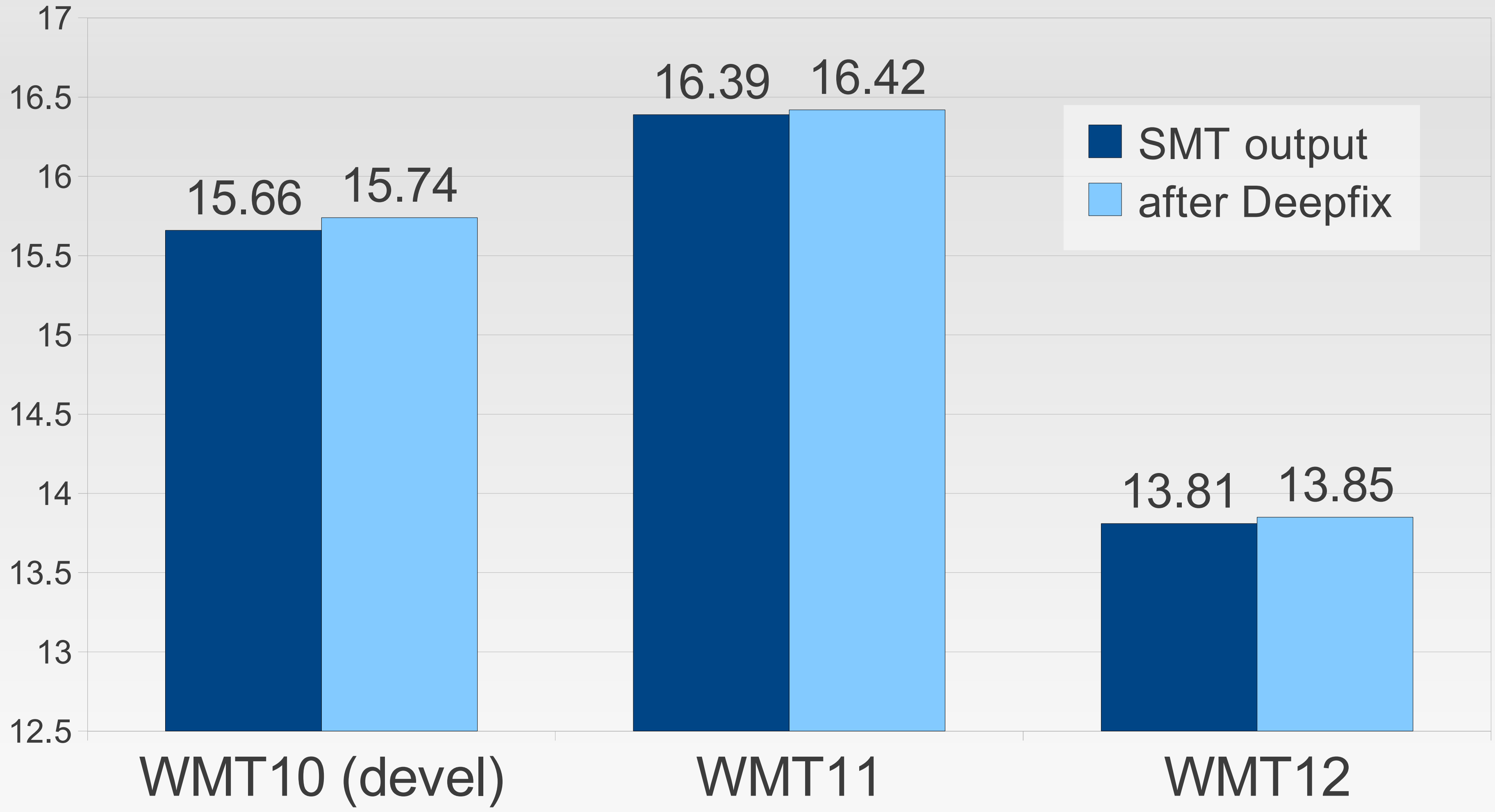
Some interesting details

- the model actually works on formemes
  - functions (EN), cases (CS), **prepositions** (EN, CS)
  - in: *The government spends on the middle schools.*
  - SMT: *Vládá utrácí střední školy.*
    - (spend, on+X) → (utrácet, 4)   P = 0.07
    - *The government destroys the middle schools.*
  - out: *Vládá utrácí za střední školy.*
    - (spend, on+X) → (utrácet, za+4)   P = 0.89
    - *The government spends on the middle schools.*

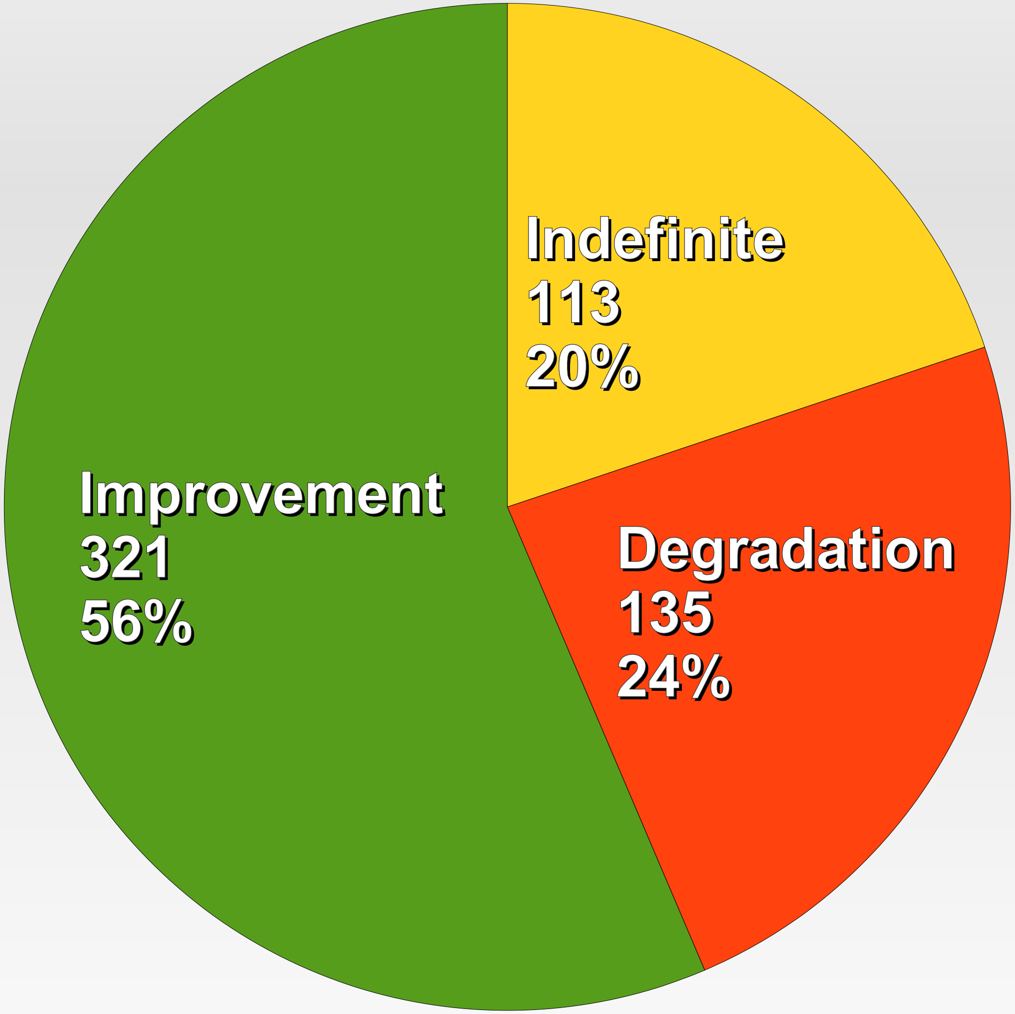
Some interesting details

- the model actually works on formemes
  - functions (EN), cases (CS), **prepositions** (EN, CS)
  - in: *The government spends on the middle schools.*
  - SMT: *Vládá utrácí střední školy.*
    - (spend, on+X) → (utrácet, 4) P = 0.07
    - *The government destroys the middle schools.*
  - out: *Vládá utrácí za střední školy.*
    - (spend, on+X) → (utrácet, za+4) P = 0.89
    - *The government spends on the middle schools.*
- we model both verb valency and noun valency

Automatic evaluation (BLEU)



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

Possible future work

- 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
- explore existing valency lexicons

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/