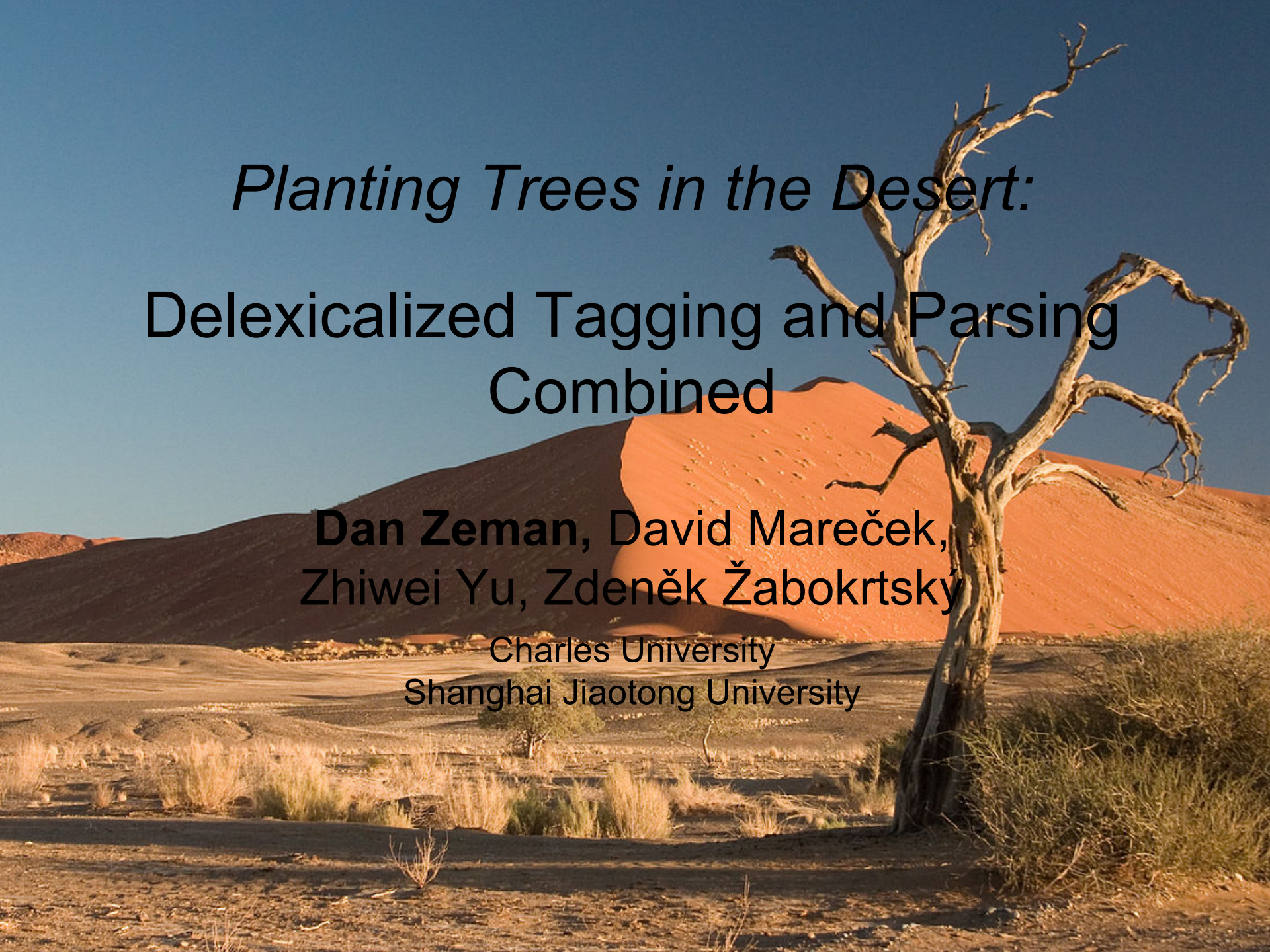


Planting Trees in the Desert:

Delexicalized Tagging and Parsing Combined

Dan Zeman, David Mareček, Zhiwei Yu, Zdeněk Žabokrtský

Charles University
Shanghai Jiaotong University

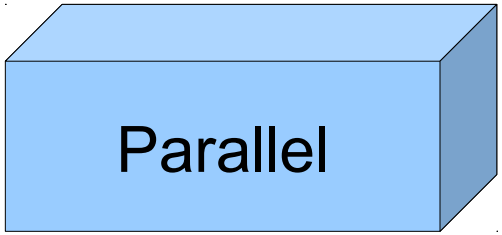
Planting Trees in the Desert:

Delexicalized Tagging and Parsing Combined

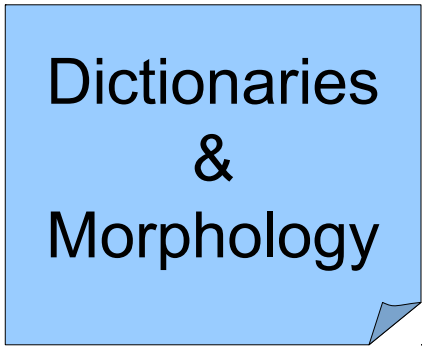
Dan Zeman, David Mareček,
Zhiwei Yu, Zdeněk Žabokrtský

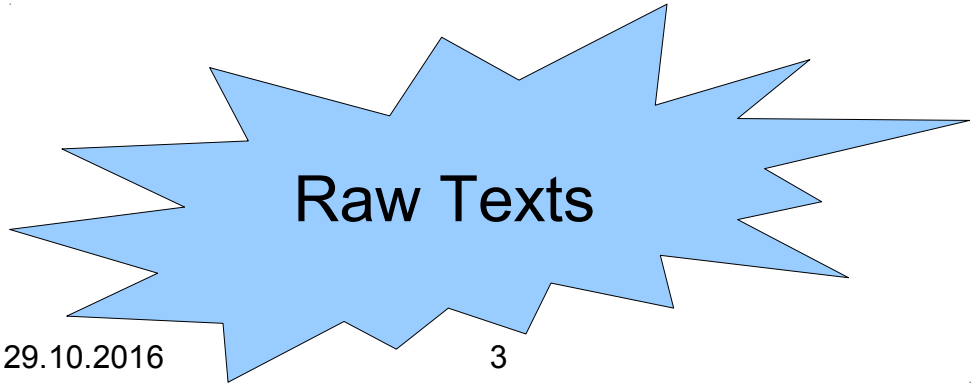
Charles University
Shanghai Jiaotong University

Language Resources

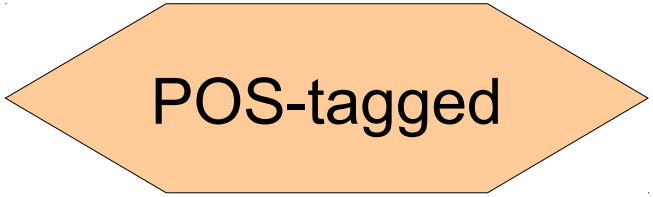


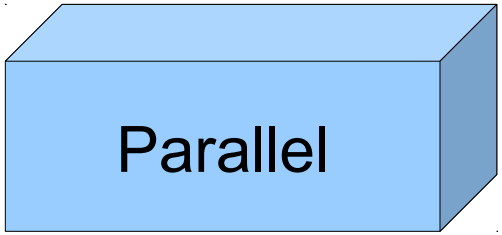
7,097 languages
398 with > 1M speakers



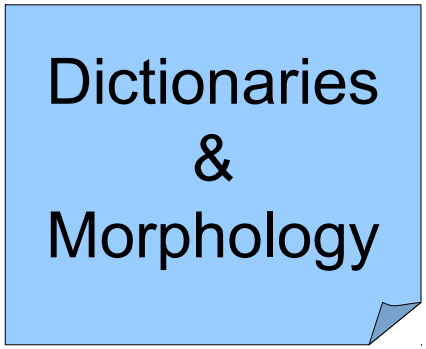


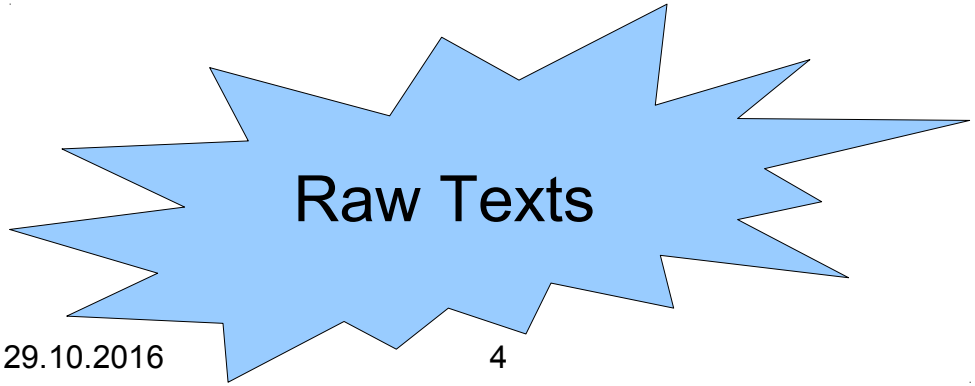
Resource-Rich Languages





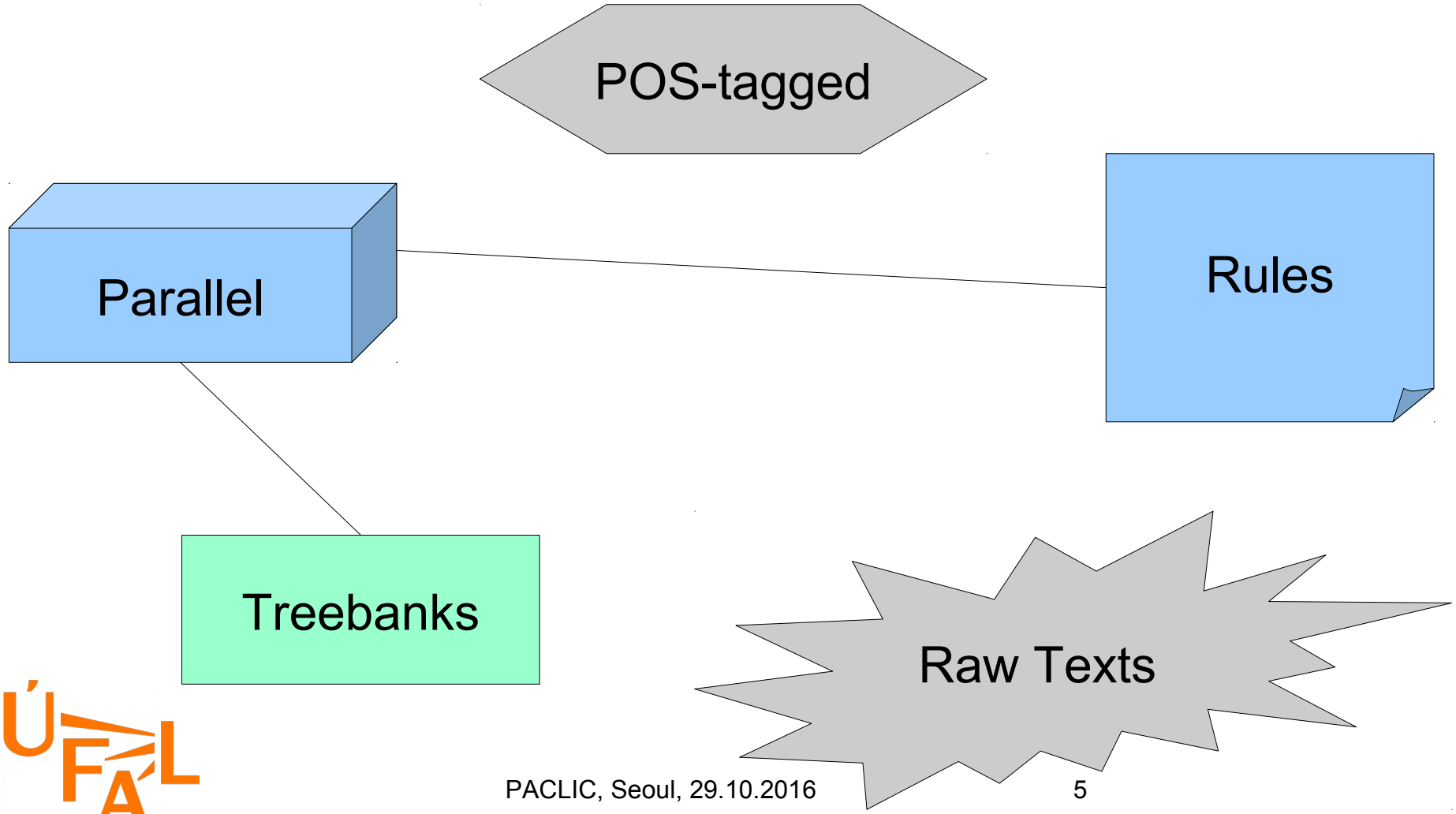
61 languages
105 treebanks
(that I know of)



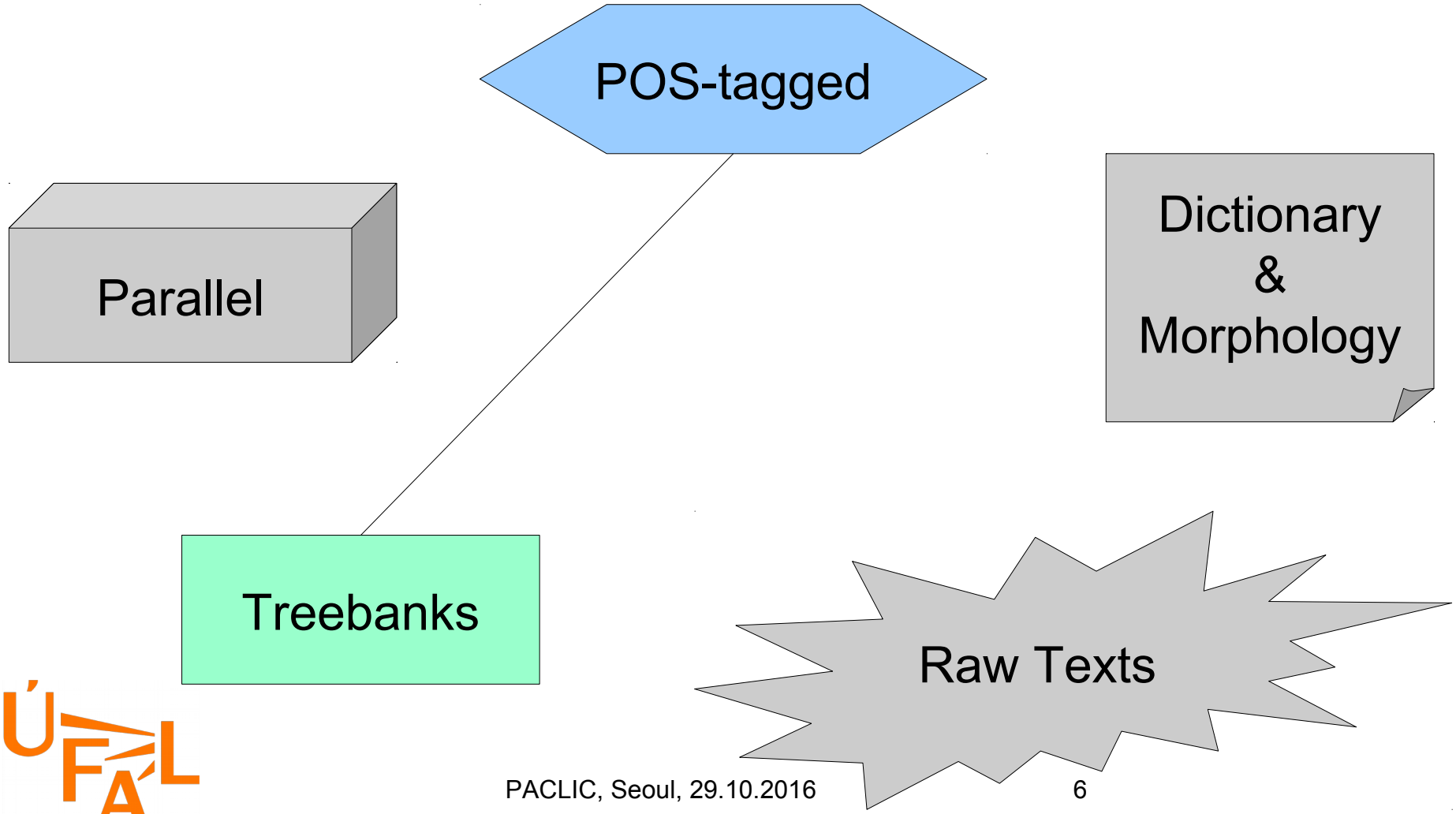


Cross-Lingual Transfer

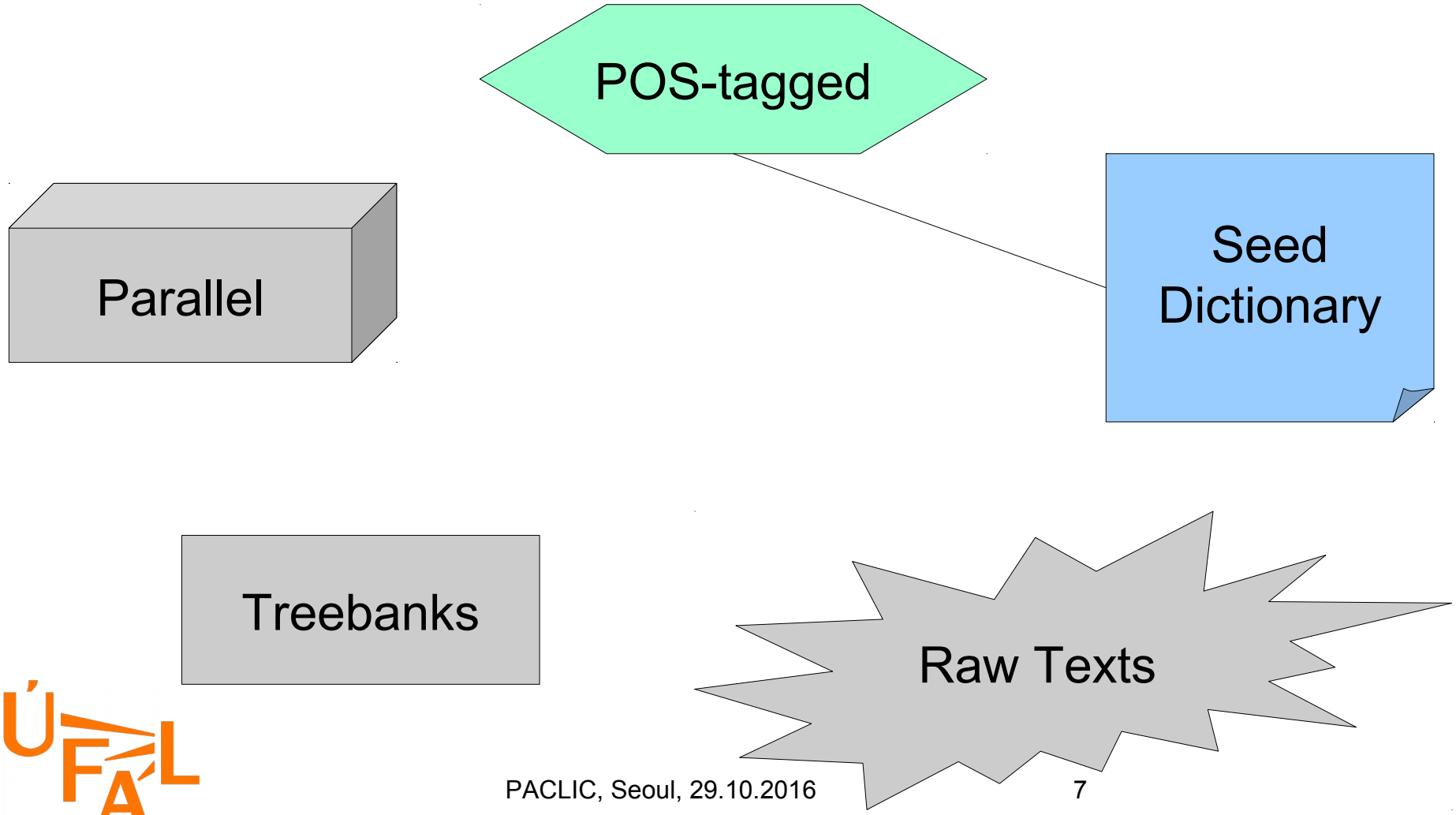
Hwa et al. (2004)



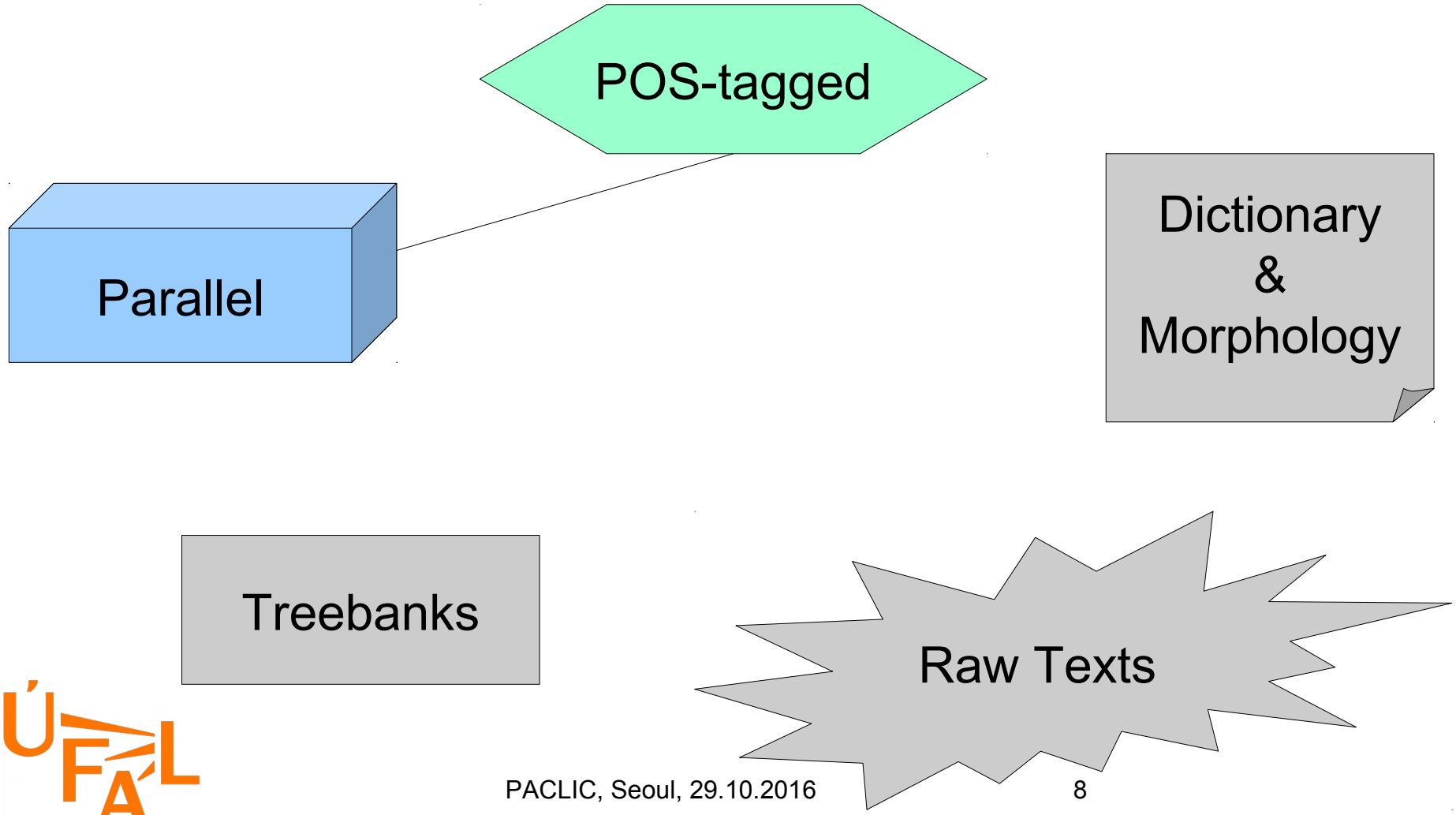
Zeman & Resnik (2008)
Delexicalized Parsing



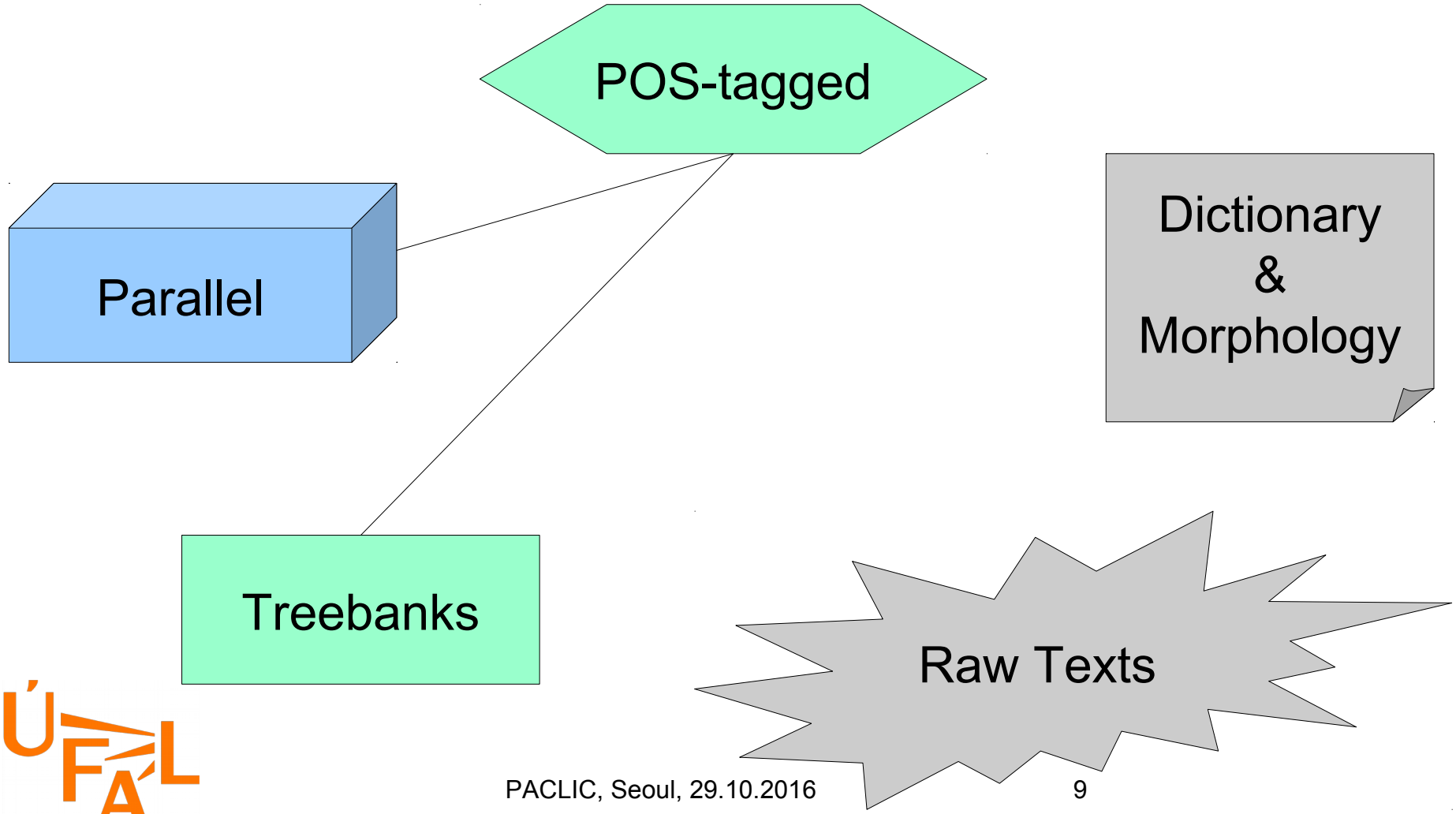
Cucerzan & Yarowsky (2002)



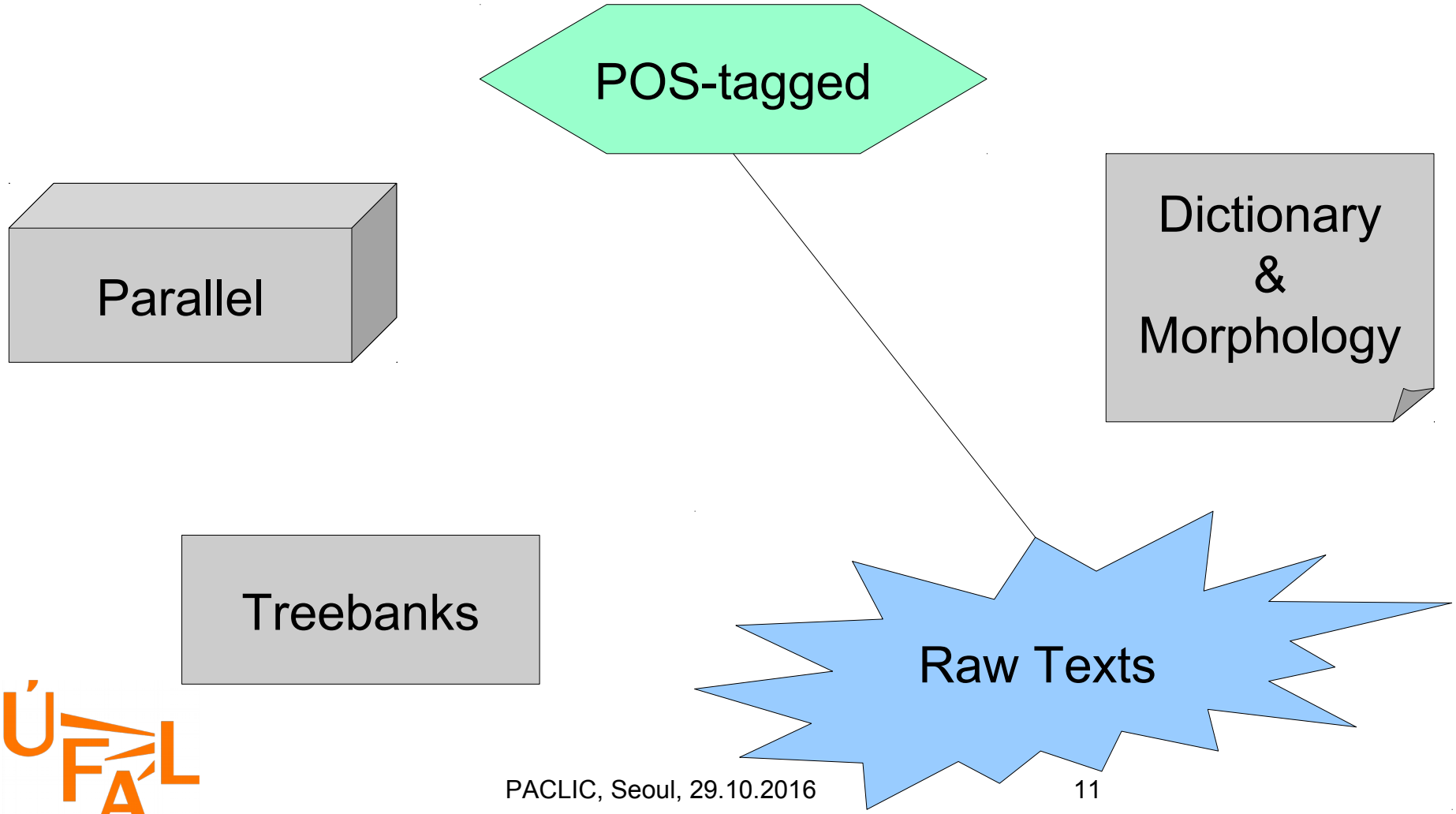
Yarowsky & Ngai (2001)



McDonald et al. (2011)



Yu et al. (LREC 2016)
Delexicalized Tagging



Delexicalized Parsing

- What if we feed the parser with tags instead of words?
  - *Ændringer i listen i bilaget offentliggøres og meddeles på samme måde.*
  - `NNS IN NN IN NN VB CC VB IN DT NN`
  - `NNS IN NN MD VB CC VB IN DT NN`
  - *Förändringar i förteckningen skall offentliggöras och meddelas på samma sätt.*

Delexicalized Parsing

- What if we feed the parser with tags instead of words?
  - *Ændringer i listen i bilaget offentliggøres og meddeles på samme måde.*
  - `((NNS (IN NN (IN NN))) ((VB CC VB) (IN (DT NN))))`
  - `((NNS (IN NN)) ((MD (VB CC VB)) (IN (DT NN))))`
  - *Förändringar i förteckningen skall offentliggöras och meddelas på samma sätt.*

Treebank Normalization

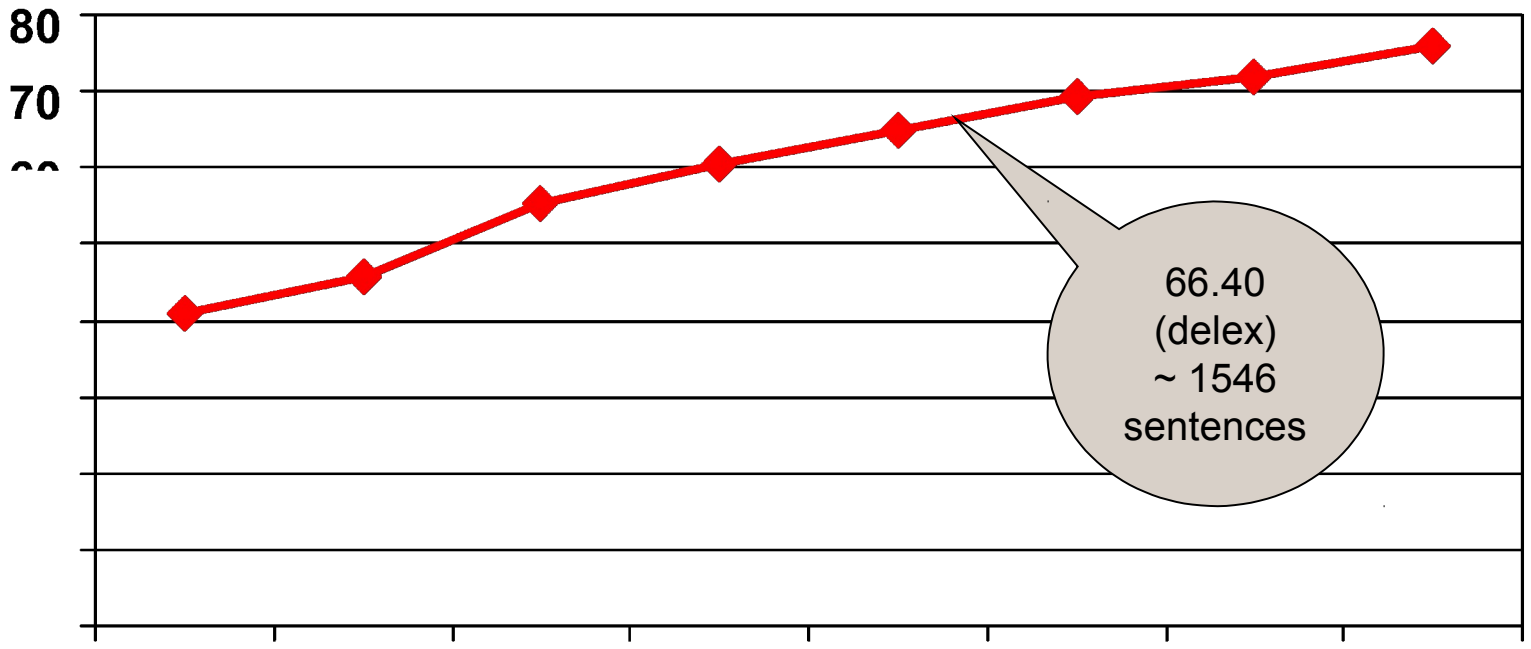
Danish

- DET governs ADJ, ADJ governs NOUN
- NUM governs NOUN
- GEN governs NOM
  Ruslands vej
  Russia's way
- COORD: last member on conjunction, everything else on first member

Swedish

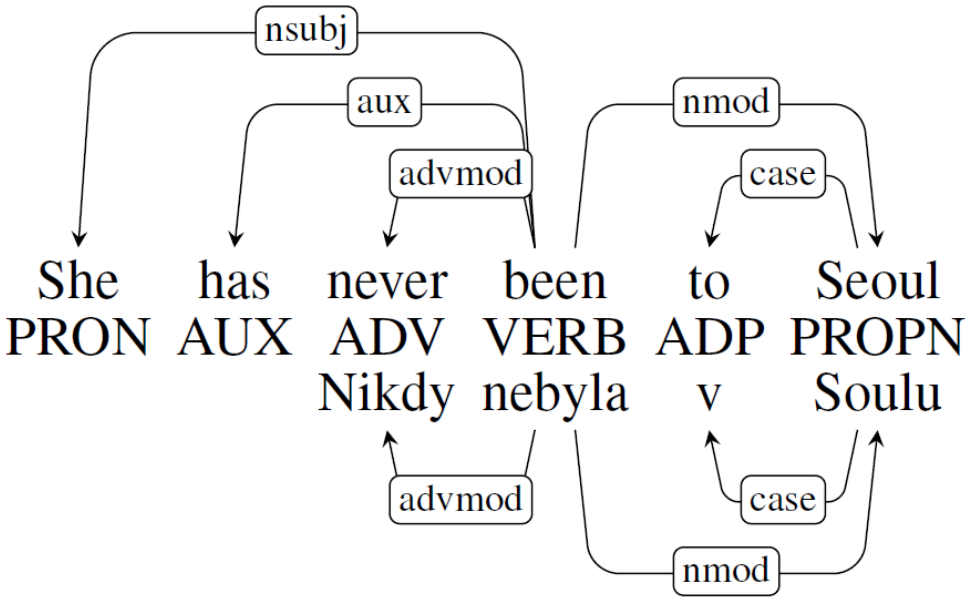
- NOUN governs both DET and ADJ
- NOUN governs NUM
- NOM governs GEN
  års inkomster
  year's income
- COORD: member on previous member, commas and conjs on next member

How Big Swedish Treebank Yields Similar Results?



Present Work

- Dependency (instead of constituency)
- Universal Dependencies
  - Same annotation style!
  - Many languages

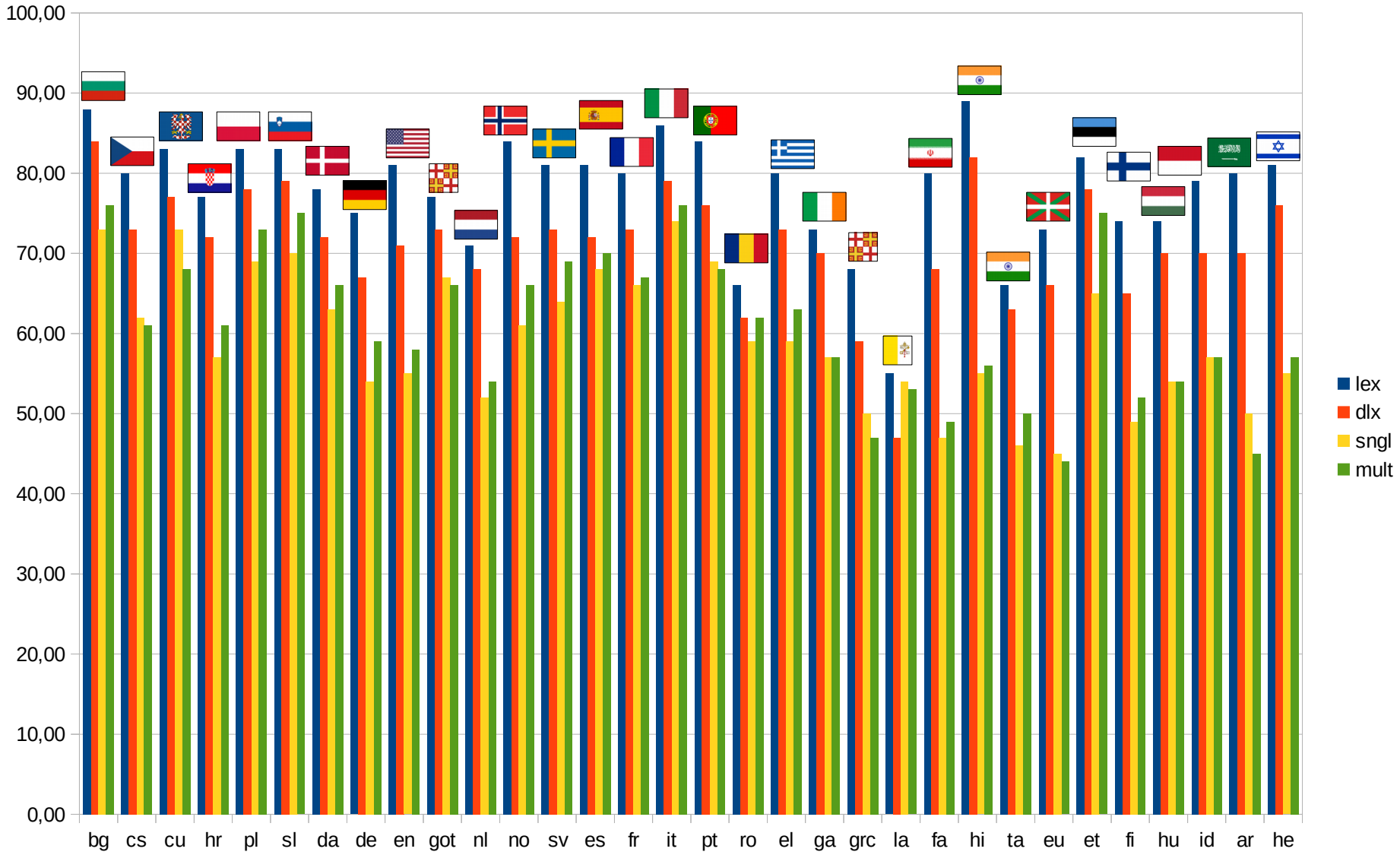


Default Setup

- Malt Parser, stack-lazy algorithm
  - same configuration for all, no optimization
  - same selection of training features for all treebanks

- Trained on the first 5000 sentences only
- Tested on the whole test set
- Default score: UAS

Malt Trained on 5000 Sents.



Who Helps Whom?

- Czech (80) ⇦ Croatian (62), Bulgarian (60)
- Polish (83) ⇦ Slovenian (69), Croatian (67)
- Croatian (77) ⇦ Slovenian (57), Czech (55)
- Slovenian (83) ⇦ Czech (70), Croatian (65)
- Bulgarian (88) ⇦ Slovenian (73), Czech (72)
- Church Slavonic (83) ⇦ Gothic (73), Ancient Greek (69)

Who Helps Whom?

- Italian (86) ⇦ Spanish (74), French (72)
- French (80) ⇦ Spanish (66), Italian (65)
- Spanish (81) ⇦ Italian (68), French (65)
- Portuguese (84) ⇦ Italian (69), Spanish (69)
- Romanian (66) ⇦ Italian (59), Indonesian (58)

Who Helps Whom?

- Swedish (81) ⇦ Norwegian (64), Danish (63)
- Danish (78) ⇦ Norwegian (63), Bulgarian (61)
- Norwegian (84) ⇦ Swedish (61), Croatian (61)
- English (81) ⇦ Swedish (55), German (54)
- German (75) ⇦ Swedish (54), Slovenian (53)
- Dutch (71) ⇦ German (52), Portuguese (52)

Delexicalized Tagging

- Language-independent features?
- Not fully unsupervised tagging!
- We still want “real” POS tags!

Universal POS Tagset

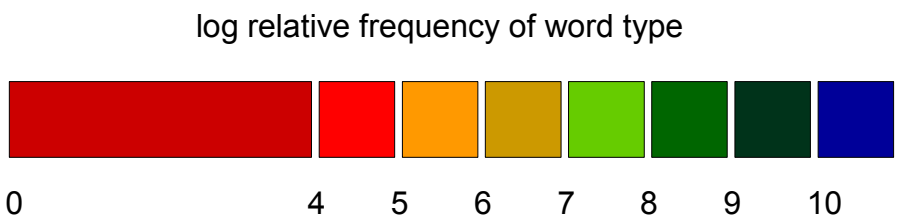
- NOUN … common noun
- PROPN … proper noun
- ADJ … adjective
- VERB … main verb
- ADV … adverb

- SYM … symbol
- PUNCT … punctuation
- X … unknown

- PRON … pronoun
- DET … determiner
- NUM … numeral
- AUX … auxiliary verb
- ADP … adposition
- CONJ … coord. conj.
- SCONJ … subord. conj.
- PART … particle
- INTJ … interjection

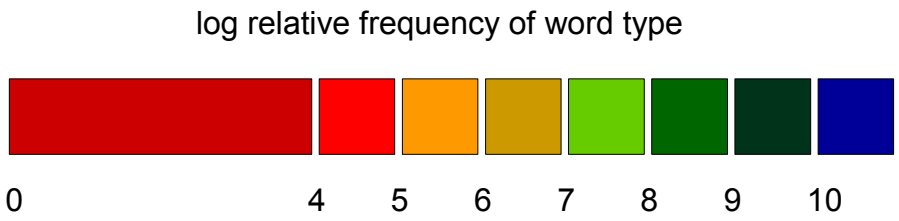
Log Frequency: English

- *These people are looking for you, Sam!*
- *President nominated two jurists in the area.*



Log Frequency: Italian

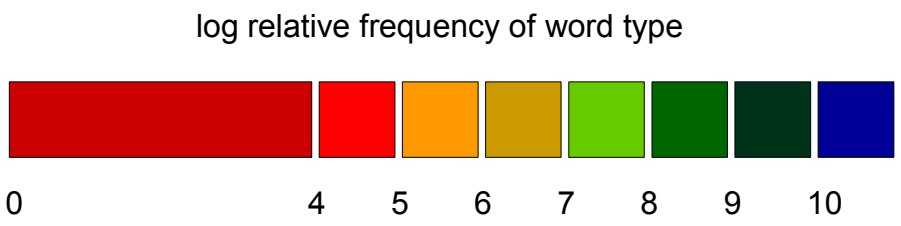
- *Dal '93 dirige il Festival di Taormina.*
- *I tre avevano da poco lasciato la cima e stavano cominciando la discesa.*



Log Frequency: Slovenian

- *Pri tem moram izpostaviti odgovornost za opravljeno delo.*

- *Pogrešajo predstavništvo, ki bi v Albaniji zastopalo njihove interese.*



Log Frequency: English

- OVERALL => NOUN (20%) VERB (15%) PRON (10%) ADP (10%) …

- 0-4 => DET (43%) ADP (17%) CONJ (17%) PRON (12%) PART (11%) …
- 4-5 => ADP (40%) PRON (28%) VERB (15%) …
- 5-6 => PRON (25%) AUX (18%) VERB (14%) …
- 6-7 => VERB (18%) ADV (18%) PRON (14%) NOUN (11%) …
- 7-8 => NOUN (29%) VERB (21%) ADV (13%) ADJ (13%) …
- 8-9 => NOUN (40%) VERB (20%) ADJ (15%) PROPN (12%) …
- 9-10 => NOUN (42%) VERB (23%) PROPN (15%) ADJ (12%) …
- ≥ 10 => NOUN (39%) PROPN (22%) VERB (18%) ADJ (11%) …

Word Length: en+it+sl

- *These people are looking for you, Sam!*
- *President nominated two jurists in the area.*
- *Dal '93 dirige il Festival di Taormina.*
- *I tre avevano da poco lasciato la cima e stavano cominciando la discesa.*
- *Pri tem moram izpostaviti odgovornost za opravljeno delo.*
- *Pogrešajo predstavništvo, ki bi v Albaniji zastopalo njihove interese.*

Word Length: en+it+sl

- OVERALL => NOUN (22%) VERB (13%) ADP (13%) DET (12%) …

- 1 => ADP (33%) DET (24%) CONJ (17%) PRON (12%) …
- 2 => ADP (35%) DET (24%) PRON (11%) …
- 3 => DET (23%) PRON (14%) ADP (11%) CONJ (10%) …
- 4 => NOUN (24%) VERB (17%) ADV (11%) …
- 5 => NOUN (35%) VERB (14%) ADJ (11%) PROPN (10%) …
- 6 => NOUN (37%) VERB (17%) PROPN (12%) ADJ (12%) …
- 7 => NOUN (42%) VERB (21%) ADJ (13%) PROPN (12%) …
- ≥ 8 => NOUN (43%) VERB (22%) ADJ (21%) …

Left and Right Neighborhood

- [en] lH << rH: *been, own, be, few, lot, same*
- [en] lH >> rH: *thank, if, because, I, let, when*

- [it] lH << rH: *suoi, sue, stata, sua, cui, su*
- [it] lH >> rH: *qual, fino, durante, repubblica, mondo*

- [sl] lH << rH: *tem, bila, bili, bil, bilo, jih*
- [sl] lH >> rH: *ki, ko, kjer, saj, vendar, če*

17 Features

- length
- log frequency

- is number
- is punctuation
- log freq after number
- log freq after punctuation

- entropy of suffixes

- number of left word types
- number of right word types
- number of subst. word types

- left entropy
- right entropy
- substituting word entropy

- weighted sum of pointwise mutual information (PMI) left
- ditto right
- PMI with most freq. word left
- ditto right

Delexicalized Transfer

- Source language: raw corpus & gold-tagged corpus
- Target language: raw corpus

- Compute features:
  - first 20M tokens from W2C for each language
  - need features for both source and target language
- Train classifer on source language(s):
  - first 30K tokens from Universal Dependencies 1.2
- Apply it to target language.

Results

- Tested on 28 languages from UD 1.2
- Baseline: 20% – 49% (average 36%)
- SVM:
  - Source = target language: 69% – 89% (average 82%)
  - Best foreign source mix: 35% – 79% (average 59%)

  - Yu et al. on easier data, 7 source languages:
    43% – 70% (average 60%)

Delex Tagging + Parsing

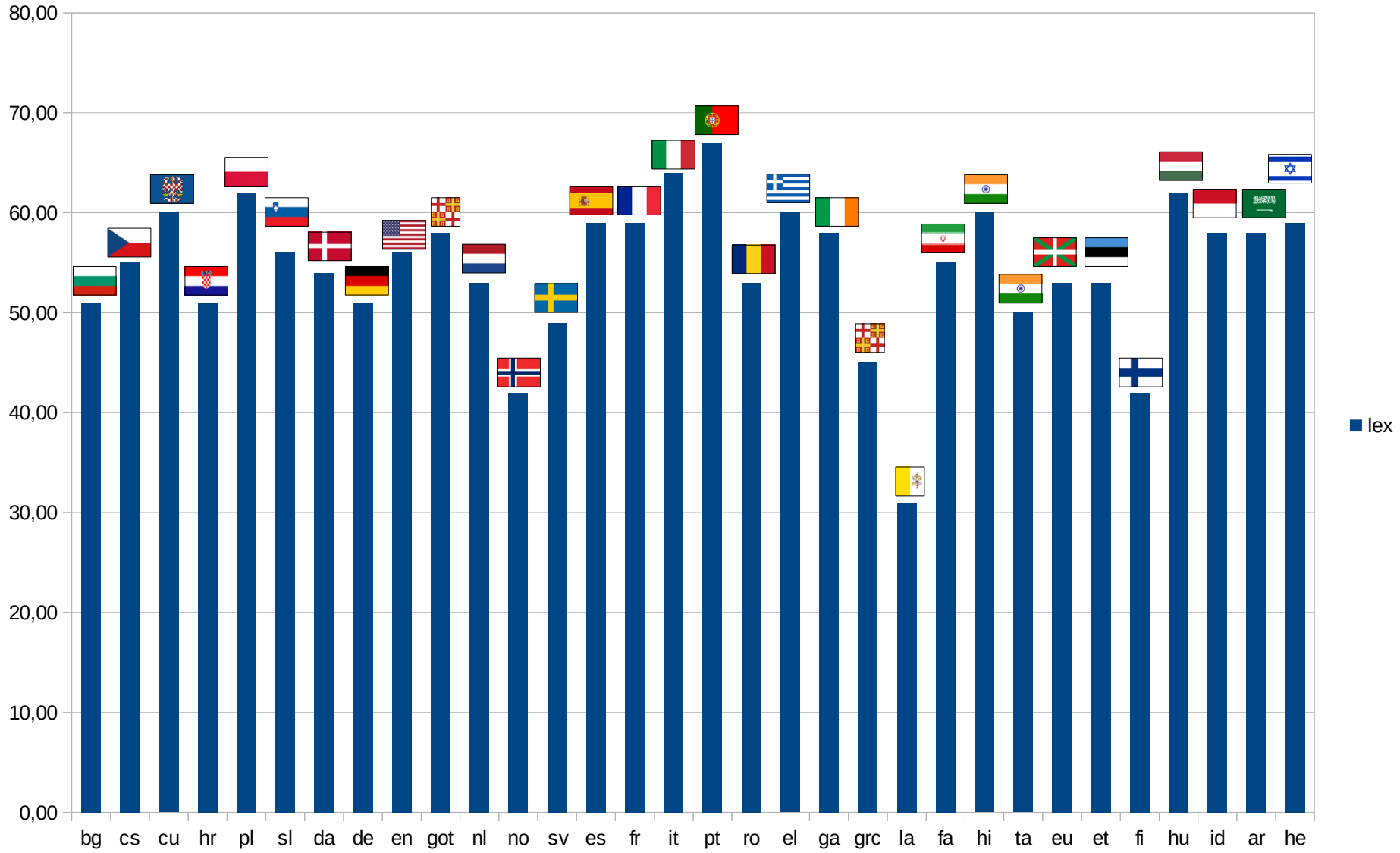
- Does not work (<20%)
- Why?
  - Distribution of error types … too random

Delex Tagging + Parsing

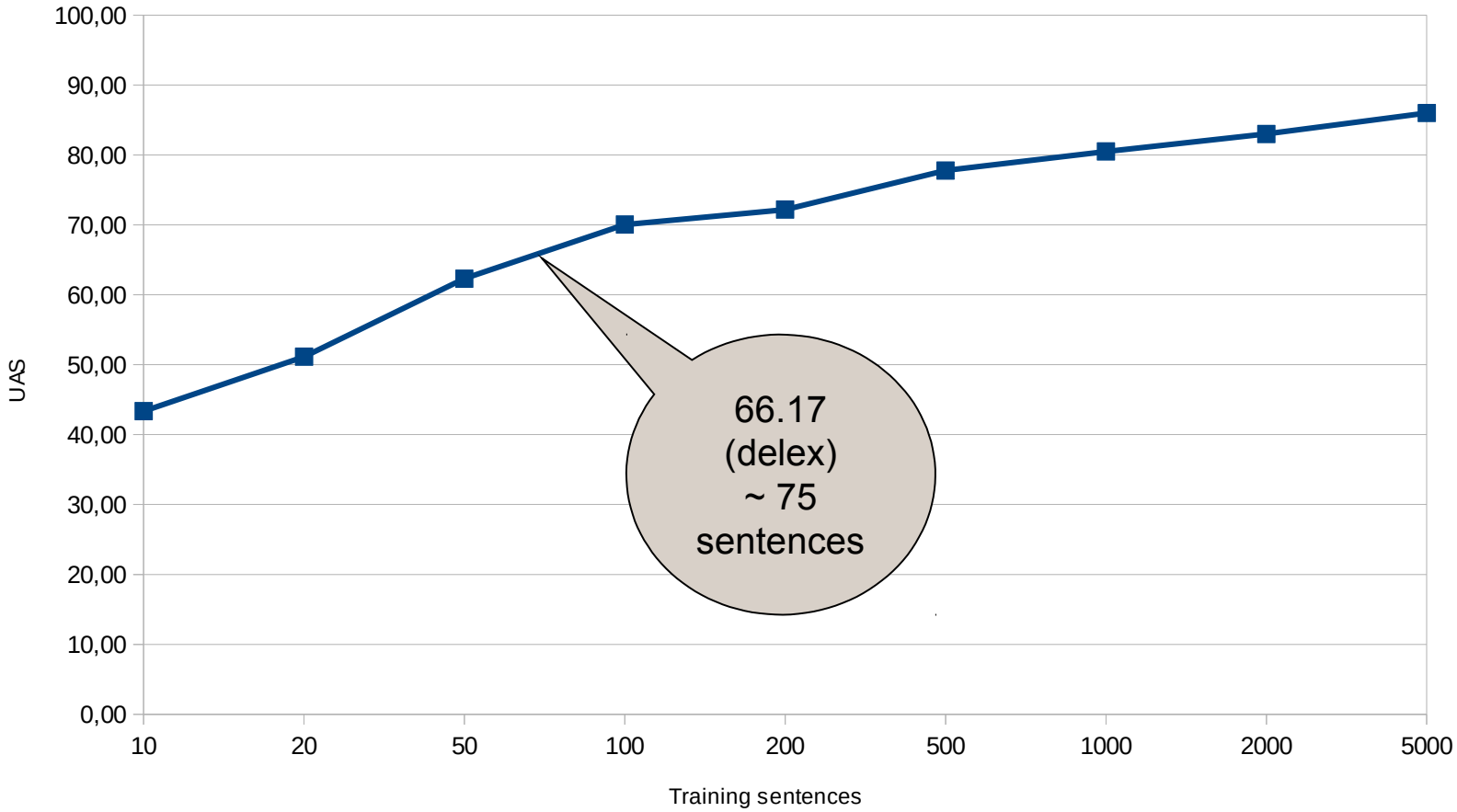
- Does not work (<20%)

- Why?
  - Distribution of error types … too random
  - Function vs. content words … OK
  - NOUN vs. VERB … very bad

Malt Trained on 20 Sentences



Learning Curve



Conclusion

- Tagging: accuracy misleading
- Parsing: 60 – 70% sounds not bad, but …
- Tagging + parsing: useless

- Native speaker available?

Hire him!


thank you
고맙습니다