Machine Translation and Discriminative Models

Tree-to-tree transfer and Discriminative learning

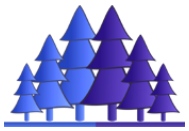
Martin Popel

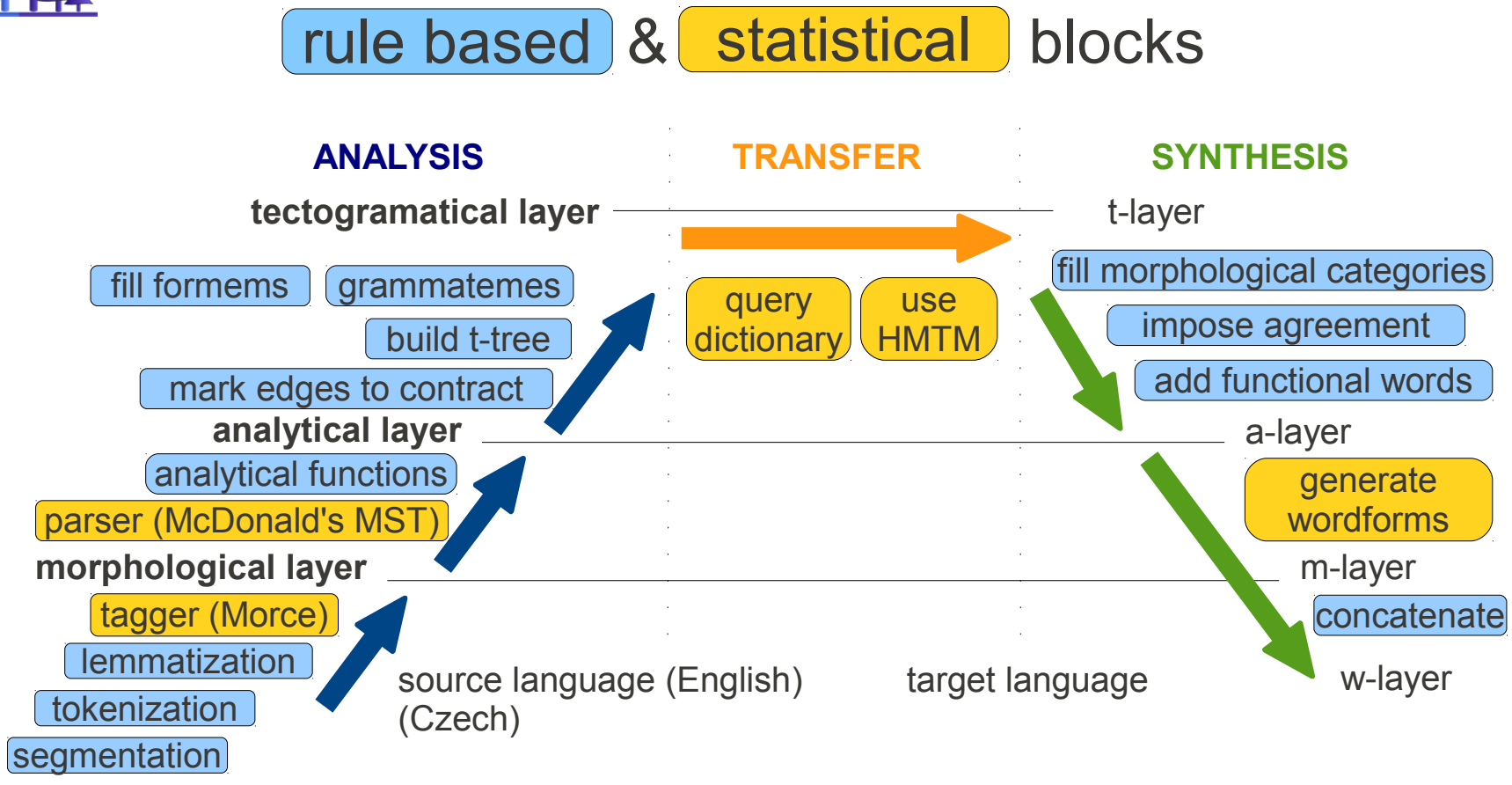
ÚFAL (Institute of Formal and Applied Linguistics)
Charles University in Prague

March 23rd 2015, Seminar of Formal Linguistics, Prague

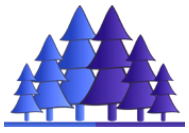
Outline

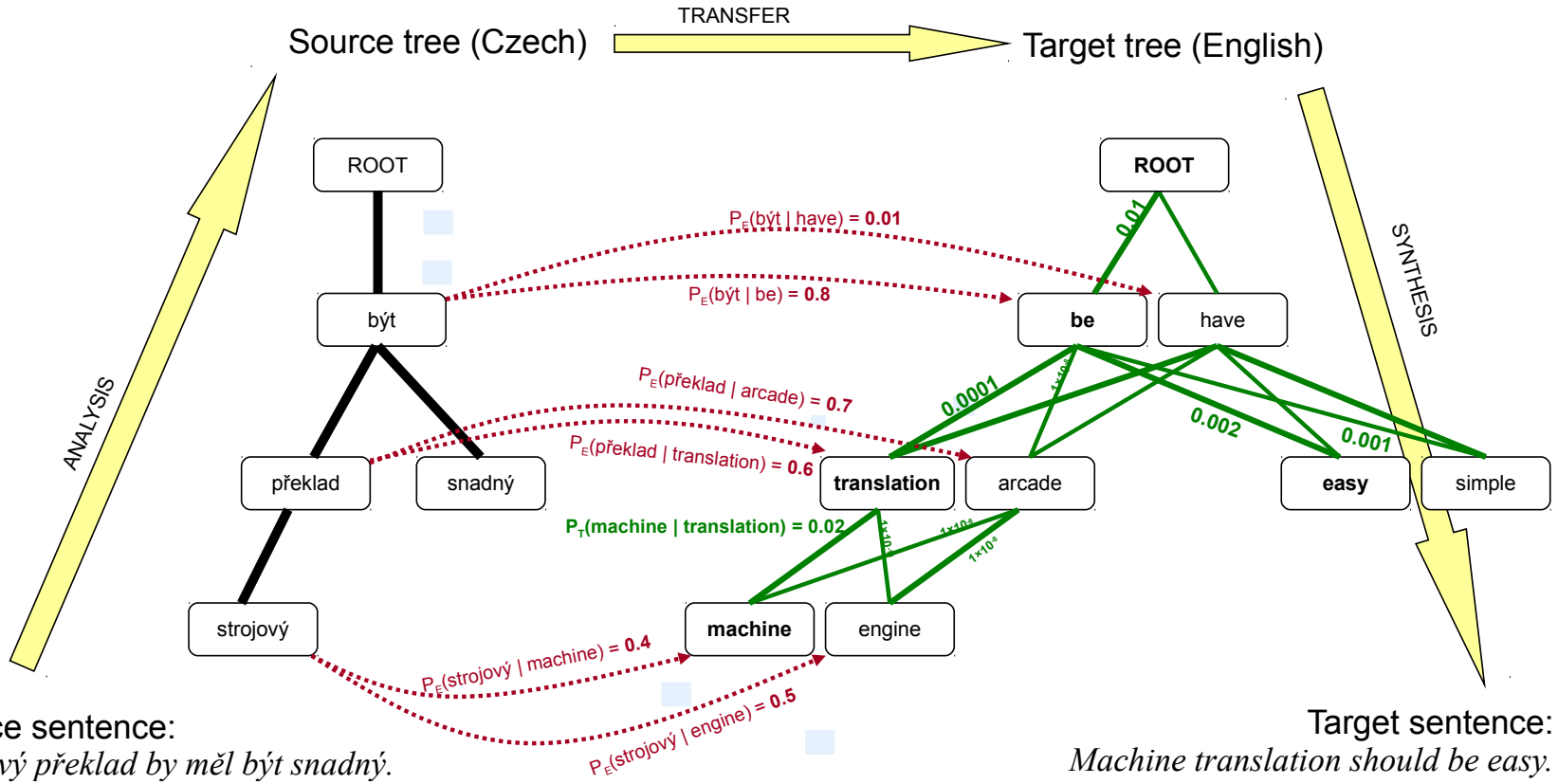
TectoMT: analysis, transfer, synthesis





TectoMT: isomorphic transfer (1-1 node mapping)





Source sentence:
Strojový překlad by měl být snadný.

Target sentence:
Machine translation should be easy.

Phrase-based Statistical Machine Translation

- currently most popular approach (Moses toolkit)
- no linguistic analysis needed (just tokenization)
- translates each phrase independently (except for LM)
- many segmentations to phrases considered, only one used

Quiz: English-Czech translation

Is it possible to translate *prime* as *vláda*?

- In which context?

Quiz: English-Czech translation

Is it possible to translate *prime* as *vláda*?

- In which context?
  prime minister
  předseda vlády

Quiz: English-Czech translation

Is it possible to translate *prime* as *vláda*?

- In which context?
  prime minister
  předseda vlády
- How to formalize such translation rule?

lemma=*prime* formeme=*adj:attr* parent_lemma=*minister*	⇒	lemma=*vláda* formeme=*n:2* parent_lemma=*předseda*

Quiz: English-Czech translation

Is it possible to translate *prime* as *vláda*?

- In which context?
  prime minister
  předseda vlády
- How to formalize such translation rule?

lemma=*prime* formeme=*adj:attr* parent_lemma=*minister*	⇒	lemma=*vláda* formeme=*n:2* parent_lemma=*předseda*

This is still isomorphic transfer, unlike *prime minister* ⇒ *premiér*.

Quiz: English-Czech translation

Is it possible to translate *find* as *přijít*?

- In which context?

Quiz: English-Czech translation

Is it possible to translate *find* as *přijít*?

- In which context? (on t-layer, *find*≠*find_out*)

Quiz: English-Czech translation

Is it possible to translate *find* as *přijít*?

- In which context? (on t-layer, *find*≠*find_out*)

Agatha found that book interesting.

Agátě přišla ta kniha zajímavá.

Quiz: English-Czech translation

Is it possible to translate *find* as *přijít*?

- In which context? (on t-layer, *find*≠*find_out*)
  Agatha found that book interesting.
  Agátě přišla ta kniha zajímavá.
- How to formalize such translation rule?

Quiz: English-Czech translation

Is it possible to translate *find* as *přijít*?

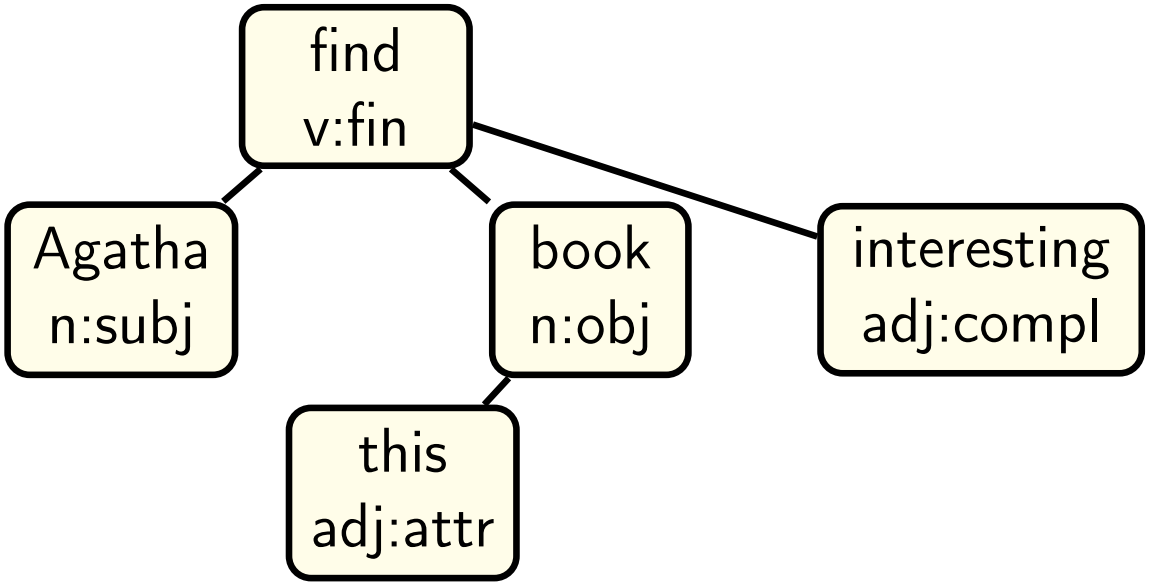
- In which context? (on t-layer, *find*≠*find_out*)
  Agatha[n:subj] found that book[n:obj] interesting[adj:compl].
  Agátě[n:3] přišla ta kniha[n:1] zajímavá[adj:1].
- How to formalize such translation rule?

child1_formeme=*n:subj* lemma=*find* child2_formeme=*n:obj* child3_formeme=*adj:compl*	⇒	child1_formeme=*n:3* lemma=*přijít* child2_formeme=*n:1* child3_formeme=*adj:1*	

(adj:compl means predicative adjective)

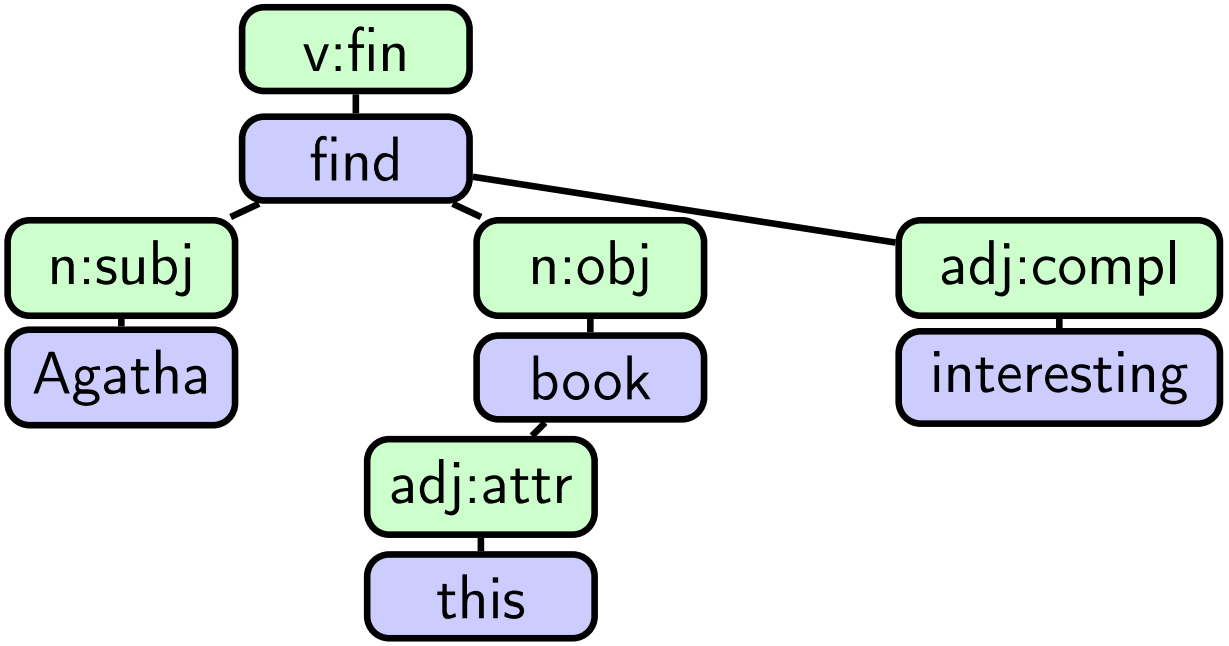
Representation of t-layer

lemma and formeme as two attributes



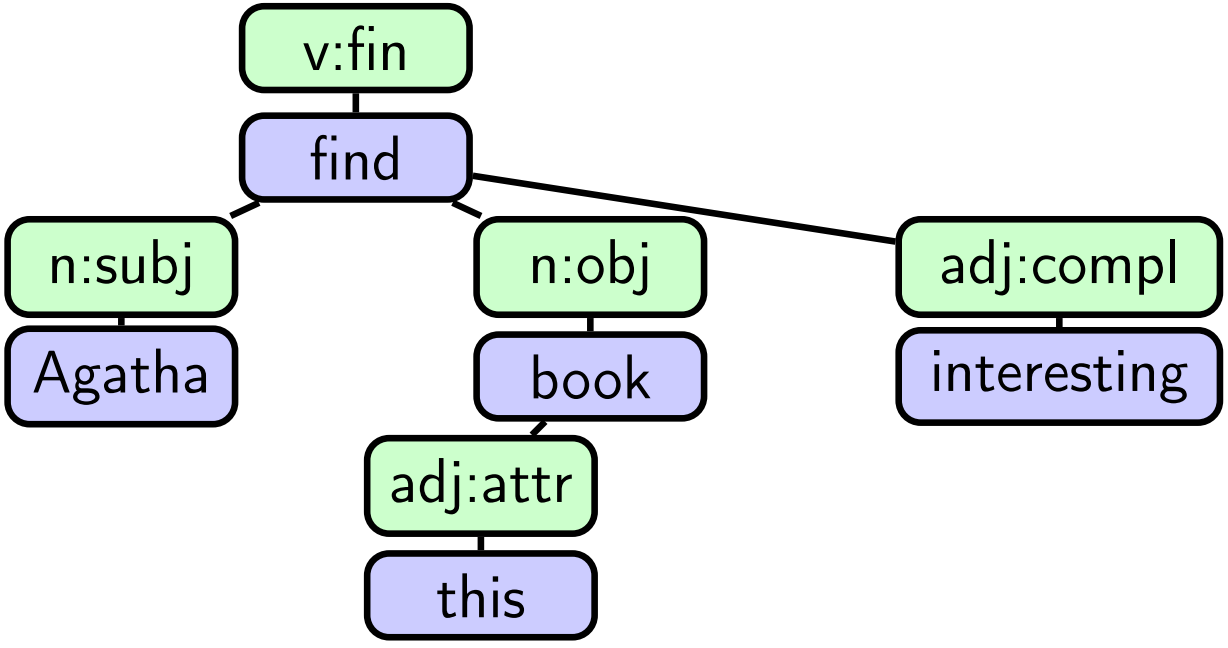
Representation of t-layer

lemma and formeme as interleaved "sub-nodes"



Representation of t-layer

lemma and formeme as interleaved "sub-nodes"



grammatemes:

- translated in postprocessing (current approach)
- as subnodes (leaves, children of lemmas)
- encoded within lemma, but only if grammateme changed

Handling non-isomorphic transfer

- preprocessing or postprocessing within transfer (current approach)
- natively in the main transfer algorithm
- convert training data to isomorphic trees [not tried yet]
  - n-1 alignment: add special [delete_node] label to the target side
  - 1-n alignment: encode added nodes (L+F) into the "main" lemma
  - encode topology change: as_child, as_sibling, as_parent

TectoMT transfer over years

year	*BLEUdiff*	*method*
2008		initial baseline
2009	+1.5	HMTM (TreeViterbi, TreeLM)
2010	+0.8	HMTM + MaxEnt
2012	-2.2	TectoMoses
2012	NA	Gibbs sampling treelets
2013	-3.0	Easy-first treelets
2013	-2.0	Interpol treelets
2014	+0.1	VowpalWabbit

TectoMT transfer over years

year	*BLEUdiff*	*method*
2008		initial baseline
2009	+1.5	HMTM (TreeViterbi, TreeLM)
2010	+0.8	HMTM + MaxEnt
2012	-2.2	TectoMoses
2012	NA	Gibbs sampling treelets
2013	-3.0	Easy-first treelets
2013	-2.0	Interpol treelets
2014	+0.1	VowpalWabbit
	+0.9	other improvements in 2010–2014

TectoMT transfer over years

year	*BLEUdiff*	*method*
2008		initial baseline
2009	+1.5	HMTM (TreeViterbi, TreeLM)
2010	+0.8	HMTM + MaxEnt
2012	-2.2	TectoMoses
2012	NA	Gibbs sampling treelets
2013	-3.0	Easy-first treelets
2013	-2.0	Interpol treelets
2014	+0.1	VowpalWabbit
	+0.9	other improvements in 2010–2014
2015	+8.6	QTLeap en→cs in two months

2008: baseline TectoMT transfer

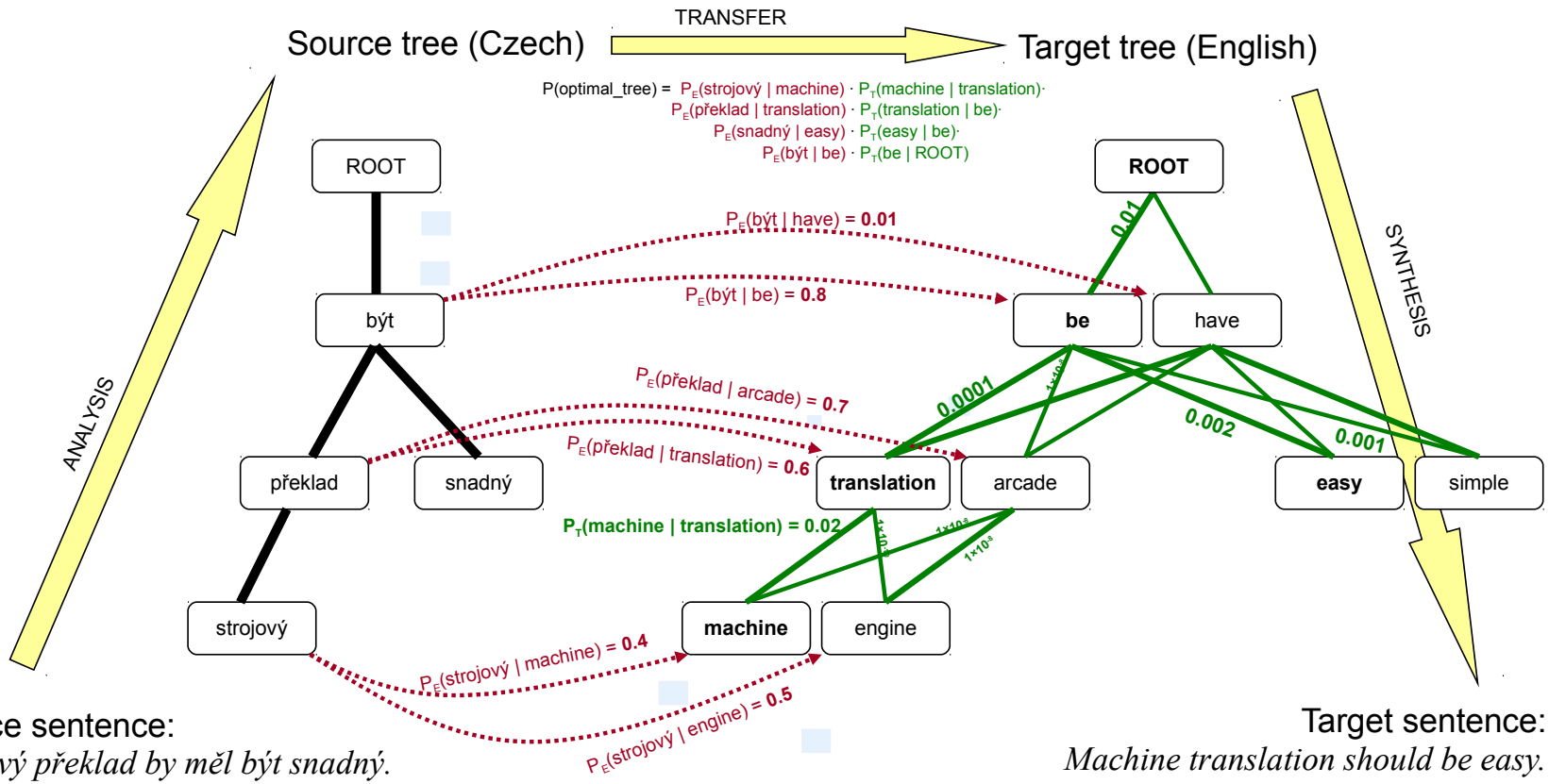
- "static" translation model $P(target|source) = \frac{\#(source,target)}{\#(source)}$
- first translate formemes, then lemmas
- use **only the top variant**

WMT 2009 en→cs results

	BLEU	human score
Moses (CUNI)	14.2	61
Google	13.6	66
Moses (UEdin)	13.5	53
Eurotran XP	9.5	67
PC Translator	9.4	67
TectoMT	7.3	48

2009: Hidden Markov Tree Model (HMTM)

- still using "static" translation models, but also
- TreeLM (target lemma-formeme and parent-child compatibility)
- best labeling is found via HMTM (Tree Viterbi)

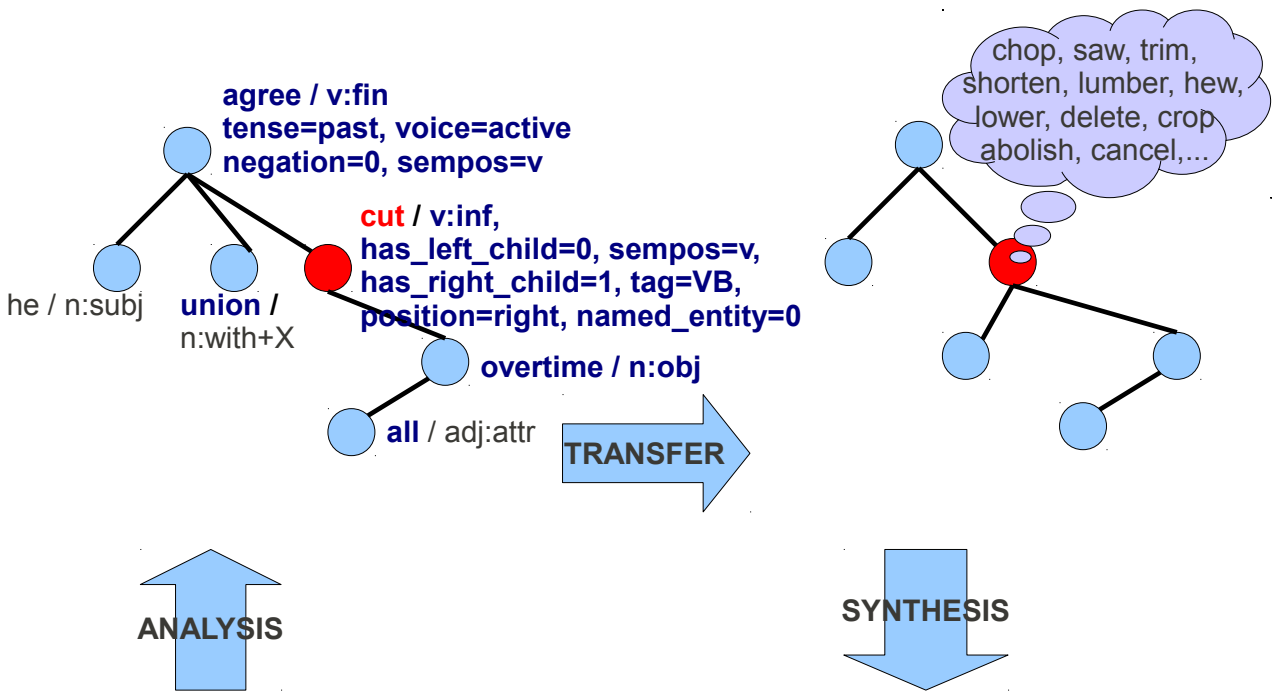


P_E(source | target) … emission probabilities … **translation model**

P_T(dependent | governing) … transition probabilities … **target-language tree model**

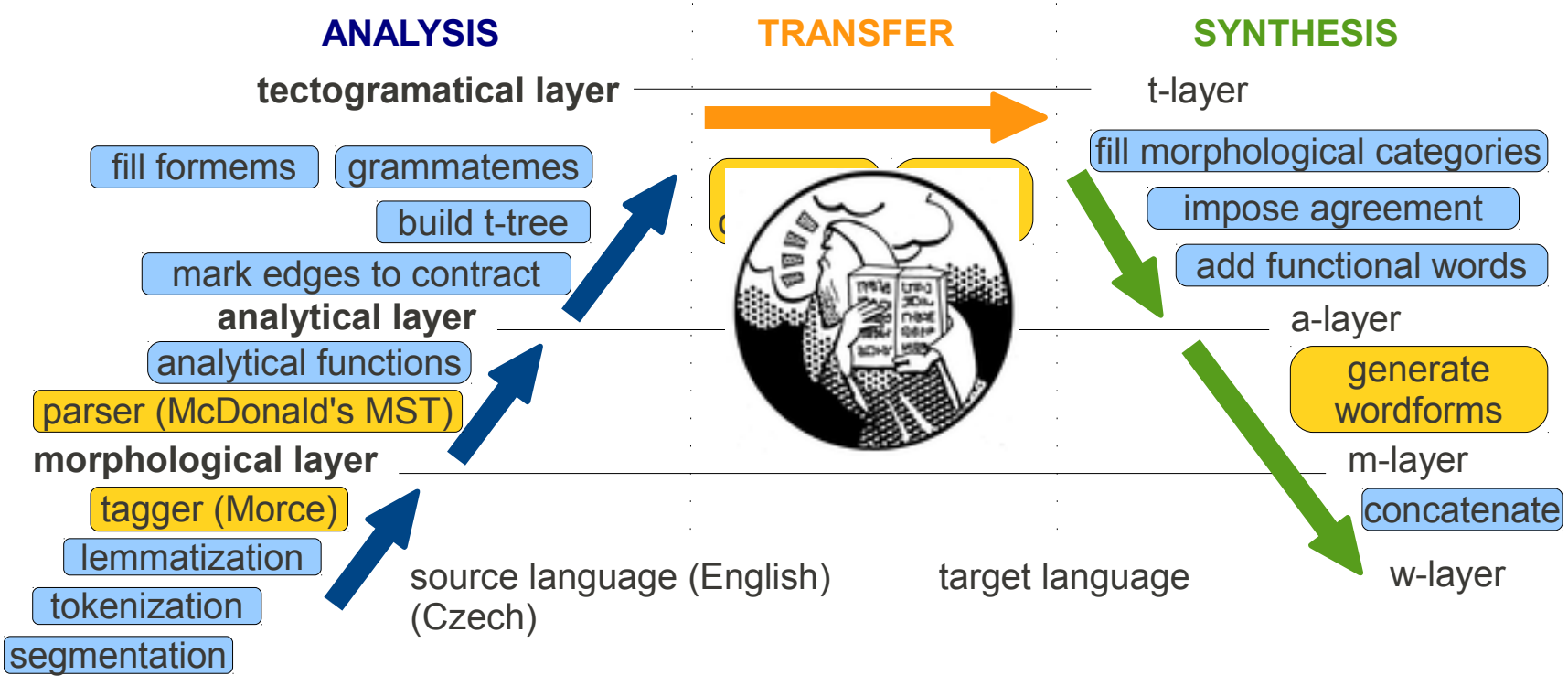
2010: Maximum Entropy translation model

- still using HMTM (and generative TreeLM),
- but the "static" model P(lemma | src_lemma) interpolated with
- context-sensitive discriminative (MaxEnt) model P(lemma | src_lemma, **other features**)



2012: TectoMoses

- substitute transfer (MaxEnt+HMTM) in TectoMT with
- **phrase-based decoding** (Moses) of linearized t-trees
- 2 factors: lemma and formeme, but joint (L+F) **n-gram LM** better
- project dependencies and use TectoMT synthesis



2012–2013: Gibbs sampling and CRP segmentations

- example of treelet-based transfer, P(trg treelet | src treelet)
- Bayesian approach
- use Chinese Restaurant Process on parallel treebank to learn
  - optimal segmentation to treelet pairs
  - optimal translations
- use Gibbs sampling both in training and decoding

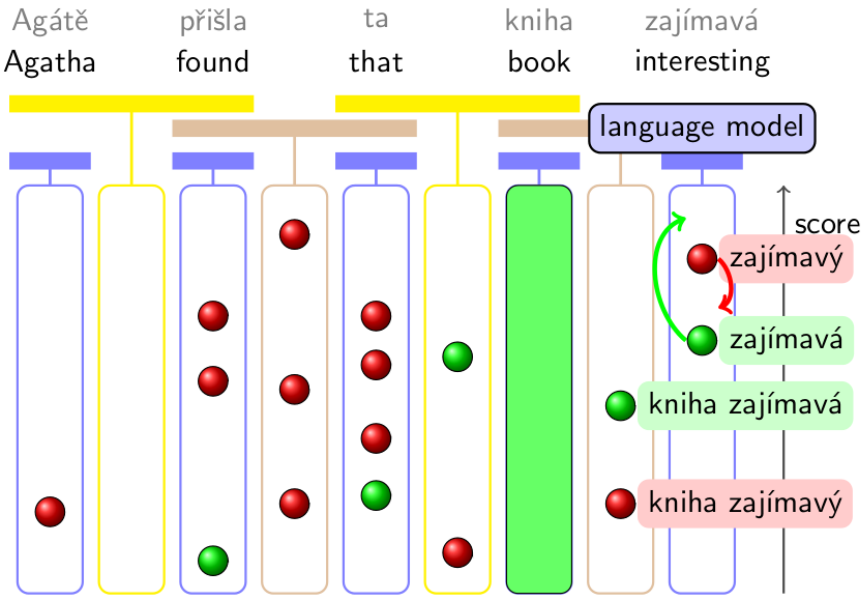
Problems

- Pitman–Yor instead of Chinese Restaurant (heavy tail)
- Slice sampling or annealing instead of Gibbs (local maxima)
- We want *cut+grass=sekat+tráva* and *cut+taxes=snížit+daně*, but CRP/PYP prefers more reusable segments *grass=tráva*, *taxes=daně* (and *cut=sekat*, *cut=snížit*).
  CRP/PYP knows nothing about translation
  (cf. Chung et al. 2013: Sampling tree fragments from forests).

2013: Easy-first treelet-based transfer

- easy-first decoding of overlapping treelets
- guided learning of both
  - how to translate each treelet
  - in which order
- Handles non-isomorphic transfer natively (not tried).
- How to integrate TreeLM (for partial translations)?

Easy-First Decoding (Phrase-Based MT)



2013: Interpolation treelet-based transfer (Interpol)

- similar to easy-first decoding
  - one feature (weight): source treelet size
  - many other features (lexicalized, PoS tags,...)
  - combinations and quantizations of features
- all matching "rules" applied, their scores interpolated
- does not handle non-isomorphism
- no guided learning

Example

score(*zajímavá*) = P_{MaxEnt}(zajímavá | interesting)
+ w_L s(interesting⇒zajímavá)
+ w_{Lf} s(interesting adj:compl⇒zajímavá adj:1)
+ w_{Lf} s(interesting adj:compl⇒zajímavá adj:attr)
+ w_{Lfl} s(interesting adj:compl find⇒zajímavá adj:1 přijít)
+ w_{Lfl} s(interesting adj:compl find⇒zajímavá adj:attr najít)
+ w_{Lflf} s(interesting adj:compl find v:fin⇒zajímavá adj:1 přijít v:fin)

How is Interpol (2013) different from MaxEnt (2010)?

How is Interpol (2013) different from MaxEnt (2010)?

The Logic, or There and Back Again

How is Interpol (2013) different from MaxEnt (2010)?

The Logic, or There and Back Again

- Interpol can be trained with MaxEnt (multinomial logit)
- HMTM can be applied on top of Interpol
- Interpol is equivalent to the 2010 approach with **additional features**

"Summary" features

- feature value precomputed on whole training data, relative frequency, or rather (#(src,trg) - 1)/#(src)
- dense feature (unlike sparse lexicalized higher-order features)
- Moses also uses "summary" features (TM-forward, TM-backward, LM)

2014: VowpalWabbit-based transfer

- VW is an ultra-fast and modular machine learning toolkit
- optimized SGD (AdaGrad, dense+sparse features,...)
- cost-sensitive one-against-all reduction to binary classification
- logistic loss enables probabilistic interpretation (for HMTM)
- all lemmas in one model, fixed memory requirements
- label-dependent features (features shared for more lemmas)

Thank you