

Deep Syntactic Machine Translation with Hidden Markov Tree Models

Martin Popel

ÚFAL, Charles University in Prague



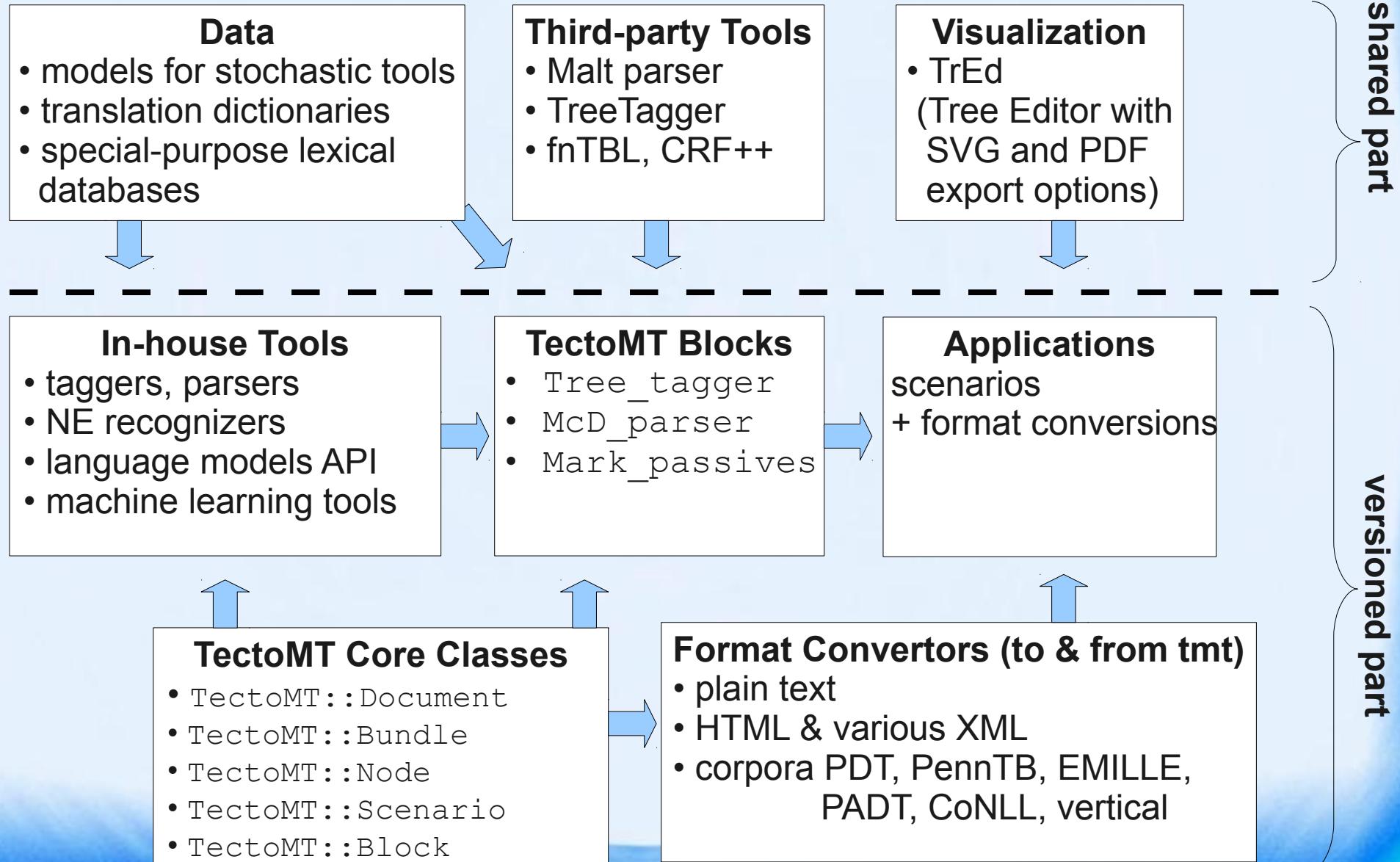
5th PIRE Meeting, Dec. 12 2009

Outline

- TectoMT – NLP framework and MT system
- Demo translation step by step
- Annotation of translation errors
- Parsing sentences with parentheses
- Hidden Markov Tree Models (HMTM)
- Conclusion

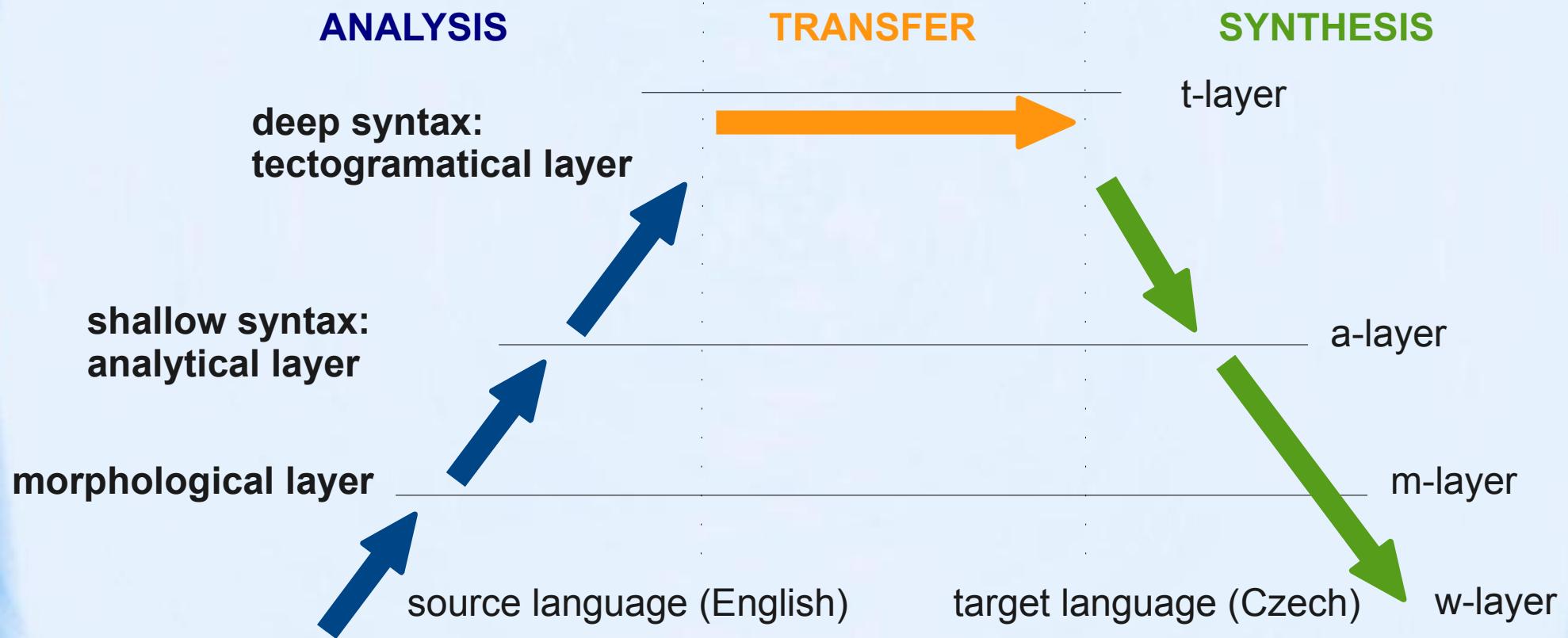
TectoMT – framework for NLP

modular, open source, Perl, Linux, OOP-style



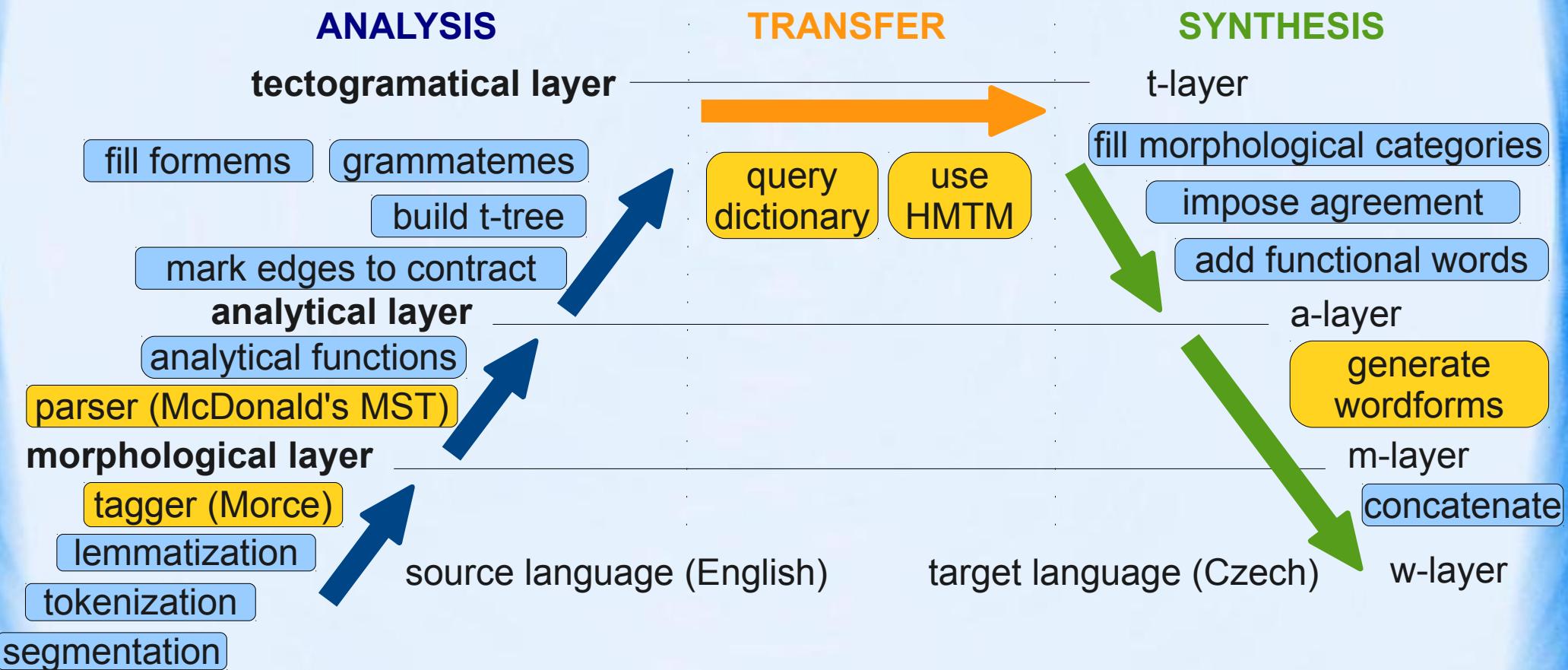
TectoMT – MT system

transfer over the tectogrammatical layer



TectoMT – MT system

rule based & statistical blocks



Demo Translation – Analysis

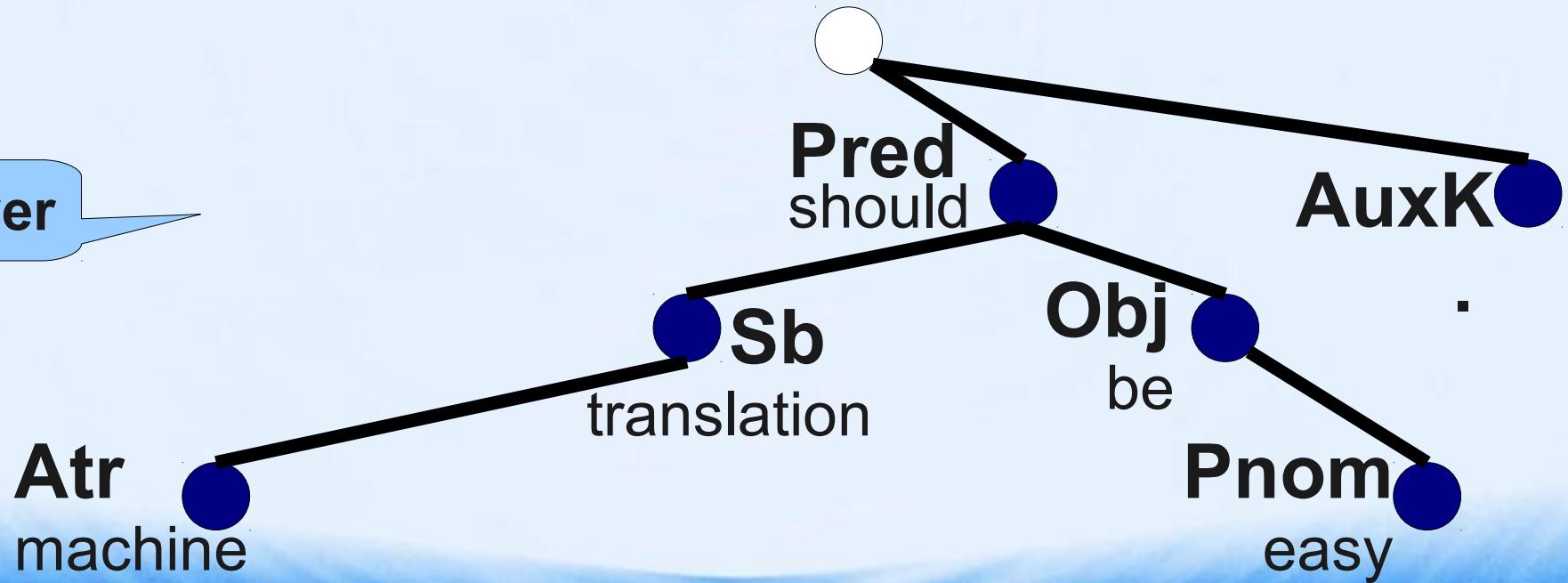
raw text

Machine translation should be easy.

m-layer

machine translation should be easy .
NN NN MD VB JJ .

a-layer



Demo Translation – Analysis

raw text

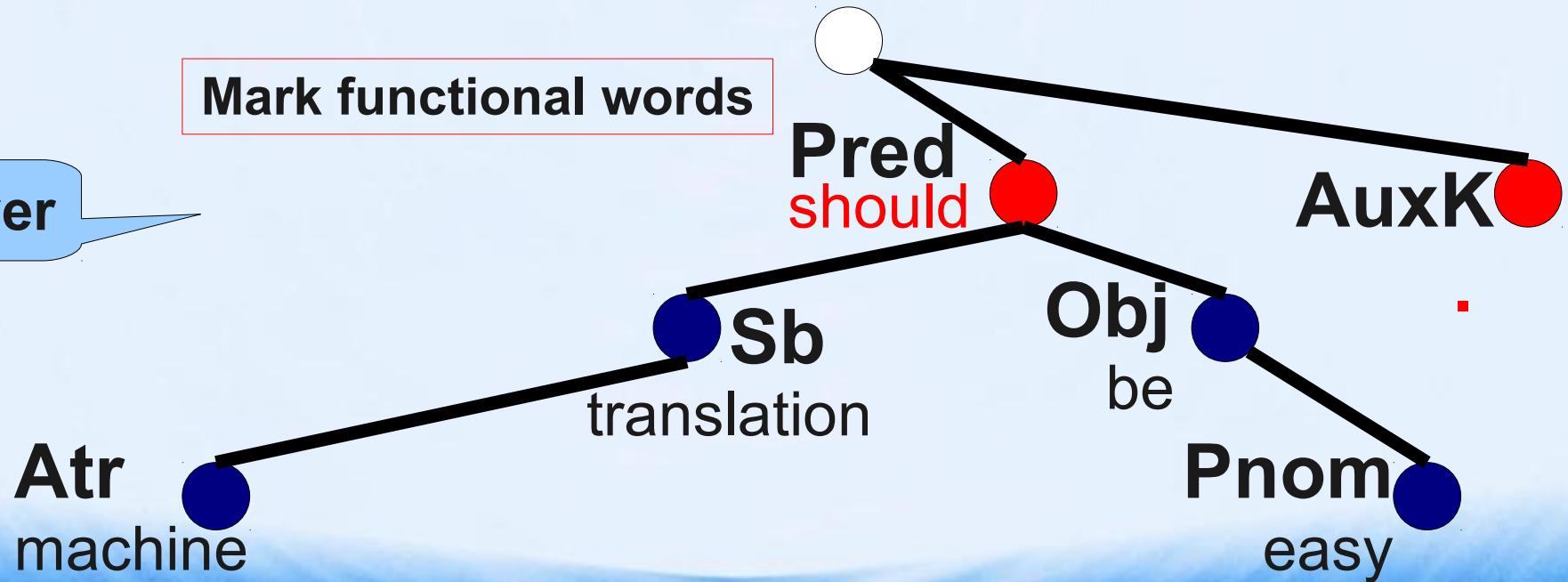
Machine translation should be easy.

m-layer

machine translation should be easy .

NN NN MD VB JJ .

a-layer



Demo Translation – Analysis

raw text

Machine translation should be easy.

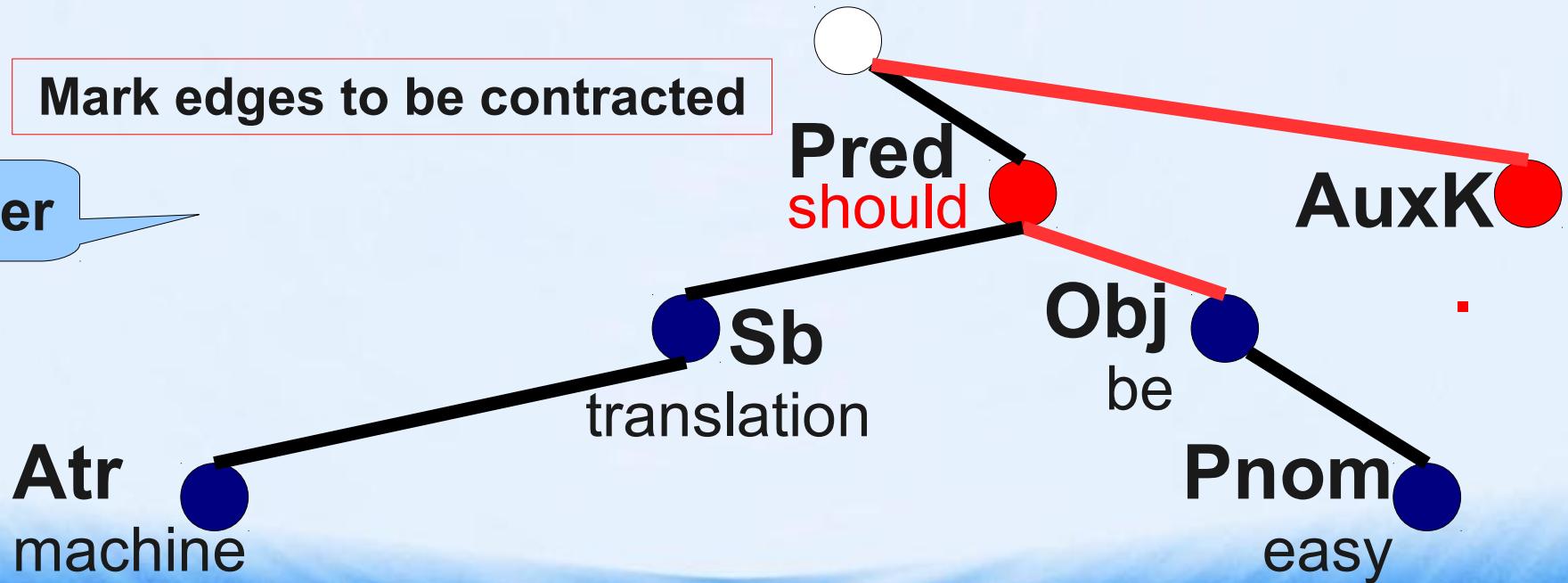
m-layer

machine translation should be easy .

NN NN MD VB JJ .

a-layer

Mark edges to be contracted



Demo Translation – Analysis

raw text

Machine translation should be easy.

m-layer

machine translation should be easy .
NN NN MD VB JJ .

Build t-tree (backbone)

t-layer

The diagram illustrates the step-by-step construction of a t-tree backbone for the sentence "Machine translation should be easy .". It begins with the raw text "Machine translation should be easy ." where each word is represented by a blue dot. Below the words are their part-of-speech tags: "NN", "NN", "MD", "VB", and "JJ". A blue speech bubble labeled "m-layer" points to this row. A red box labeled "Build t-tree (backbone)" contains the text "be". A white circle is positioned above the word "be". A black line connects this white circle to the blue dot representing "be". Another black line extends from the same white circle to the blue dot representing "easy". A third black line extends from the same white circle to the blue dot representing "translation". A fourth black line extends from the same white circle to the blue dot representing "machine". A blue speech bubble labeled "t-layer" points to the word "machine".

machine translation should be easy .

NN NN MD VB JJ .

Demo Translation – Analysis

raw text

Machine translation should be easy.

m-layer

machine translation should be easy .
NN NN MD VB JJ .

t-layer

Fill formems

n:attr
machine

n:subj
translation

be
v:fin

easy
adj:compl

Demo Translation – Analysis

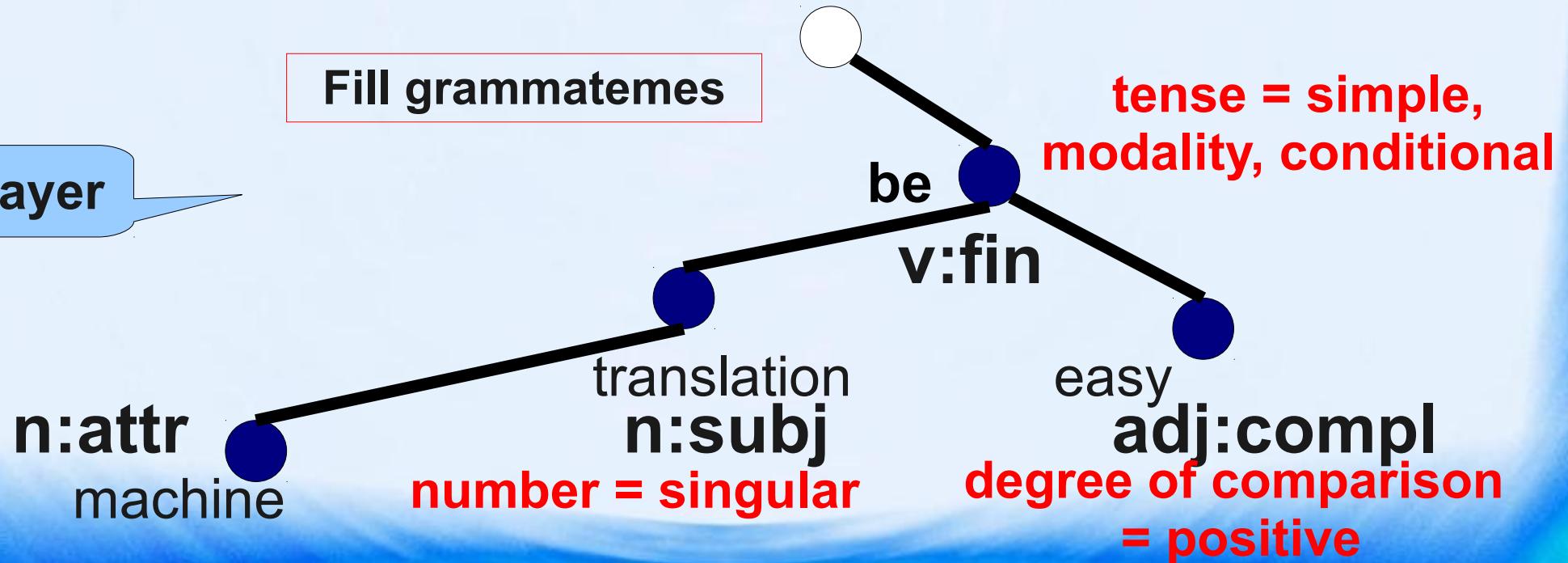
raw text

Machine translation should be easy.

m-layer

machine translation should be easy .
NN NN MD VB JJ .

t-layer



Demo Translation – Transfer

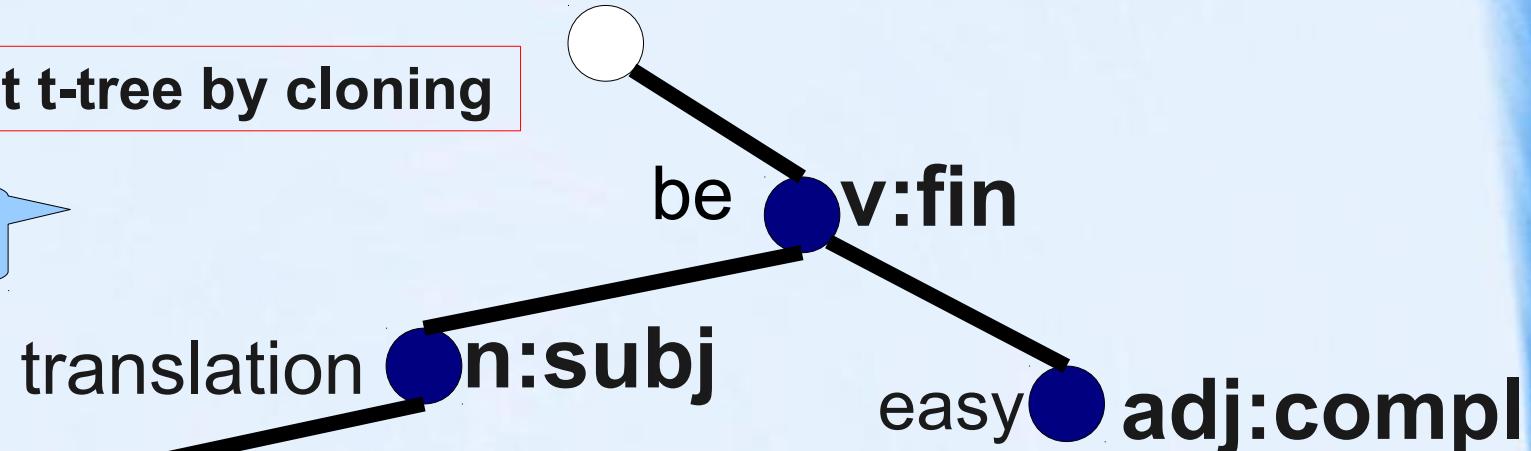
Build target t-tree by cloning

source t-layer

machine n:attr

target t-layer

machine n:attr



be v:fin

easy adj:compl

be v:fin

easy adj:compl

Demo Translation – Transfer

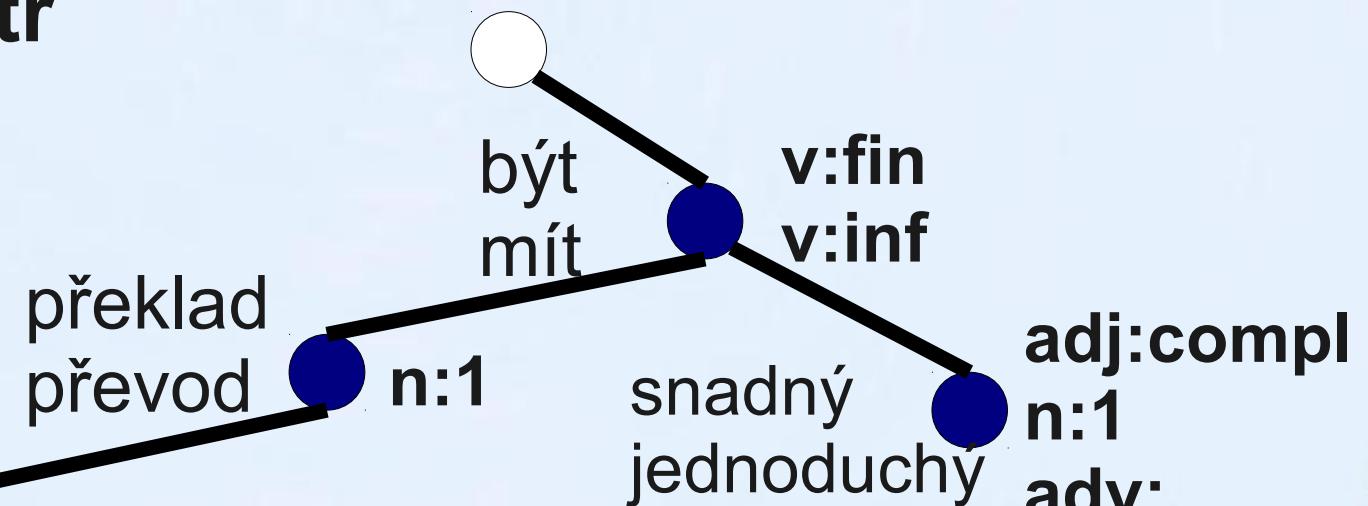
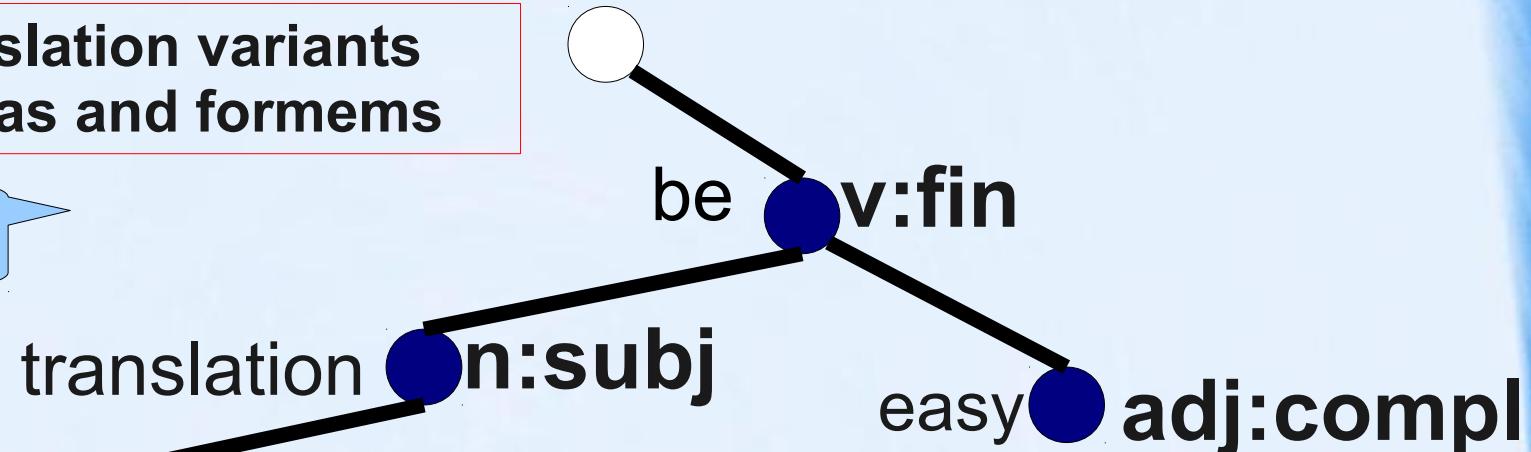
Get translation variants
for lemmas and formems

source t-layer

machine n:attr

target t-layer

počítač
stroj
strojový
n:2
n:attr
adj:attr



Demo Translation – Transfer

Select the best combination
of lemmas and formems

source t-layer

machine

target t-layer

počítač
stroj
strojový

translation

překlad
převod

n:2
n:attr
adj:attr

be v:fin

easy adj:compl

být
mít v:fin
v:inf

snadný
jednoduchý adj:compl
n:1
adv:

n:1

Demo Translation – Synthesis

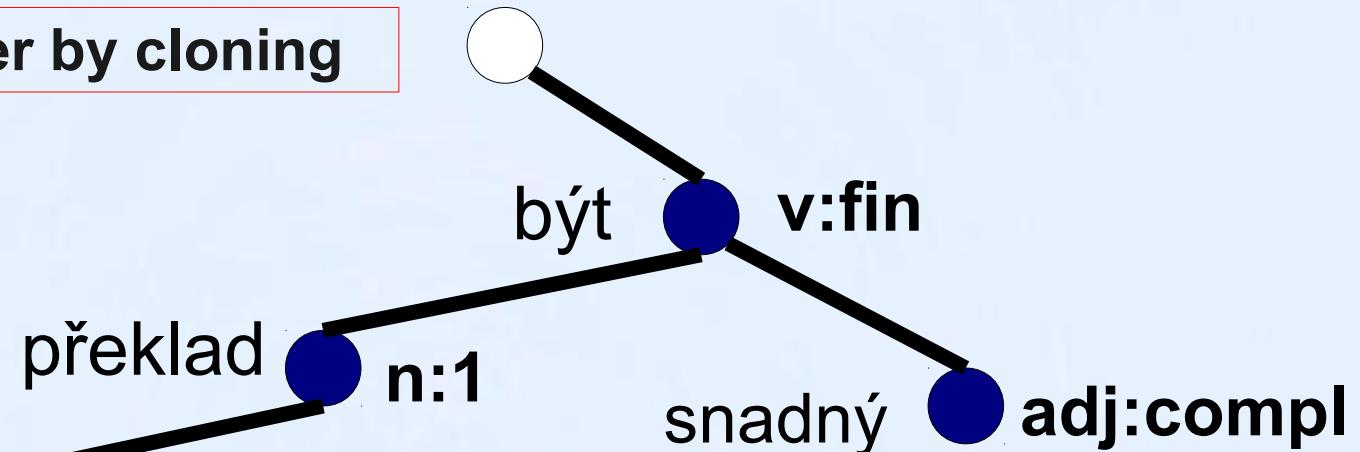
Build target a-layer by cloning

target t-layer

strojový

target a-layer

strojový



Demo Translation – Synthesis

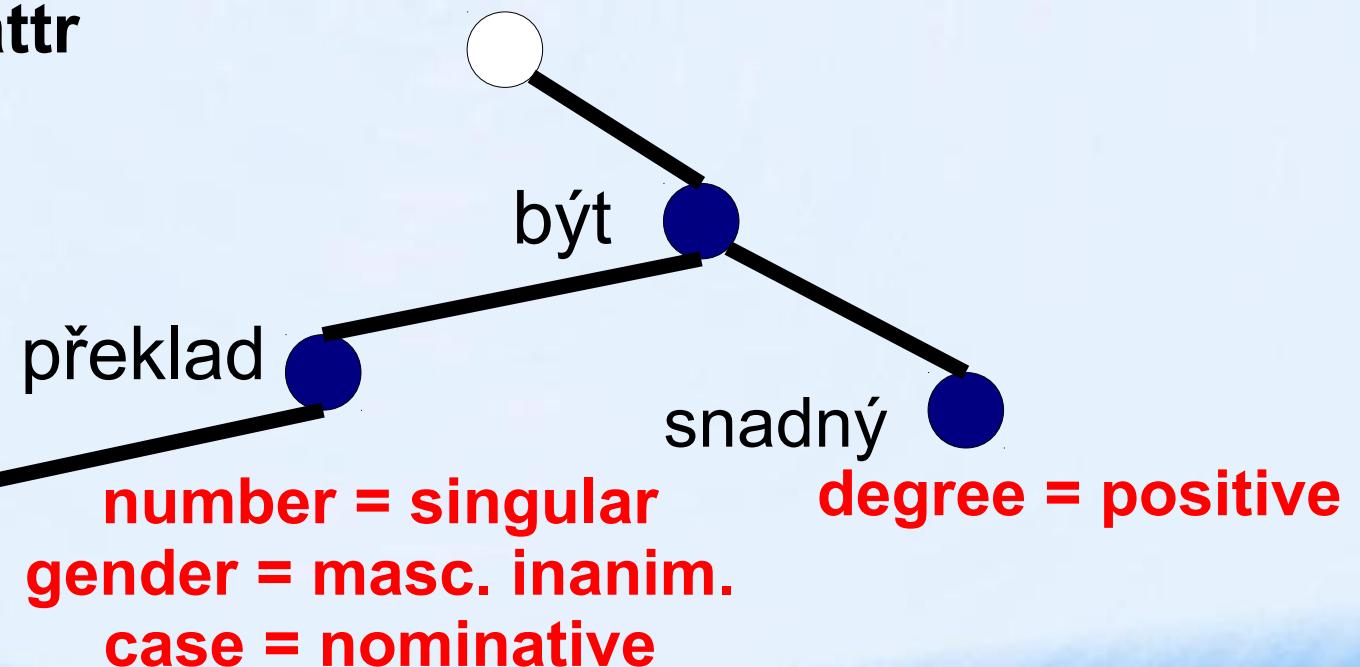
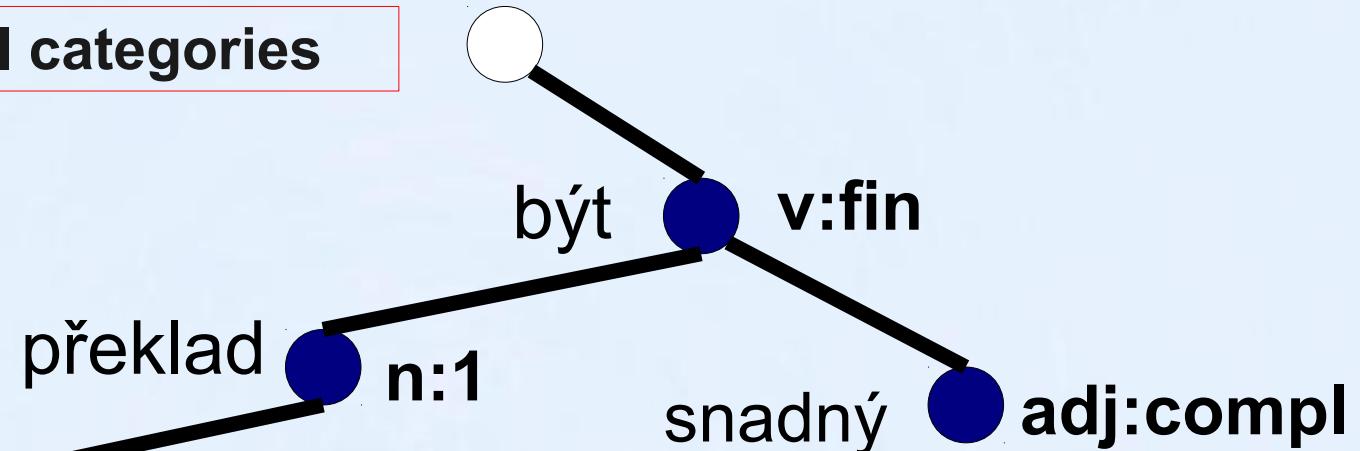
Fill morphological categories

target t-layer

strojový adj:attr

target a-layer

strojový
degree = positive



Demo Translation – Synthesis

Impose agreement

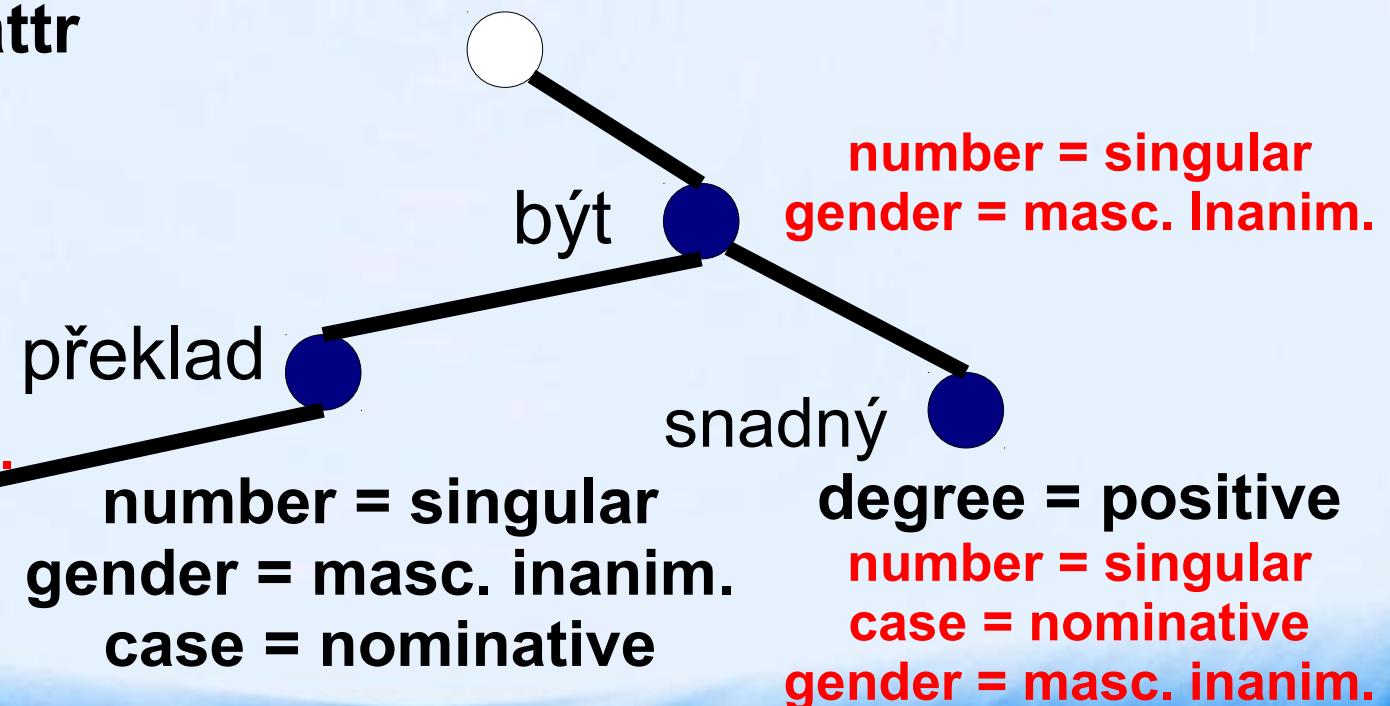
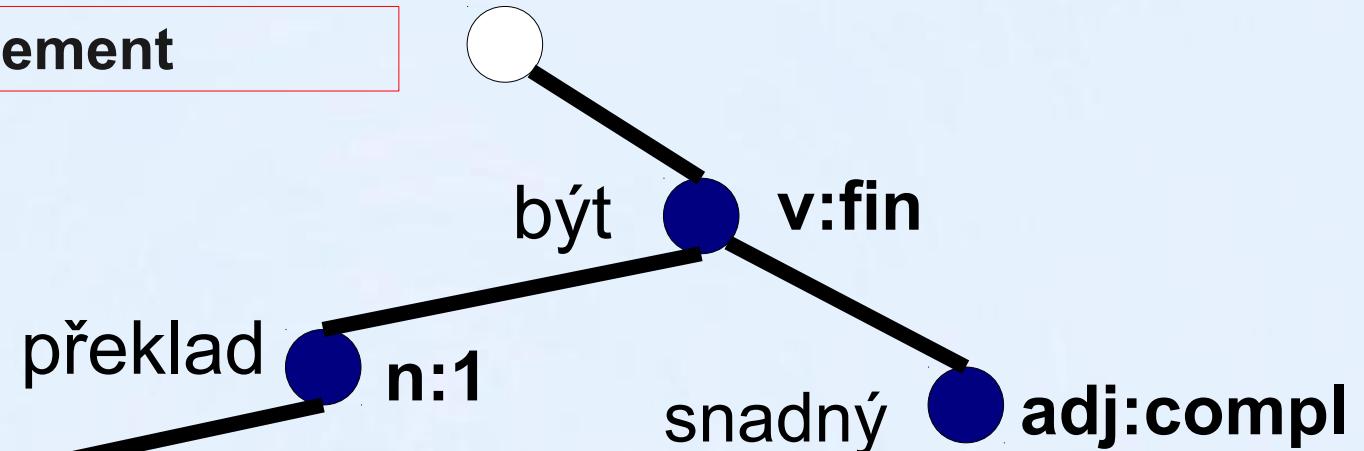
target t-layer

strojový adj:attr

target a-layer

number = singular
case = nominative
gender = masc. inanim.

strojový
degree = positive



Demo Translation – Synthesis

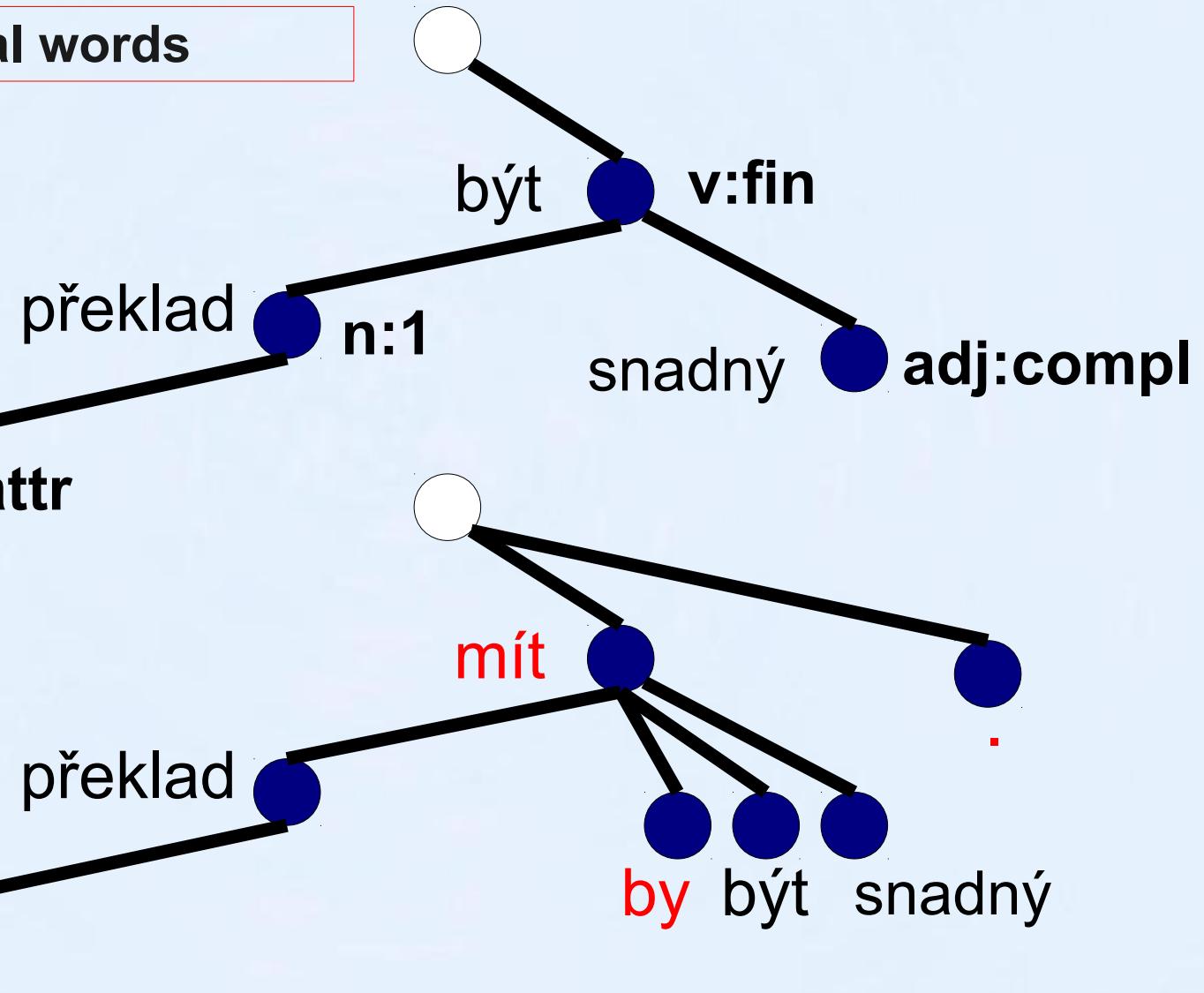
Add functional words

target t-layer

strojový adj:attr

target a-layer

strojový



Demo Translation – Synthesis

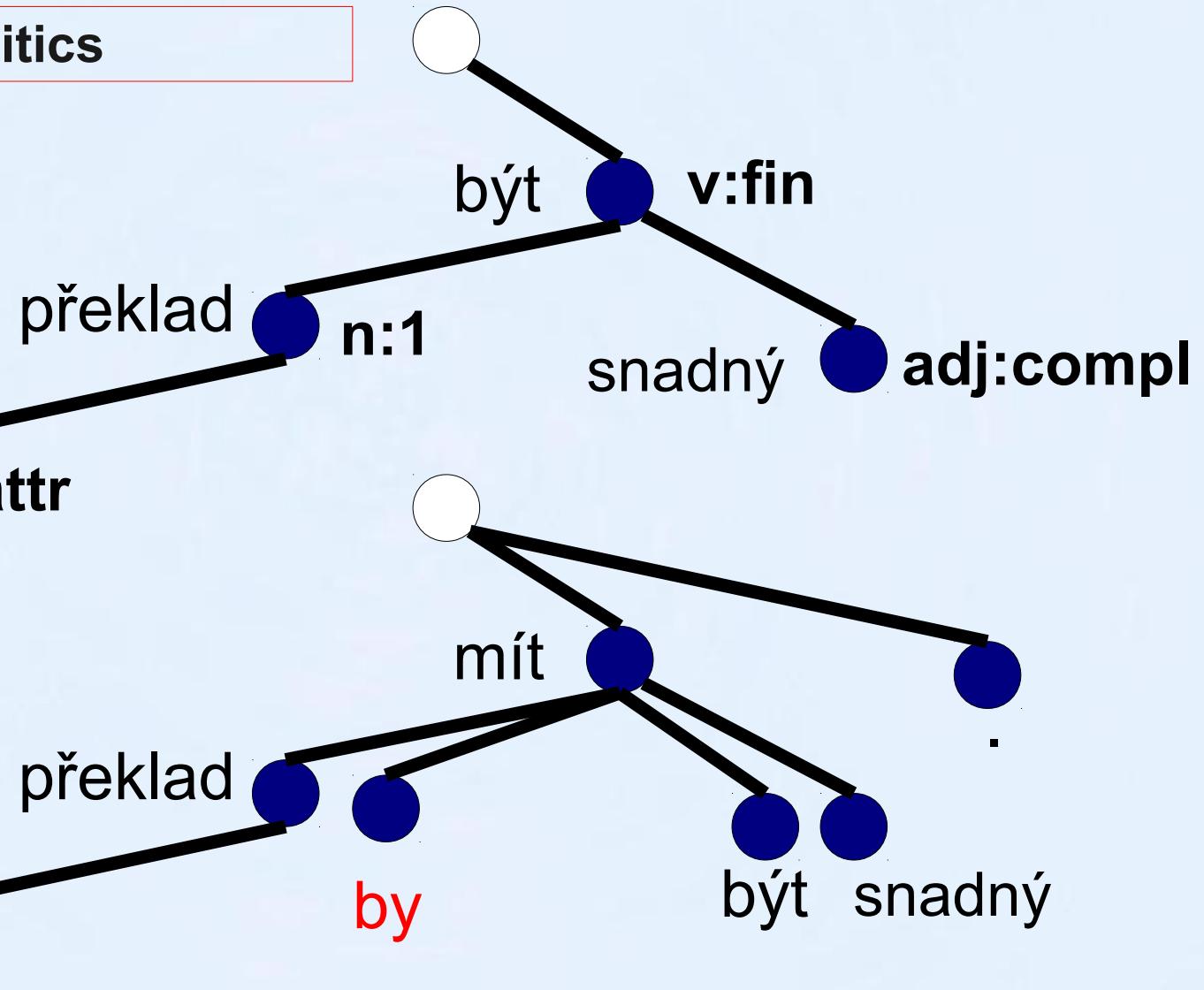
Reorder clitics

target t-layer

strojový adj:attr

target a-layer

strojový



Demo Translation – Synthesis

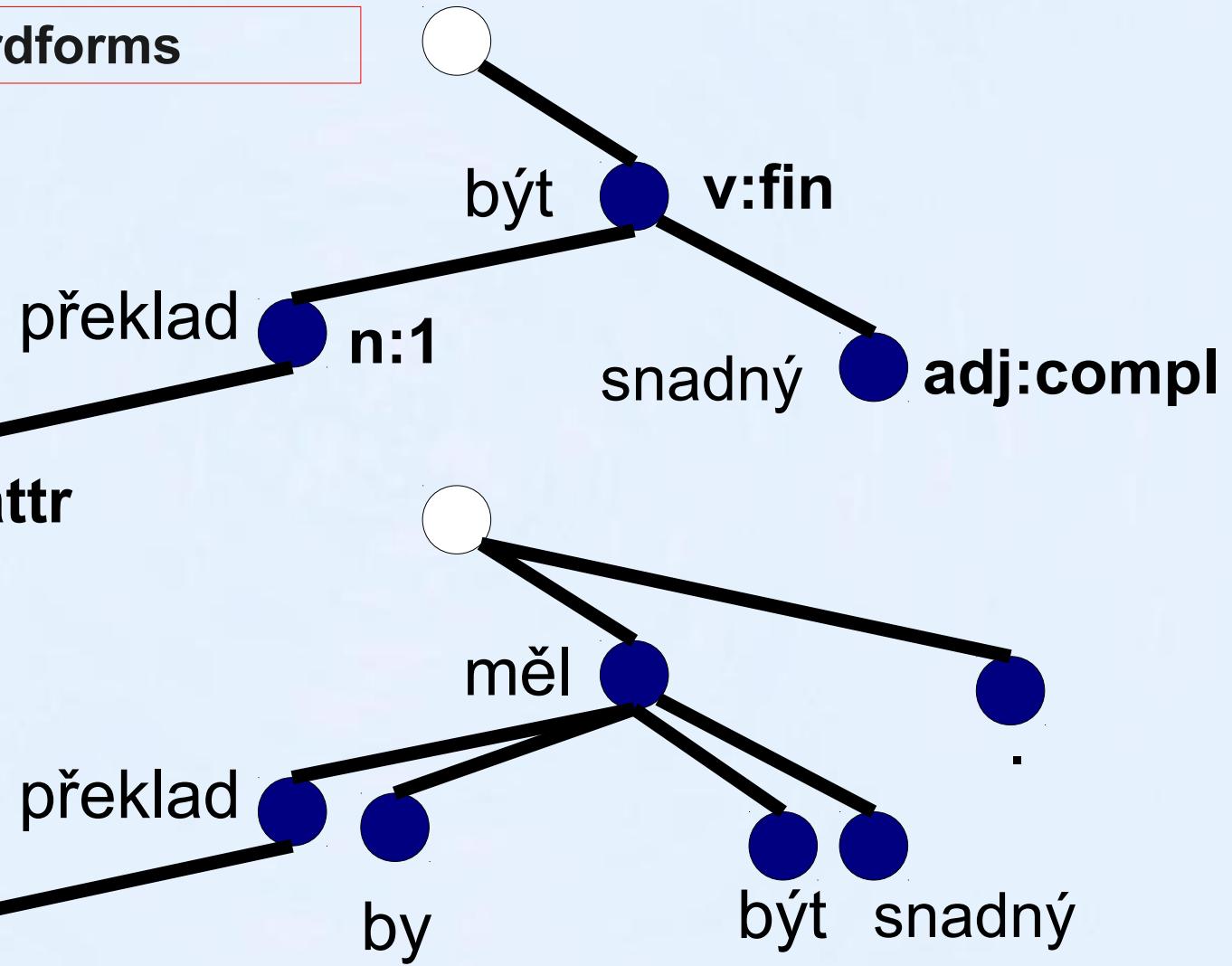
Generate wordforms

target t-layer

strojový adj:attr

target a-layer

strojový



Demo Translation – Synthesis

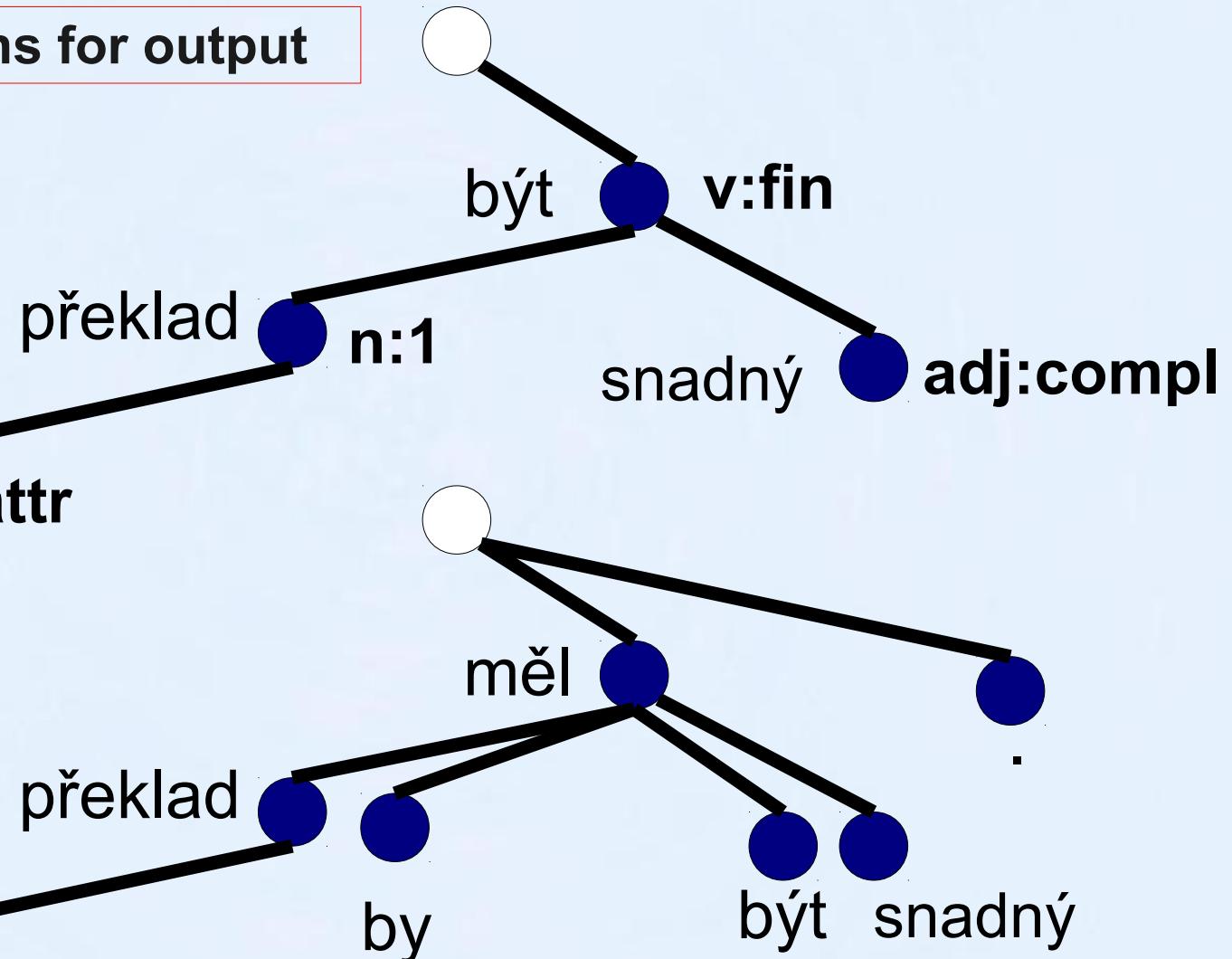
Concatenate tokens for output

target t-layer

strojový adj:attr

target a-layer

strojový

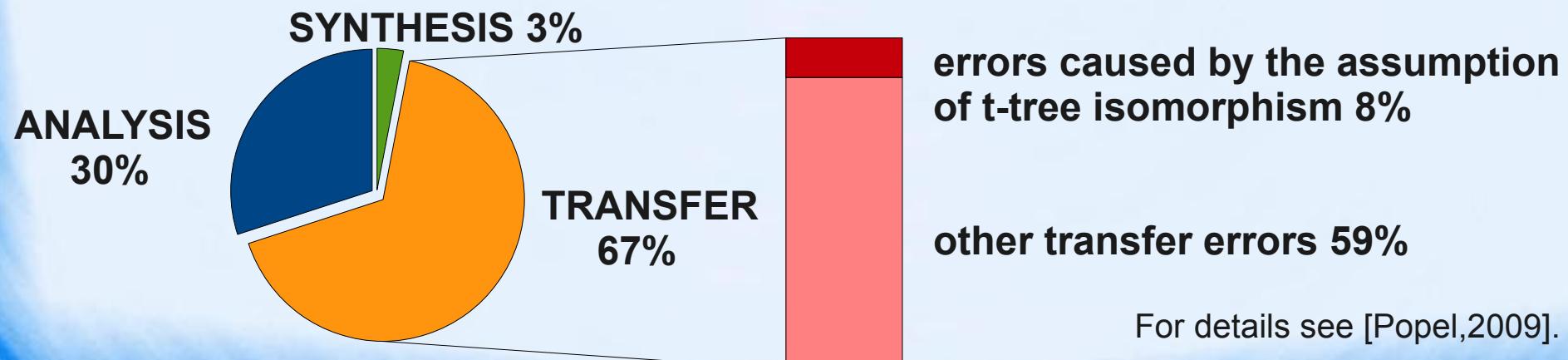


Strojový překlad by měl být snadný.

Annotation of Translation Errors

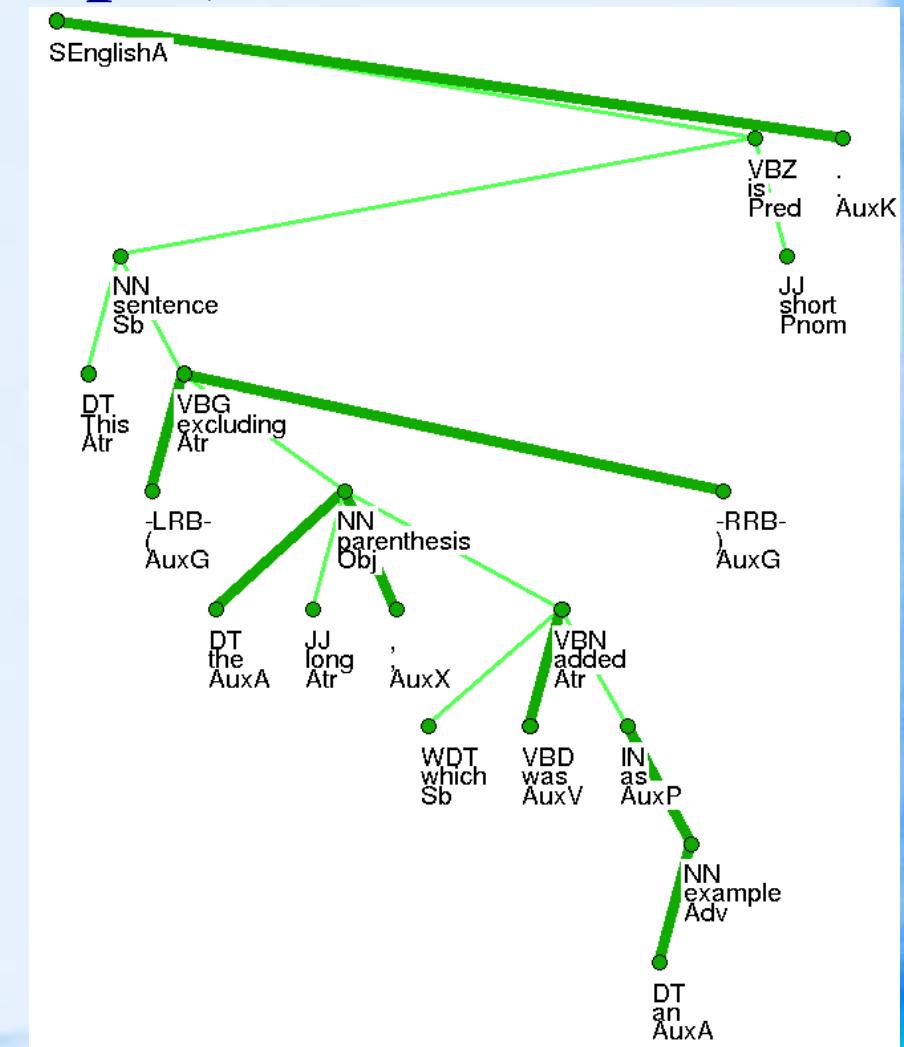
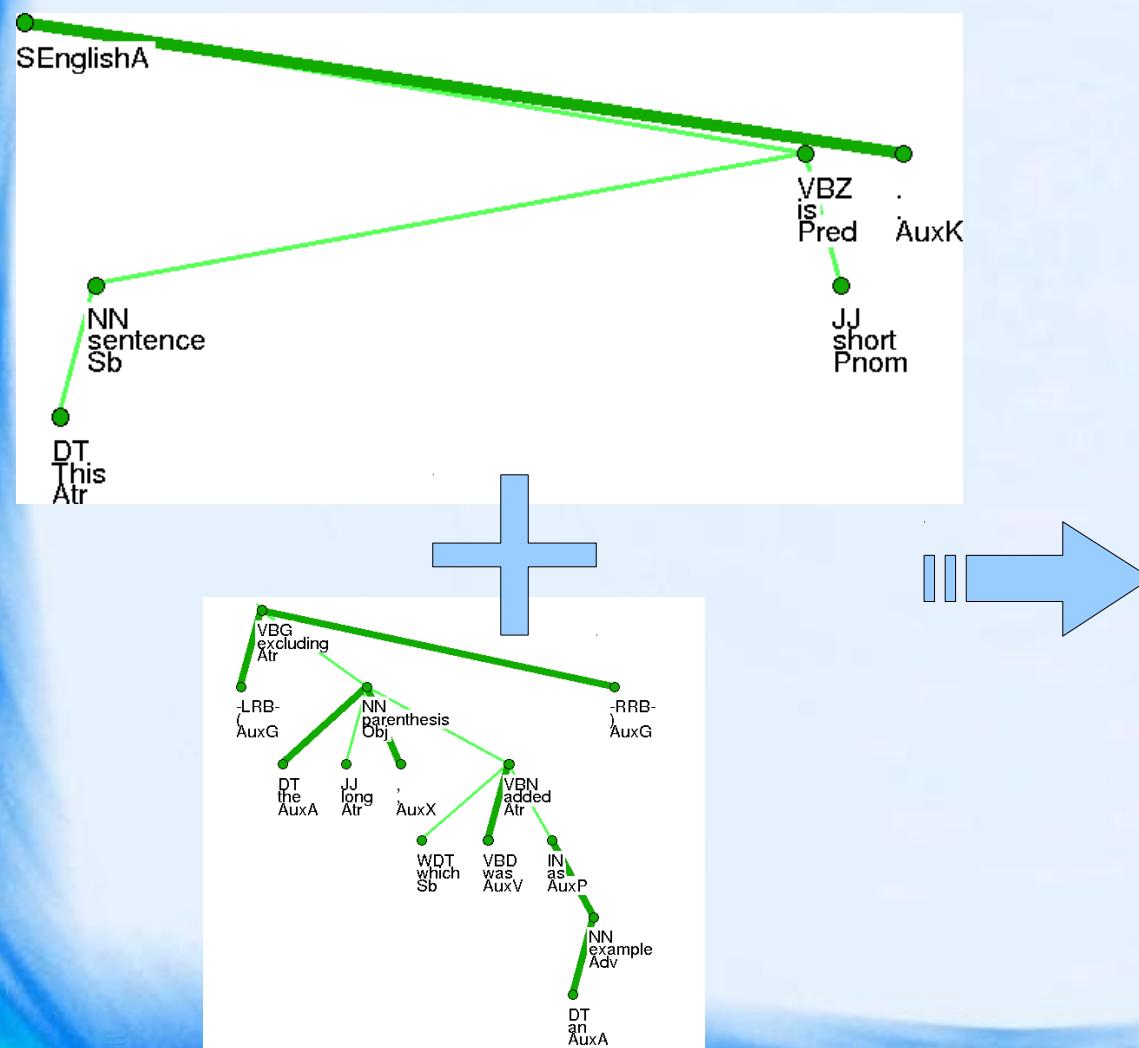
sample of 250 sentences, 1463 errors in total

| | |
|---------------|--|
| Type | lemma, formeme, gram., w. order, ... |
| Subtype | gram: gender, person, tense, ... |
| Seriousness | serious, minor |
| Circumstances | coordination, named entity, numbers |
| Source | tok, lem, tagger, parser, tecto, trans , x , syn , ? |



Parsing Parentheses

This sentence (excluding the long parenthesis, which was added as an example) is short.



HMTM – Motivation

Select the best combination
of lemmas and formems

source t-layer

machine

target t-layer

počítač
stroj
strojový

translation

překlad
převod

n:2
n:attr
adj:attr

be v:fin

easy adj:compl

být
mít v:fin
v:inf

snadný
jednoduchý

adj:compl
n:1
adv:

HMTM – Motivation

Select the best label
for each node

source t-layer

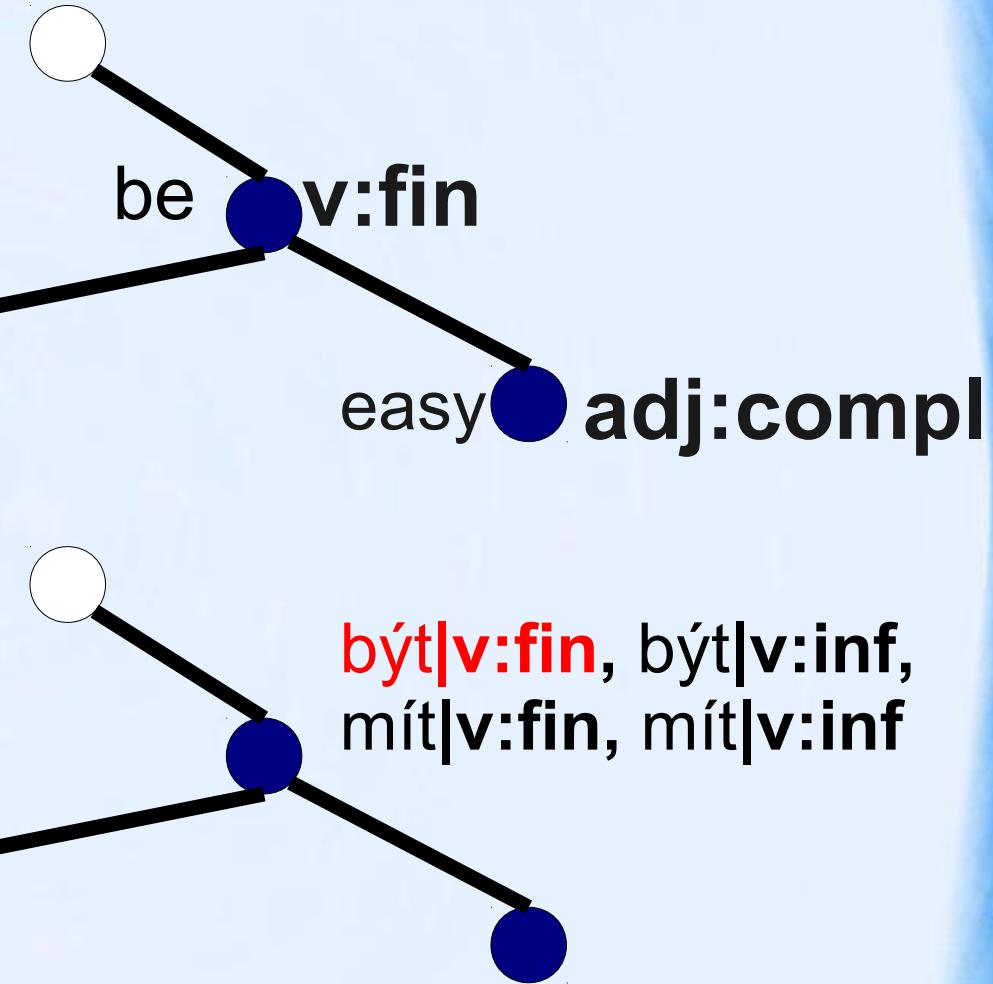
machine

target t-layer

počítač|n:2,
počítač|n:attr,
strojový|adj:attr, ...

translation
n:subj

překlad|n:1,
převod|n:1



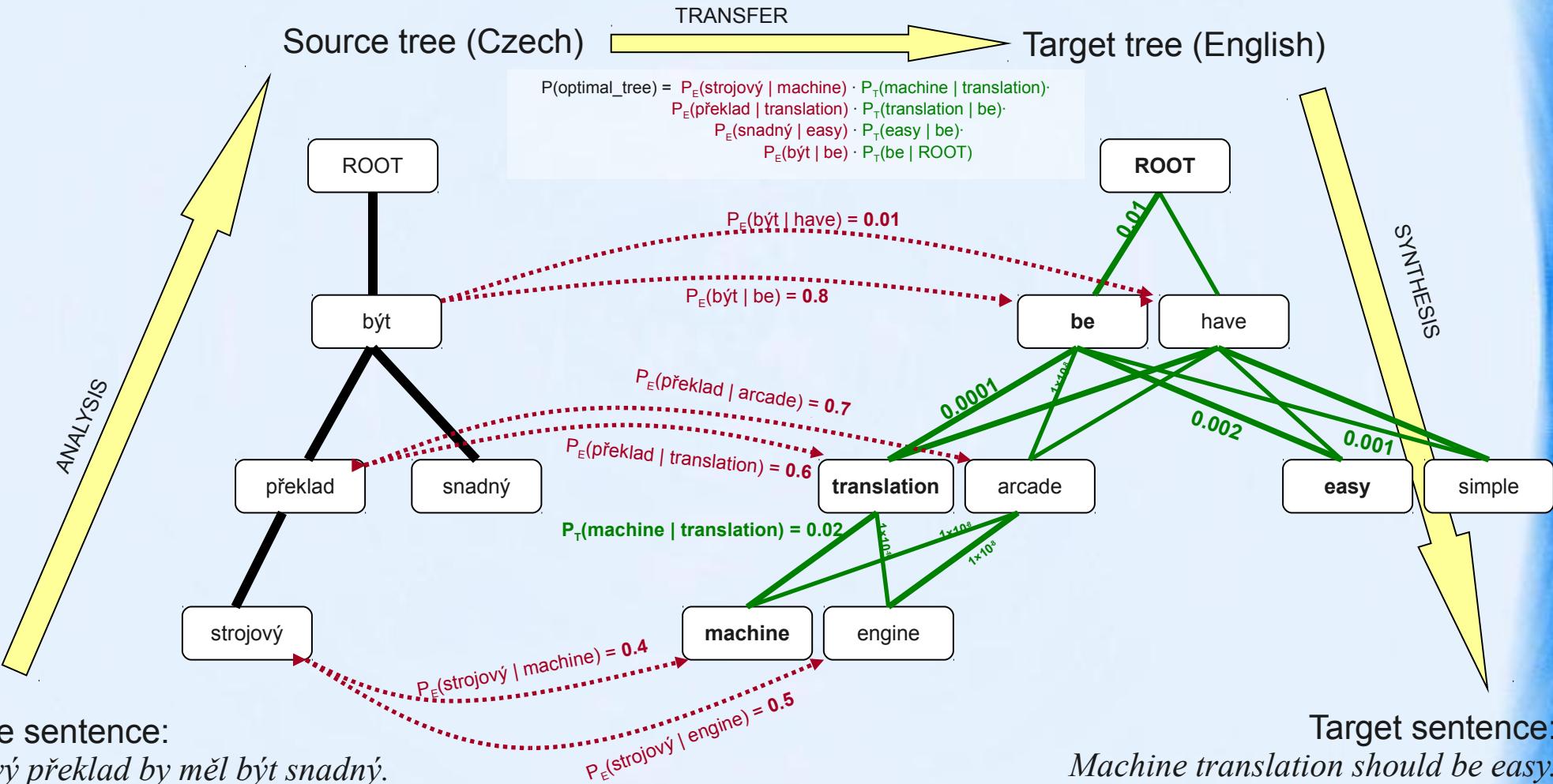
být|v:fin, být|v:inf,
mít|v:fin, mít|v:inf

snadný|adj:compl,
jednoduchý|adj:compl, ...

HMTM in Theory

- Introduced by [Crouse,1998], used in signal processing, image segmentation etc. (See [Durand,2004].)
- (V, E) – rooted tree
- X – sequence of random variables (hidden states) for V
- Y – sequence of random variables (observable symbols)
- $P(X_v | X_{\text{parent}(v)})$ – transition probabilities
- $P(Y_v | X_v)$ – emission probabilities
- Tree-Markov property:
 $\forall v \in V \setminus \{\text{root}\}, \forall w \in V \setminus \text{subtree}(v) :$
 $P(X_{\text{subtree}(v)} | X_{\text{parent}(v)}, X_w) = P(X_{\text{subtree}(v)} | X_{\text{parent}(v)})$
i.e. given $X_{\text{parent}(v)}$, all hidden states of the subtree rooted in v are conditionally independent of any other nodes.
- The most probable hidden tree labeling X^* can be obtained given the observed tree labeling Y can be obtained using tree-modified Viterbi algorithm.

HMTM in MT



$P_E(\text{source} \mid \text{target})$... emission probabilities ... **translation model**

$P_T(\text{dependent} \mid \text{governing})$... transition probabilities ... **target-language tree model**

Conclusion – Results

| | NIST | BLEU |
|------------------|-------|-------|
| baseline (WMT09) | 3.974 | 0.066 |
| modified | 4.716 | 0.098 |

- 2777 sentences from WMT2009 (news-test2009)
- 1 reference translation
- Most helpful improvements: HMTM & parsing

Future plans

- Better language models for dependency trees
- Machine learning techniques for tuning TM & LM

Examples of Translation

A miss by an inch
is a miss by a mile.

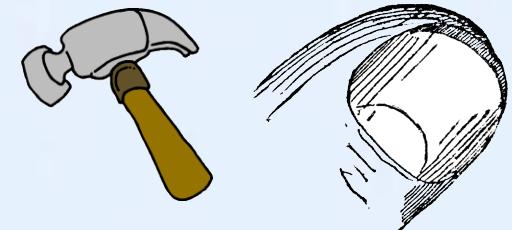
I'd rather be a hammer
than a nail.

A bird in the hand is worth
two in the bush.

Slečna palec je slečna miliónu.



Spíše bych byl kladivo než nehét.



Pták v ruce je cenný
dvakrát v Bushovi.



References

- TectoMT: <http://ufal.mff.cuni.cz/tectomt>
- [Popel,2009] Martin Popel: Ways to Improve the Quality of English-Czech Machine Translation. Master's thesis, ÚFAL, MFF UK, Prague, 2009.
- [Crouse,1998] Matthew Crouse, Robert Nowak, and Richard Baraniuk: Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Transactions on Signal Processing*, 46(4):886–902.1998.
- [Durand,2004] Jean-Baptiste Durand, Paulo Gonçalvès, Yann Guédon: Computational Methods for Hidden Markov Tree Models – An Application to Wavelet Trees *IEEE Transactions on Signal Processing*, 2004.