Strojový překlad přes tektogramatickou rovinu v systému TectoMT

Martin Popel
ÚFAL, MFF UK

Pondělní seminář, 22. března 2010

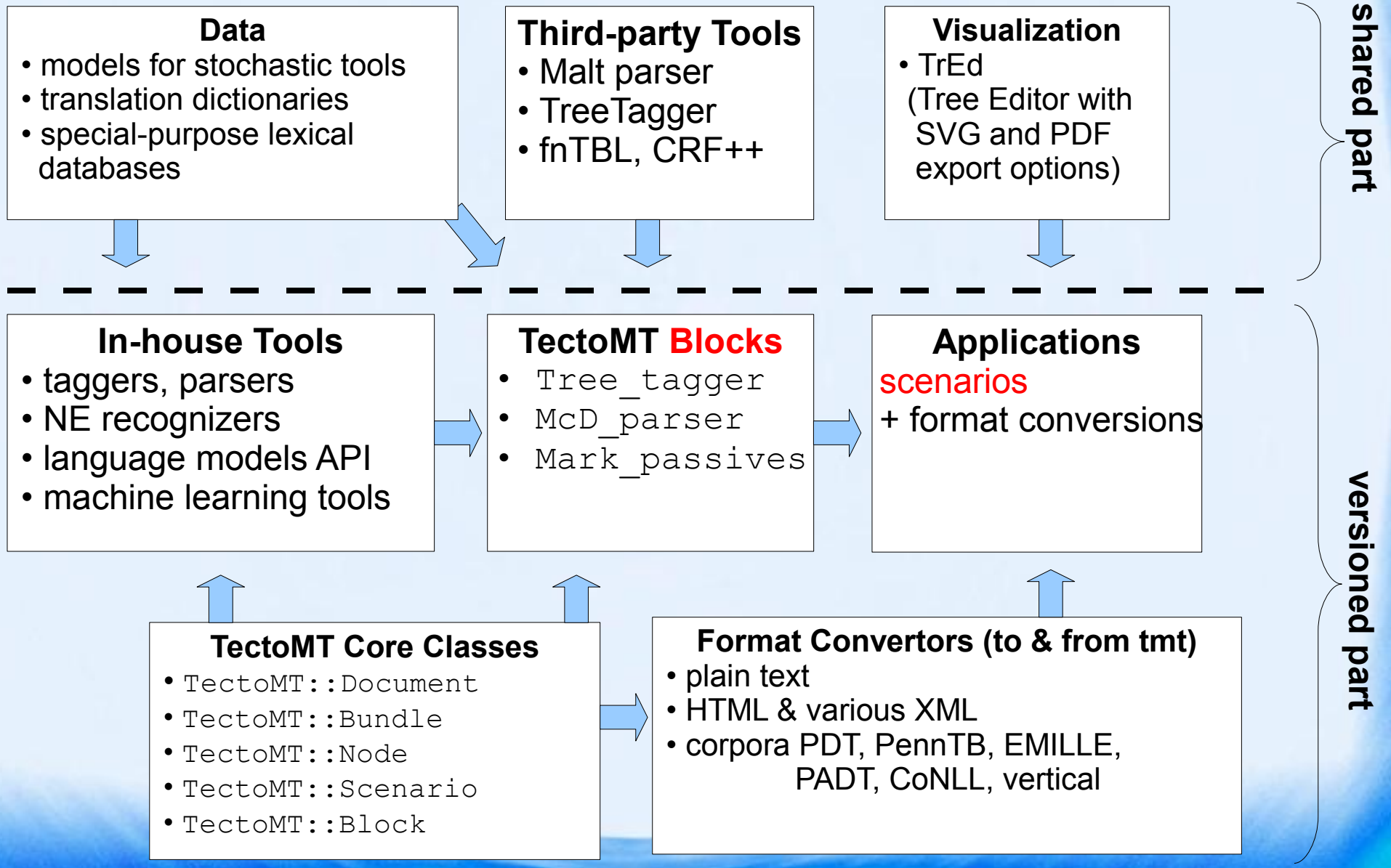
Osnova

TectoMT

- Ukázka překladu krok za krokem
- Anotace překladových chyb
- Novinky v TectoMT
  - Hidden Markov Tree Models (HMTM)
  - nové slovníky (Maximum Entropy)
- Výsledky a zhodnocení

TectoMT jako framework

modulární, open source, objektový, Perl, Linux



Ukázka překladu - Schéma

transfer přes tektogramatickou rovinu

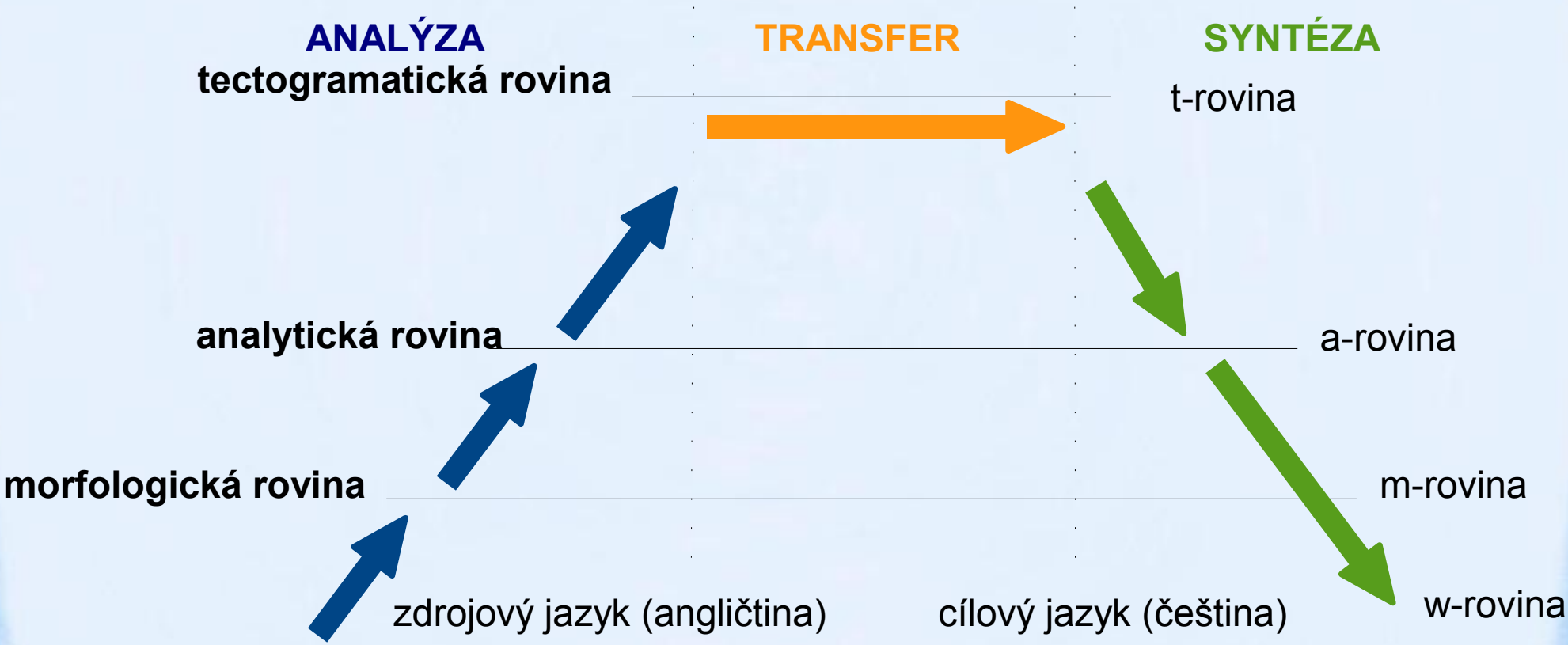
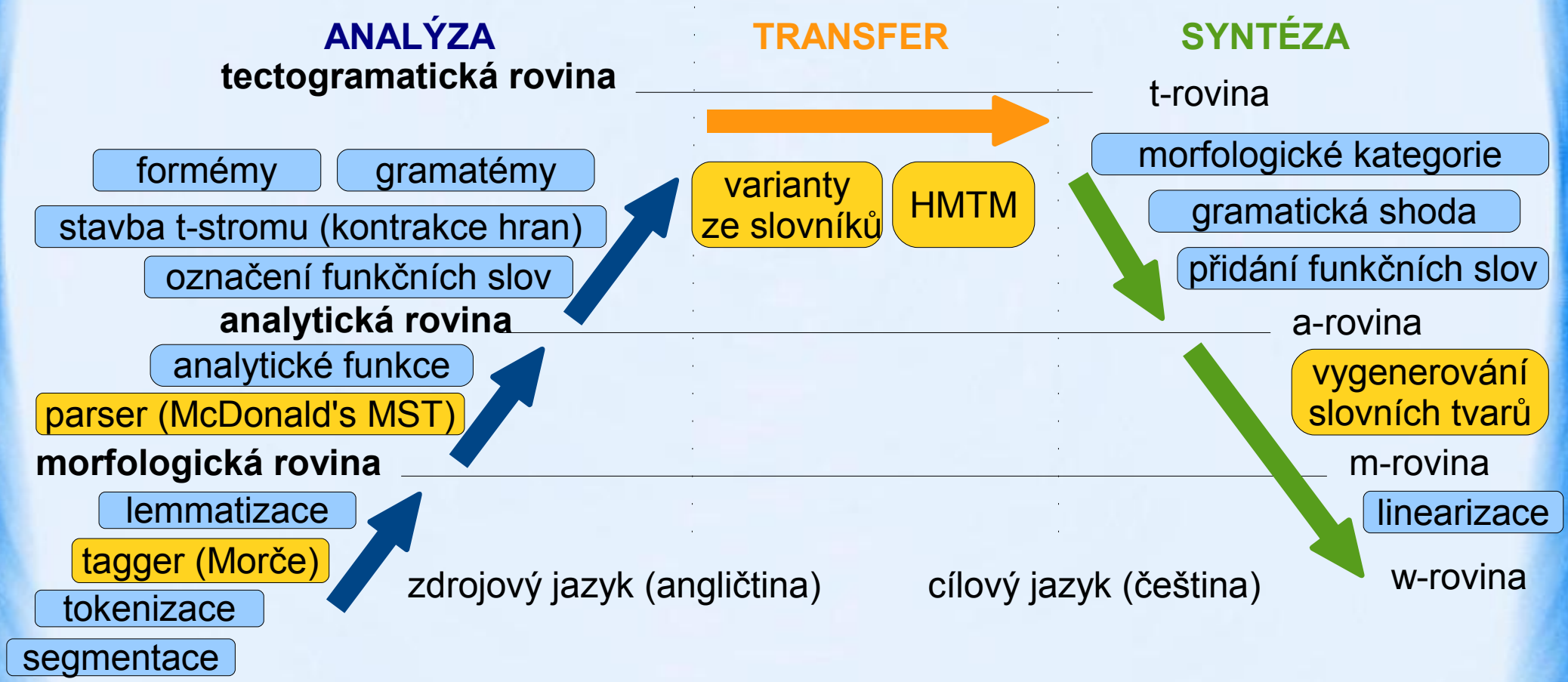


Schéma překladu v TectoMT

pravidlové a statistické bloky



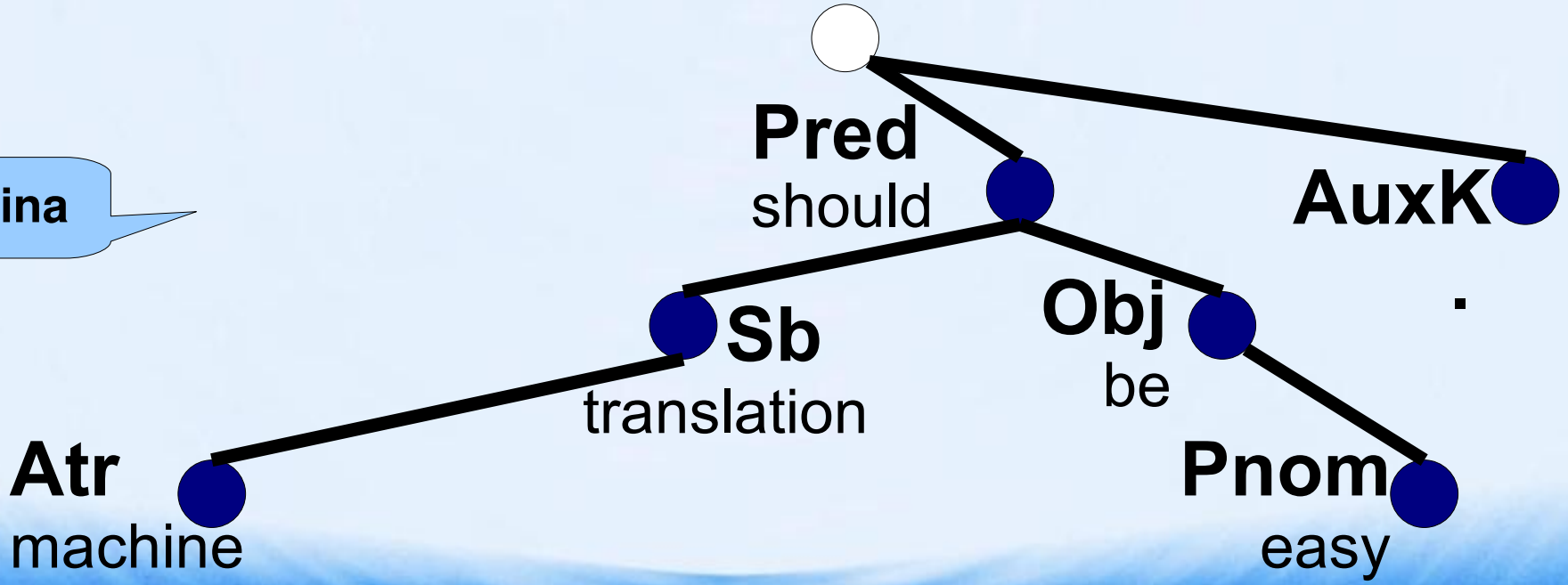
Ukázka překladu – Analýza

vstupní text Machine translation should be easy.

m-rovina

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

a-rovina



Ukázka překladu – Analýza

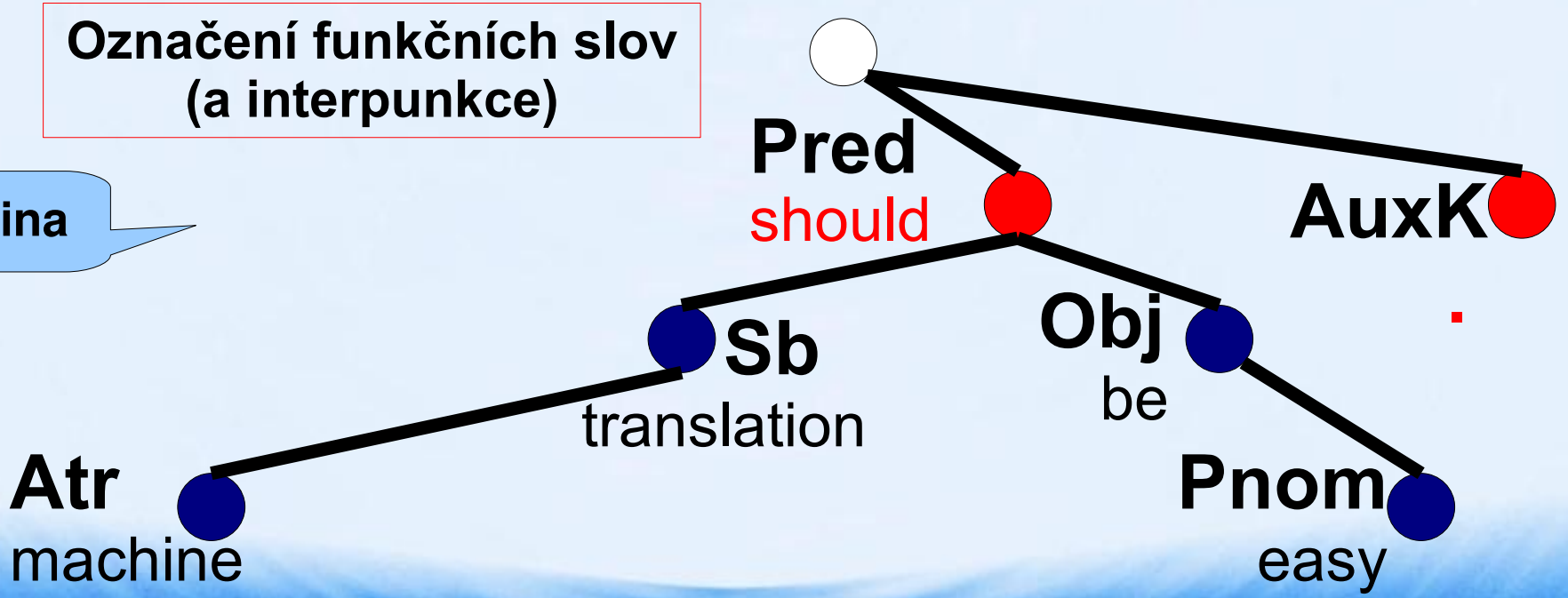
vstupní text Machine translation should be easy.

m-rovina

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

Označení funkčních slov (a interpunkce)

a-rovina



Ukázka překladu – Analýza

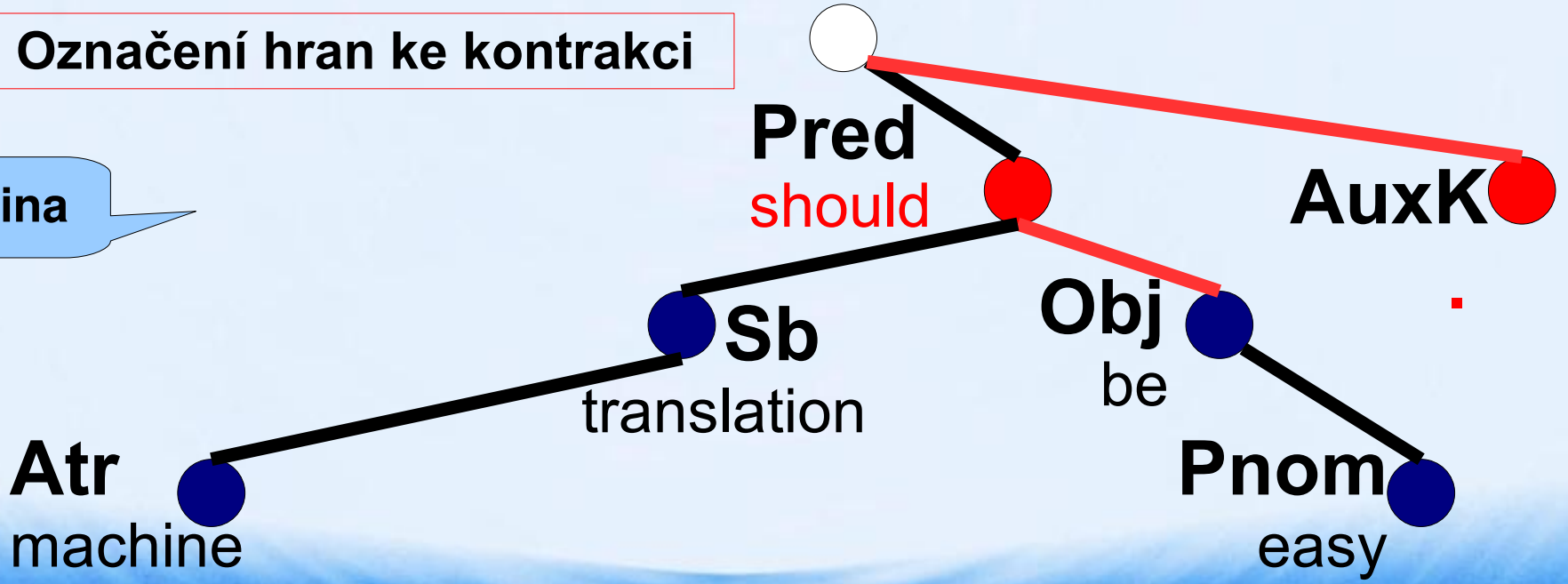
vstupní text Machine translation should be easy.

m-rovina

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

Označení hran ke kontrakci

a-rovina



Ukázka překladu – Analýza

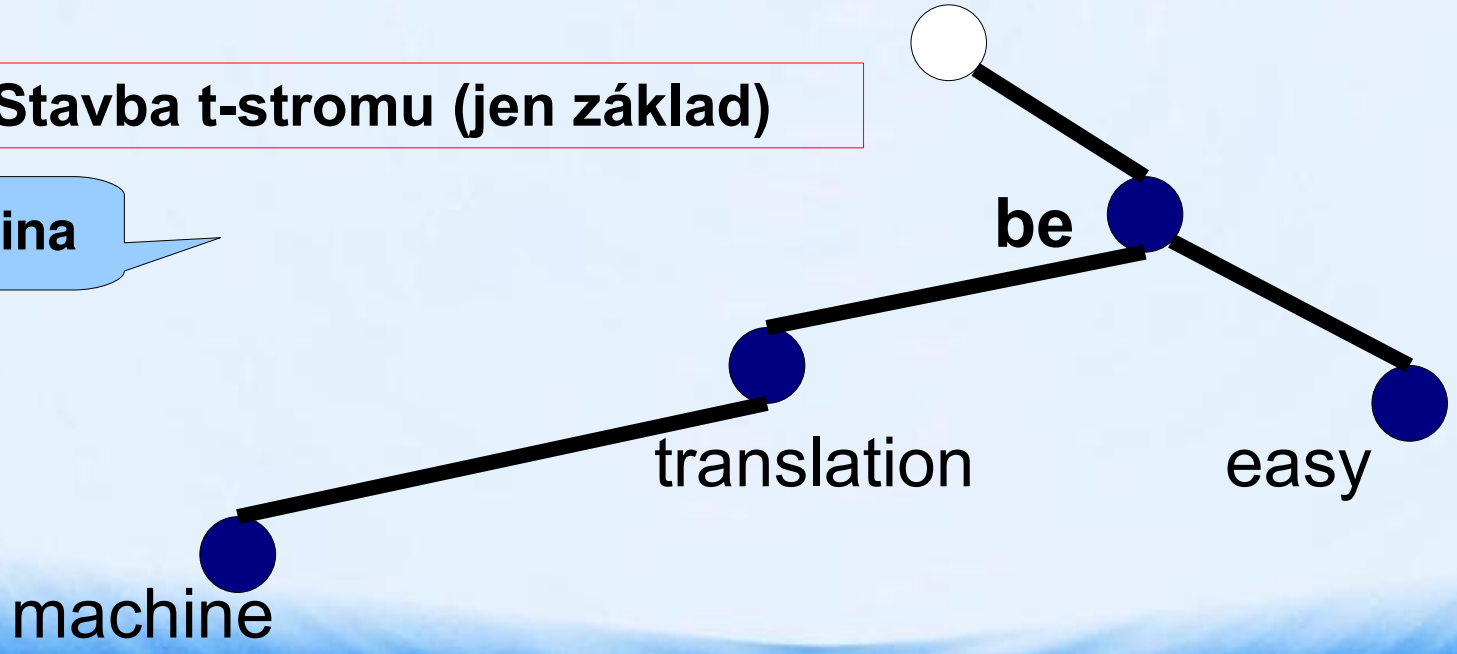
vstupní text Machine translation should be easy.

m-rovina

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

Stavba t-stromu (jen základ)

a-rovina



Ukázka překladu – Analýza

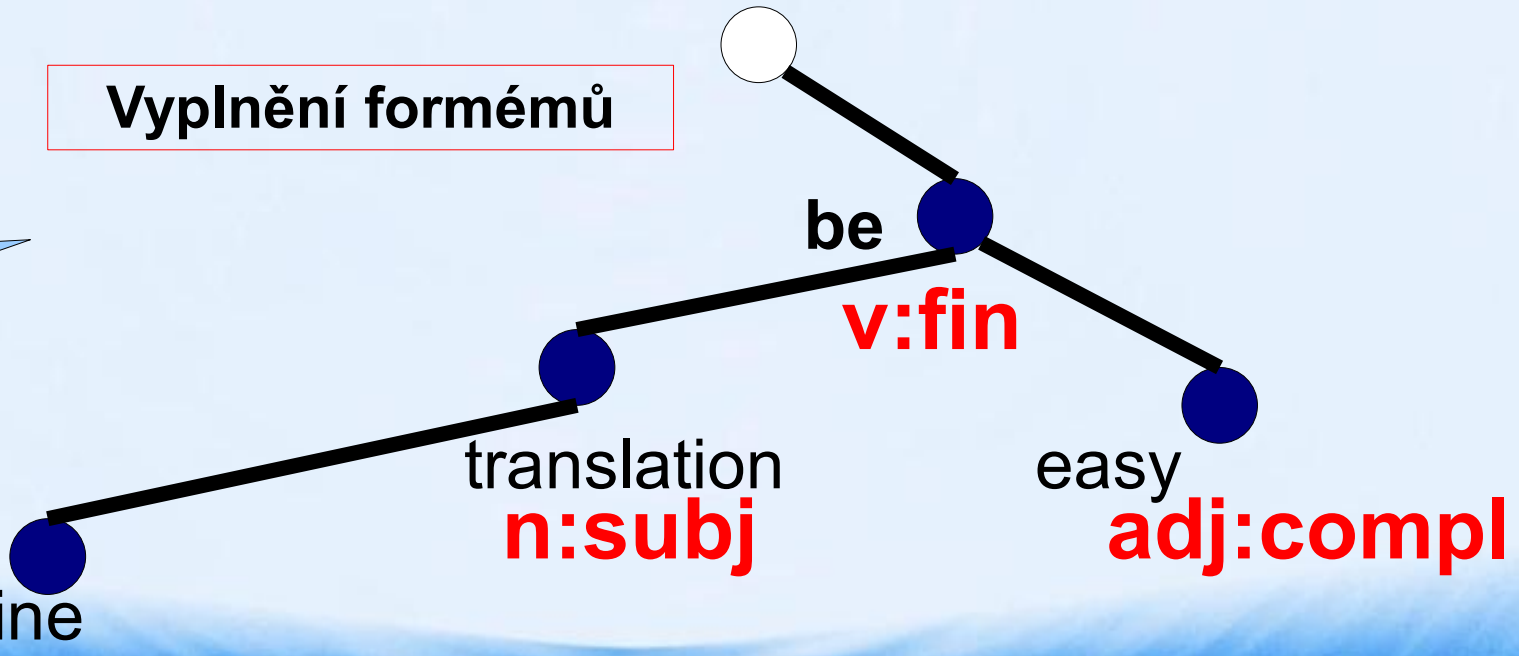
vstupní text Machine translation should be easy.

m-rovina

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

a-rovina

Vyplnění formémů



Ukázka překladu – Analýza

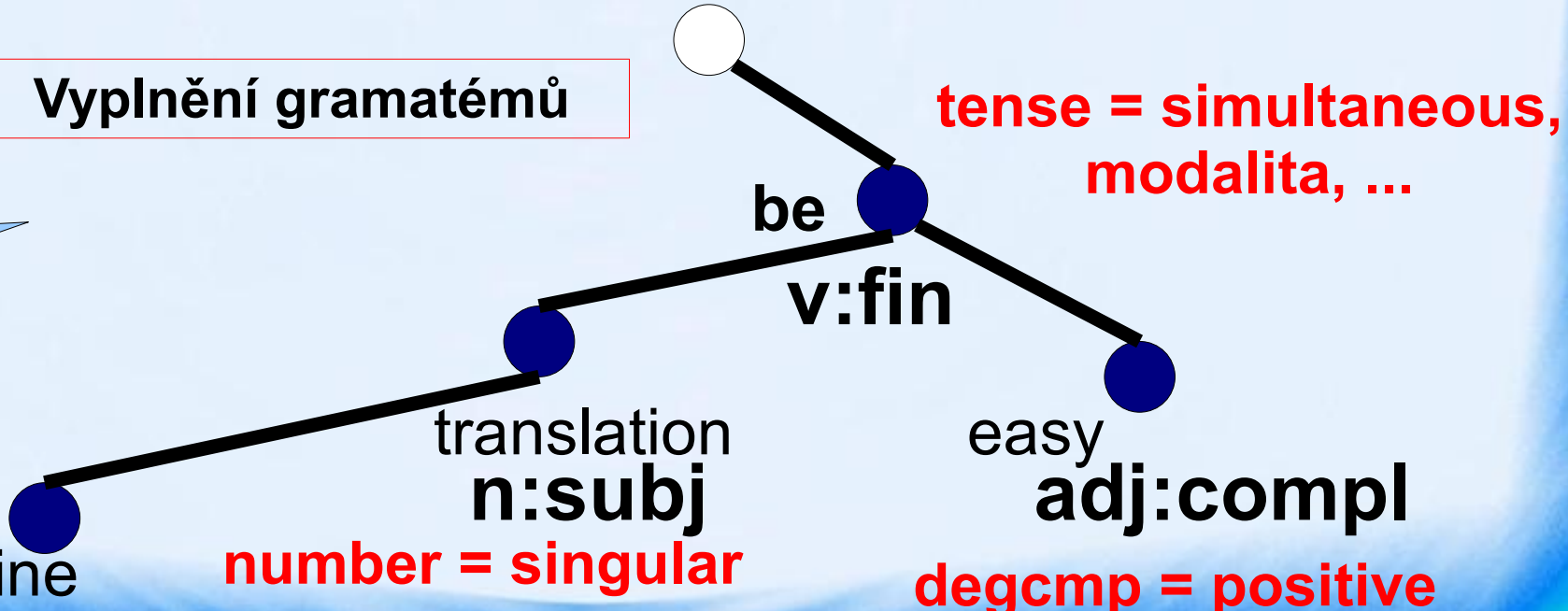
vstupní text Machine translation should be easy.

m-rovina

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

Vyplnění gramatémů

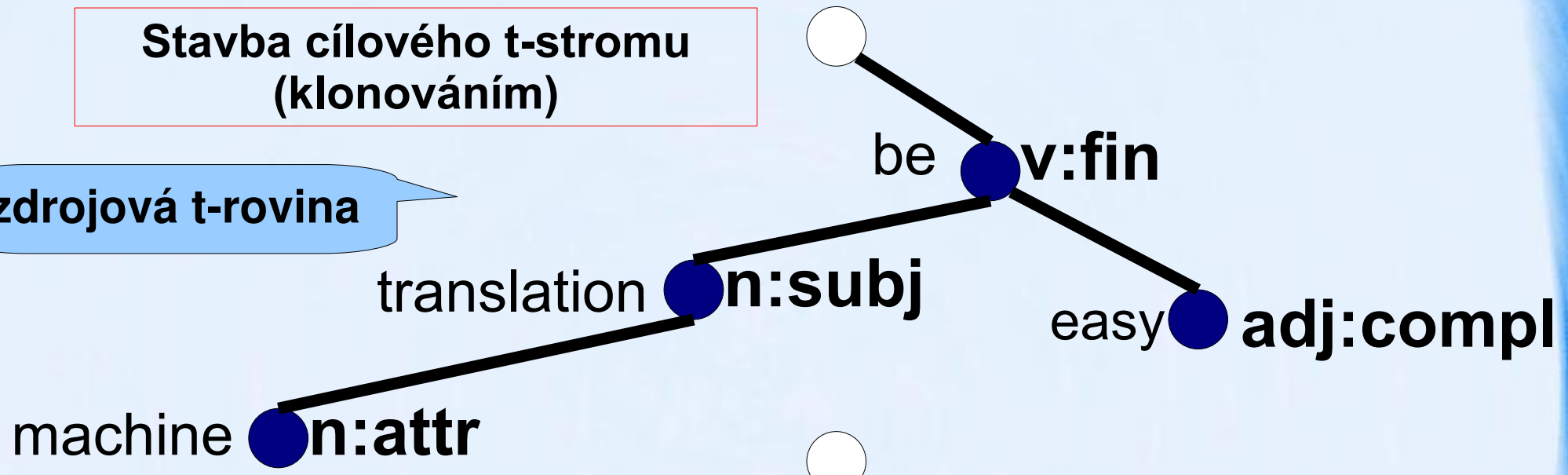
a-rovina



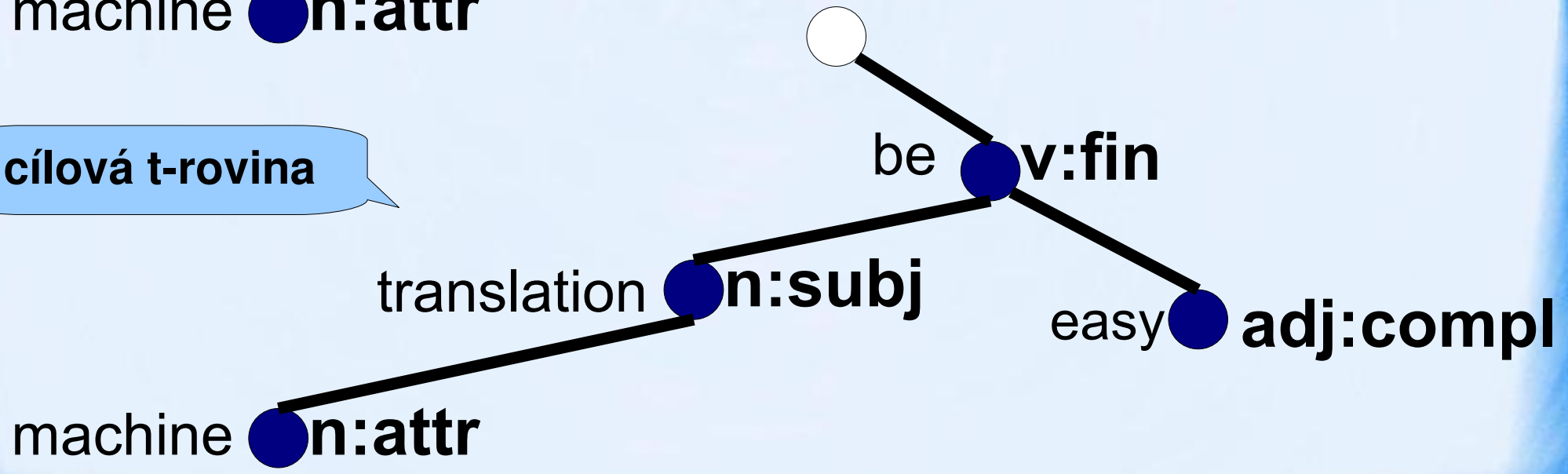
Ukázka překladu – Transfer

Stavba cílového t-stromu (klonováním)

zdrojová t-rovina

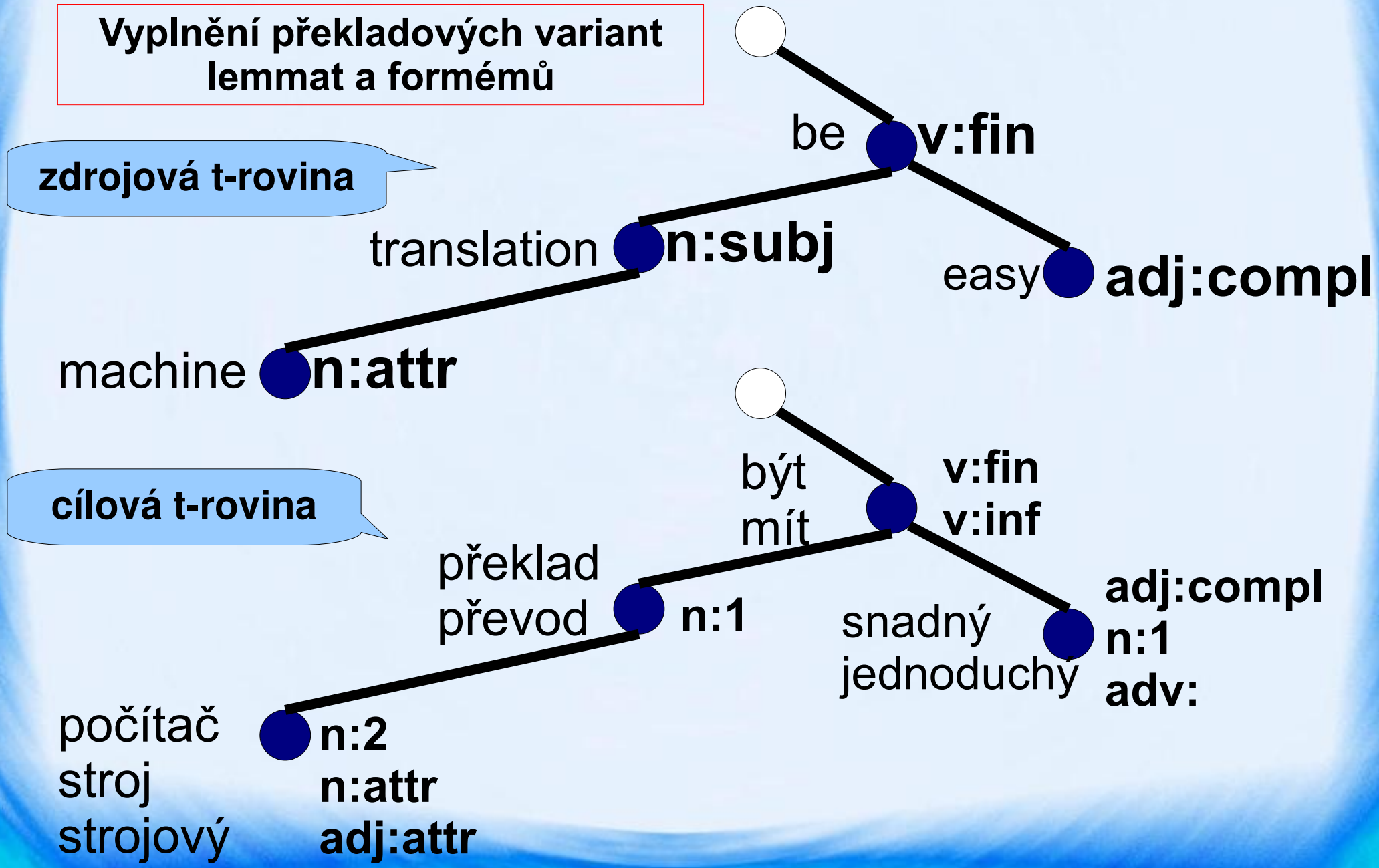


cílová t-rovina



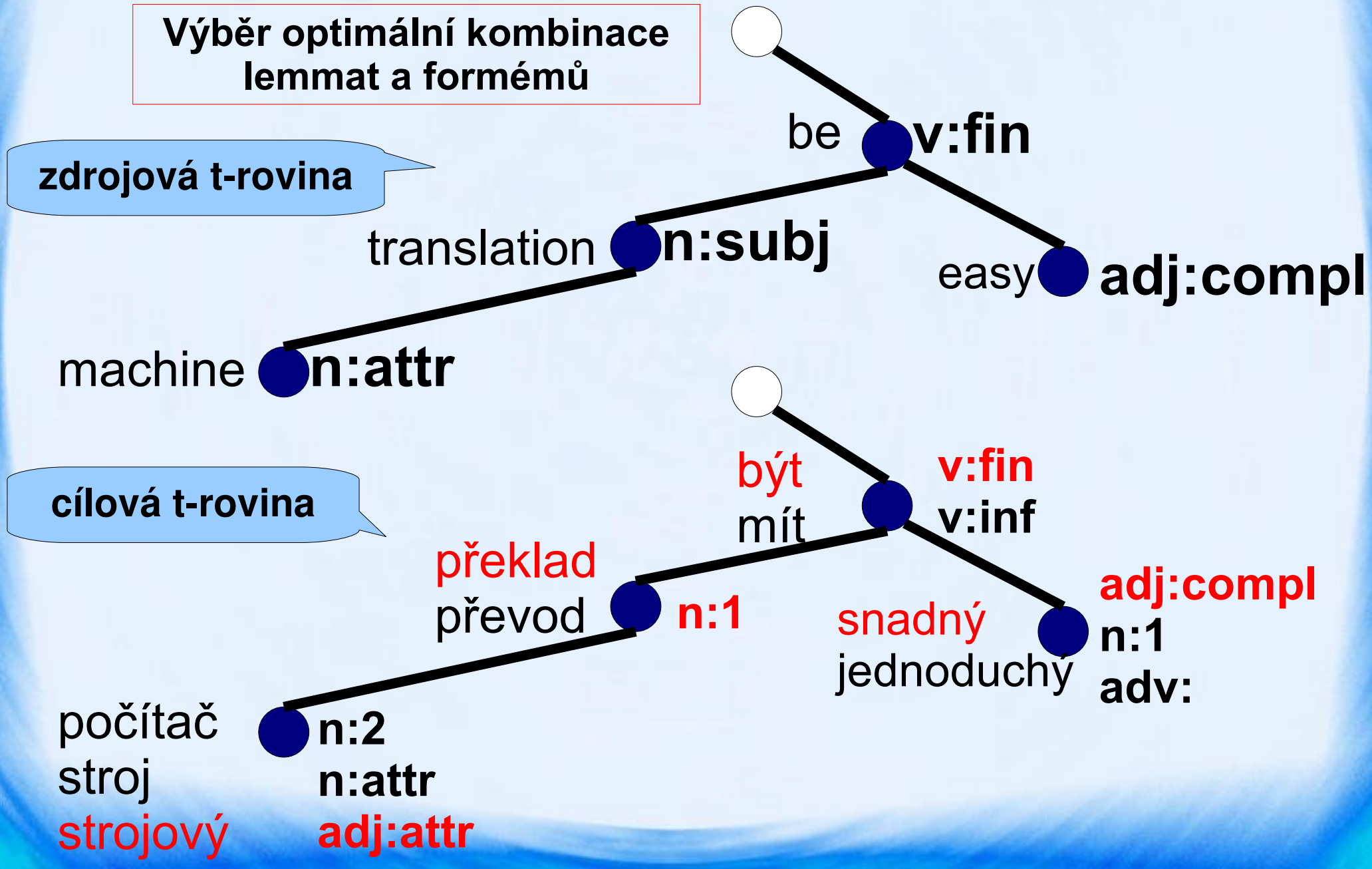
Ukázka překladu – Transfer

Vyplnění překladových variant lemmat a formémů



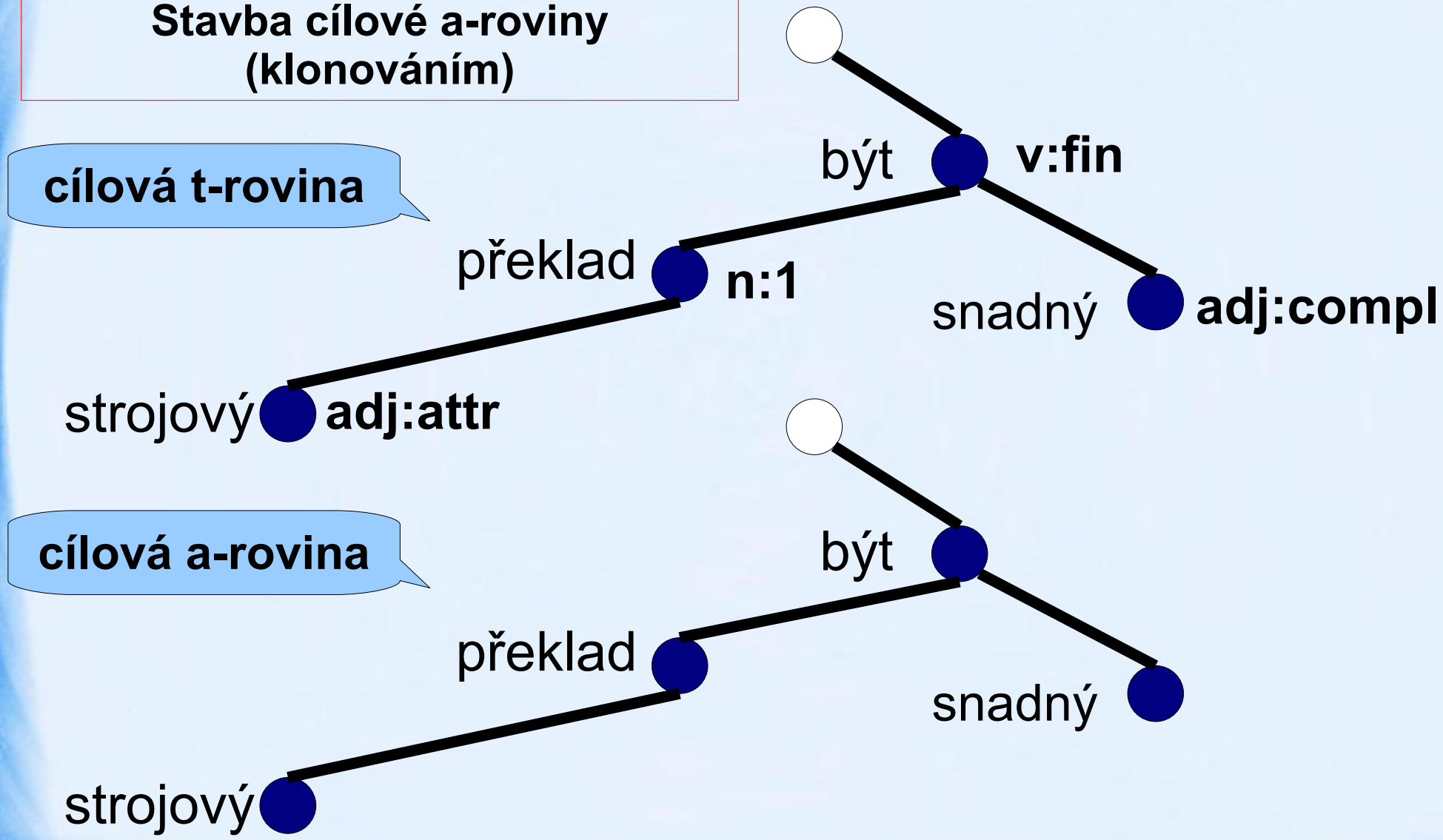
Ukázka překladu – Transfer

Výběr optimální kombinace lemmat a formémů



Ukázka překladu – Syntéza

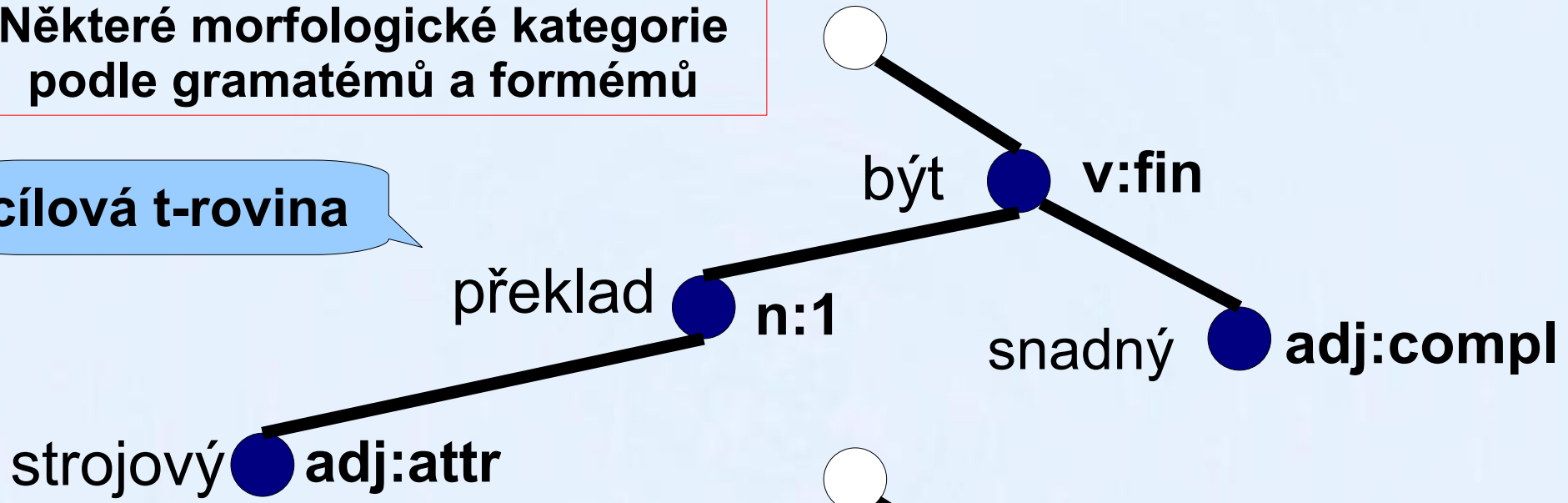
Stavba cílové a-roviny (klonováním)



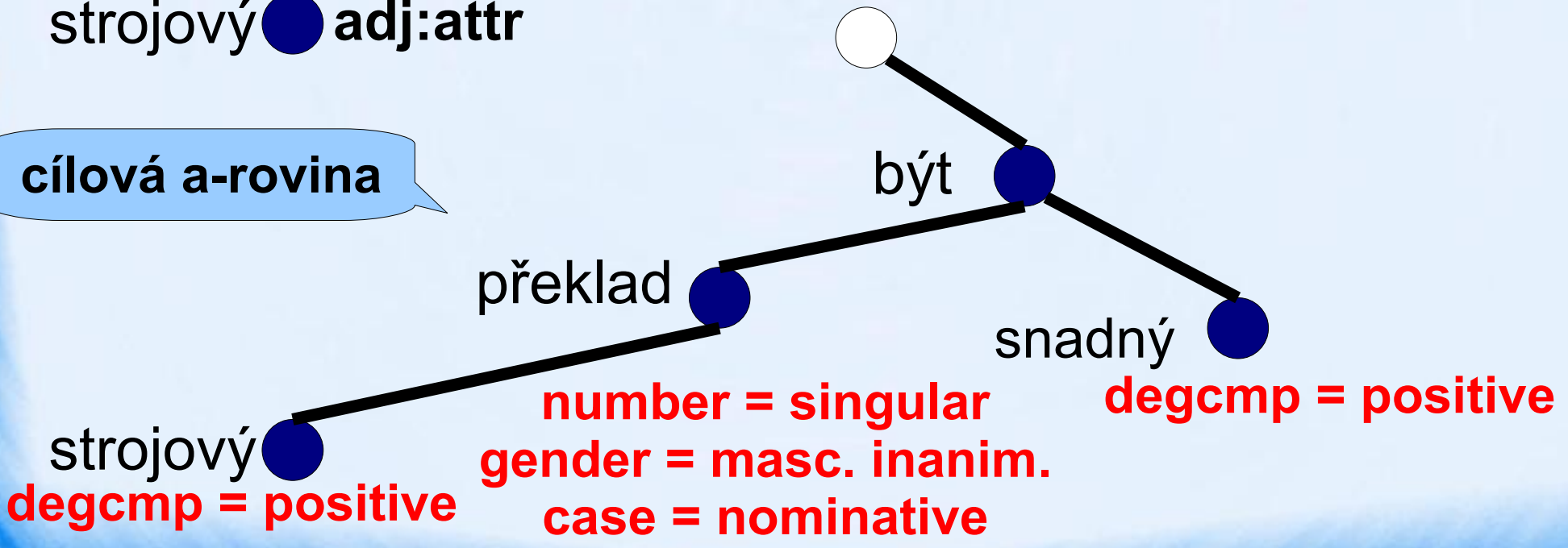
Ukázka překladu – Syntéza

Některé morfologické kategorie podle gramatémů a formémů

cílová t-rovina

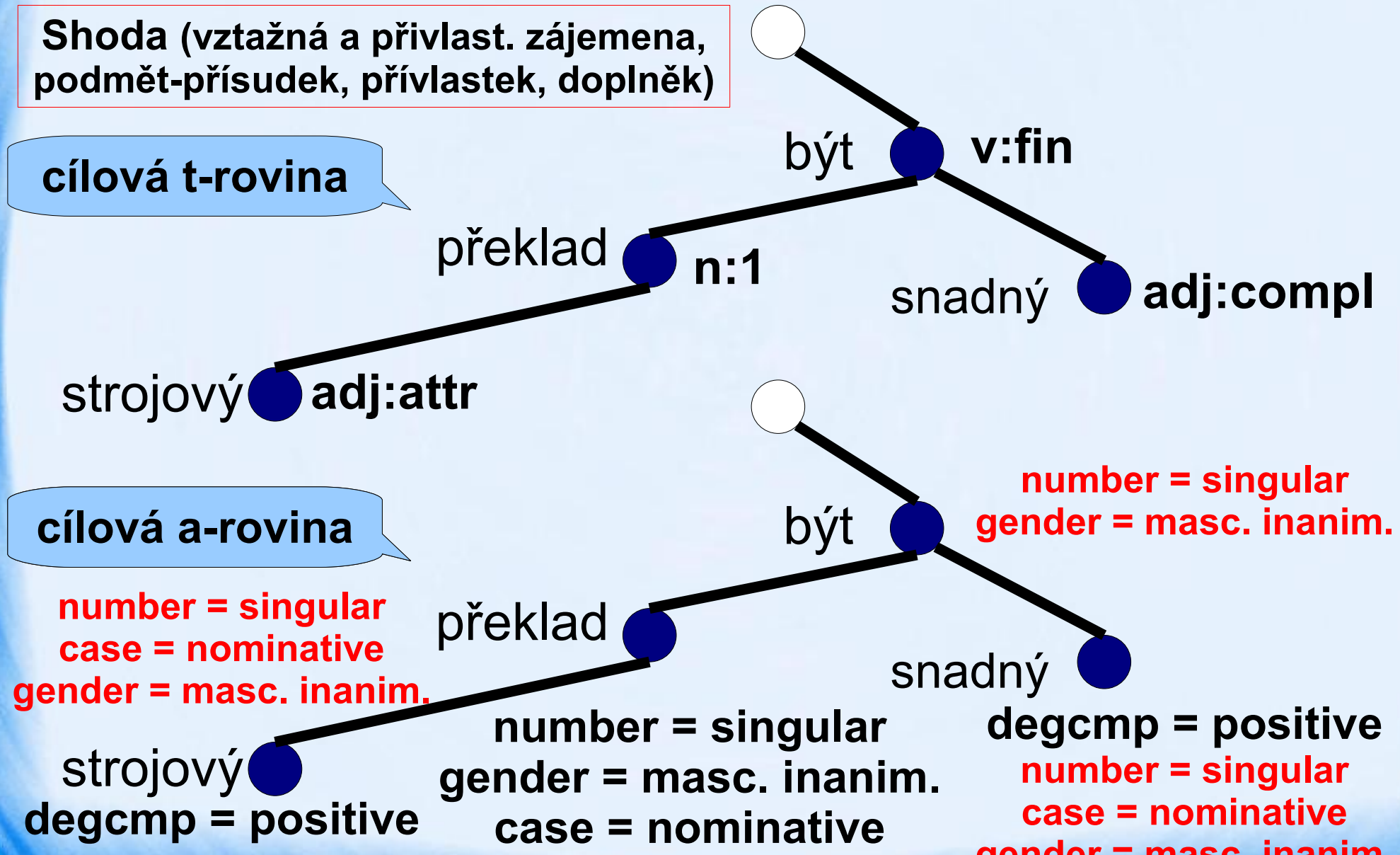


cílová a-rovina



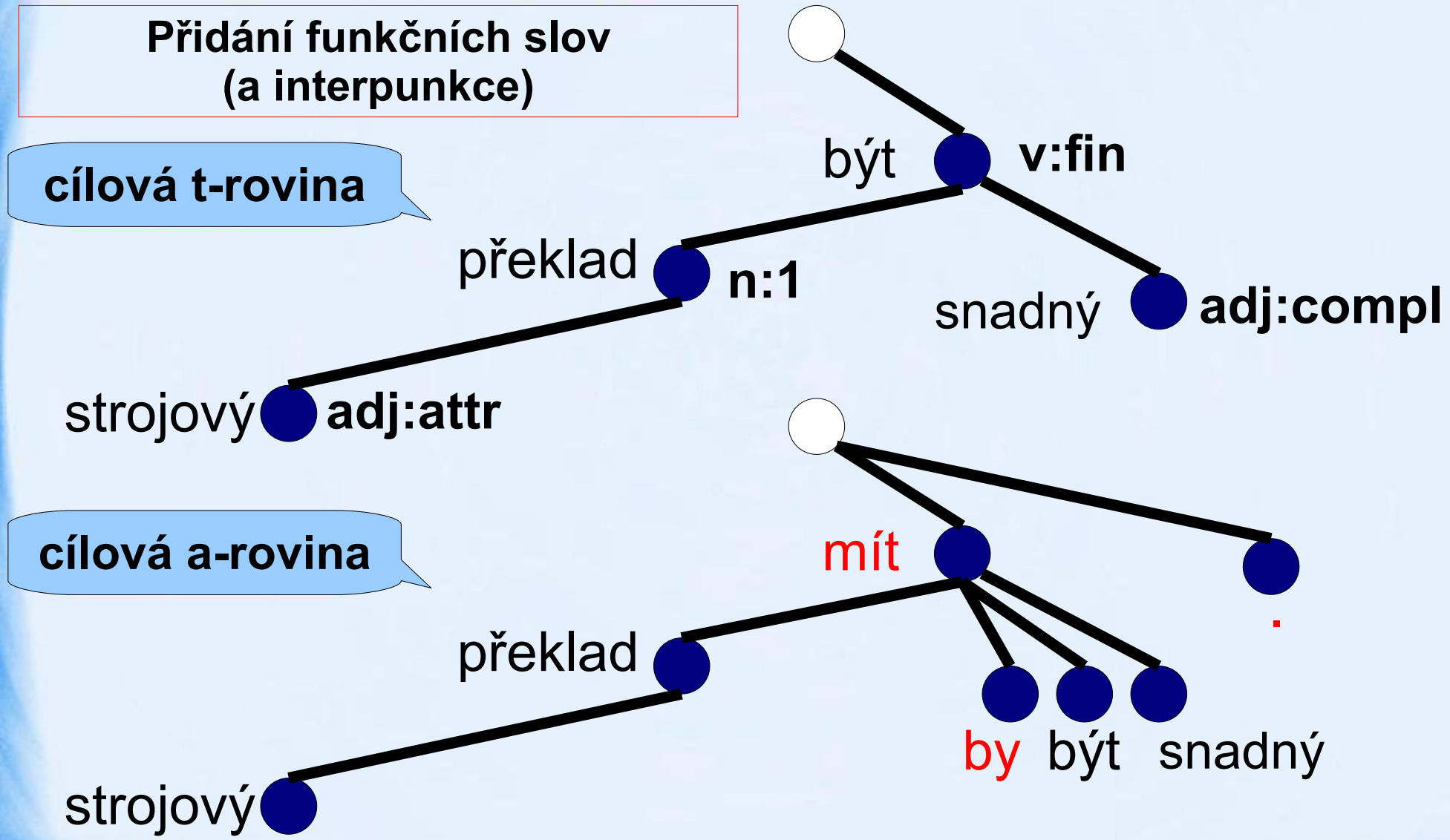
Ukázka překladu – Syntéza

Shoda (vztažná a přivlast. zájemena, podmět-přísudek, přívlastek, doplněk)



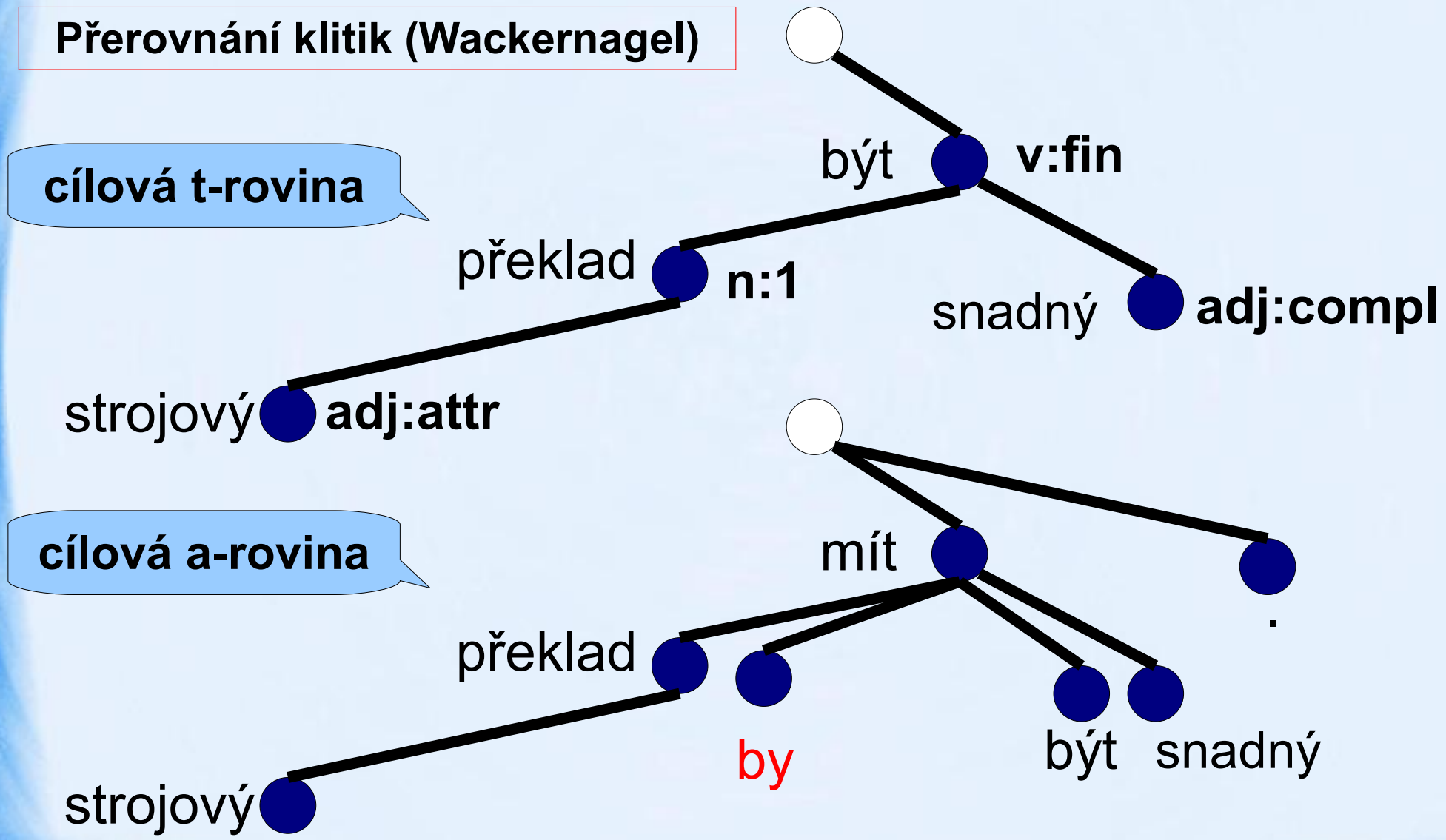
Ukázka překladu – Syntéza

Přidání funkčních slov (a interpunkce)

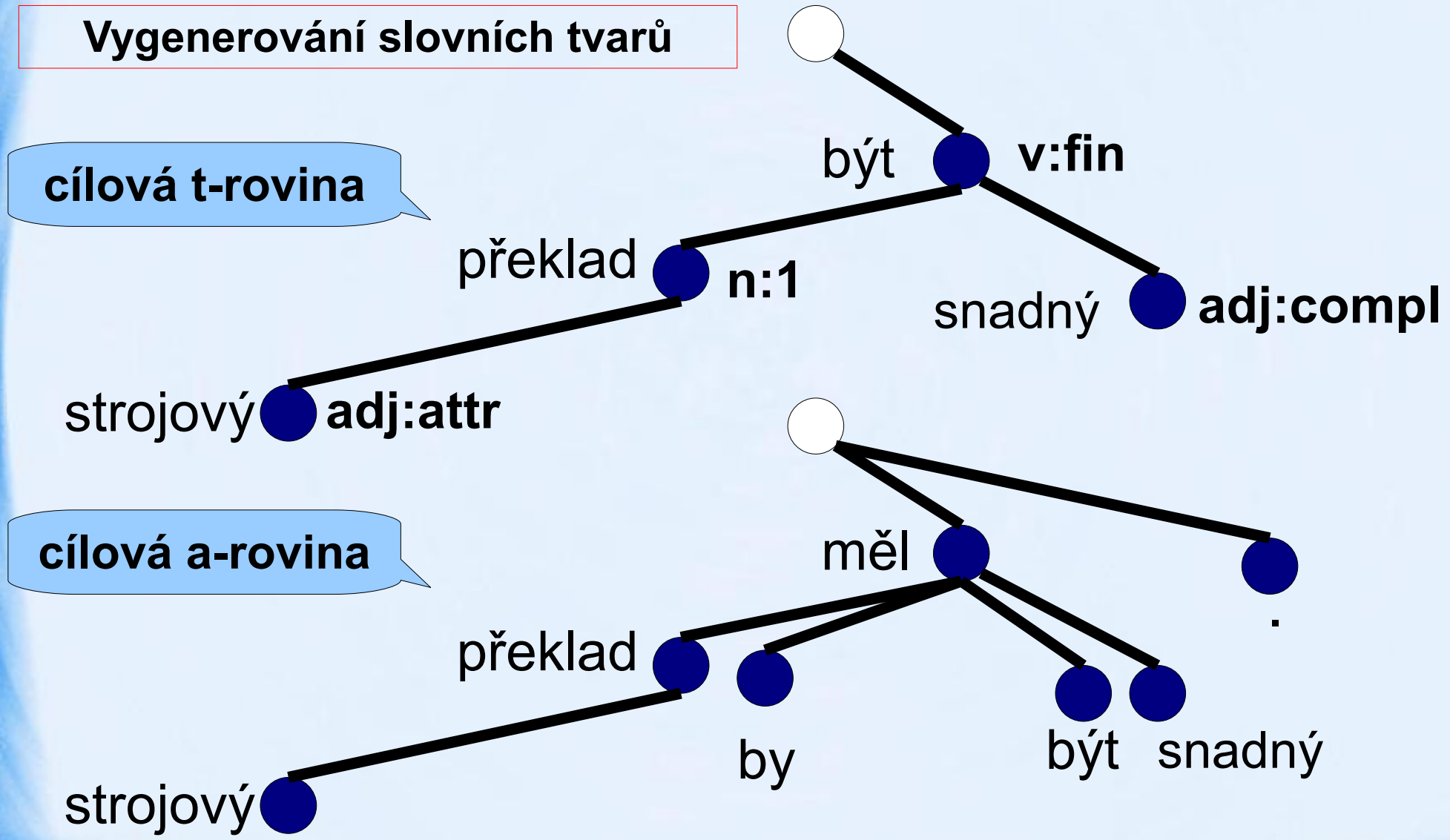


Ukázka překladu – Syntéza

Přerovnání klitik (Wackernagel)

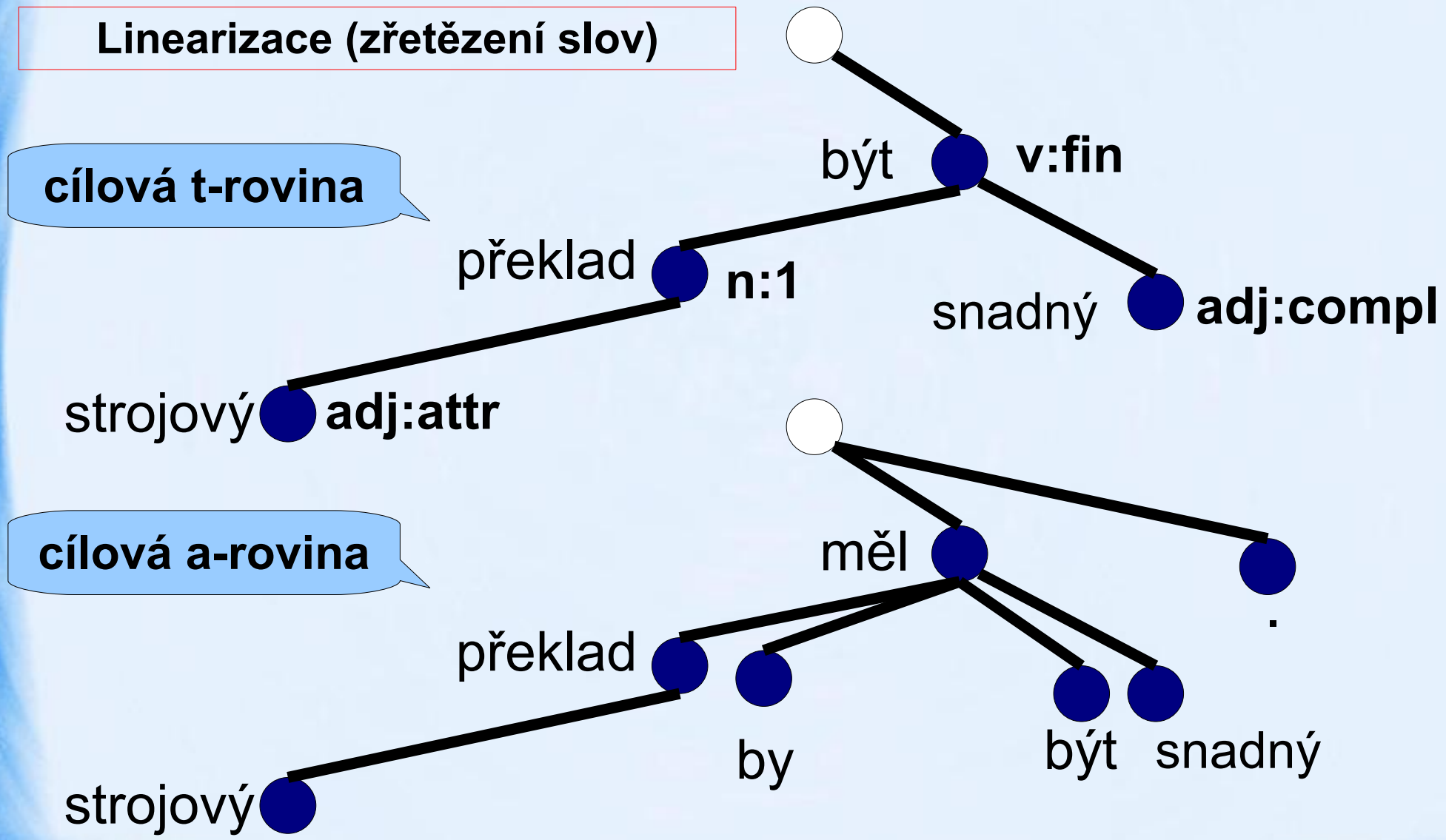


Ukázka překladu – Syntéza



Ukázka překladu – Syntéza

Linearizace (zřetězení slov)



Strojový překlad by měl být snadný.

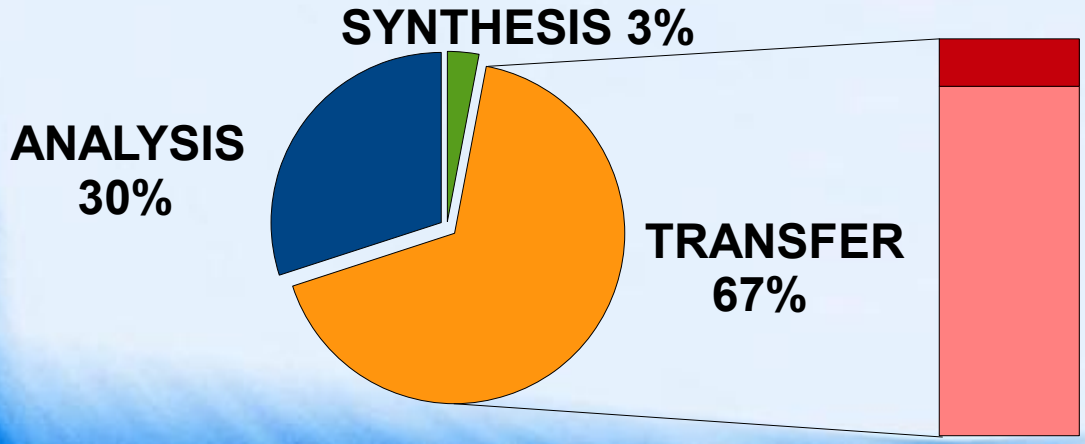
Ukázka překladu – Skutečný scénář

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

Anotace překladových chyb

vzorek 250 vět, celkem 1463 označených chyb

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, non-iso, syn, ?



chyby způsobené předpokladem Izomorfismu t-stromů 8%

ostatní chyby v transferu 59%

Detaily viz [Popel,2009].

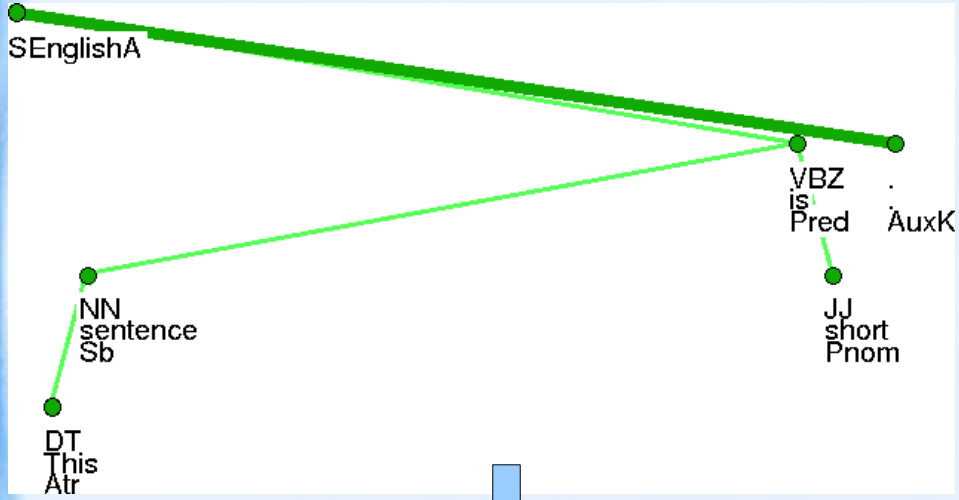
Novinky v TectoMT – Analýza

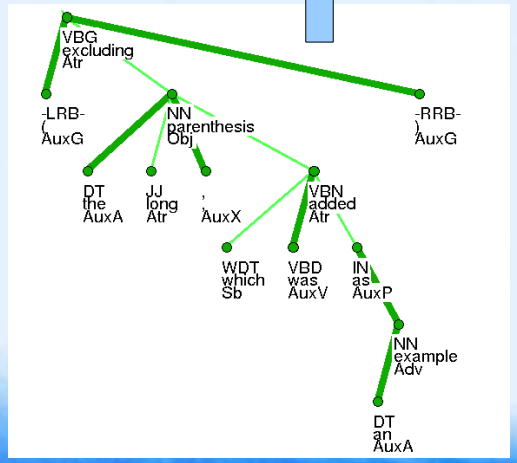
Analýza angličtiny

- Lemmatizace (70krát zrychlena)
- Parsing – pravidlové opravy
  – oddělený parsing parentezí v závorkách

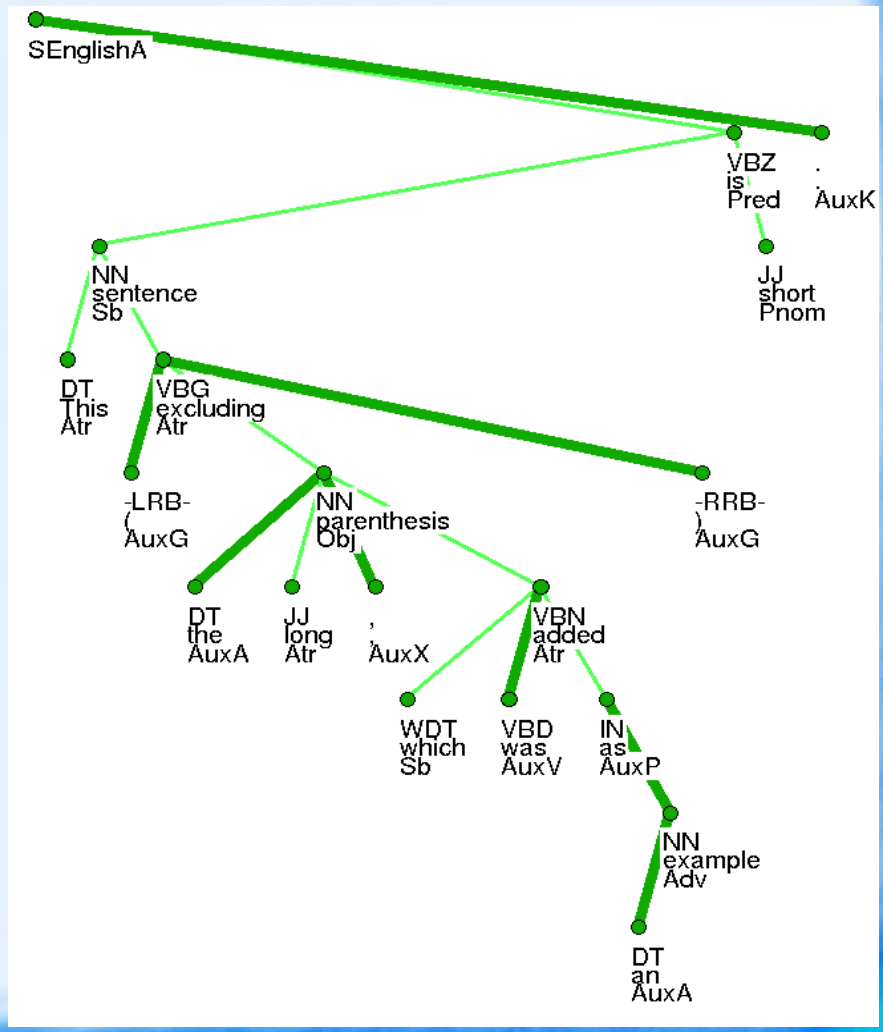
Parsing parentezí

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





Rozdíl:
0,3 bodu BLEU



Novinky v TectoMT – Analýza

Analýza angličtiny

- Lemmatizace (70krát zrychlena)
- Parsing – pravidlové opravy
  – oddělený parsing parentezí v závorkách
- Analytické funkce (pravidlový blok, chybí manuál)

I do AuxV not Neg want to AuxV make up AuxV an AuxA example.

Novinky v TectoMT – Analýza

Analýza angličtiny

- Lemmatizace (70krát zrychlena)
- Parsing – pravidlové opravy
  – oddělený parsing parentezí v závorkách
- Analytické funkce (pravidlový blok, chybí manuál)
- Budování t-roviny – vydělena jazykově nezávislá část
- Pojmenované entity ve zvláštním stromě
- Rozpoznávání ženských a mužských jmen
- Koreference

Novinky v TectoMT - Transfer

- nové slovníky (Maximum Entropy)
- Hidden Markov Tree Models (HMTM)
- časté fráze (neizomorfní t-stromy), např.

  take place → konat_se, proběhnout

  prime minister → premiér
- přechylování ženských příjmení
- pravidla pro slovesný vid, číslovky,...

Novinky v TectoMT - Syntéza

- Upraveno dělení věty na klauze, vkládání interpunkce, přesun klitik
- Přidán morfologický model (trénován na SYNu)
  - nalezení slovního tvaru pro dané lemma s daným omezením na tag
  - některé pozice tagu po překladu neznáme, netřeba je specifikovat, vybere se nejčastější tvar
- Potíže s morfologií omezeny, byť ne zcela

Slovníky - MaxEnt

- Slovník natrénován na paralelním korpusu CzEng pomocí metody Maximum Entropy
- Pro slovník lemmat použit kontext (features):

pro daný uzel a jeho rodiče:

tlemma, formeme, voice, negation, tense, number, degcmp, sempos, short_sempos, person, is_capitalized

pro daný uzel:

position (před/za rodičem), is_member, tag, has_left_child, has_right_child, prev_node_tlemma, next_node_tlemma, child_formem_*, child_tlemma_*, determiner (a/the)

Slovníky – Nové rozhraní

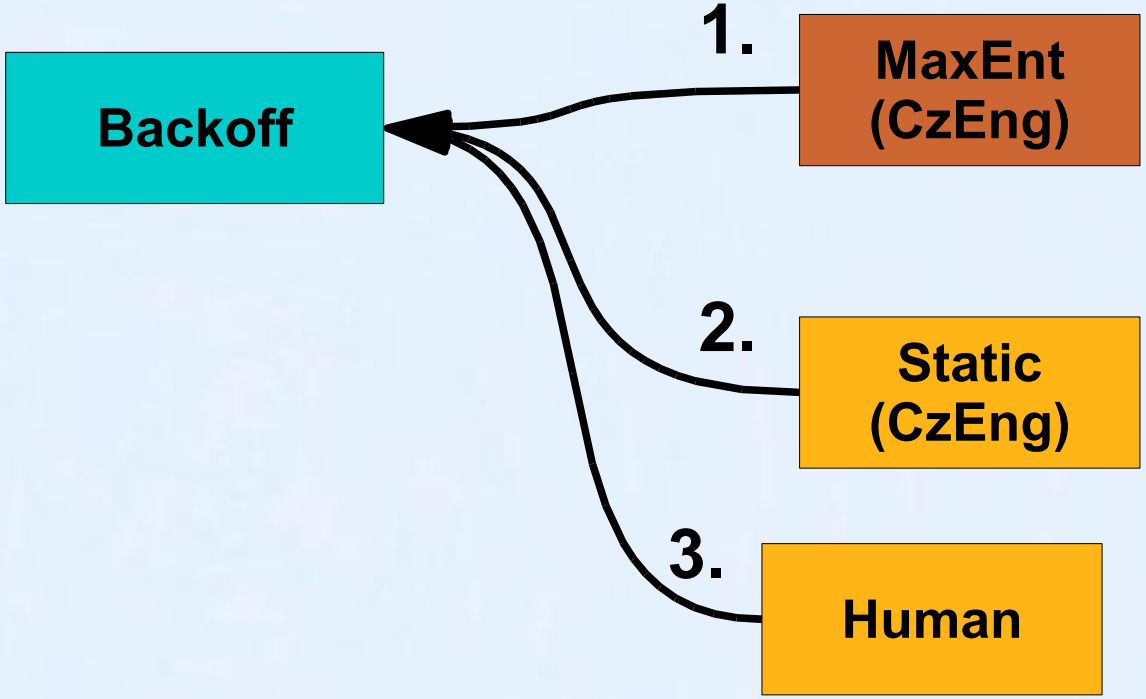
- obecné – totéž rozhraní pro lemmata i formémy

  `$dict->get_translations($input_label, $features)`

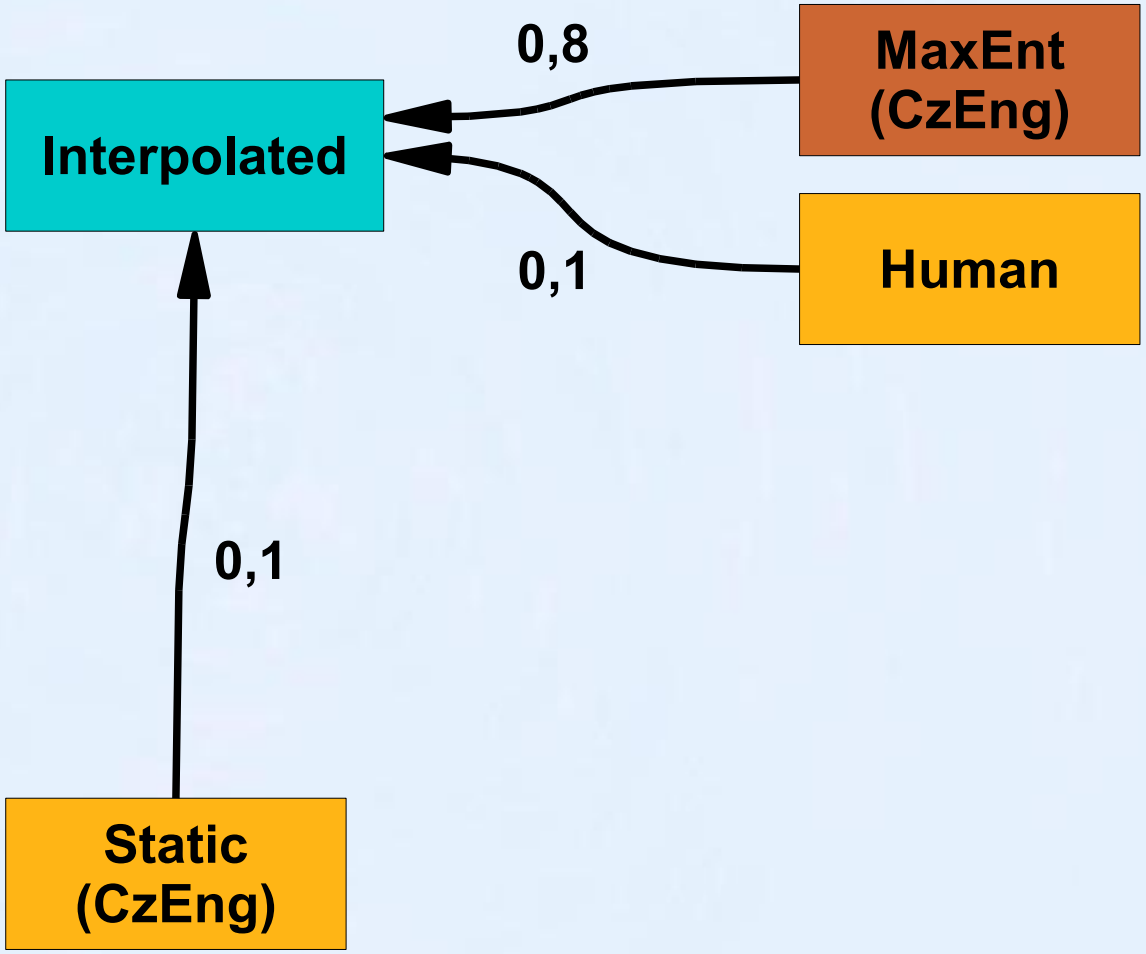
  vrátí seznam překladových variant včetně pravděpodobnosti
- Slovníky jsou objekty, v konstruktoru lze zadat jeden či více jiných slovníků – hierachie
- Základní typy slovníků:

Statický	data ze souboru, „lemma → lemma“
Kontextový	data ze souboru, „lemma,features → lemma“
Derivační	překlady odvozeny dynamicky, vstupní slovník
Kombinační	kombinace více vstupních slovníků

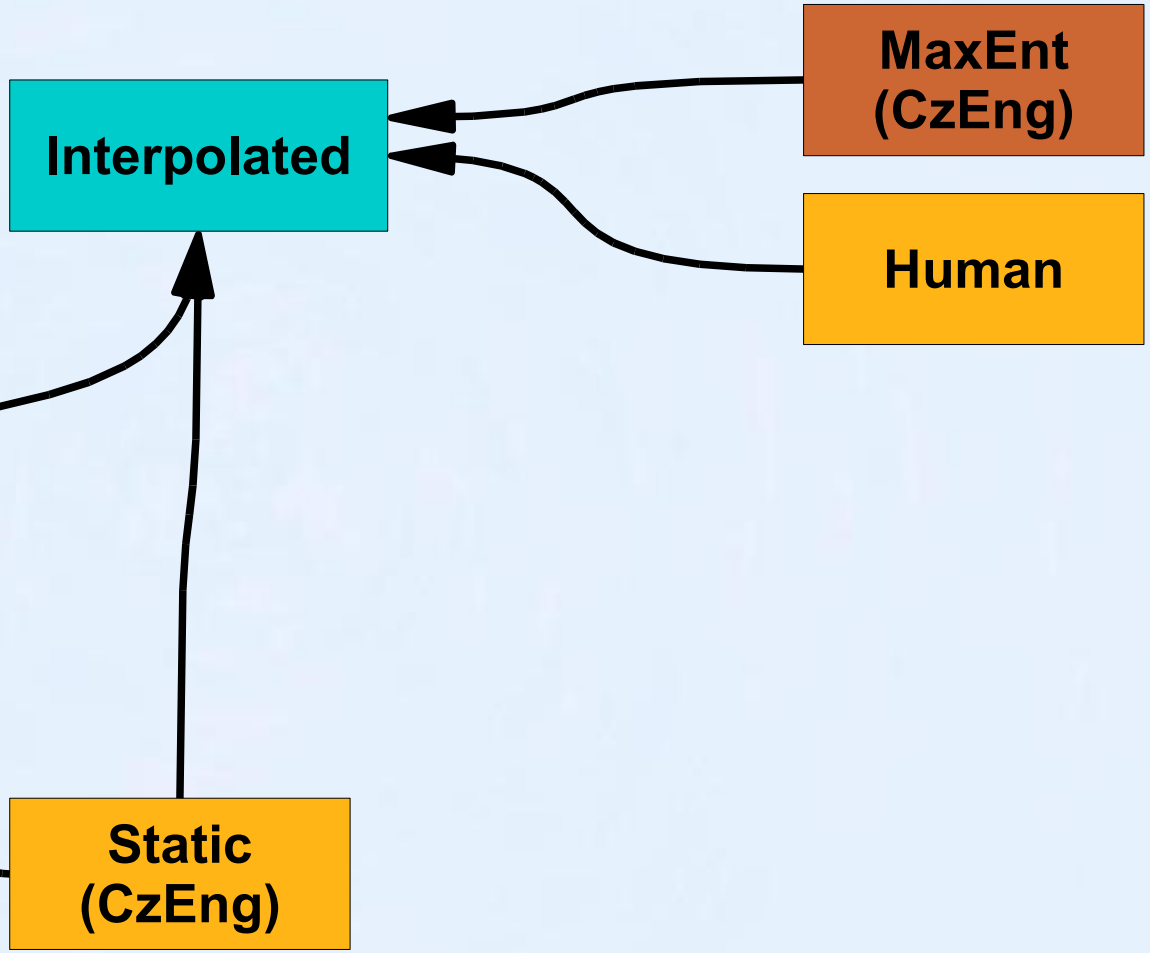
Slovníky – Hierarchie (lemmata)



Slovníky – Hierarchie (lemmata)

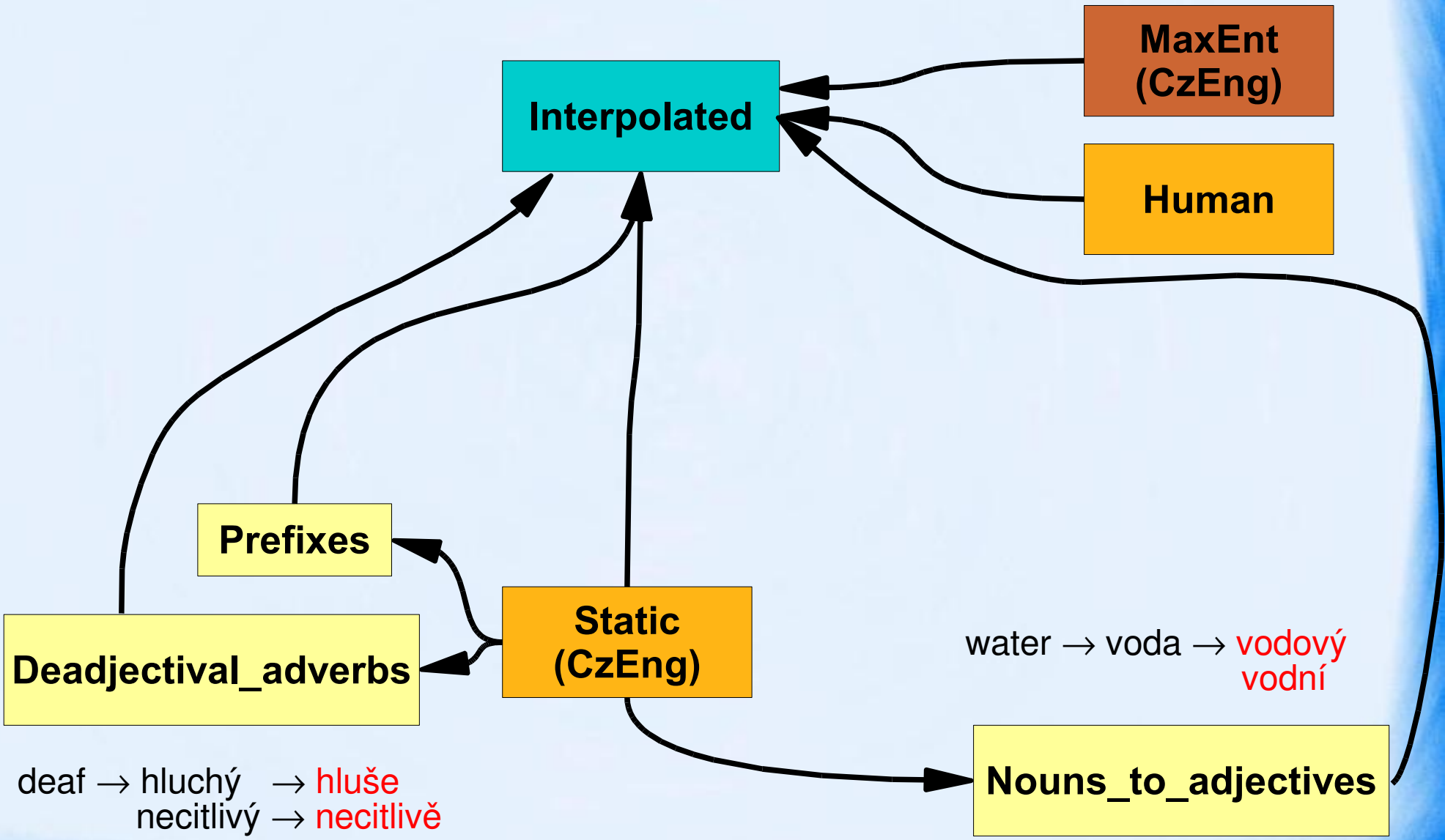


Slovníky – Hierarchie (lemmata)

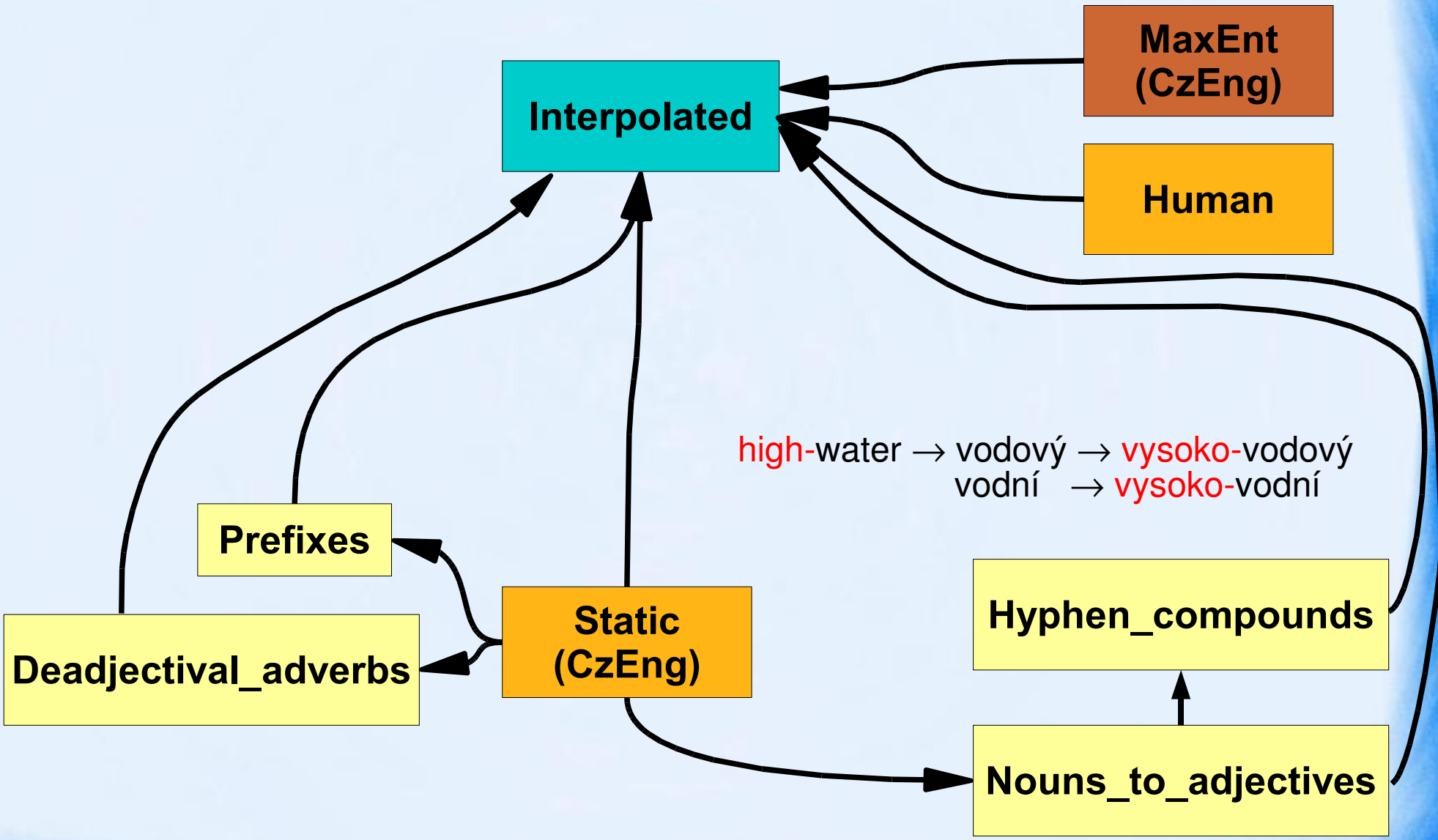


multi-core → více-jádrový
více-jádro
multi-jádrový
multi-jádro

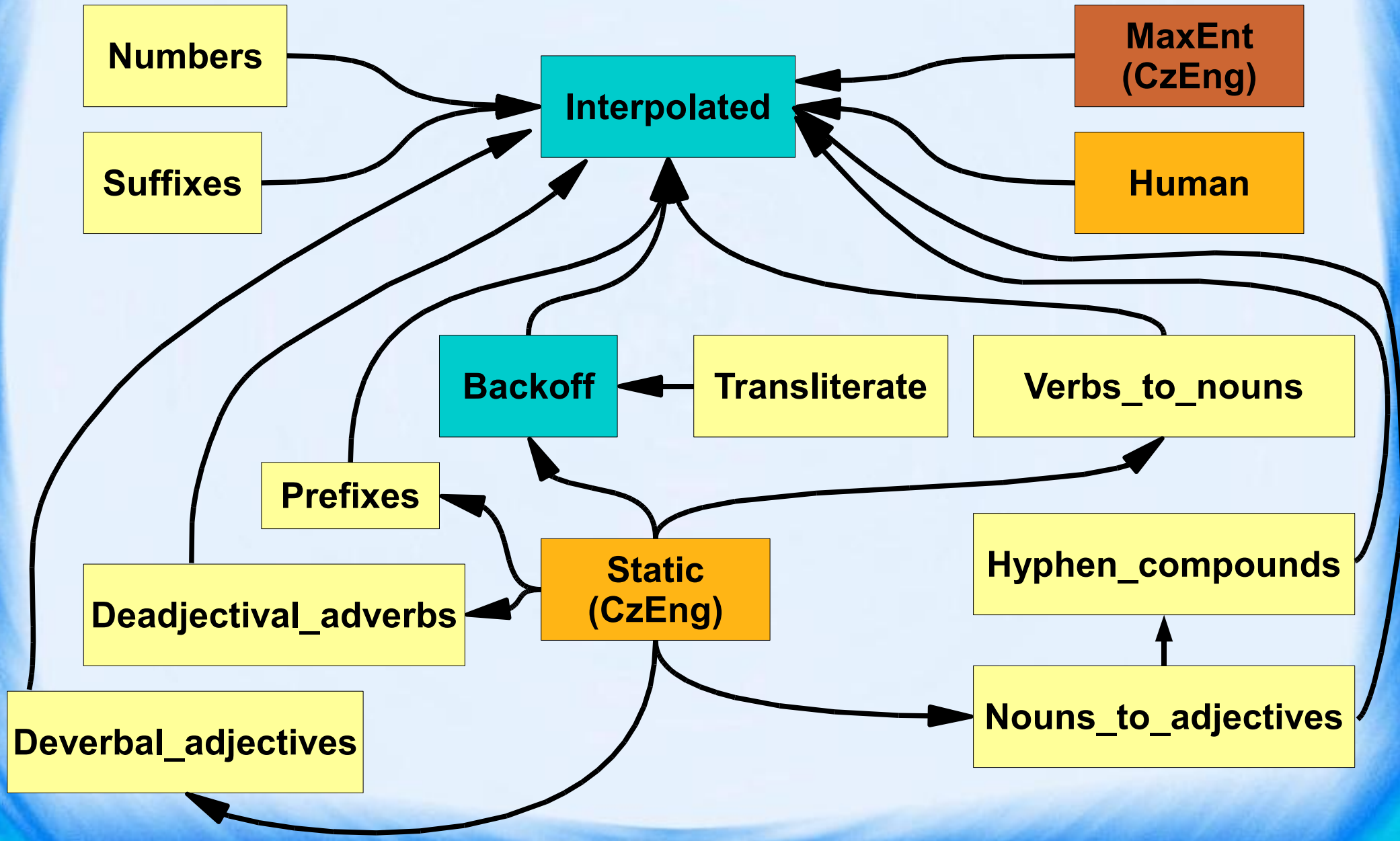
Slovníky – Hierarchie (lemmata)



Slovníky – Hierarchie (lemmata)

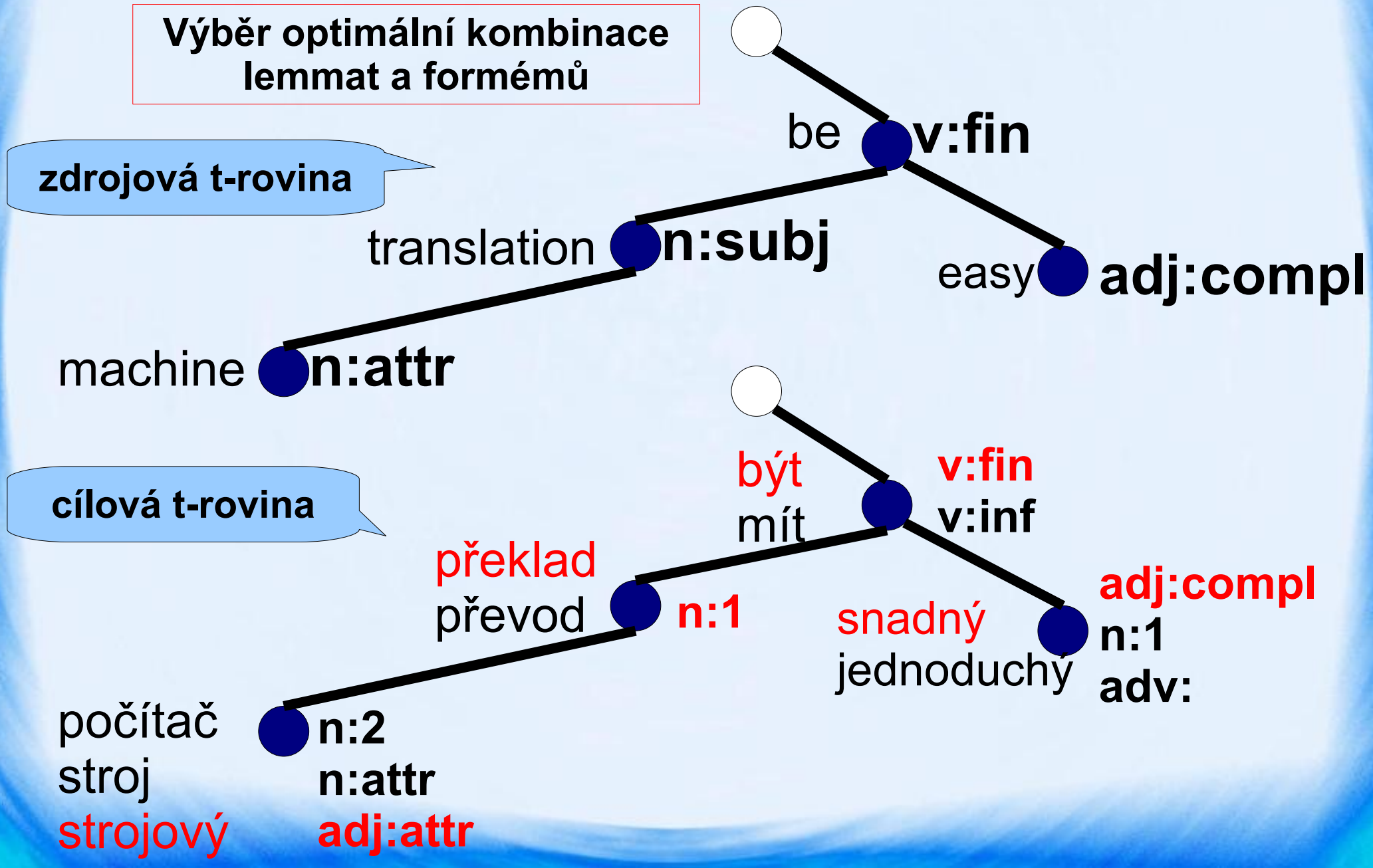


Slovníky – Hierarchie (lemmata)



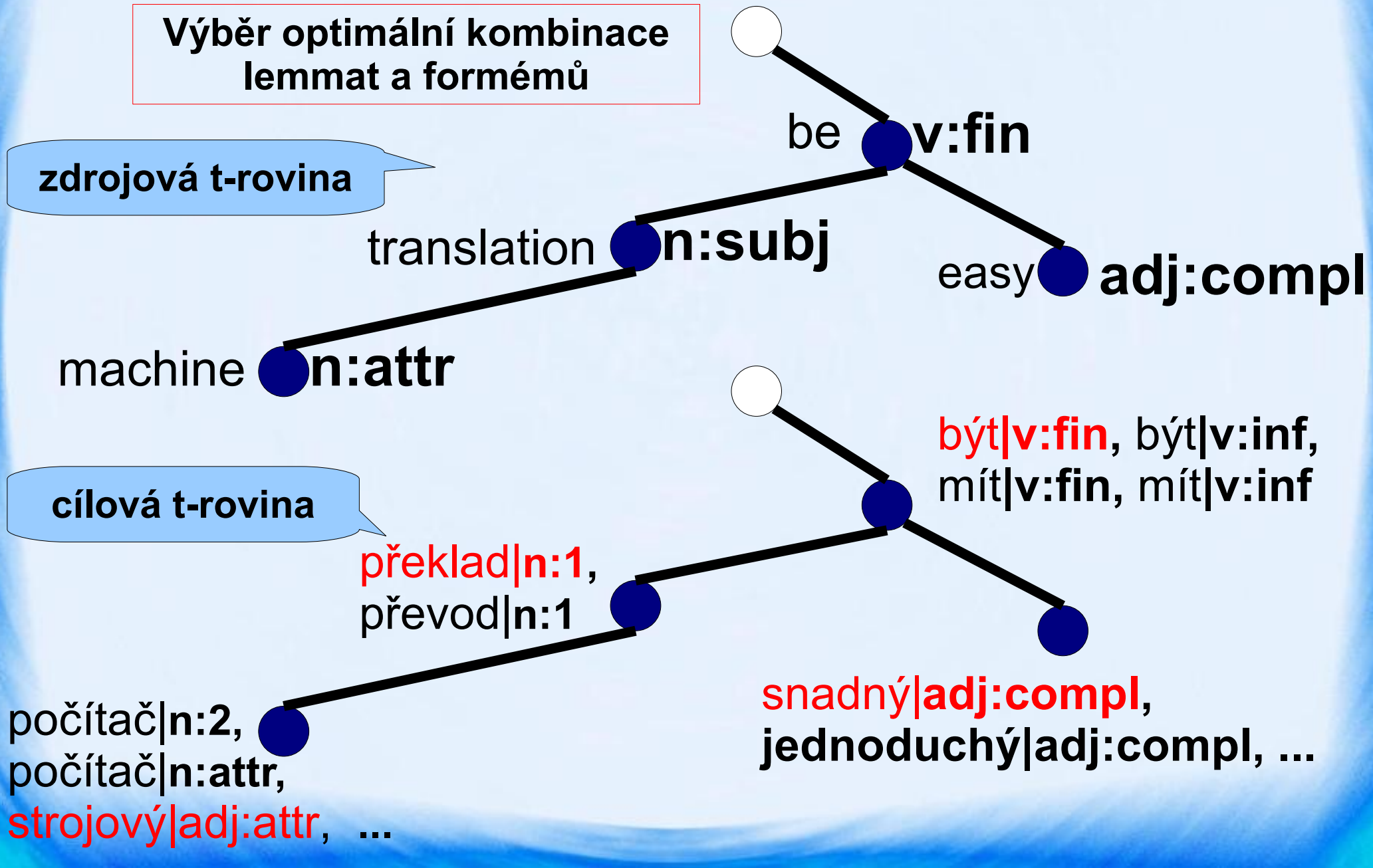
HMTM – Motivace

Výběr optimální kombinace lemmat a formémů



HMTM – Motivace

Výběr optimální kombinace lemmat a formémů

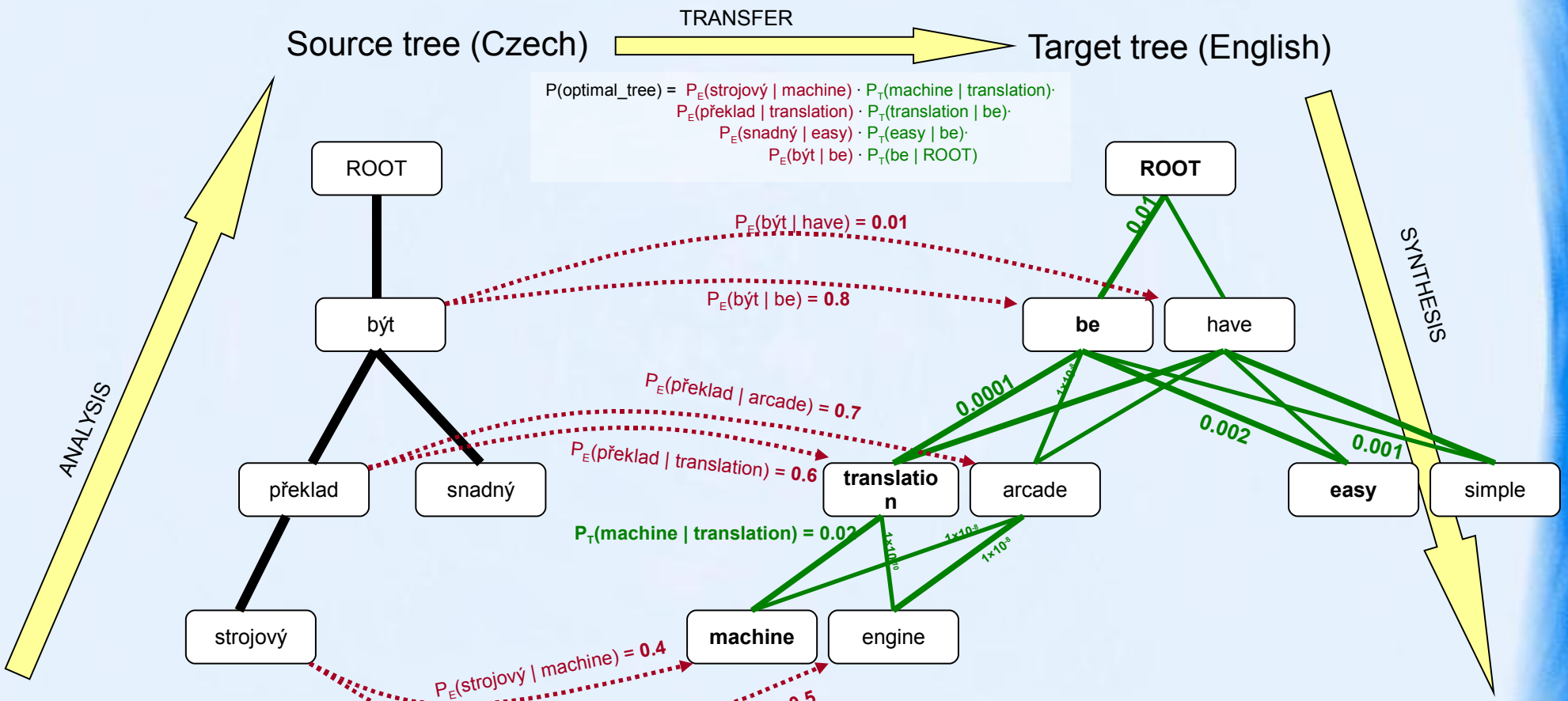


HMTM - Teorie

- HMTM zavedl [Crouse,1998], používáno pro signal processing segmentaci obrazu apod., viz [Durand,2004].
- (V,E) – zakořeněný strom
- **X** – sekvence náhodných proměnných (skryté stavy vrcholů V)
- **Y** – sekvence náhodných proměnných (viditelné symboly)
- $P(X_v \mid X_{rodič(v)})$ – přechodová pravděpodobnost (transition prob.)
- $P(Y_v \mid X_v)$ – emisní pravděpodobnost (emission prob.)
- Stromová Markovova vlastnost (podmínka nezávislosti):
  $\forall v \in V \setminus \{kořen\}, \forall w \in V \setminus podstrom(v)$:

  $$P(\mathbf{X}_{podstrom(v)} \mid X_{rodič(v)}, X_w) = P(\mathbf{X}_{podstrom(v)} \mid X_{rodič(v)})$$
- Známe-li **Y**, můžeme najít nejpravděpodobnější sekvenci skrytých stavů pomocí stromového Viterbiho algoritmu.

HMTM – v překladu

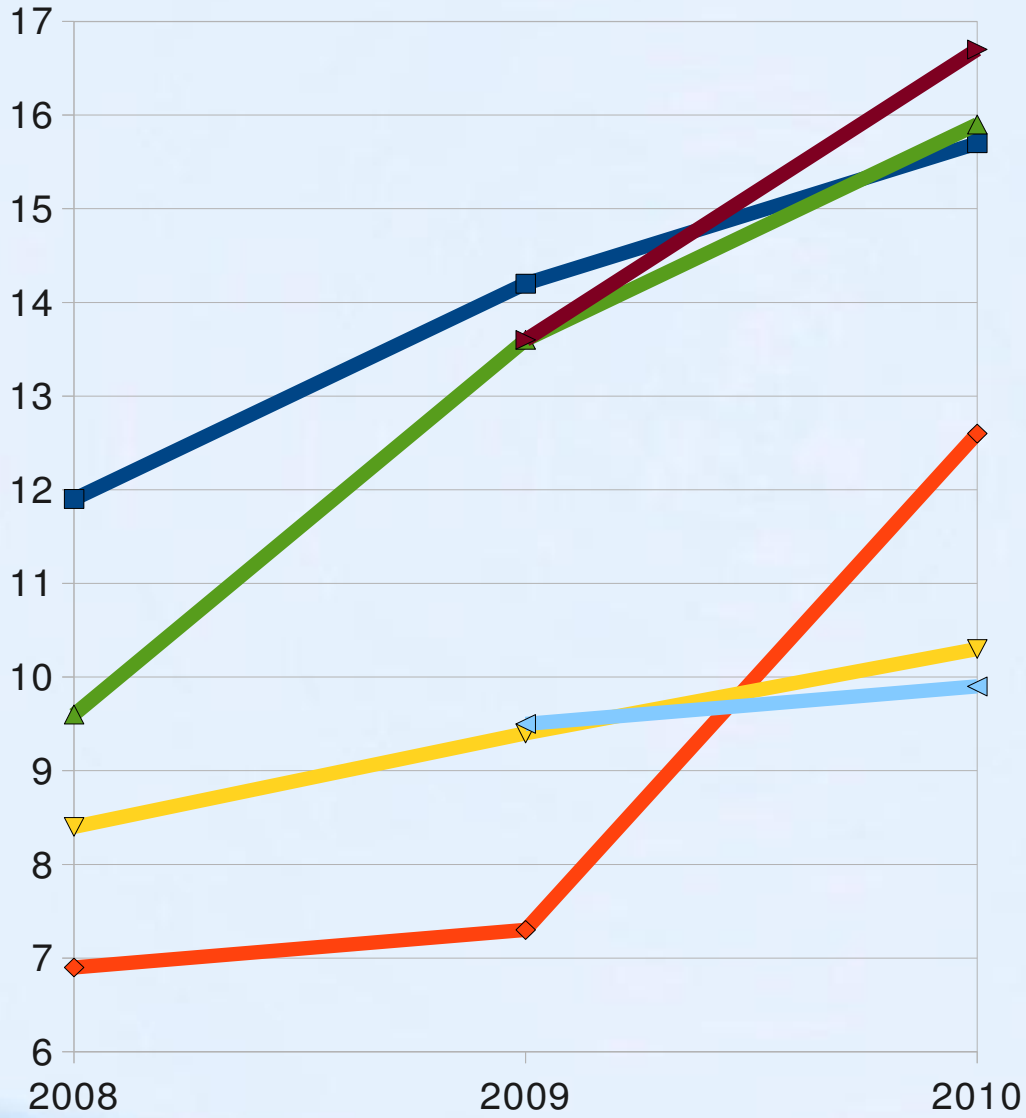


Source sentence:
Strojový překlad by měl být snadný.

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**

Výsledky – WMT (BLEU)



TectoMT:

Rok	BLEU
2008 WMT	6,9
2009 WMT	7,3
2009 září	10,2
2010 leden	10,4
2010 únor	11,3
2010 WMT	12,6

TectoMT před 3 lety

Translation sample

- Original text in the source language:
  - *A Turkish girl has died from bird flu, days after her brother and sister died from the disease. The girl, 11, who lived on a poultry farm in eastern Turkey's Van province, was being treated in hospital after her siblings became infected with bird flu. The cases are the first human deaths from bird flu outside Asia, where the virus has killed more than 70 people. The hospital in Van is treating 15 others, three of whom are in a critical condition, according to a doctor there. The latest victim, Hulya Kocyigit, died early on Friday at the hospital.*

- Automatically translated text in the target language:
  - *Turecká dívka zemřela z ptačí chřipky dny after, že její bratr a sestra zemřeli z nemoci. Dívka 11, kdo žilo v drůbeží farmě ve van provincii východního Turecka, jsoucno zacházet v nemocnici, že její sourozenci slušeli nakažený s ptačí chřipkou. Případy jsou přední lidské smrti z ptačí chřipky mimo Asii, kde virus zabilo than 70 lid. Nemocnice ve Van zachází 15 zbývajících, whom three of v kritické podmínce souzvuk lékaře tam. Nejpozdnější oběť Kocyigit Hulya zemřela brzy v pátku v nemocnici.*

TectoMT před 3 lety a dnes

SRC: A Turkish girl has died from bird flu, days after her brother and sister died from the disease.
2007: Turecká dívka zemřela z ptačí chřipky dny after, že její bratr a sestra zemřeli z nemoci.
2010: Turecká dívka zemřela ptačí chřipkou, dny, ona, bratr a sestra zemřela nemocí.

SRC: The latest victim, Hulya Kocyigit, died early on Friday at the hospital.
2007: Nejpozdnější oběť Kocyigit Hulya zemřela brzy v pátku v nemocnici.
2010: Poslední oběť Hulya Kocyigit zemřela brzy v pátek v nemocnici.

Ukázky překladu

Birds of a feather flock together.	Ptáci v bederním hejnu spolu.
Great talkers are little doers.	Velcí řečníci jsou malí vrazi.
As good be an addled egg as an idle bird.	Dobré je feťácké vejce jako činný pták.
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.
Bread is the staff of life.	Chléb je zaměstnanec života.
I'll come a bit later on my own.	Sem čelist ještě na své milé.

Literatura

- 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.