English-Czech Machine Translation Using TectoMT

Martin Popel
ÚFAL, Charles University in Prague

19th Week of Doctoral Students, June 3, 2010

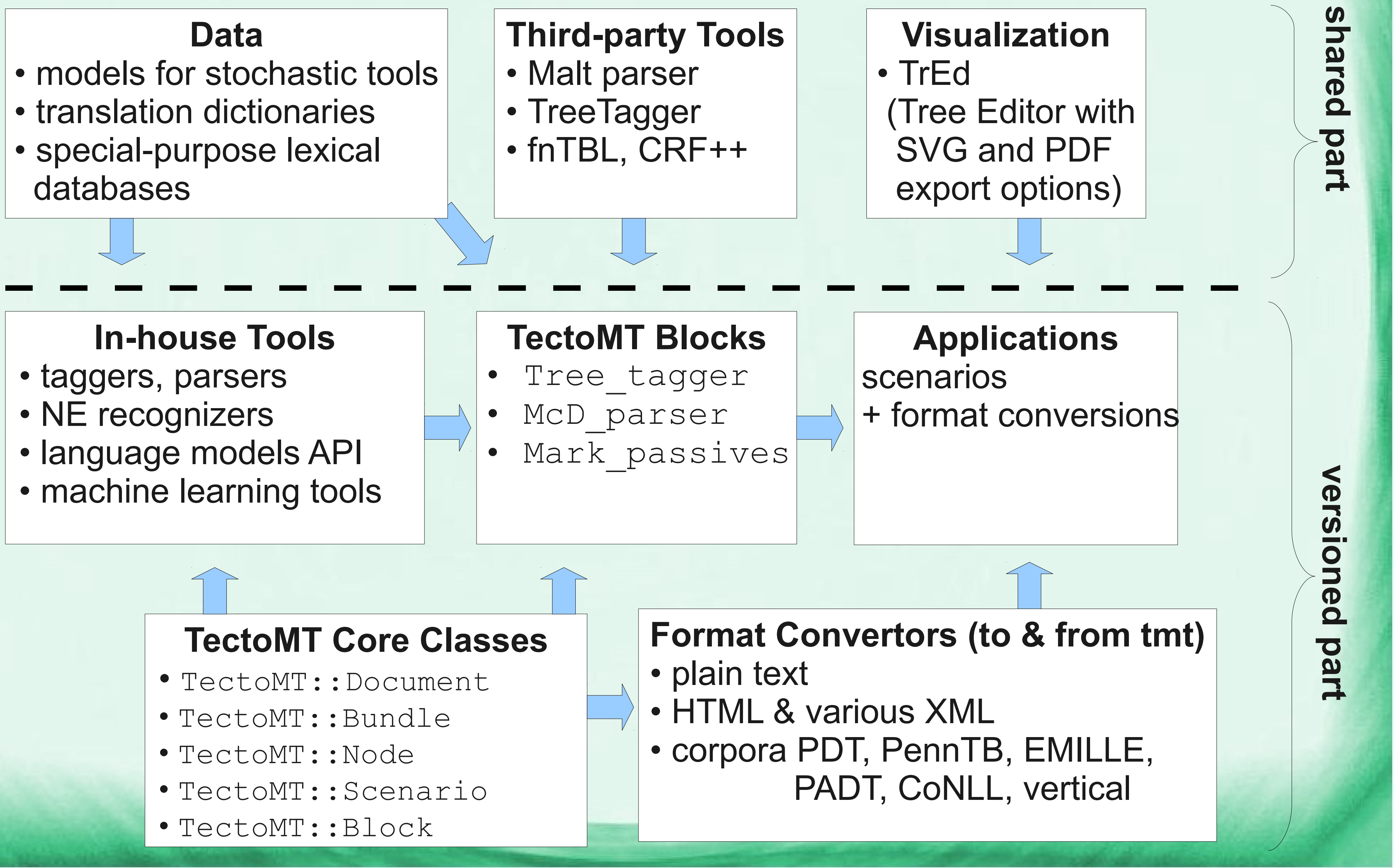
Outline

- TectoMT – NLP framework and MT system
- Demo translation step by step
- Hidden Markov Tree Models (HMTM)
- Results – three metrics of translation quality

TectoMT – framework for NLP

TectoMT

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



Data
- models for stochastic tools
- translation dictionaries
- special-purpose lexical databases

Third-party Tools
- Malt parser
- TreeTagger
- fnTBL, CRF++

Visualization
- TrEd (Tree Editor with SVG and PDF export options)

shared part

In-house Tools
- taggers, parsers
- NE recognizers
- language models API
- machine learning tools

TectoMT Blocks
- `Tree_tagger`
- `McD_parser`
- `Mark_passives`

Applications
scenarios
+ format conversions

TectoMT Core Classes
- `TectoMT::Document`
- `TectoMT::Bundle`
- `TectoMT::Node`
- `TectoMT::Scenario`
- `TectoMT::Block`

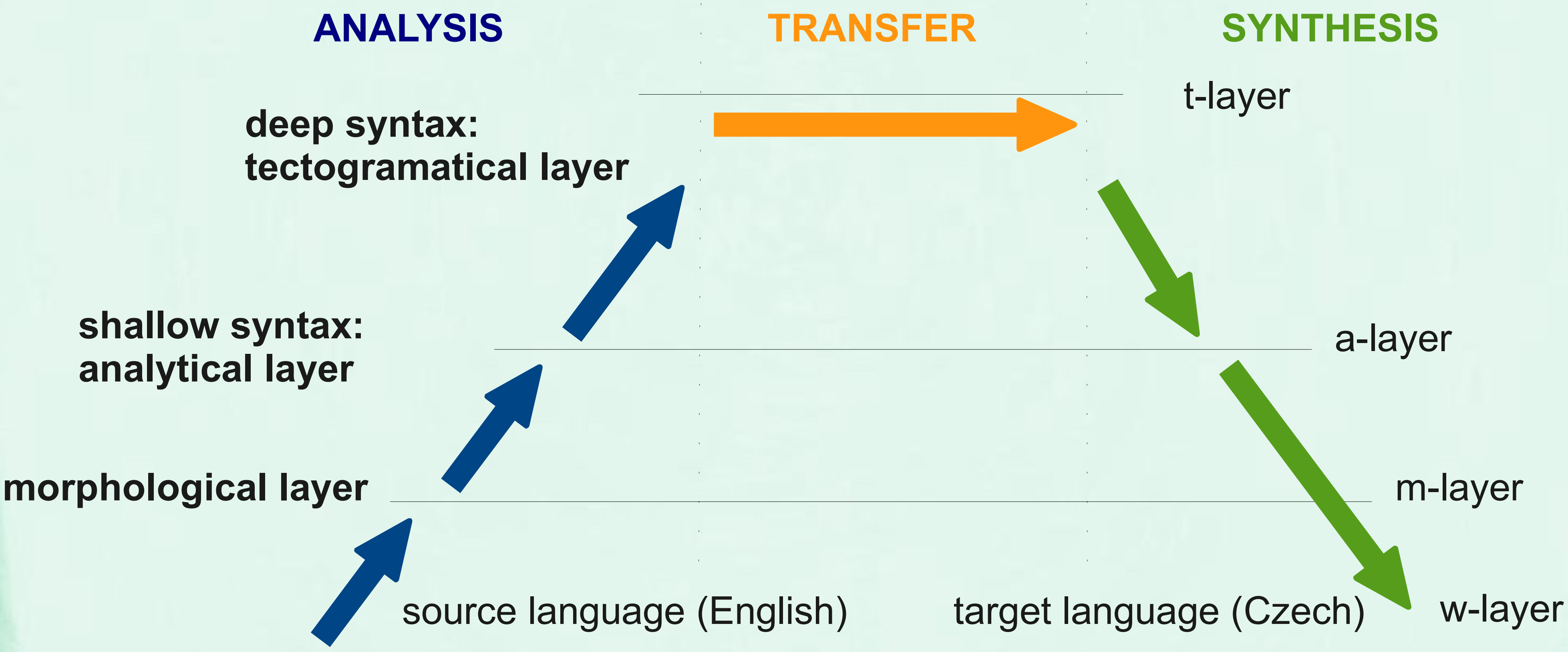
Format Convertors (to & from tmt)
- plain text
- HTML & various XML
- corpora PDT, PennTB, EMILLE, PADT, CoNLL, vertical

versioned part

TectoMT – MT system

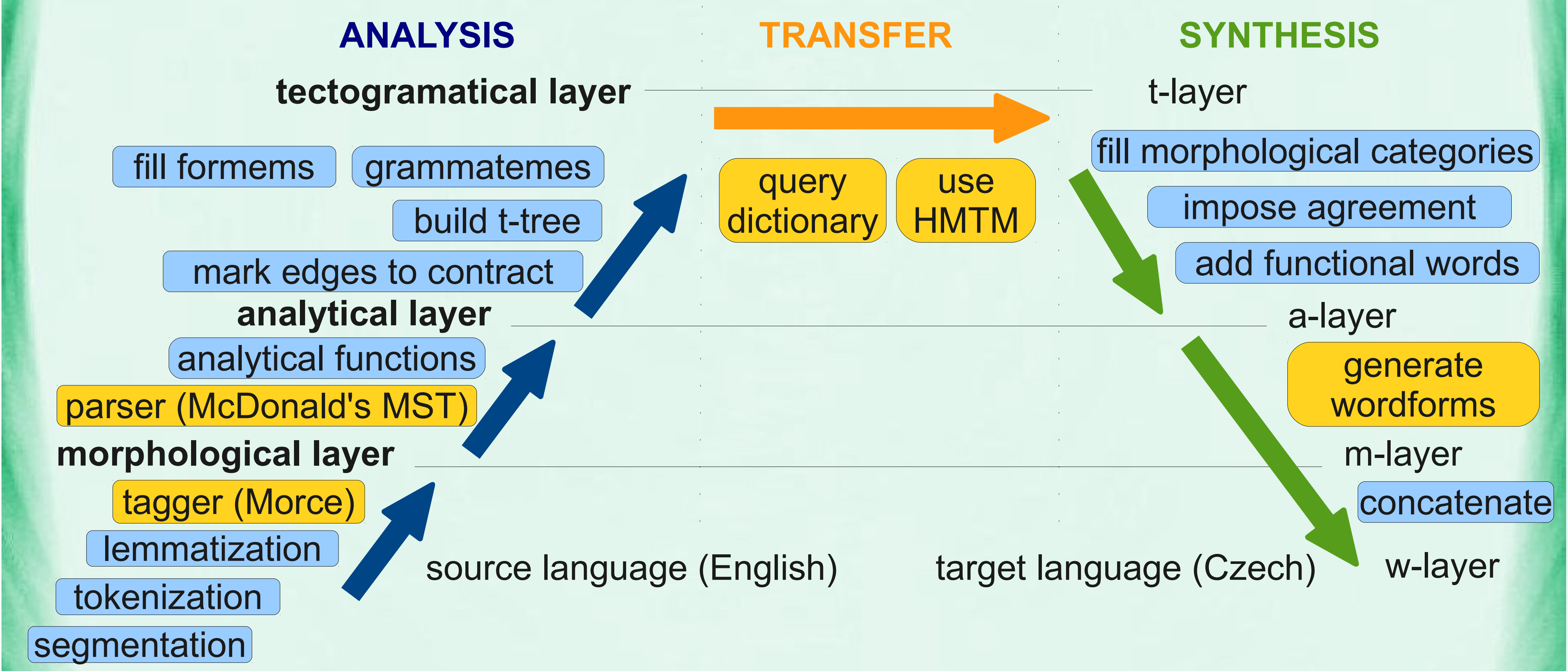
TectoMT

transfer over the tectogrammatical layer



TectoMT – MT system

rule based & statistical blocks



ANALYSIS

tectogramatical layer

fill formems

grammatemes

build t-tree

mark edges to contract

analytical layer

analytical functions

parser (McDonald's MST)

morphological layer

tagger (Morce)

lemmatization

tokenization

segmentation

source language (English)

TRANSFER

query dictionary

use HMTM

SYNTHESIS

t-layer

fill morphological categories

impose agreement

add functional words

a-layer

generate wordforms

m-layer

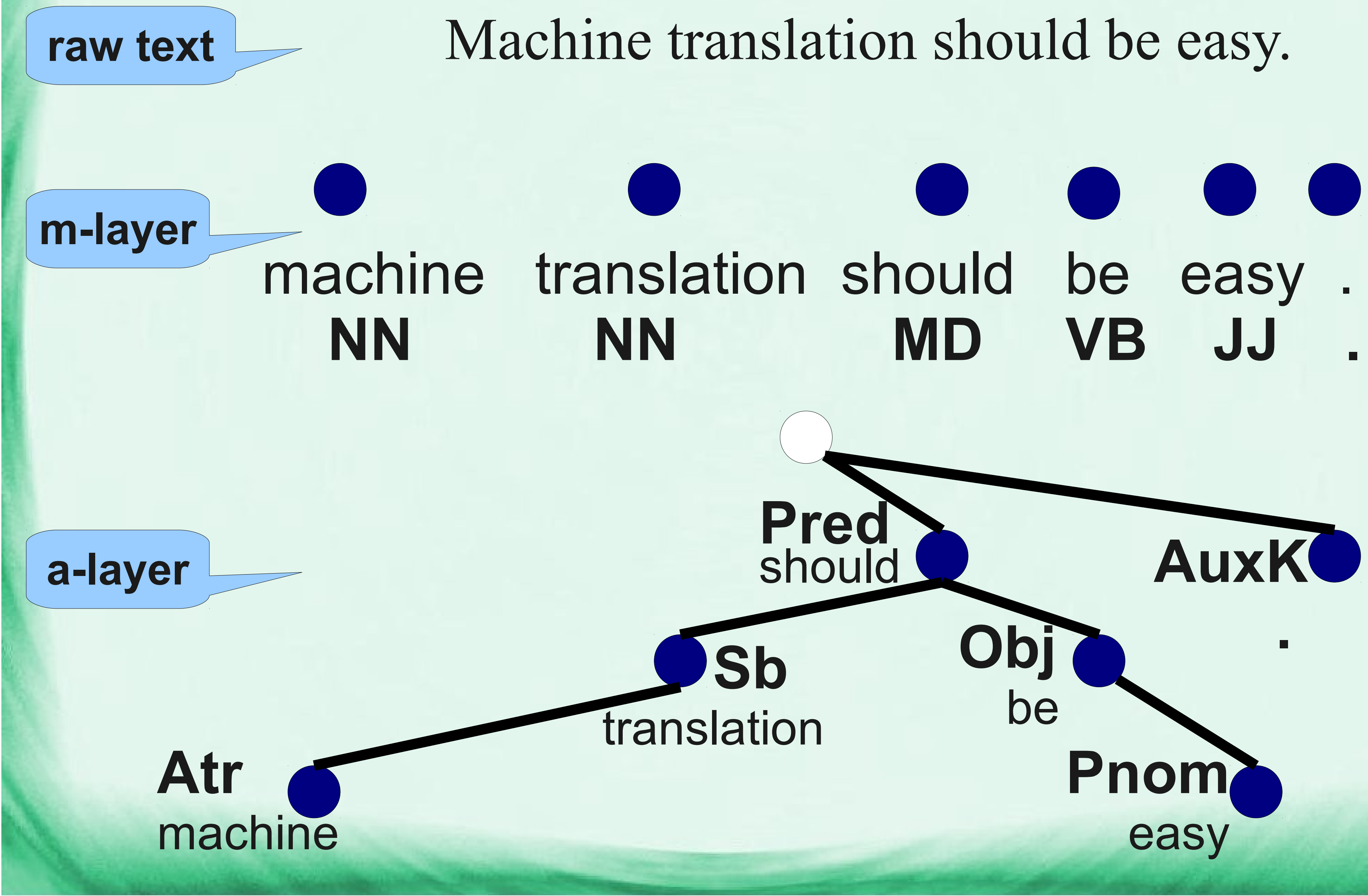
concatenate

w-layer

target language (Czech)

Demo Translation – Analysis

TectoMT



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

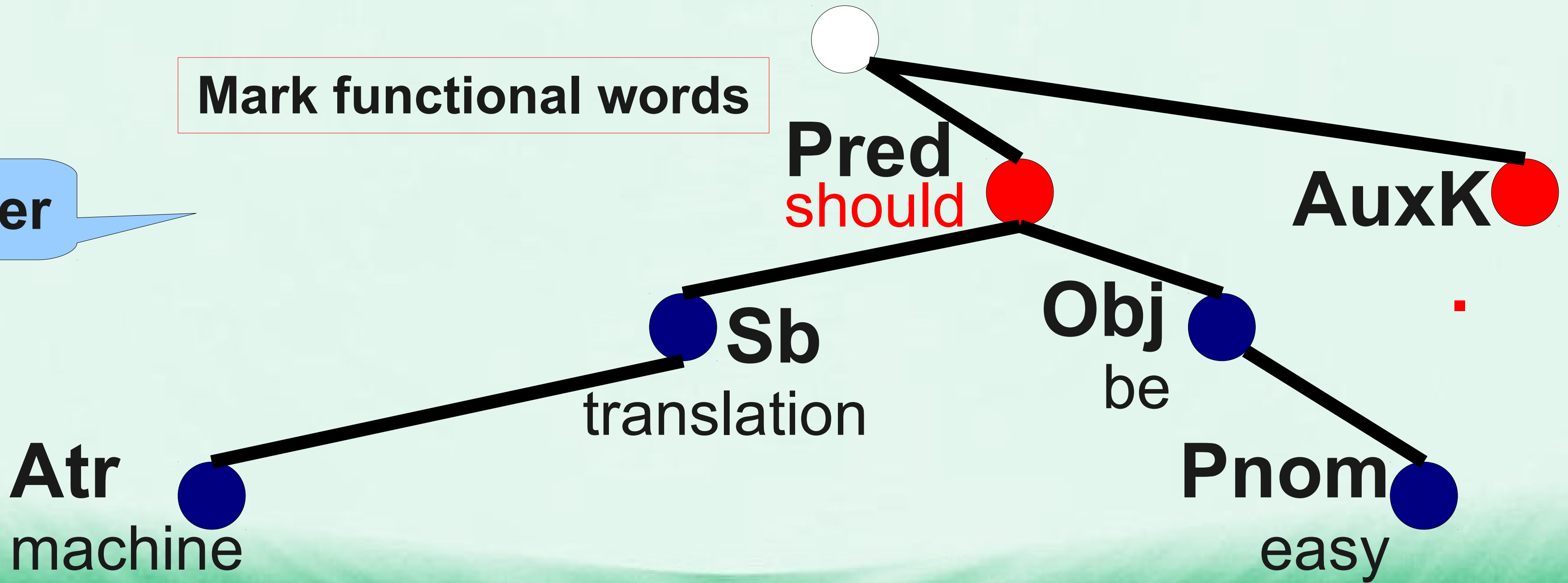
Machine translation should be easy.

m-layer

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

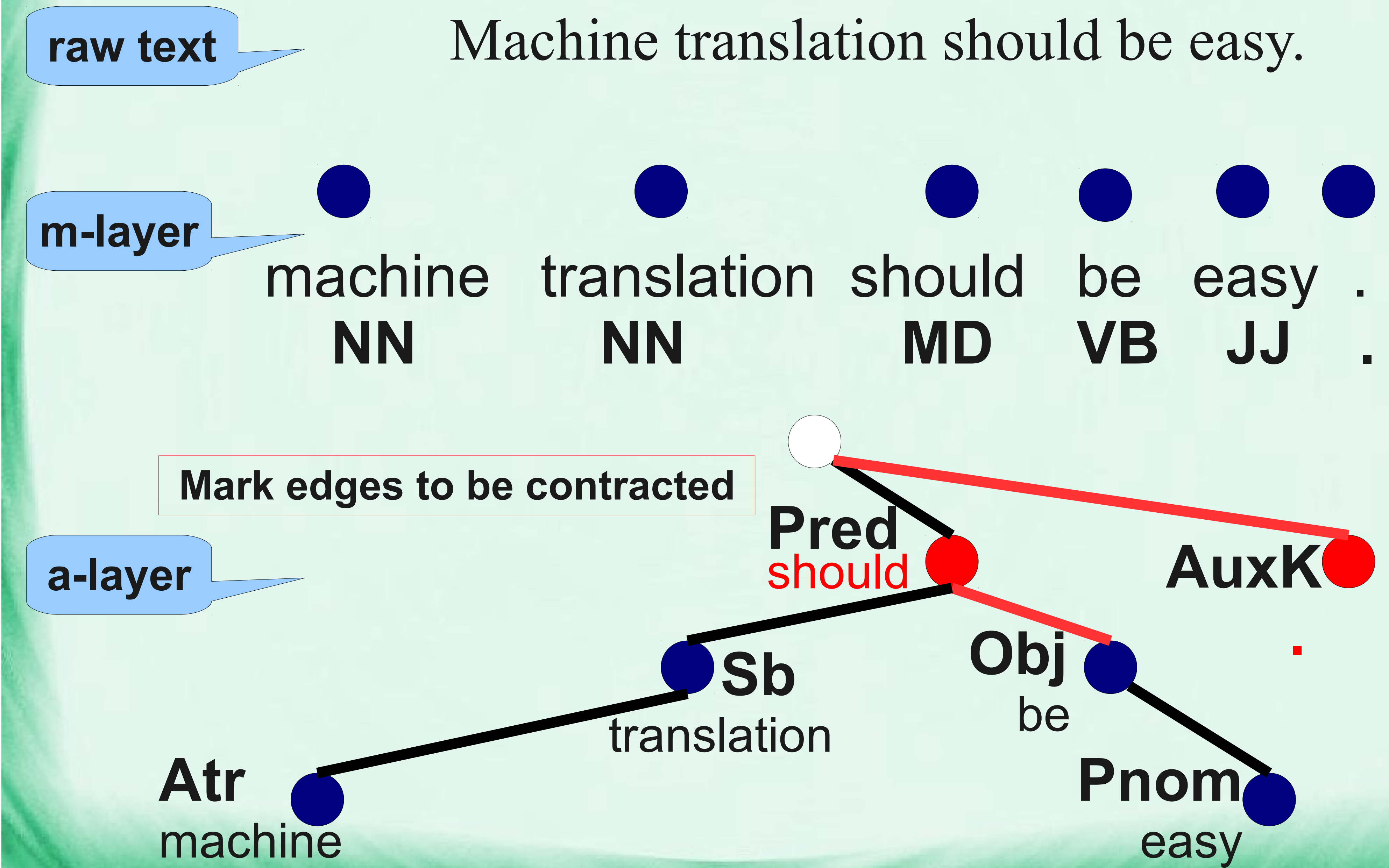
a-layer

Mark functional words



Demo Translation – Analysis

TectoMT



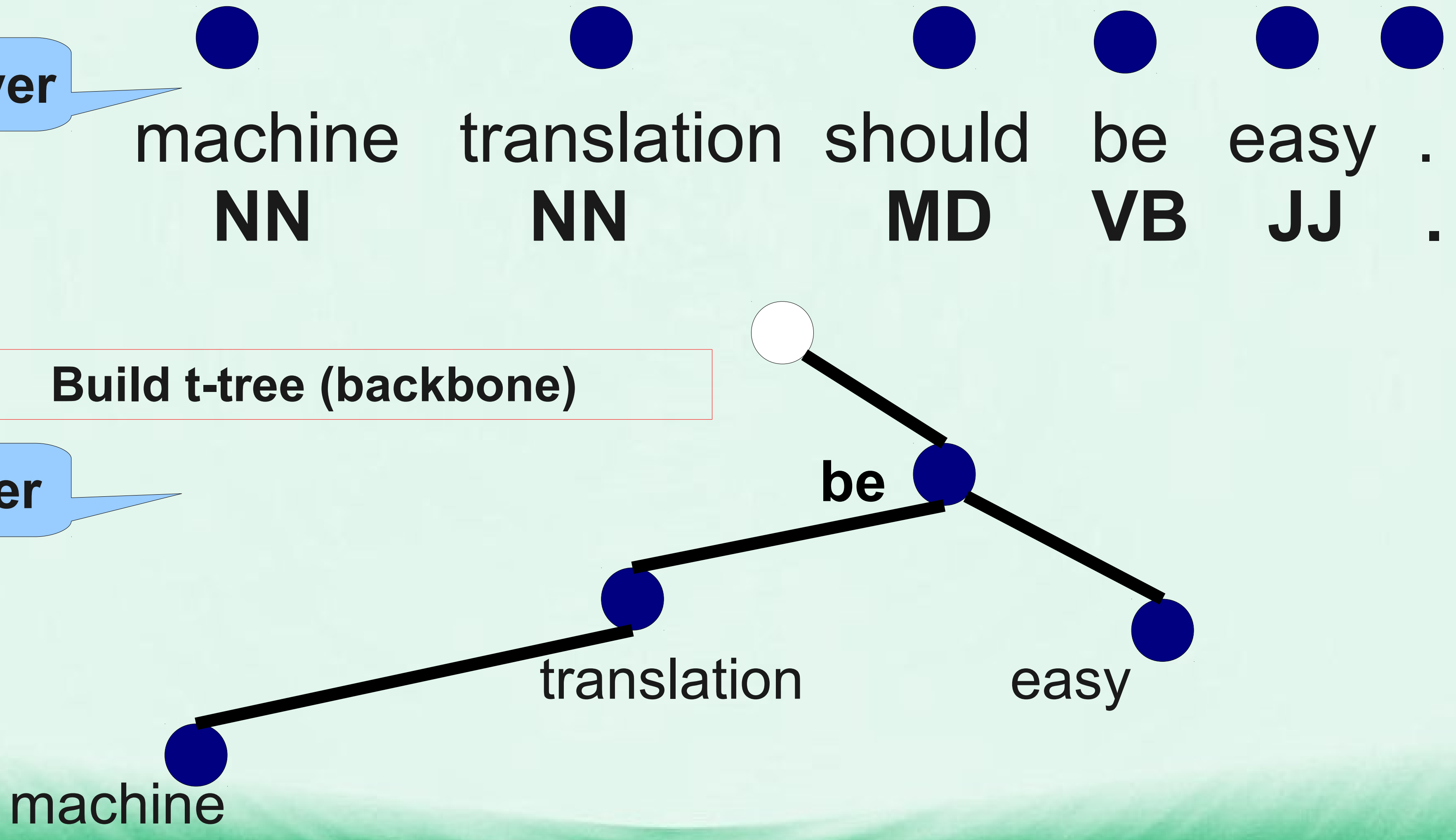
Demo Translation – Analysis

raw text

Machine translation should be easy.

m-layer

t-layer



machine NN

translation NN

should MD

be VB

easy JJ

. .

Build t-tree (backbone)

be

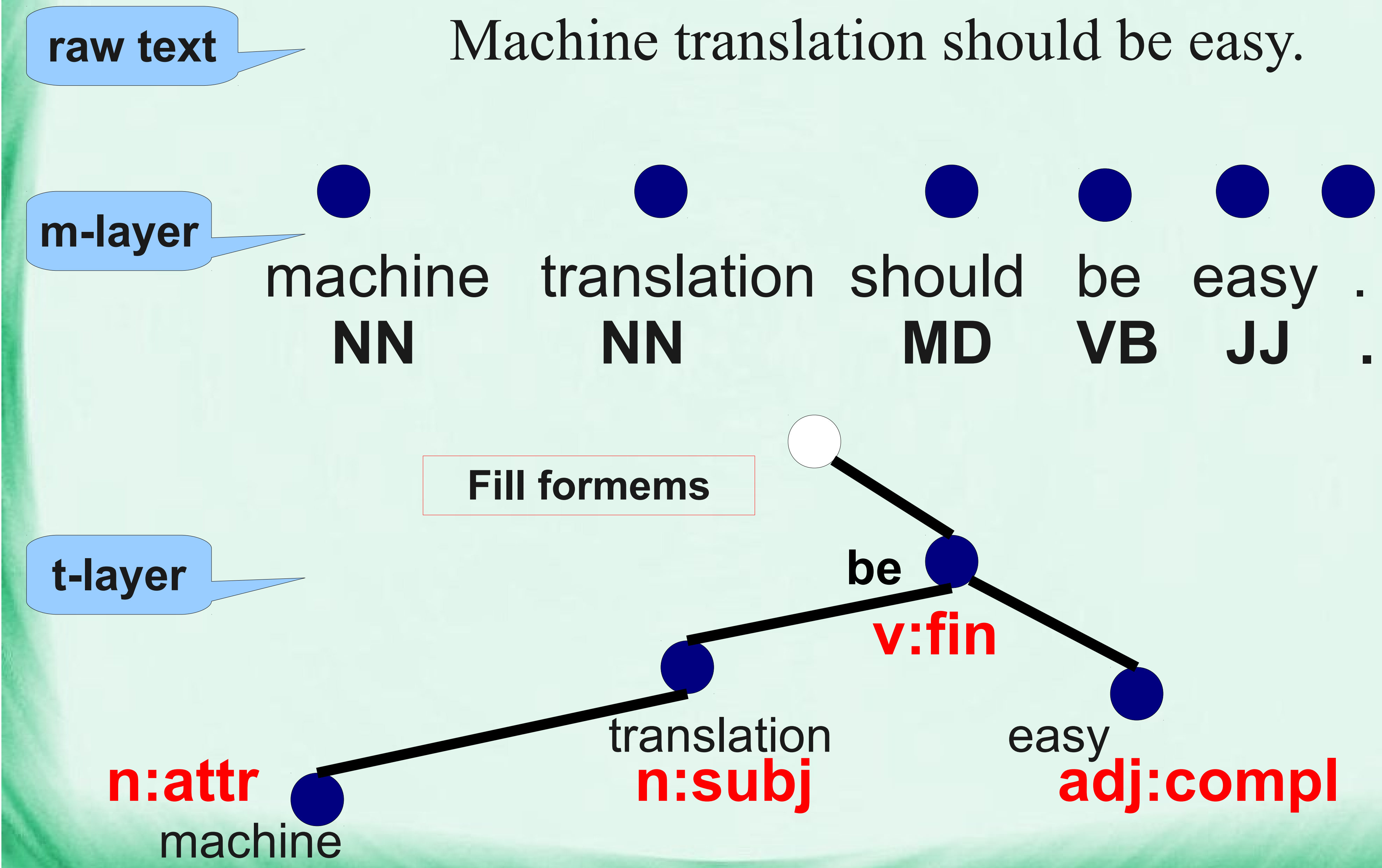
translation

easy

machine

Demo Translation – Analysis

TectoMT



raw text

Machine translation should be easy.

m-layer

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

t-layer

Fill formems

be
v:fin

translation
n:subj

easy
adj:compl

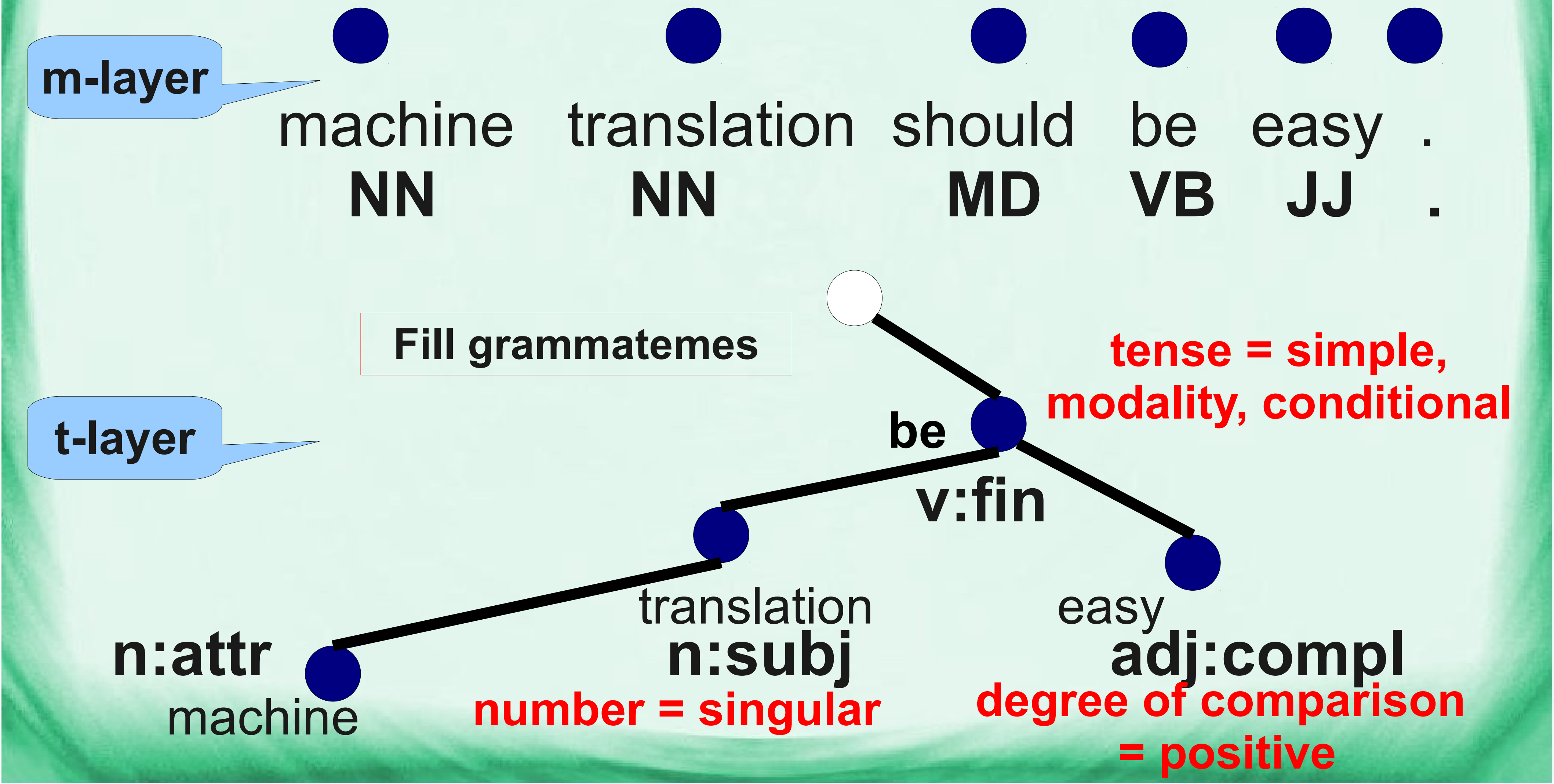
n:attr
machine

Demo Translation – Analysis

TectoMT

raw text

Machine translation should be easy.



m-layer

machine NN

translation NN

should MD

be VB

easy JJ

. .

Fill grammatemes

t-layer

be v:fin

tense = simple, modality, conditional

translation n:subj

number = singular

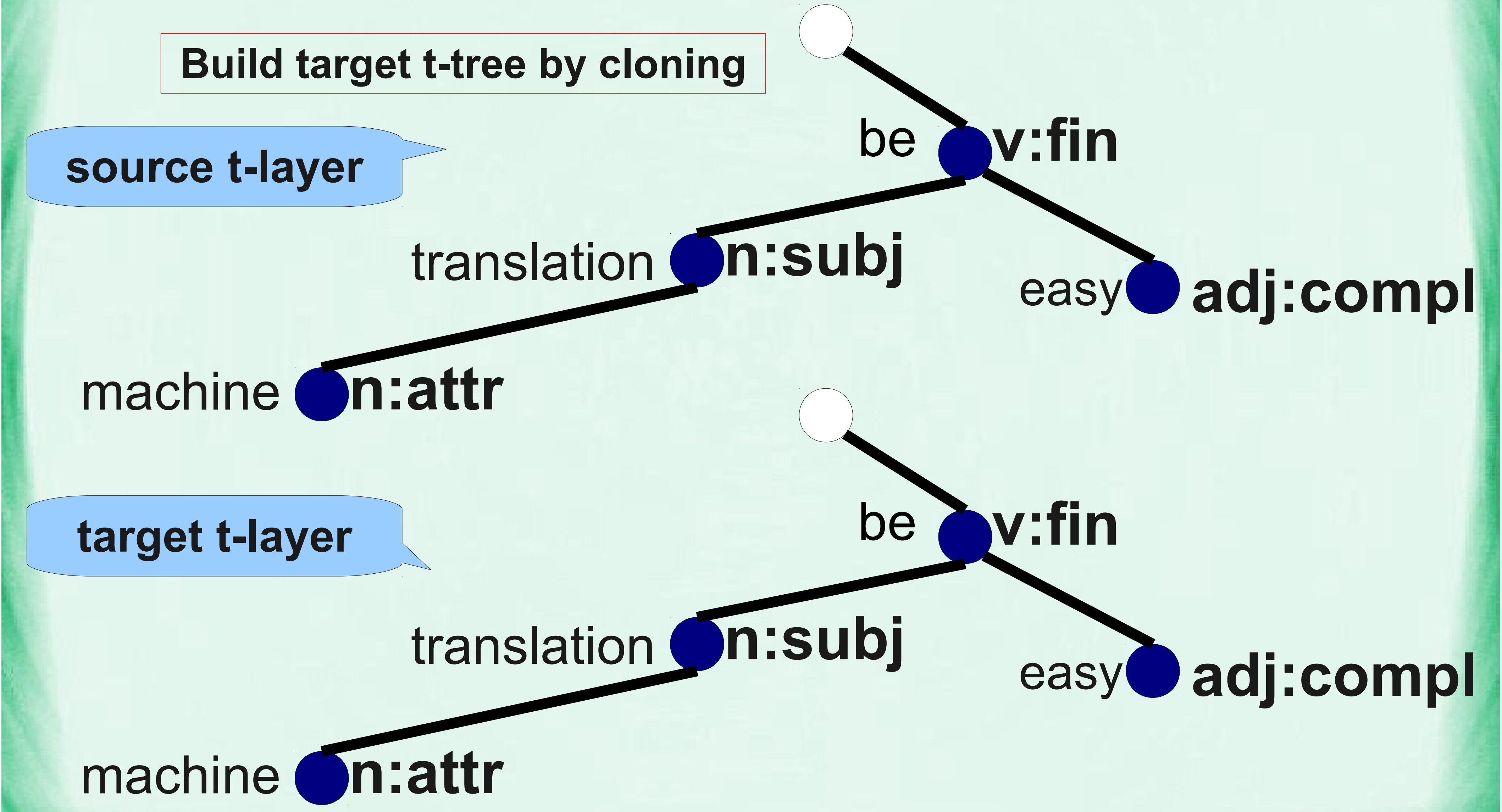
machine n:attr

easy adj:compl

degree of comparison = positive

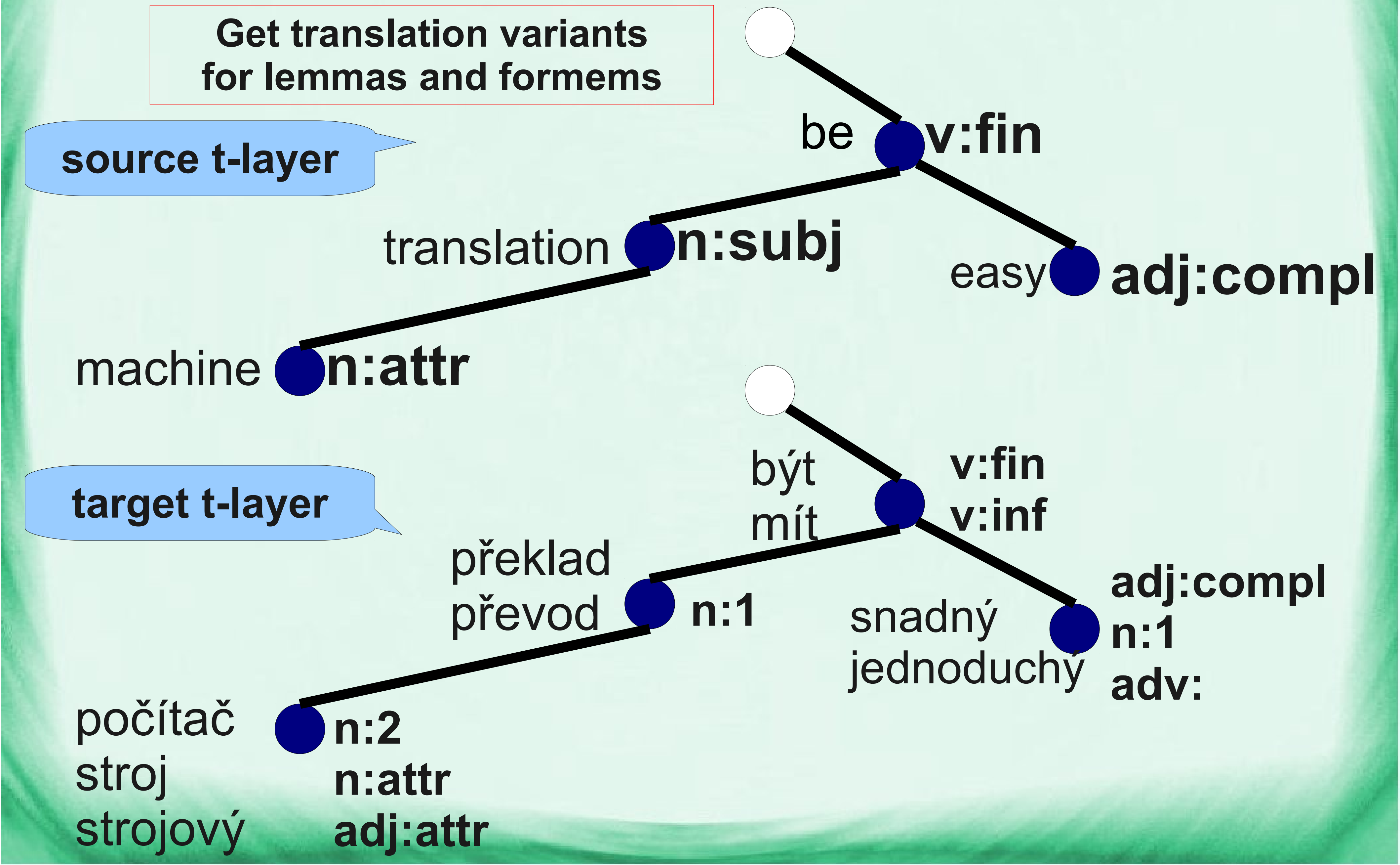
Demo Translation – Transfer

Build target t-tree by cloning



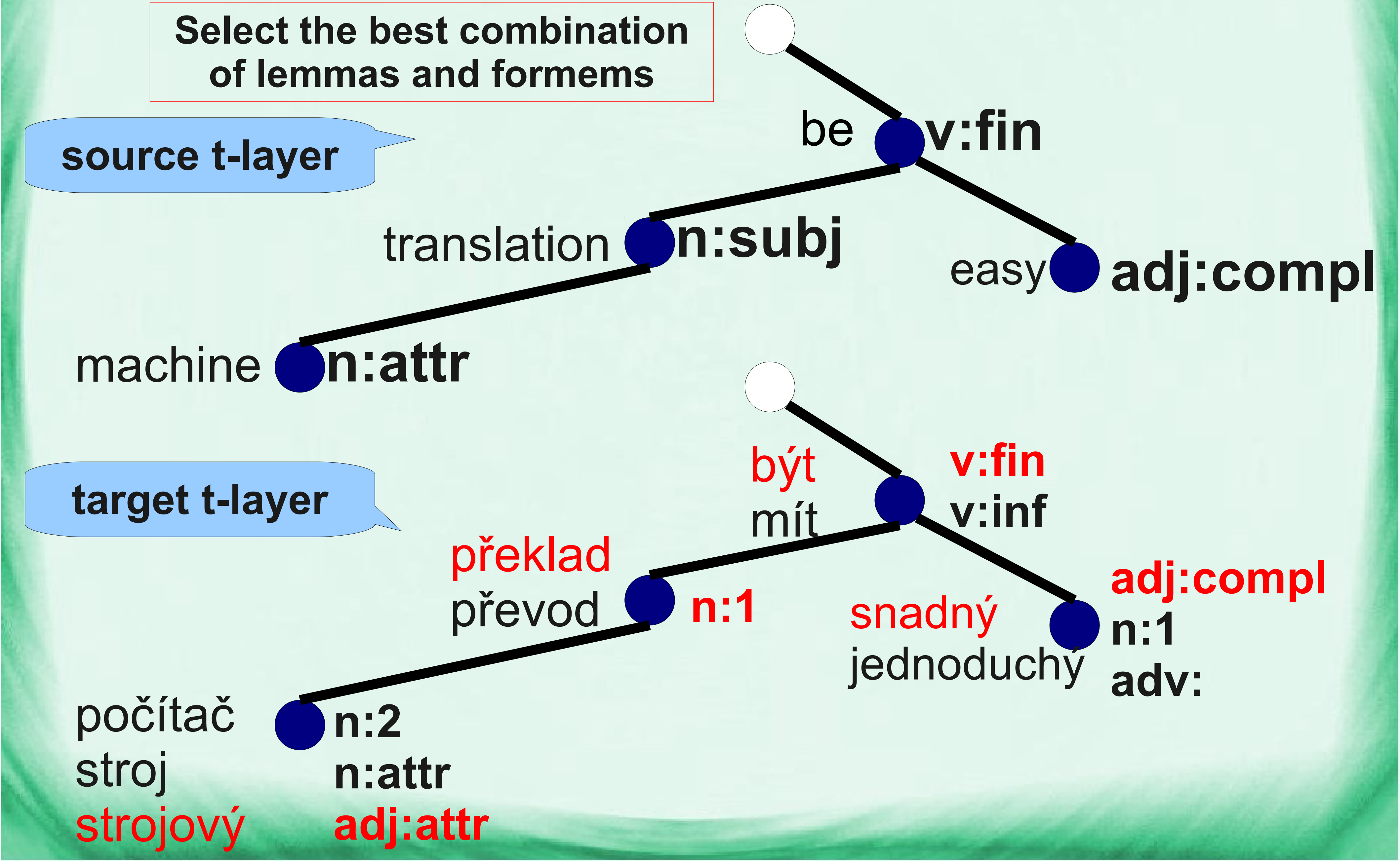
Demo Translation – Transfer

Get translation variants for lemmas and formems



Demo Translation – Transfer

Select the best combination of lemmas and formems



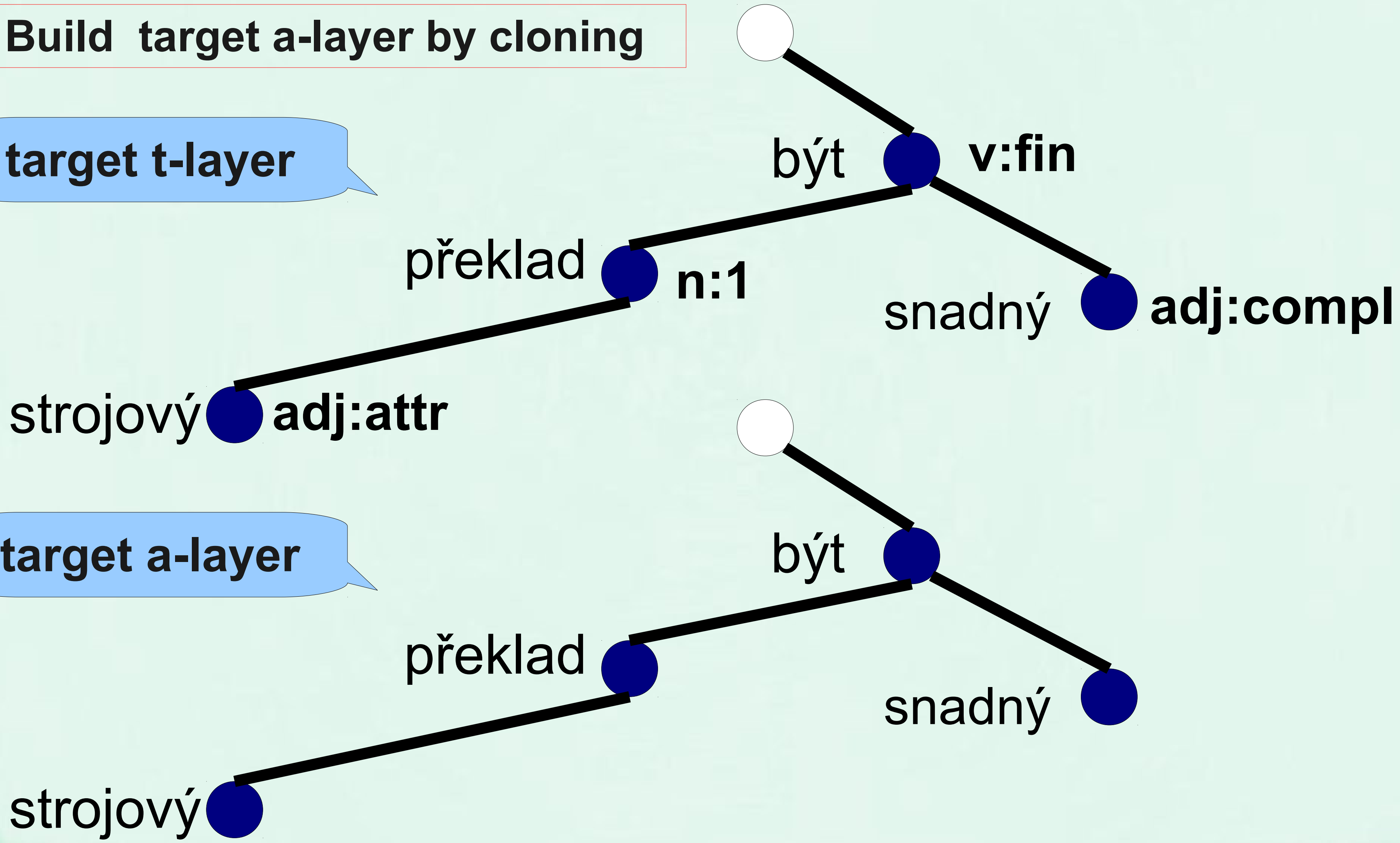
Demo Translation – Synthesis

TectoMT

Build target a-layer by cloning

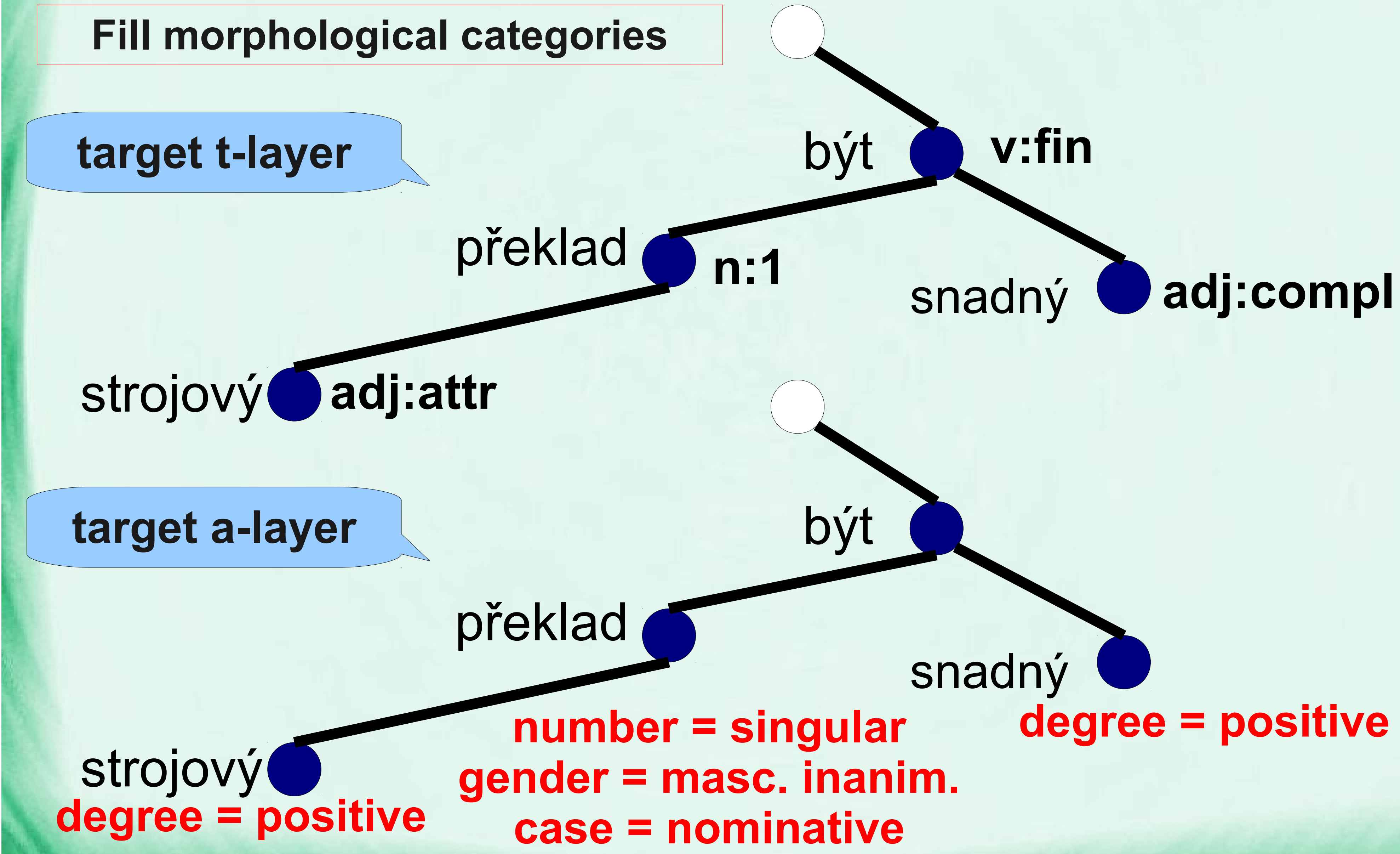
target t-layer

target a-layer



Demo Translation – Synthesis

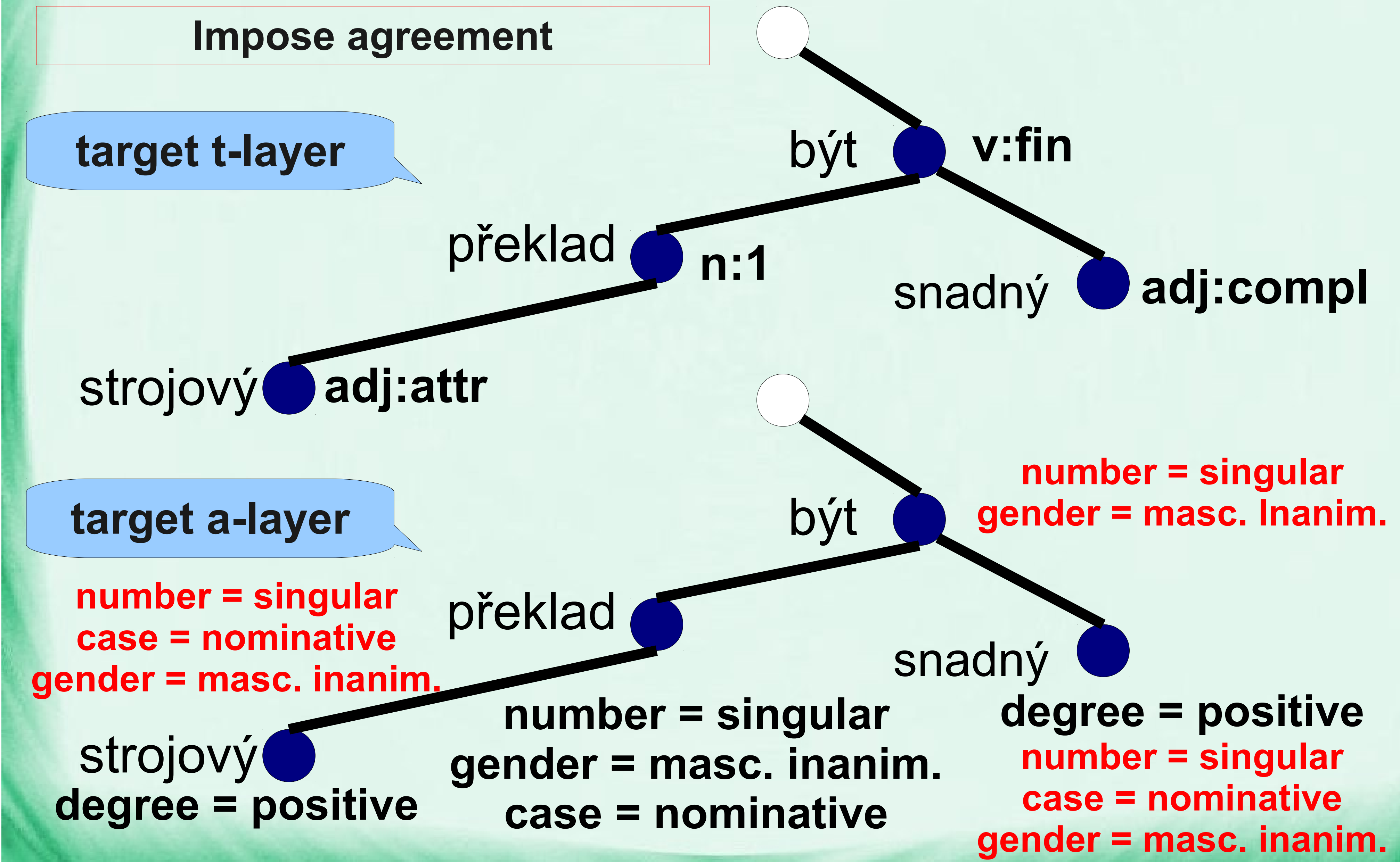
Fill morphological categories



Demo Translation – Synthesis

TectoMT

Impose agreement



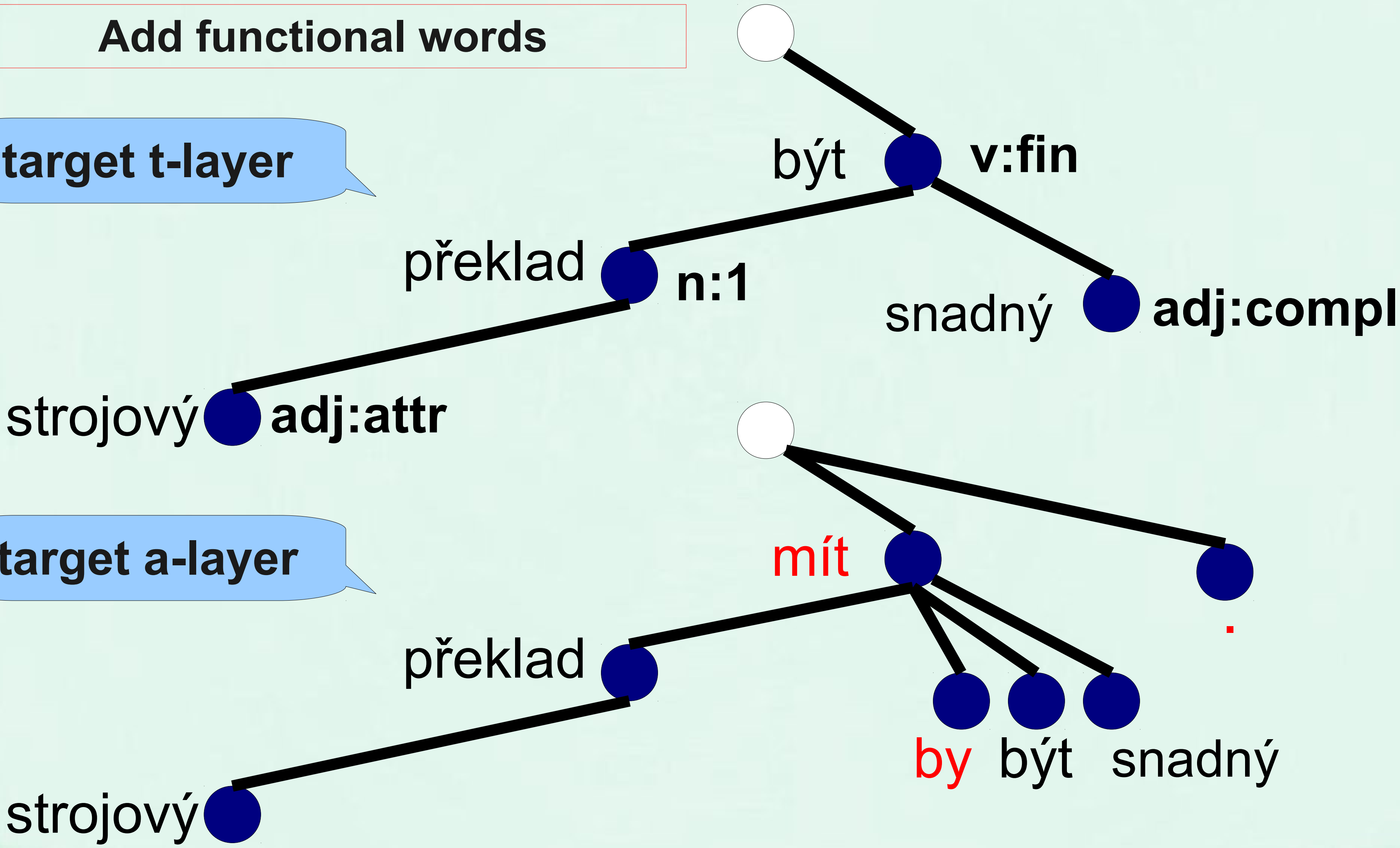
Demo Translation – Synthesis

TectoMT

Add functional words

target t-layer

target a-layer



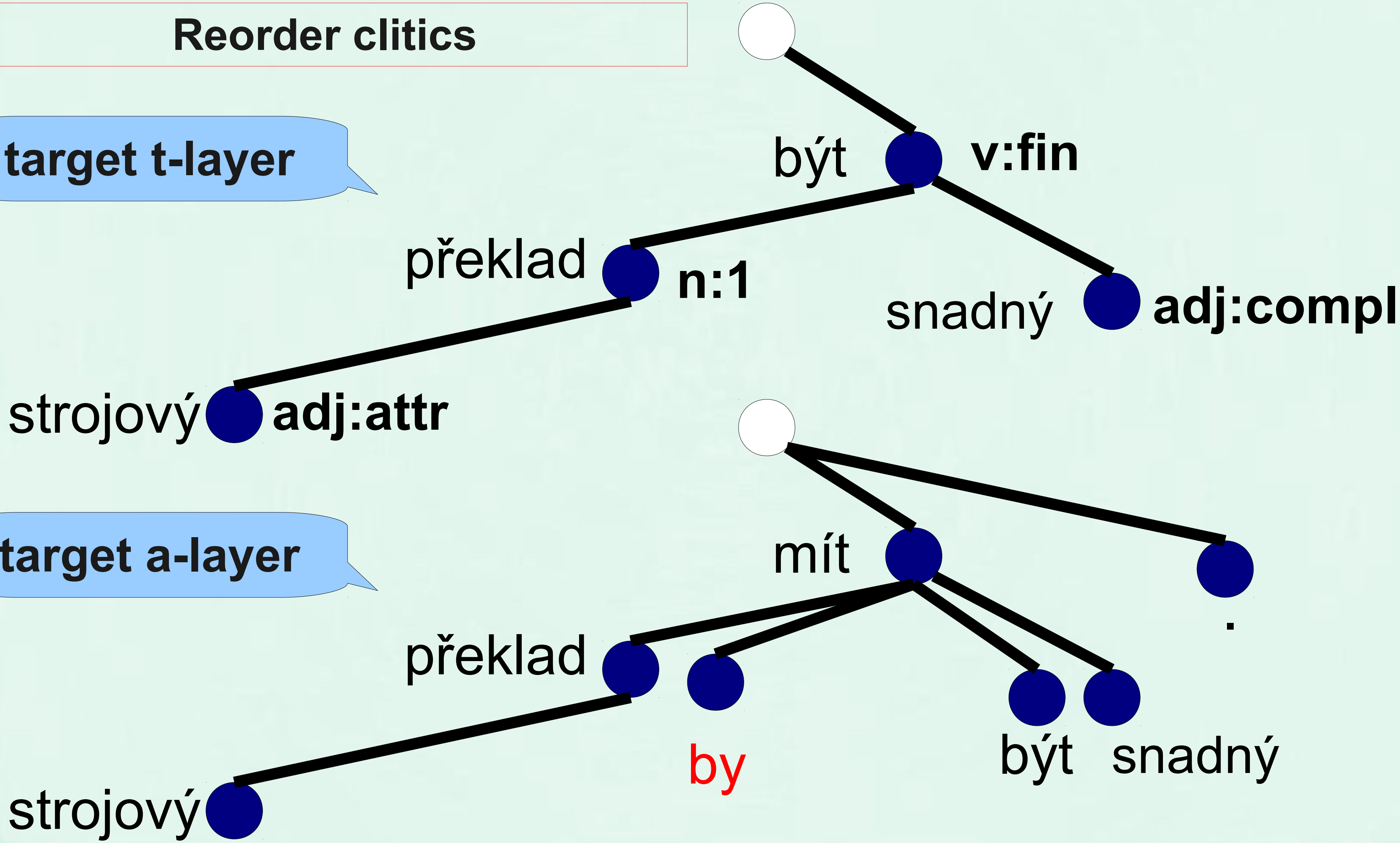
Demo Translation – Synthesis

TectoMT

Reorder clitics

target t-layer

target a-layer



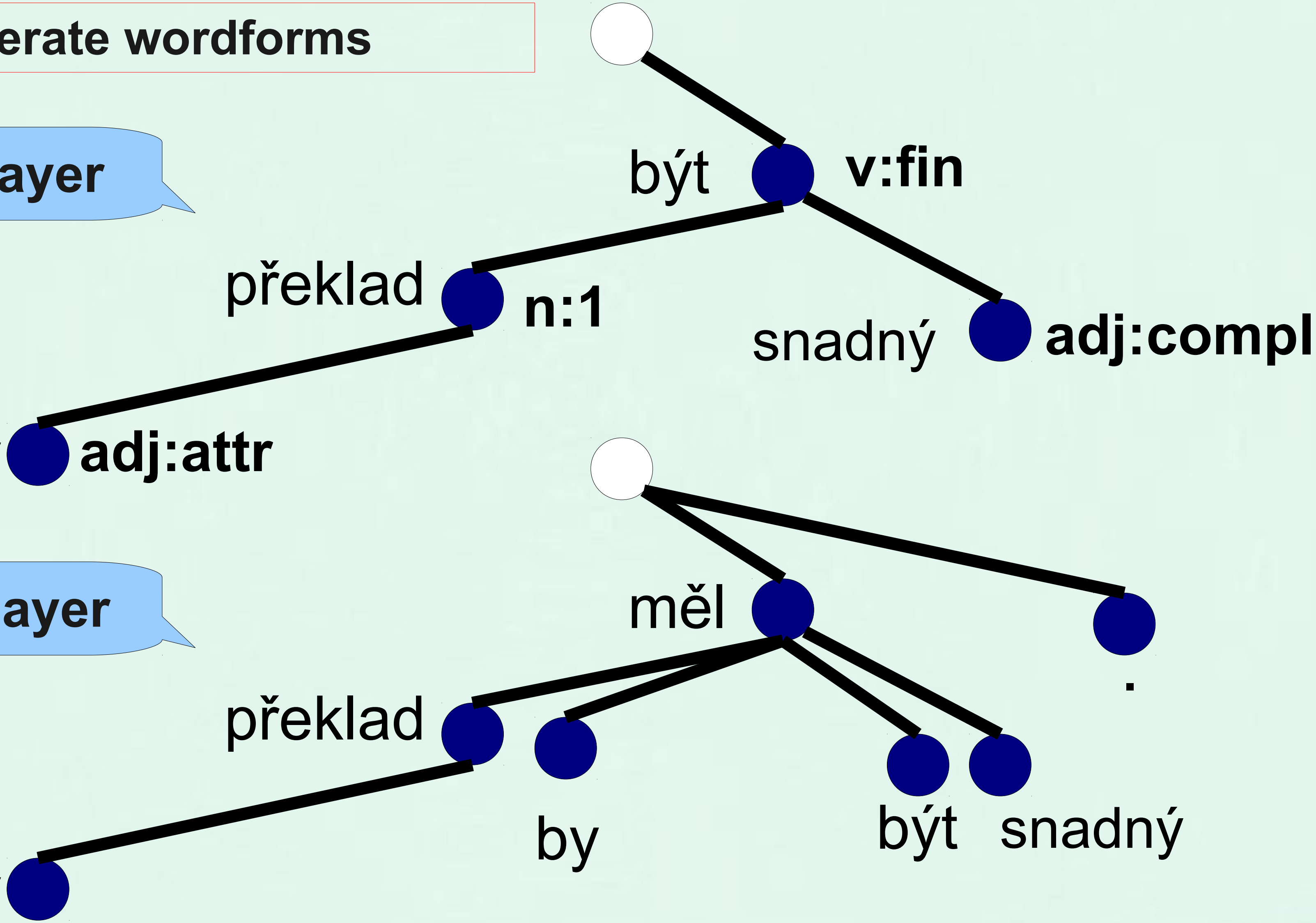
Demo Translation – Synthesis

TectoMT

Generate wordforms

target t-layer

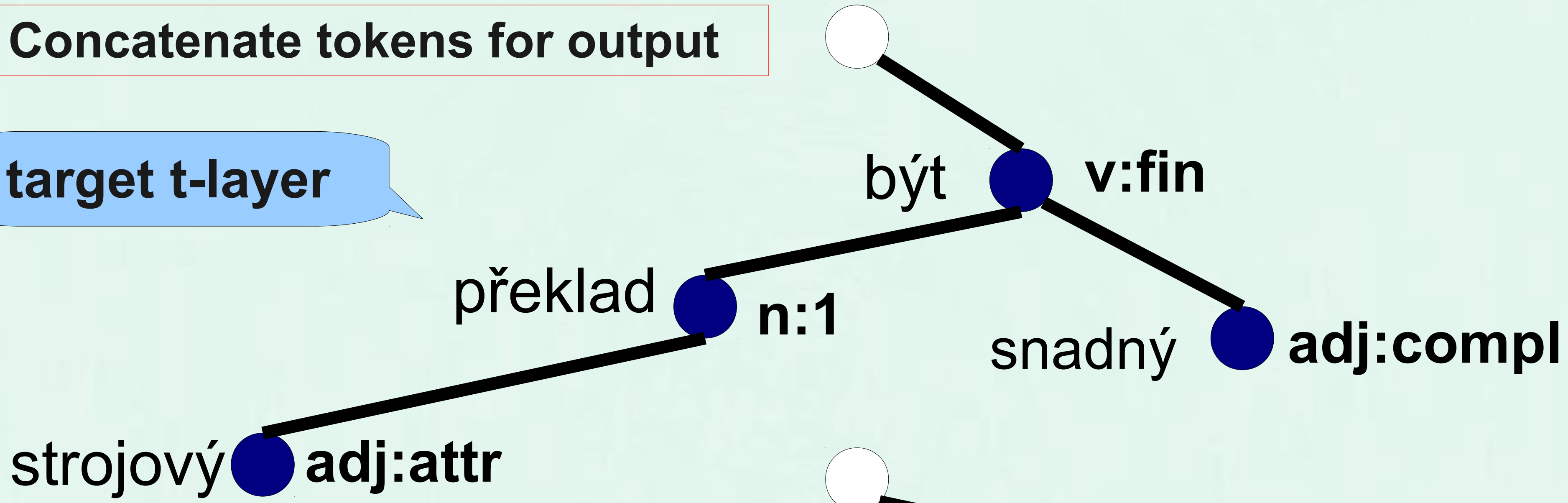
target a-layer



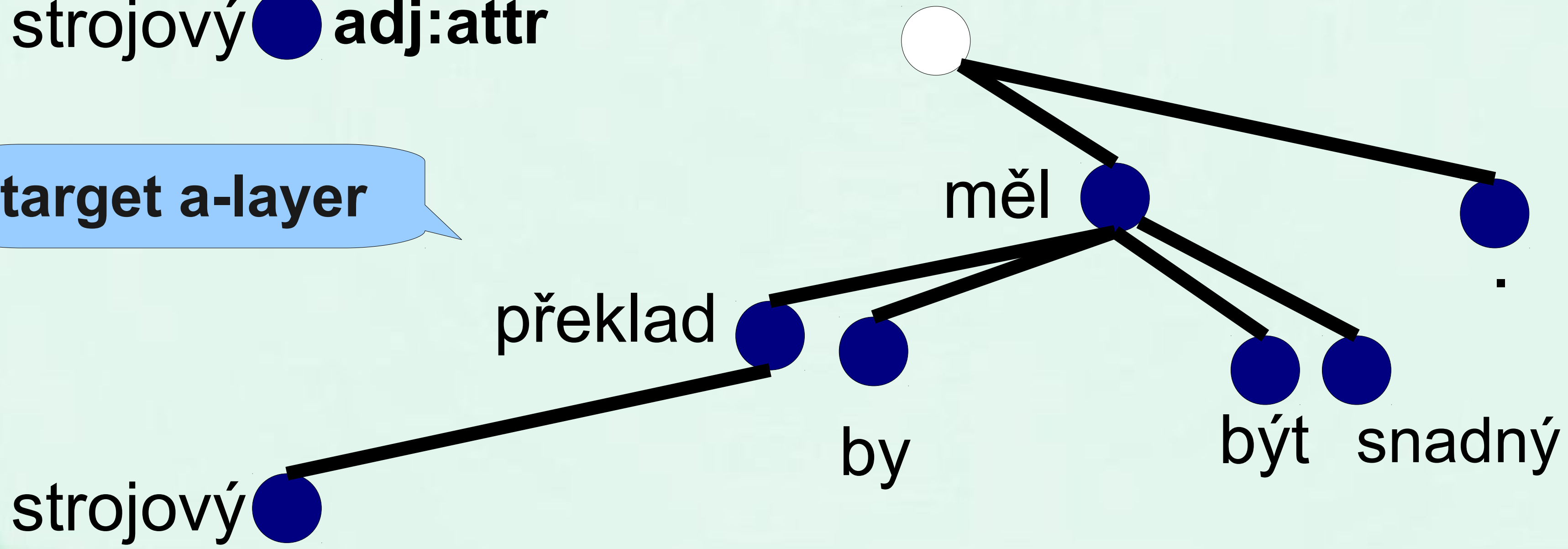
Demo Translation – Synthesis

Concatenate tokens for output

target t-layer



target a-layer



Strojový překlad by měl být snadný.

Demo Translation – Real Scenario

TectoMT

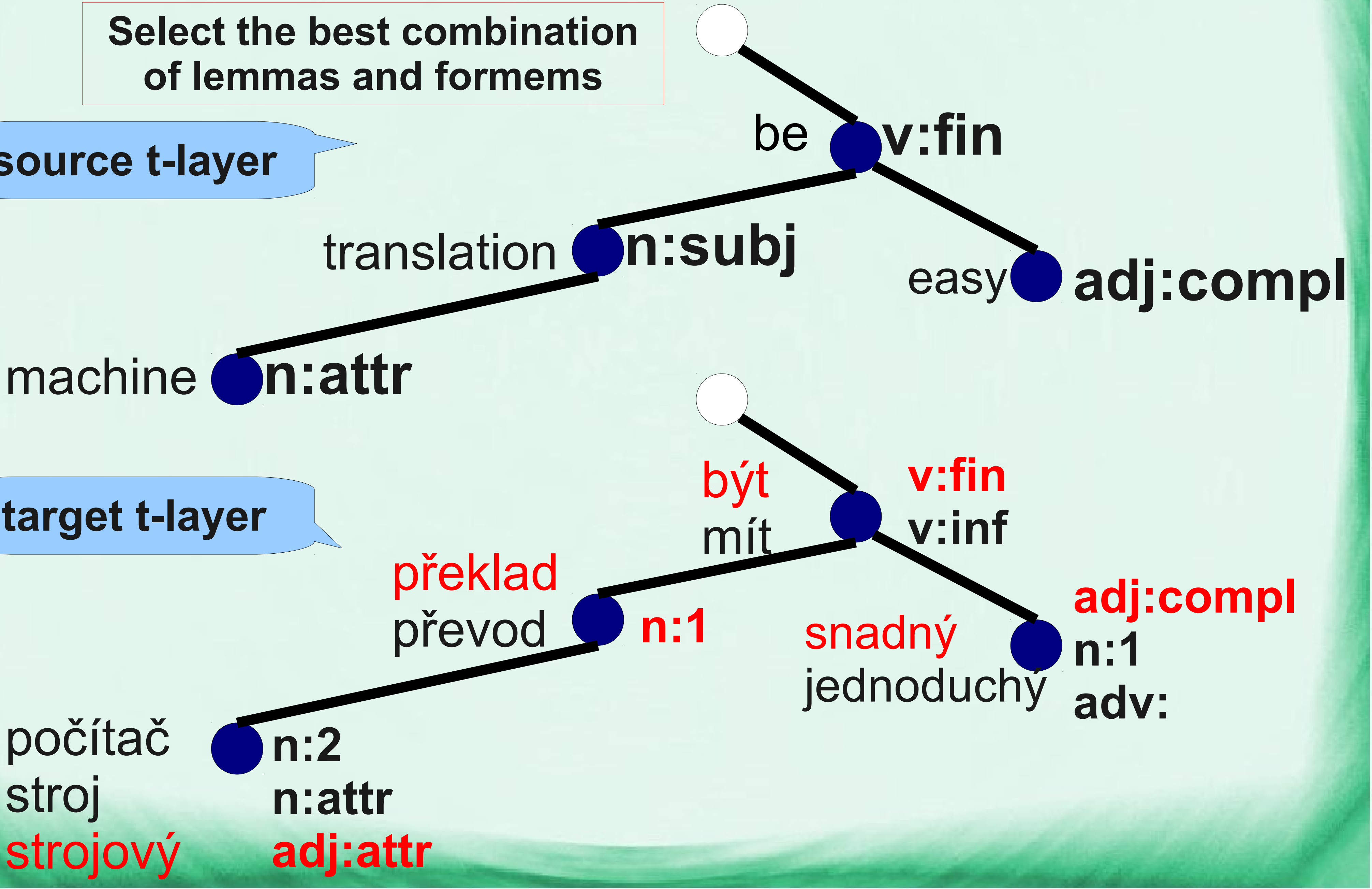
SEnglishW_to_SEnglishM::
Tokenization
Normalize_forms
Fix_tokenization
TagMorce
Fix_mtags
Lemmatize_mtree
SEnglishM_to_SEnglishN::
Stanford_named_entities
Distinguish_personal_names
SEnglishM_to_SEnglishA::
McD_parser
Fill_is_member_from_deprel
Fix_tags_after_parse
McD_parser REPARSE=1
Fill_is_member_from_deprel
Fix_McD_topology
Fix_nominal_groups
Fix_is_member
Fix_atree
Fix_multiword_prep_and_conj
Fix_dicendi_verbs
Fill_afun_AuxCP_Coord
Fill_afun
SEnglishA_to_SEnglishT::
Mark_edges_to_collapse
Mark_edges_to_collapse_neg
Build_ttree
Fill_is_member
Move_aux_from_coord_to_members
Fix_tlemmas
Assign_coap_functors
Fix_either_or
Fix_is_member
Mark_clause_heads
Mark_passives
Assign_functors
Mark_infin
Mark_relclause_heads
Mark_relclause_coref
Mark_dsp_root
Mark_parentheses
Recompute_deepord
Assign_nodetype
Assign_grammatemes
Detect_formeme
Rehang_shared_attr
Detect_voice
Fix_imperatives
Fill_is_name_of_person
Fill_gender_of_person
Add_cor_act
Find_text_coref
SEnglishT_to_TCzechT::
Clone_ttree
Translate_LF_phrases
Translate_LF_joint_static
Delete_superfluous_tnodes
Translate_F_try_rules
Translate_F_add_variants
Translate_F_rerank
Translate_L_try_rules
Translate_L_add_variants
Translate_LF_numerals_by_rules
Translate_L_filter_aspect
Transform_passive_constructions
Prune_personal_name_variants
Remove_unpassivizable_variants
Translate_LF_compounds
Cut_variants
Rehang_to_eff_parents
Translate_LF_tree_Viterbi
Rehang_to_orig_parents
Fix_transfer_choices
Translate_L_female_surnames
Add_noun_gender
Add_relpron_below_rc
Change_Cor_to_PersPron
Add_PersPron_below_vfin
Add_verb_aspect
Fix_date_time
Fix_grammatemes_after_transfer
Fix_negation
Move_adjectives_before_nouns
Move_genitives_to_postposit
Move_relclause_to_postposit
Move_dicendi_closer_to_dsp
Move_PersPron_next_to_verb
Move_enough_before_adj
Fix_money
Recompute_deepord
Find_gram_coref_for_refl_pron
Neut_PersPron_gender_from_antec
Override_pp_with_phrase_translation
Valency_related_rules
Fill_clause_number
Turn_text_coref_to_gram_coref
TCzechT_to_TCzechA::
Clone_atree
Distinguish_homonymous_mlemmas
Reverse_number_noun_dependency
Init_morphcat
Fix_possessive_adjectives
Mark_subject
Impose_pron_z_agr
Impose_rel_pron_agr
Impose_subjpred_agr
Impose_attr_agr
Impose_compl_agr
Drop_subj_pers_prons
Add_prepositions
Add_subconjs
Add_reflex_particles
Add_auxverb_compound_passive
Add_auxverb_modal
Add_auxverb_compound_future
Add_auxverb_conditional
Add_auxverb_compound_past
Add_clausal_expletive_pronouns
Resolve_verbs
Project_clause_number
Add_parentheses
Add_sent_final_punct
Add_subord_clause_punct
Add_coord_punct
Add_apposition_punct
Choose_mlemma_for_PersPron
Generate_wordforms
Move_clitics_to_wackernagel
Recompute_ordering
Delete_superfluous_prepos
Delete_empty_nouns
Vocalize_prepositions
Capitalize_sent_start
Capitalize_named_entities
TCzechA_to_TCzechW::
Concatenate_tokens
Ascii_quotes
Remove_repeated_tokens

HMTM – Motivation

Select the best combination of lemmas and formems

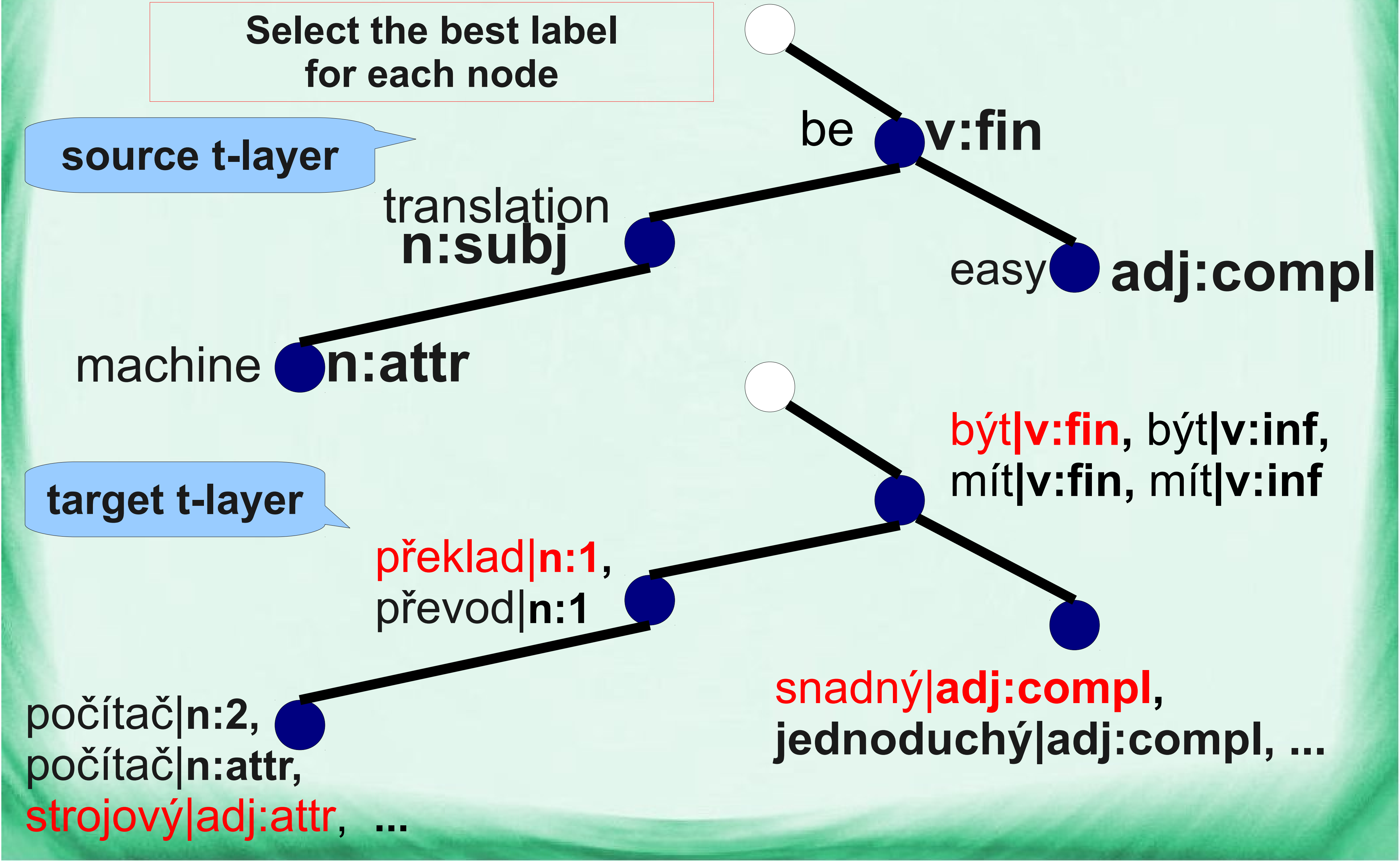
source t-layer

target t-layer



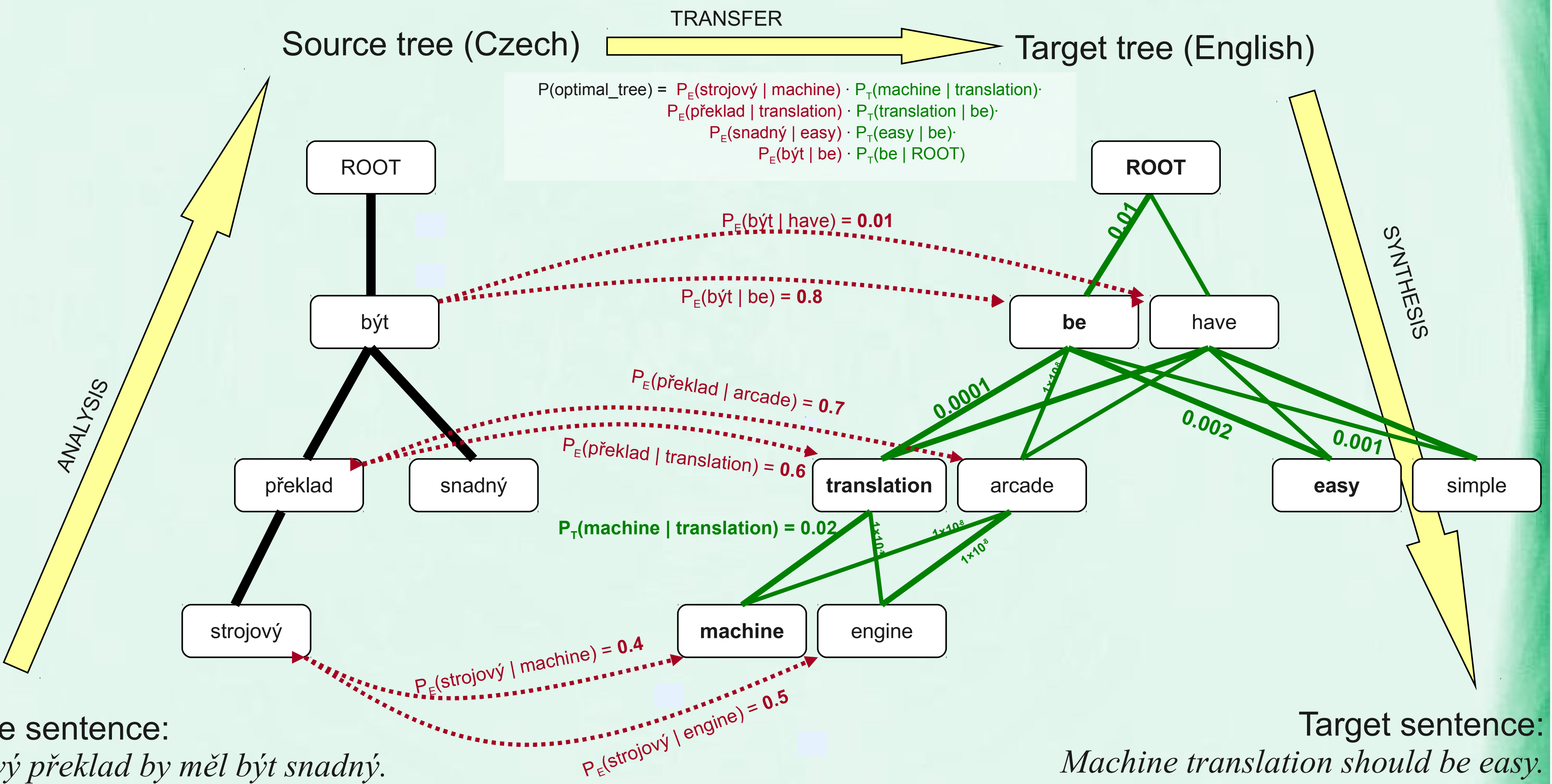
HMTM – Motivation

Select the best label for each node



HMTM in MT

TectoMT



P(optimal_tree) = P_E(strojový | machine) · P_T(machine | translation)·
P_E(překlad | translation) · P_T(translation | be)·
P_E(snadný | easy) · P_T(easy | be)·
P_E(být | be) · P_T(be | ROOT)

Source sentence:
Strojový překlad by měl být snadný.

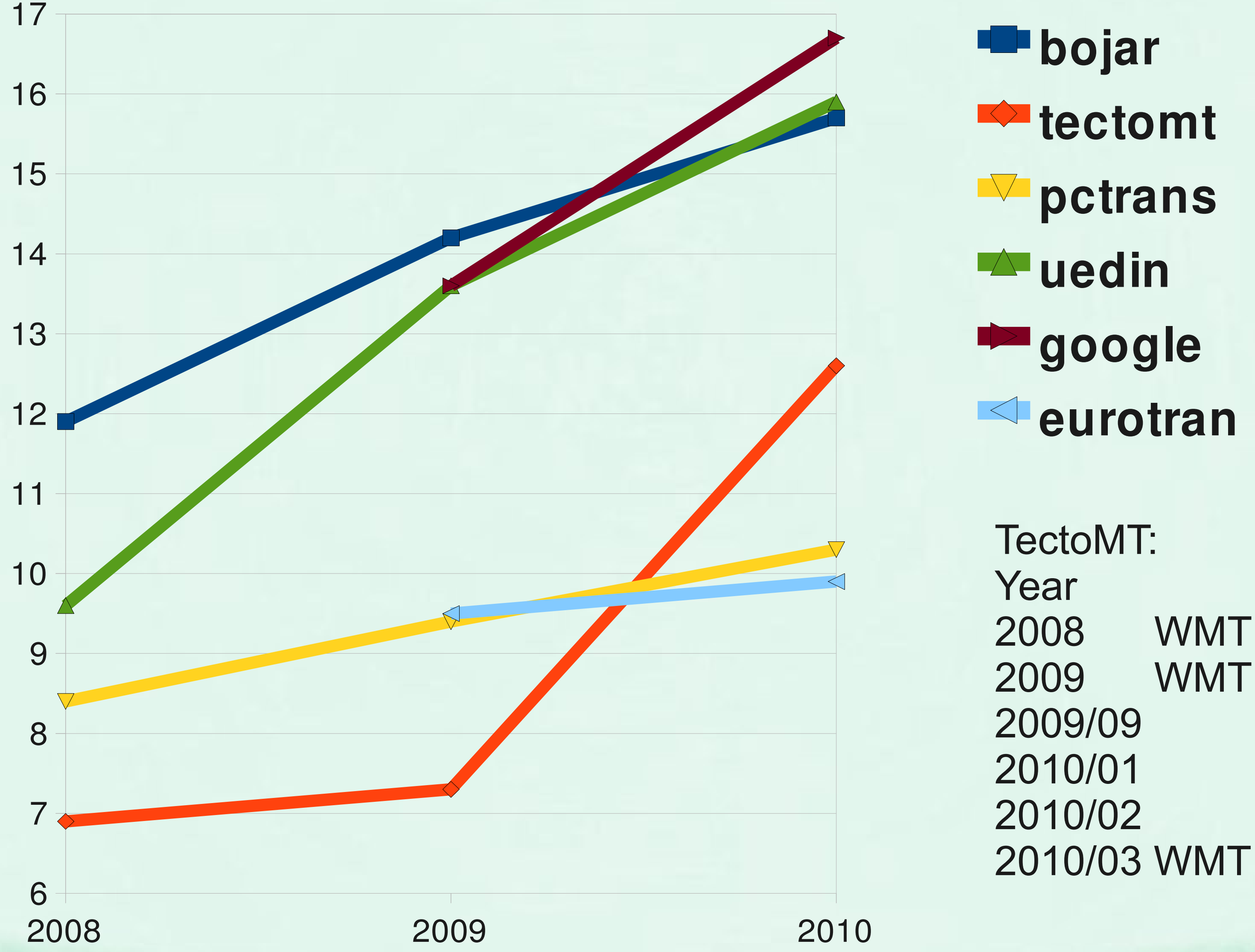
Target sentence:
Machine translation should be easy.

P_E(source | target) … emission probabilities … **translation model**
P_T(dependent | governing) … transition probabilities … **target-language tree model**

Results – BLEU

TectoMT

WMT = Workshop on Statistical Machine Translation



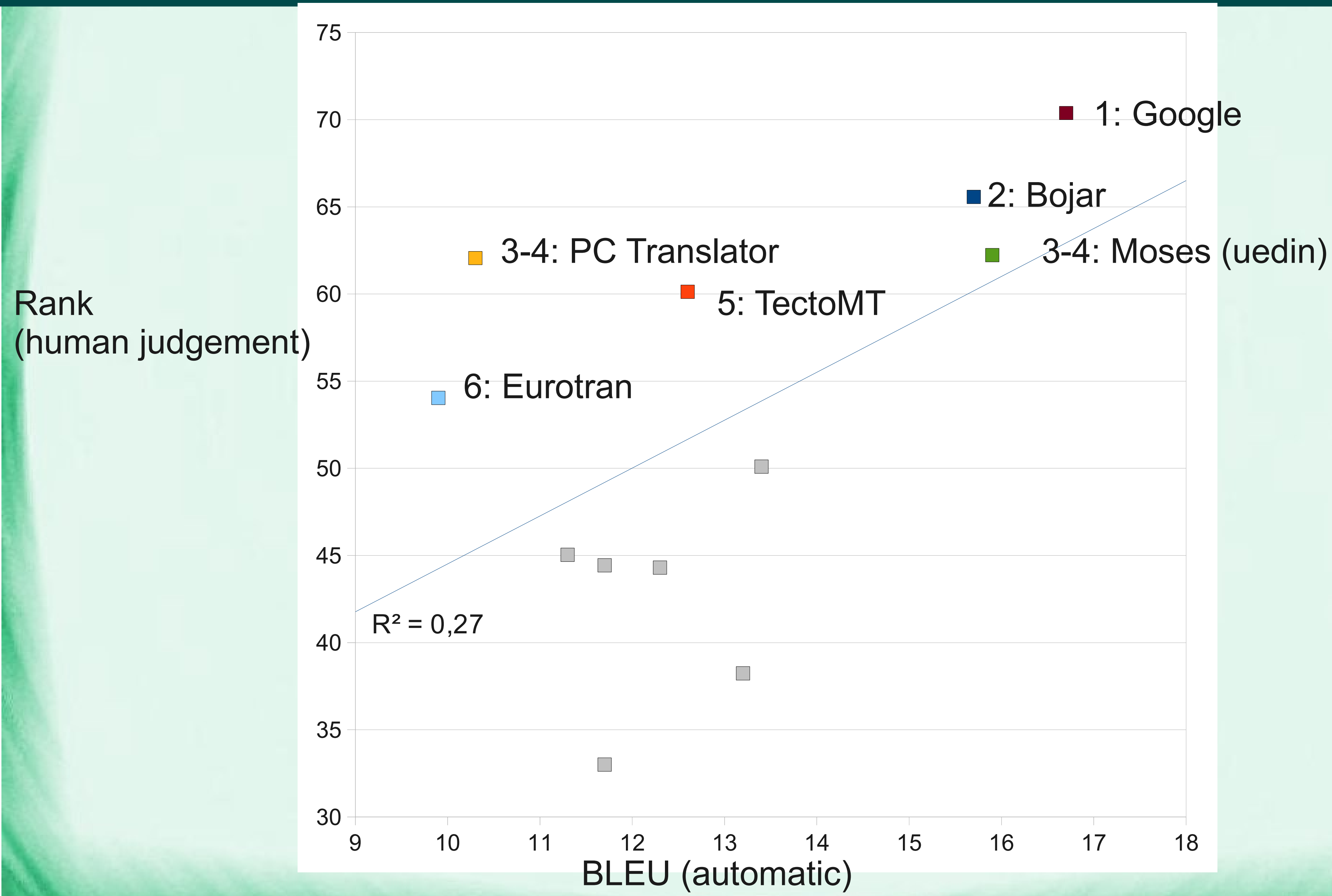
TectoMT:

Year		BLEU
2008	WMT	6,9
2009	WMT	7,3
2009/09		10,2
2010/01		10,4
2010/02		11,3
2010/03	WMT	12,6

WMT results: http://matrix.statmt.org

Results – BLEU vs. Ranks

TectoMT



Results – task-based evaluation

Example

- **Translated text:**
  Bílý dům náčelník štábu Rahm Emanuel je naplánovaný k pojídaný vsedě v budově Kongresu se sněmovnou diskutující Nancy Pelosi v 9 p.m.
- **Yes/No Statements:**
  - *Nancy Pelosi je zaměstnankyní Bílého domu.*
  - *Schůzka se koná dopoledne.*
  - *Schůzka se koná v Bílem domě.*

Results

TectoMT	PC Translator	Google	Moses
80.56	80.23	78.68	73.58

Jan Berka, Martin Černý, 2010: 18 annotators, 1905 * 3 statements, 4 systems evaluated
[https://svn.ms.mff.cuni.cz/projects/NPFL087/export/598/projects/task-based-evaluation/tbe.pdf]

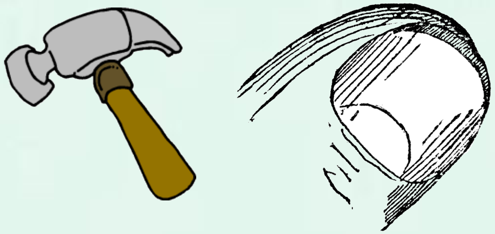
Examples of Translation

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.