Deep Syntactic Machine Translation with Hidden Markov Tree Models

Martin Popel
ÚFAL, Charles University in Prague

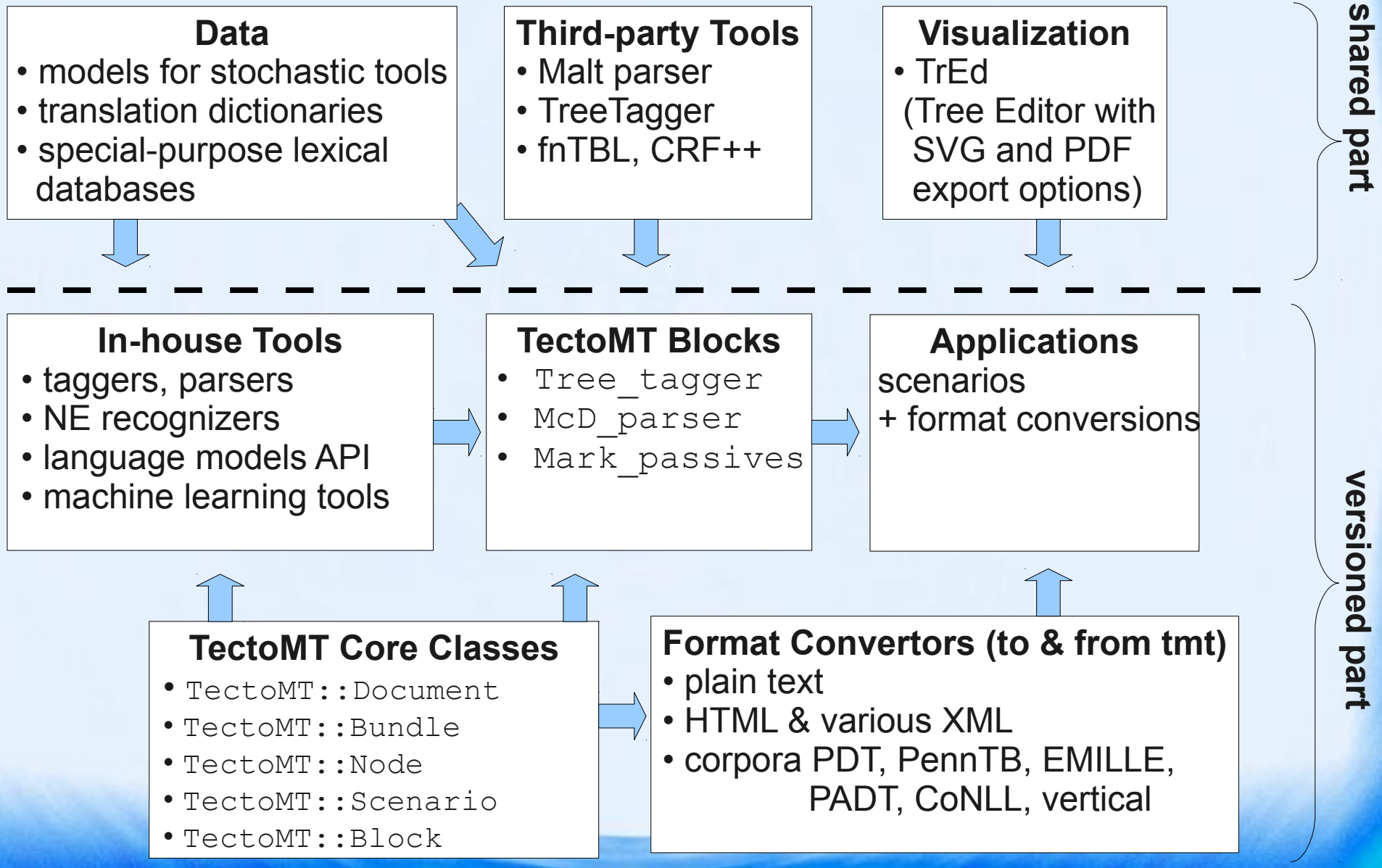
5th PIRE Meeting, Dec. 12 2009

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

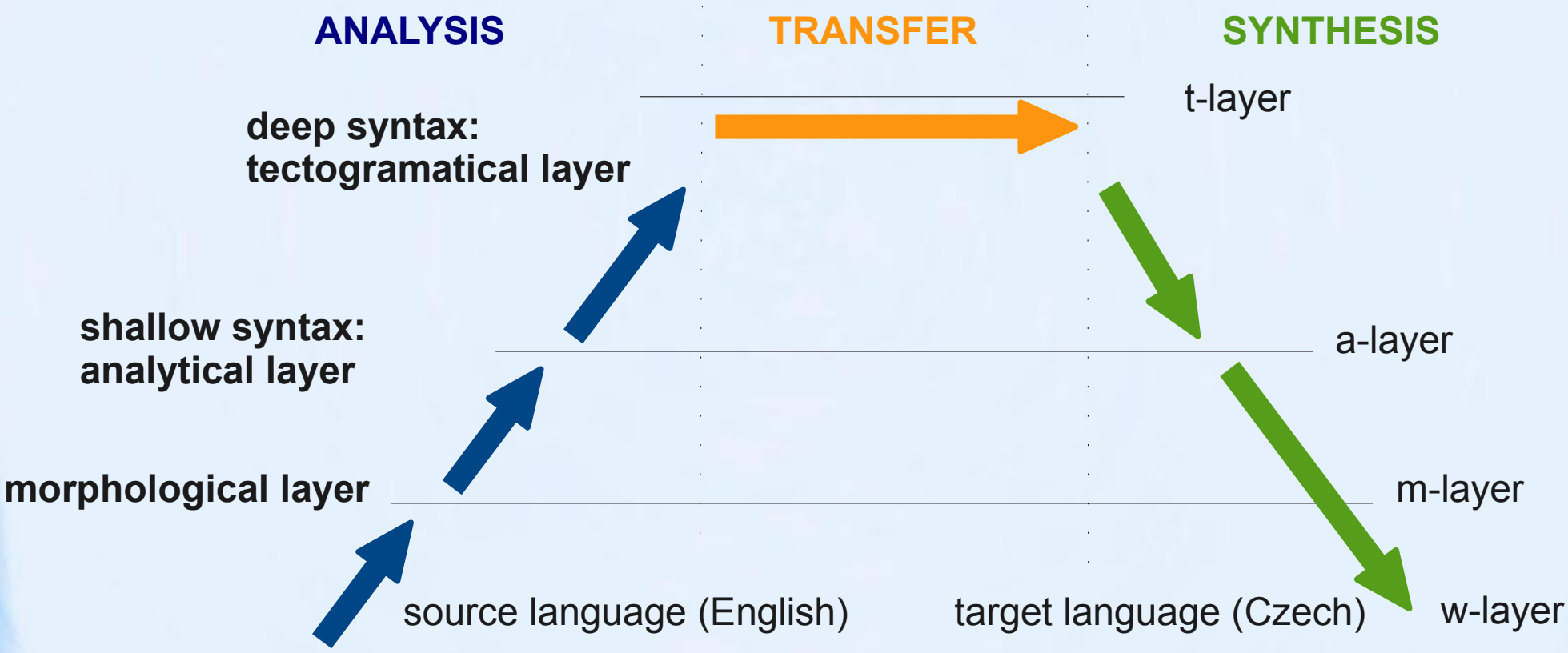
TectoMT – framework for NLP

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



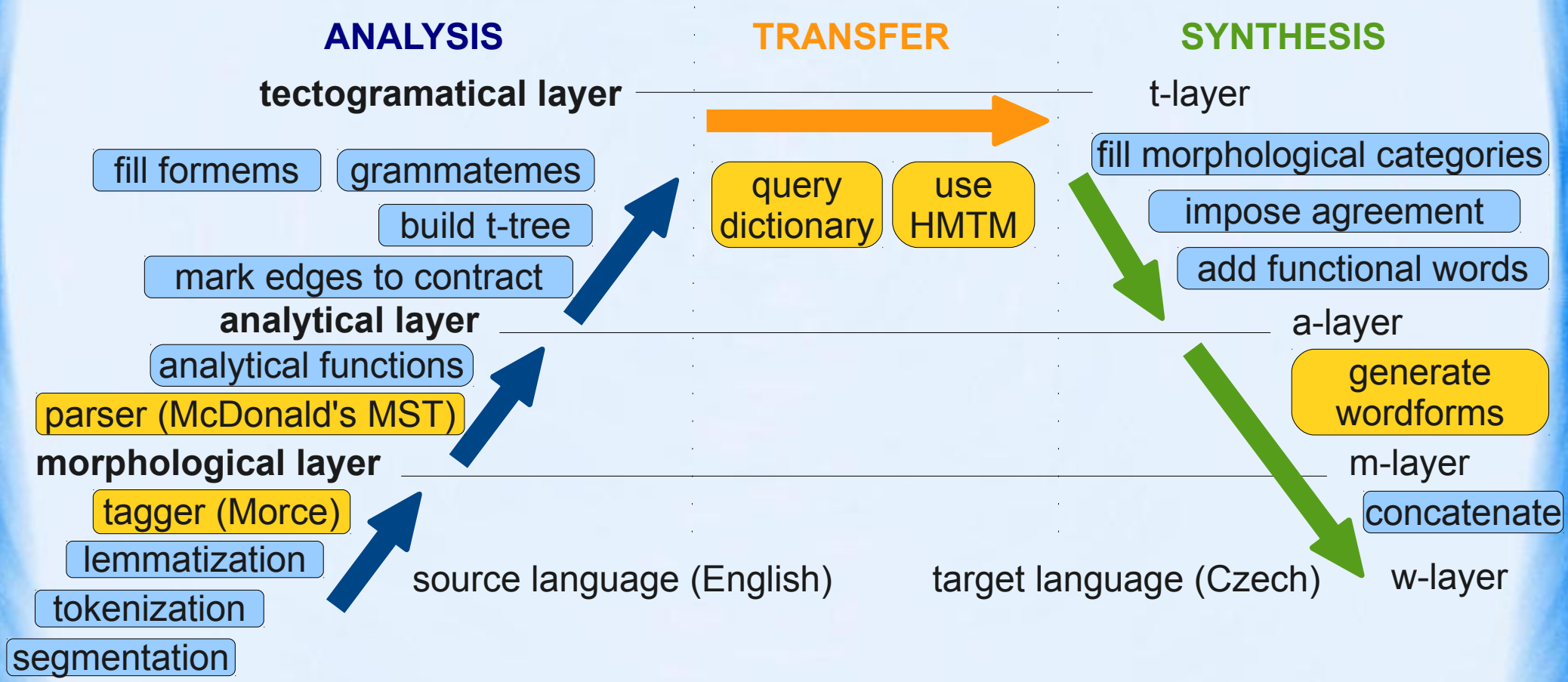
TectoMT – MT system

transfer over the tectogrammatical layer

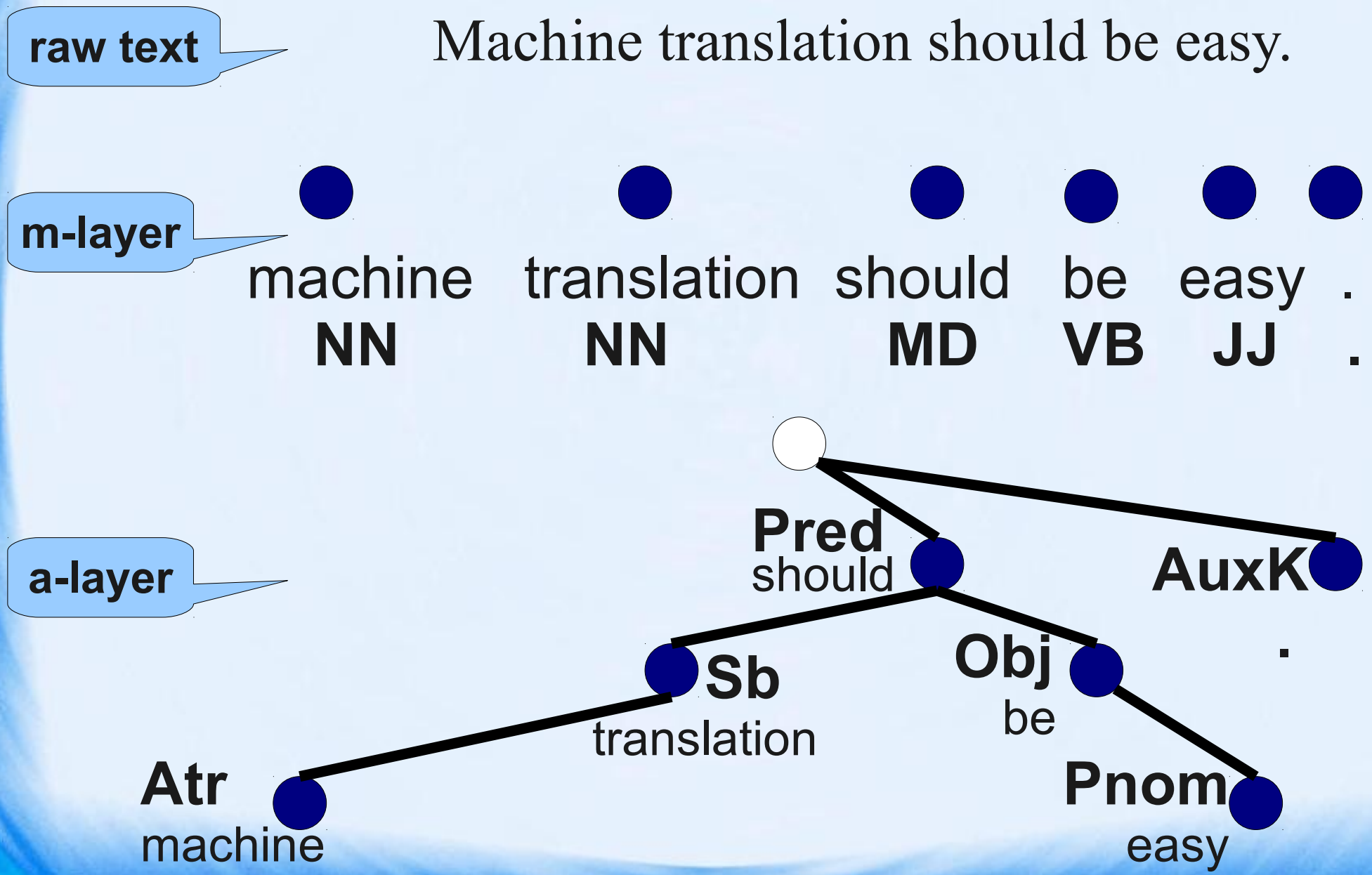


TectoMT – MT system

rule based & statistical blocks



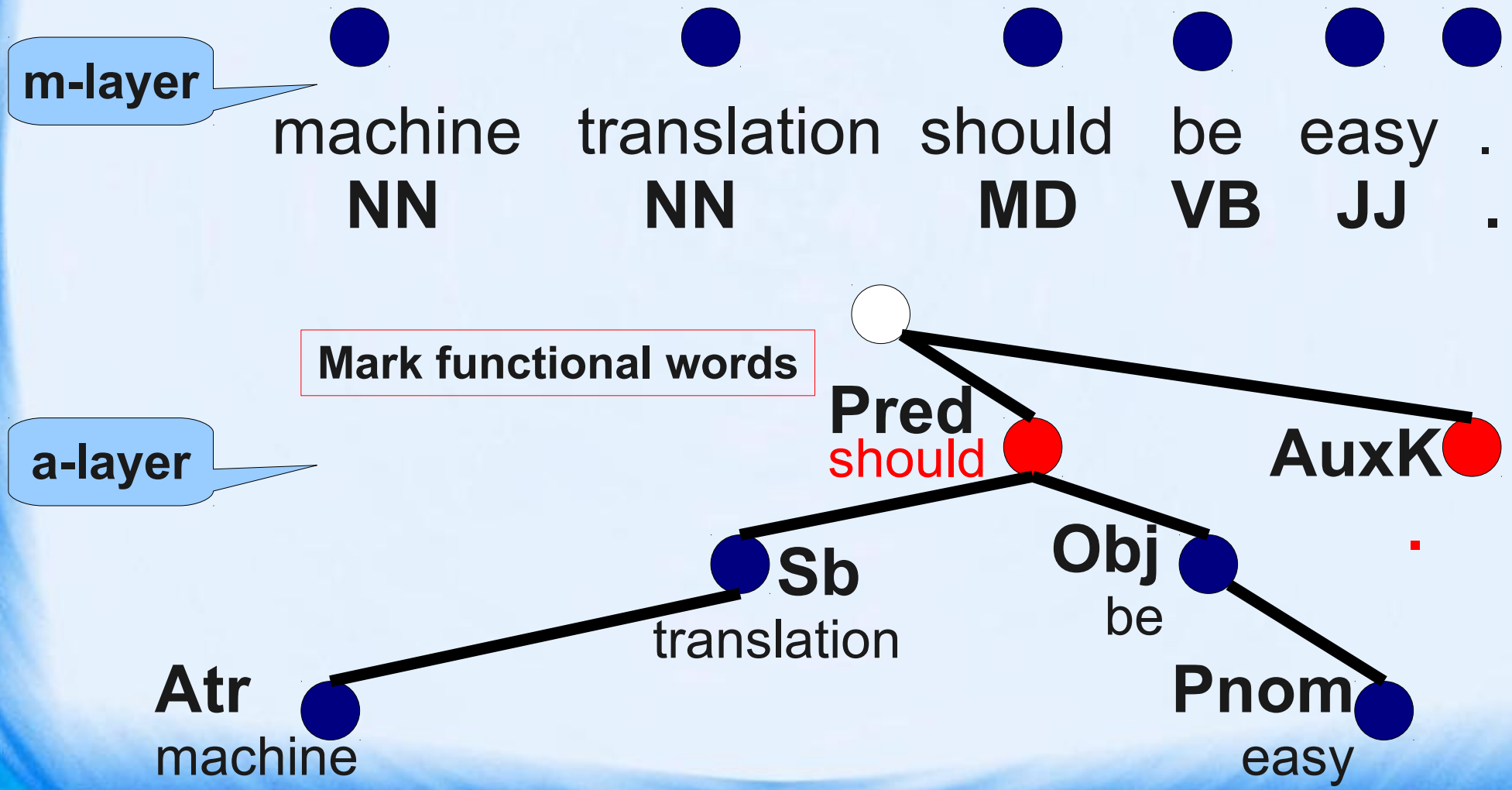
Demo Translation – Analysis



Demo Translation – Analysis

raw text

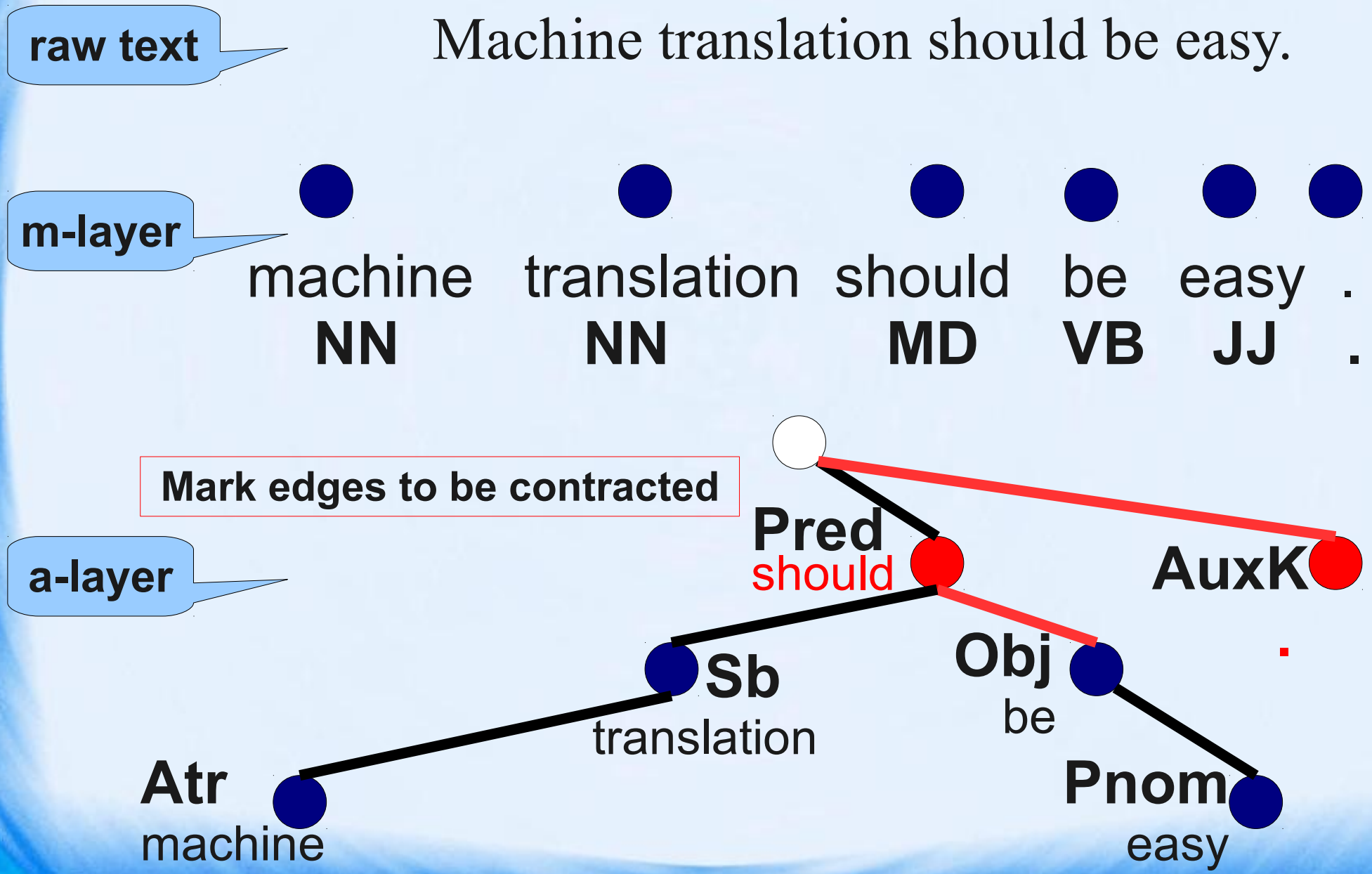
Machine translation should be easy.



Demo Translation – Analysis

raw text

Machine translation should be easy.



Demo Translation – Analysis

raw text

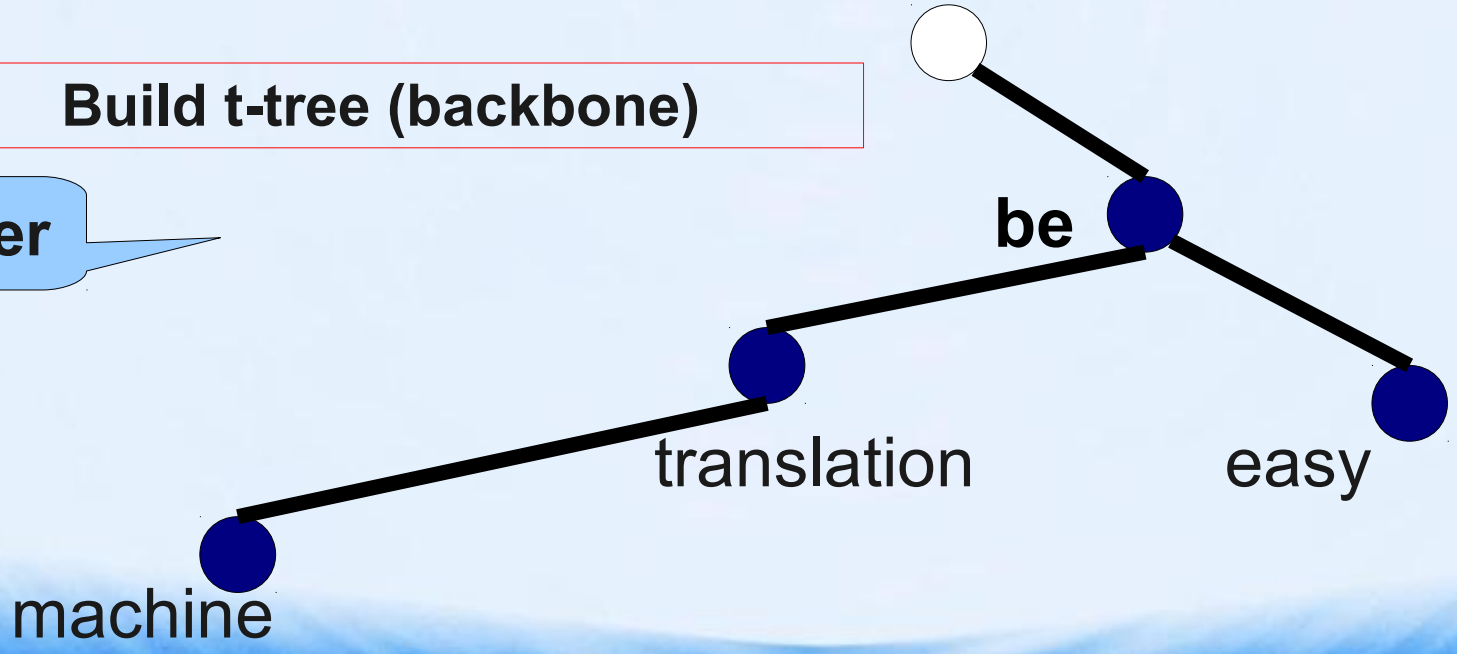
Machine translation should be easy.

m-layer

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

Build t-tree (backbone)

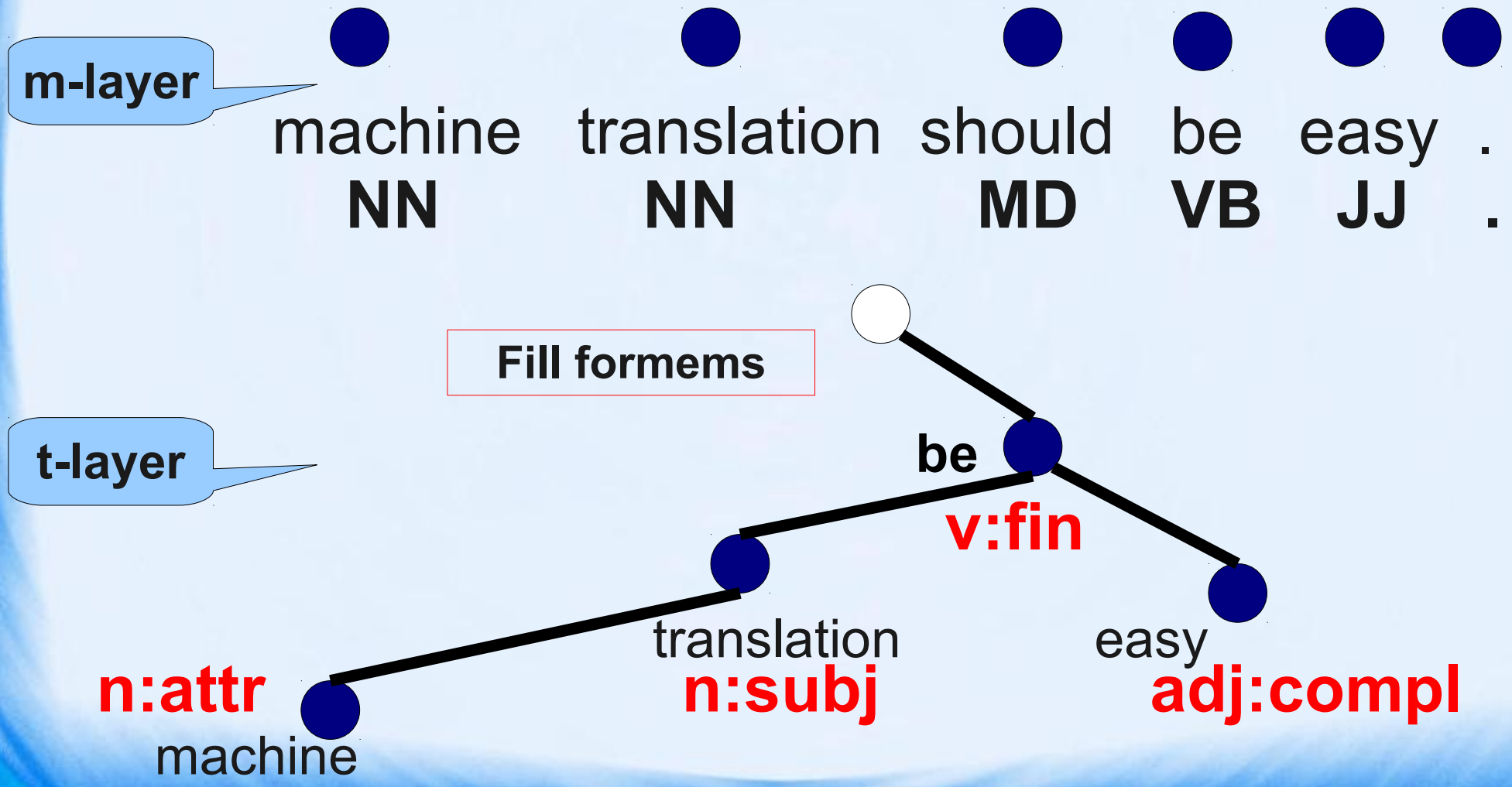
t-layer



Demo Translation – Analysis

raw text

Machine translation should be easy.



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 grammatemes

tense = simple,
modality, conditional

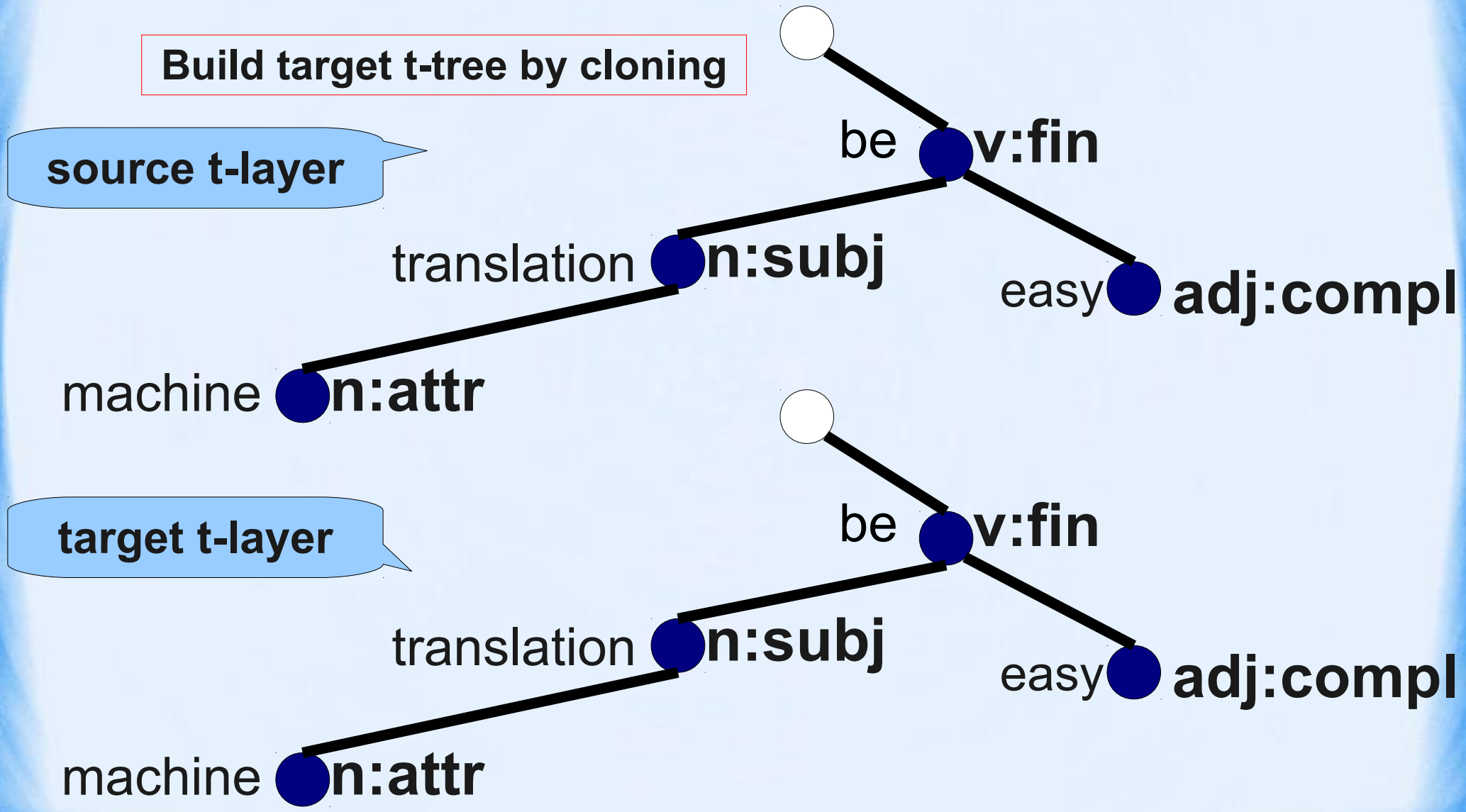
be
v:fin

translation
n:subj
number = singular

easy
adj:compl
degree of comparison
= positive

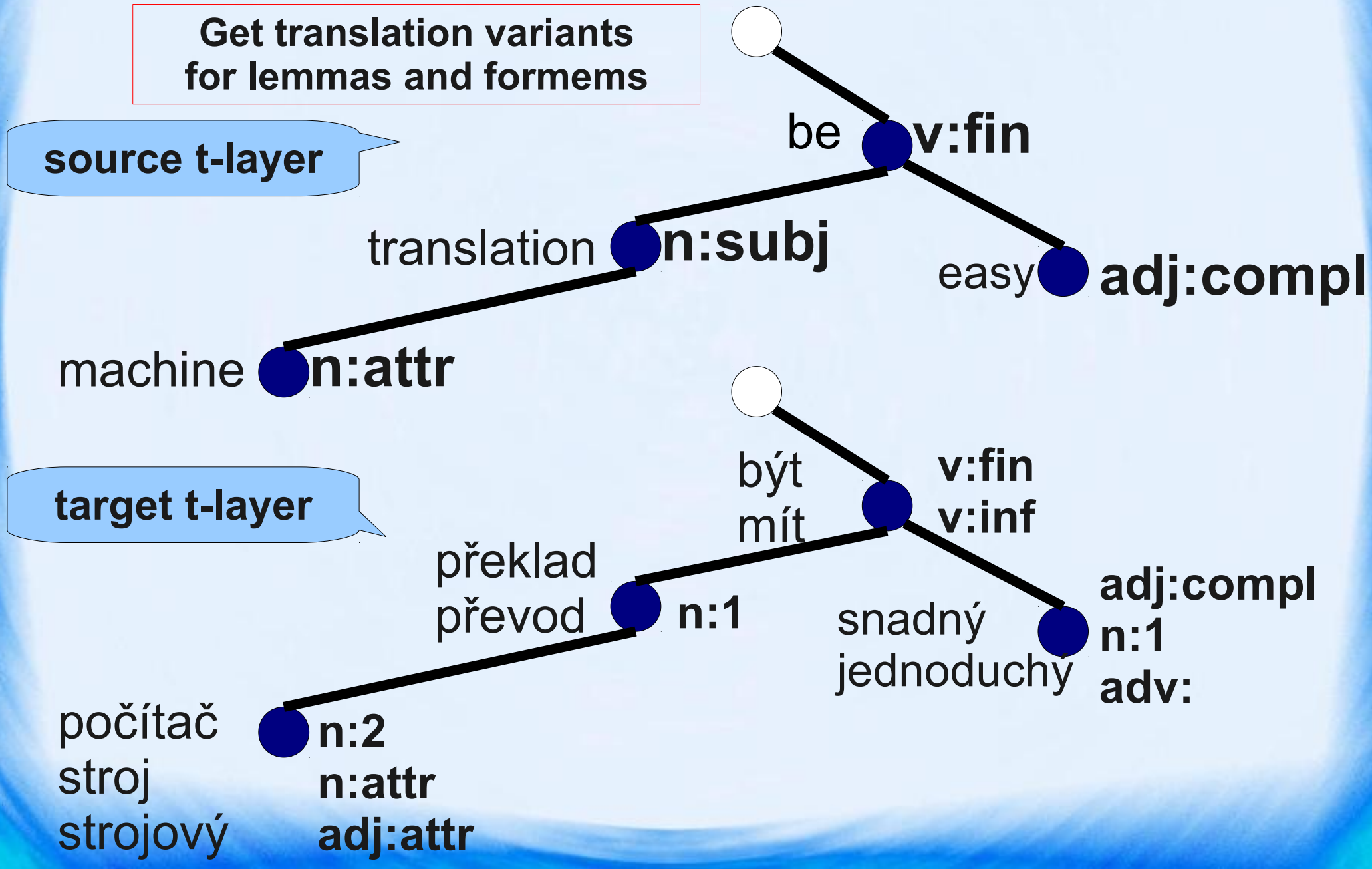
n:attr
machine

Demo Translation – Transfer

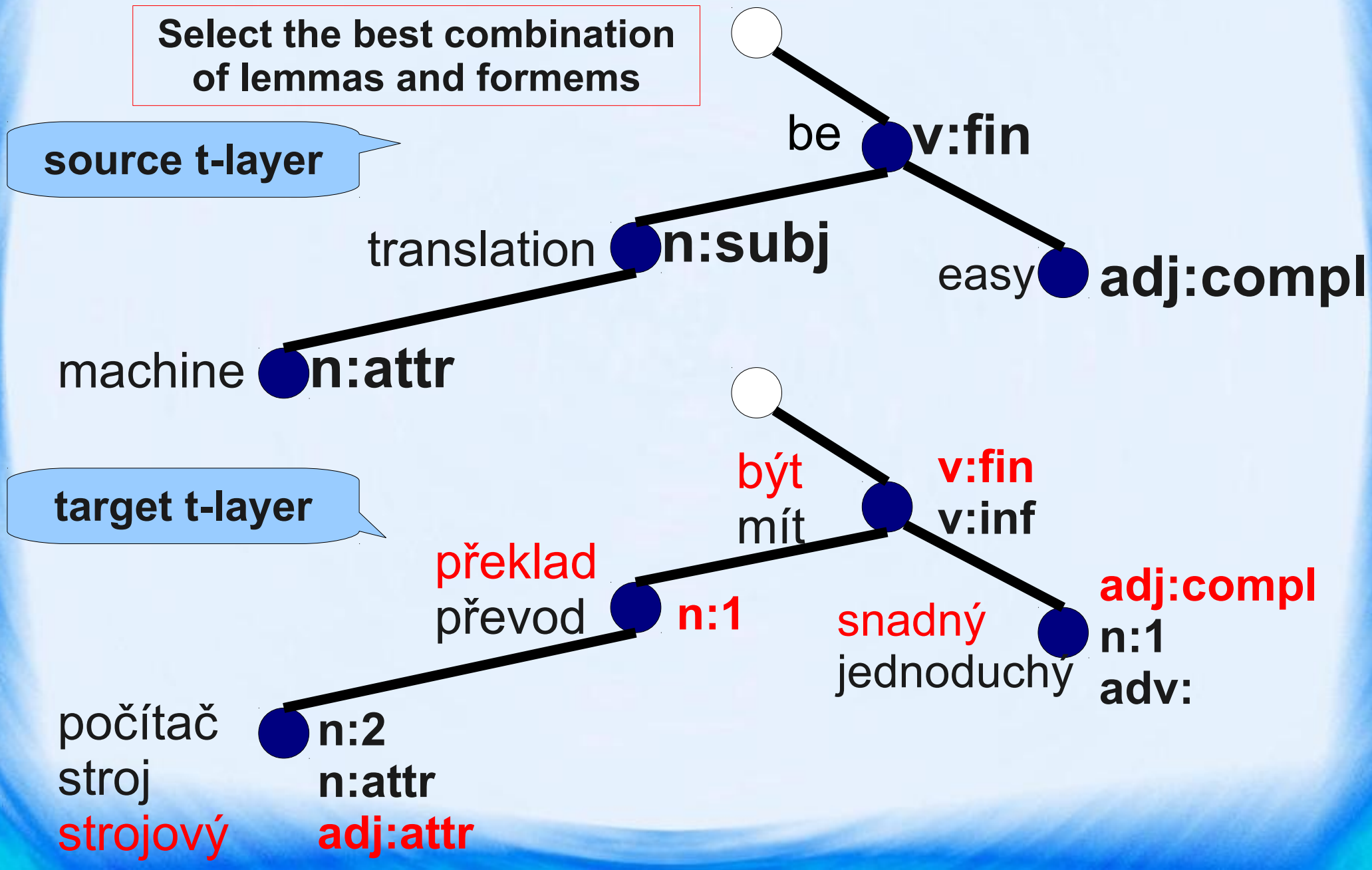


Demo Translation – Transfer

Get translation variants for lemmas and formems

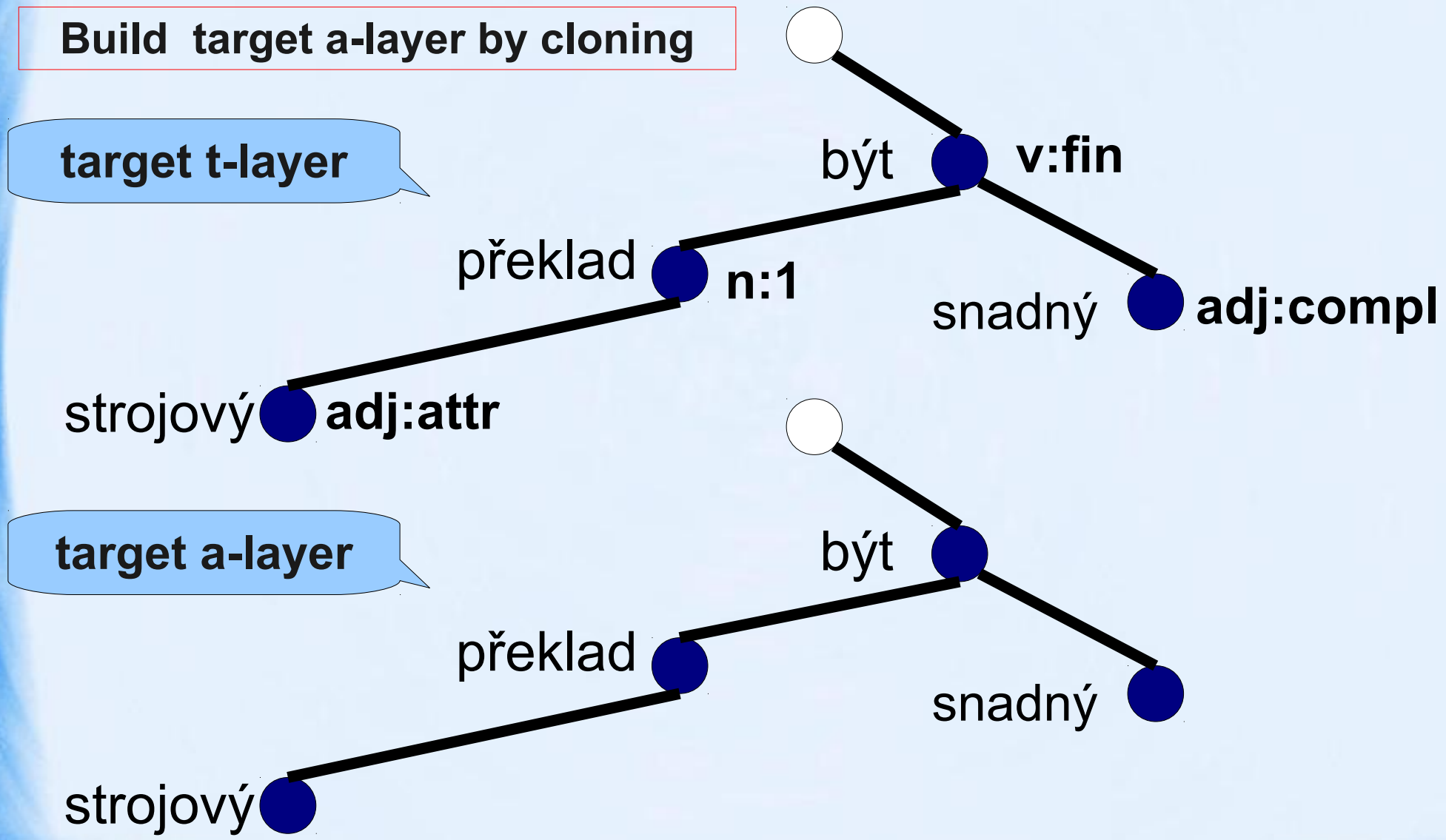


Demo Translation – Transfer



Demo Translation – Synthesis

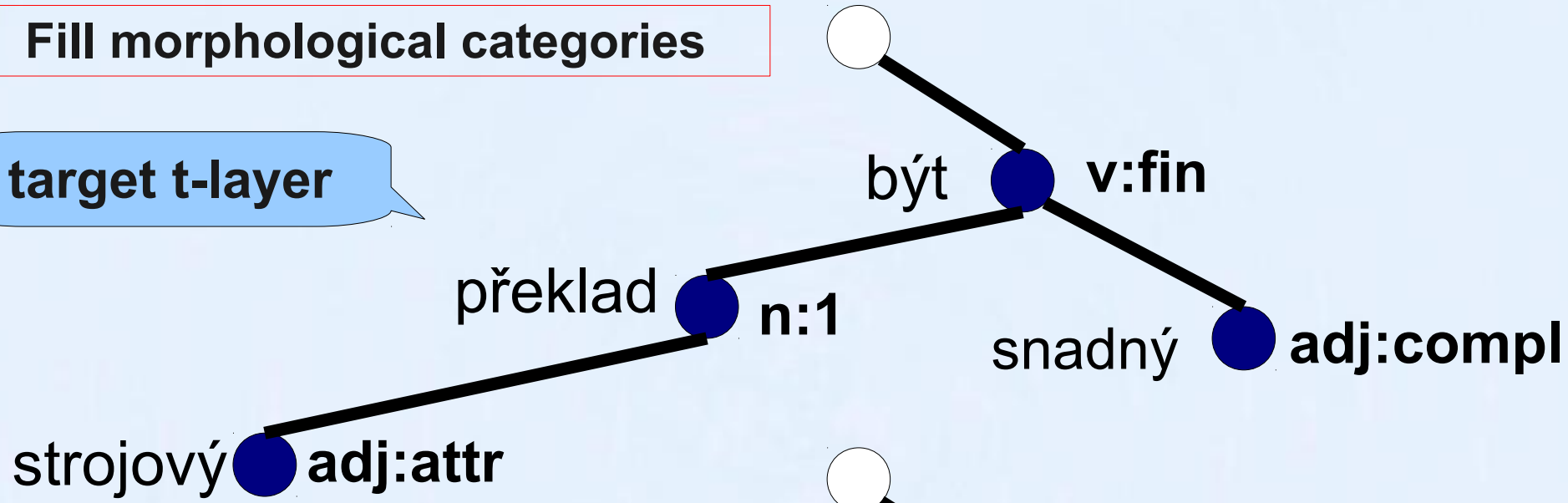
Build target a-layer by cloning



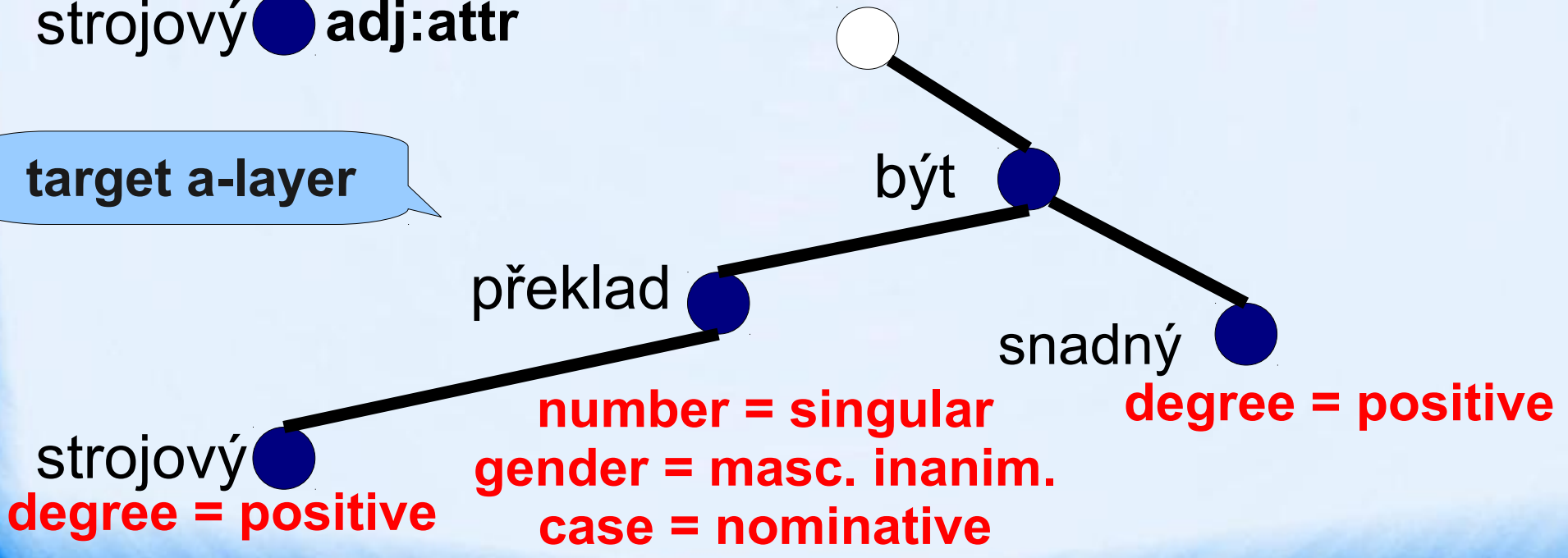
Demo Translation – Synthesis

Fill morphological categories

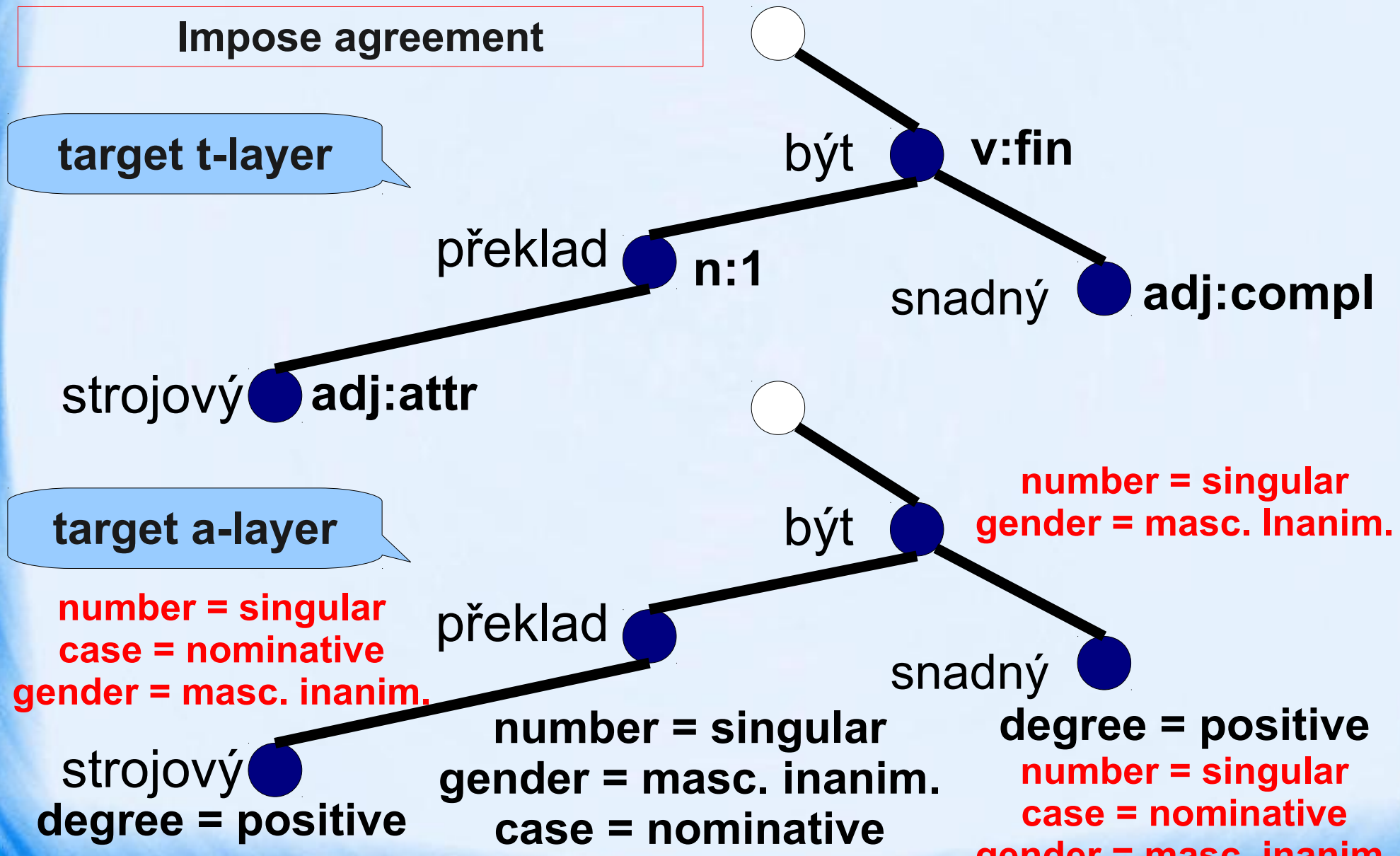
target t-layer



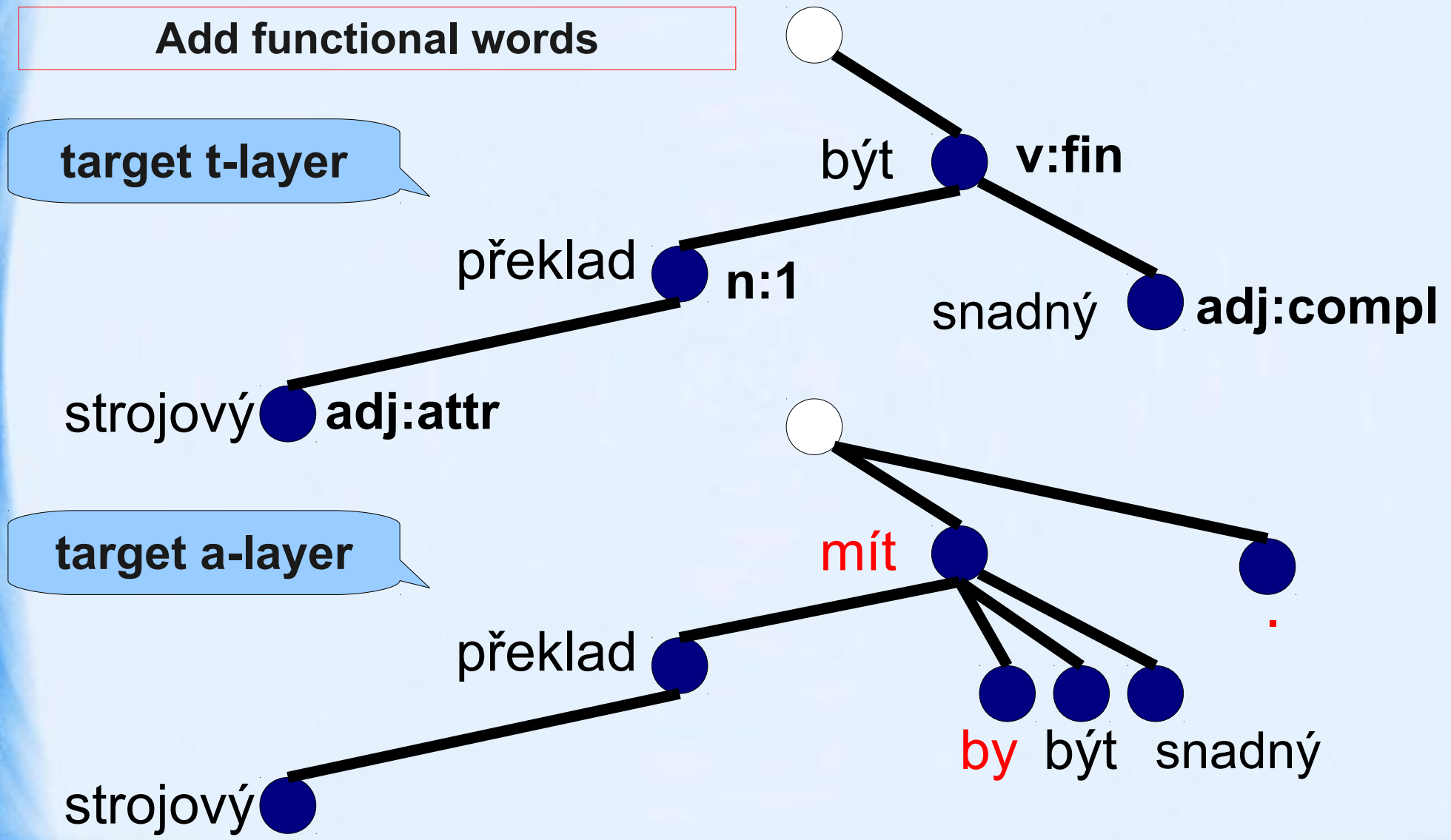
target a-layer



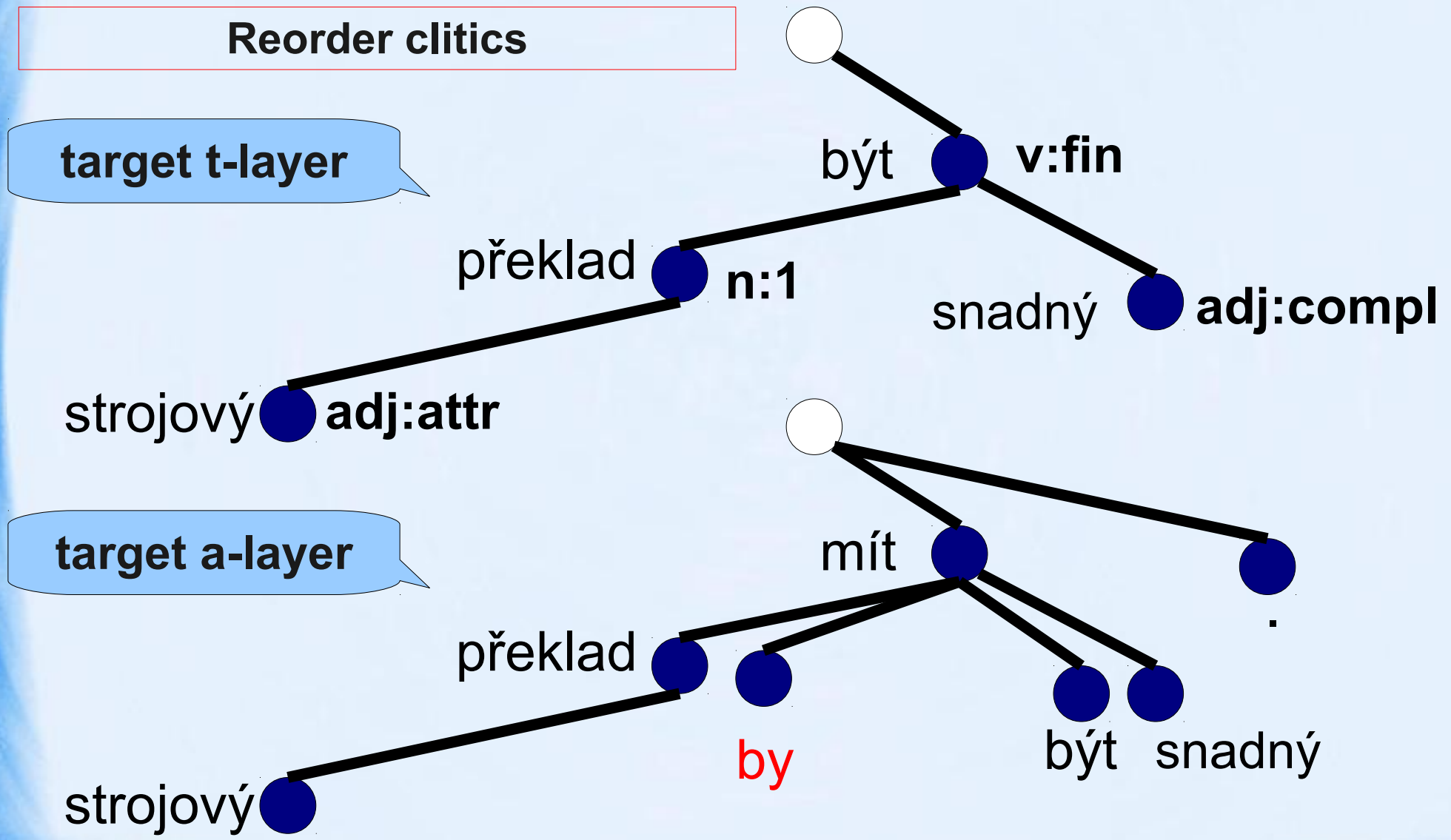
Demo Translation – Synthesis



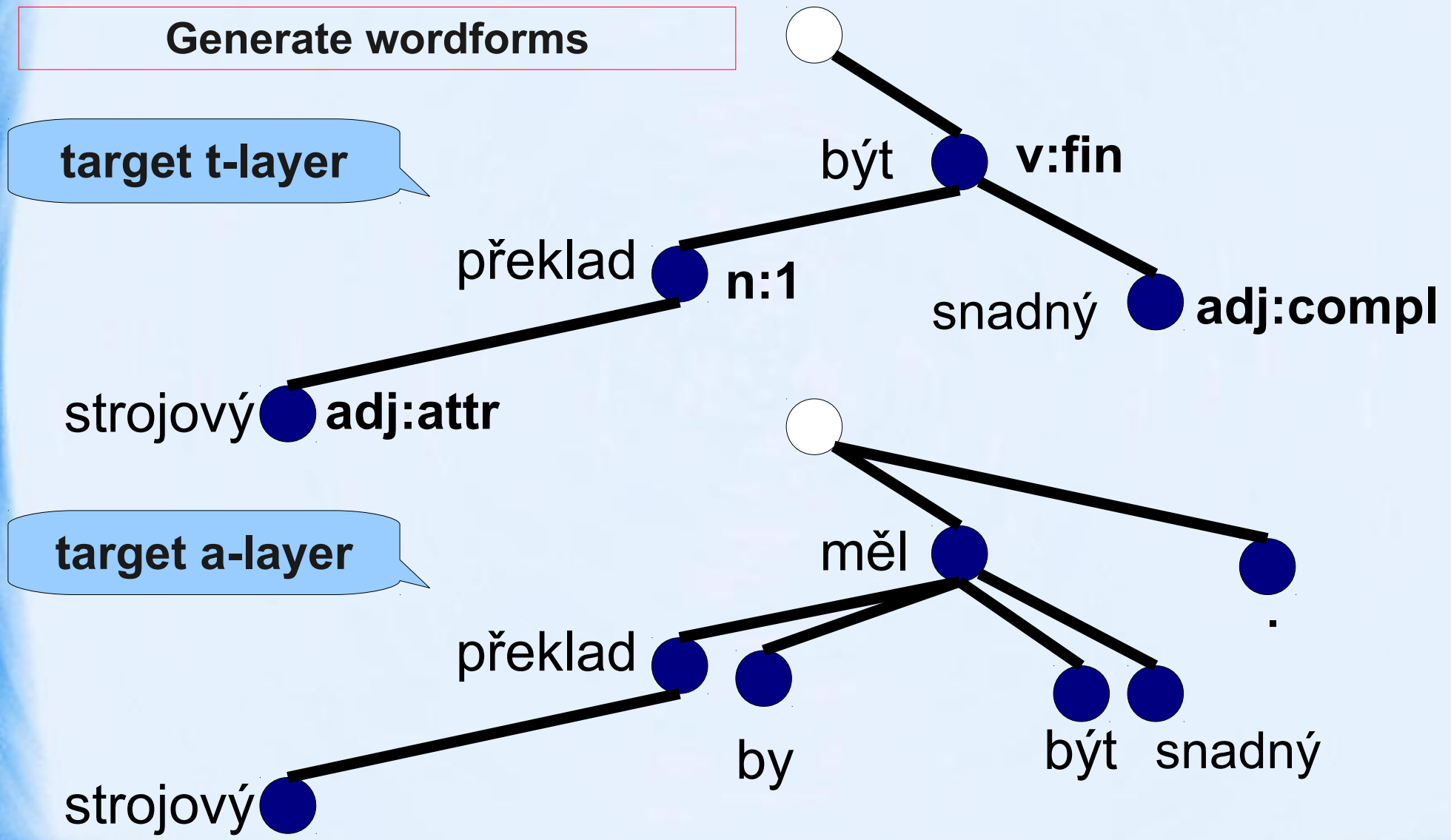
Demo Translation – Synthesis



Demo Translation – Synthesis



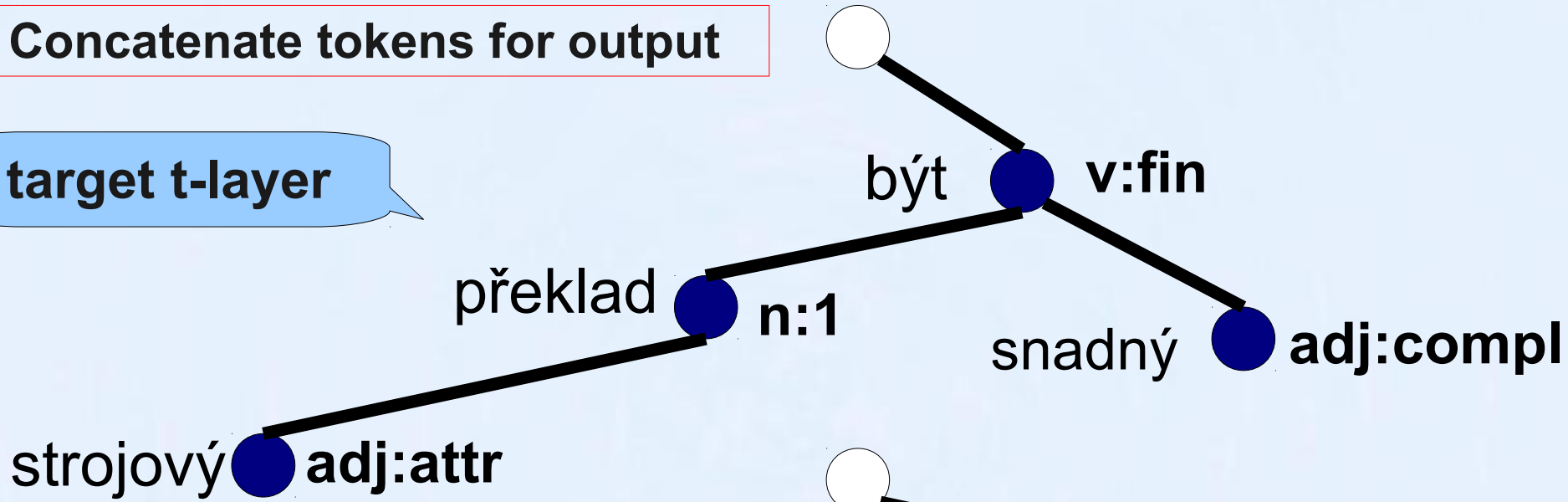
Demo Translation – Synthesis



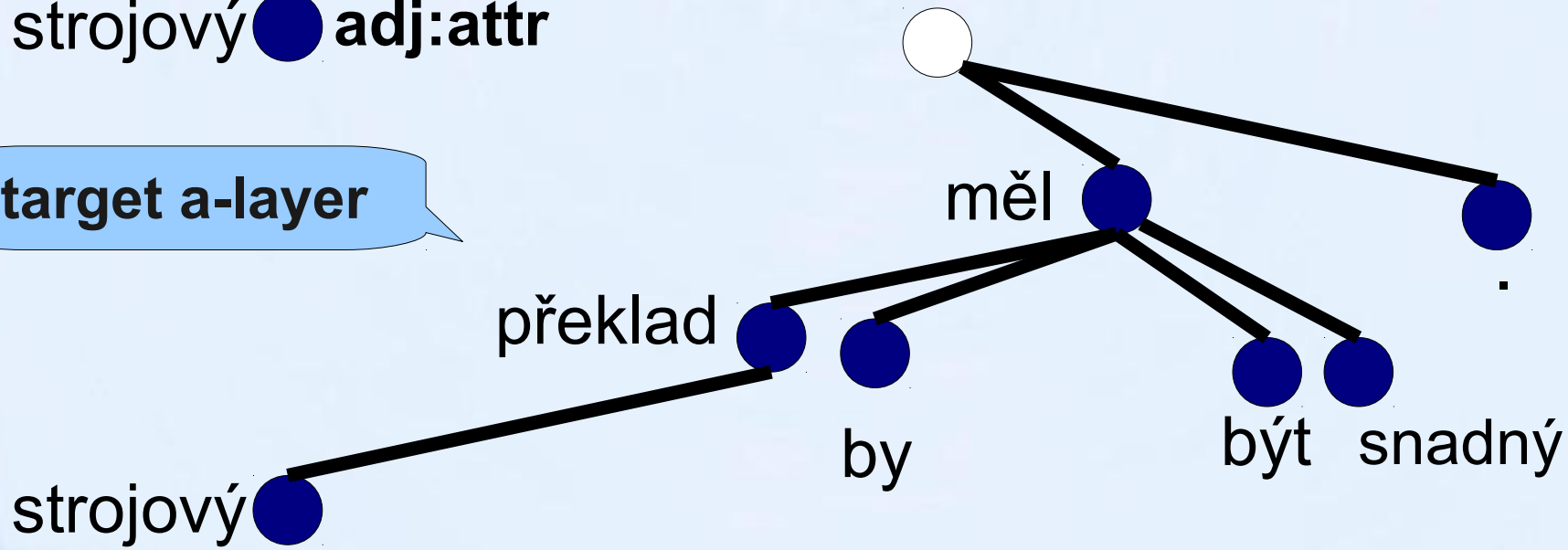
Demo Translation – Synthesis

Concatenate tokens for output

target t-layer



target a-layer

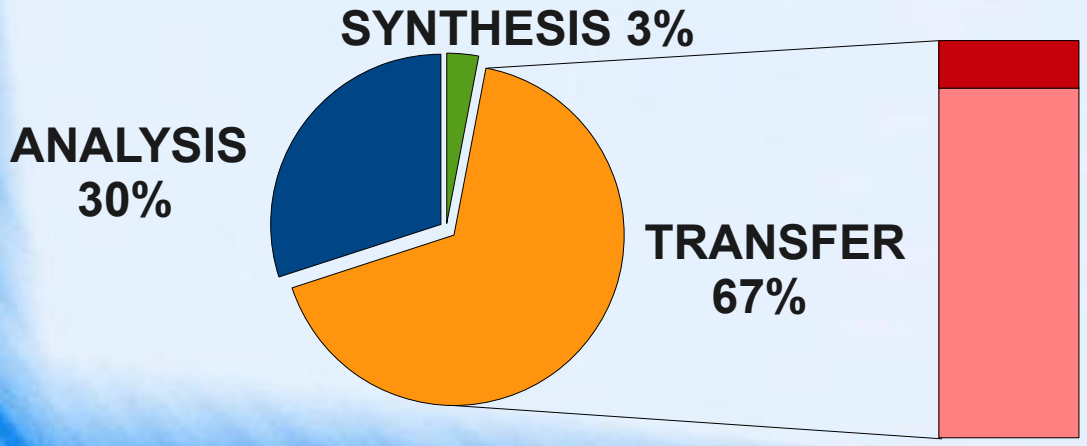


Strojový překlad by měl být snadný.

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, ?



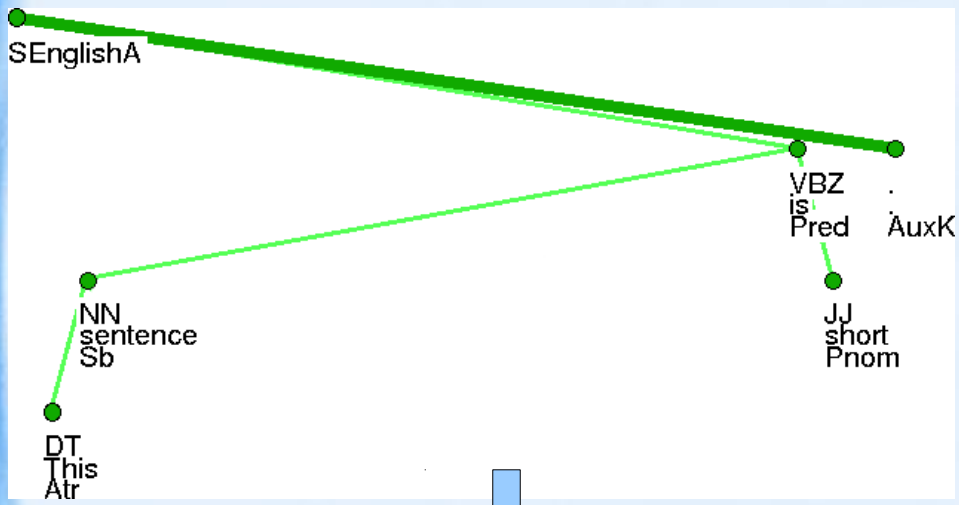
errors caused by the assumption of t-tree isomorphism 8%

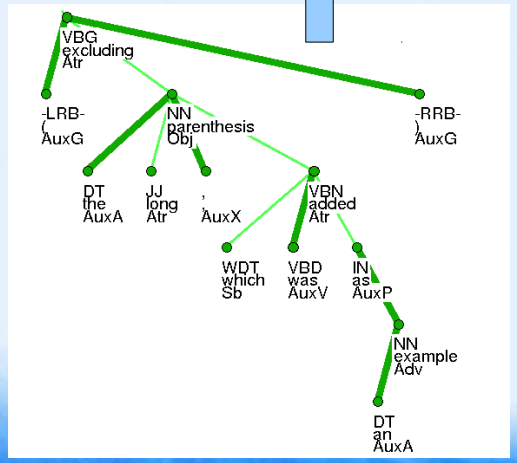
other transfer errors 59%

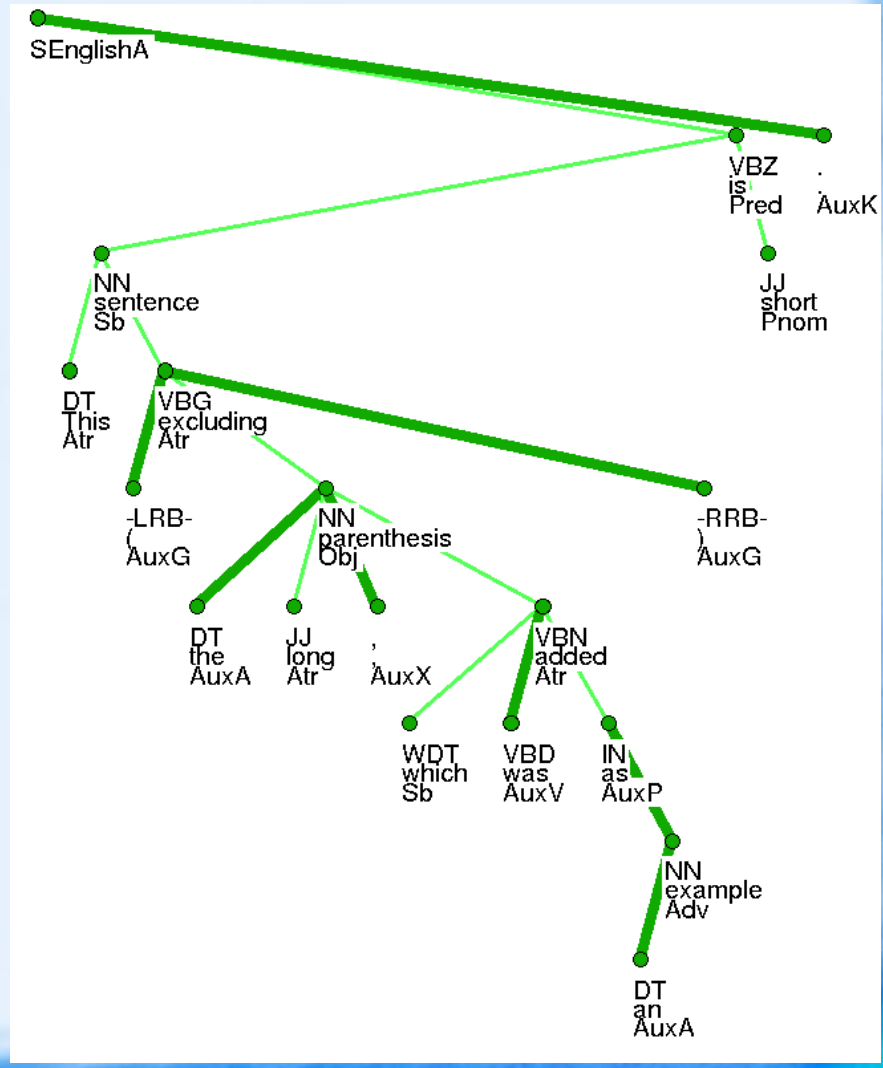
For details see [Popel,2009].

Parsing Parentheses

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

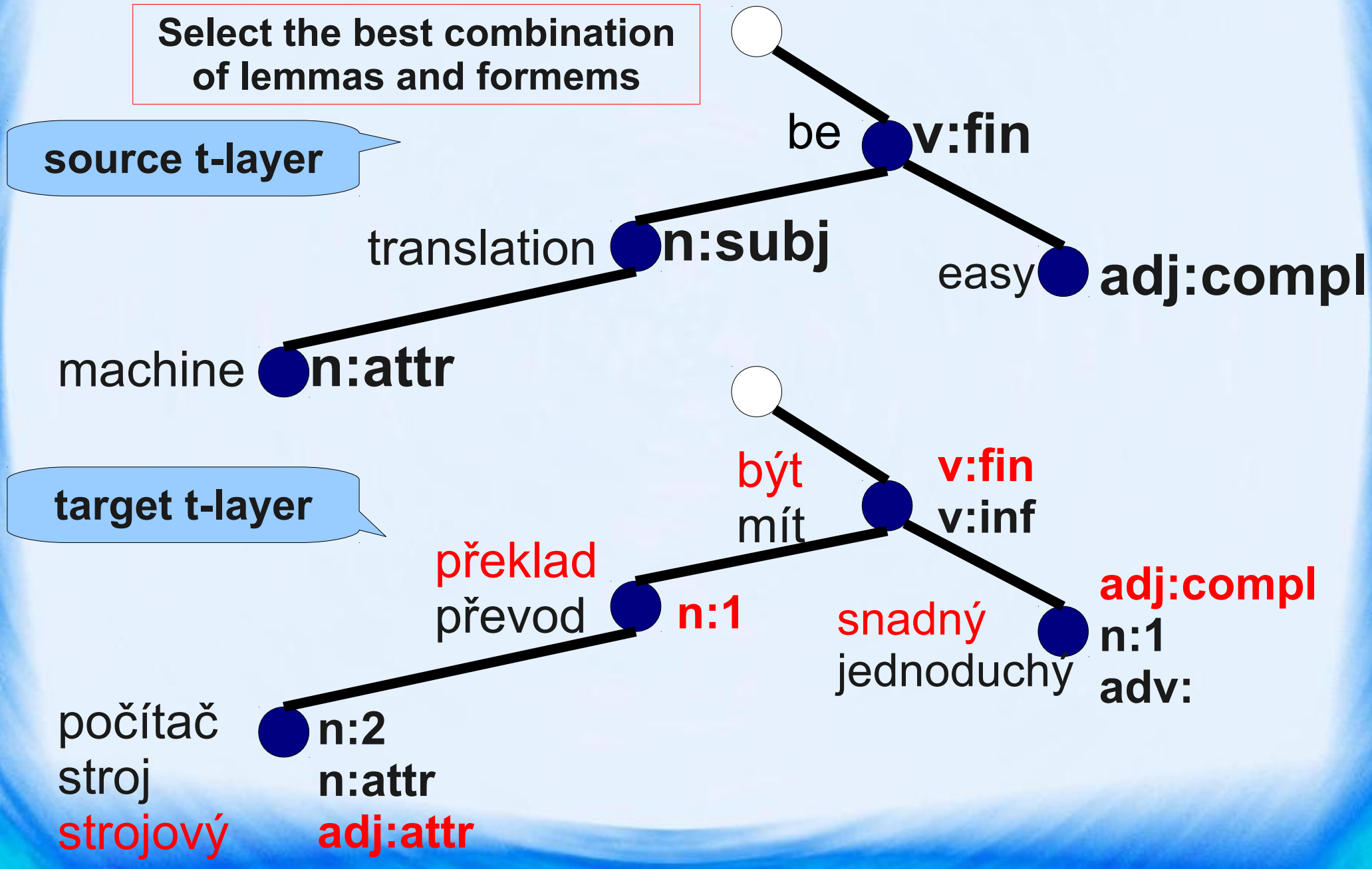






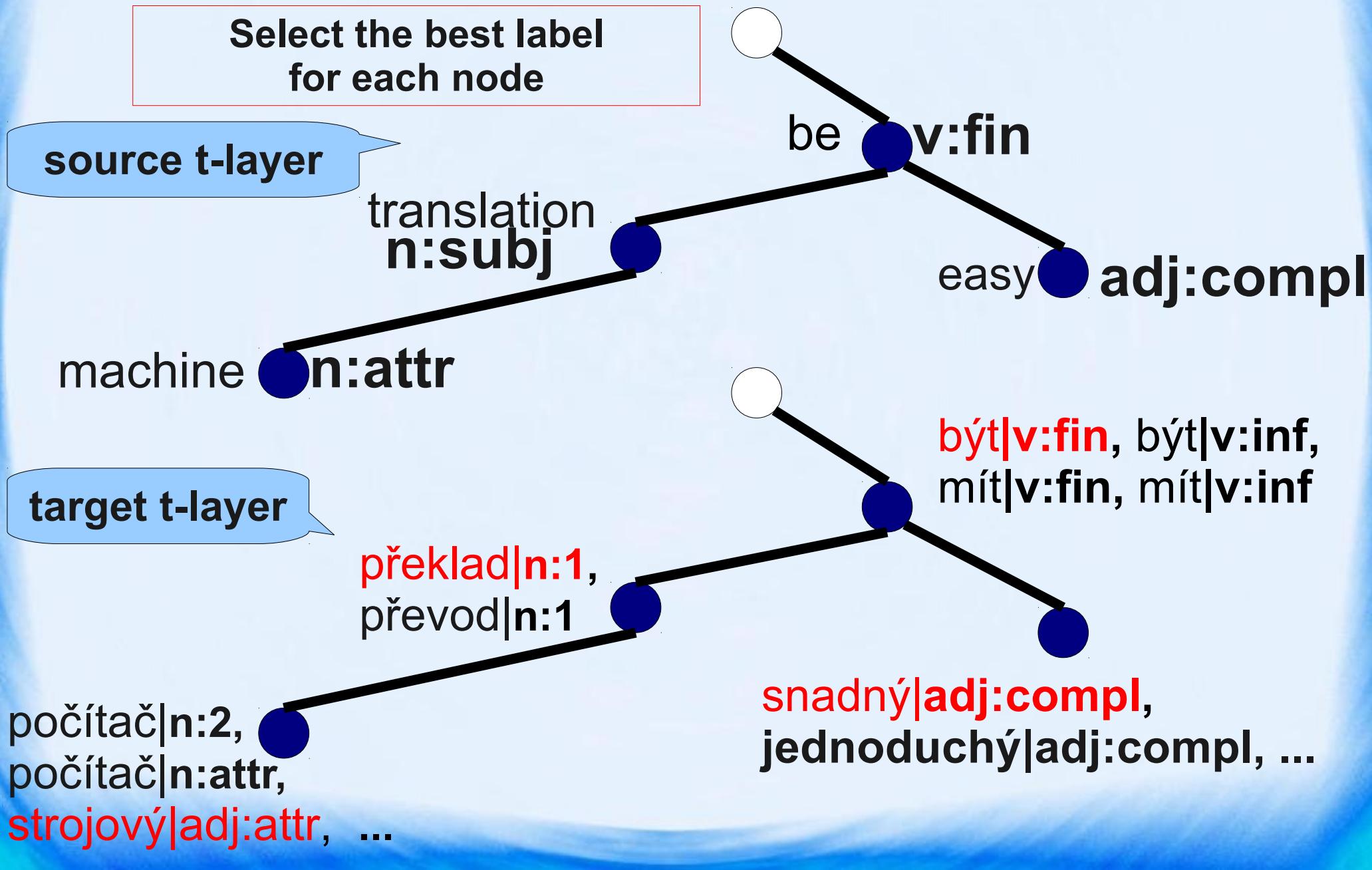
HMTM – Motivation

Select the best combination of lemmas and formems



HMTM – Motivation

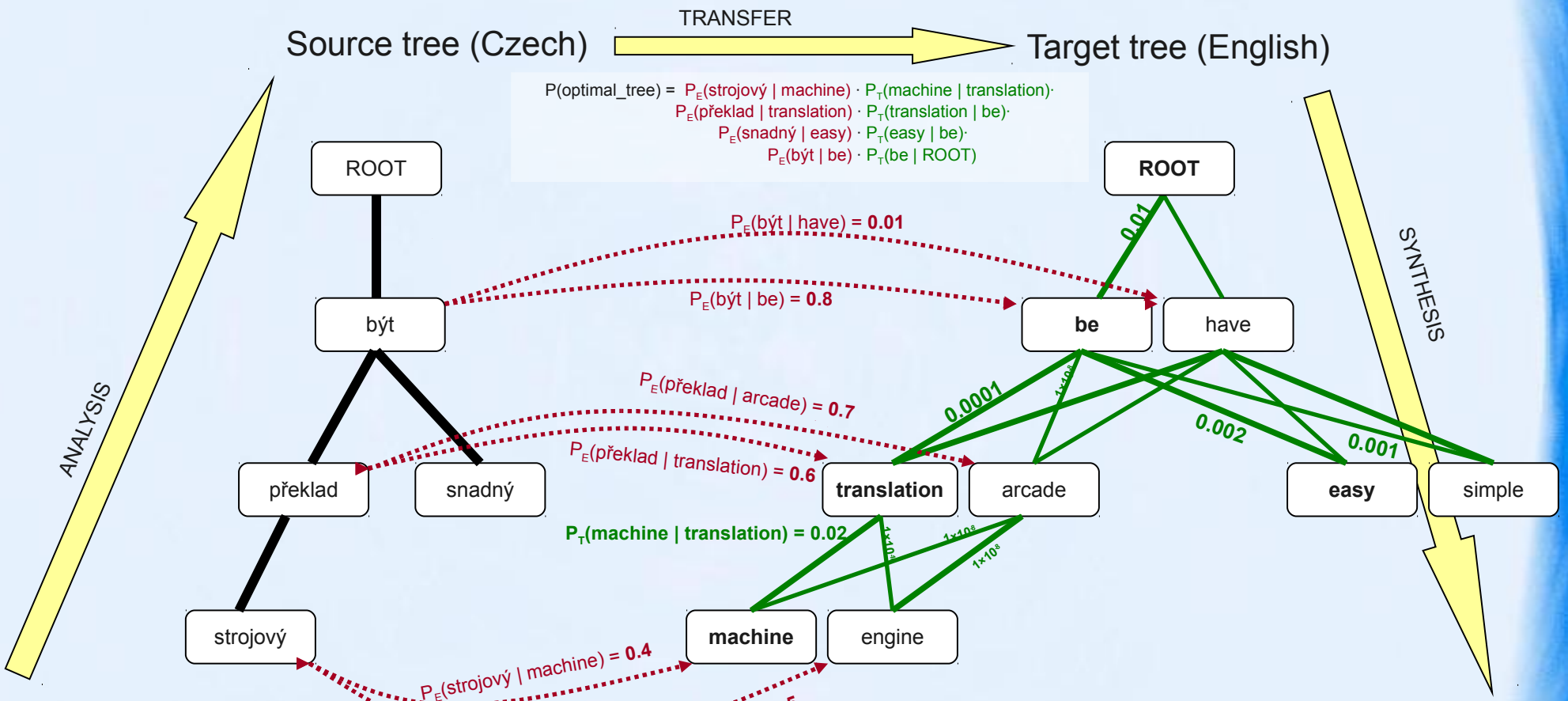
Select the best label for each node



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)
- $P(X_v \mid X_{parent(v)})$ – transition probabilities
- $P(Y_v \mid X_v)$ – emission probabilities
- Tree-Markov property:
  $\forall v \in V \setminus \{root\}, \forall w \in V \setminus subtree(v)$:
  $P(\mathbf{X}_{subtree\ (v)} \mid X_{parent(v)}, X_w) = P(\mathbf{X}_{subtree\ (v)} \mid X_{parent(v)})$
  i.e. given $X_{parent(v)}$, all hidden states of the subtree rooted in v are conditionally independent of any other nodes.
- The most probable hidden tree labeling $\mathbf{X}^*$ can be obtained given the observed tree labeling **Y** can be obtained using tree-modified Viterbi algorithm.

HMTM in MT



Source sentence:
Strojový překlad by měl být snadný.

Target sentence:
Machine translation should be easy.

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

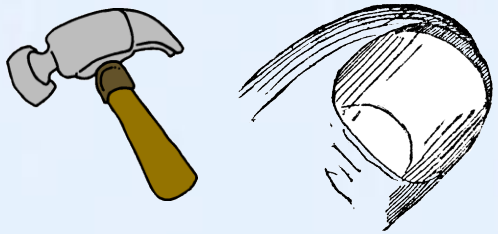
Examples of Translation

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

Slečna palec je slečna miliónu.

I'd rather be a hammer
than a nail.

Spíše bych byl kladivo než nehet.



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

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.