Training parsers for low-resourced languages: improving cross-lingual transfer with monolingual knowledge

Lauriane Aufrant – PhD Defense

April 6, 2018

Supervisor: François Yvon
Co-supervisor: Guillaume Wisniewski

They ate pizza with anchovies

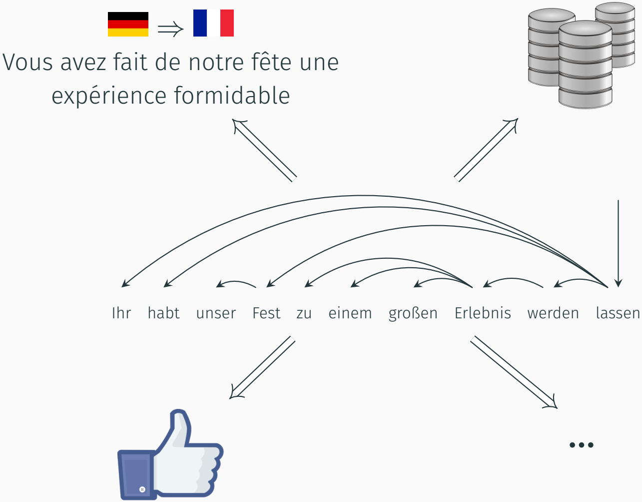
They ate pizza with anchovies

They ate pizza with anchovies They ate pizza with anchovies

↪ Natural Language Processing

↪ **Dependency parsing**

Dependency parsing: downstream tasks



Transition-based dependency parsing [ArcEager system]

$\perp$ | They ate pizza with anchovies

stack buffer

They ate pizza with anchovies

CLASSIFIER

Transition-based dependency parsing [ArcEager system]

They | ate pizza with anchovies

stack buffer

They ate pizza with anchovies

Classifier

Shift

Transition-based dependency parsing [ArcEager system]

$\perp$ | ate pizza with anchovies

stack buffer

They ate pizza with anchovies

CLASSIFIER

LEFT

Transition-based dependency parsing [ArcEager system]

ate | pizza with anchovies

stack buffer

They ate pizza with anchovies

CLASSIFIER

SHIFT

Transition-based dependency parsing [ArcEager system]

$\boxed{\text{ate}}$ | pizza with anchovies

stack buffer

$\uparrow \; \uparrow \qquad \uparrow \qquad \uparrow$

$s_1 \; s_0 \qquad n_0 \qquad n_1$

$$
\begin{aligned}
s_1 &= \varnothing \\
s_0 &= \text{ATE} \\
n_0 &= \text{PIZZA} \\
n_1 &= \text{WITH}
\end{aligned}
$$

$\downarrow$

CLASSIFIER

They ate pizza with anchovies

Transition-based dependency parsing [ArcEager system]

ate | pizza with anchovies

stack buffer

s_1 s_0 n_0 n_1

$s_1 = \varnothing$
$s_0 =$ ATE
$n_0 =$ PIZZA
$n_1 =$ WITH

CLASSIFIER

RIGHT

They ate pizza with anchovies

Transition-based dependency parsing [ArcEager system]

ate pizza | with anchovies

stack buffer

They ate pizza with anchovies

CLASSIFIER

RIGHT

Transition-based dependency parsing [ArcEager system]

ate pizza with | anchovies

stack buffer

They ate pizza with anchovies

CLASSIFIER

SHIFT

Transition-based dependency parsing [ArcEager system]

ate pizza | anchovies

stack buffer

They ate pizza with anchovies

CLASSIFIER

LEFT

Transition-based dependency parsing [ArcEager system]

ate pizza anchovies | $\perp$

stack buffer

They ate pizza with anchovies

Classifier

Right

Transition-based dependency parsing [ArcEager system]

ate pizza | ⊥

stack buffer

They ate pizza with anchovies

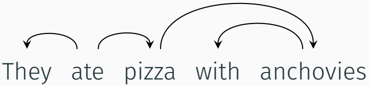
CLASSIFIER

REDUCE

Transition-based dependency parsing [ArcEager system]

ate | $\perp$

stack buffer



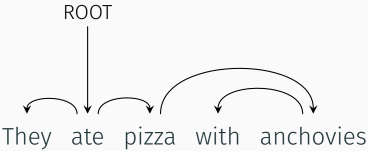
Classifier

Reduce

Transition-based dependency parsing [ArcEager system]

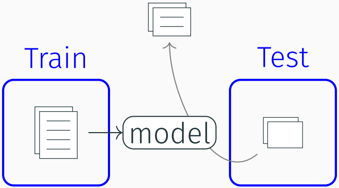
$\boxed{\text{ate}} \mid \perp$

stack buffer



Classifier

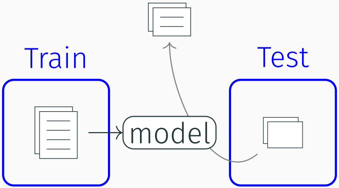
Data requirements of modern NLP



Machine learning $\Longleftrightarrow$ annotated data

$\Longleftrightarrow$ time and money

Data requirements of modern NLP



Machine learning $\iff$ annotated data

$\iff$ time and money

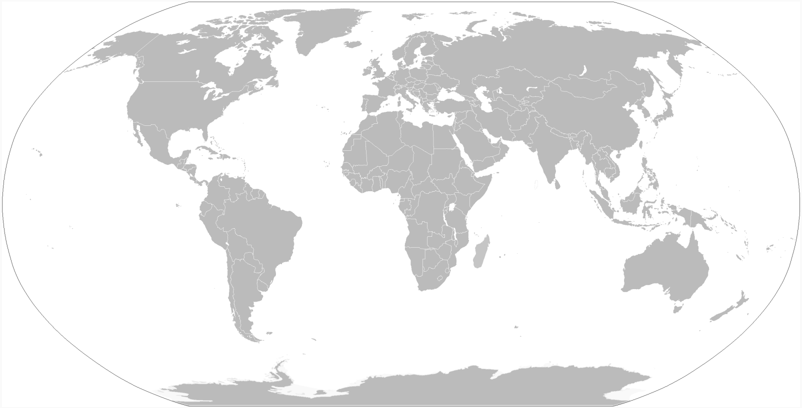
Dependency parsing

- Penn Treebank (English): 43k sentences, 10 years, 1 M$
- Prague Dependency Treebank (Czech): 87k sentences
- 500M tweets per day $\Rightarrow$ only a few thousands annotated

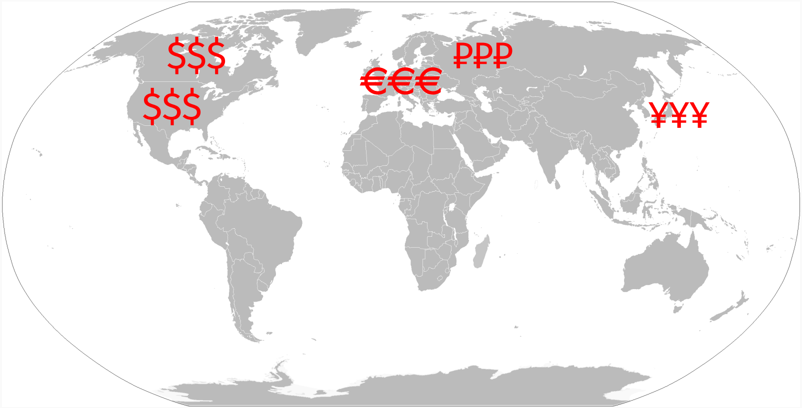
Machine Translation

- 52,000,000 Czech-English translated sentences
- 3,000,000,000 English sentences

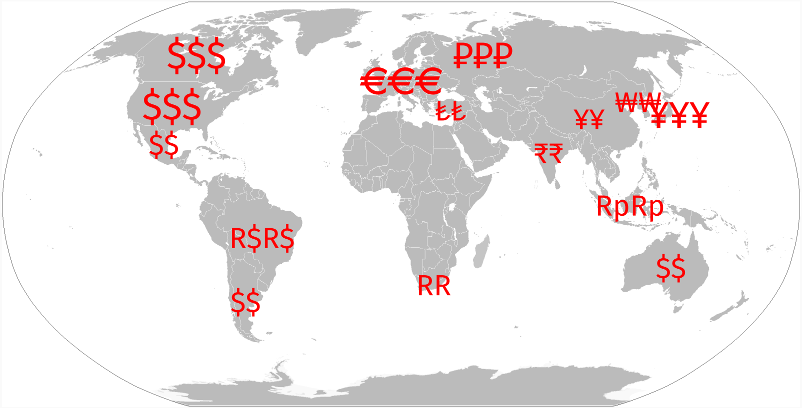
Time and money: where are they?



Time and money: where are they?



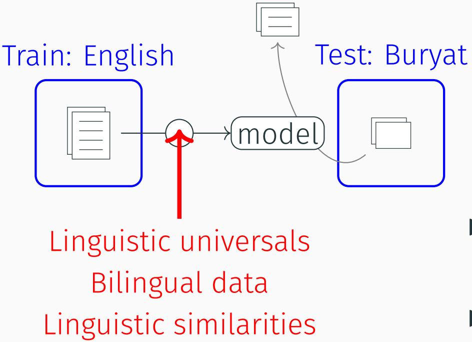
Time and money: where are they?



The Buryat language

The Buryat language

Cross-lingual transfer



- Transfer of knowledge
  $\rightsquigarrow$ model parameters
- Transfer of data
  $\rightsquigarrow$ annotations

Worst-case scenario:

$$\left\{\begin{array}{l}\text{No annotated data}\\ \text{No bilingual data}\\ \text{No raw data}\end{array}\right. \implies \text{zero-resource scenario}$$

Cross-lingual transfer

- PoS tagging and morphology
  - [Yarowsky *et al.*, 2001]
  - [Das & Petrov, 2011; Täckström *et al.*, 2013; Agić *et al.*, 2015; Yu *et al.*, 2016]
- Dependency parsing
  - [Hwa *et al.*, 2002; Zeman & Resnik, 2008; McDonald *et al.*, 2011; Naseem *et al.*, 2012]
  - [McDonald *et al.*, 2013; Ma & Xia, 2014; Tiedemann *et al.*, 2014; Rosa & Zabokrtsky, 2015; Duong *et al.*, 2015; Rasooli & Collins, 2015; Agić *et al.*, 2016]
- Opinion and subjectivity
  - [Banea *et al.*, 2008; Wan, 2009; Wei & Pal, 2010; Lu *et al.*, 2011; Klinger & Cimiano, 2015]
- Named Entity Recognition
  - [Täckström *et al.*, 2012; Wang & Manning, 2014]
- Coreferences [Martins, 2015]
- Semantic parsing [Kozhevnikov & Titov, 2014]
- Speech recognition [Ghoshal *et al.*, 2013]
- Document classification [Rigutini *et al.*, 2005; Klementiev *et al.*, 2012]

Problem statement

✓ Low-resourced NLP $\Rightarrow$ cross-lingual transfer

✗ Not always applicable: specific requirements of cross-lingual resources

$\hookrightarrow$ Give up on other languages?

Purpose:

- Make more resources usable
- Make transfer methods more flexible regarding resources

$\Longrightarrow$ How to combine those sources/resources at fine grain?

Contributions [11 publications, 2 shared tasks, 1 award]

- **A new transfer framework**: multi-(re)source combination based on a cascading architecture

- **PanParser**: a modular and open source parser
  - unified formalism for several parsing algorithms
  - global dynamic oracle, sampling bias, non-projective training data, non-arc-decomposable cases of ArcEager...
- Assessment of **transfer usefulness**
- Avoid **systematic errors**, using typological knowledge
- Evaluation of **cross-linguistic divergences**

- In-depth analysis of the **inner workings** of parsers
  - feature-level interactions, complexity of a dependency, quantification of available knowledge...
- Improved **cross-lingual generalization** of taggers/parsers
- Transfer of **bilingual knowledge**: word alignments

Contributions [11 publications, 2 shared tasks, 1 award]

- **A new transfer framework**: multi-(re)source combination based on a cascading architecture

- **PanParser**: a modular and open source parser
  - unified formalism for several parsing algorithms
  - global dynamic oracle, sampling bias, non-projective training data, non-arc-decomposable cases of ArcEager...
- Assessment of **transfer usefulness**
- Avoid **systematic errors**, using typological knowledge
- Evaluation of **cross-linguistic divergences**

- In-depth analysis of the **inner workings** of parsers
  - feature-level interactions, complexity of a dependency, quantification of available knowledge...
- Improved **cross-lingual generalization** of taggers/parsers
- Transfer of **bilingual knowledge**: word alignments

Outline

Cross-lingual transfer

Leveraging typological knowledge

Extensions to the parsing framework

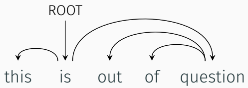
A new transfer framework: multi-(re)source combination

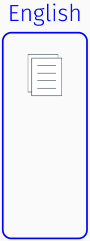
Conclusions

Outline

Delexicalized transfer [Zeman & Resnik, 2008]

↪ Identical PoS tags behave similarly in both languages

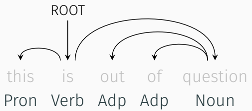


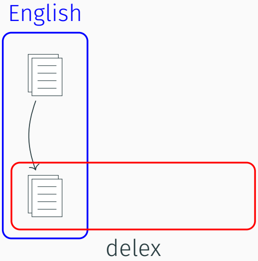


Reuse of source model

Delexicalized transfer [Zeman & Resnik, 2008]

↪ Identical PoS tags behave similarly in both languages

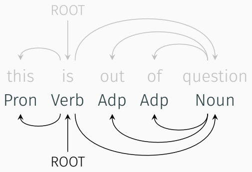


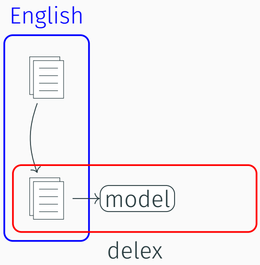


Reuse of source model

Delexicalized transfer [Zeman & Resnik, 2008]

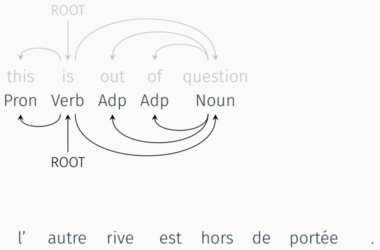
↪ Identical PoS tags behave similarly in both languages

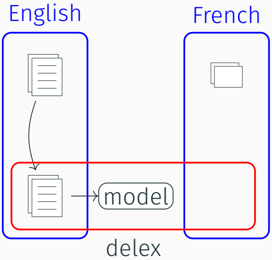




Reuse of source model

↪ Identical PoS tags behave similarly in both languages

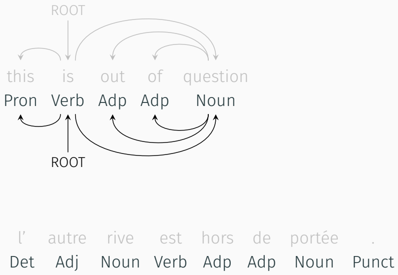


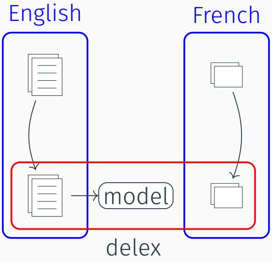


Reuse of source model

Delexicalized transfer [Zeman & Resnik, 2008]

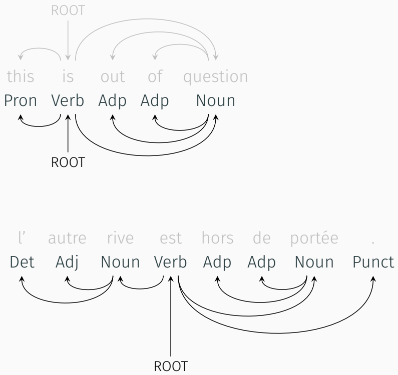
↪ Identical PoS tags behave similarly in both languages

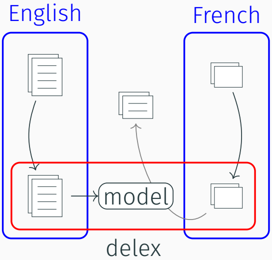




Reuse of source model

↪ Identical PoS tags behave similarly in both languages

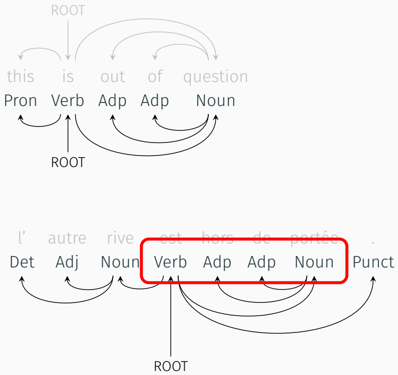


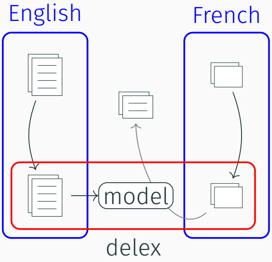


Reuse of source model

Delexicalized transfer [Zeman & Resnik, 2008]

↪ Identical PoS tags behave similarly in both languages

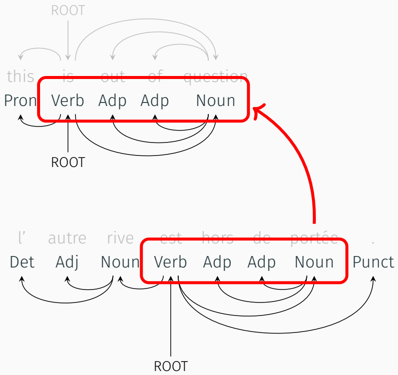


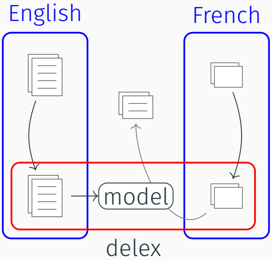


Reuse of source model

Delexicalized transfer [Zeman & Resnik, 2008]

↪ Identical PoS tags behave similarly in both languages





Reuse of source model

$\hookrightarrow$ Aligned words behave similarly in both languages

Pron	Verb	Noun	Adp	Det	Noun	Noun
They	took	part	in	the	vaccination	campaign

Ils ont participé à la campagne de vaccination

Annotation projection [Yarowsky *et al.*, 2001]

↪ Aligned words behave similarly in both languages

Pron Verb Noun Adp Det Noun Noun

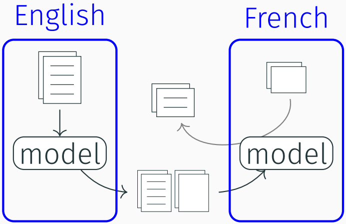
They took part in the vaccination campaign

Ils ont participé à la campagne de vaccination

Pron Verb/Noun Adp Det Noun Noun

↪ Aligned words behave similarly in both languages

Pron	Verb	Noun	Adp	Det	Noun		Noun	
They	took	part	in	the	vaccination		campaign	
Ils	ont	participé	à	la	campagne	de	vaccination	
Pron		Verb/Noun	Adp	Det	Noun		Noun	

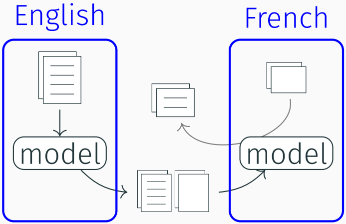


Creation of annotated data

Annotation projection [Yarowsky *et al.*, 2001]

↪ **Aligned words behave similarly in both languages**

Pron	Verb	Noun	Adp	Det	Noun		Noun
They	took	part	in	the	vaccination		campaign
Ils	ont	participé	à	la	campagne	de	vaccination
Pron		Verb/Noun	Adp	Det	Noun		Noun



Creation of annotated data

✓ Also works with distant languages

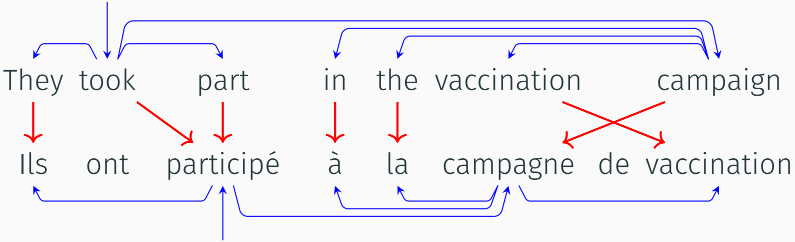
✓ High accuracy

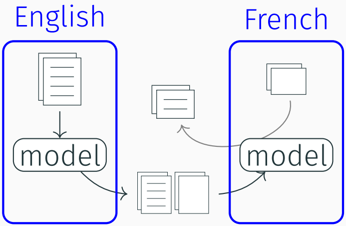
✗ Completion heuristics

✗ Parallel data: availability? domain? quality?

Annotation projection [Yarowsky *et al.*, 2001]

↪ **Aligned words behave similarly in both languages**





Creation of annotated data

- ✓ Also works with distant languages
- ✓ High accuracy
- ✗ Completion heuristics
- ✗ Parallel data: availability? domain? quality?

Cross-lingual resources

- Consistent annotation schemes
  - UPOS [Petrov *et al.*, 2012]
  - UDT [McDonald *et al.*, 2013]
  - UD [Nivre *et al.*, 2016]

- Cross-lingual datasets
  - UD v1.0 (January 2015): 10 treebanks, 10 languages

    ...
  - UD v2.1 (November 2017): 102 treebanks, 60 languages

$\hookrightarrow$ mostly UD v2.0 here (73 treebanks, 54 languages)

Summary: cross-lingual transfer

- Extending NLP methods to **more than the 100 usual languages** (out of 7,000)
- Leverage **bilingual data** or **linguistic similarities** with better-resourced languages
- Main methods: **delexicalized transfer** and **annotation projection**
  - but also: feature mapping, training guidance, joint learning, multilingual models...
- Growing datasets with **consistent annotation schemes**

Outline

Cross-lingual transfer

Leveraging typological knowledge

- Impact of word order
- WALS-based rewriting [COLING'16]

Extensions to the parsing framework

A new transfer framework: multi-(re)source combination

Conclusions

An adjective close to a noun depends on this noun.

An adjective close to a noun depends on this noun.

An adjective close to a noun depends on this noun.

True in...

- ✓ English
- ✓ French
- ✓ Hebrew
- ✓ Bulgarian

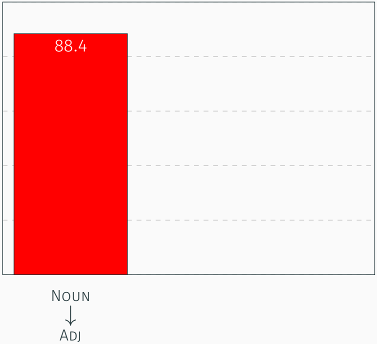
An adjective close to a noun depends on this noun.

True in...

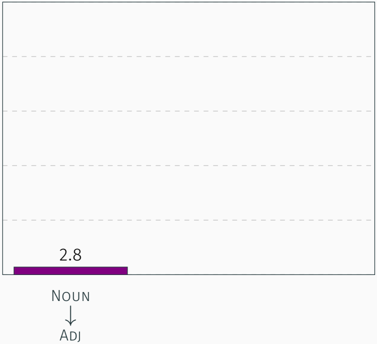
✓ English ✓ French

✓ Hebrew ✓ Bulgarian

Hebrew (monolingual)



Hebrew → Bulgarian

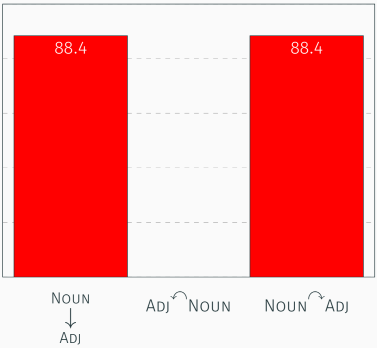


An adjective close to a noun depends on this noun.

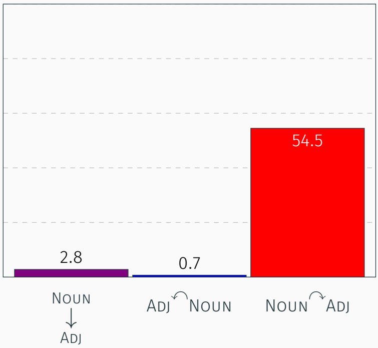
True in...

- ✓ English
- ✓ French
- ✓ Hebrew
- ✓ Bulgarian

Hebrew (monolingual)

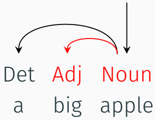


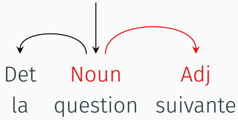
Hebrew → Bulgarian



Impact of word order

At data level:



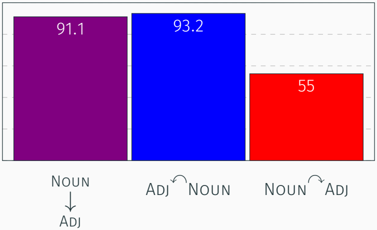


At model level:

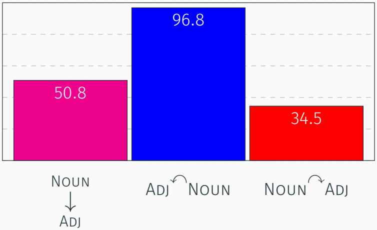
$(s_0 = \textsc{Adj} \wedge n_0 = \textsc{Noun}) \Rightarrow \textsc{Left}$ $(s_0 = \textsc{Noun} \wedge n_0 = \textsc{Adj}) \Rightarrow \textsc{Right}$

On accuracy (UAS):

English (monolingual)

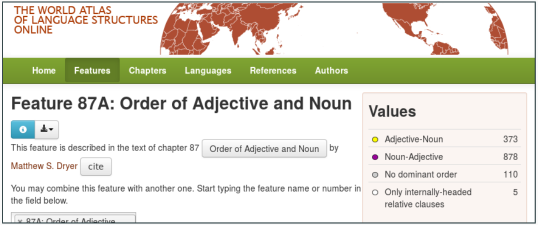


English → French



WALS: a database of typological features for 2,679 languages
[`http://wals.info`]

↪ Over 1,000 languages with word order features



English	Adjective-Noun			
French	Noun-Adjective	Harris 1988: 227		

Using WALS to preprocess training data

Heuristic rule extraction for **switching** and **deleting** words

$$\text{87A} \quad \begin{cases} \text{[English] Adjective-Noun} \\ \text{[French] Noun-Adjective} \end{cases}$$

$\Longrightarrow$ [English$\rightarrow$French] switch ADJ-NOUN into NOUN-ADJ

Using WALS to preprocess training data

Heuristic rule extraction for **switching** and **deleting** words

$$\text{87A} \quad \begin{cases} \text{[English] Adjective-Noun} \\ \text{[French] Noun-Adjective} \end{cases}$$

$\Longrightarrow$ [English$\rightarrow$French] switch ADJ-NOUN into NOUN-ADJ

- ✓ just a preprocessing step: easy to perform & to extend
- ✓ most work already done by linguists
- ✓ readily available for 1,000 languages

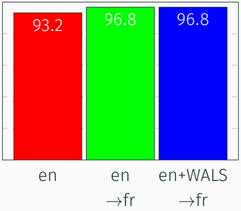
Reshaping training instances: examples

English – training data			French – desired output
Baseline		Proposal	
Det Adj Noun a big apple	⇒	Det Noun Adj a apple big	Det Noun Adj la question suivante
Det Adj Noun the whole world	⇒	Det Noun Adj the world whole	Det Noun Adj la flotte romaine
Det Noun Noun an investment firm	⇒	Det Noun Noun an firm investment	Det Noun Adp Noun la plate-forme de distribution

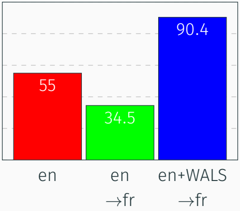
Experimental results

English → French

Adj ↷ Noun

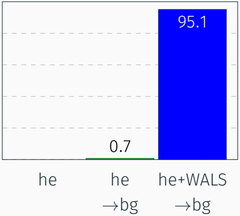


Noun ↷ Adj

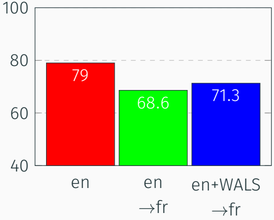


Hebrew → Bulgarian

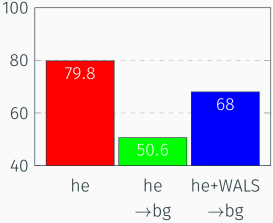
Adj ↷ Noun



Overall score: +2.7%



Overall score: +17.4%



Systematic experiments

Fine-grained analysis across various language pairs

↪ 6,000+ experiments on 40 languages & 4 methods

Many transfer errors are easy to avoid

↪ regular divergences between both languages

↪ word order issues, non-existing PoS

Proposal: leveraging previous works in linguistics (WALS)

↪ +3% accuracy on average

↪ very efficient on some error types: up to +90% accuracy

Summary: leveraging typological knowledge

- **Extension of linguistic coverage**: zero-resource transfer targeting 1,000 languages
- Identification of **typological differences** as the main cause of many failures: consistent annotations do not suffice
- Preprocessing using **linguistic knowledge** boosts the systems
- A way to exploit **additional resources** during the transfer process

Outline

Cross-lingual transfer

Leveraging typological knowledge

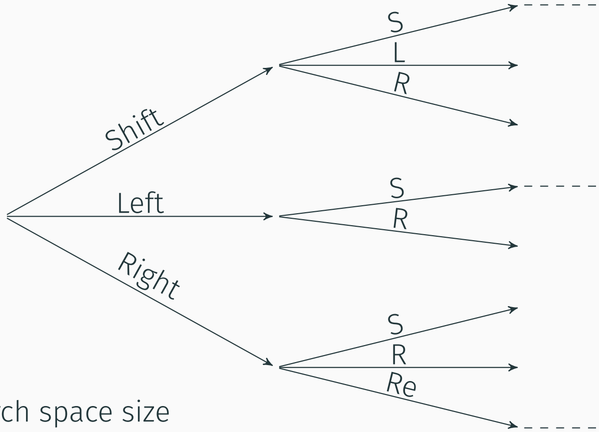
Extensions to the parsing framework

- Dynamic oracle and beam search
- Global dynamic oracle with restart [EACL'17]
- PanParser

A new transfer framework: multi-(re)source combination

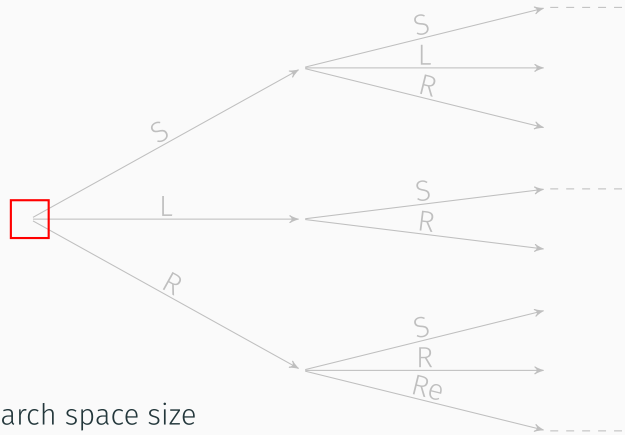
Conclusions

Greedy inference



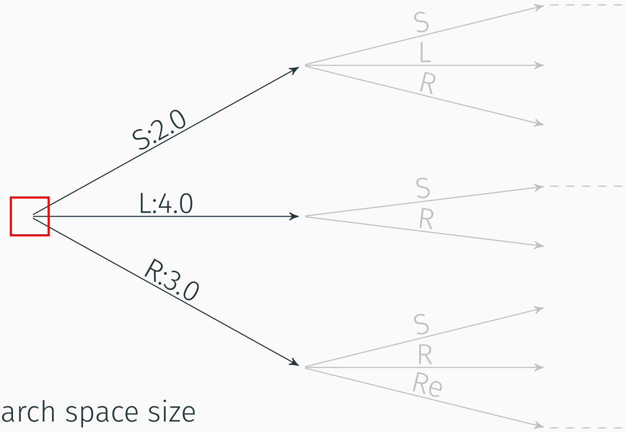
Search space size
$\approx 3^{2n}$

Greedy inference



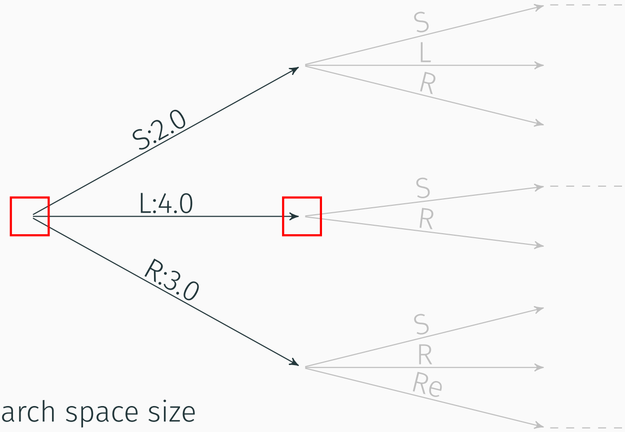
Search space size
$\approx 3^{2n}$

Greedy inference



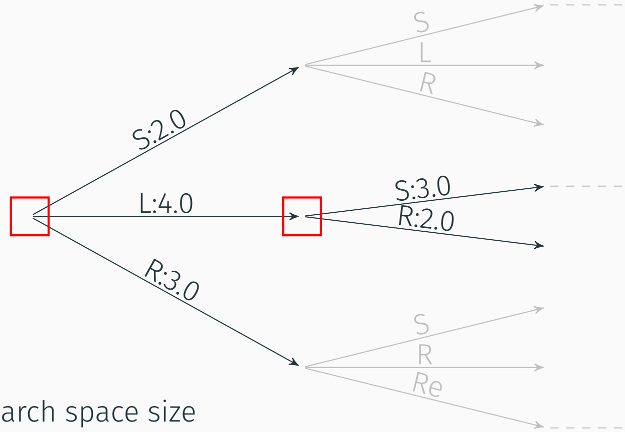
Search space size
$\approx 3^{2n}$

Greedy inference



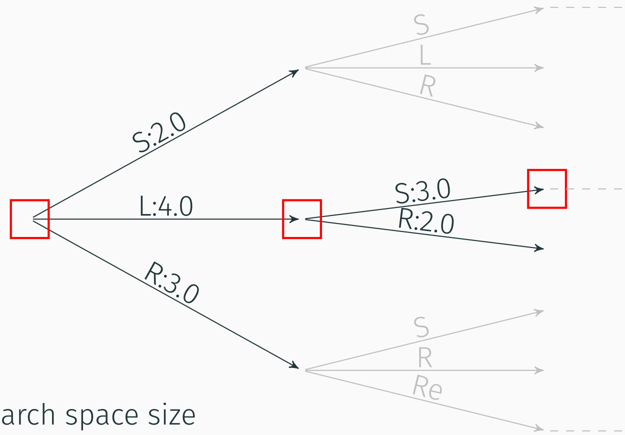
Search space size
$\approx 3^{2n}$

Greedy inference



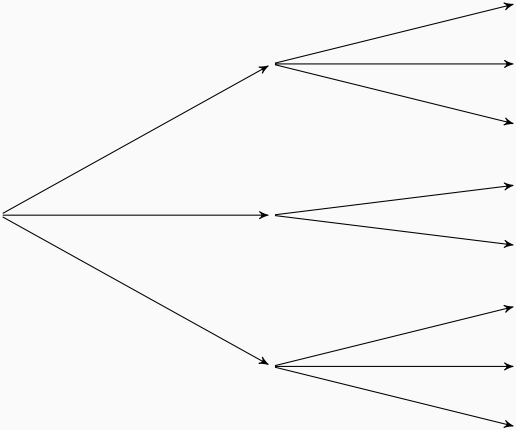
Search space size
$\approx 3^{2n}$

Greedy inference

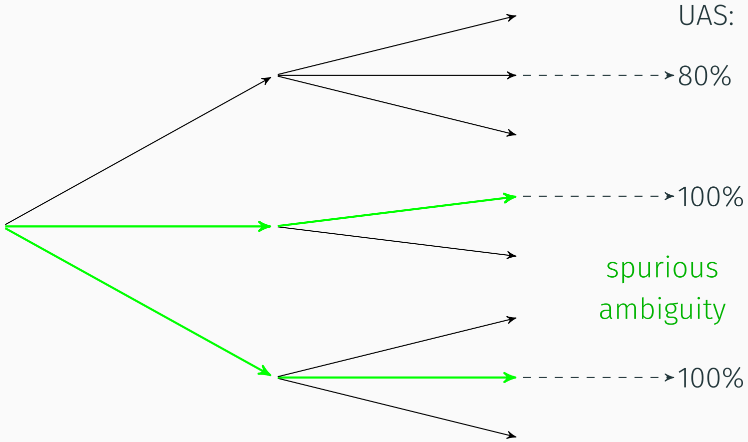


Search space size
$\approx 3^{2n}$

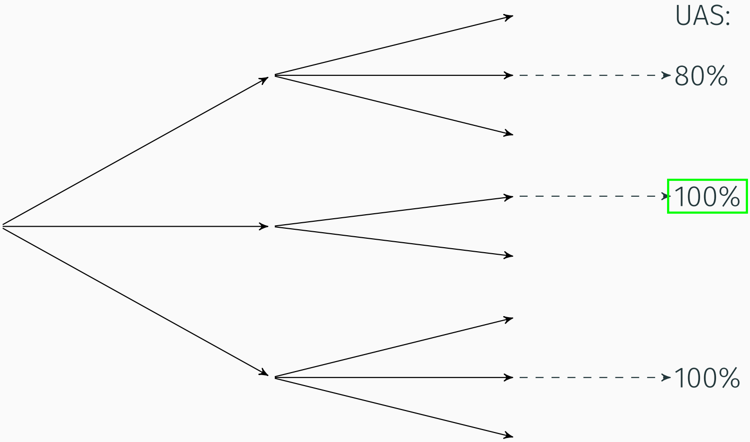
Greedy training...



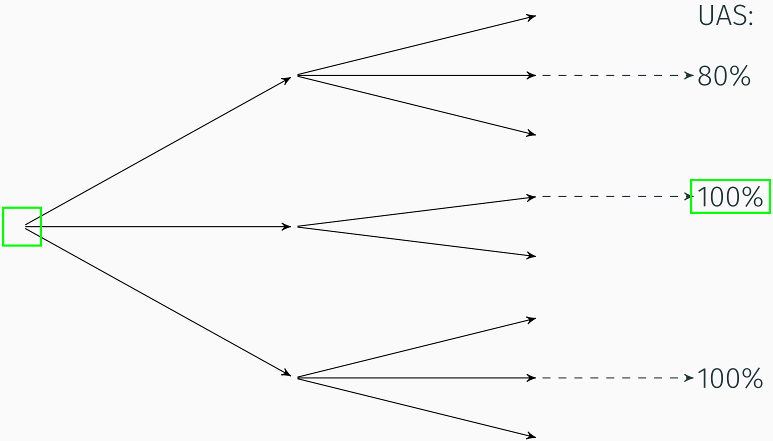
Greedy training...



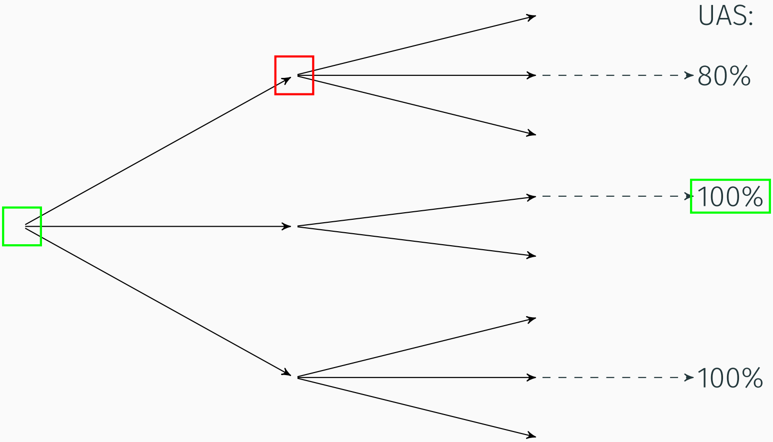
Greedy training...



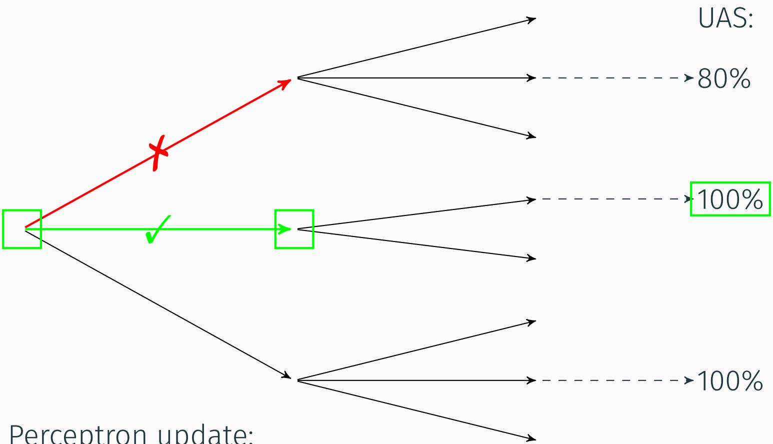
Greedy training...



Greedy training...



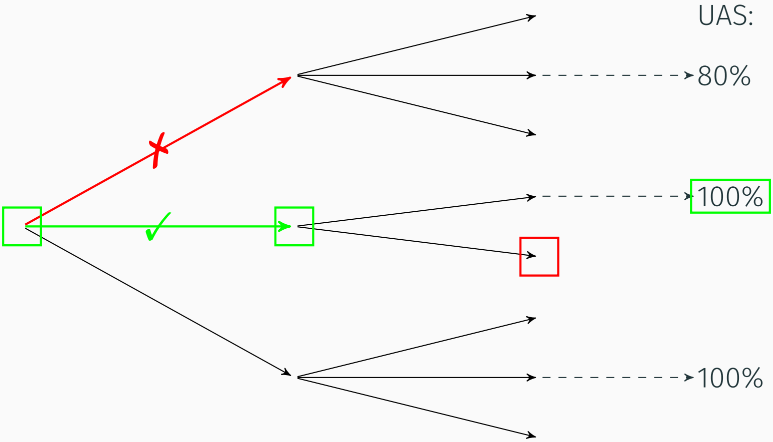
Greedy training...



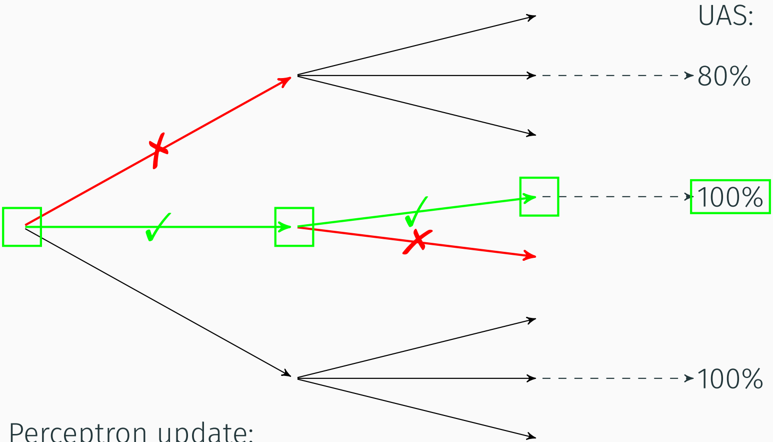
Perceptron update:

$\mathbf{w} \leftarrow \mathbf{w} - \phi + \phi^*$

Greedy training...



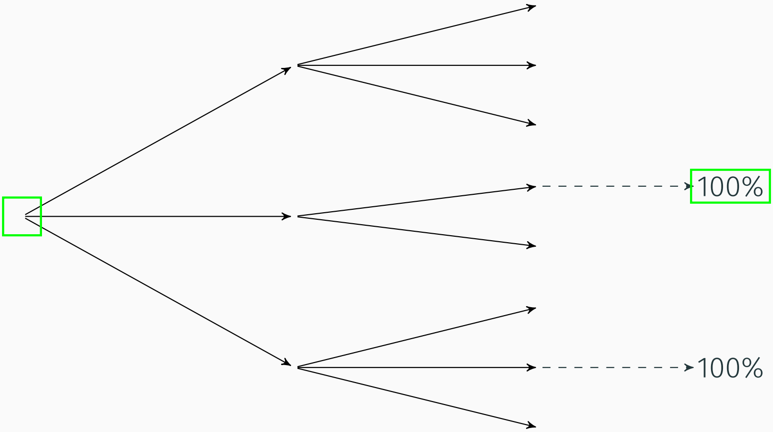
Greedy training...



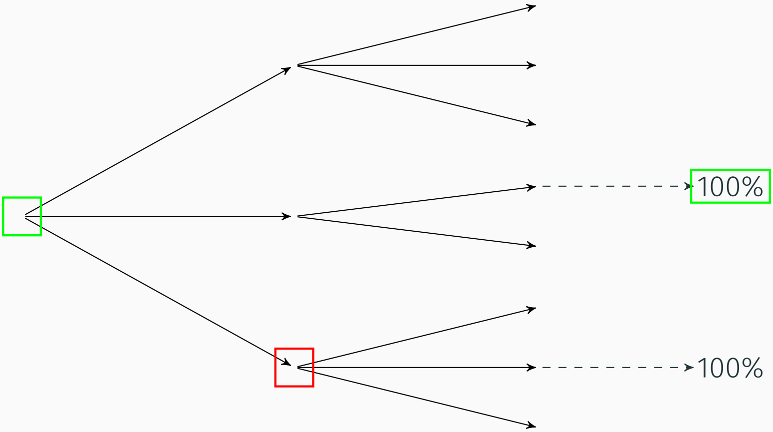
Perceptron update:

$\mathbf{w} \leftarrow \mathbf{w} - \phi + \phi^*$

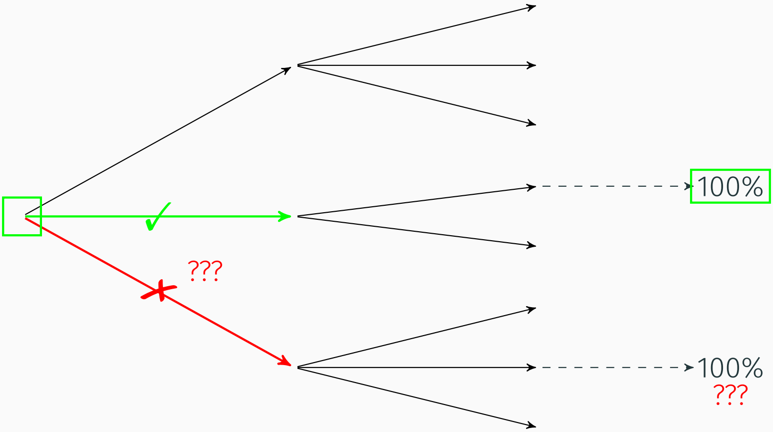
Greedy training with non-determinism?



Greedy training with non-determinism?

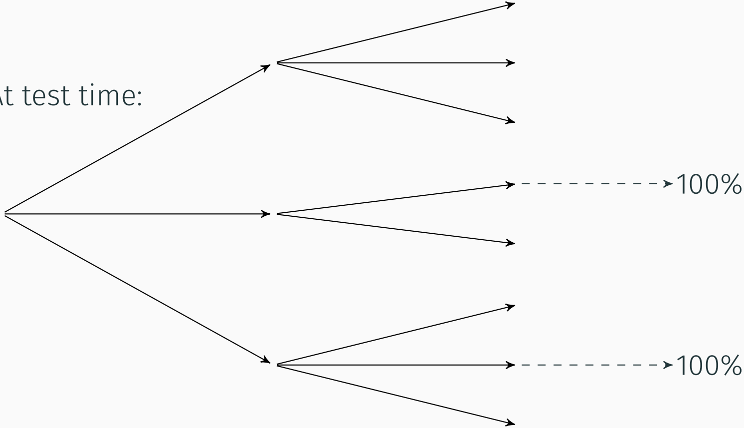


Greedy training with non-determinism?



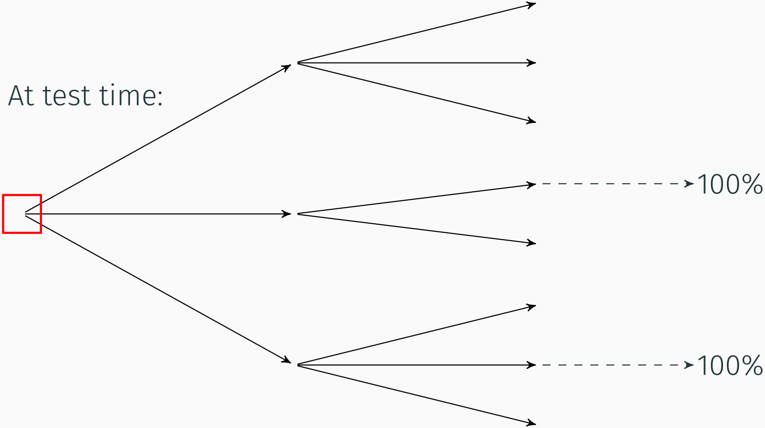
Greedy training in the suboptimal space?

At test time:

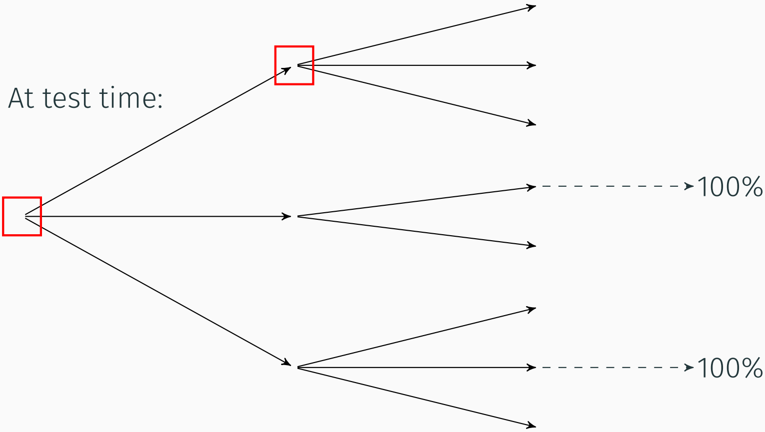


Greedy training in the suboptimal space?

At test time:

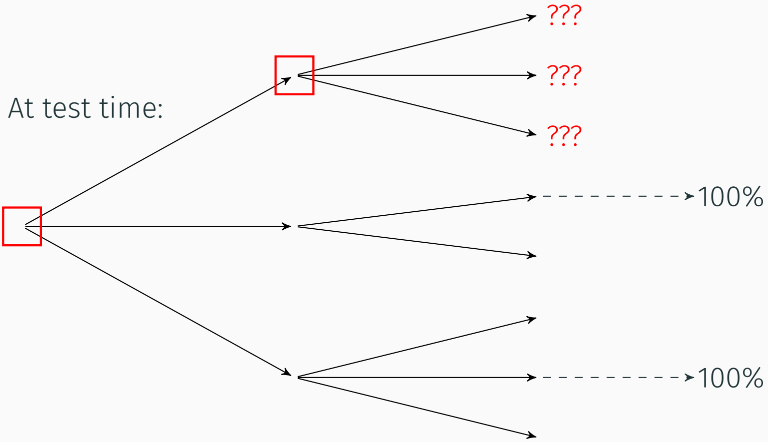


Greedy training in the suboptimal space?

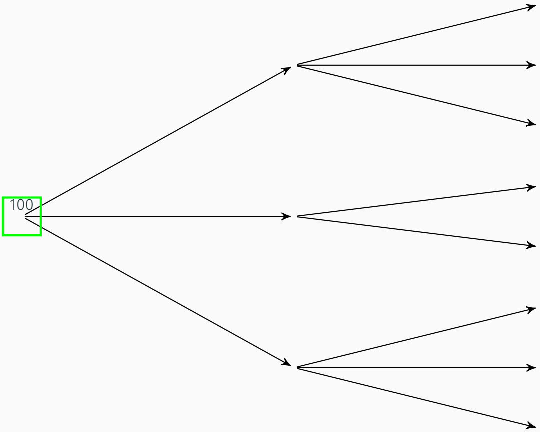


Greedy training in the suboptimal space?

At test time:

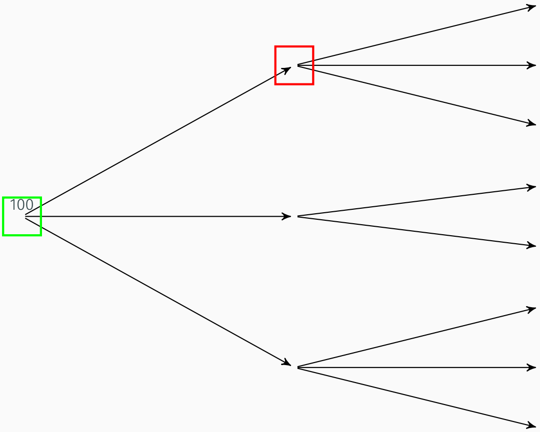


Greedy training with a dynamic oracle



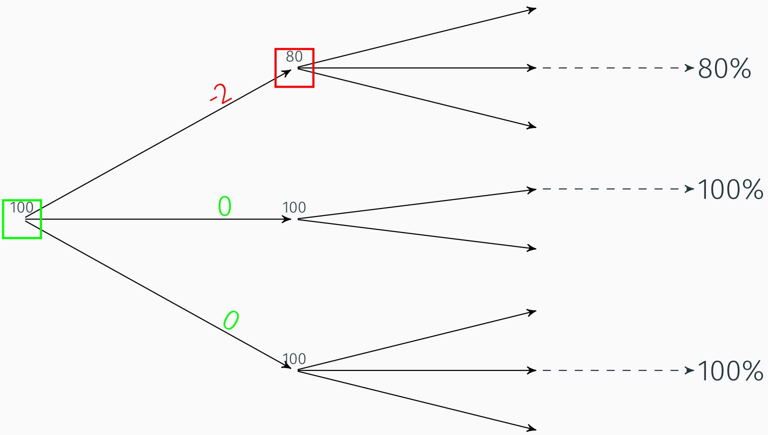
COST(*action*) [Goldberg & Nivre, 2012]:
Δ expected UAS over the sentence

Greedy training with a dynamic oracle



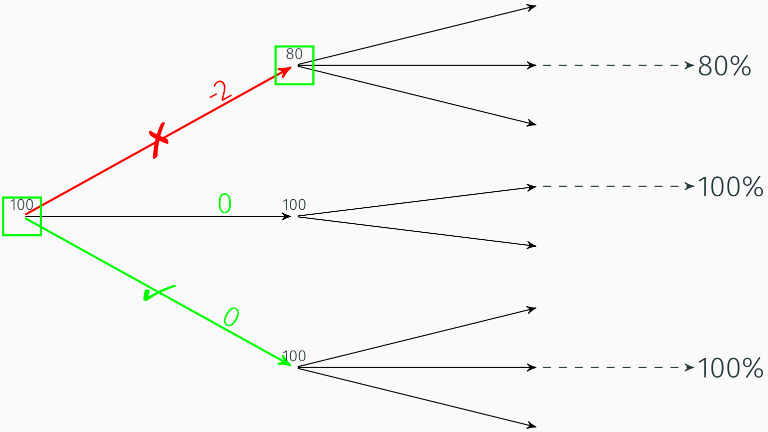
COST(*action*) [Goldberg & Nivre, 2012]:
Δ expected UAS over the sentence

Greedy training with a dynamic oracle



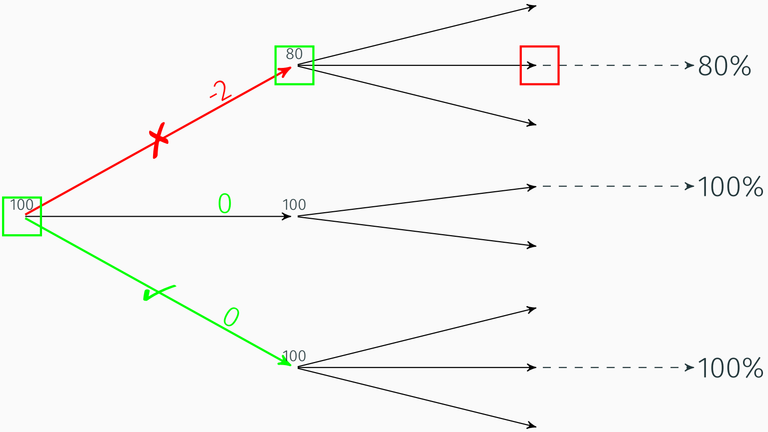
COST(*action*) [Goldberg & Nivre, 2012]:
Δ expected UAS over the sentence

Greedy training with a dynamic oracle



COST(*action*) [Goldberg & Nivre, 2012]:

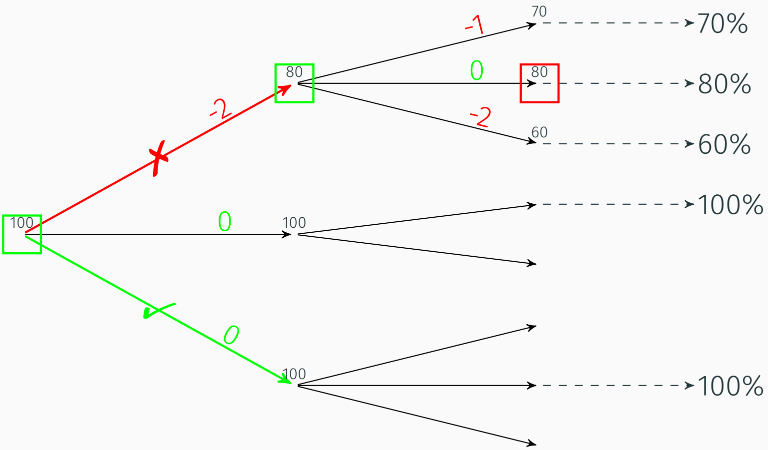
Δ expected UAS over the sentence



Cost(*action*) [Goldberg & Nivre, 2012]:

Δ expected UAS over the sentence

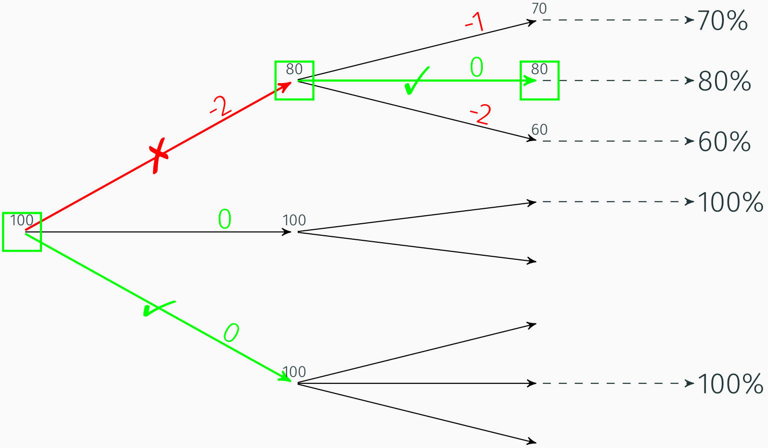
Greedy training with a dynamic oracle



COST(*action*) [Goldberg & Nivre, 2012]:

Δ expected UAS over the sentence

Greedy training with a dynamic oracle



COST(*action*) [Goldberg & Nivre, 2012]:
Δ expected UAS over the sentence

Greedy dynamic oracle [Goldberg & Nivre, 2012]

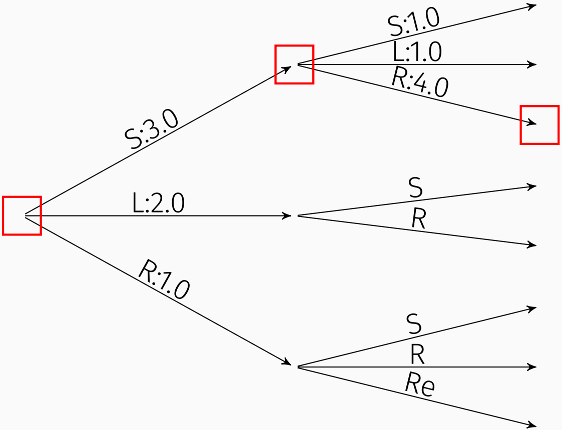
References = zero-cost actions (Cost function)

No deterministic precomputation of the reference

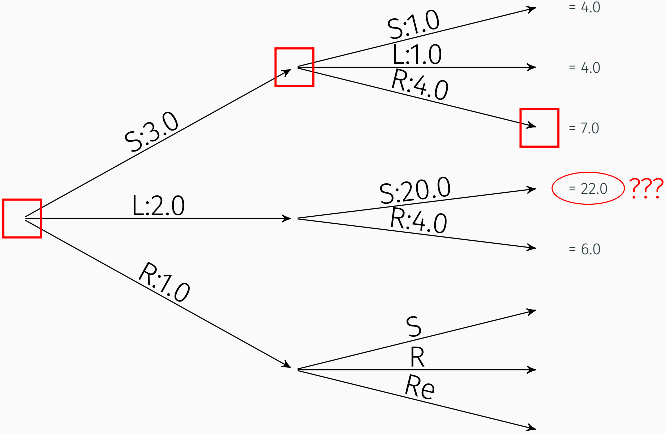
Abstract from past errors: enable exploration

Experimental gain: +1 to +2 UAS

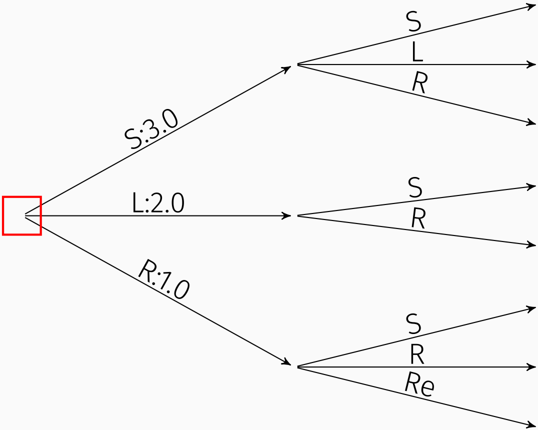
Beam search: why?



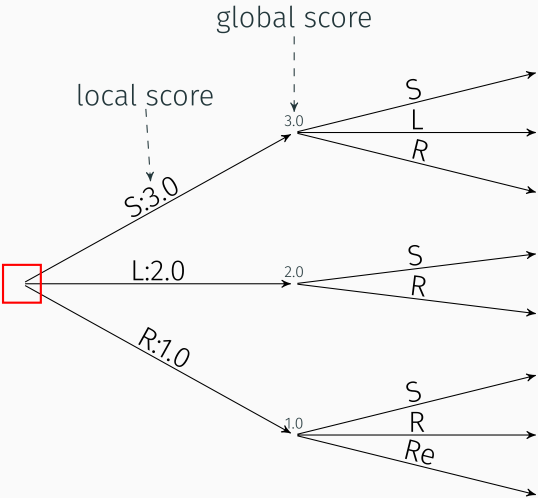
Beam search: why?

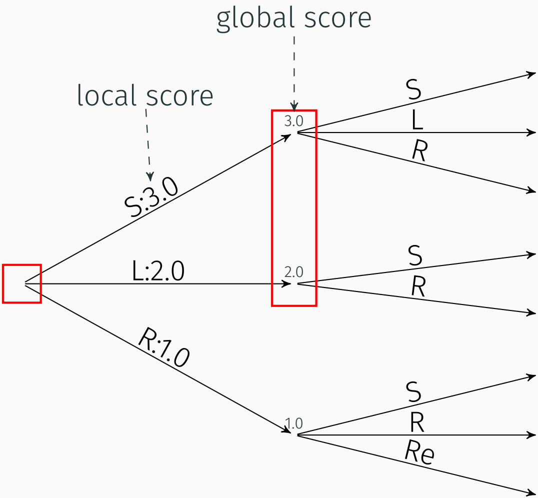


Beam search

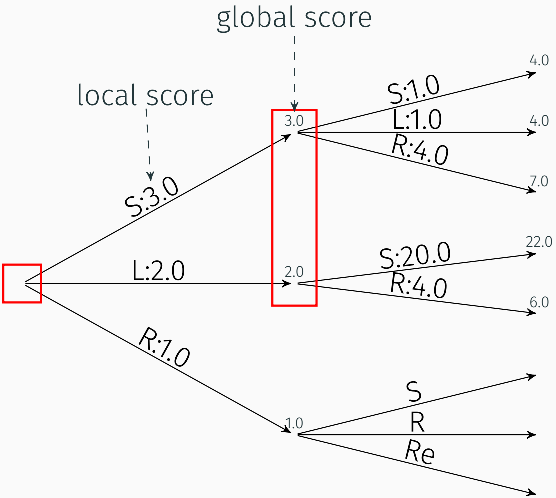


Beam search



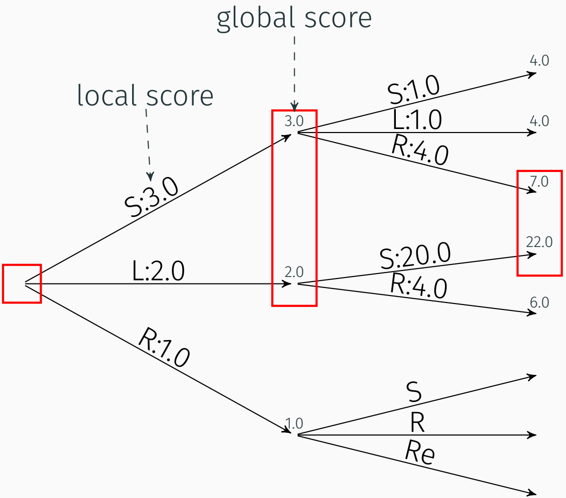

global score
local score
S:3.0
L:2.0
R:1.0
3.0
2.0
1.0
S
L
R
S
R
S
R
Re

Beam search

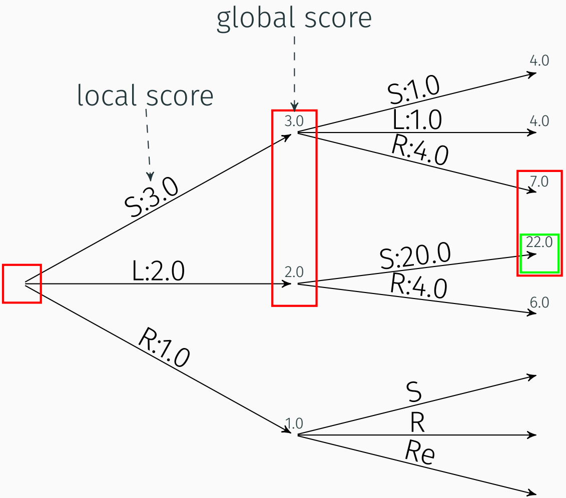


global score
local score
3.0
2.0
1.0
S:3.0
L:2.0
R:1.0
S:1.0
L:1.0
R:4.0
S:20.0
R:4.0
S
R
Re
4.0
4.0
7.0
22.0
6.0

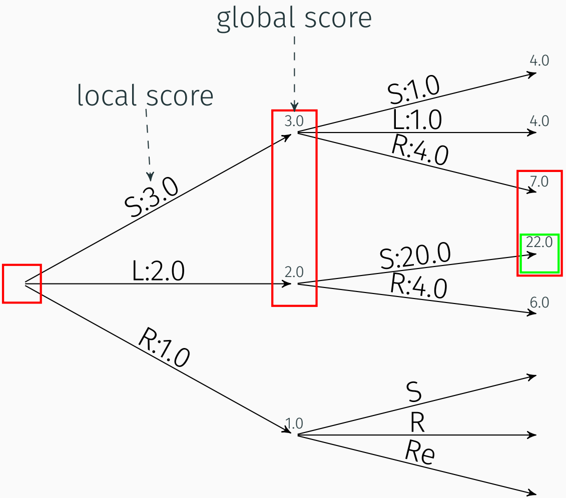
Beam search


global score
local score
S:3.0
L:2.0
R:1.0
3.0
2.0
1.0
S:1.0
L:1.0
R:4.0
S:20.0
R:4.0
S
R
Re
4.0
4.0
7.0
22.0
6.0

Beam search

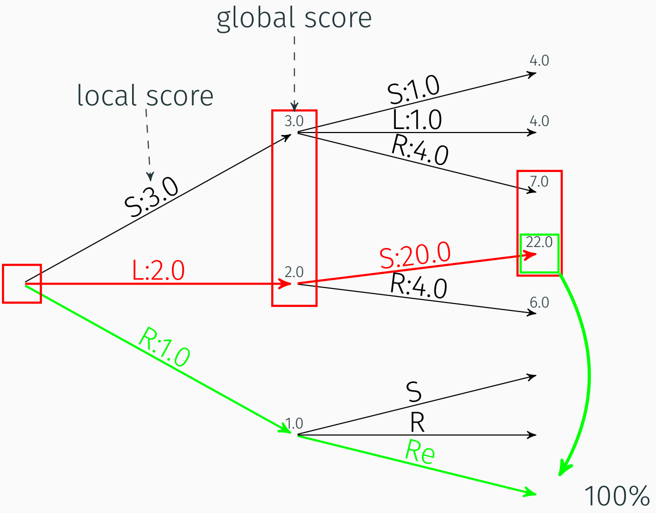

global score
local score
S:3.0
L:2.0
R:1.0
3.0
2.0
1.0
S:1.0
L:1.0
R:4.0
S:20.0
R:4.0
S
R
Re
4.0
4.0
7.0
22.0
6.0

Beam search



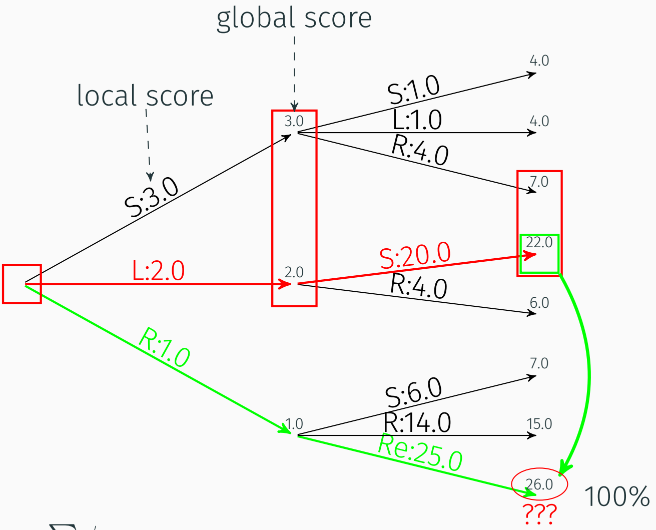
$\Phi_{global} = \sum \phi_{local}$

Global training



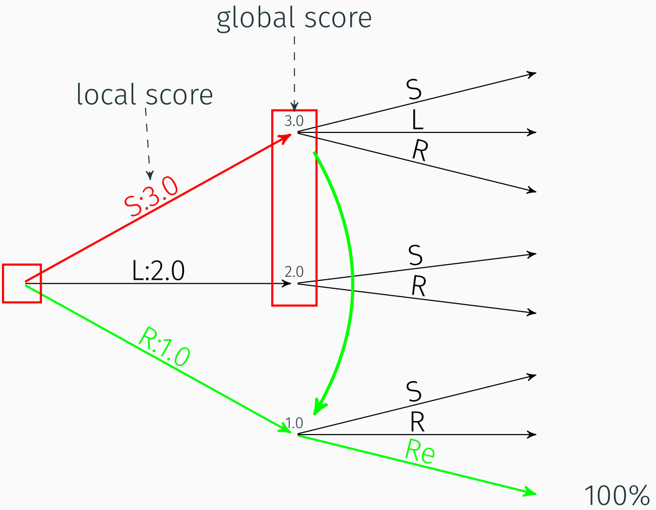
$\Phi_{global} = \sum \phi_{local}$

Global training



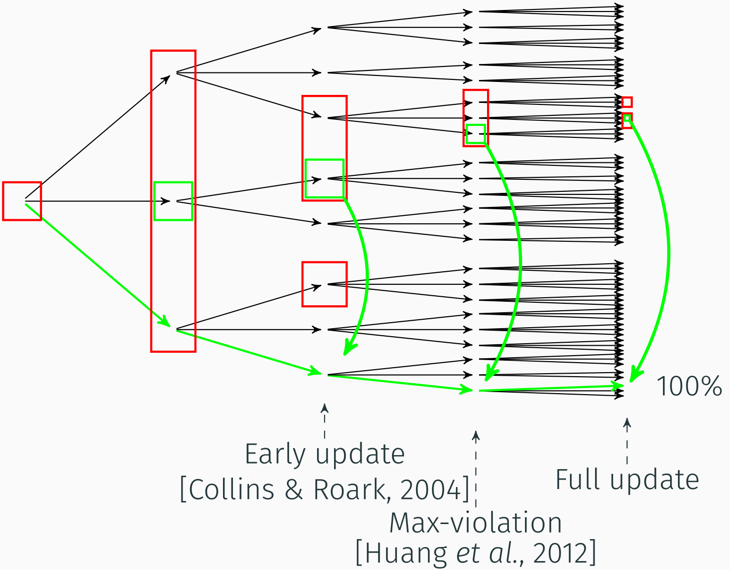
$\Phi_{global} = \sum \phi_{local}$

Global training

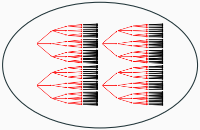


$\Phi_{global} = \sum \phi_{local}$

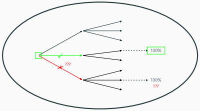
Global training: update strategies



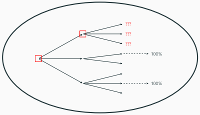
Global dynamic oracle: why?



Bias towards beginnings of derivations

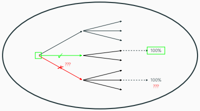


Deterministic oracle

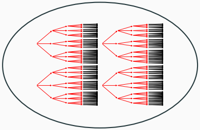


References always gold

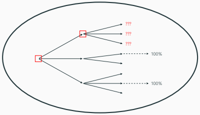
Global dynamic oracle: why?



Deterministic oracle



Bias towards beginnings of derivations

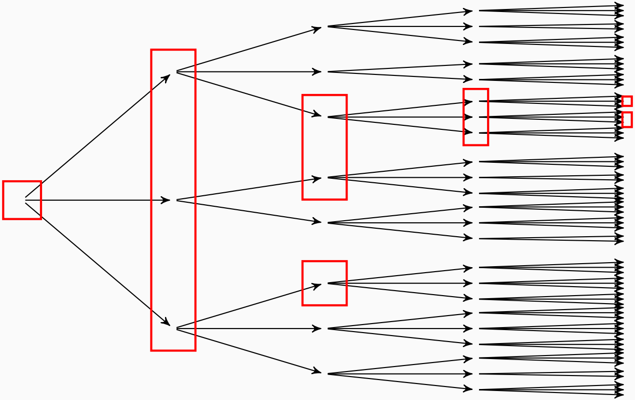


References always gold

Combine both lines of research

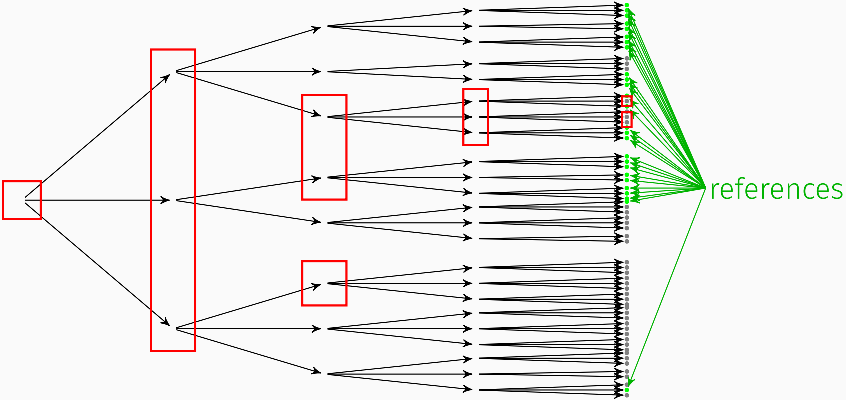
Global dynamic oracle

Old criterion: the reference falls out of the beam



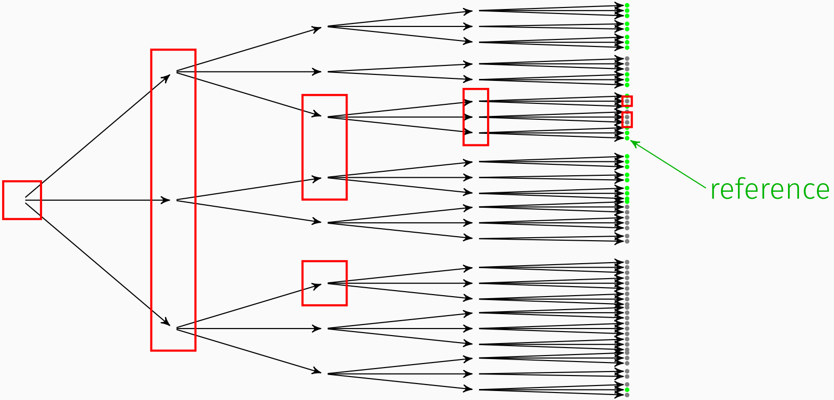
Global dynamic oracle

Old criterion: the reference falls out of the beam



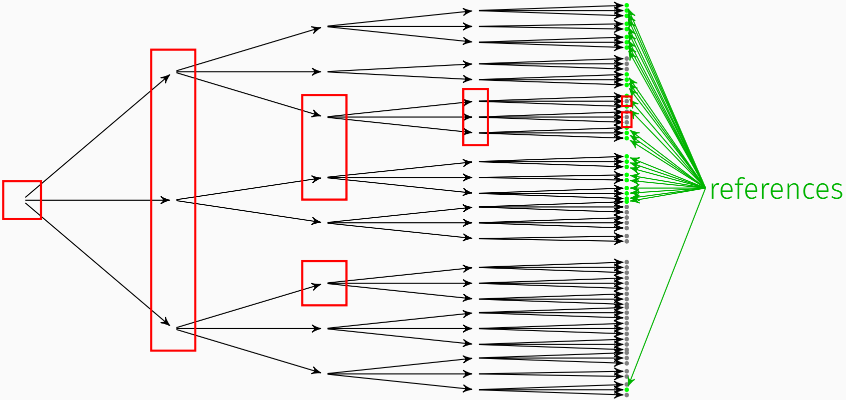
Global dynamic oracle

Old criterion: the reference falls out of the beam



Global dynamic oracle

Old criterion: the reference falls out of the beam

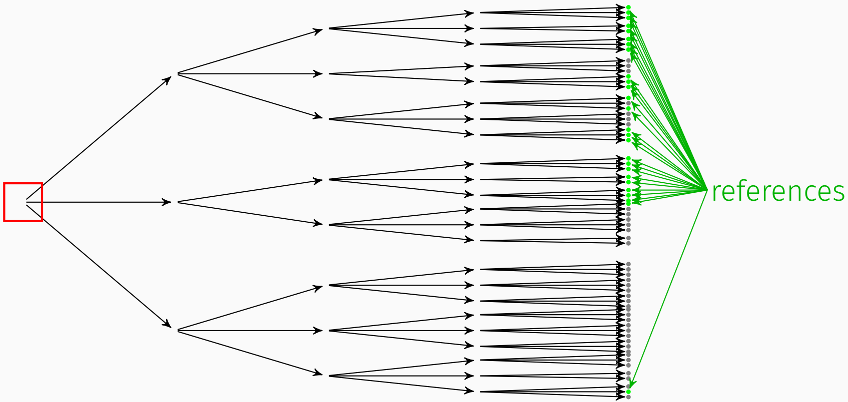


New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Old criterion: the reference falls out of the beam



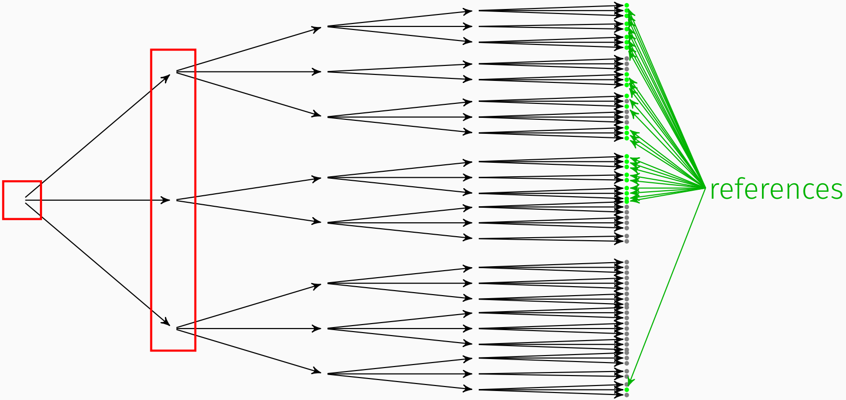
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



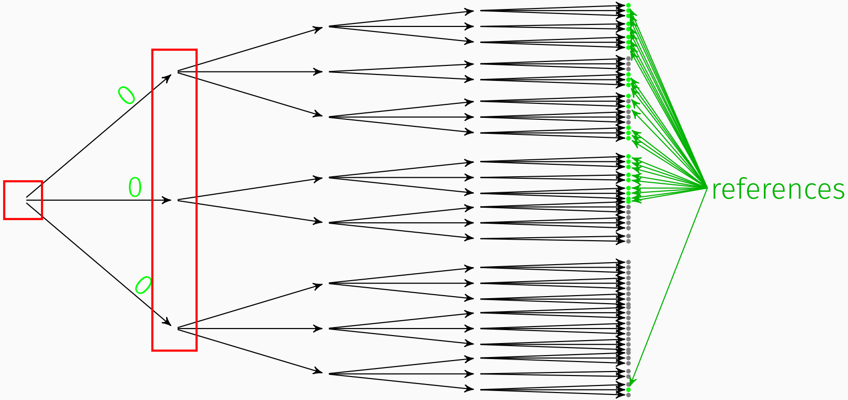
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



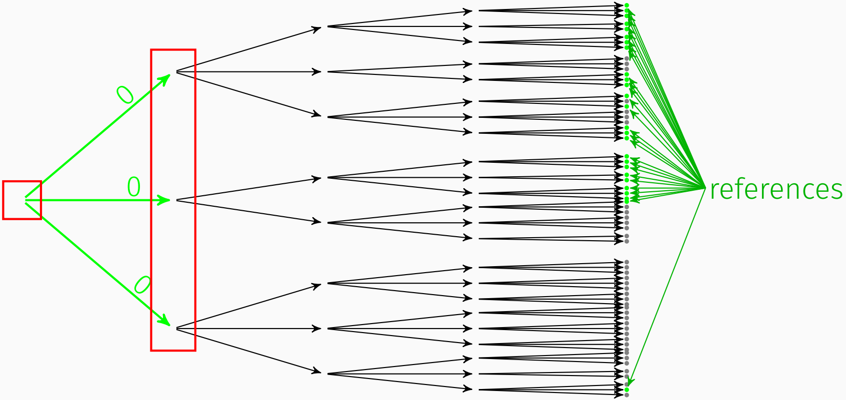
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



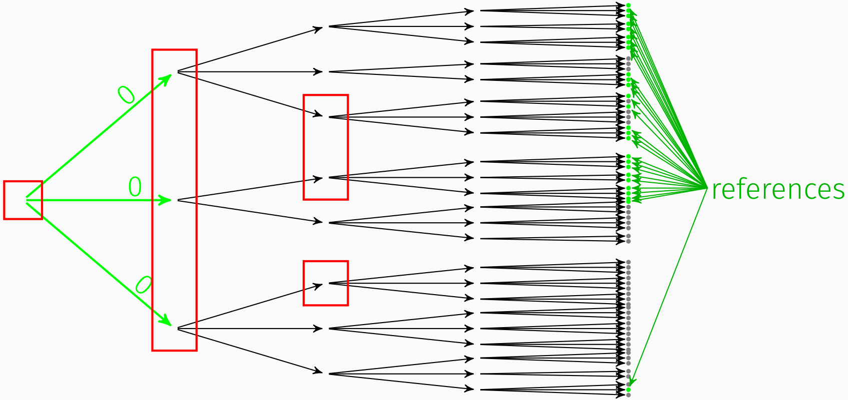
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



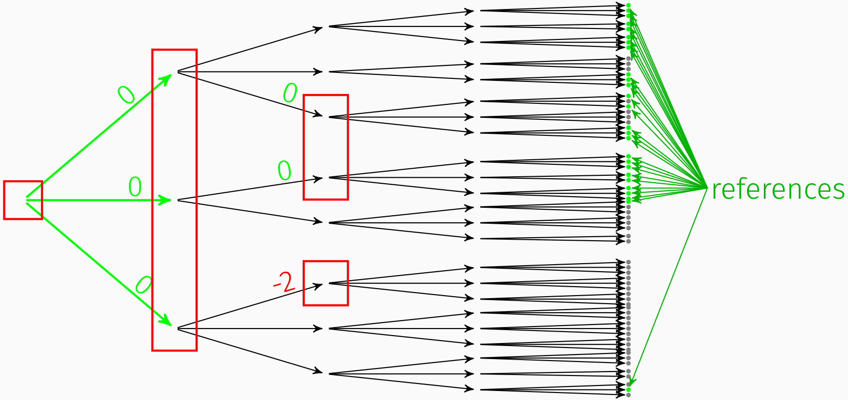
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



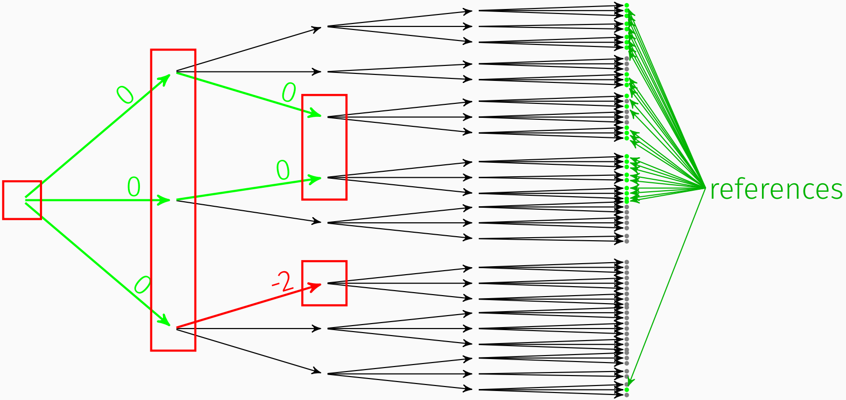
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



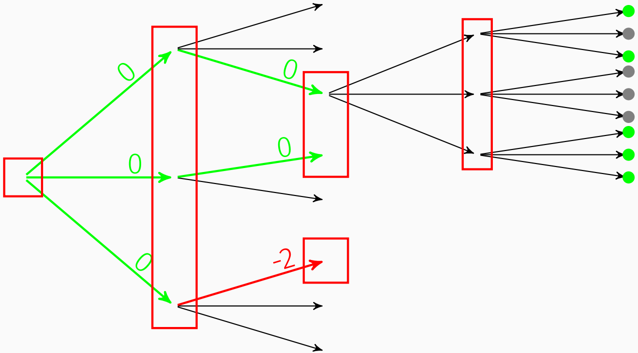
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam

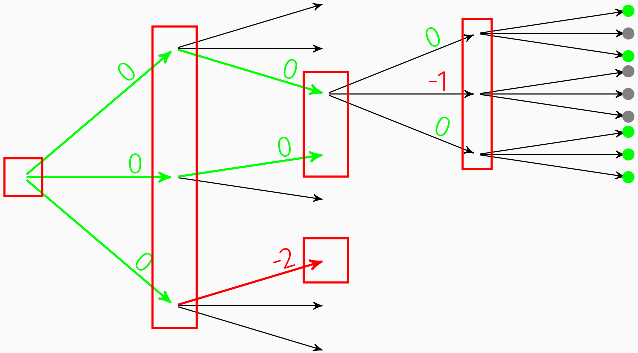


New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Old criterion: the reference falls out of the beam



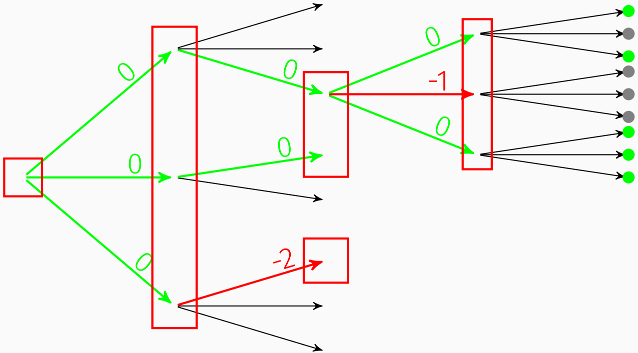
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



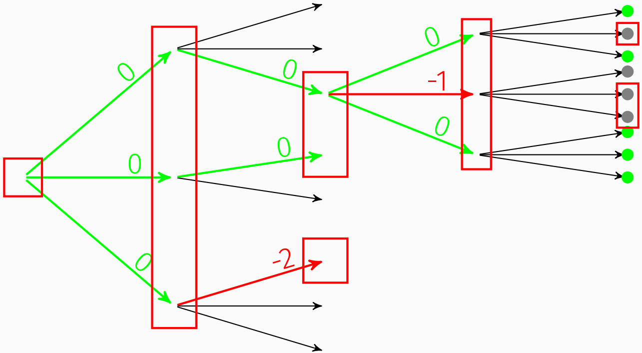
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



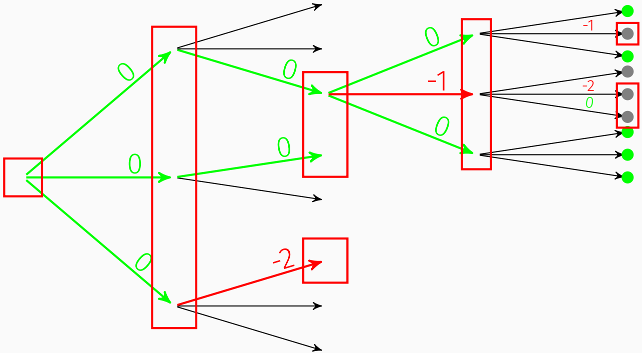
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \Longleftrightarrow \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



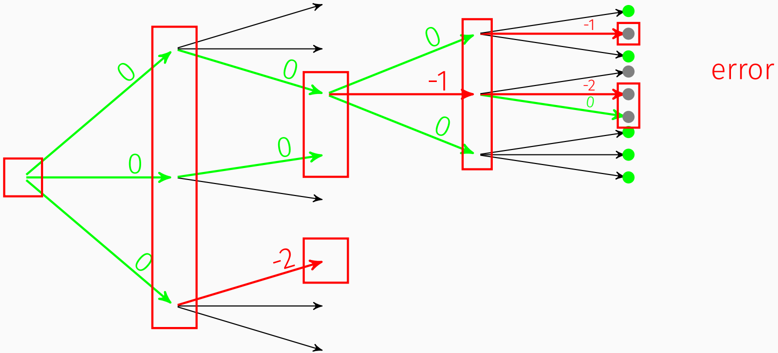
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



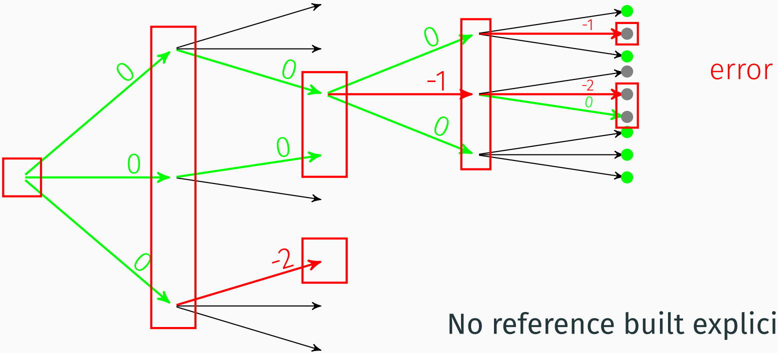
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



No reference built explicitly

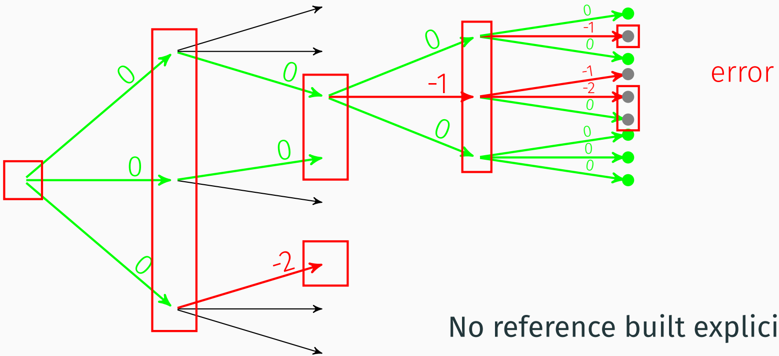
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



No reference built explicitly

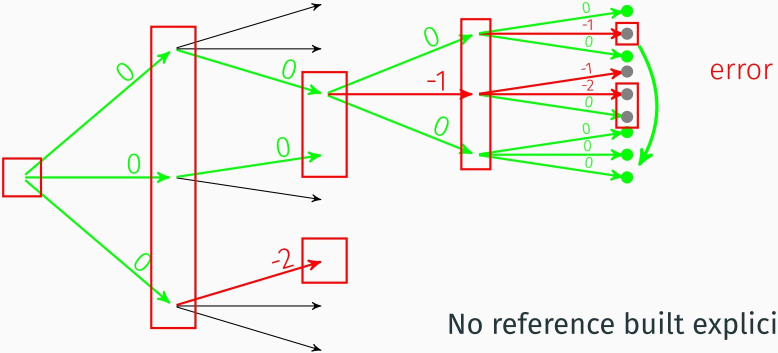
New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle

Old criterion: the reference falls out of the beam



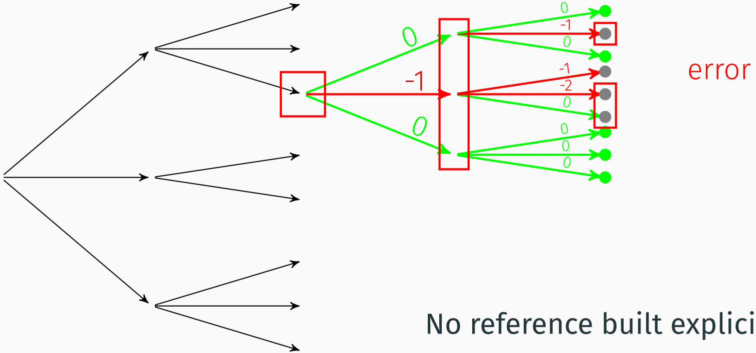
No reference built explicitly

New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle: completeness



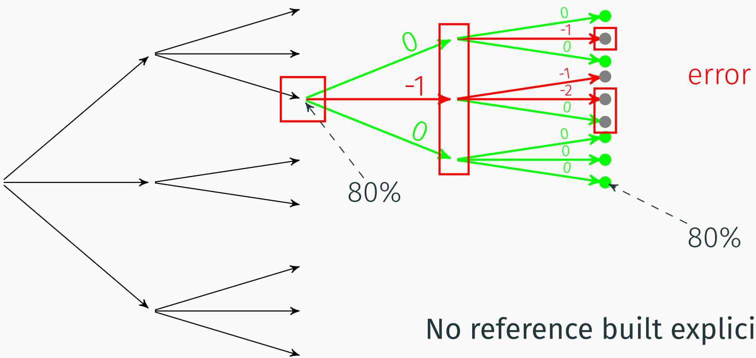
No reference built explicitly

New criterion: no beam hypothesis can produce the reference tree y

For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

Global dynamic oracle: completeness



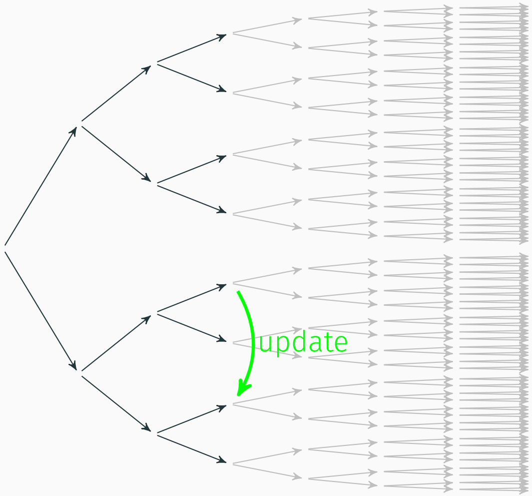
No reference built explicitly

New criterion: no beam hypothesis can produce the reference tree y

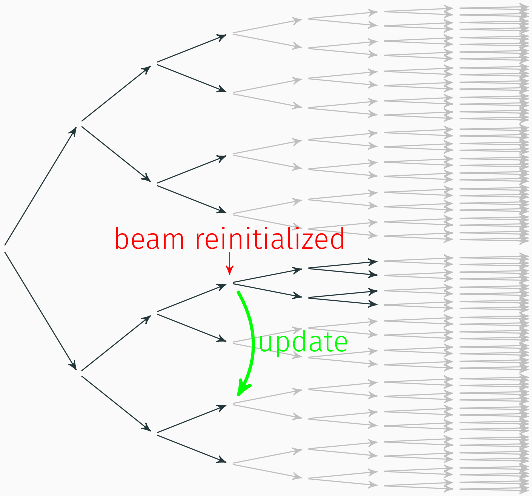
For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$$\textsc{Correct}_y(c'|c) \iff \textsc{Cost}_y(t_1) = \cdots = \textsc{Cost}_y(t_n) = 0$$

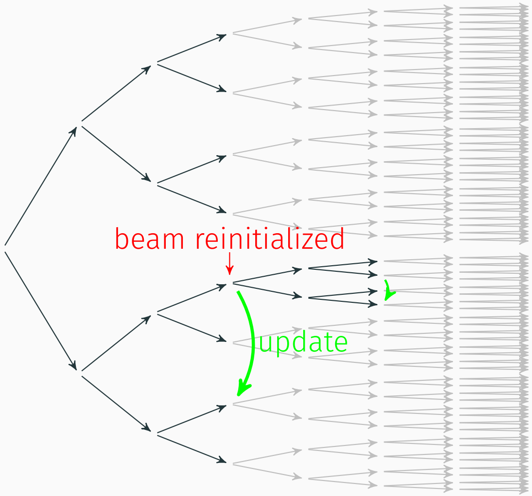
Restart: in suboptimal space



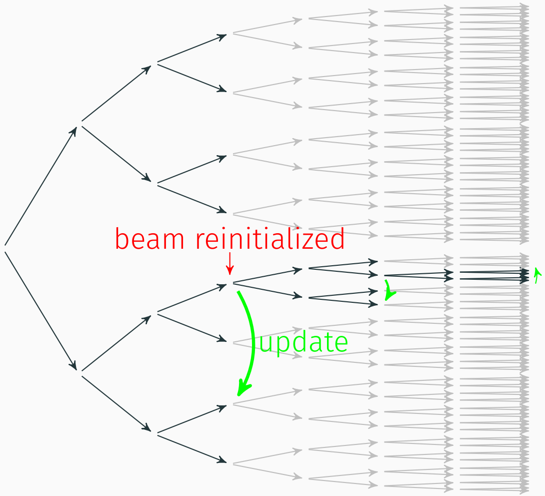
Restart: in suboptimal space



Restart: in suboptimal space



Restart: in suboptimal space



SPMRL (9 languages)

ΔUAS	min	max	average
EARLY	-0.05	+0.45	+0.21
MAXV	-0.02	+0.70	+0.20

French, early update

Quarter	1st	2nd	3rd	4th
Baseline	90.0	85.4	83.1	84.7
Improved	90.0	85.3	**84.2**	**85.1**

Improved accuracy

SPMRL (9 languages)

ΔUAS	min	max	average
EARLY	-0.05	+0.45	+0.21
MAXV	-0.02	+0.70	+0.20

French, early update

Quarter	1st	2nd	3rd	4th
Baseline	90.0	85.4	83.1	84.7
Improved	90.0	85.3	**84.2**	**85.1**

Improved accuracy

SPMRL (9 languages)

ΔUAS	min	max	average
Early	-0.05	+0.45	+0.21
MaxV	-0.02	+0.70	+0.20

French, early update

Quarter	1st	2nd	3rd	4th
Baseline	90.0	85.4	83.1	84.7
Improved	90.0	85.3	**84.2**	**85.1**

on derivation endings

Improved accuracy

SPMRL (9 languages)

ΔUAS	min	max	average
Early	-0.05	+0.45	+0.21
MaxV	-0.02	+0.70	+0.20

French, early update

Quarter	1st	2nd	3rd	4th
Baseline	90.0	85.4	83.1	84.7
Improved	90.0	85.3	**84.2**	**85.1**

on derivation endings

✓ Better convergence

Improved accuracy

SPMRL (9 languages)

ΔUAS	min	max	average
Early	-0.05	+0.45	+0.21
MaxV	-0.02	+0.70	+0.20

French, early update

Quarter	1st	2nd	3rd	4th
Baseline	90.0	85.4	83.1	84.7
Improved	90.0	85.3	**84.2**	**85.1**

on derivation endings

- ✓ Better convergence
- ✓ Better sampling of training configurations

Improved accuracy

SPMRL (9 languages)

ΔUAS	min	max	average
EARLY	-0.05	+0.45	+0.21
MAXV	-0.02	+0.70	+0.20

French, early update

Quarter	1st	2nd	3rd	4th
Baseline	90.0	85.4	83.1	84.7
Improved	90.0	85.3	**84.2**	**85.1**

on derivation endings

- ✓ Better convergence
- ✓ Better sampling of training configurations

- ✓ Unified formalism:

$$\text{Greedy training} = \begin{cases} \text{Beam of size 1} \\ \text{Global dynamic oracle} \\ \text{Restart} \end{cases}$$

Additional benefits of dynamic oracles: partial parses

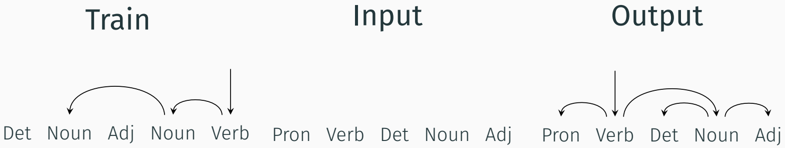
Train

Input

Pron Verb Det Noun Adj

Output

Additional benefits of dynamic oracles: partial parses



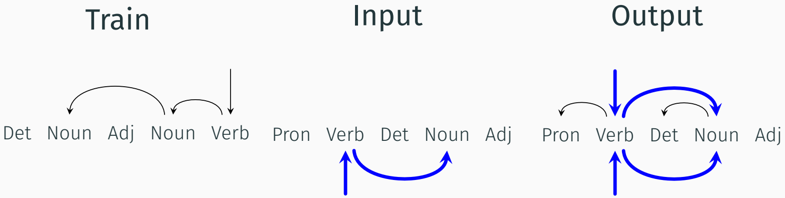
✓ Partial training [NAACL'16]

Additional benefits of dynamic oracles: partial parses

Train | Input | Output

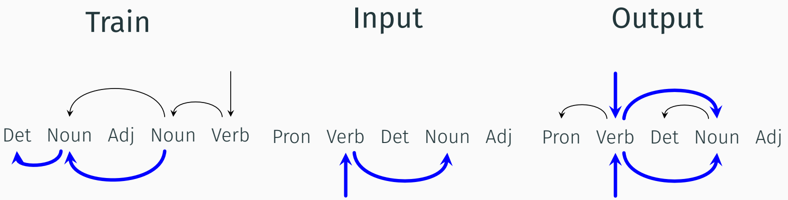
- ✓ Partial training [NAACL'16]
- ✓ Partial prediction

Additional benefits of dynamic oracles: partial parses



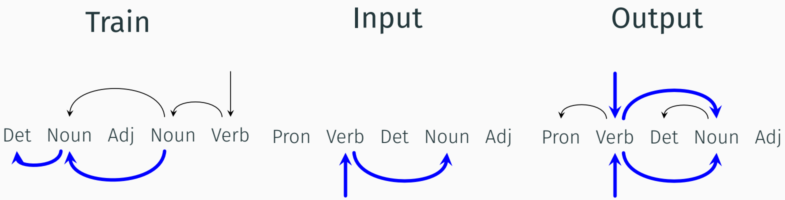
- ✓ Partial training [NAACL'16]
- ✓ Partial prediction
- ✓ Constrained prediction

Additional benefits of dynamic oracles: partial parses



- ✓ Partial training [NAACL'16]
- ✓ Partial prediction
- ✓ Constrained prediction
- ✓ Constrained training

Additional benefits of dynamic oracles: partial parses



- ✓ Partial training [NAACL'16]
- ✓ Partial prediction
- ✓ Constrained prediction
- ✓ Constrained training

... and many other benefits!

↪ training with non-projectivity [NAACL'18]

PanParser

- Extensive use of global dynamic oracles
- Modular architecture

$\hookrightarrow$ Classifier $\times$ transition system $\times$ search strategy $\times$ update strategy $\times$ feature representation $\times \cdots$

- Fair benchmarking: single out each hyperparameter
- State-of-the-art: several strategies already built-in
- Generic framework for structured prediction

$\hookrightarrow$ PoS tagging, semantic parsing, joint predictions...

► `https://perso.limsi.fr/aufrant` ◄

Summary: extensions to the parsing framework

- **Dynamic oracles** make structured training exact
- Identification of **new benefits** of dynamic oracles
- Extension to **global dynamic oracles** with restart
- PanParser: a new **modular** implementation based on a **unified framework**

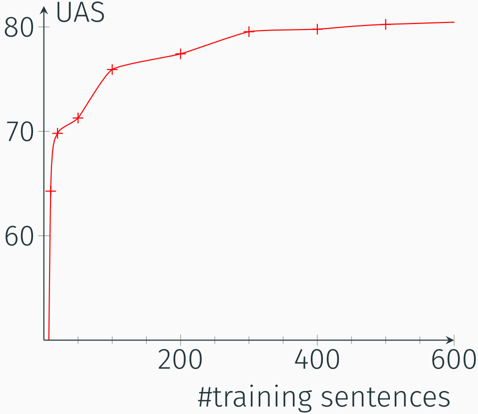
Case study [LREC'16]

Multi-source transfer [McDonald *et al.*, 2011]
$\hookrightarrow$ delexicalized transfer + raw data + parallel data

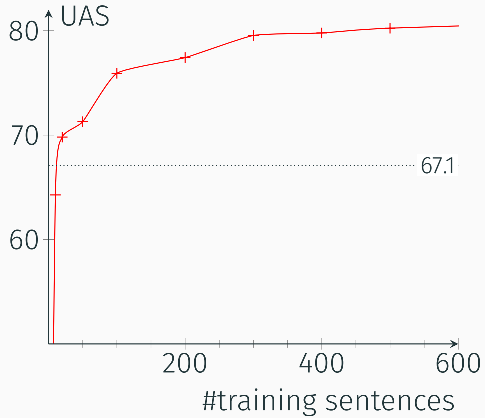
Romance languages → Romanian

Source	fr	it	es	fr+it+es
Delexicalized	60.8	61.5	61.2	61.7
Full transfer	67.0	66.9	67.1	**67.1**
Supervised		82.7		

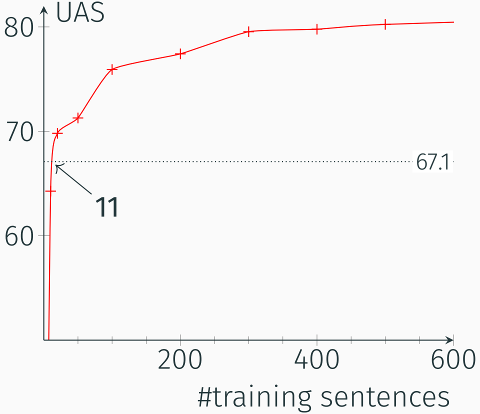
Is transfer really useful?



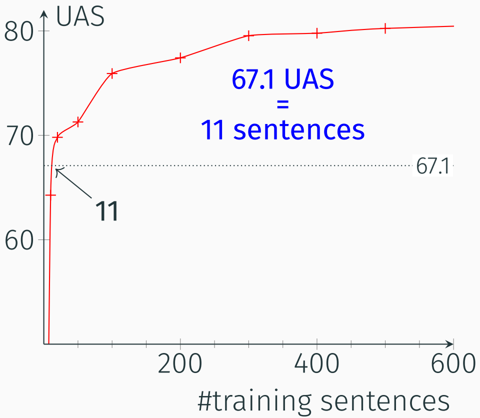
Is transfer really useful?



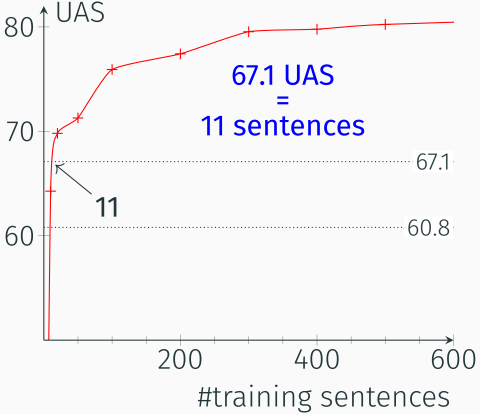
Is transfer really useful?



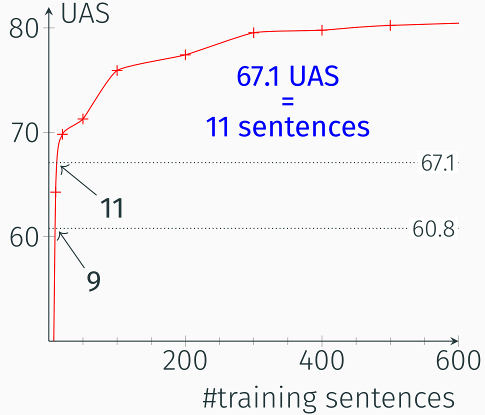
Is transfer really useful?



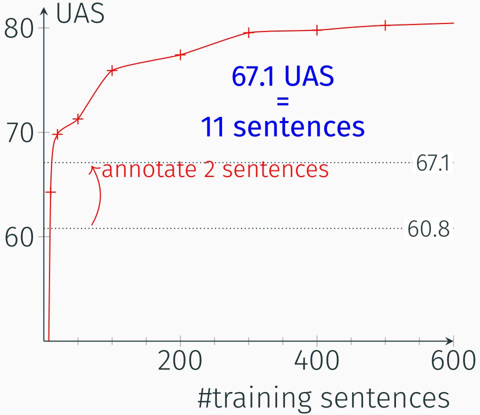
Is transfer really useful?



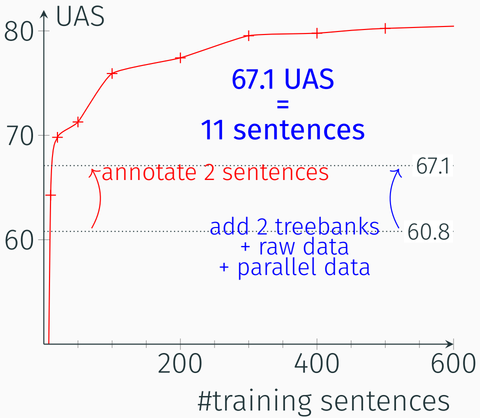
Is transfer really useful?

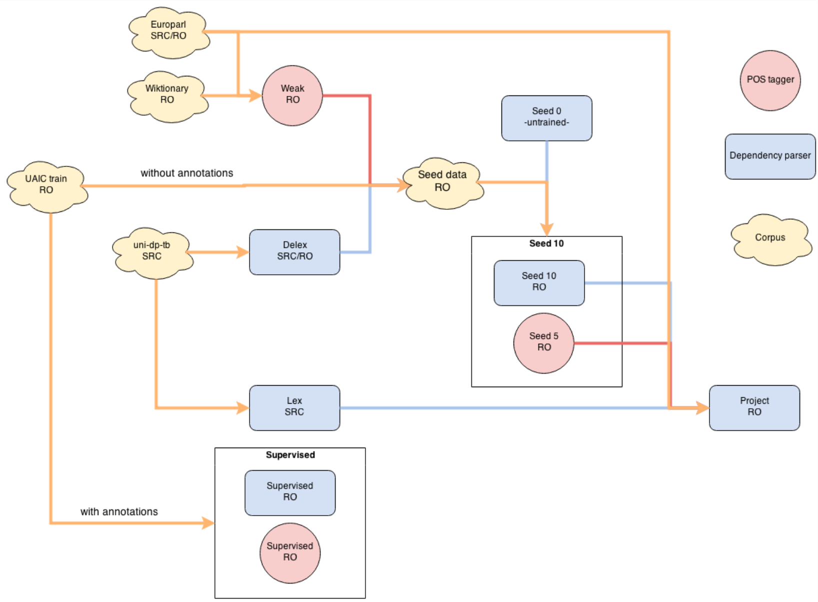


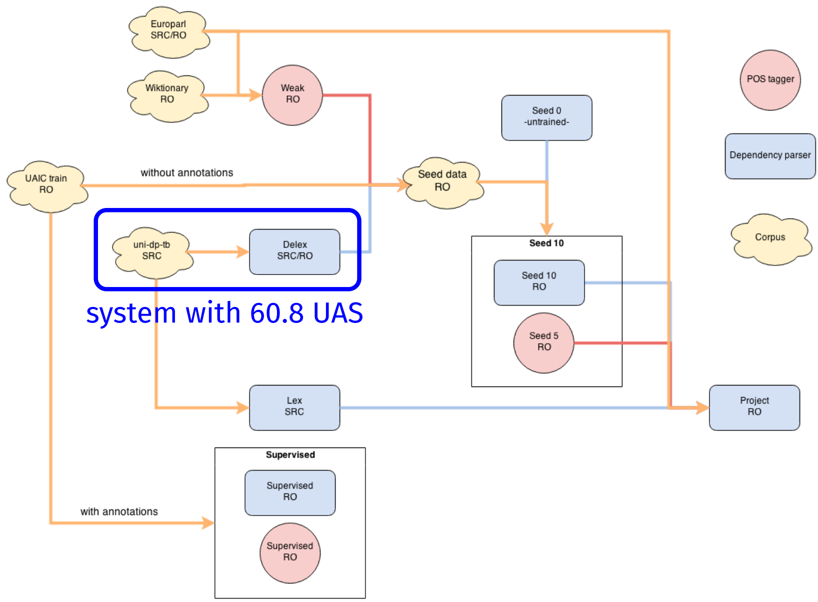
Is transfer really useful?

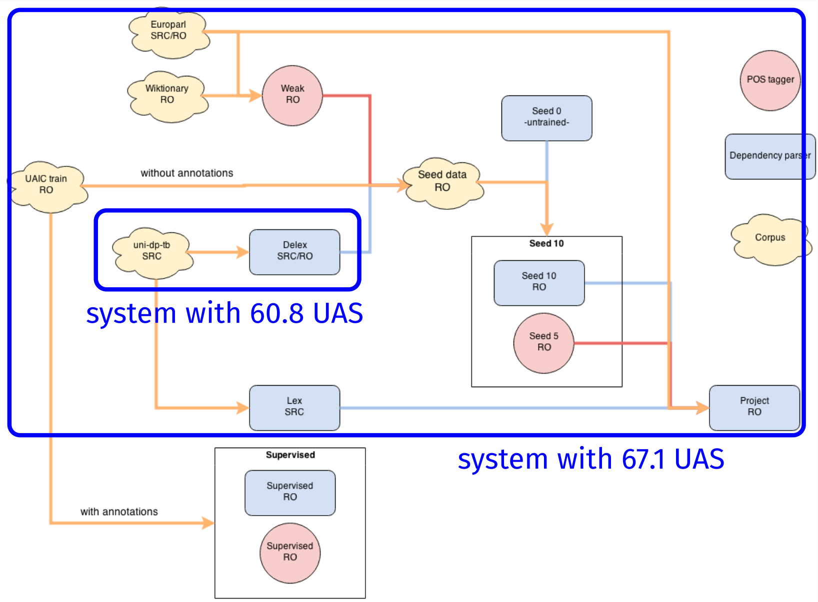


Is transfer really useful?

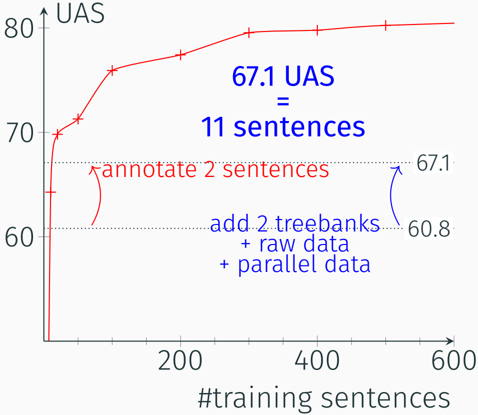



Europarl SRC/RO
Wiktionary RO
Weak RO
Seed 0 -untrained-
POS tagger
Dependency parser
UAIC train RO
without annotations
Seed data RO
Corpus
uni-dp-tb SRC
Delex SRC/RO
Seed 10
Seed 10 RO
Seed 5 RO
Lex SRC
Project RO
Supervised
Supervised RO
Supervised RO
with annotations


Europarl SRC/RO
Wiktionary RO
Weak RO
Seed 0 -untrained-
POS tagger
Dependency parser
UAIC train RO
without annotations
Seed data RO
uni-dp-tb SRC
Delex SRC/RO
Corpus
Seed 10
Seed 10 RO
Seed 5 RO
system with 60.8 UAS
Lex SRC
Project RO
Supervised
Supervised RO
Supervised RO
with annotations


Europarl
SRC/RO
Wiktionary
RO
Weak
RO
POS tagger
Seed 0
-untrained-
Dependency parser
UAIC train
RO
without annotations
Seed data
RO
uni-dp-tb
SRC
Delex
SRC/RO
Seed 10
Seed 10
RO
Corpus
system with 60.8 UAS
Seed 5
RO
Lex
SRC
Project
RO
Supervised
Supervised
RO
system with 67.1 UAS
with annotations
Supervised
RO

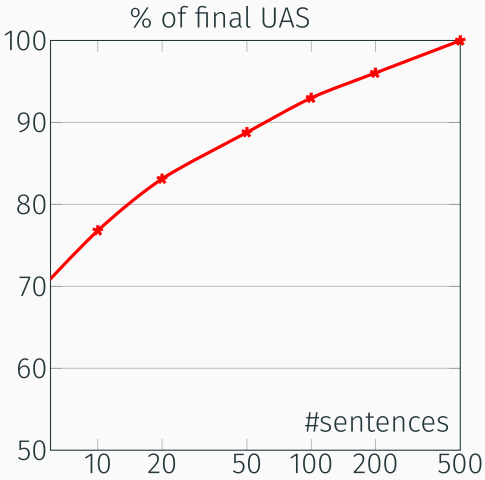
Is transfer really useful?



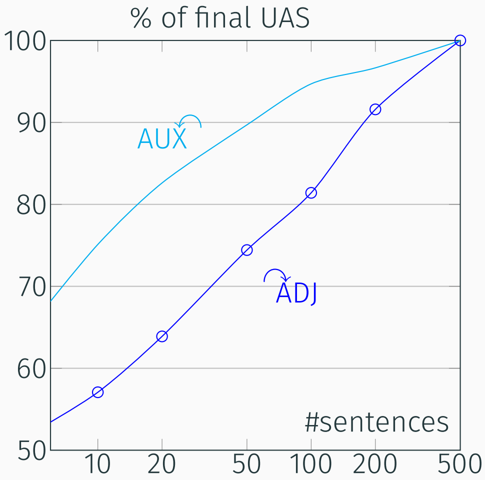
- Better to annotate 11 sentences than using complex transfer methods
- Similar findings in PoS tagging

⇒ Have we underestimated the benefits of monolingual data?

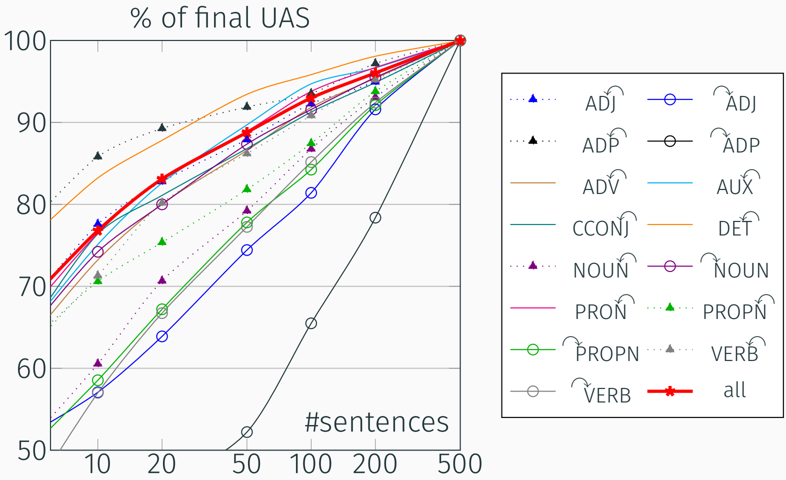
Simple to learn, complex to learn



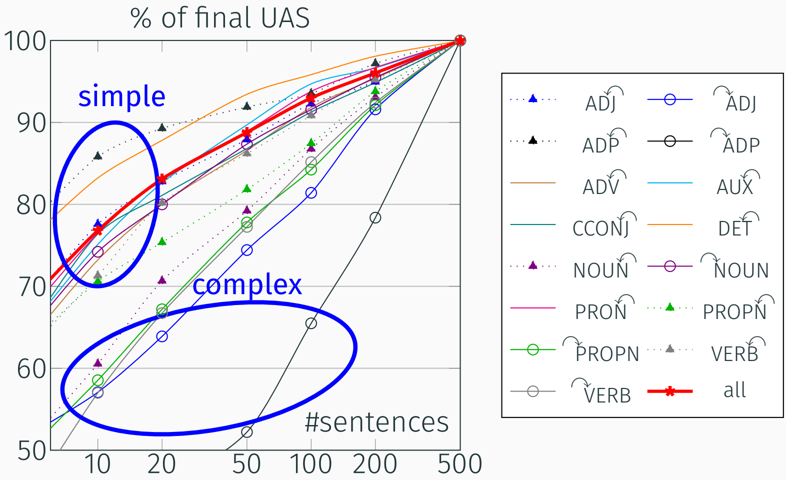
Simple to learn, complex to learn



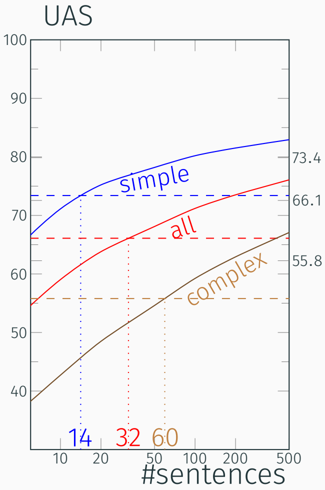
Simple to learn, complex to learn



Simple to learn, complex to learn



Transfer is useful... for complex classes!



- Systematic experiments
  - 56 languages
  - multi-source transfer

- Transfer efficiency can depend:
  - on the language
  - on the type of dependency

↪ **Cross-lingual transfer conveys non-trivial information on complex classes**

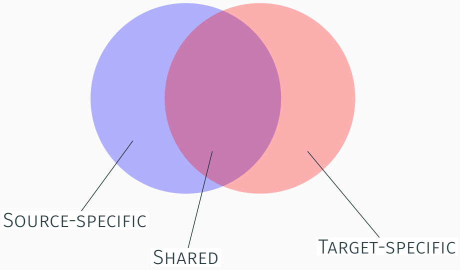
Typology of syntactic information

- 1 language $\rightsquigarrow$ multiple aspects, various influences
- Example: Romanian syntax
  - Word order $\Rightarrow$ as in Romance languages
  - Clitic doubling $\Rightarrow$ as in Spanish
  - Prepositional phrases, subjunctive $\Rightarrow$ as in Bulgarian
  - Double marking of possession $\Rightarrow$ unique property

Typology of syntactic information

- 1 language $\rightsquigarrow$ multiple aspects, various influences

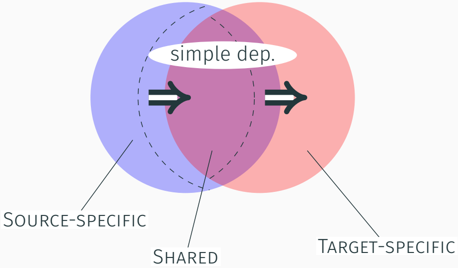
- **Example: Romanian syntax**
  - Word order $\Rightarrow$ as in Romance languages
  - Clitic doubling $\Rightarrow$ as in Spanish
  - Prepositional phrases, subjunctive $\Rightarrow$ as in Bulgarian
  - Double marking of possession $\Rightarrow$ unique property



Typology of syntactic information

- 1 language $\rightsquigarrow$ multiple aspects, various influences

- **Example: Romanian syntax**
  - Word order ⇒ as in Romance languages
  - Clitic doubling ⇒ as in Spanish
  - Prepositional phrases, subjunctive ⇒ as in Bulgarian
  - Double marking of possession ⇒ unique property



Cascading: an example

Her	boyfriend	broke	up	on	February	14
DET	NOUN	VERB	ADP	ADP	NOUN	NUM

Submodels:

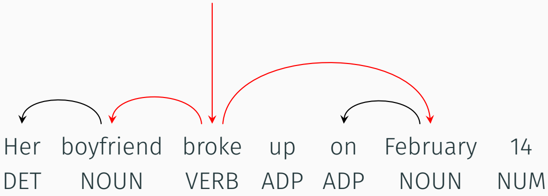
Cascading: an example

Her	boyfriend	broke	up	on	February	14
DET	NOUN	VERB	ADP	ADP	NOUN	NUM

Submodels:

✓ target bootstrap: **simple dependencies** (determiner, preposition)

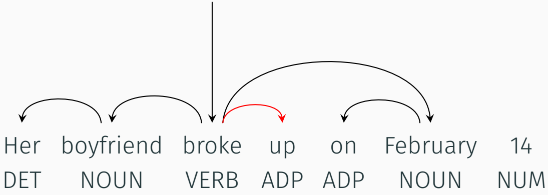
Cascading: an example



Submodels:

- ✓ target bootstrap: **simple dependencies** (determiner, preposition)
- ✓ transfer from French: **main structure** (subject, verb modifier)

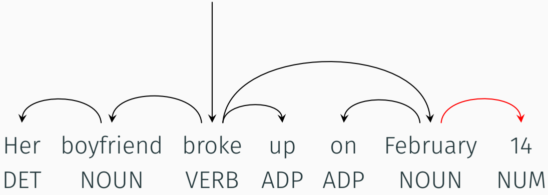
Cascading: an example



Submodels:

- ✓ target bootstrap: **simple dependencies** (determiner, preposition)
- ✓ transfer from French: **main structure** (subject, verb modifier)
- ✓ transfer from German: **influences** (verbal postposition)

Cascading: an example



Submodels:

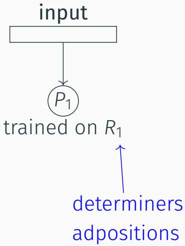
- ✓ target bootstrap: **simple dependencies** (determiner, preposition)
- ✓ transfer from French: **main structure** (subject, verb modifier)
- ✓ transfer from German: **influences** (verbal postposition)
- ✓ target-side **tuning**

Adapting an ensembling method: the cascading architecture

input

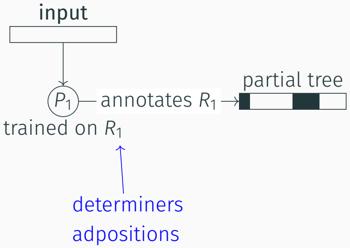
- 1 parser $\rightsquigarrow$ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - $\hookrightarrow$ by annotating a target sample
  - $\hookrightarrow$ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - $\hookrightarrow$ specialized models
  - $\hookrightarrow$ no redundancy

Adapting an ensembling method: the cascading architecture



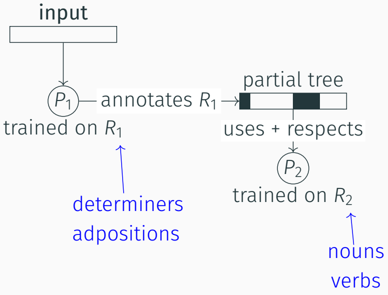
- 1 parser $\rightsquigarrow$ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - ↪ by annotating a target sample
  - ↪ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - ↪ specialized models
  - ↪ no redundancy

Adapting an ensembling method: the cascading architecture



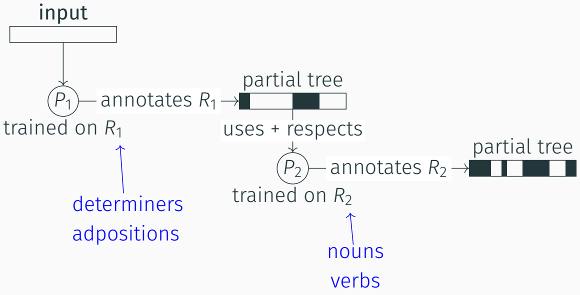
- 1 parser $\rightsquigarrow$ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - $\hookrightarrow$ by annotating a target sample
  - $\hookrightarrow$ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - $\hookrightarrow$ specialized models
  - $\hookrightarrow$ no redundancy

Adapting an ensembling method: the cascading architecture



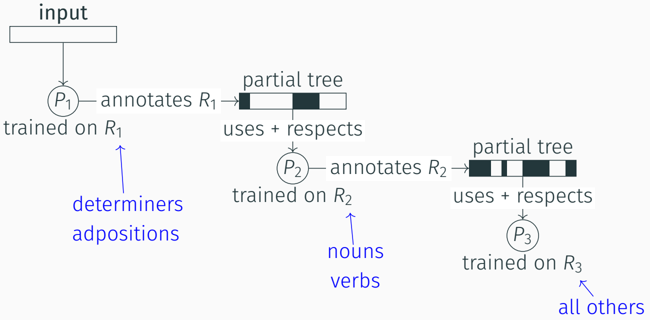
- 1 parser ⇝ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - ↪ by annotating a target sample
  - ↪ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - ↪ specialized models
  - ↪ no redundancy

Adapting an ensembling method: the cascading architecture



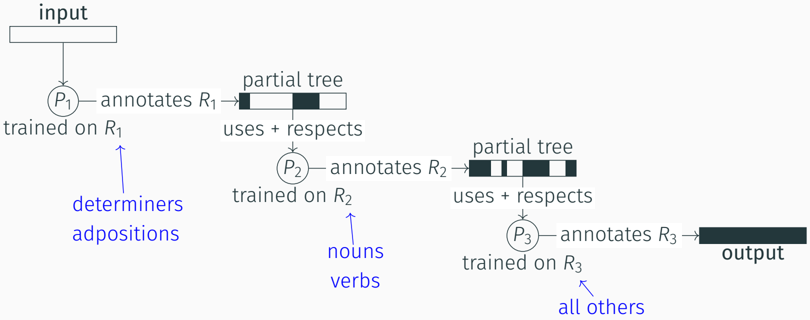
- 1 parser $\rightsquigarrow$ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - ↪ by annotating a target sample
  - ↪ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - ↪ specialized models
  - ↪ no redundancy

Adapting an ensembling method: the cascading architecture



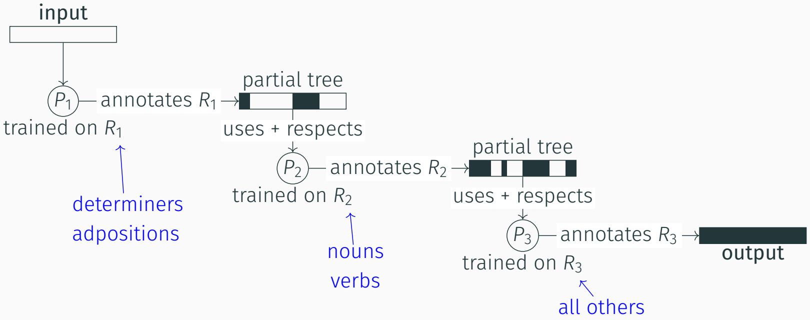
- 1 parser $\rightsquigarrow$ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - ↪ by annotating a target sample
  - ↪ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - ↪ specialized models
  - ↪ no redundancy

Adapting an ensembling method: the cascading architecture



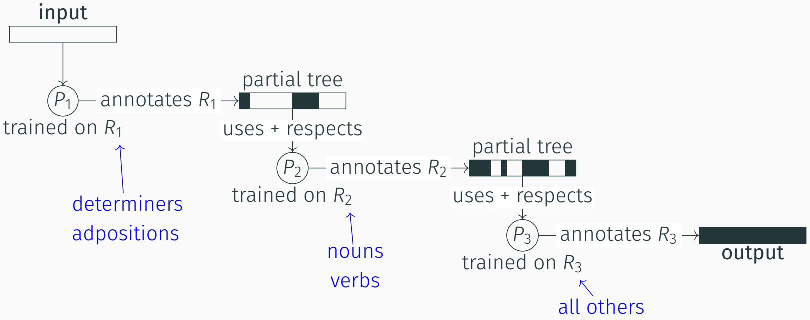
- 1 parser ⇝ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - ↪ by annotating a target sample
  - ↪ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - ↪ specialized models
  - ↪ no redundancy

Adapting an ensembling method: the cascading architecture



- 1 parser $\rightsquigarrow$ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - ↪ by annotating a target sample
  - ↪ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - ↪ specialized models
  - ↪ no redundancy

Adapting an ensembling method: the cascading architecture

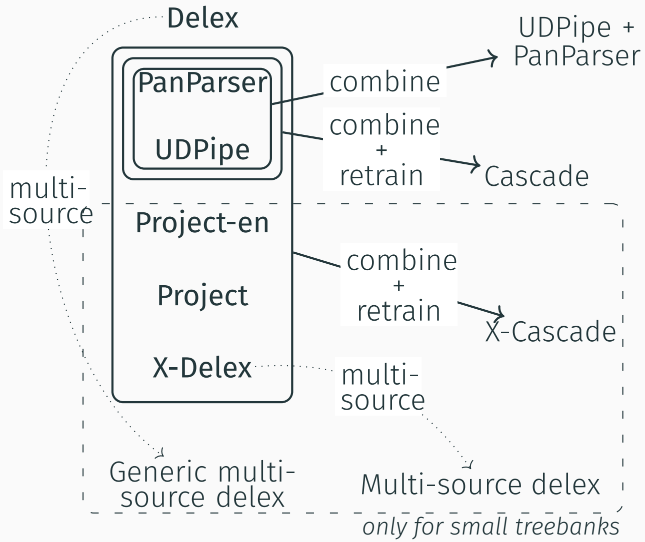


- 1 parser ⇝ a sequence of **partial parsers** (P_1, P_2, P_3)
- Estimating **regions of competence** (R_1, R_2, R_3)
  - ↪ by annotating a target sample
  - ↪ using similarity metrics
- **Optimized training** thanks to dynamic oracles
  - ↪ specialized models
  - ↪ no redundancy

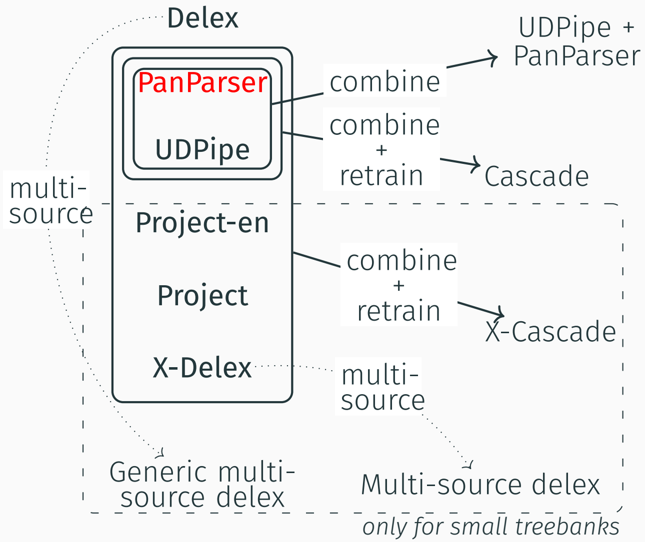
Shared task evaluation [CoNLL'17]

- End-to-end parsing: from raw text to dependencies
- Multilingual dataset (UD)
  ↪ diverse language families, domains, treebank sizes
- Evaluation in realistic conditions
  ↪ blind test, surprise languages
- 33 teams: highly competitive
- Our focus: small treebanks

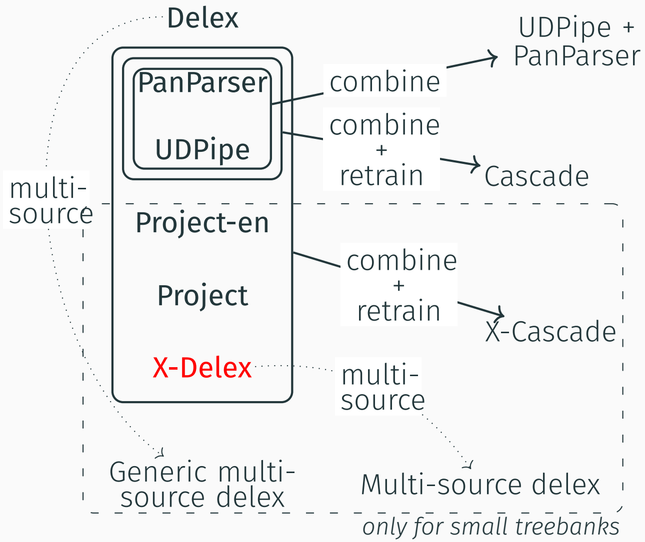
All-in-one system



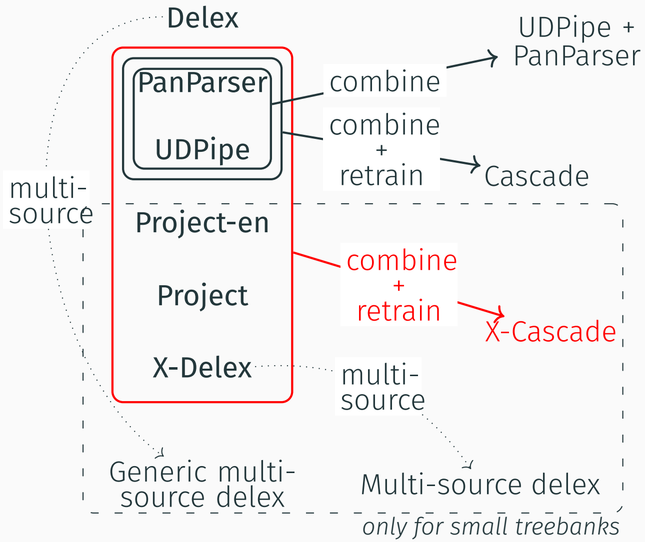
All-in-one system



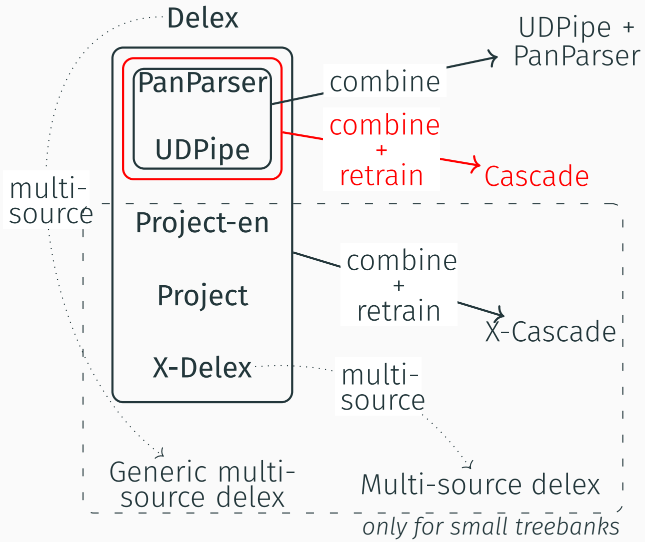
All-in-one system



All-in-one system



All-in-one system



Shared task results

Positive impact of...

- ✓ PanParser
- ✓ WALS-based transfer
- ✓ Transfer cascades
- ✓ Monolingual cascades

Error analysis: perspectives for improvements

- Tiny target samples: poor estimation of regions
- Unreliable PoS: can delexicalized models still contribute?
- Unveiled remaining annotation inconsistencies

Summary: a new transfer framework

- The benefits of **target samples** have been underestimated
- **Characterize the information** conveyed by target samples and by each source
- Cascading architecture: **sequential combination** of partial parsers
- Shared task evaluation: **validates** all contributions (PanParser, WALS, cascades)

Outline

Conclusions

- **Main purpose: improve the coverage** of cross-lingual transfer
  - ↪ by adding more flexibility regarding leveraged resources

✓ Make **new resources** usable ($\rightsquigarrow$ typological knowledge)
  - ↪ avoid systematic errors
  - ↪ extend candidate sources

✓ Make **any resource combination** possible ($\rightsquigarrow$ cascading)
  - ↪ including target samples, distant sources...
  - ↪ fine-grained targeting

- Additional improvements in **transition-based parsing**
  - ↪ to reach the required degree of flexibility

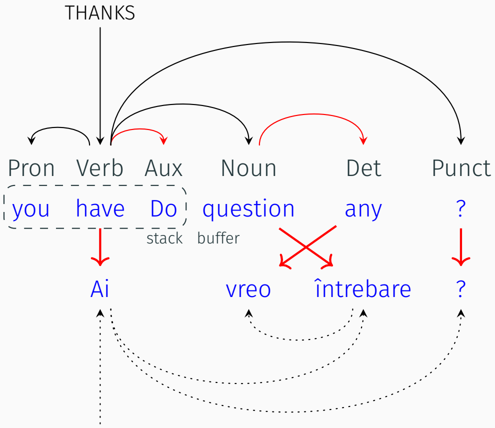
Perspectives

Cross-lingual transfer

- Cascading experiments with other metrics
- Application to other tasks
- Better use of lexical similarities

Transition-based parsing

- Deriving new dynamic oracles
- Better control on information extracted at training time
- Divide-and-conquer cascades

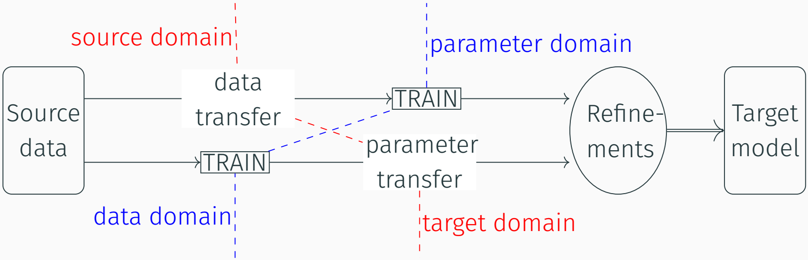

THANKS
Pron
Verb
Aux
Noun
Det
Punct
you
have
Do
question
any
?
stack
buffer
Ai
vreo
întrebare
?

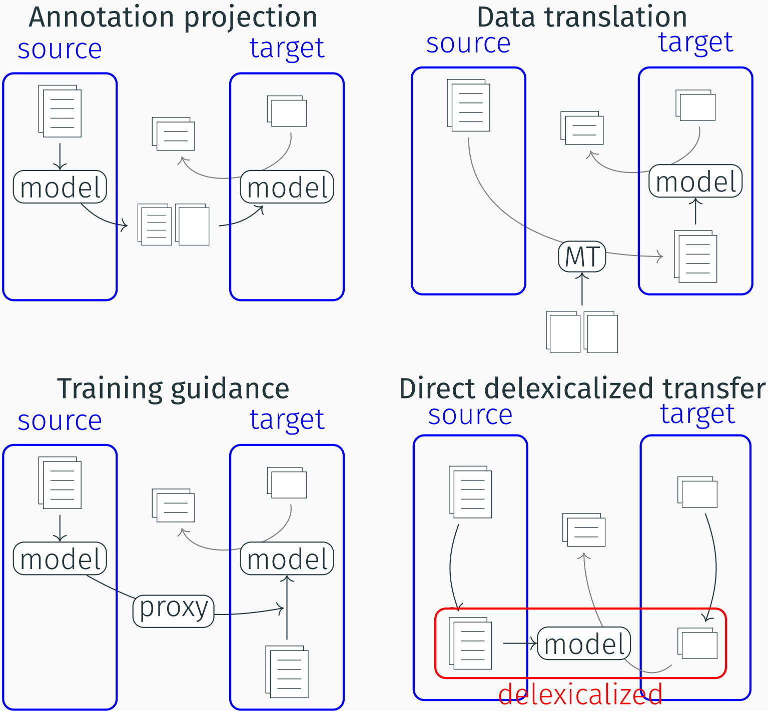
Take-home messages

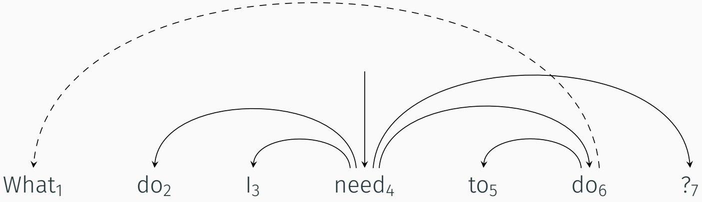
- Modern NLP: many successful systems... for **a handful** of languages
- **Cross-lingual transfer**: a promising approach, yet not always the best one
- The key to low-resourced NLP: exploit **all resources** together (typology, samples...)
- **Dynamic oracles** have taken transition-based parsing to the next level

Additional tables and figures

Chapter 2


source domain
parameter domain
Source data
data transfer
TRAIN
TRAIN
parameter transfer
Refine-ments
Target model
data domain
target domain


Annotation projection
source
target
model
model
Data translation
source
target
model
MT
Training guidance
source
target
model
model
proxy
Direct delexicalized transfer
source
target
model
delexicalized



Indices	1	2	3	4	5	6	7
Words	What	do	I	need	to	do	?
Heads	6	4	4	0	6	4	4
Labels	dobj	aux	nsubj	root	mark	xcomp	punct

Chapter 3

ArcStandard

Shift	$(\sigma,$	$b\vert\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert b,$	$\beta,$	$P)$	
Left	$(\sigma\vert s'\vert s,$	$\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert s,$	$\beta,$	$P + (s \to s'))$	if s' is a word
Right	$(\sigma\vert s'\vert s,$	$\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert s',$	$\beta,$	$P + (s' \to s))$	

ArcEager

Shift	$(\sigma,$	$b\vert\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert b,$	$\beta,$	$P)$	if b is a word
Left	$(\sigma\vert s,$	$b\vert\beta,$	$P)$	$\Rightarrow$	$(\sigma,$	$b\vert\beta,$	$P + (b \to s))$	if s is a word and s is unattached
Right	$(\sigma\vert s,$	$b\vert\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert s\vert b,$	$\beta,$	$P + (s \to b))$	
Reduce	$(\sigma\vert s,$	$\beta,$	$P)$	$\Rightarrow$	$(\sigma,$	$\beta,$	$P)$	if s is attached

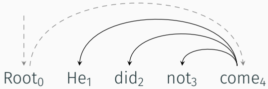
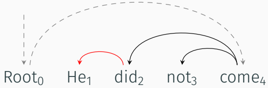
ArcHybrid

Shift	$(\sigma,$	$b\vert\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert b,$	$\beta,$	$P)$	if b is a word
Left	$(\sigma\vert s,$	$b\vert\beta,$	$P)$	$\Rightarrow$	$(\sigma,$	$b\vert\beta,$	$P + (b \to s))$	if s is a word
Right	$(\sigma\vert s'\vert s,$	$\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert s',$	$\beta,$	$P + (s' \to s))$	

SwapStandard

Shift	$(\sigma,$	$b\vert\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert b,$	$\beta,$	$P)$	
Left	$(\sigma\vert s'\vert s,$	$\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert s,$	$\beta,$	$P + (s \to s'))$	if s' is a word
Right	$(\sigma\vert s'\vert s,$	$\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert s',$	$\beta,$	$P + (s' \to s))$	
Swap	$(\sigma\vert s'\vert s,$	$\beta,$	$P)$	$\Rightarrow$	$(\sigma\vert s,$	$s'\vert\beta,$	$P)$	if s' is a word and $s < s'$

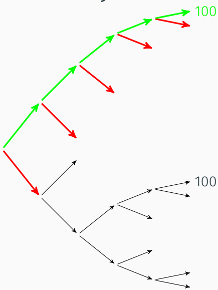
UAS	ArcEager	ArcStandard
No Root	84.35	84.41
Root in first position	83.67	84.44
Root in last position	84.35	84.38

Derivation						Resulting parse
Shift$_1$	**Shift$_2$**	Shift$_3$	Left$_{3\leftarrow 4}$	Left$_{2\leftarrow 4}$	Left$_{1\leftarrow 4}$	
Shift$_1$	Left$_{1\leftarrow 2}$	Shift$_2$	Shift$_3$	Left$_{3\leftarrow 4}$	Left$_{2\leftarrow 4}$	
Shift$_1$	Right$_{1\rightarrow 2}$	Reduce$_2$	Shift$_3$	Left$_{3\leftarrow 4}$	Left$_{1\leftarrow 4}$	

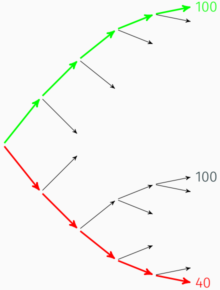
Classifier	UAS	Speed (sent/s)
Averaged perceptron (MaltParser)	89.9	560
Feed-forward neural network	92.0	1,013

	Standard templates
1 word	w, p and wp for S_0, N_0, N_1, N_2
2 words	$wp{\cdot}wp$, $wp{\cdot}w$, $w{\cdot}wp$, $wp{\cdot}p$, $p{\cdot}wp$, $w{\cdot}w$ and $p{\cdot}p$ for $S_0{\cdot}N_0$; $N_0p{\cdot}N_1p$
3 words	$p{\cdot}p{\cdot}p$ for $N_0{\cdot}N_1{\cdot}N_2$, $S_0{\cdot}N_0{\cdot}N_1$, $S_{0h}{\cdot}S_0{\cdot}N_0$, $S_0{\cdot}S_{0l}{\cdot}N_0$, $S_0{\cdot}S_{0r}{\cdot}N_0$, $S_0{\cdot}N_0{\cdot}N_{0l}$
	New templates with rich non-local features
Distance	$S_0w{\cdot}d$, $S_0p{\cdot}d$, $N_0w{\cdot}d$, $N_0p{\cdot}d$; $S_0w{\cdot}N_0w{\cdot}d$, $S_0p{\cdot}N_0p{\cdot}d$
Valency	S_0wv_l, S_0pv_l, S_0wv_r, S_0pv_r, N_0wv_l, N_0pv_l
Unigrams	w and p for S_{0h}, S_{0l}, S_{0r}, N_{0l}; l for S_0, S_{0l}, S_{0r}, N_{0l}
Third-order	w and p for S_{0h2}, S_{0l2}, S_{0r2}, N_{0l2}; l for S_{0h}, S_{0l2}, S_{0r2}, N_{0l2};
	$p{\cdot}p{\cdot}p$ for $S_0{\cdot}S_{0h}{\cdot}S_{0h2}$, $S_0{\cdot}S_{0l}{\cdot}S_{0l2}$, $S_0{\cdot}S_{0r}{\cdot}S_{0r2}$, $N_0{\cdot}N_{0l}{\cdot}N_{0l2}$
Label set	S_0ws_l, S_0ps_l, S_0ws_r, S_0ps_r, N_0ws_l, N_0ps_l

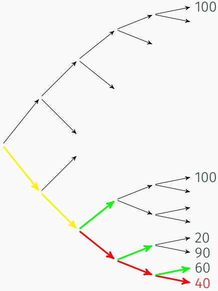
Greedy static



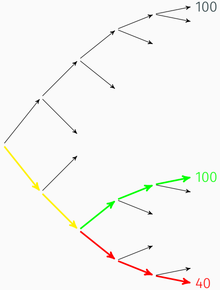
Beam static



Greedy dynamic



Beam non-deterministic



UAS	Local [train]	Global [train]
Local [test]	89.04	87.07
Global [test]	79.34	92.27

Update criterion	Convergence time		UAS
Full update	1 it.	0.4 h	79.14
Early update	38 it.	15.4 h	92.09
Max-violation	12 it.	5.5 h	92.18

UAS	Locally normalized	Globally normalized
Beam size = 1	92.95	–
Beam size = 32	93.59	94.61

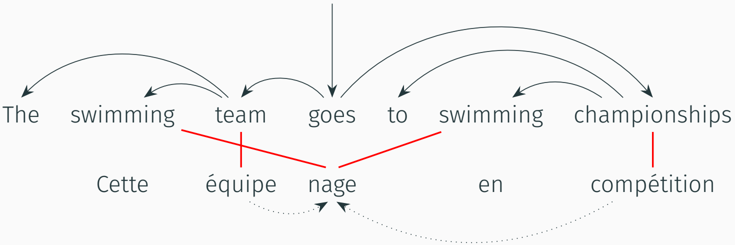
UAS	Static oracle	Dynamic oracle
Gold space training	89.88	90.18
Suboptimal space training	–	90.96

SHIFT	$(\sigma,\quad b \mid \beta)$	$\sigma \curvearrowright b$	$\rightsquigarrow$	b if h_b^* is in stack
	$(\sigma,\quad b \mid \beta)$	$\sigma \curvearrowleft b$	$\rightsquigarrow$	children of b that are in stack and unattached
LEFT	$(\sigma \mid s, b \mid \beta)$	$s \curvearrowleft \beta$	$\rightsquigarrow$	s if h_s^* is in buffer but not on top
	$(\sigma \mid s,\quad \beta)$	$s \curvearrowright \beta$	$\rightsquigarrow$	children of s that are in buffer
RIGHT	$(\sigma,\quad b \mid \beta)$	$b \curvearrowleft \beta$	$\rightsquigarrow$	b if h_b^* is in buffer but not on top
	$(\sigma \mid s, b \mid \beta)$	$\sigma \curvearrowright b$	$\rightsquigarrow$	b if h_b^* is in stack but not on top
	$(\sigma,\quad b \mid \beta)$	$\sigma \curvearrowleft b$	$\rightsquigarrow$	children of b that are in stack and unattached
REDUCE	$(\sigma \mid s,\quad \beta)$	$s \curvearrowright \beta$	$\rightsquigarrow$	children of s that are in buffer

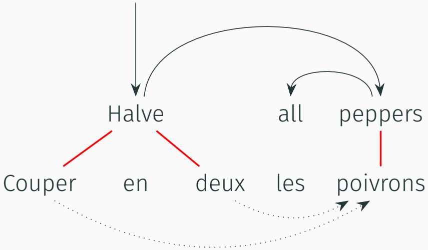
UAS	ArcStandard	ArcHybrid
SLSTM – Static	93.04	92.78
SLSTM – Dynamic	–	93.56

Chapter 4

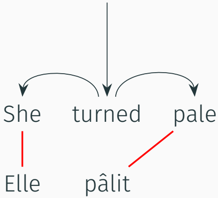
Many-to-one alignment



One-to-many alignment



Unaligned word



Data space transfer

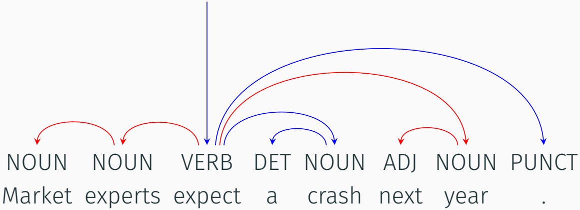
	Target	de	en	es	fr	sv
Supervised	standard	80.34	92.11	83.65	82.17	85.97
	coarse PoS	78.38	91.46	82.30	82.30	84.52
Direct delexicalized transfer (coarse PoS)	de	70.84	45.28	48.90	49.09	52.24
	en	48.60	82.44	56.25	58.47	59.42
	es	47.16	47.31	71.45	62.39	54.63
	fr	46.77	47.94	62.66	73.71	54.89
	sv	52.53	48.24	52.95	55.02	74.55
Annotation projection	de	–	53.80	61.34	62.32	68.20
	en	63.52	–	63.18	67.04	67.74
	es	60.65	50.10	–	68.81	65.79
	fr	62.49	53.88	68.15	–	64.83
	sv	63.83	52.36	63.29	66.12	–
Treebank translation	de	–	58.60	61.00	63.45	67.88
	en	62.67	–	64.58	68.45	68.16
	es	57.13	52.65	–	69.37	63.55
	fr	61.41	56.83	68.97	–	62.56
	sv	61.73	52.13	62.34	64.50	–

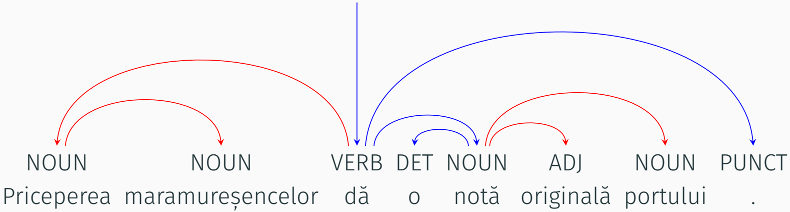
Parameter space transfer (with a target treebank and a bilingual lexicon)

Target	cs	de	es	fi	fr	ga	hu	it	sv	μ
Target only	43.1	47.3	60.3	46.4	56.2	59.4	48.4	65.4	52.6	53.2
Guidance	49.6	59.2	66.4	49.5	63.2	59.5	50.5	69.9	61.4	58.8
Joint learning	55.2	61.2	69.1	51.4	65.3	60.6	51.2	71.2	61.4	60.7
Joint + guidance	55.7	61.8	70.5	51.5	67.2	61.1	51.0	71.3	62.5	61.4

Parameter space transfer (with parallel and raw data)

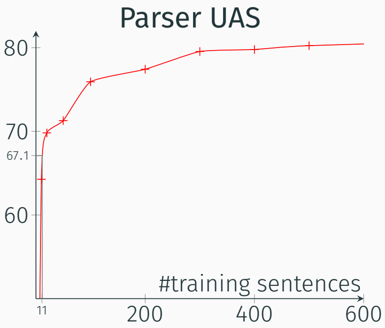
Target	de	es	fr	it	ko	pt	sv	μ
Supervised	81.65	83.92	83.51	85.47	90.42	85.67	85.59	85.18
Direct transfer	58.56	68.72	71.13	70.74	38.55	69.82	70.59	64.02
Guidance	73.92	75.21	76.14	77.55	59.71	76.30	78.91	73.96
Guidance + unlabeled	74.30	75.53	76.53	77.74	59.89	76.65	79.27	74.27

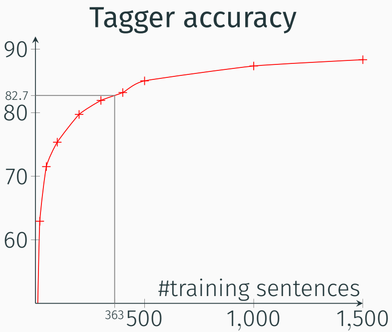

NOUN NOUN VERB DET NOUN ADJ NOUN PUNCT
Market experts expect a crash next year .


NOUN NOUN VERB DET NOUN ADJ NOUN PUNCT
Priceperea maramureșencelor dă o notă originală portului .

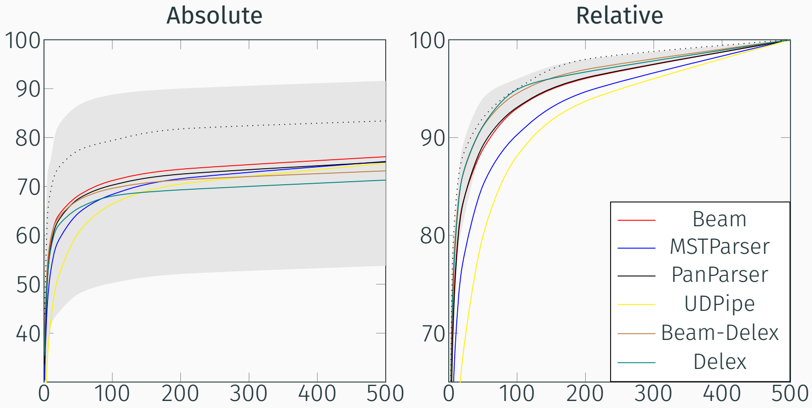
Chapter 5

Source	fr	it	es	fr+it+es
Delexicalized	60.8	61.5	61.2	61.7
Full transfer	67.0	66.9	67.1	67.1
Supervised		82.7		


Parser UAS
80
70
67.1
60
#training sentences
11
200
400
600


Tagger accuracy
90
82.7
80
70
60
#training sentences
363
500
1,000
1,500

Trainset	10 sentences	500 sentences	Full UD
UDPipe	22.4‖55.5‖66.6‖**42.5**	53.0‖84.8‖90.2‖**74.7**	66.4‖89.0‖92.7‖**83.2**
PanParser	41.4‖69.3‖75.6‖**57.7**	53.8‖83.4‖91.6‖**75.0**	58.0‖87.5‖93.4‖**81.2**
Delex	41.3‖70.6‖75.1‖**57.2**	50.9‖81.7‖85.7‖**71.3**	51.0‖83.8‖87.7‖**74.3**
MSTParser	38.1‖62.7‖68.2‖**52.8**	57.6‖81.2‖86.9‖**75.1**	65.8‖86.7‖90.6‖**83.4**
Beam	42.4‖69.8‖76.8‖**59.0**	56.0‖84.2‖91.1‖**76.1**	61.5‖88.2‖93.7‖**82.6**
Beam-Delex	41.5‖70.6‖77.3‖**59.5**	53.8‖83.5‖87.1‖**73.2**	55.9‖85.6‖88.6‖**76.8**



Absolute
Relative
100
90
80
70
60
50
40
0
100
200
300
400
500
100
90
80
70
0
100
200
300
400
500
Beam
MSTParser
PanParser
UDPipe
Beam-Delex
Delex

UAS	30	40	50	60	70	75
Parsing capacity (sentences)	1	2	4	12	77	401
Annotation cost (euros)	10	20	40	120	770	4,010
Romanian trainset size	1	2	3	9	53	410

	All	ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	NUM	PART	PRON	PROPN	SCONJ	VERB
KL-Beam	66.1	72.1	73.4	66.3	72.7	63.7	84.9	59.4	68.3	65.6	72.8	65.0	70.6	55.9
Beam$_{10}$	59.0	64.5	74.6_{+}	50.6	64.5	55.7	75.2	52.5	52.8	63.5	61.4	47.3	48.6	44.4
Beam$_{50}$	68.1_{+}	72.9_{+}	82.9_{+}	60.9	75.4_{+}	65.1_{+}	84.2	61.9_{+}	61.9	73.4_{+}	71.6	59.2	65.3	55.6
Beam$_{100}$	71.2_{+}	75.7_{+}	85.1_{+}	64.9	78.9_{+}	68.6_{+}	86.3_{+}	65.1_{+}	65.5	76.1_{+}	75.4_{+}	62.9	71.2_{+}	59.6_{+}

	All	CORE	NON-CORE	FUN	MWE
KL-Beam	66.1	70.2	60.1	74.2	45.9
Beam$_{10}$	59.0	58.3	51.2	71.2	36.9
Beam$_{50}$	68.1_{+}	68.8	60.6_{+}	79.4_{+}	47.2_{+}
Beam$_{100}$	71.2_{+}	72.3_{+}	63.9_{+}	81.9_{+}	51.2_{+}

Double marking of possession uses both genitive and ‘a’

Casa	este	a	Anei	.
NOUN_{Def}	AUX	DET_{Poss}	$\text{PROPN}_{\text{Gen}}$	PUNCT
The-house	*is*	–	*Anas’s*	.

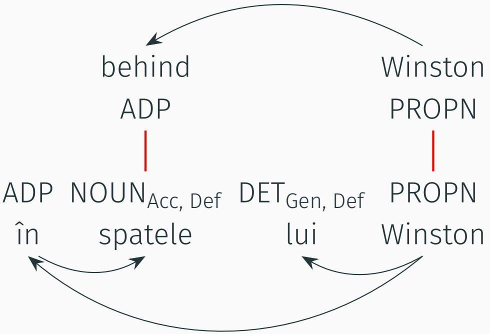
Preposition ‘de’ occurs together with the infinitive marker ‘a’

dorința	irezistibilă	de	a	o	avea
$\text{NOUN}_{\text{Nom, Def}}$	ADJ_{Nom}	ADP	PART	PRON_{Acc}	VERB_{Inf}
the-desire	*irresistible*	*of*	*to*	*it*	*have*

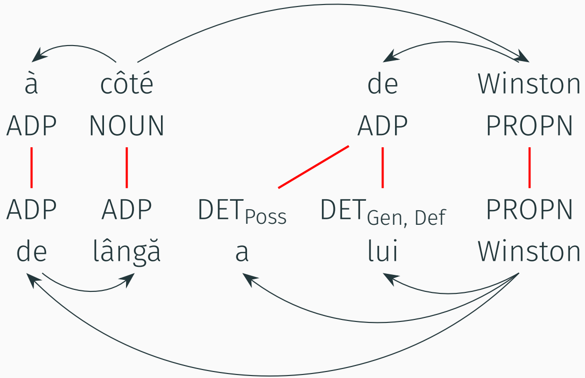
Postnominal demonstrative ‘asta’ is placed mandatorily just after the noun ‘clipa’

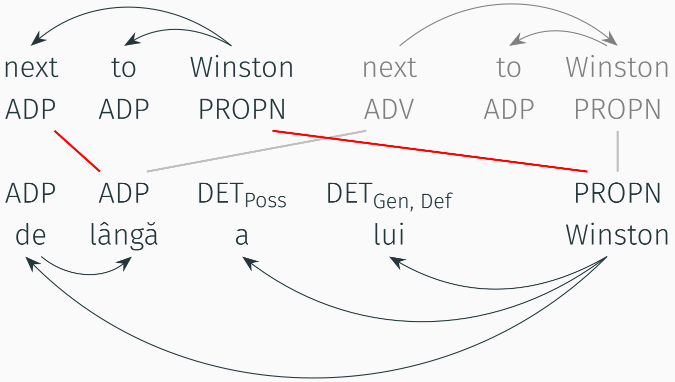
Îmi	scapă	numele	lui	în	clipa	asta
PRON_{Dat}	VERB	$\text{NOUN}_{\text{Nom, Def}}$	DET_{Poss}	ADP	$\text{NOUN}_{\text{Acc, Def}}$	$\text{DET}_{\text{Acc, Dem}}$
To-me	*is-missing*	*the-name*	*his*	*at*	*moment-the*	*this*

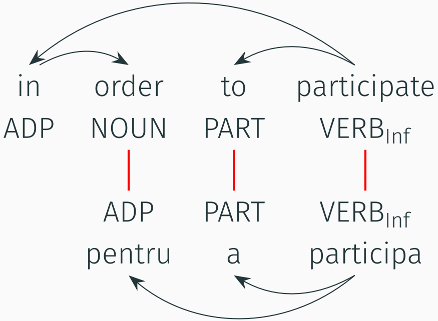
A syntactically inconsistent example of semantics-driven alignment



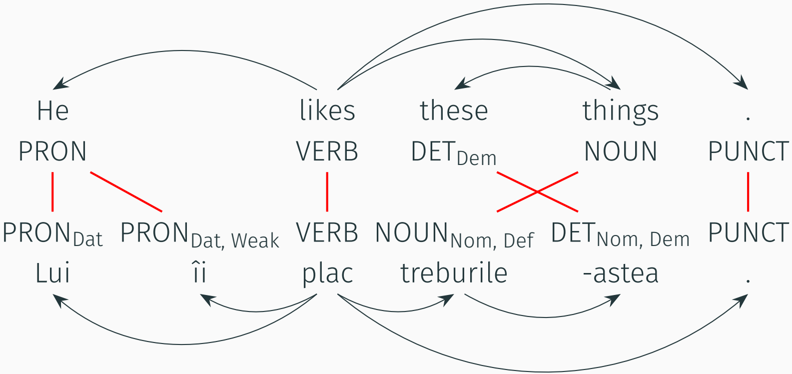
Word sequences are semantically similar, but PoS tags and dependencies differ



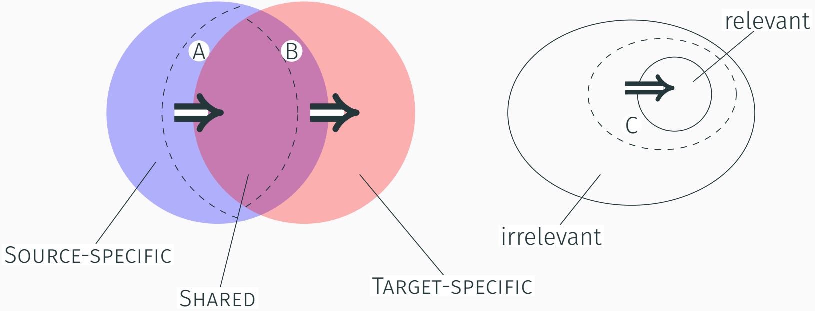

next to Winston
ADP ADP PROPN
next to Winston
ADV ADP PROPN
ADP ADP DET$_{Poss}$ DET$_{Gen, Def}$ PROPN
de lângă a lui Winston


in order to participate
ADP NOUN PART VERB$_{Inf}$
ADP PART VERB$_{Inf}$
pentru a participa

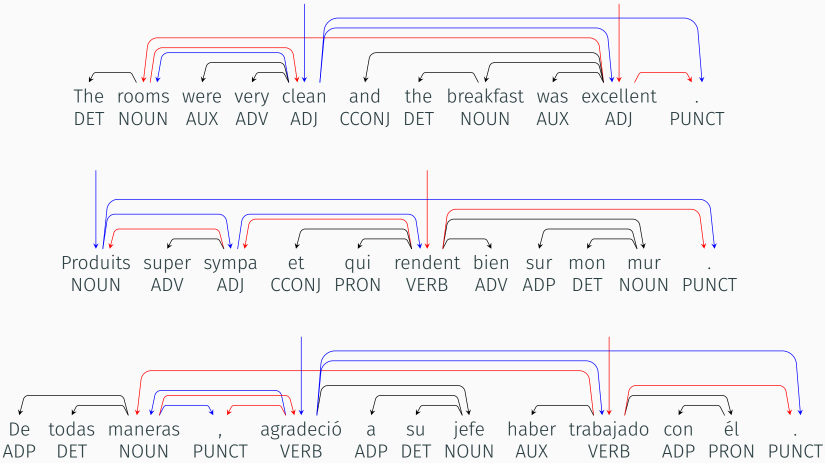
Semantic, PoS and edge correspondence, but diverging relation labels



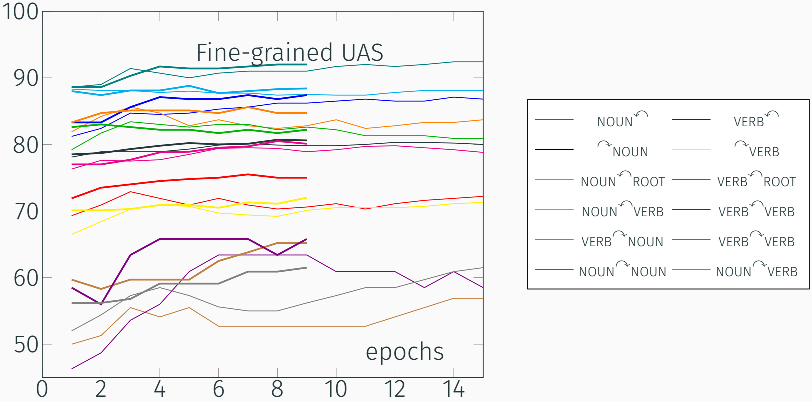
Chapter 6


A
B
Source-specific
Shared
Target-specific
relevant
C
irrelevant

	ρ(root UAS, leaves UAS)			ρ(overall UAS, root UAS)
	10 snt.	500 snt.	Full UD	10 snt.
UDPipe	.134	.519	.709	.249
PanParser	.146	.382	.595	.293
MSTParser	.017	.159	.475	.152
Beam	.360	.577	.716	.477


The rooms were very clean and the breakfast was excellent .
DET NOUN AUX ADV ADJ CCONJ DET NOUN AUX ADJ PUNCT
Produits super sympa et qui rendent bien sur mon mur .
NOUN ADV ADJ CCONJ PRON VERB ADV ADP DET NOUN PUNCT
De todas maneras , agradeció a su jefe haber trabajado con él .
ADP DET NOUN PUNCT VERB ADP DET NOUN AUX VERB ADP PRON PUNCT

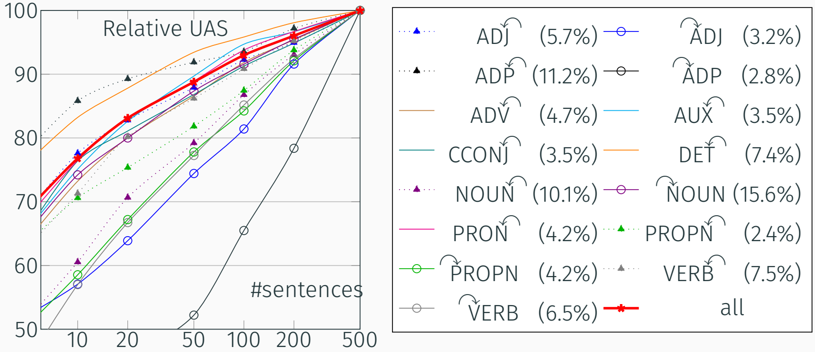
	UAS	Norm		Dist. to Lex		Dist. to Delex	Significant features	
		delex.	lex.	delex.	lex.	delex.	delex.	lex.
Lex	88.31	1,054	3,193	0	0	1,118	5,034	34,148
Delex	85.44	1,517	0	1,118	3,193	0	8,122	0
Delex(Lex)	83.73	1,054	0	0	3,193	1,118	5,034	0
X-Delex	69.68	1,403	0	1,460	3,193	1,729	7,558	0
Delex(X-Lex)	70.10	1,094	0	1,206	3,193	1,557	5,537	0
Delex + Lex	88.50	1,131	3,572	502	1,863	1,129	5,824	50,804
Delex(Lex) + Lex	88.73	1,354	2,490	491	1,824	1,126	8,202	14,640
X-Delex + Lex	88.82	1,545	3,006	1,160	1,753	1,444	9,099	27,511
Delex(X-Lex) + Lex	88.84	1,315	2,898	884	1,752	1,289	7,329	24,178


Fine-grained UAS
epochs
100
90
80
70
60
50
0
2
4
6
8
10
12
14
NOUN
VERB
NOUN
VERB
NOUN ROOT
VERB ROOT
NOUN VERB
VERB VERB
VERB NOUN
VERB VERB
NOUN NOUN
NOUN VERB

	Child PoS						
	ADV	NOUN	PROPN	VERB	SCONJ	Others	
Delex	84.0	73.8	81.1	69.9	86.4	92.8	
Delex(Lex)	79.6	70.2	76.2	66.7	82.6	92.6	
Δ UAS	-5.5	-3.6	-4.9	-3.2	-3.8	-0.2	
	CORE	NON-CORE			MWE	FUN	
	nsubj	acl	advmod	nmod	fixed	mark	Others
Delex	89.0	60.0	85.2	81.9	38.2	92.2	87.7
Delex(Lex)	83.3	51.8	80.6	70.9	31.5	87.4	88.5
Δ UAS	-5.7	-8.2	-4.6	-11.0	-6.7	-4.8	+0.8

	Head PoS			Child PoS							
	NOUN	VERB	Others	DET	ADV	ADP	SCONJ	PRON	NOUN	PROPN	Others
X-Delex	74.5	74.0	55.7	93.5	68.1	81.9	51.5	79.6	60.8	43.4	60.0
Delex(X-Lex)	70.1	79.4	56.2	94.7	71.8	84.2	56.1	86.5	54.1	36.6	60.3
Δ UAS	-4.4	+5.4	+0.5	+1.2	+3.7	+2.3	+4.6	+6.9	-6.7	-6.8	+0.3

	CORE			NON-CORE				MWE	FUN	
	xcomp	nsubj	obj	advmod	advcl	obl	nmod	flat	mark	Others
X-Delex	82.2	69.7	88.4	70.7	45.7	62.5	67.7	28.6	57.6	71.7
Delex(X-Lex)	93.3	76.4	89.9	75.1	51.1	69.8	44.7	16.0	68.8	72.4
Δ UAS	+11.1	+6.7	+5.5	+4.4	+5.4	+7.3	-23.0	-12.6	+11.2	+0.7

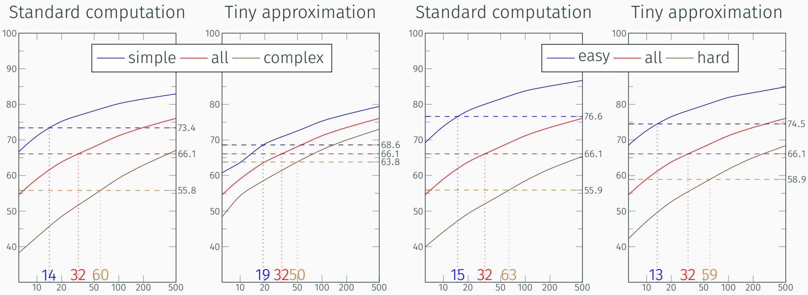


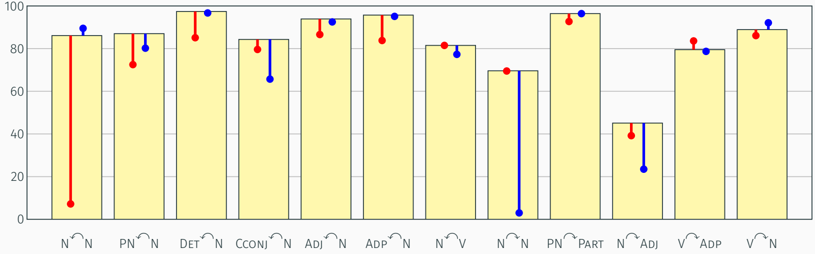
Relative UAS
100
90
80
70
60
50
10
20
50
100
200
500
#sentences
ADJ (5.7%)
ADJ (3.2%)
ADP (11.2%)
ADP (2.8%)
ADV (4.7%)
AUX (3.5%)
CCONJ (3.5%)
DET (7.4%)
NOUN (10.1%)
NOUN (15.6%)
PRON (4.2%)
PROPN (2.4%)
PROPN (4.2%)
VERB (7.5%)
VERB (6.5%)
all

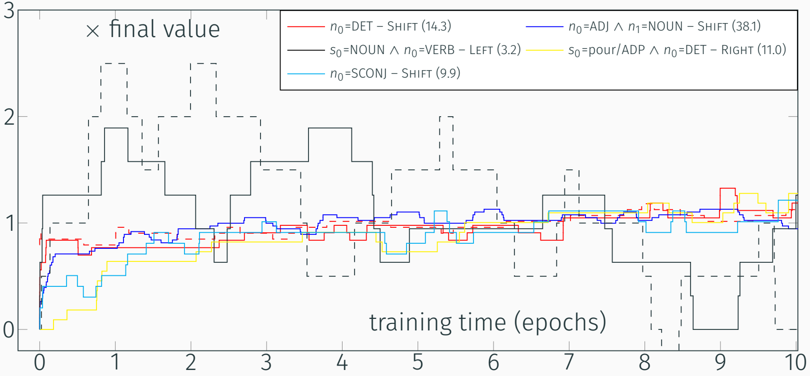
LEARNABILITY	DET 91.3	ADP 89.0	AUX 83.9	PRON 82.4	SCONJ 80.2	ADJ 80.0	CCONJ 77.1	ADV 76.1	
COMPLEXITY	ADP -18.8	DET -18.7	PRON -0.6	AUX 0.2	ADJ 1.9	CCONJ 6.3	N 7.6	ADV 9.6	
HARDNESS	DET -79.6	ADP -72.4	AUX -33.2	PRON -27.2	ADJ -20.1	CCONJ -0.4	ADV 7.5	SCONJ 11.0	

LEARNABILITY	V 75.1	PN 69.0	PN 68.4	N 68.2	N 67.9	ADJ 60.6	V 56.4	AUX 52.8	ADP 48.0
COMPLEXITY	V 12.6	PN 23.4	SCONJ 35.0	N 42.0	PN 49.5	V 52.5	ADJ 57.7	AUX 68.0	ADP 131.2
HARDNESS	V 13.3	N 35.8	PN 45.4	N 59.9	PN 62.6	ADJ 90.8	V 108.7	AUX 110.0	ADP 159.6

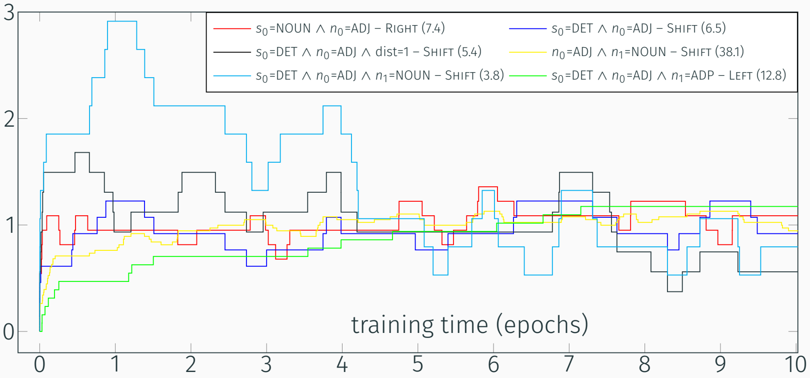
	UAS_{10}		UAS_{500}		$\mathrm{UAS}_{\mathrm{full\ UD}}$	
	simple	complex	simple	complex	simple	complex
UDPipe	56.4	28.0	82.1	66.8	88.0	78.1
PanParser	70.6	40.1	82.2	65.2	86.3	74.3
Delex	69.1	41.8	78.5	62.0	80.8	66.2
MSTParser	68.0	36.9	83.5	66.1	89.1	77.4
Beam	71.1	42.7	82.9	67.1	87.3	76.4
Beam-Delex	70.5	44.2	79.9	64.1	82.6	68.9


Standard computation
Tiny approximation
Standard computation
Tiny approximation
simple
all
complex
easy
all
hard
73.4
66.1
55.8
68.6
66.1
63.8
76.6
66.1
55.9
74.5
66.1
58.9
14
32
60
19
32
50
15
32
63
13
32
59
100
90
80
70
60
50
40
10
20
50
100
200
500

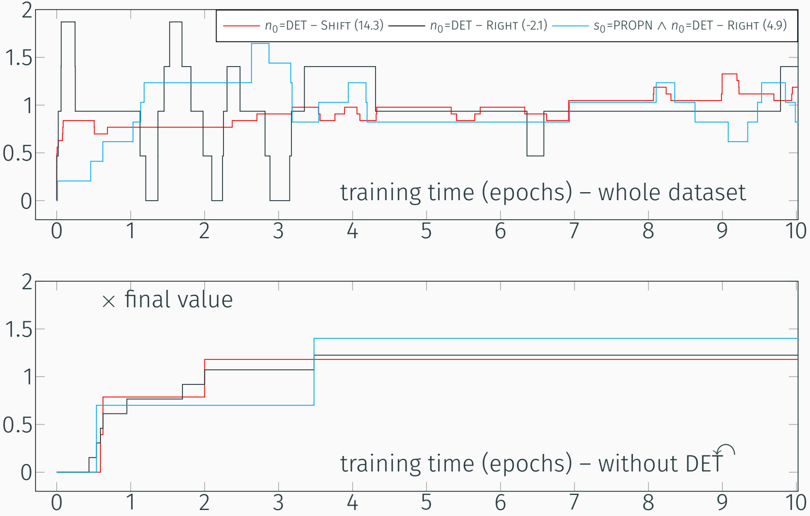

100
80
60
40
20
0
N N
PN N
DET N
CCONJ N
ADJ N
ADP N
N V
N N
PN PART
N ADJ
V ADP
V N



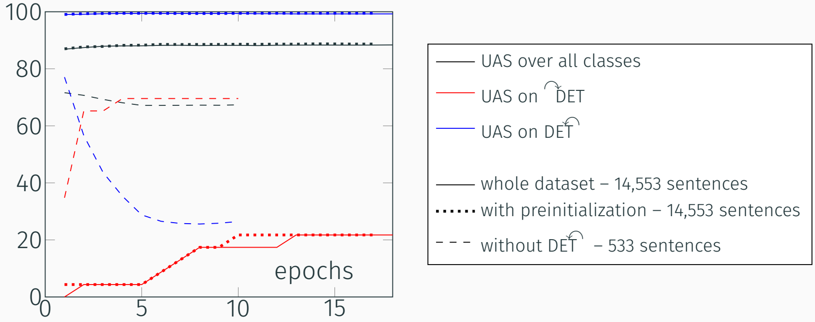
× final value
n_0=DET – SHIFT (14.3)
n_0=ADJ ∧ n_1=NOUN – SHIFT (38.1)
s_0=NOUN ∧ n_0=VERB – LEFT (3.2)
s_0=pour/ADP ∧ n_0=DET – RIGHT (11.0)
n_0=SCONJ – SHIFT (9.9)
training time (epochs)
0
1
2
3
0
1
2
3
4
5
6
7
8
9
10



s_0=NOUN ∧ n_0=ADJ – RIGHT (7.4)
s_0=DET ∧ n_0=ADJ – SHIFT (6.5)
s_0=DET ∧ n_0=ADJ ∧ dist=1 – SHIFT (5.4)
n_0=ADJ ∧ n_1=NOUN – SHIFT (38.1)
s_0=DET ∧ n_0=ADJ ∧ n_1=NOUN – SHIFT (3.8)
s_0=DET ∧ n_0=ADJ ∧ n_1=ADP – LEFT (12.8)
training time (epochs)
0
1
2
3
4
5
6
7
8
9
10



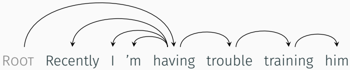
n_0=DET – SHIFT (14.3)
n_0=DET – RIGHT (-2.1)
s_0=PROPN ∧ n_0=DET – RIGHT (4.9)
training time (epochs) – whole dataset
× final value
training time (epochs) – without DET
0 1 2 3 4 5 6 7 8 9 10
0 0.5 1 1.5 2



100
80
60
40
20
0
0
5
10
15
epochs
UAS over all classes
UAS on DET
UAS on DET
whole dataset – 14,553 sentences
with preinitialization – 14,553 sentences
without DET – 533 sentences

	all	ADJ		ADP		ADV		AUX	CCONJ	DET		NOUN		NUM		PRON		PROPN		SCONJ		VERB	
		↶	↷	↶	↷	↶	↷	↶	↶	↶	↷	↶	↷	↶	↷	↶	↷	↶	↷	↶	↷	↶	↷
Size (×1,000)	317.1	5.8	14.4	55.2	2.0	9.1	3.6	12.2	9.0	54.4	0.3	13.9	52.5	4.8	4.6	14.5	1.5	3.9	23.3	1.9	0.8	11.7	16.1
Baseline UAS	88.3	91.1	93.0	96.6	40.3	89.0	81.3	96.7	88.1	99.3	21.7	72.2	80.0	93.5	74.7	96.5	77.0	80.8	86.3	88.9	75.8	86.8	71.4
Freq.-based	88.3	91.7	93.0	96.2	**48.1**	87.8	80.7	97.0	89.3	98.4	30.4	**76.9**	78.4	**95.7**	75.8	96.3	80.3	**86.3**	85.8	**91.9**	72.7	87.7	71.0
Acc.-based	87.5	91.7	90.1	94.1	**61.0**	88.1	82.0	**97.5**	86.9	95.5	**65.2**	73.8	79.3	92.8	75.8	95.5	75.4	**87.7**	85.4	88.9	75.8	84.5	**75.4**
Dyn. acc.-based	88.5	91.7	92.5	96.4	**49.4**	89.6	**83.3**	97.0	88.9	98.7	34.8	**74.0**	79.9	94.2	73.7	96.3	77.0	**89.0**	85.6	88.9	72.7	**86.0**	72.7

Chapter 7



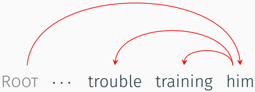
Root trouble | training him

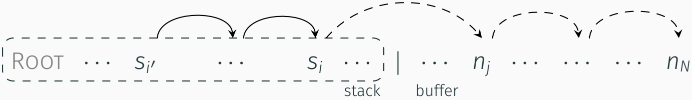
stack buffer

Right

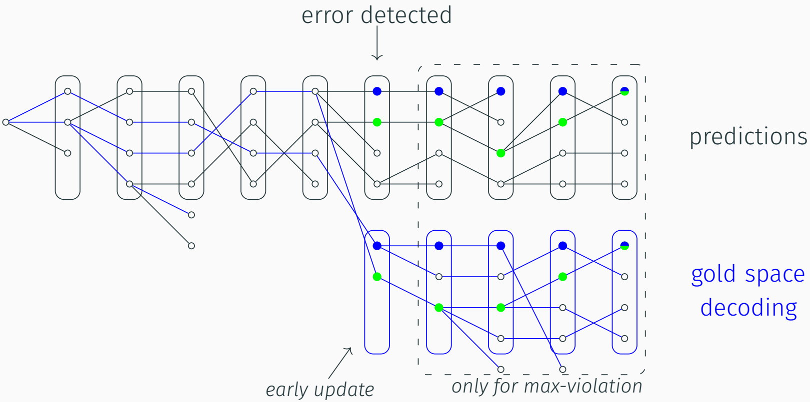
Left

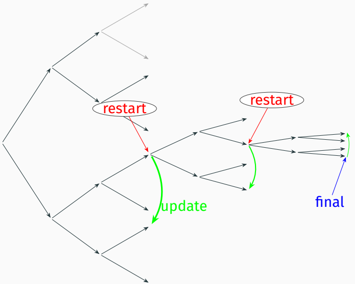
Shift

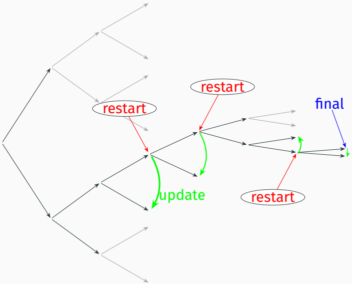



ROOT ⋯ $s_{i'}$ ⋯ s_i ⋯ | ⋯ n_j ⋯ ⋯ ⋯ n_N
stack
buffer

		% non-projective sentences				# training sentences	
	μ	$> 50\%$	25-50%	10-25%	$< 10\%$	> 500	< 500
PanParser – greedy ArcEager	78.28	56.23	76.22	75.48	82.47	81.34	67.36
+ dynamic oracle (only projective snt.)	78.94	57.74	76.98	76.25	82.96	81.92	68.34
+ dynamic oracle + pseudo-proj. snt.	+0.26	+2.01	+1.49	+0.20	-0.07	+0.46	-0.46
+ dynamic oracle + non-projective snt.	+0.48	+2.45	+1.83	+0.45	+0.08	+0.51	+0.36
PanParser – greedy ArcHybrid	75.70	53.08	73.66	73.19	79.63	78.29	66.50
+ dynamic oracle (only projective snt.)	76.50	54.22	74.61	73.95	80.40	79.22	66.81
+ dynamic oracle + non-projective snt.	+0.55	+3.08	+2.16	+0.34	+0.22	+0.53	+0.61
MaltParser (only projective snt.)	72.88	57.87	71.74	69.99	76.68	76.81	58.87
+ pseudo-projectivized sentences	+0.37	+5.84	+1.40	+0.19	+0.07	+0.48	-0.02
+ pseudo-proj. + deprojectivized output	+0.45	+6.84	+1.69	+0.25	+0.09	+0.59	-0.05


error detected
predictions
gold space
decoding
early update
only for max-violation


restart
restart
update
final


restart
restart
final
update
restart

System	Root position	Greedy	Greedy dynamic	Early update	Max-violation
ArcEager	First	77.89	78.97	80.29	80.36
	Last	78.63	79.43	80.35	80.40
ArcHybrid	First	75.72	76.54	79.39	79.78
	Last	76.02	77.05	79.70	79.86
MaltParser	72.88				
MSTParser	79.52				
UDPipe	79.47				

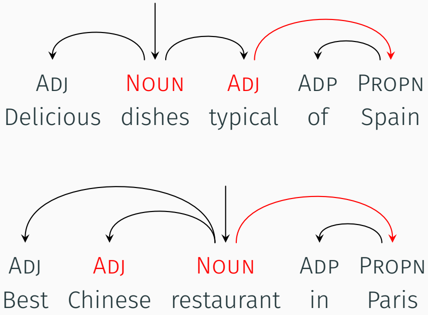
	M11	MX14	RC15		ours		sup.
Target			partial	100%	partial	100%	
de	69.77	74.30	74.32	70.56	73.40	69.36	84.43
es	73.22	75.53	78.17	75.69	77.05	73.98	85.51
fr	74.75	76.53	79.91	77.03	77.44	75.89	85.81
it	76.08	77.74	79.46	77.35	77.74	75.50	86.97
sv	75.87	79.27	82.11	78.68	82.13	77.26	87.89

Criterion	Measure	Std training	Ill-typed	Partial training	Partial parser
Easy on average	%tokens (ref: 27.4%)	28.9%	33.9%	35.6%	27.1%
	precision	86.88	69.99	68.89	85.98
	std precision	86.88	86.43	86.22	88.31
	common (26.7%)	88.61	85.14	87.28	86.81
Length 1	%tokens (ref: 42.9%)	44.4%	61.8%	80.1%	43.5%
	precision	87.42	62.78	50.61	87.06
	std precision	87.42	83.37	80.77	87.68
	common (41.7%)	88.76	87.44	88.03	88.34
Length $\leq$2	%tokens (ref: 63.4%)	65.0%	78.7%	80.9%	64.0%
	precision	85.31	69.89	69.93	85.30
	std precision	85.31	82.01	80.90	85.49
	common (61.6%)	86.54	85.04	85.93	86.46

Constraints	Gold			Standard parser			Partial parser		
Training	Constrained	Const.-pred	Std	Constrained	Const.-pred	Std	Constrained	Const.-pred	Std
Easy on average	76.73	75.82	76.00	74.50	75.04	75.40	72.46	73.52	73.99
Length 1	77.39	74.28	70.46	71.14	70.76	69.99	69.60	69.79	70.09
Length $\leq$2	74.80	71.25	64.87	64.30	64.60	62.94	62.99	63.83	62.76

Chapter 8

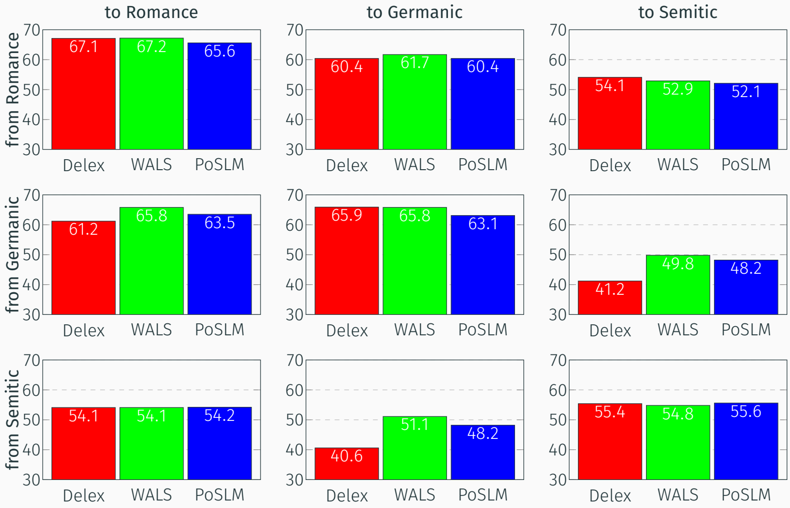
	delicious $\mid$ dishes stack buffer	dishes $\mid$ typical of Spain stack buffer
Conceptual level	Adjectives depend on nouns	
Data level	ADJ NOUN	NOUN ADJ
Classifier level	Feature ($s_0 = \text{ADJ} \wedge n_0 = \text{NOUN}$) has a high weight for LEFT	Feature ($s_0 = \text{NOUN} \wedge n_0 = \text{ADJ}$) has a high weight for RIGHT


ADJ NOUN ADJ ADP PROPN
Delicious dishes typical of Spain
ADJ ADJ NOUN ADP PROPN
Best Chinese restaurant in Paris

Source feature	Target feature	Transformation rule
any	no DEF-DET	remove all definite DETS
any	no IND-DET	remove all indefinite DETS
PR $= 0\%$	PR $\geq 50\%$	switch subtrees to reach PR $= 50\%$ (with 5% error margin)
PR $= 100\%$	PR $\leq 50\%$	switch subtrees to reach PR $= 50\%$ (with 5% error margin)
PR $= 50\%$	PR $= 100\%$	switch subtrees to reach PR $= 75\%$ (with 5% error margin)
PR $= 50\%$	PR $= 0\%$	switch subtrees to reach PR $= 25\%$ (with 5% error margin)

	min	med	max	avg
Delexicalized	23.7	52.0	68.2	49.2
PoSLM selection	23.3	52.0	68.1	-0.1
PoSLM reordering	31.8	53.5	65.6	+2.3
WALS rewrite rules	27.9	55.2	68.3	**+2.9**
Multi-delex		66.9		
Multi-WALS		67.4		

		Target language					
		Romance	Germanic	Slavic	Finno-Ugric	Semitic	Ancient
Source language	Romance	67.1‖65.6‖67.2	60.4‖60.4‖61.7	63.1‖63.5‖63.0	46.4‖50.8‖52.5	54.1‖52.1‖52.9	56.7‖56.5‖54.9
	Germanic	61.2‖63.5‖65.8	65.9‖63.1‖65.8	61.3‖62.2‖63.2	57.2‖58.6‖58.5	41.2‖48.2‖49.8	54.5‖57.1‖56.7
	Slavic	63.5‖61.7‖66.0	63.8‖60.5‖64.3	72.6‖68.4‖71.8	53.2‖57.0‖58.4	54.7‖53.6‖56.8	59.0‖59.2‖60.1
	Finno-Ugric	46.3‖51.9‖52.3	57.1‖56.2‖57.6	53.8‖58.6‖56.9	64.1‖63.0‖64.2	30.0‖43.6‖41.5	50.8‖55.7‖56.1
	Semitic	54.1‖54.2‖54.1	40.6‖48.2‖51.1	42.5‖54.6‖56.1	30.8‖41.2‖44.1	55.4‖55.6‖54.8	53.7‖55.9‖54.4
	Ancient	56.1‖49.2‖55.9	56.7‖51.5‖56.1	60.9‖57.5‖60.6	52.2‖54.9‖56.0	51.1‖47.0‖50.6	62.7‖60.0‖62.6


to Romance
to Germanic
to Semitic
from Romance
from Germanic
from Semitic
70
60
50
40
30
Delex
WALS
PoSLM
67.1
67.2
65.6
60.4
61.7
60.4
54.1
52.9
52.1
61.2
65.8
63.5
65.9
65.8
63.1
41.2
49.8
48.2
54.1
54.1
54.2
40.6
51.1
48.2
55.4
54.8
55.6

Appendix A

Function *StructuredTraining(x,y)*

$c \leftarrow \textsc{Initial}(x)$
$c^{+}, c^{-} \leftarrow \textsc{Oracle}(c, y, \theta)$
$\theta \leftarrow \textsc{Update}(\theta, c^{+}, c^{-})$

Function *StructuredTrainingRestart(x,y)*

$c \leftarrow \textsc{Initial}(x)$
while $\neg\textit{Final}(c)$ **do**
  $c^{+}, c^{-} \leftarrow \textsc{Oracle}(c, y, \theta)$
  $\theta \leftarrow \textsc{Update}(\theta, c^{+}, c^{-})$
  $c \leftarrow c^{-}$

```
Function FindViolation(c_0, y, θ)
    Beam ← {c_0}
    while ∃c ∈ Beam, ¬Final(c) do
        Succ ← ∪_{c∈Beam} Next(c)
        Beam ← k-best(Succ, θ)
        if ∀c ∈ Beam, ¬Correct_y(c|c_0) then
            gold ← {c ∈ Succ | Correct_y(c|c_0)}
            return gold, Beam
    gold ← {c ∈ Beam | Correct_y(c|c_0)}
    return gold, Beam
```

Function *EarlyUpdateOracle*(c_0, y, θ)

$\mathrm{gold}, \mathrm{Beam} \leftarrow \textsc{FindViolation}(c_0, y, \theta)$;

return $top_\theta(\mathrm{gold})$, $top_\theta(\mathrm{Beam})$;

Function *MaxViolationOracle*(c_0, y, θ)

$\mathrm{gold}, \mathrm{Beam} \leftarrow \textsc{FindViolation}(c_0, y, \theta)$;

$\mathrm{candidates} \leftarrow \{(top_\theta(\mathrm{gold}), top_\theta(\mathrm{Beam}))\}$;

while $\exists c \in Beam, \neg\textsc{Final}(c)$ **do**

$\mathrm{Succ} \leftarrow \cup_{c \in \mathrm{Beam}} \textsc{Next}(c)$;

$\mathrm{Beam} \leftarrow k$-best($\mathrm{Succ}$, θ);

$\mathrm{Succ}^+ \leftarrow \cup_{c \in \mathrm{gold}} \{c' \in \textsc{Next}(c) | \textsc{Correct}_y(c'|c_0)\}$;

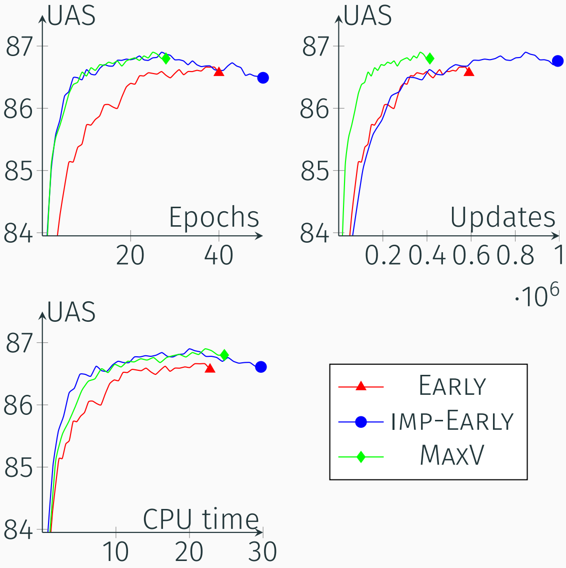
$\mathrm{gold} \leftarrow k$-best($\mathrm{Succ}^+$, θ);

$\mathrm{candidates} \leftarrow \mathrm{candidates} + (top_\theta(\mathrm{gold}), top_\theta(\mathrm{Beam}))$;

return $\mathrm{argmax}_{c^+, c^- \in \mathrm{candidates}}(score_\theta\,(c^-) - score_\theta\,(c^+))$;

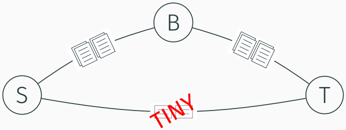
	ar	de	eu	fr	he	hu	ko	pl	sv	μ
Greedy dyn	83.98	90.73	84.00	84.23	83.78	84.33	82.79	87.66	86.35	85.32
Early	85.03	92.74	84.42	86.02	85.39	85.63	82.73	89.60	87.00	86.51
imp-Early	**85.27**	92.89	84.59	86.26	**85.84**	**85.74**	**82.98**	89.55	**87.37**	86.72
MaxV	85.06	92.77	84.59	86.10	85.53	85.57	82.68	89.42	87.16	86.54
imp-MaxV	85.04	**92.90**	**84.68**	**86.26**	85.83	85.55	82.94	**90.12**	87.31	**86.74**

KL div	Baseline	Improved
EARLY	0.350	0.280
MAXV	0.357	0.277

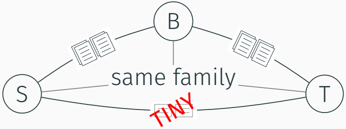

UAS
87
86
85
84
Epochs
20
40
UAS
87
86
85
84
Updates
0.2 0.4 0.6 0.8 1
·10^6
UAS
87
86
85
84
CPU time
10
20
30
EARLY
IMP-EARLY
MAXV

Appendix B

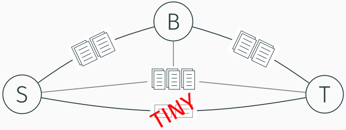
BRIDGE



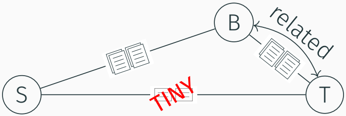
RELATED



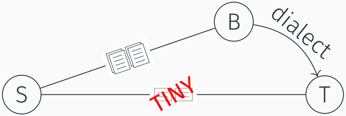
MULTIPARALLEL



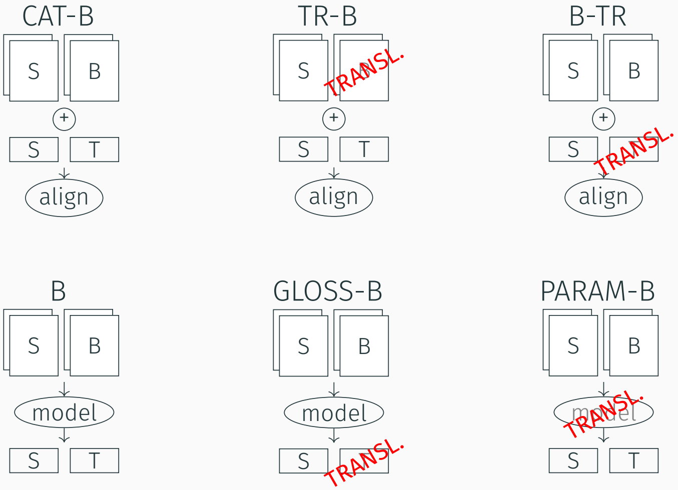
DIRECTED BRIDGE



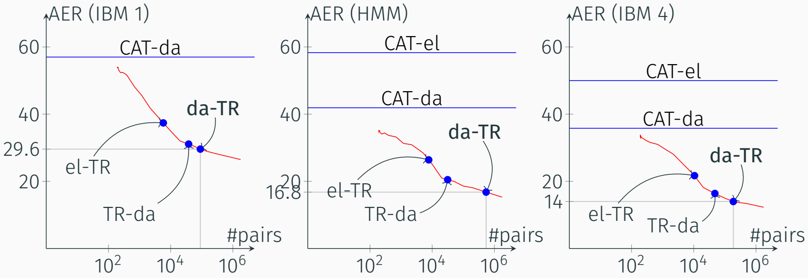
DIALECT



Data space	
CAT-B	concatenate S-B and test data; train
TR-B	word-for-word translate S-B data; concatenate with test data; train
B-TR	word-for-word translate test data in B; concatenate with S-B data; train
Parameter space	
B	train an S-B model; apply on test data
GLOSS-B	train an S-B model; apply on test data word-for-word translated in B
PARAM-B	train an S-B model; translate the parameters; apply on test data


CAT-B
S
B
+
S
T
align
TR-B
S
B
TRANSL.
+
S
T
align
B-TR
S
B
+
S
T
TRANSL.
align
B
S
B
model
S
T
GLOSS-B
S
B
model
S
T
TRANSL.
PARAM-B
S
B
model
TRANSL.
S
T

		Swedish only		Danish data			Greek data			Danish parameters		
		baseline	CAT-sv	CAT-da	TR-da	da-TR	CAT-el	TR-el	el-TR	da	GLOSS-da	PARAM-da
A	IBM 1	53.9	**26.5**	57.0	31.1	**29.6**	74.3	**35.9**	37.4	66.0	**28.3**	33.3
E	HMM	35.3	**15.3**	41.9	20.5	**16.8**	58.3	26.9	**26.4**	46.7	**16.4**	25.8
R	IBM 4	33.9	**12.3**	35.8	16.4	**14.0**	50.0	**20.6**	21.7	49.1	**14.8**	24.3
P	IBM 1	68.7	**73.3**	58.7	73.8	**74.0**	47.4	**71.9**	71.5	67.0	**72.2**	71.1
O	HMM	69.9	**73.8**	71.9	73.5	**73.6**	66.6	**73.4**	71.9	69.5	**73.4**	72.4
S	IBM 4	73.0	**74.7**	74.0	73.9	**74.9**	72.0	73.4	**73.5**	66.7	**73.6**	72.0


AER (IBM 1)
60
CAT-da
40
29.6
20
da-TR
el-TR
TR-da
#pairs
10^2
10^4
10^6
AER (HMM)
60
CAT-el
CAT-da
40
20
16.8
da-TR
el-TR
TR-da
#pairs
10^2
10^4
10^6
AER (IBM 4)
60
CAT-el
40
CAT-da
20
14
da-TR
el-TR
TR-da
#pairs
10^2
10^4
10^6

Agić v., Hovy D. & Søgaard A. (2015). If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, p. 268–272, Beijing, China: Association for Computational Linguistics.

Agić v., Johannsen A., Plank B., Martínez Alonso H., Schluter N. & Søgaard A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, **4**, 301–312.

Banea C., Mihalcea R., Wiebe J. & Hassan S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 127–135, Honolulu, Hawaii: Association for Computational Linguistics.

Collins M. & Roark B. (2004). Incremental parsing with the perceptron algorithm. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, p. 111–118, Barcelona, Spain.

Das D. & Petrov S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, p. 600–609, Portland, Oregon, USA: Association for Computational Linguistics.

Duong L., Cohn T., Bird S. & Cook P. (2015). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, p. 845–850, Beijing, China: Association for Computational Linguistics.

Ghoshal A., Swietojanski P. & Renals S. (2013). Multilingual training of deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7319–7323: IEEE.

Goldberg Y. & Nivre J. (2012). A dynamic oracle for arc-eager dependency parsing. In *Proceedings of COLING 2012*, p. 959–976, Mumbai, India: The COLING 2012 Organizing Committee.

Huang L., Fayong S. & Guo Y. (2012). Structured perceptron with inexact search. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 142–151, Montréal, Canada: Association for Computational Linguistics.

Hwa R., Resnik P., Weinberg A. & Kolak O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, p. 392–399, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Klementiev A., Titov I. & Bhattarai B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, p. 1459–1474, Mumbai, India: The COLING 2012 Organizing Committee.

Klinger R. & Cimiano P. (2015). Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, p. 153–163, Beijing, China: Association for Computational Linguistics.

Kozhevnikov M. & Titov I. (2014). Cross-lingual model transfer using feature representation projection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 579–585, Baltimore, Maryland: Association for Computational Linguistics.

Lu B., Tan C., Cardie C. & K. Tsou B. (2011). Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, p. 320–330, Portland, Oregon, USA: Association for Computational Linguistics.

Ma X. & Xia F. (2014). Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1337–1348, Baltimore, Maryland: Association for Computational Linguistics.

Martins A. F. T. (2015). Transferring coreference resolvers with posterior regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 1427–1437, Beijing, China: Association for Computational Linguistics.

McDonald R., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K., Petrov S., Zhang H., Täckström O., Bedini C., Bertomeu Castelló N. & Lee J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 92–97, Sofia, Bulgaria: Association for Computational Linguistics.

McDonald R., Petrov S. & Hall K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 62–72, Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Naseem T., Barzilay R. & Globerson A. (2012). Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 629–637, Jeju Island, Korea: Association for Computational Linguistics.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R. & Zeman D. (2016). Universal dependencies v1: A multilingual treebank collection. In N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis, Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France: European Language Resources Association (ELRA).

Petrov S., Das D. & McDonald R. (2012). A universal part-of-speech tagset. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA).

Rasooli M. S. & Collins M. (2015). Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 328–338, Lisbon, Portugal: Association for Computational Linguistics.

Rigutini L., Maggini M. & Liu B. (2005). An em based training algorithm for cross-language text categorization. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, p. 529–535: IEEE.

Rosa R. & Zabokrtsky Z. (2015). Klcpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, p. 243–249, Beijing, China: Association for Computational Linguistics.

Täckström O., Das D., Petrov S., McDonald R. & Nivre J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, **1**, 1–12.

Täckström O., McDonald R. & Uszkoreit J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 477–487, Montréal, Canada: Association for Computational Linguistics.

Tiedemann J., Agić v. & Nivre J. (2014). Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, p. 130–140, Ann Arbor, Michigan: Association for Computational Linguistics.

References v

WAN X. (2009). Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, p. 235–243, Suntec, Singapore: Association for Computational Linguistics.

WANG M. & MANNING C. D. (2014). Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association of Computational Linguistics*, **2**(1), 55–66.

WEI B. & PAL C. (2010). Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the ACL 2010 Conference Short Papers*, p. 258–262, Uppsala, Sweden: Association for Computational Linguistics.

YAROWSKY D., NGAI G. & WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, p. 1–8: Association for Computational Linguistics.

YU Z., MAREČEK D., ŽABOKRTSKÝ Z. & ZEMAN D. (2016). If you even don't have a bit of bible: Learning delexicalized pos taggers. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France: European Language Resources Association (ELRA).

ZEMAN D. & RESNIK P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, p. 35–42.