

TectoMT: Deep-Syntactic Machine Translation

Martin Popel

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



MT Marathon
September 10, 2015, Prague

Motivation

Motivation

TectoMT was one of the **worst** systems in WMT2015.

TectoMT is **difficult** to train for a new language pair.

Motivation

TectoMT was one of the **worst** systems in WMT2015.

TectoMT is **difficult** to train for a new language pair.

Chimera **won** EN→CS WMT 2013–2015.

Chimera = Moses + TectoMT + Depfix



Motivation

TectoMT	✓			✓		✓		✓	53.5
Moses	✓		✓		✓			✓	57.7
Moses+TectoMT	✓		✓	✓			✓		59.2
% of reference tokens	44.7	32.9	8.6	4.5	3.6	3.5	1.4	0.8	100

adapted from [Tamchyna and Bojar, ACL 2015](#)

TectoMT translations are very different from phrase-based systems (8% of reference tokens are only in TectoMT).
TectoMT is **essential** for Chimera's success.

Motivation

TectoMT was one of the **worst** systems in WMT2015.

TectoMT is **difficult** to train for a new language pair.

BUT

TectoMT is **essential** for Chimera's **success**.

We have many plans for improving TectoMT.

TectoMT is now **less difficult** to adapt:
prototypes of CS→EN, EN↔ES, EN↔NL, EN↔PT, EN↔EU
created within the QTLep project.

Outline

- Treex (framework) vs. TectoMT (MT system)
- Demo translation step by step
- Annotation of translation errors
- Details
 - Hidden Markov Tree Models (HMTM)
 - Combining dictionaries
 - Maximum Entropy dictionary
- Examples of translation

Treex vs. TectoMT



2005

...

2011

NLP framework
TectoMT

MT system
TectoMT

lemmatization

tagging

parsing

Main author:
Zdeněk Žabokrtský

**multi-purpose
NLP framework**
Treex

MT system
TectoMT

lemmatization

tagging

parsing

coreference

PEDT preprocessing

CzEng analysis

treebank conversions

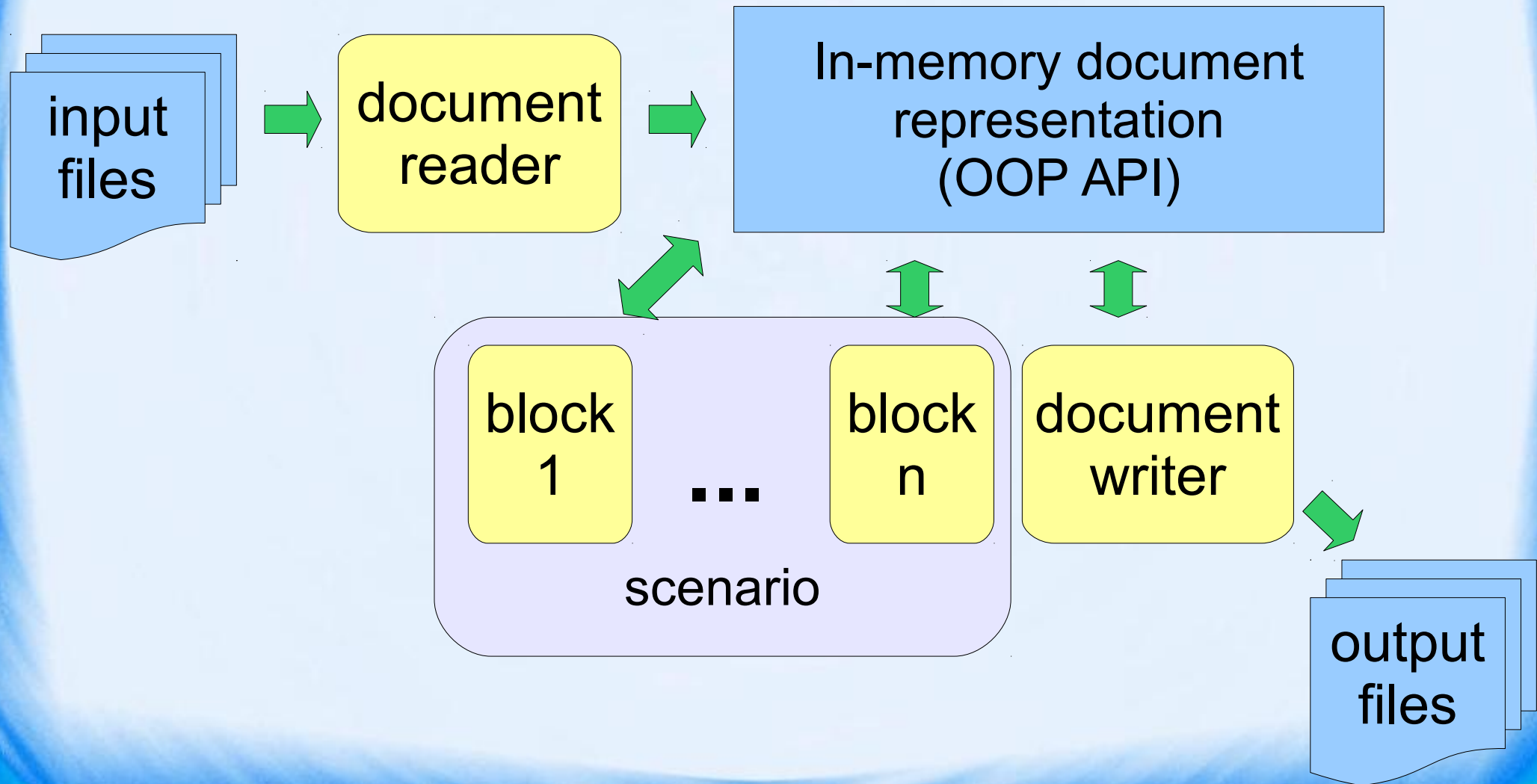
named entity r.

alignment (word,tree)

SMT preproc.

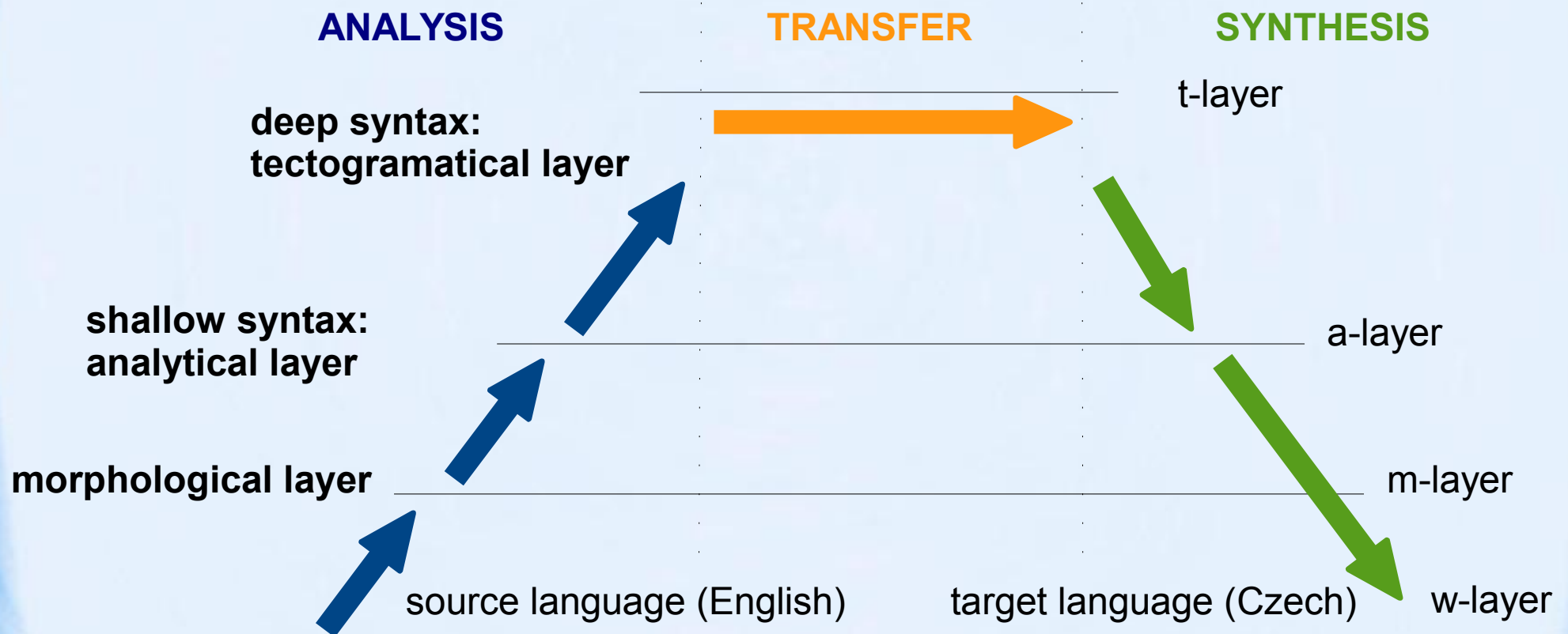
etc.

Treex architecture



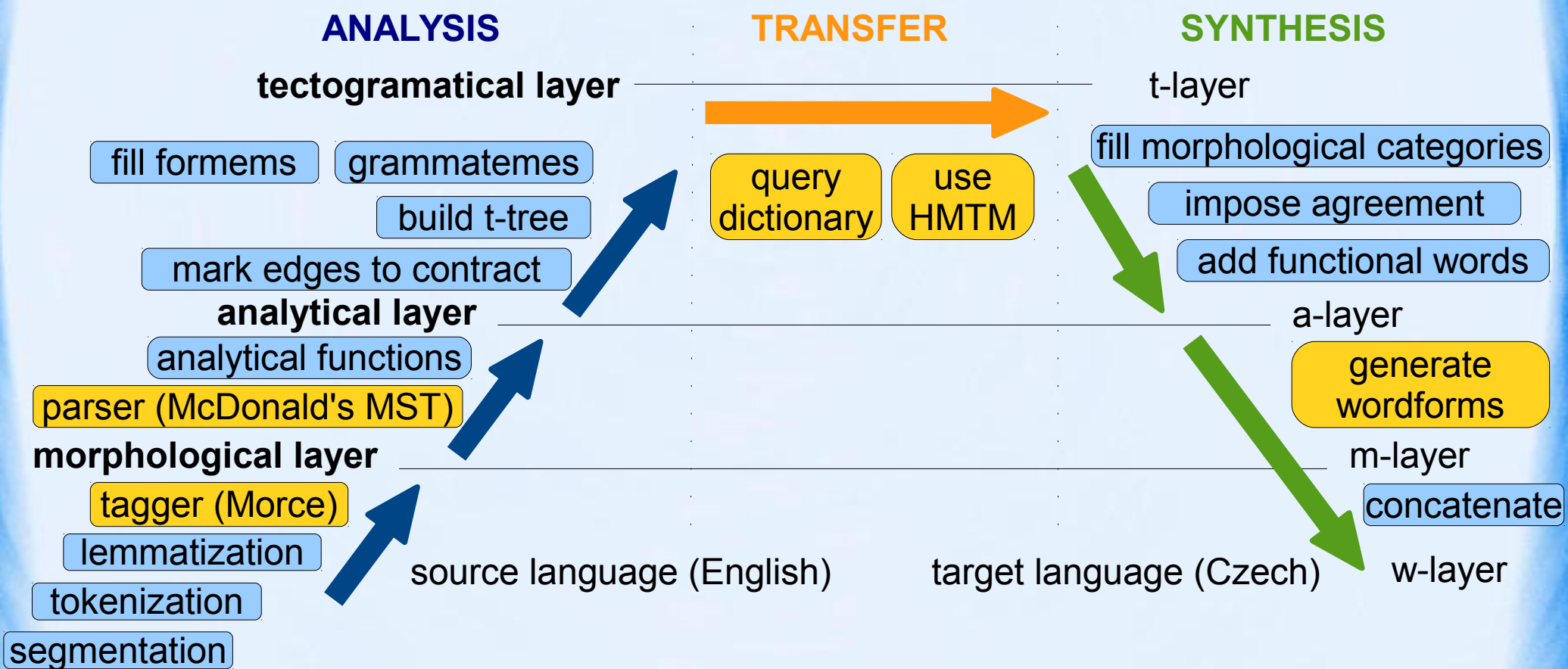
Translation scheme

transfer over the tectogrammatical layer



Translation scheme

rule based & statistical blocks



Demo Translation – Analysis

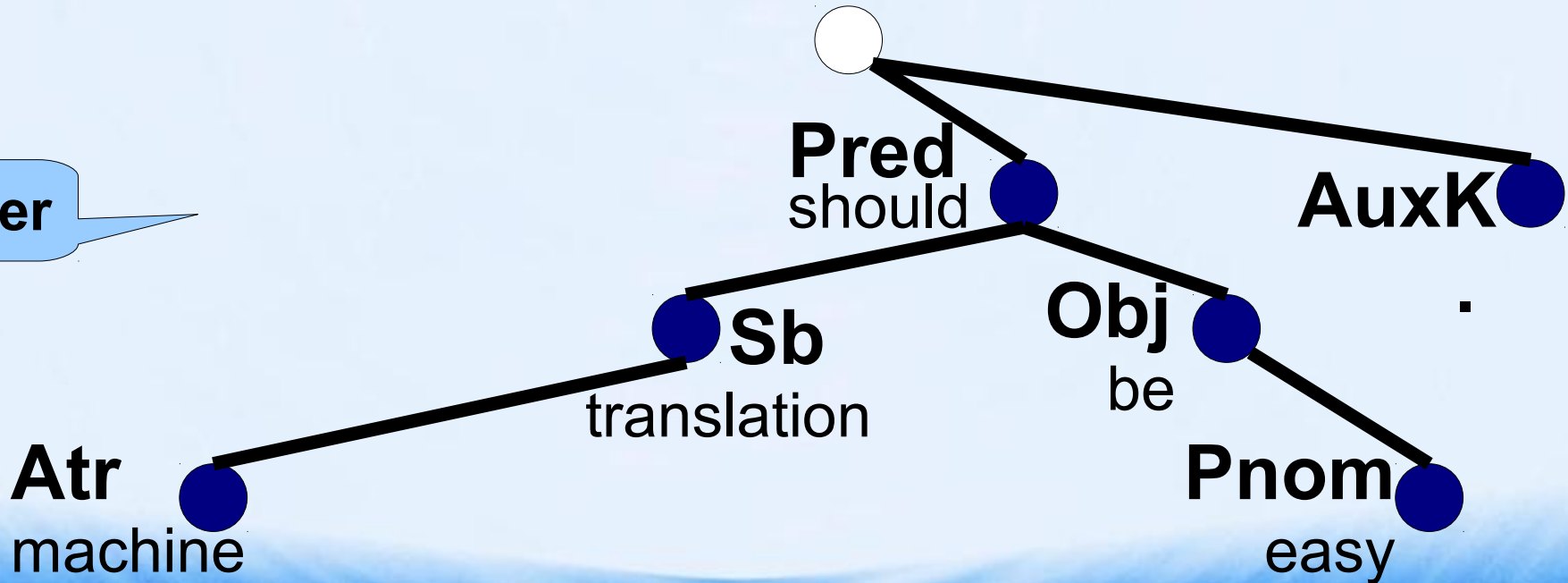
raw text

Machine translation should be easy.

m-layer

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

a-layer



Demo Translation – Analysis

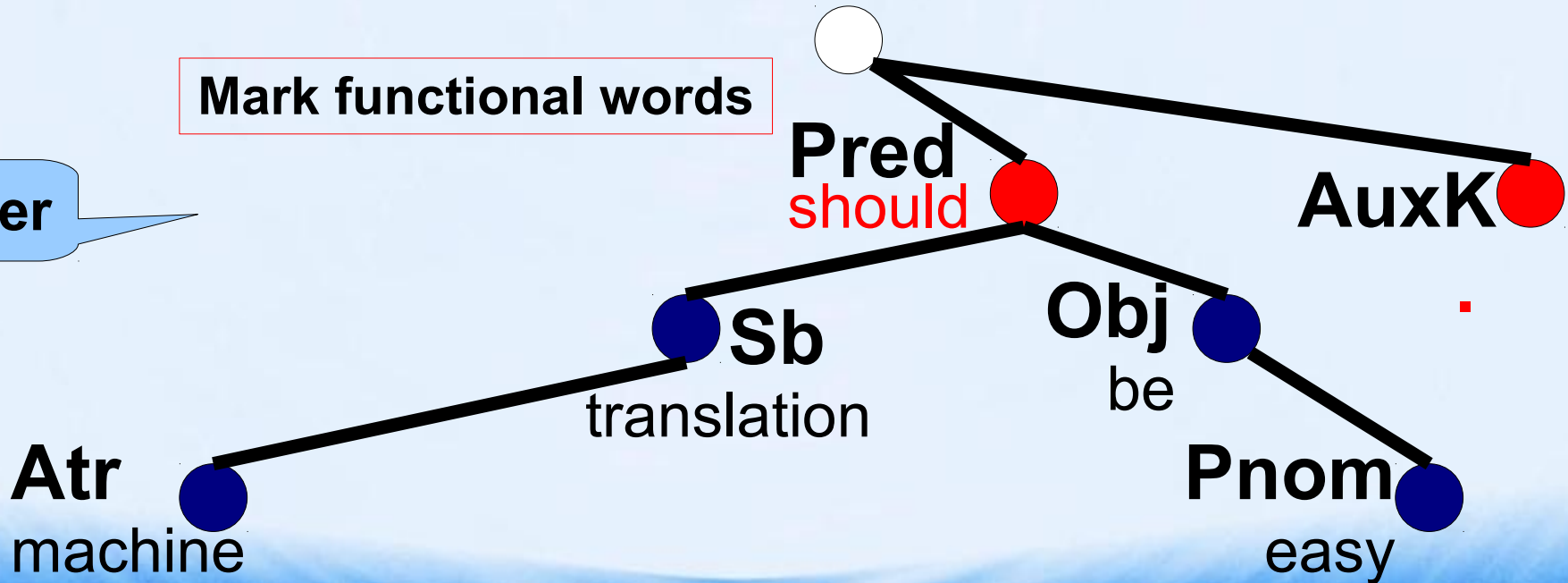
raw text

Machine translation should be easy.

m-layer

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

a-layer



Demo Translation – Analysis

raw text

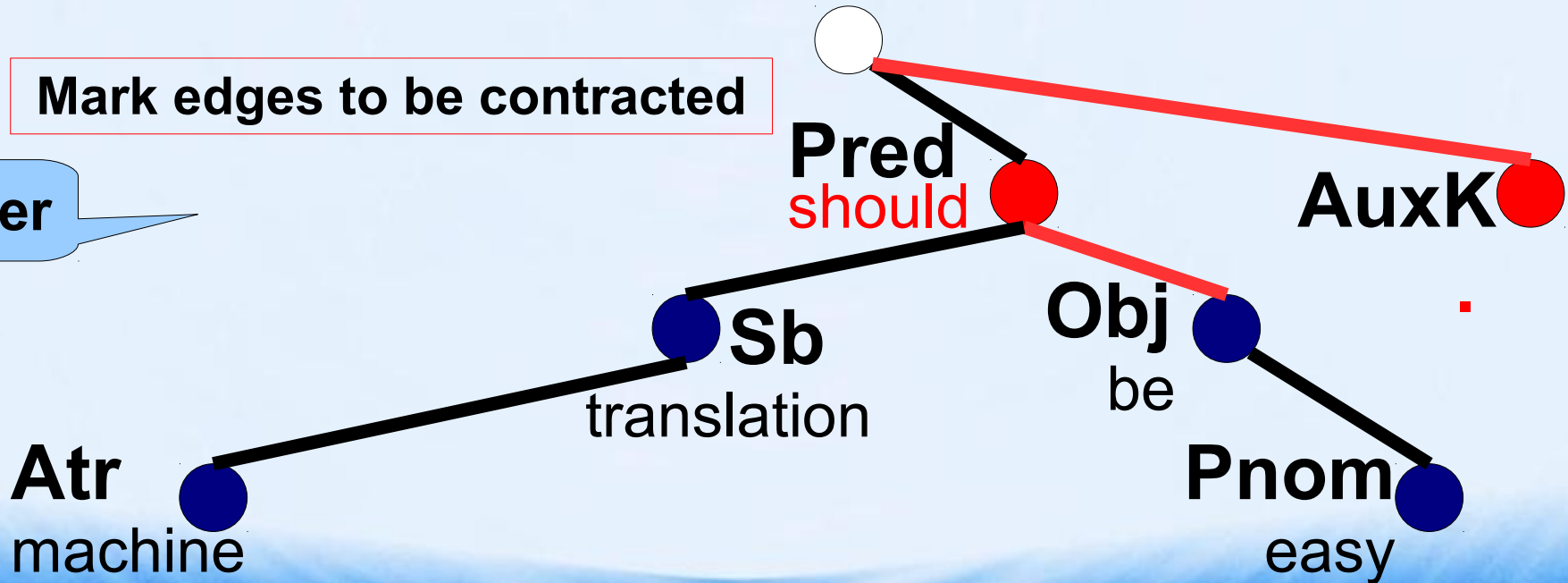
Machine translation should be easy.

m-layer

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

Mark edges to be contracted

a-layer



Demo Translation – Analysis

raw text

Machine translation should be easy.

m-layer

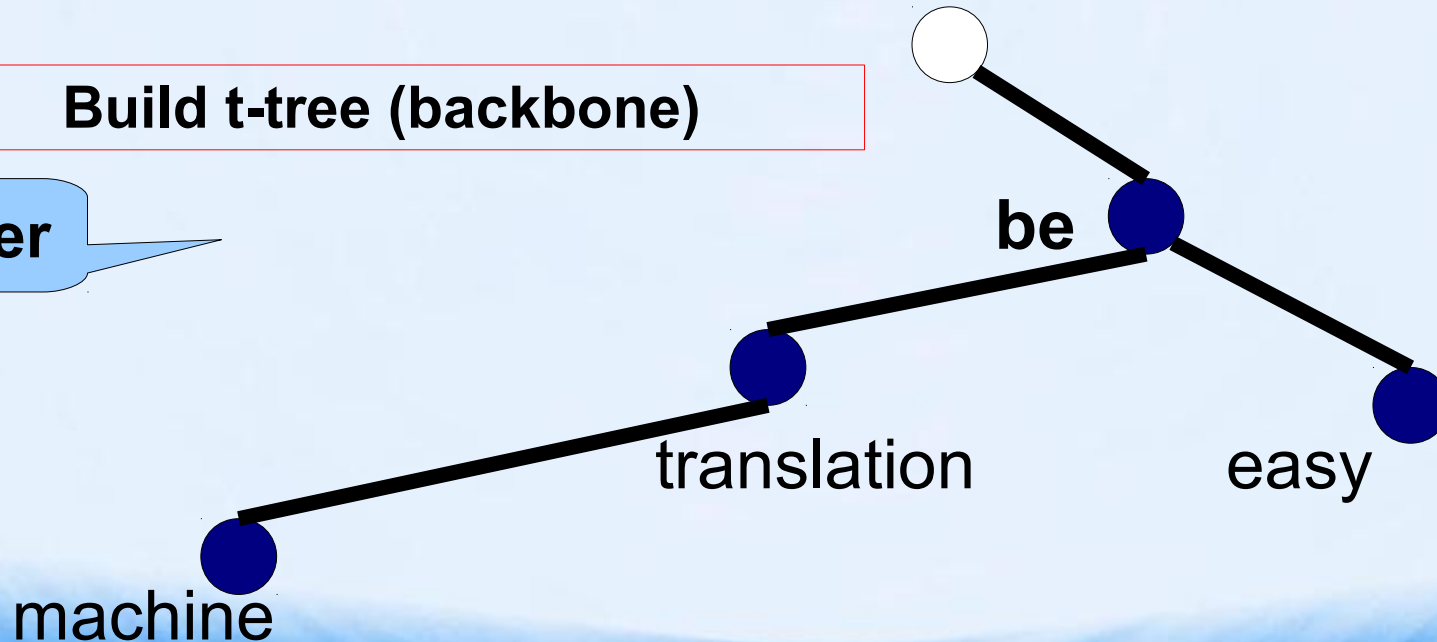
● ● ● ● ● ●

machine translation should be easy .

 NN NN MD VB JJ .

Build t-tree (backbone)

t-layer



Demo Translation – Analysis

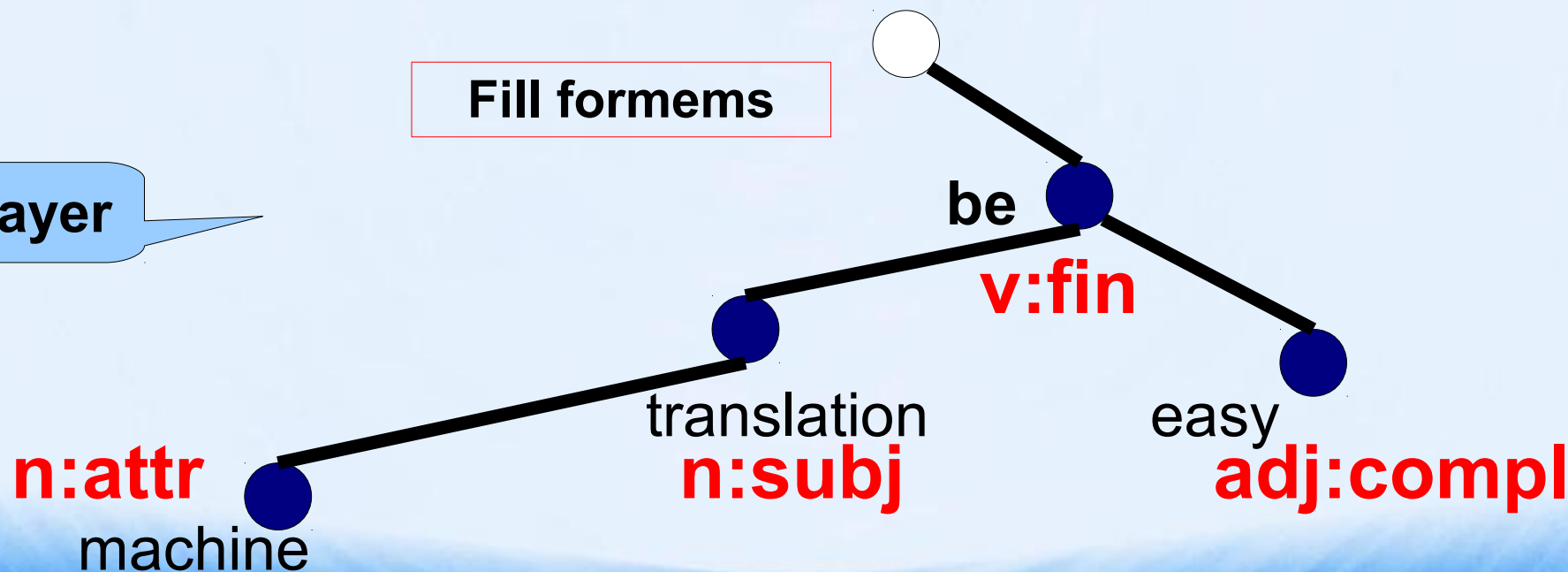
raw text

Machine translation should be easy.

m-layer

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

t-layer



Demo Translation – Analysis

raw text

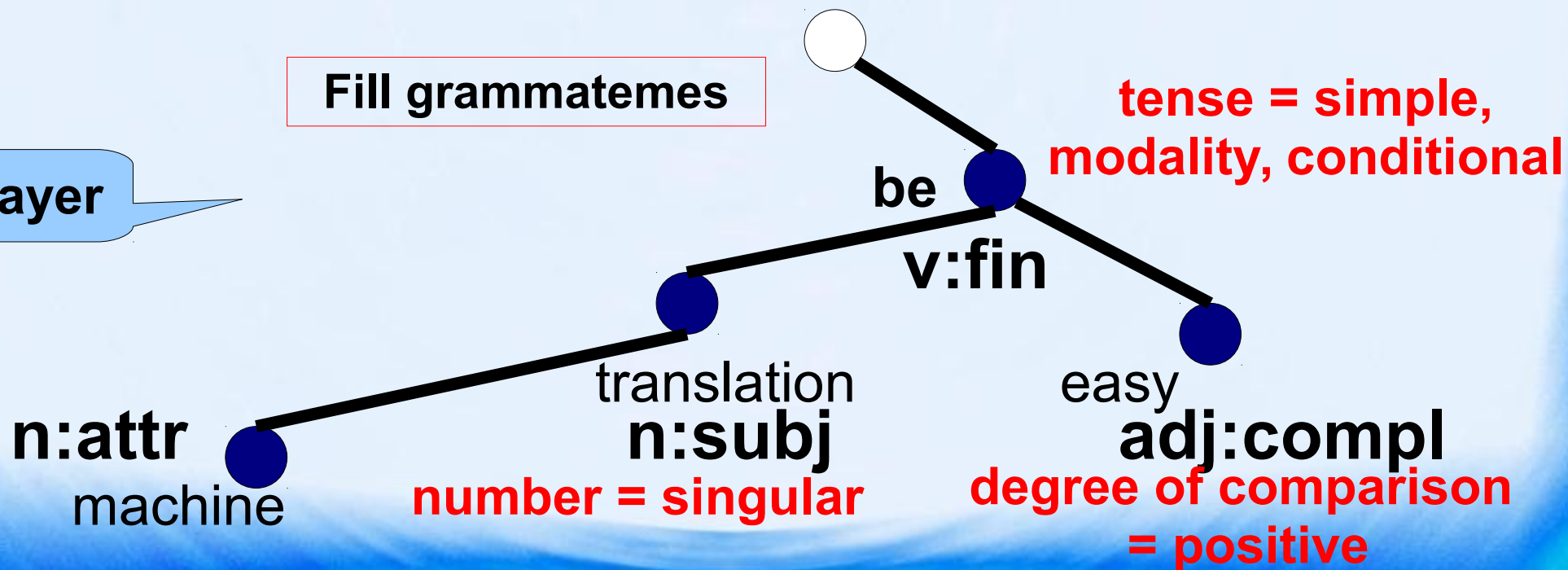
Machine translation should be easy.

m-layer

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

Fill grammatememes

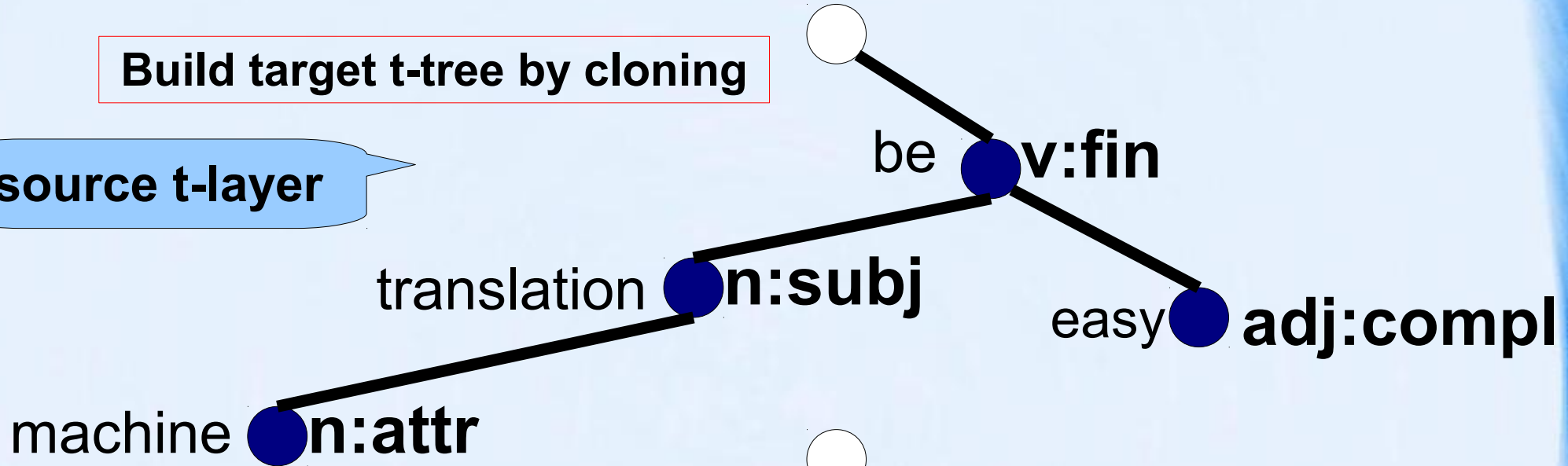
t-layer



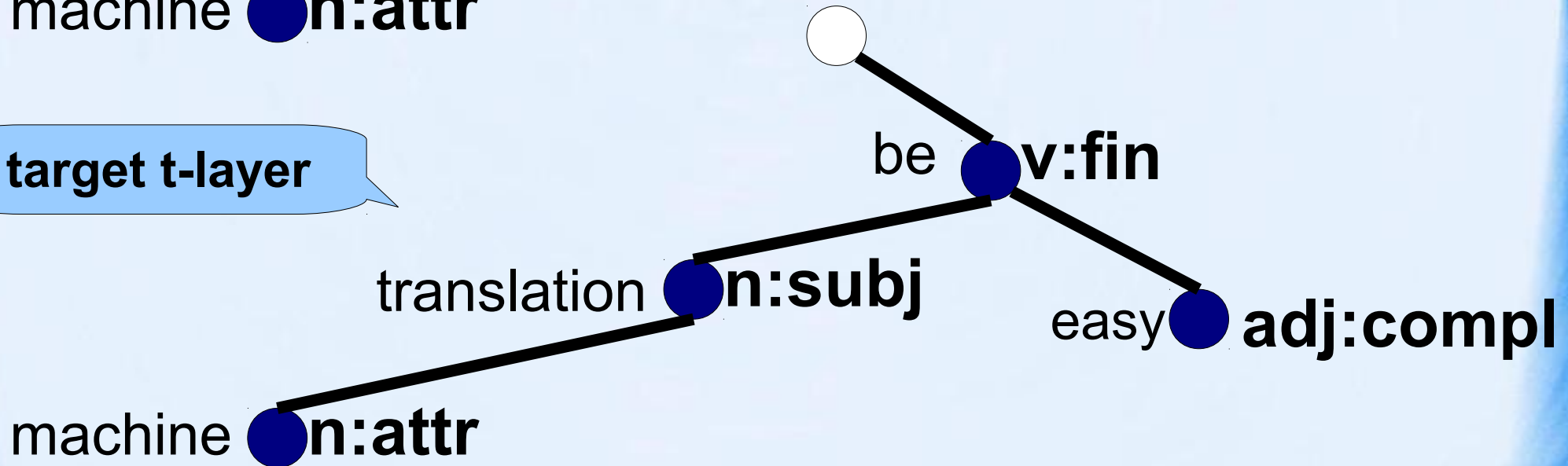
Demo Translation – Transfer

Build target t-tree by cloning

source t-layer



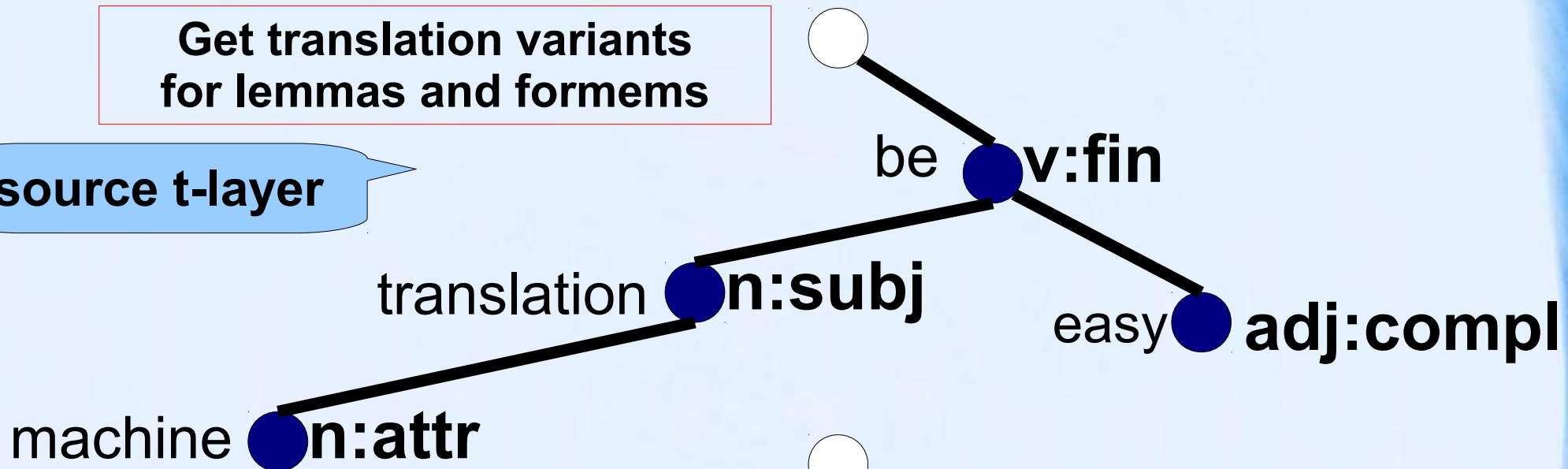
target t-layer



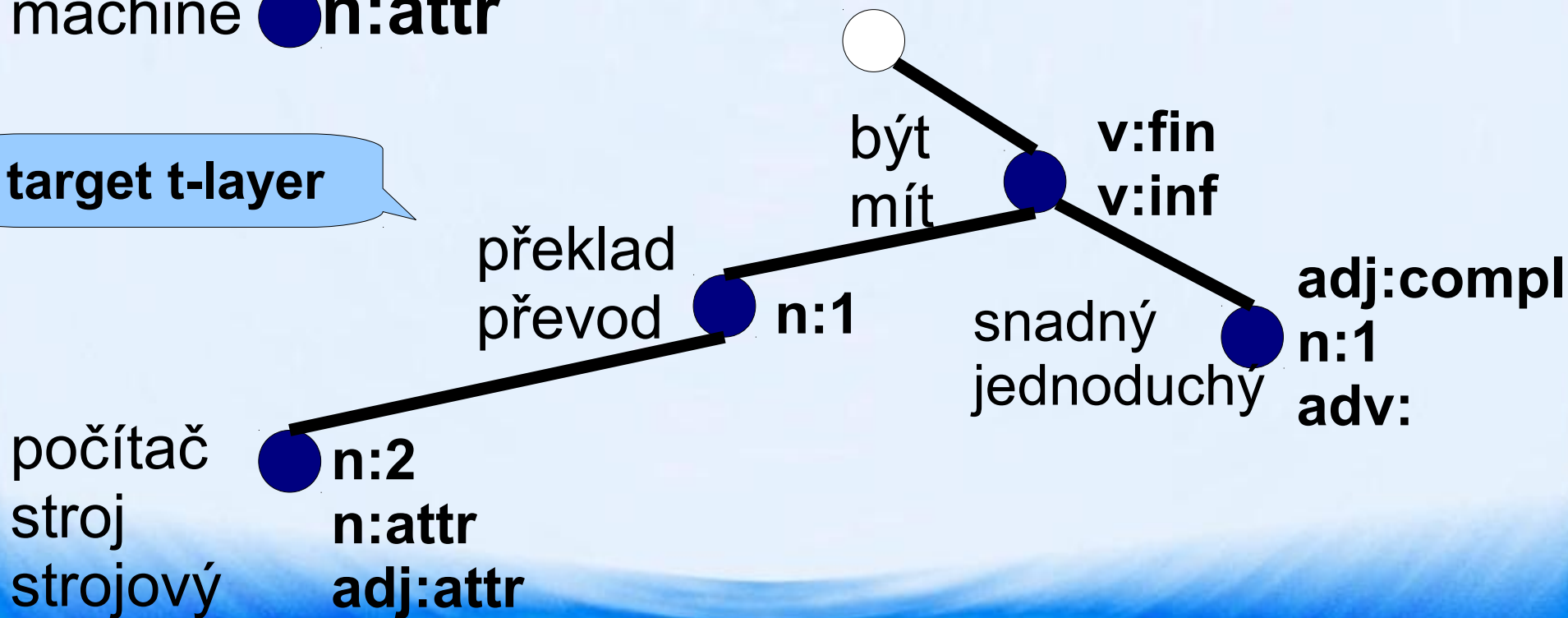
Demo Translation – Transfer

Get translation variants for lemmas and formems

source t-layer



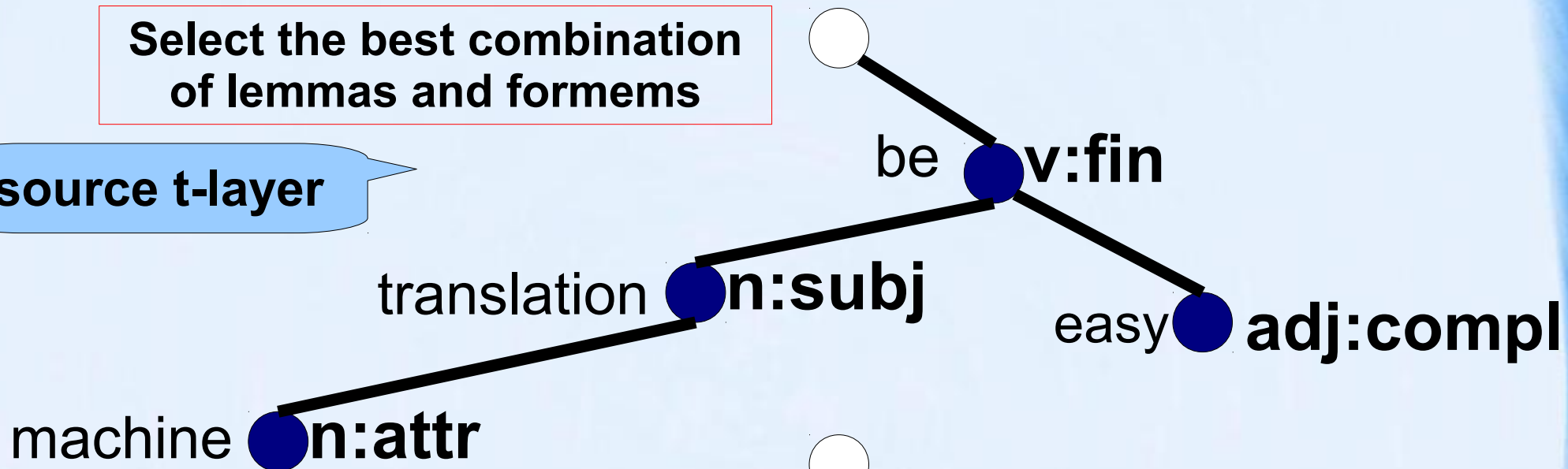
target t-layer



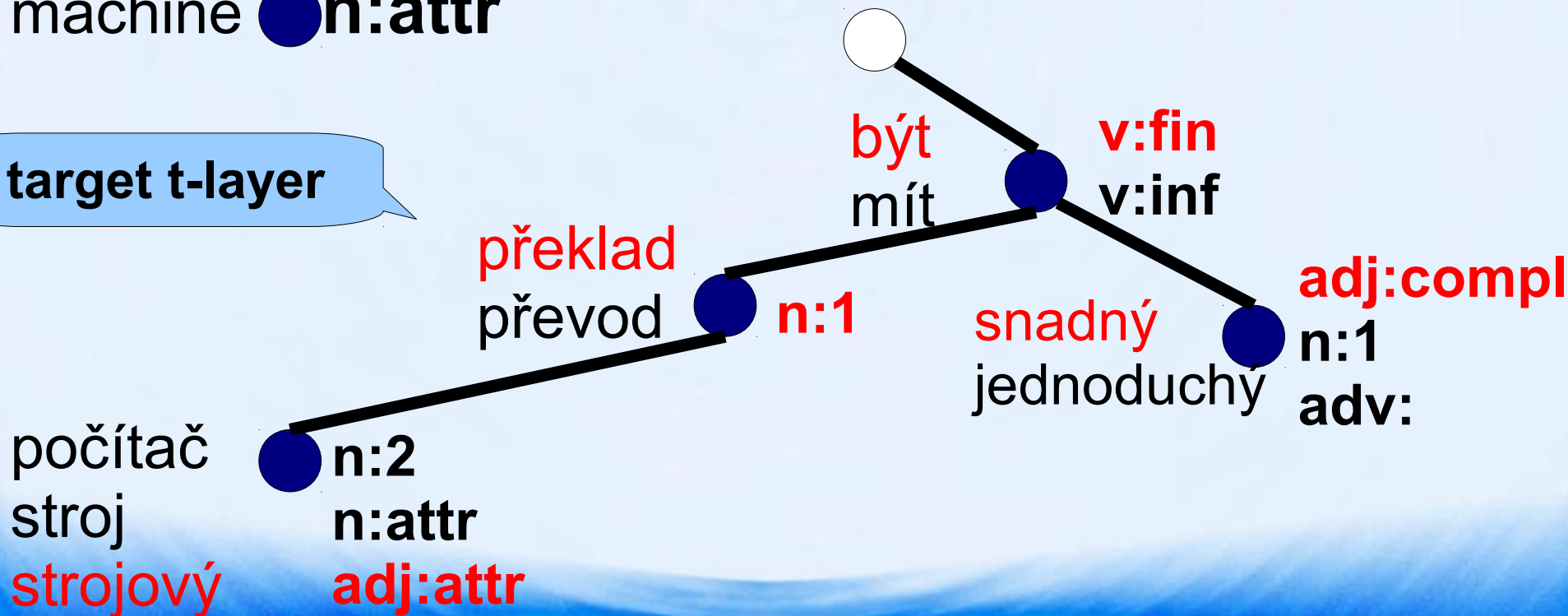
Demo Translation – Transfer

Select the best combination of lemmas and formems

source t-layer



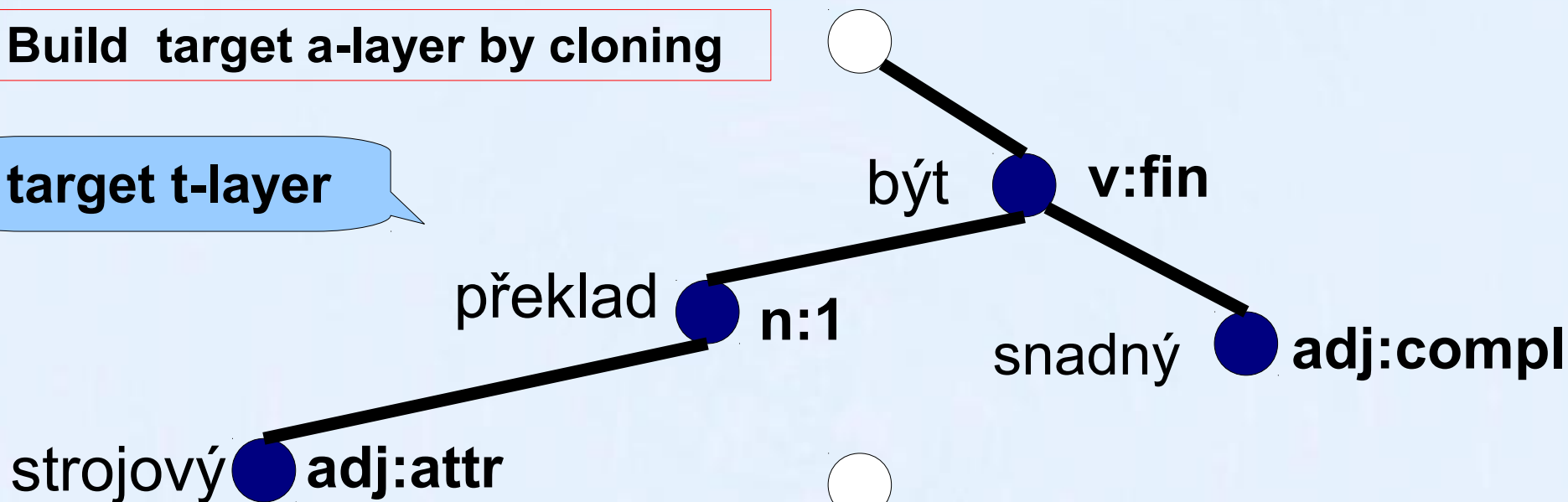
target t-layer



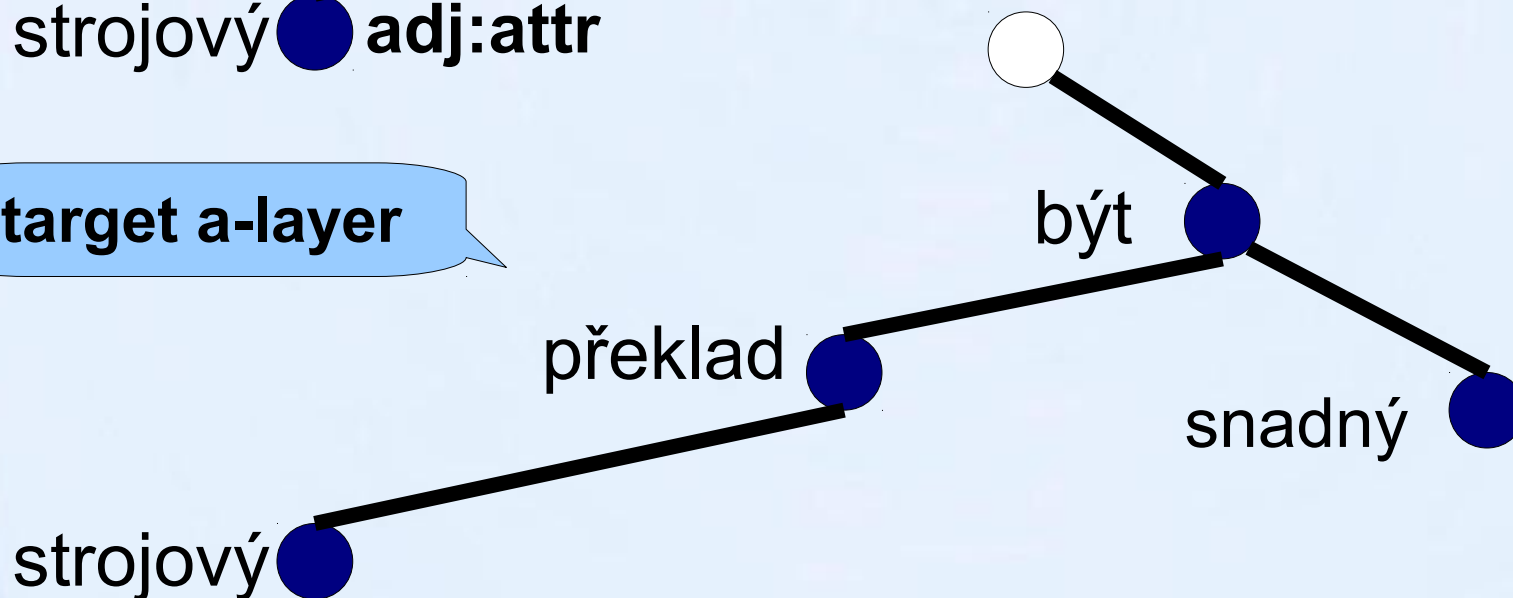
Demo Translation – Synthesis

Build target a-layer by cloning

target t-layer



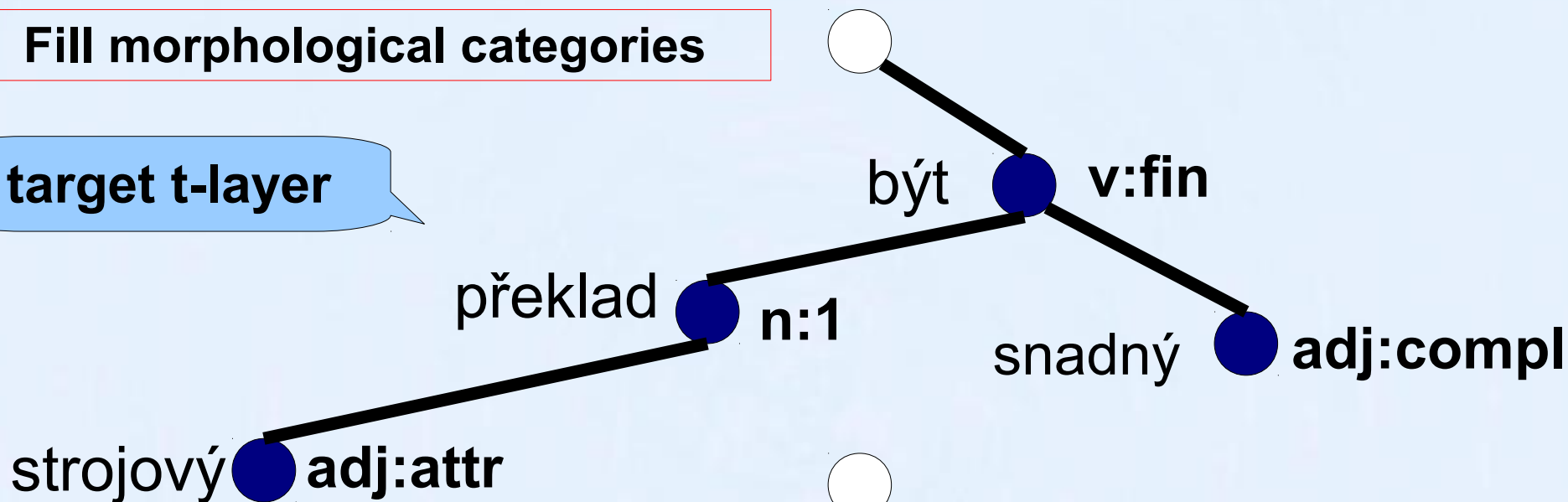
target a-layer



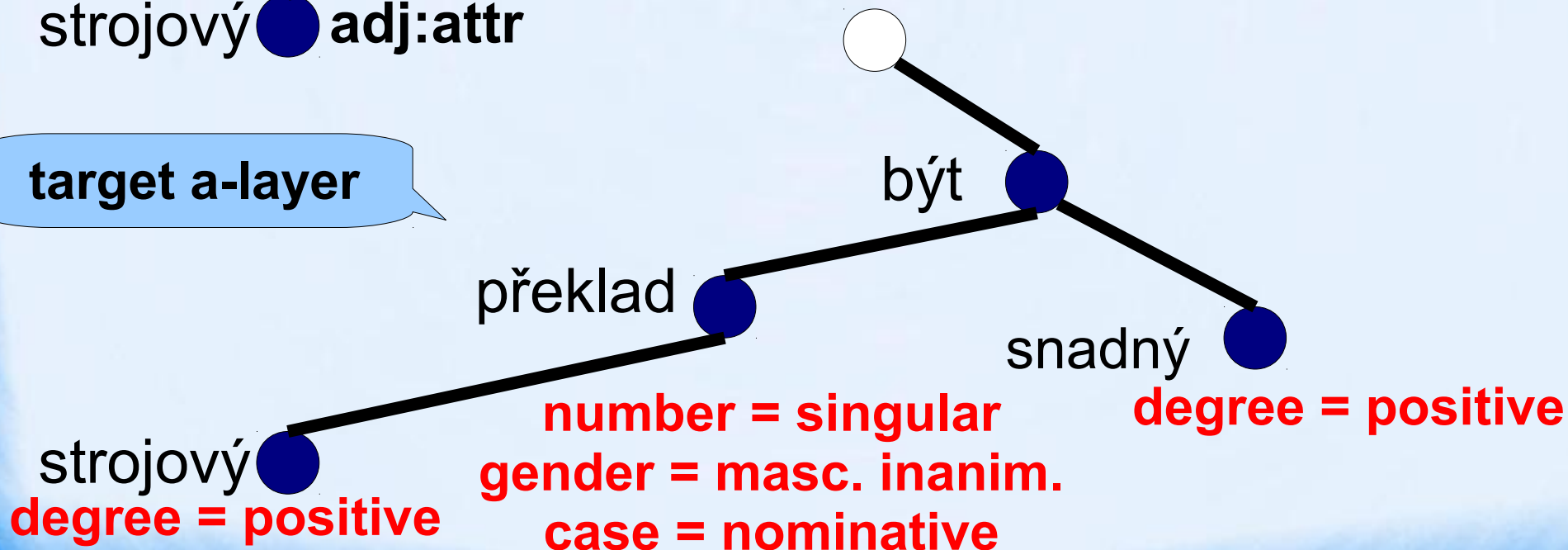
Demo Translation – Synthesis

Fill morphological categories

target t-layer



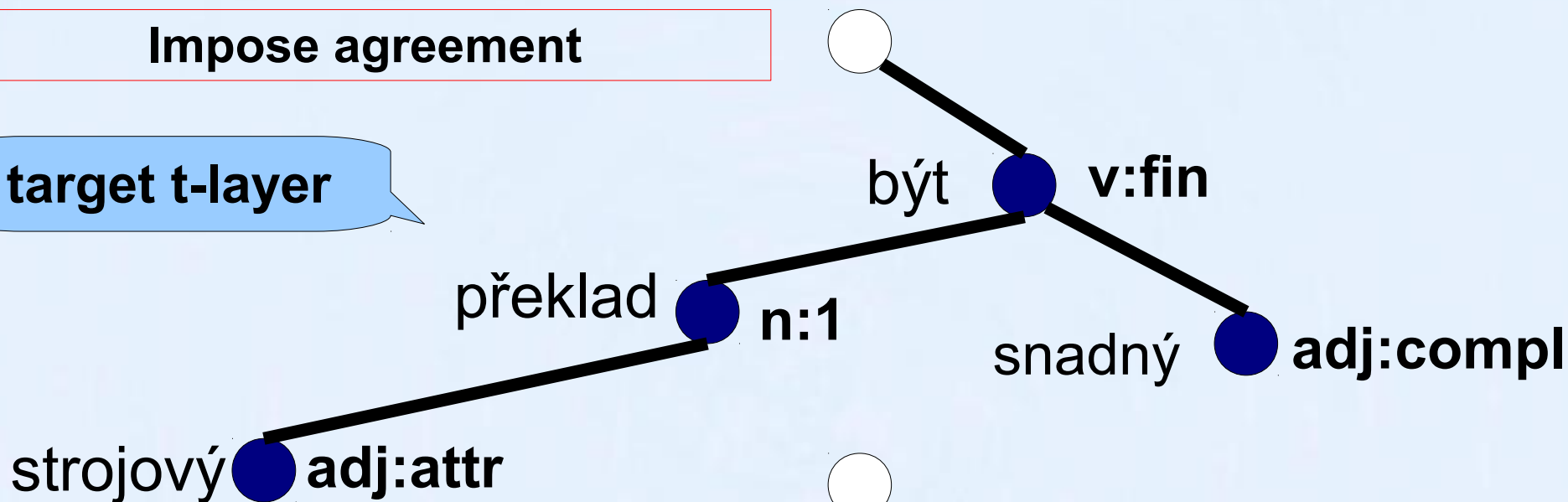
target a-layer



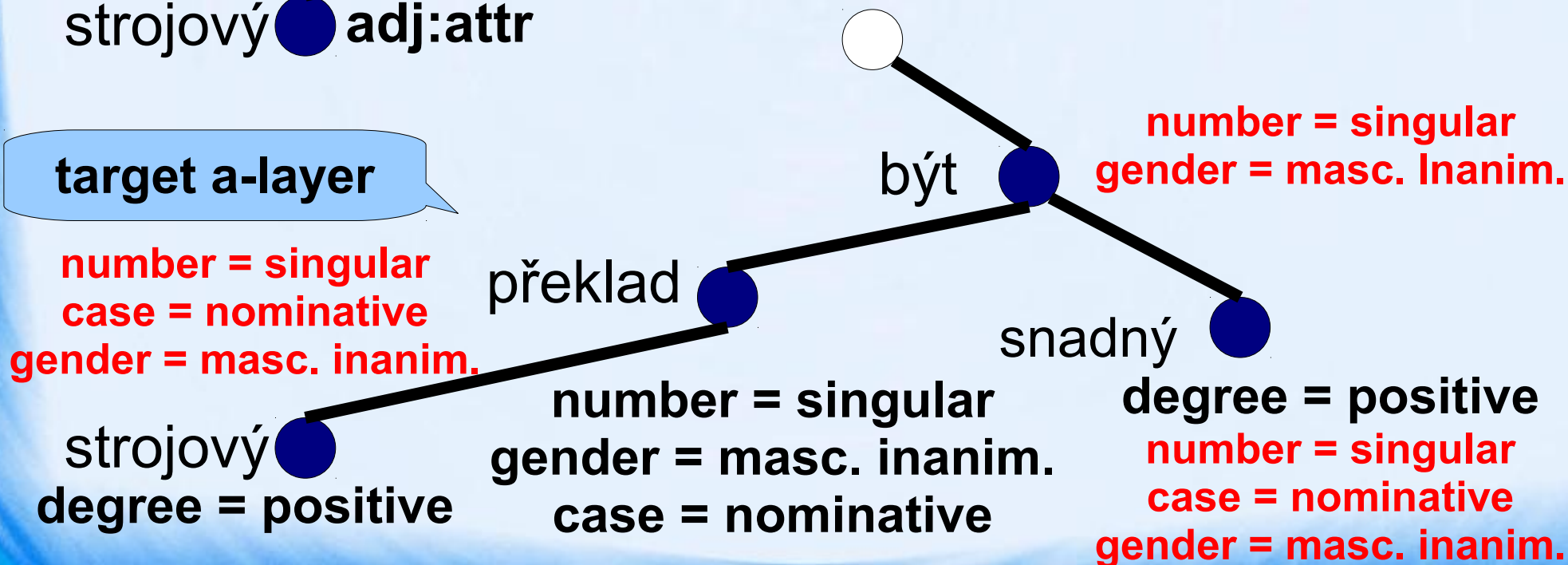
Demo Translation – Synthesis

Impose agreement

target t-layer



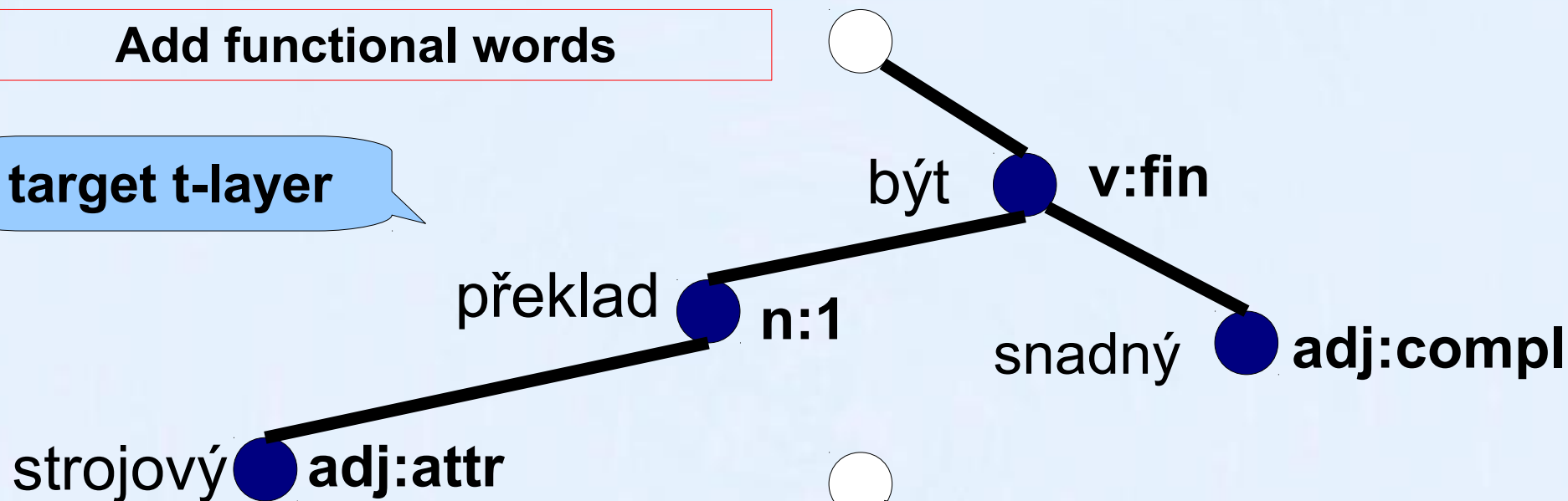
target a-layer



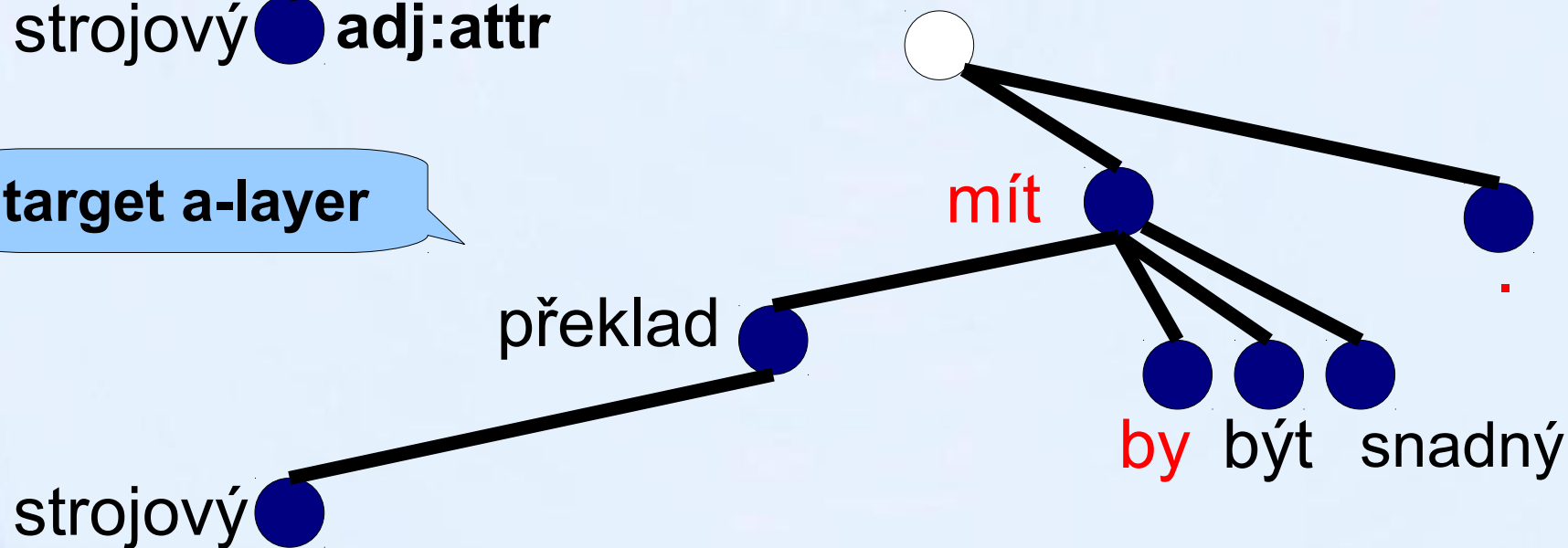
Demo Translation – Synthesis

Add functional words

target t-layer



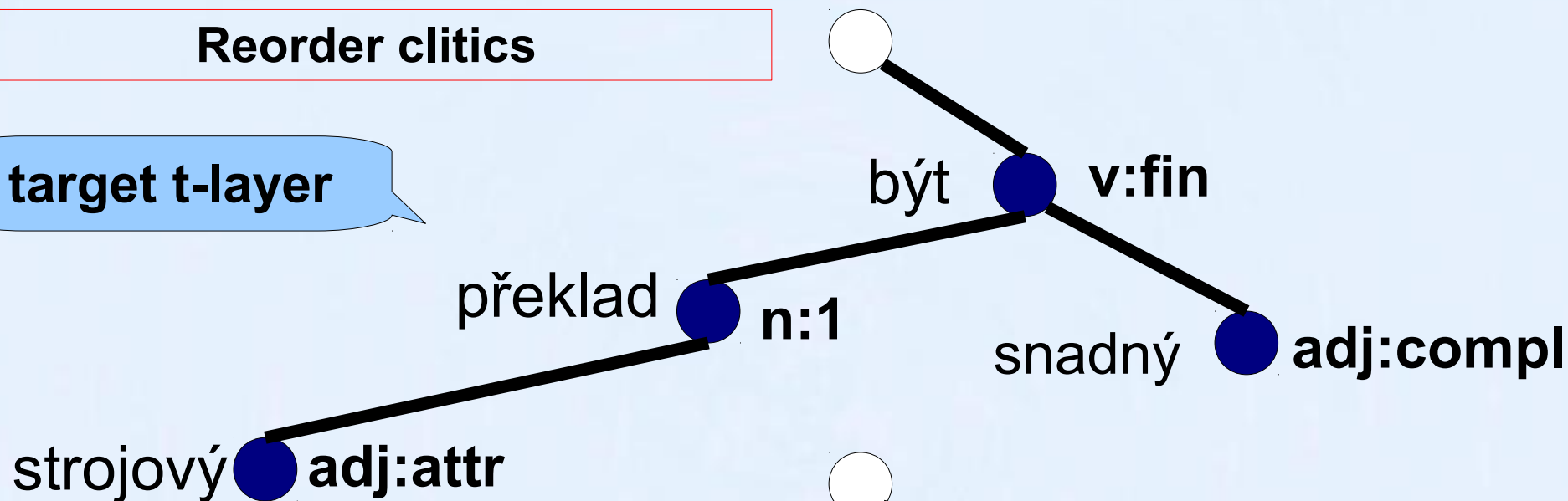
target a-layer



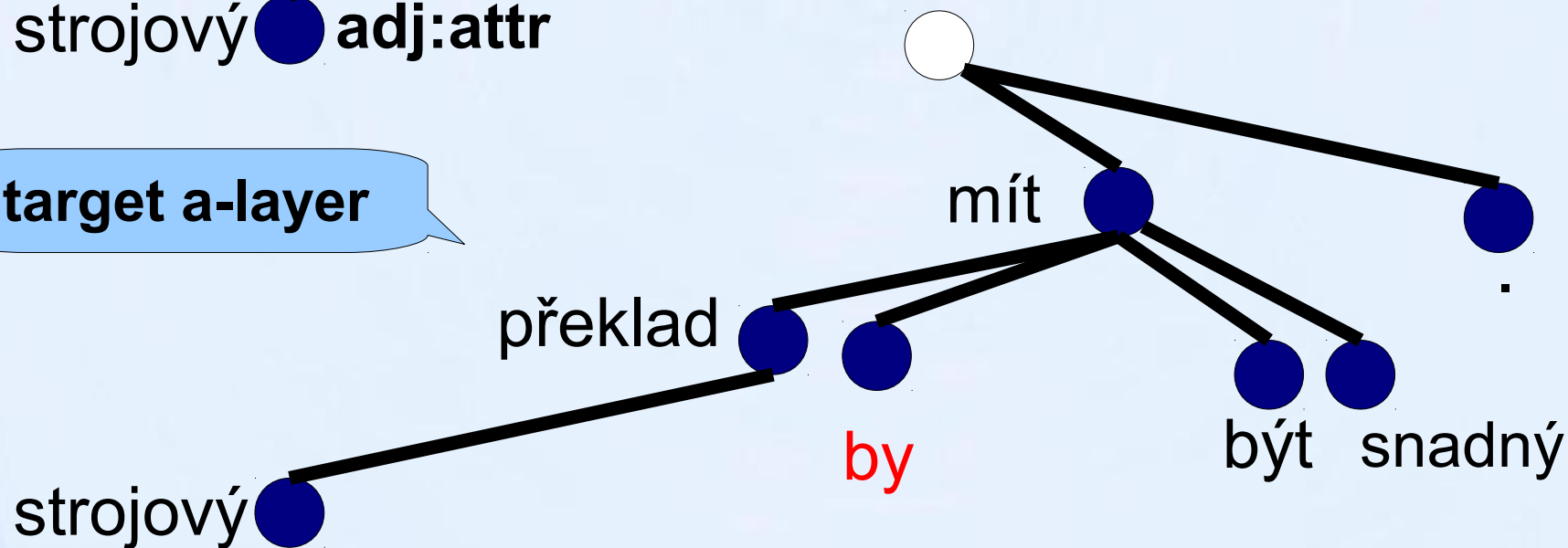
Demo Translation – Synthesis

Reorder clitics

target t-layer



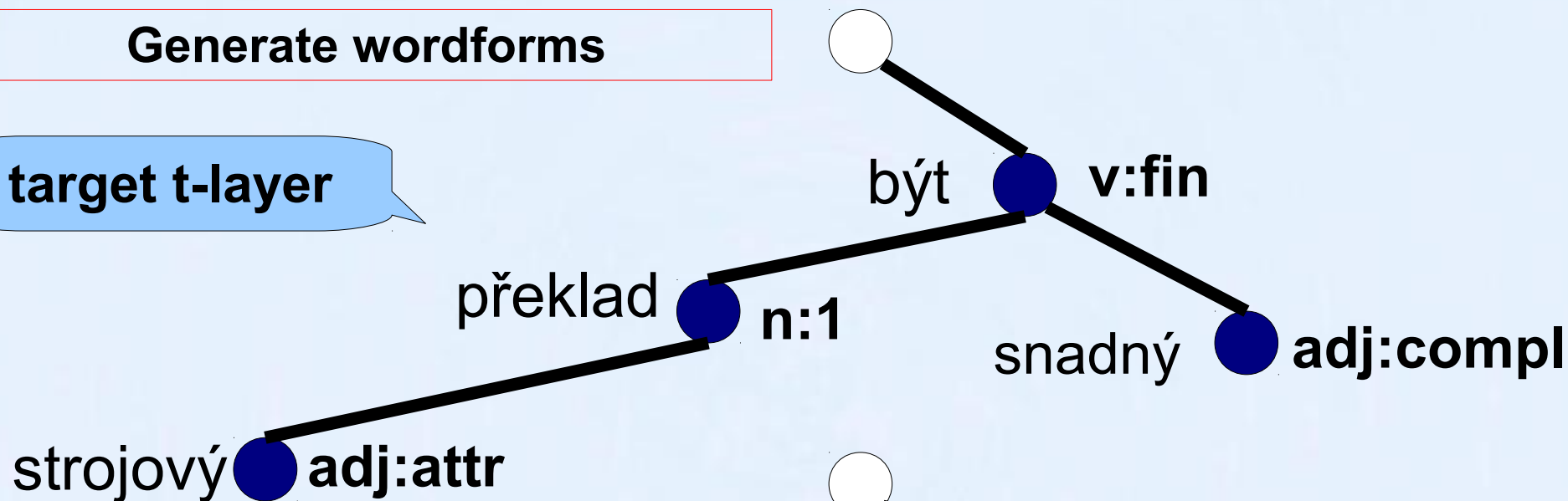
target a-layer



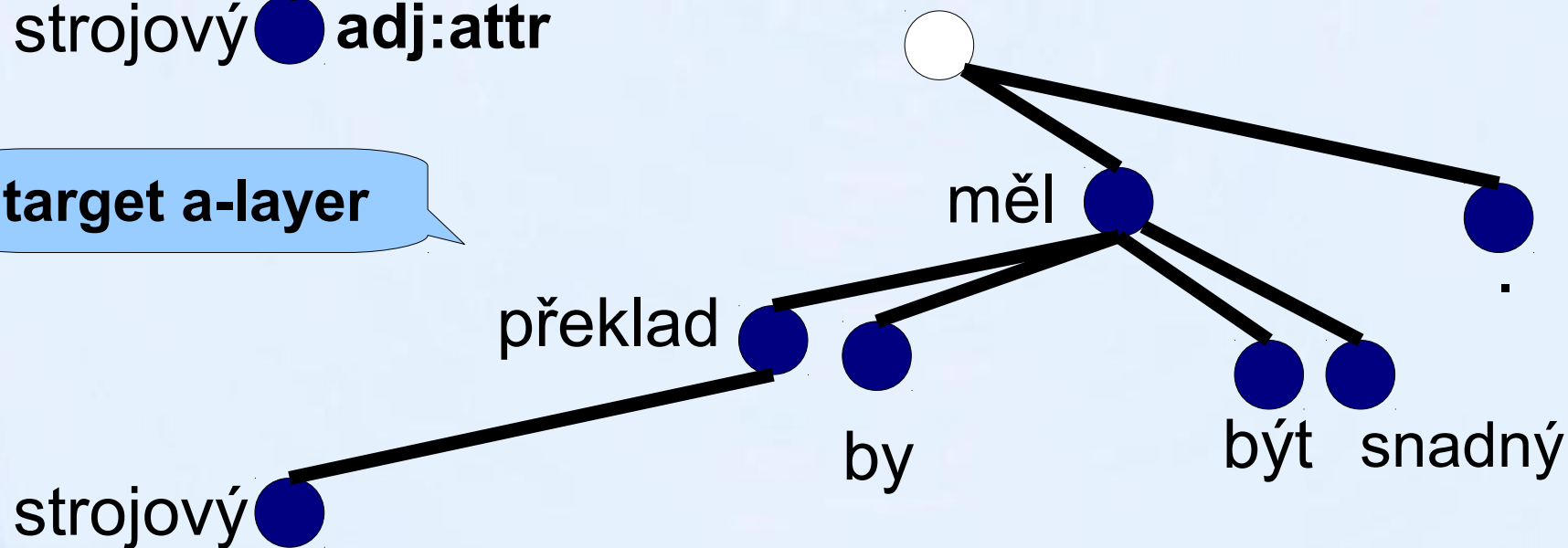
Demo Translation – Synthesis

Generate wordforms

target t-layer



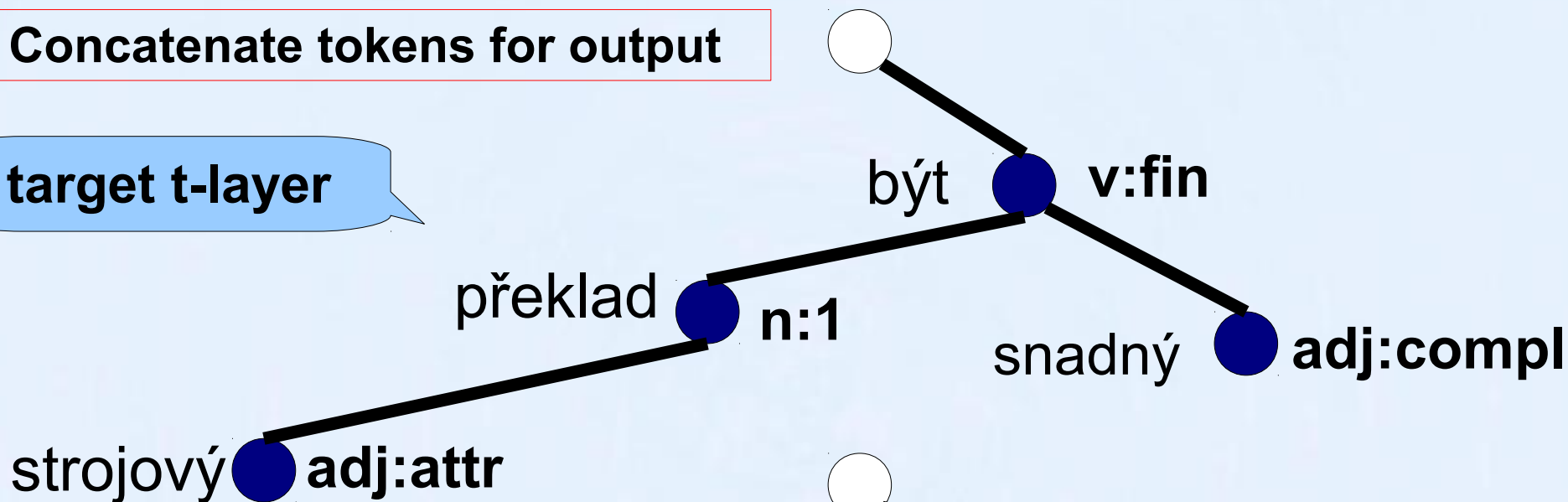
target a-layer



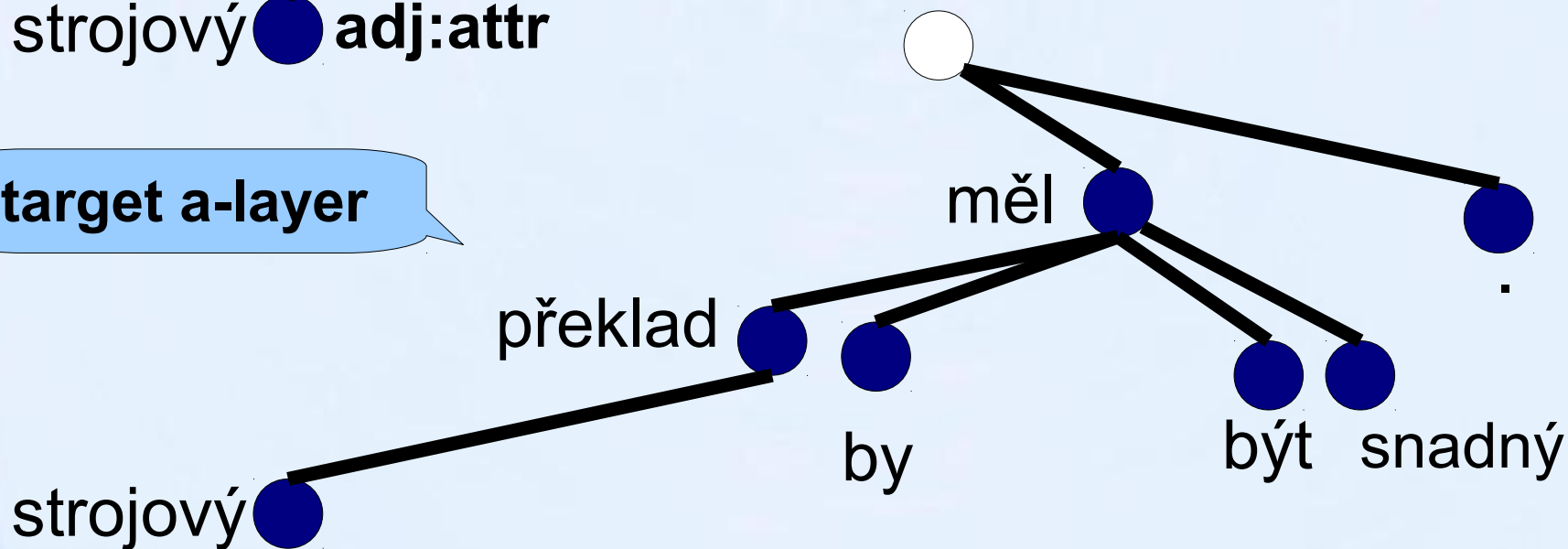
Demo Translation – Synthesis

Concatenate tokens for output

target t-layer



target a-layer



Strojový překlad by měl být snadný.

Demo Translation – Real Scenario



MORPHOLOGY:

ResegmentSentences

Tokenize

NormalizeForms

FixTokenization

TagMorce

FixTags

Lemmatize

NAMED ENTITIES:

StanfordNamedEntities

DistinguishPersonalNames

A-LAYER:

MarkChunks

ParseMST

SetIsMemberFromDeprel

RehangConllToPdtStyle

FixNominalGroups

FixIsMember

FixAtree

FixMultiwordPrepAndConj

FixDicendiVerbs

SetAfunAuxCPCoord

SetAfun

T-LAYER:

MarkEdgesToCollapse

MarkEdgesToCollapseNeg

BuildTtree

SetIsMember

MoveAuxFromCoordToMembers

FixTlemmas

SetCoapFunctors

FixEitherOr

FixIsMember

MarkClauseHeads

MarkPassives

SetFunctors

MarkInfin

MarkRelClauseHeads

MarkRelClauseCoref

MarkDspRoot

MarkParentheses

SetNodetype

SetGrammatemes

SetFormeme

RehangSharedAttr

SetVoice

FixImperatives

SetIsNameOfPerson

SetGenderOfPerson

AddCorAct

FindTextCoref

TRANSFER:

CopyTtree

TrLFPhrases

TrLFJointStatic

DeleteSuperfluousTnodes

TrFTryRules

TrFAddVariants

TrFRerank

TrLTryRules

TrLAddVariants

TrLFNumeralsByRules

TrLFilterAspect

TransformPassiveConstructions

PrunePersonalNameVariants

RemoveUnpassivizableVariants

TrLFCompounds

CutVariants

RehangToEffParents

TrLFTreeViterbi

RehangToOrigParents

CutVariants

FixTransferChoices

ReplaceVerbWithAdj

DeletePossPronBeforeVlastni

TrLFemaleSurnames

AddNounGender

MarkNewRelClauses

AddRelpronBelowRc

ChangeCorToPersPron

AddPersPronBelowVfin

AddVerbAspect

FixDateTime

FixGrammatemesAfterTransfer

FixNegation

MoveAdjsBeforeNouns

MoveGenitivesRight

MoveRelClauseRight

MoveDicendiCloserToDsp

MovePersPronNextToVerb

MoveEnoughBeforeAdj

MoveJesteBeforeVerb

FixMoney

OverridePpWithPhraseTr

FindGramCoreffForReflPron

NeutPersPronGenderFromAntec

ValencyRelatedRules

SetClauseNumber

TurnTextCorefToGramCoref

SYNTHESIS TO A-LAYER:

CopyTtree

DistinguishHomonymous.

ReverseNumberNounDep.

InitMorphcat

FixPossessiveAdjs

MarkSubject

ImposePronZAgr

ImposeRelPronAgr

ImposeSubjpredAgr

ImposeAttrAgr

ImposeComplAgr

DropSubjPersProns

AddPrepos

AddSubconj

AddReflexParticles

AddAuxVerbCompoundPassive

AddAuxVerbModal

AddAuxVerbCompoundFuture

AddAuxVerbConditional

AddAuxVerbCompoundPast

AddClausalExpletivePronouns

ResolveVerbs

ProjectClauseNumber

AddParentheses

AddSentFinalPunct

AddSubordClausePunct

AddCoordPunct

AddAppositionPunct

ChooseMlemmaForPersPron

GenerateWordforms

MoveCliticsToWackernagel

DeleteSuperfluousPrepos

DeleteEmptyNouns

VocalizePrepos

CapitalizeSentStart

CapitalizeNamedEntities.

FillTagFromMorphcat

SYNTHESIS TO TEXT:

ConcatenateTokens

ApplySubstitutions

DetokenizeUsingRules

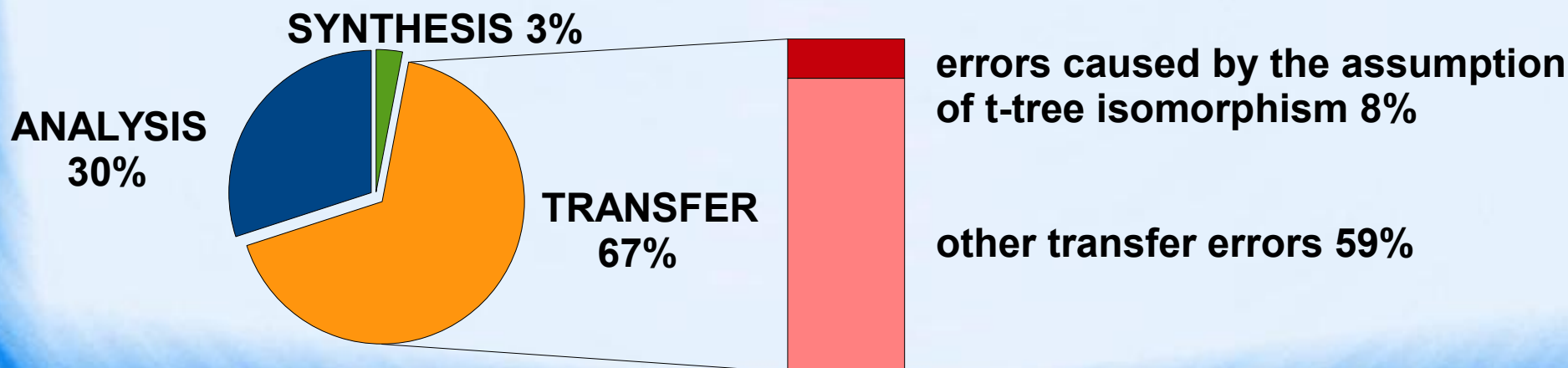
RemoveRepeatedTokens

NormalizePunctuationForWMT

Annotation of Translation Errors

sample of 250 sentences, 1463 errors in total

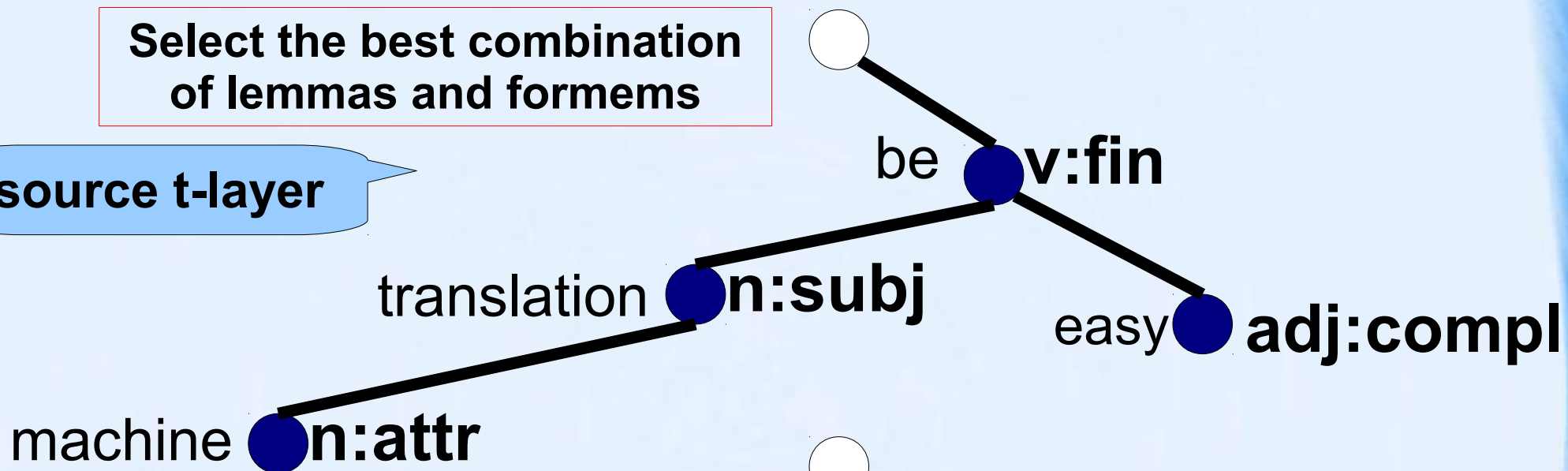
Type	lemma, formeme, gram., w. order,...
Subtype	gram: gender, person, tense,...
Seriousness	serious, minor
Circumstances	coordination, named entity, numbers
Source	tok, lem, tagger, parser, tecto, trans, x, syn, ?



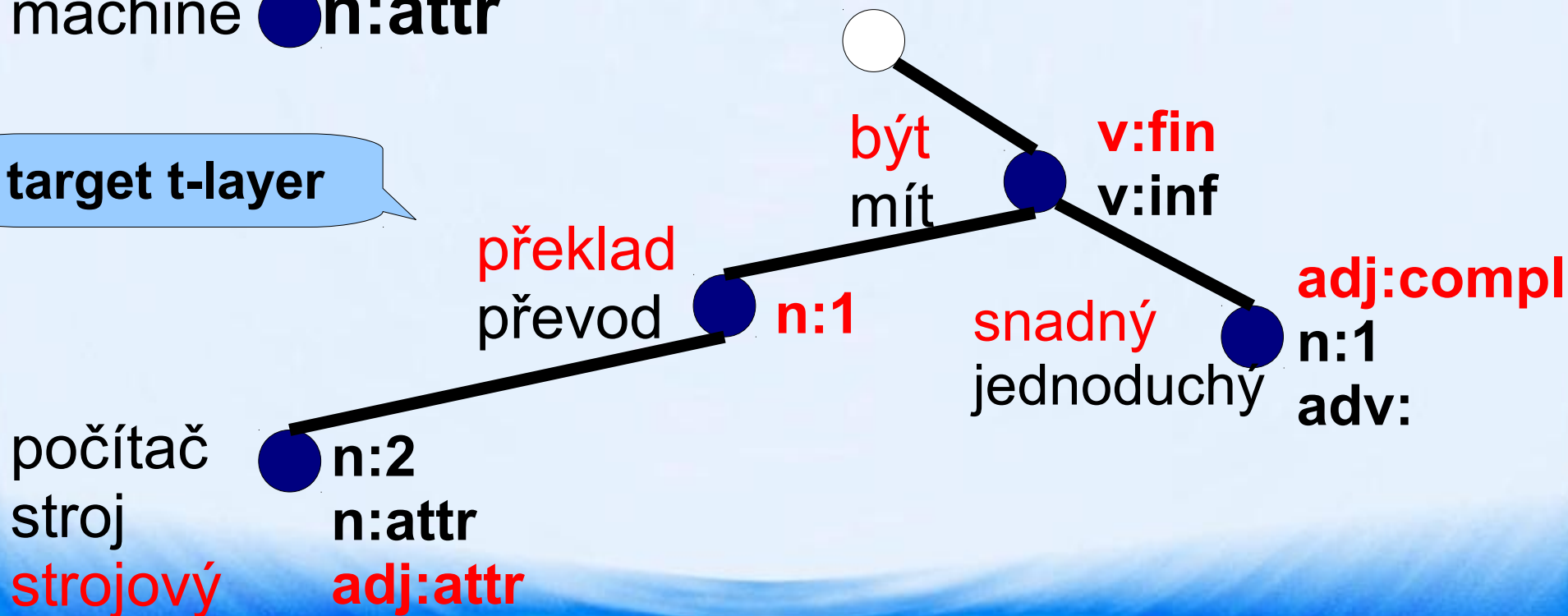
HMTM – Motivation

Select the best combination of lemmas and formems

source t-layer



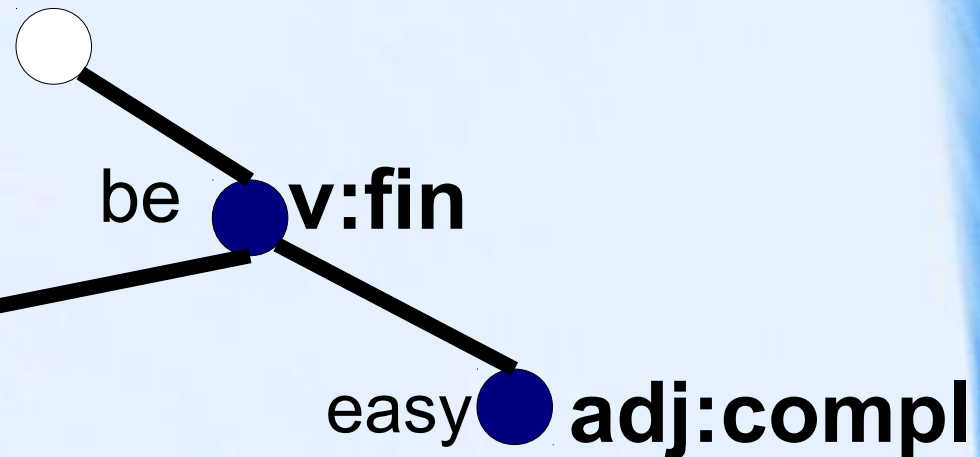
target t-layer



HMTM – Motivation

Select the best label for each node

source t-layer

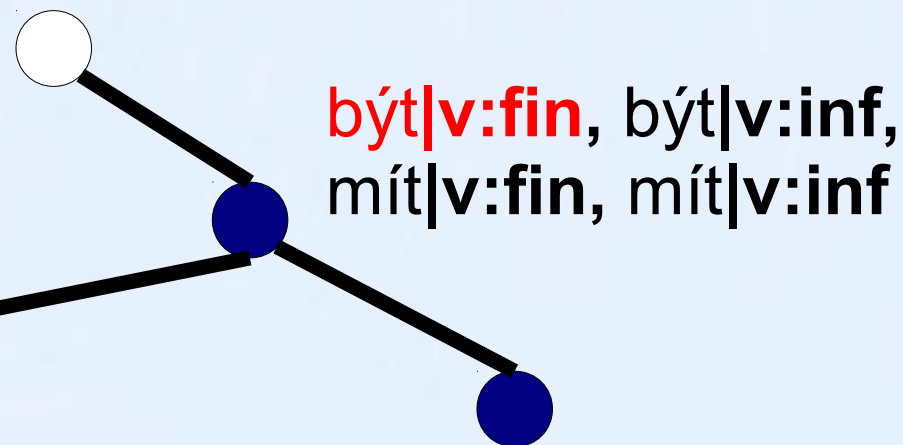


machine n:attr

translation
n:subj

easy adj:compl

target t-layer



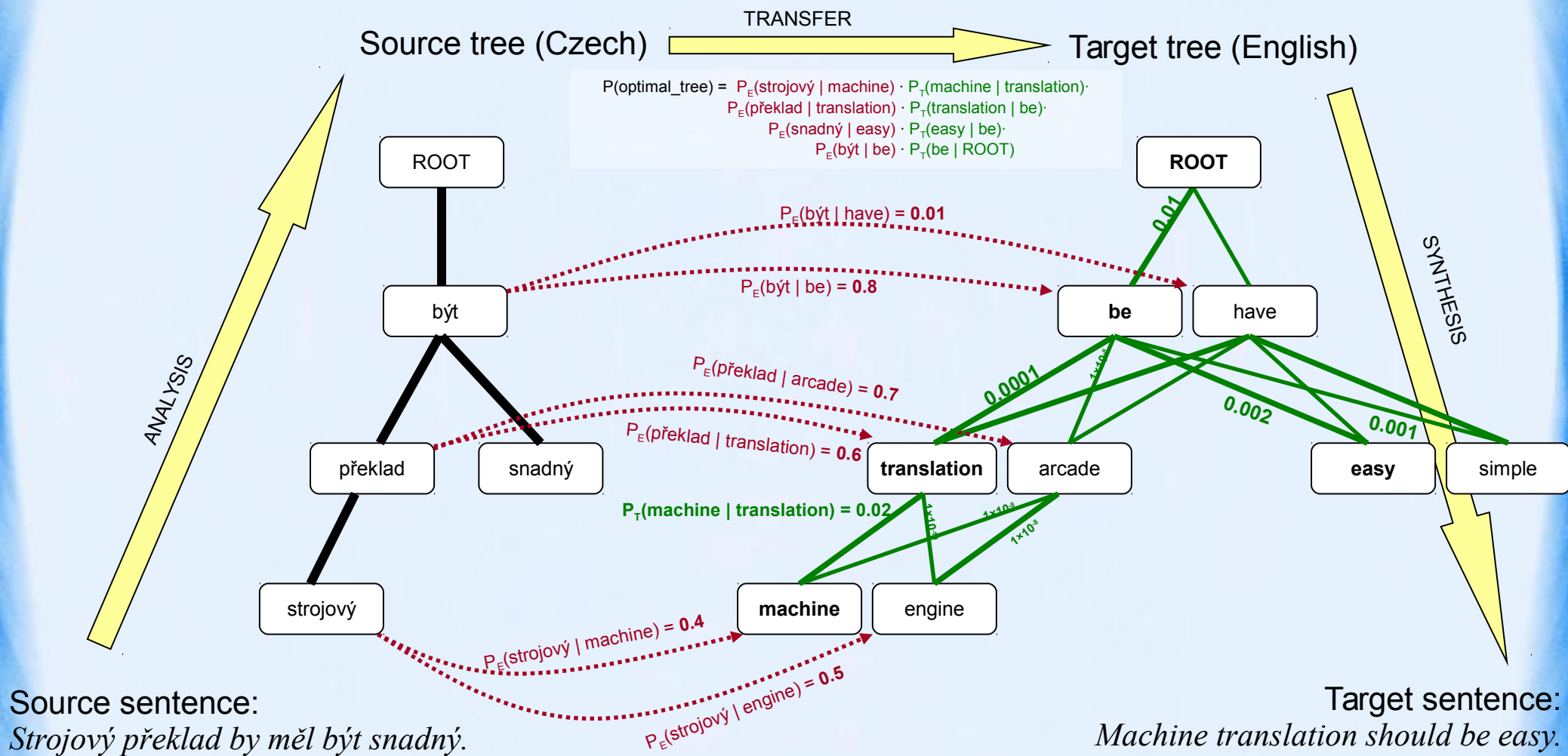
být|v:fin, být|v:inf,
mít|v:fin, mít|v:inf

překlad|n:1,
převod|n:1

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

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

HMTM in MT



$P_E(\text{source | target})$... emission probabilities ... **translation model**
 $P_T(\text{dependent | governing})$... transition probabilities ... **target-language tree model**

Combining Dictionaries

- new general interface (for lemmas and formems)
`$dict->get_translations($input_label, $features)`
returns a list of translation variants including probabilities
- OOP style, dictionary constructor can take another dictionary (or more) as a parameter → hierachy
- Four basic types of dictionaries:

Static plain

loaded from a file „lemma → lemma“

Context

loaded from a file „lemma,features → lemma“

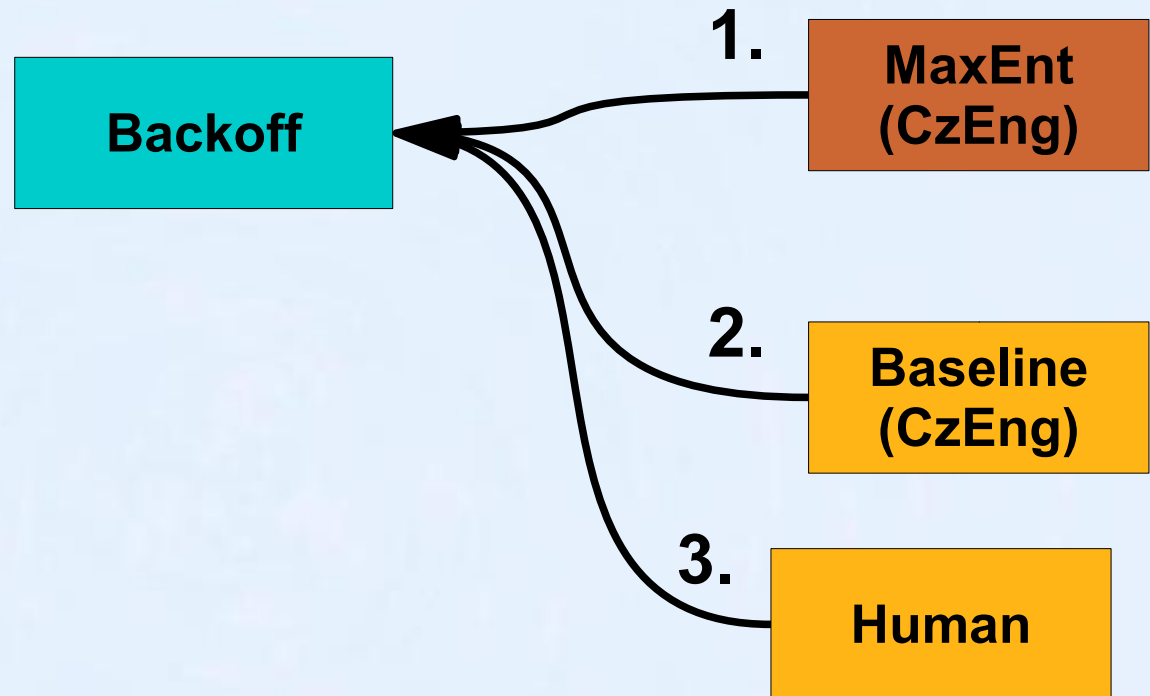
Derivational

translations derived dynamicaly, input dictionary

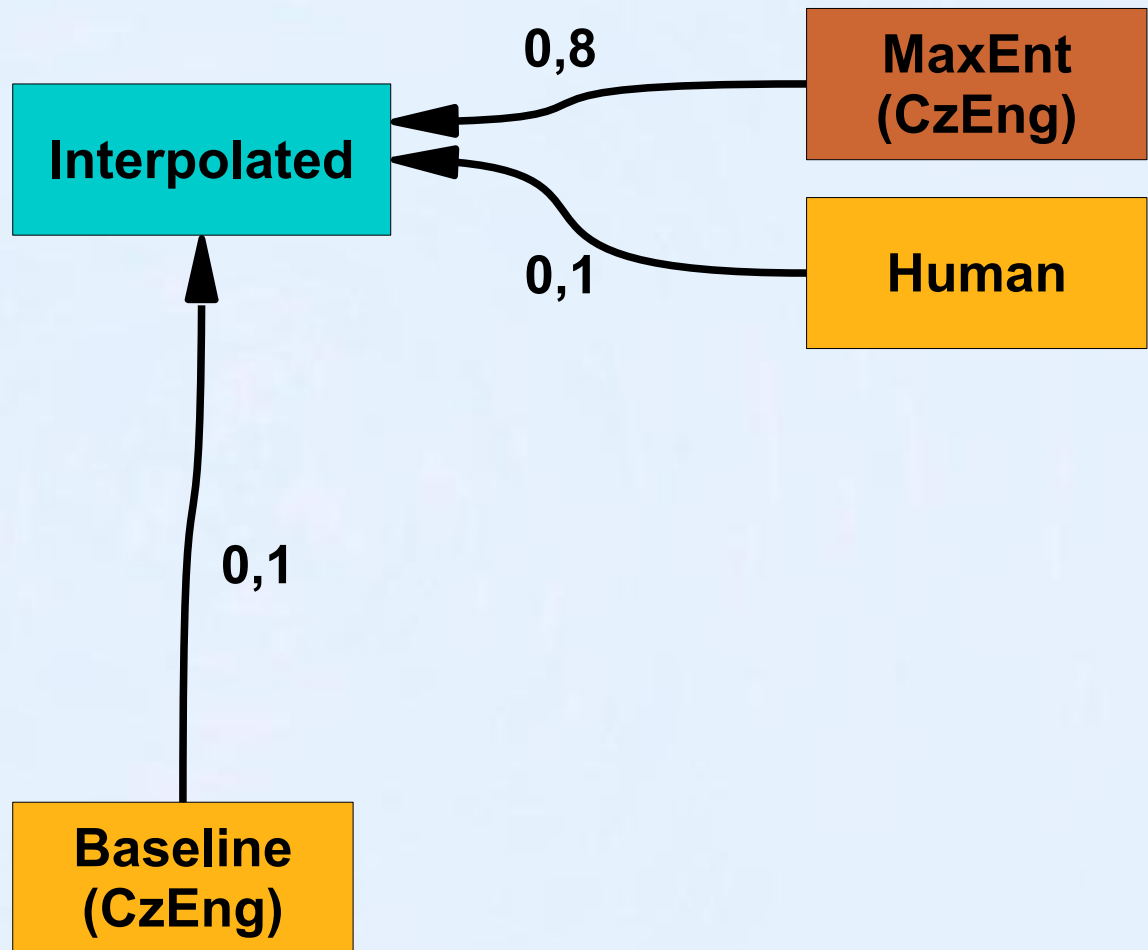
Combinaional

combination of more input dictionaries

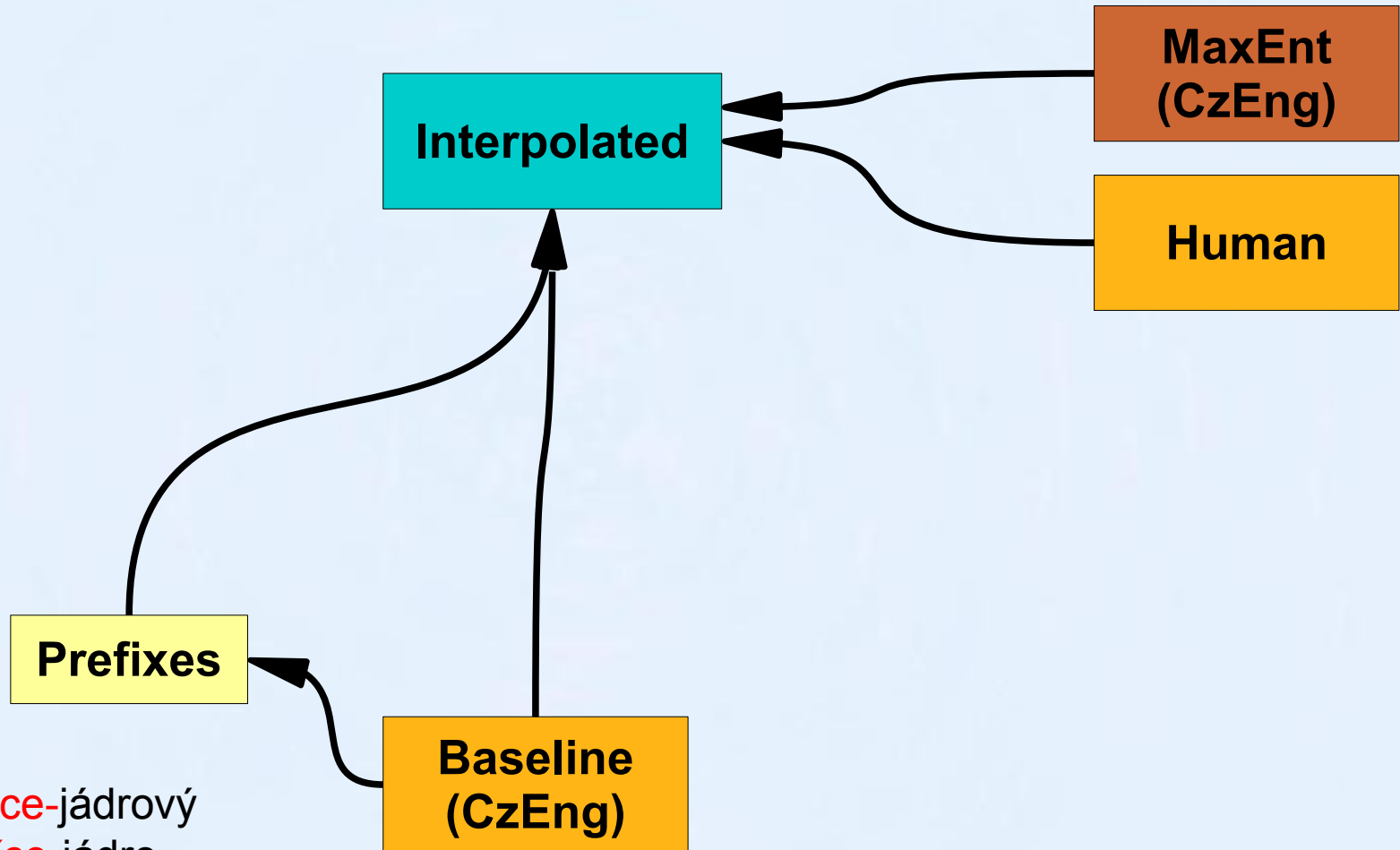
Hierarchy of lemma dictionaries



Hierarchy of lemma dictionaries

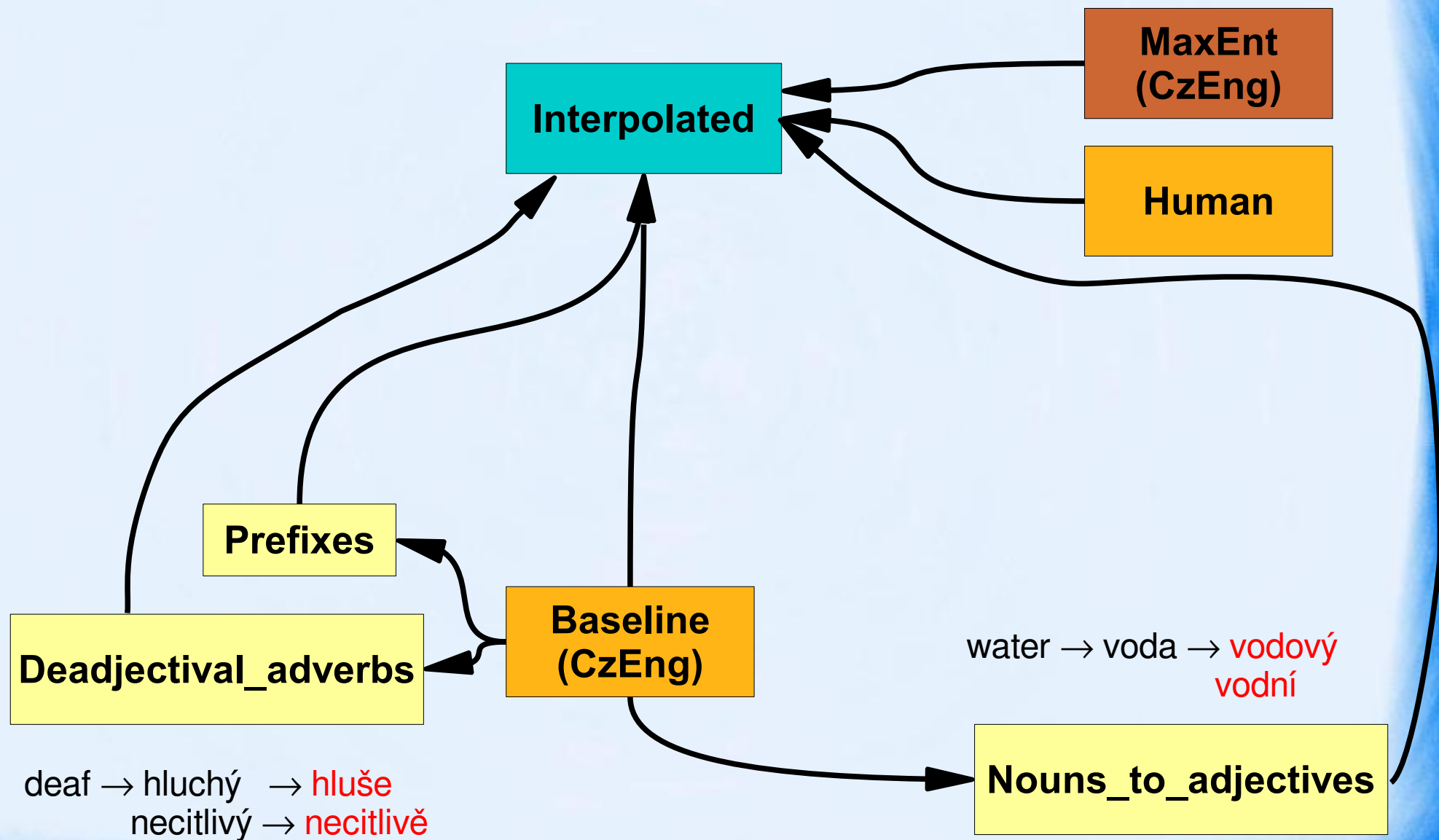


Hierarchy of lemma dictionaries

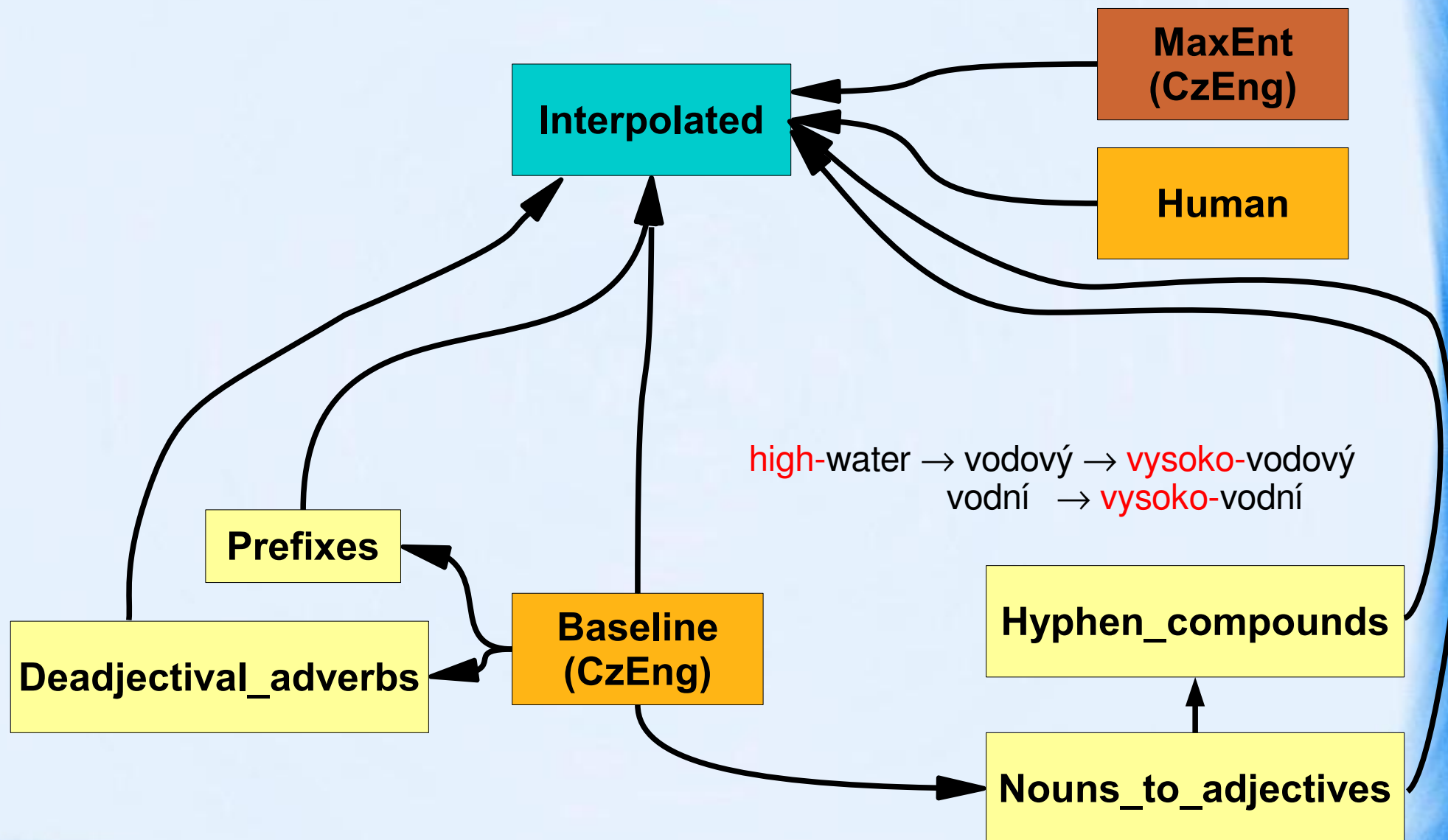


multi-core → více-jádrový
více-jádro
multi-jádrový
multi-jádro

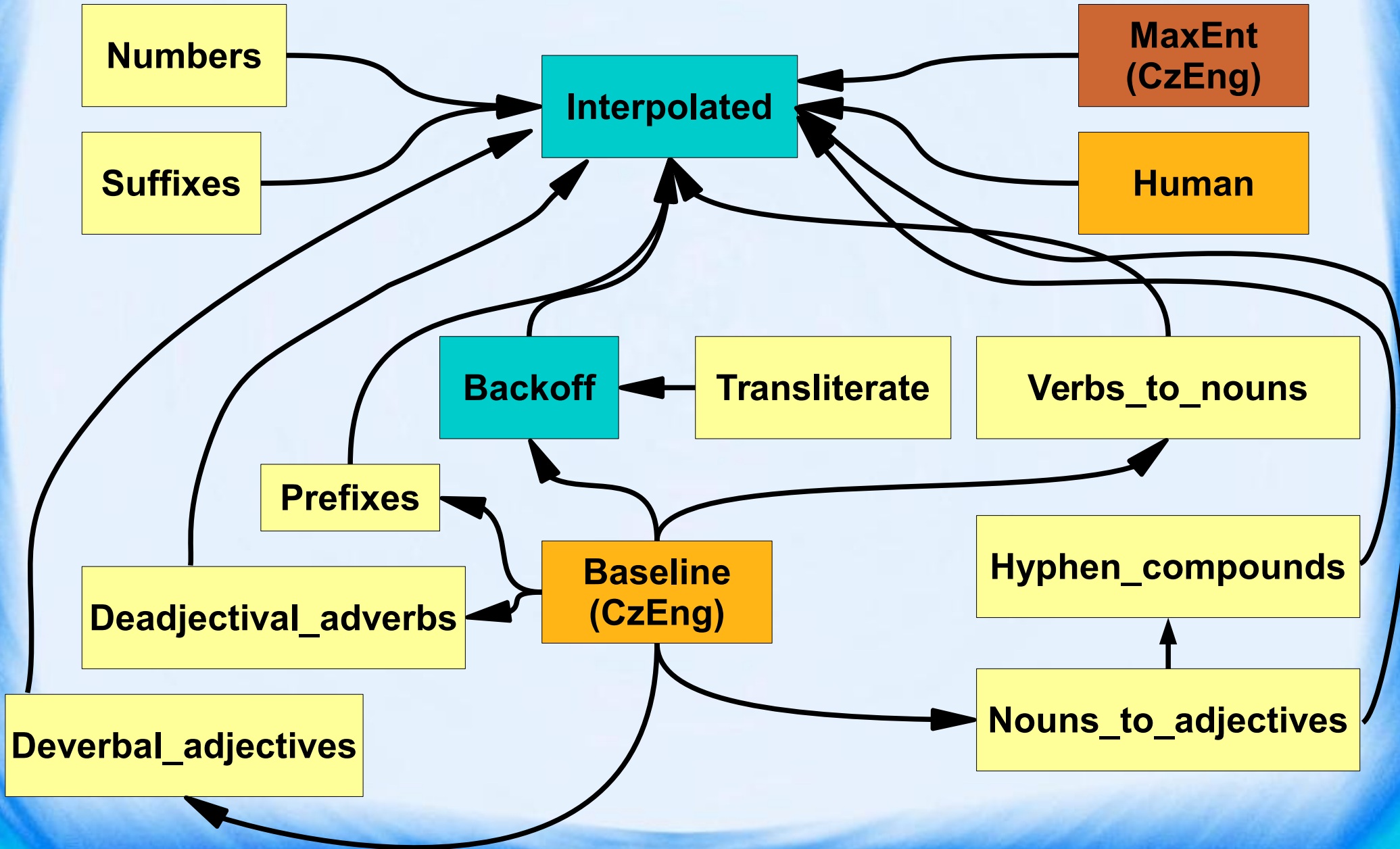
Hierarchy of lemma dictionaries



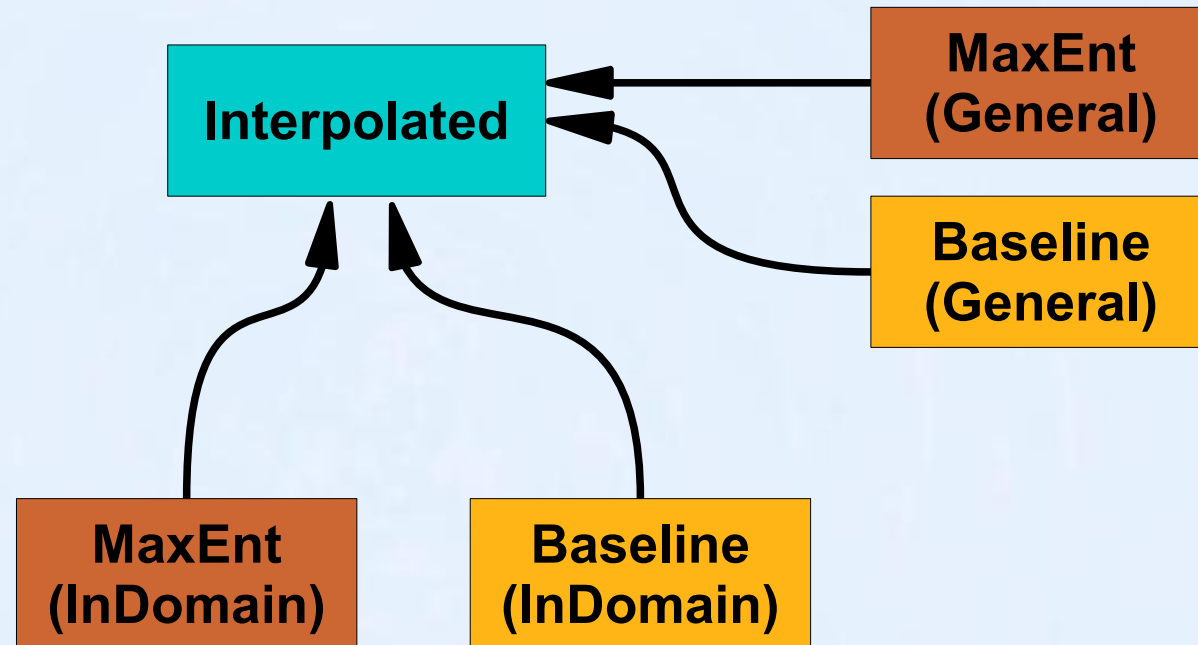
Hierarchy of lemma dictionaries



Hierarchy of lemma dictionaries



Domain adaptation by TM interpolation



```
T2T::TrLAddVariantsInterpol model_dir=data/models/translation/en2nl models='
static 0.5 20150217_tlemma.static.gz
maxent 1.0 20150217_tlemma.maxent.gz
static 0.5 IT/20150725_batch1a-tlemma.static.gz
maxent 1.0 IT/20150725_batch1a-tlemma.maxent.gz'
```


Maximum Entropy Dictionary

Baseline Dictionary

$$p(y|x) = \frac{\text{count}(x, y)}{\text{count}(x)}$$

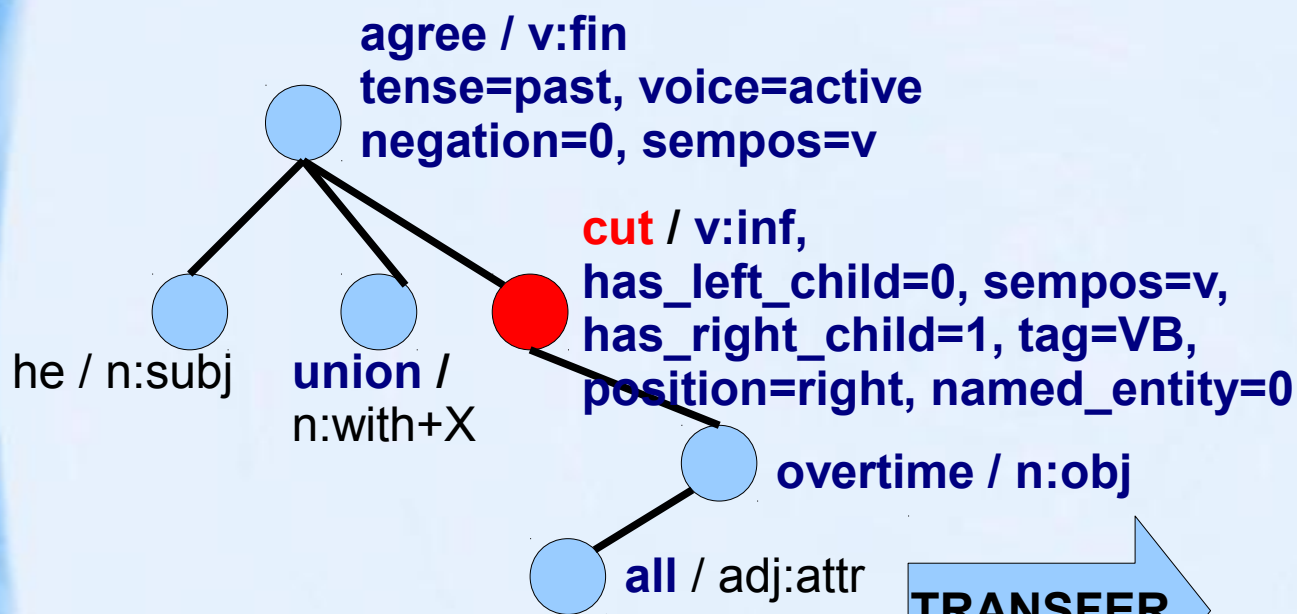
- Maximum likelihood estimates (from the training sections of CzEng)
- Pruned by thresholds on $p(x|y)$ and $p(y|x)$
- No context used
x = source lemma
y = target lemma

MaxEnt Dictionary

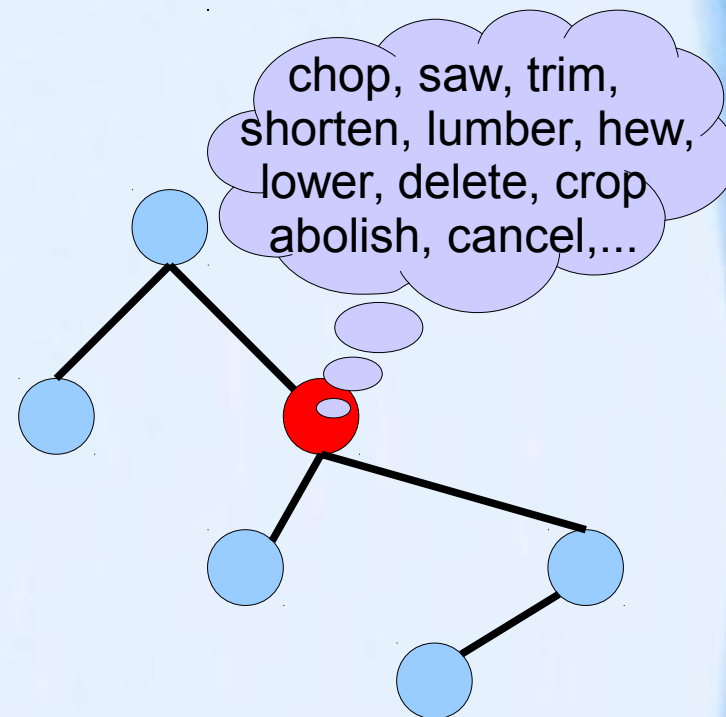
$$p(y|x) = \frac{1}{Z(x)} \exp \sum_i \lambda_i f_i(x, y)$$

- One MaxEnt model for each source lemma (same training data as for the Baseline Dict.)
- Interpolated with Baseline Dict. (due to pruning)
- Context features used (x = source context)
 - local tree context
 - local linear context
 - morphological & syntactic categories
 - ...

Maximum Entropy Dictionary



TRANSFER



ANALYSIS

SYNTHESIS

He agreed with the unions to cut all overtime.

Dohodl se s odbory na zrušení všech přesčasů.

Examples of Translation (2009)

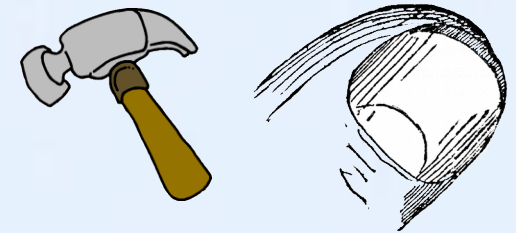
A miss by an inch
is a miss by a mile.

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.

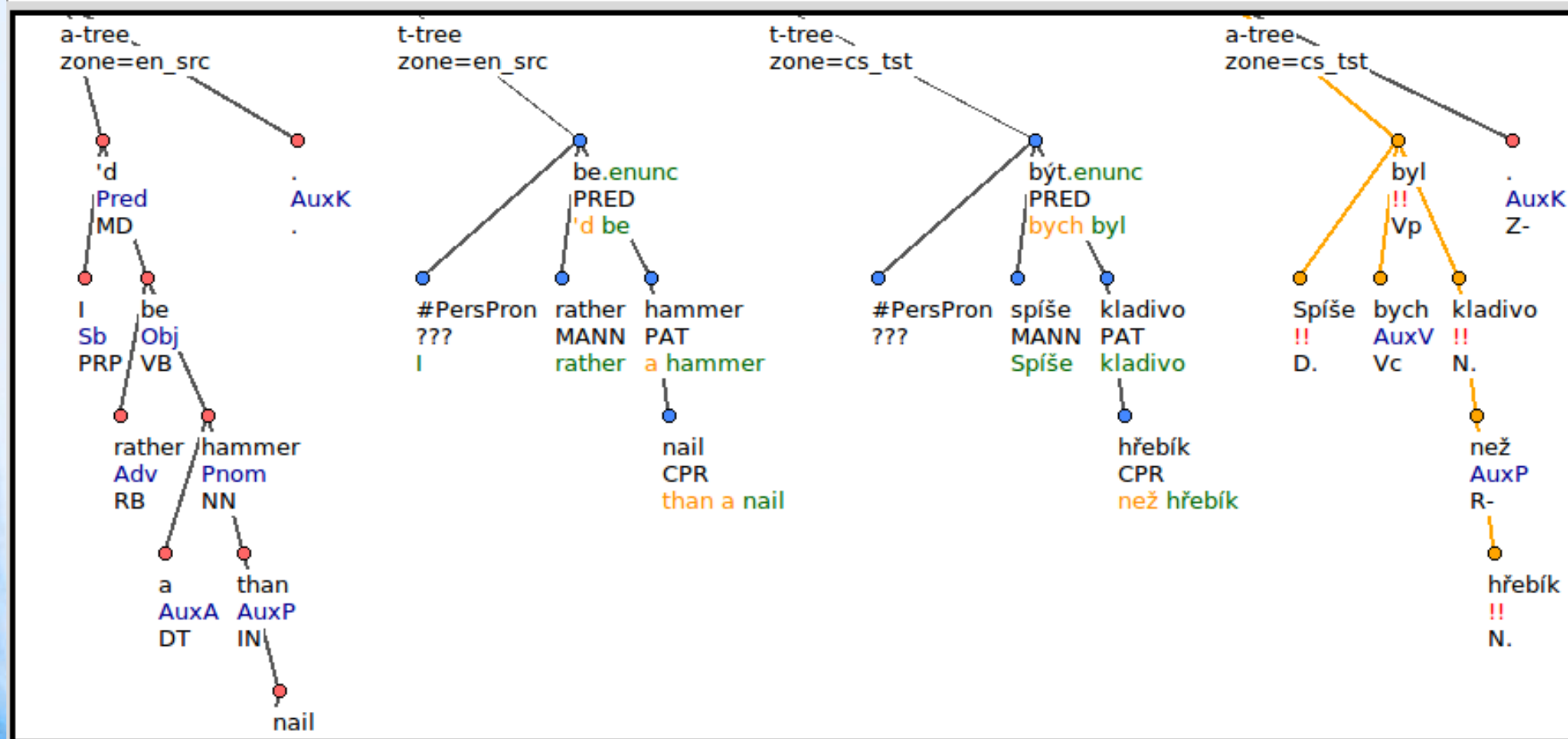
Pták v ruce je cenný
dvakrát v Bushovi.



Example of Translation (2011)

File Node Tree View Macros Setup Help

[cs_tst] Spíše bych byl kladivo než hřebík .
 [en_src] I'd rather be a hammer than a nail.



The image displays four syntax trees illustrating the translation process:

- Tree 1 (a-tree, zone=en_src):** Represents the English source sentence. The root node is 'd (Pred MD), which branches into 'I (Sb PRP) and 'be (Obj VB). 'be branches into 'rather (Adv RB) and 'hammer (Pnom NN). 'hammer branches into 'a (AuxA DT) and 'than (AuxP IN). 'than branches into 'nail (CPR).
- Tree 2 (t-tree, zone=en_src):** Represents the English target sentence. The root node is 'be.enunc (PRED), which branches into '#PersPron (MANN) and 'rather (PAT). '#PersPron branches into 'I (MANN). 'rather branches into 'rather (MANN). 'rather branches into 'hammer (PAT) and 'nail (CPR). 'hammer branches into 'a (MANN) and 'hammer (PAT).
- Tree 3 (t-tree, zone=cs_tst):** Represents the Czech target sentence. The root node is 'být.enunc (PRED), which branches into '#PersPron (MANN) and 'spíše (PAT). '#PersPron branches into '??? (MANN). 'spíše branches into 'spíše (MANN). 'spíše branches into 'kladivo (PAT) and 'hřebík (CPR). 'kladivo branches into 'kladivo (PAT).
- Tree 4 (a-tree, zone=cs_tst):** Represents the Czech source sentence. The root node is 'byl (Vp), which branches into 'Spíše (D.), 'bych (AuxV Vc), and 'kladivo (N.). 'byl branches into 'byl (AuxV Vc). 'kladivo branches into 'než (AuxP R-) and 'hřebík (N.). 'než branches into 'než (AuxP R-). 'hřebík branches into 'hřebík (N.).

Sample of MaxEnt Features

input_label=nail

output_label=hřebík#N (metal nail)

child_formeme_n:in+X=1	1.64483855116042
is_member=1	1.30042900630692
child_formeme_v:fin=1	1.04422203176176
next_node_tlemma=down	0.838961007712912
is_capitalized=1	0.792130821958927
position=right	0.747785245407306
tense_g=post	0.744919903760696
voice_g=active	0.659489975893991
prev_node_tlemma=drive	0.655357850937254
parent_capitalized=1	0.622953832124697
formeme=n:from+X	0.599348506643414
prev_node_tlemma=hammer	0.592276691427986
child_tlemma_few=1	0.553464629114697
child_tlemma_remove=1	0.546698831608057
sempos=n.denot	0.504719359514573
next_node_tlemma=and	0.502529618088752
formeme_g=v:until+fin	0.491064112122981
child_tlemma_rusty=1	0.428884558837039
tag_g=VBP	0.422967377093101
next_node_tlemma=screw	0.344701934524519

...

output_label=nehet#N (fingernail or toenail)

child_formeme_n:poss=1	1.32717038827268
child_tlemma_finger=1	1.07509772743853
child_formeme_n:of+X=1	0.982021327950337
position=left	0.886912864256063
prev_node_tlemma=black	0.770671304450658
child_tlemma_broken=1	0.761077744287099
child_formeme_v:attr=1	0.700099311992958
formeme=n:at+X	0.674547829214778
formeme_g=n:attr	0.673367412957367
child_tlemma_long=1	0.673158400394094
next_node_tlemma=file	0.600496248030202
child_tlemma_false=1	0.584236638145312
prev_node_tlemma=false	0.584236638145312
number=sg	0.563056142428995
formeme=n:obj	0.533943098032196
formeme=n:by+X	0.528852315800188

...

Cooperation is welcome

- Exploit English a-layer or t-layer for your project (e.g. Extra features/factors for Moses)
- Adapt TectoMT for a new language pair
- Try your machine learning (CRF, SVM, NN,...) instead of our MaxEnt dictionary
- Suggest a better algorithm for the transfer (treelet → treelet)
- Use SMT for the deep-syntactic transfer (TectoMoses: linearize trees & project dependencies)

Thank you

