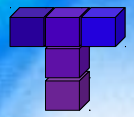
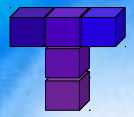


Deep-syntactic Machine Translation in the Treex NLP Framework

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

March 7, 2012, Prague, Czech Republic



Outline

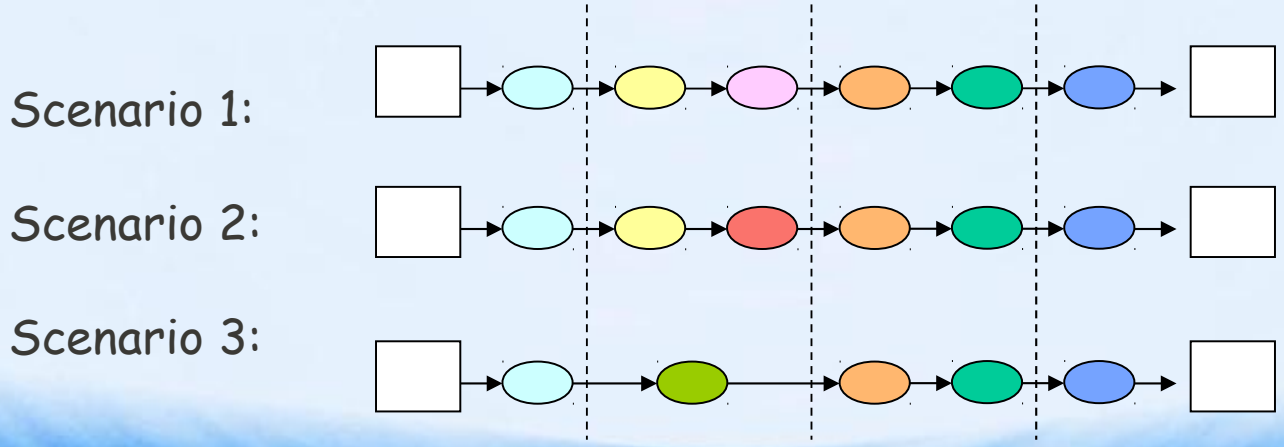
- Treex (Natural Language Processing framework)
  - Motivation
  - Applications
  - Treex architecture
- TectoMT (deep-syntactic Machine Translation)
  - Translation scenario overview
  - Hidden Markov Tree Models (HMTM)
  - Maximum Entropy dictionary
  - Results and translation examples

Treex

Motivation
Goals of Treex

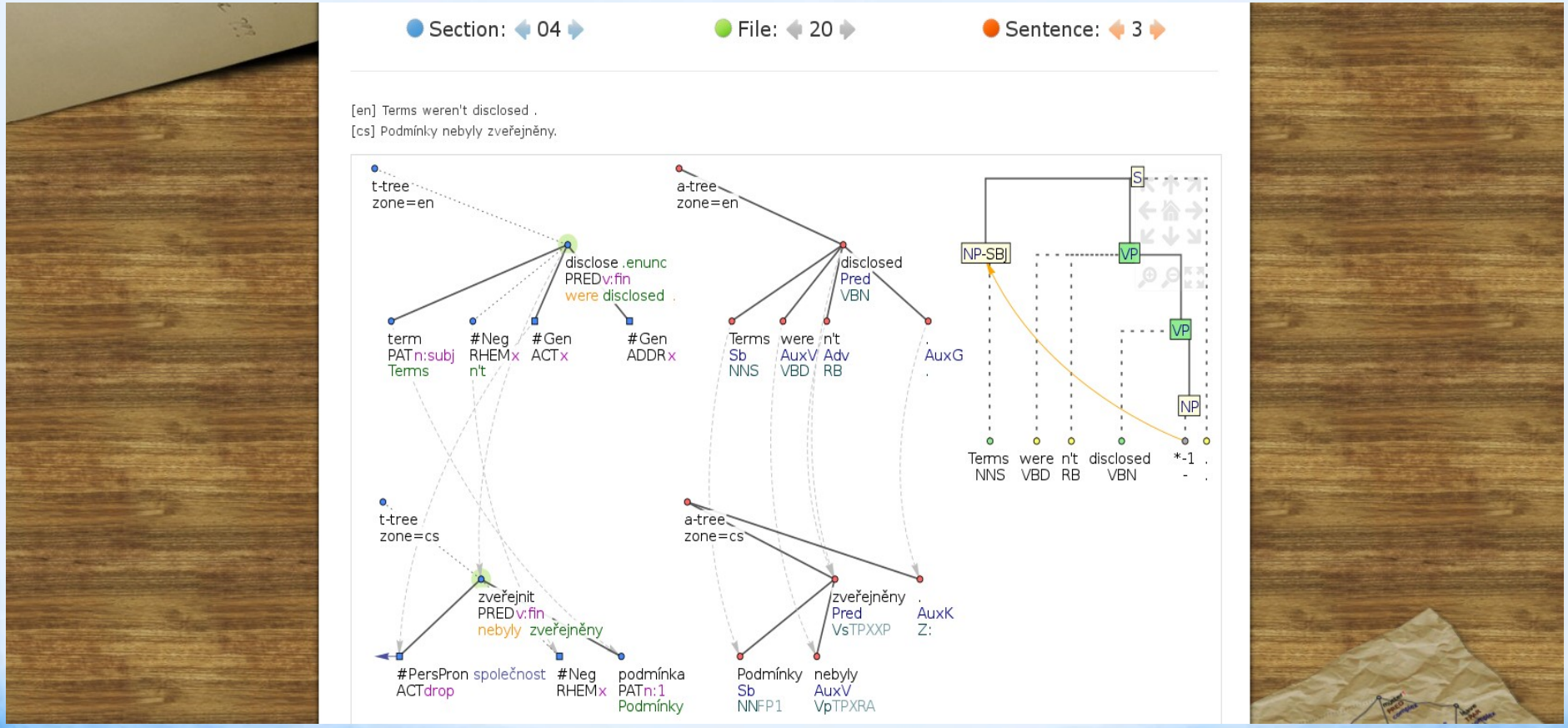
- elegant integration of in-house and third-party NLP tools
- modularity, reusability, cooperation
- ability to easily modify and add code in a full-fledged programming language

- lab for NLP experiments (fast prototyping)



Preprocessing for manual annotation
PEDT 2.0 and PCEDT 2.0

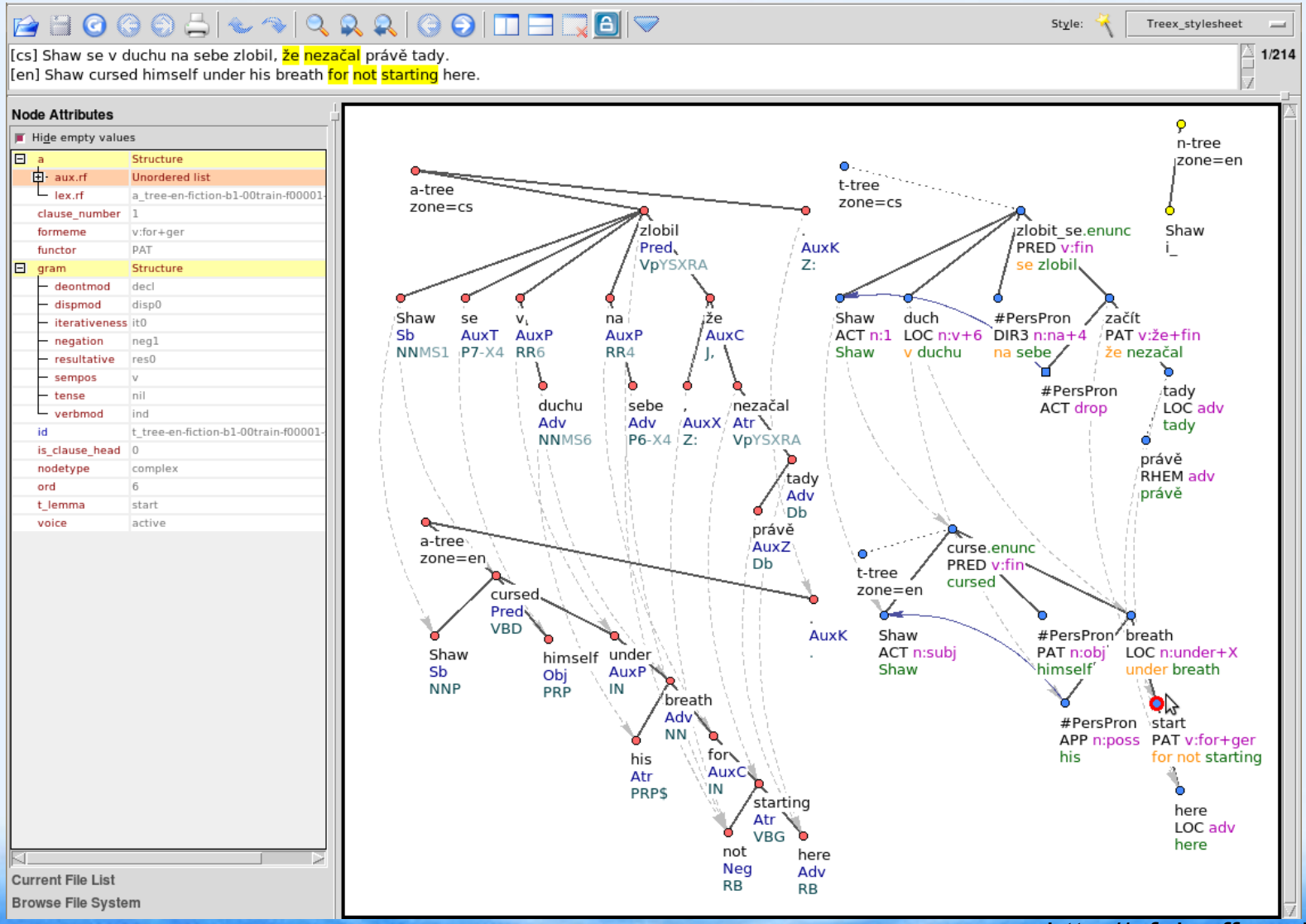
- WSJ data, PennTB phrase structure,
  two layers of dependency structure, semantic labeling, ...
- Treex used for converting phrase trees to dependencies,
  syntactic labeling, word alignment, pre-annotation



Fully automatic annotation

CzEng 1.0 (Czech-English parallel treebank)

- 15 million parallel sentences (> 200 MW), free for research purposes
- morphology, shallow and deep syntactic layer, rich annotation



http://ufal.mff.cuni.cz/czeng/czeng10

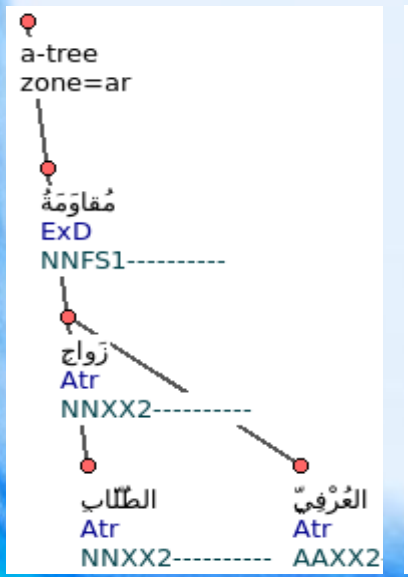
Treebank conversion

Treex

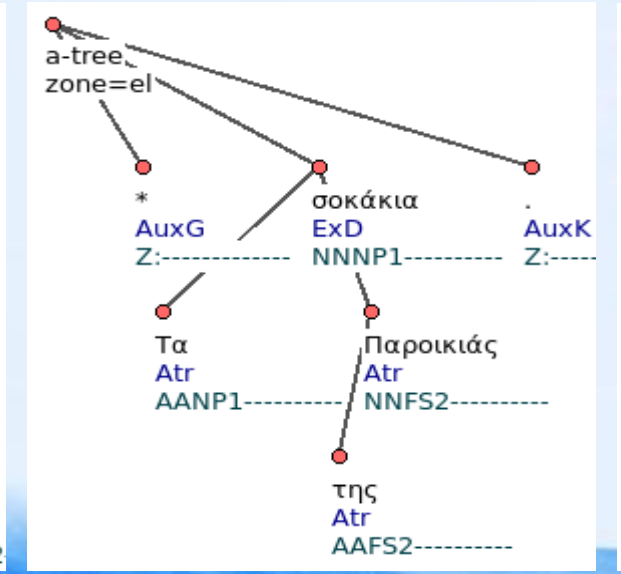
HamleDT – HArmonized Multi-LanguagE Dependency Treebank

- Dependency treebanks for 30 languages
- Normalized to a unified style and format
- Scripts for converting among several annotations styles (e.g. treatment of coordination structures)
- Basque, Bengali, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, German, ancient Greek, Hindi, Hungarian, Italian, Japanese, Persian, Portuguese, Romanian, Slovene, Spanish, Swedish, Tamil, Telugu, Turkish

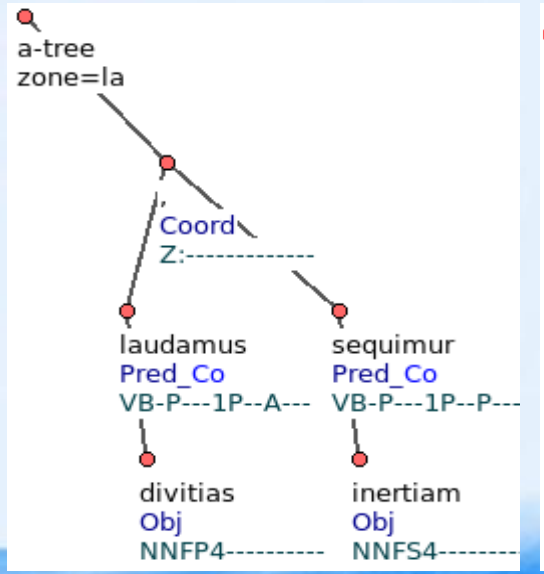
Arabic



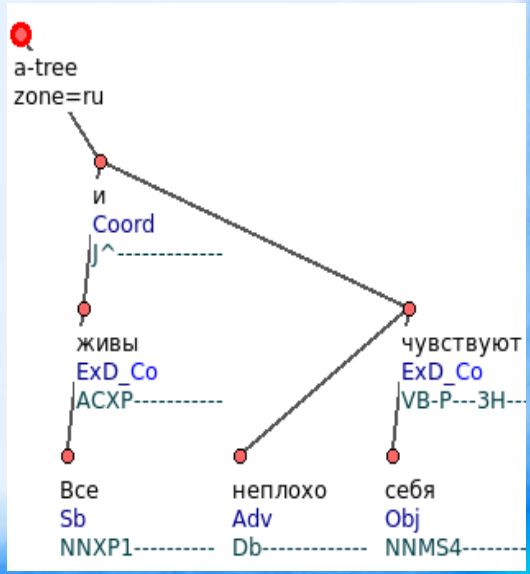
modern Greek



Latin



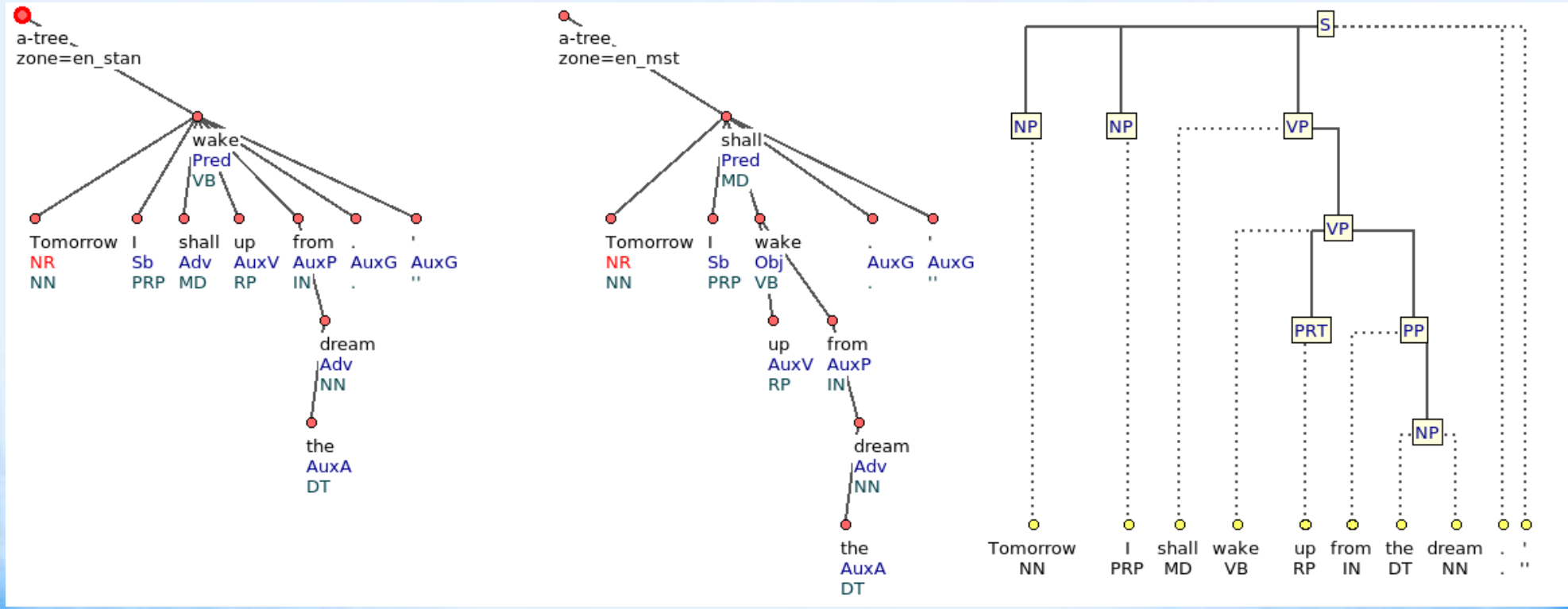
Russian



Evaluating and combining parsers

Parsing BNC

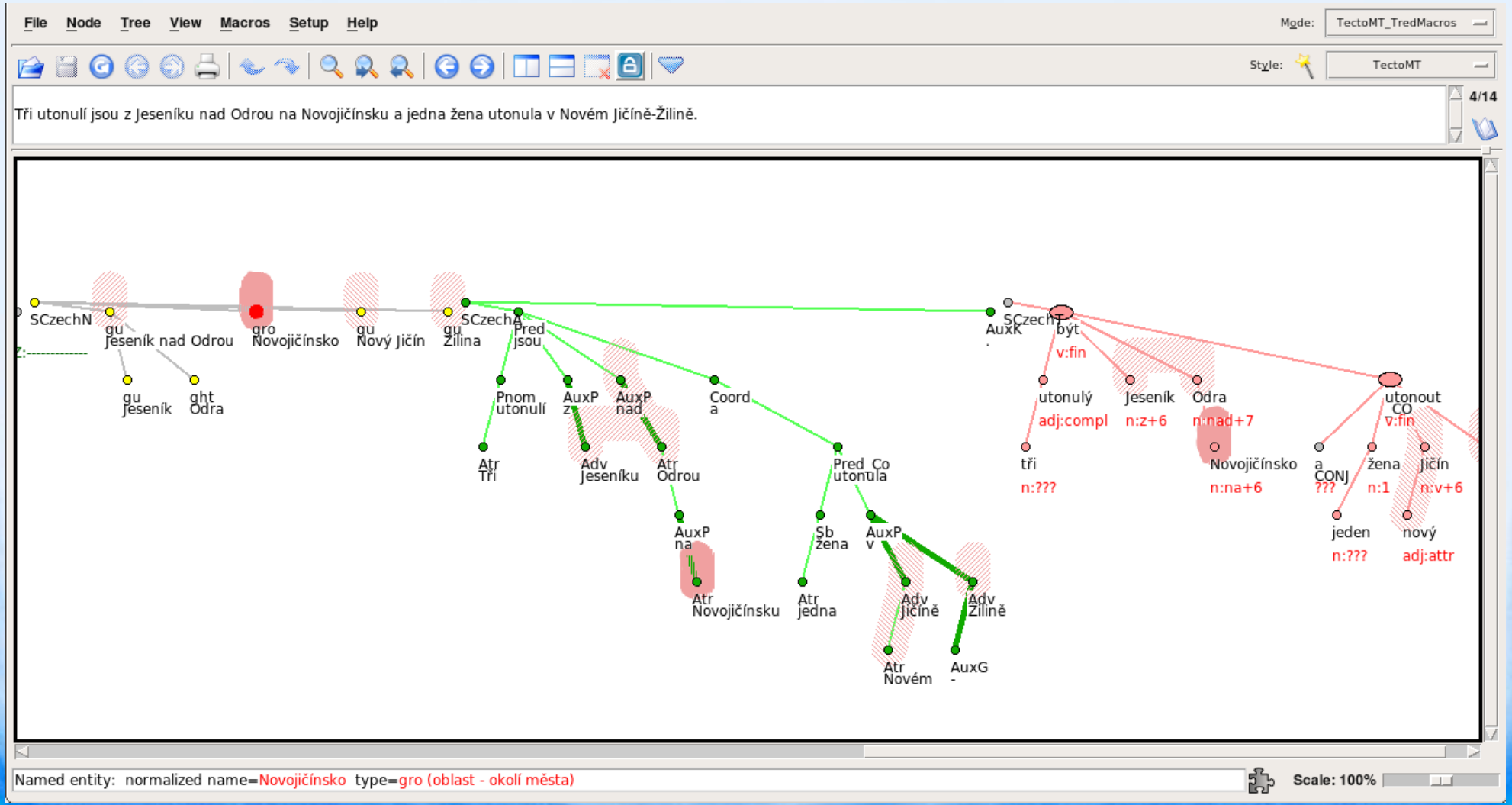
- Treex offers 3 constituency parsers (Stanford, Charniak's, Collins'), 2 phrase-to-dep converters, several dependency parsers (MST, Malt,...), ensemble combination
- British National Corpus (BNC) parsed to create training data for Lexical Disambiguation of English Verbs



Treex

Named entity recognition

- Wrapper for Stanford NER
- Support for nested named entities and visualization



Treex

Semantic processing

Dialog Manager in “Companions” Project

- Automatic Speech Recognition,
  Natural Language Understanding,
  Natural Language Generation,
  Text to Speech Synthesis

- Treex used for Semantic parsing tasks
  - Assignment semantic roles
  - Coordinations
  - Argument structure
  - Partial ellipsis resolution
  - Pronominal anaphora resolution

Comparing two MT systems

ComparEval – a tool for MT analysis

Treex

Show: Sentences ☑ info ☑ names ☑ src ☑ ref ☑ GOOGLE ☑ TECTOMT

N-gram stats ☑ confirmed ☑ unconfirmed

1-gram			
GOOGLE wins		TECTOMT wins	
433	,	-18	který
132	na	-18	;
106	v	-16	je
83	se	-12	by
65	-	-11	byla
50	o	-9	nebo
42	pro	-9	liber
37	let	-9	USA
35	za	-9	také
30	do	-7	už

2-gram			
GOOGLE wins		TECTOMT wins	
65	" .	-16	, který
39	, že	-13	, říká
32	, a	-6	že by
20	. 000	-5	. "
18	, kteří	-5	) a
12	al -	-4	Je to
11	, což	-4	pátek ,

n-grams unconfirmed by the reference

1-gram			
GOOGLE loses		TECTOMT loses	
635	,	-142	je
424	-	-50	roků
206	se	-47	který
175	na	-42	byla
123	v	-39	jsou
74	pro	-31	miliónů
70	k	-31	jejich
67	bude	-31	roku
52	o	-29	která

2-gram			
GOOGLE loses		TECTOMT	
194	, a	-88	, j
45	, že	-76	.
44	, "	-38	že
40	, se	-37	a
36	, v	-27	, js
34	a to	-25	, kt
30	- rok	-24	, kt
28	rok -	-22	Je
26	, co	-21	a
5	, což	-20	to

sentence #711 ID: ? Matching n-grams: GOOGLE - TECTOMT = -32

SRC:
This is part of the reason why I have decided to join the big march -- to pass on the word and to appeal to the world's leaders to deliver a fair, ambitious and binding deal, she said.

REF:
To je částečný důvod toho , proč jsme se rozhodli připojit se k tomuto velkému pochodu - - projít světem a apelovat na vůdčí osobnosti světa , aby předložili spravedlivou , ambiciózní a závaznou dohodu , řekla .

GOOGLE: 1-grams: 13, 2-grams: 4, 3-grams: 1, 4-grams: 0
To je jedním z důvodů , proč jsem se rozhodl vstoupit do velké pochod - předat slovo a odvolat se světoví vůdci poskytovat spravedlivé , ambiciózní a závazná dohoda , dodala .

TECTOMT: 1-grams: 21, 2-grams: 14, 3-grams: 9, 4-grams: 6
Je to součást důvodu , proč jsem se rozhodl připojit se k velkému pochodu - předat slovo a podat vůdcům světa , dodají spravedlivou , ambiciózní a závaznou dohodu , řekla .

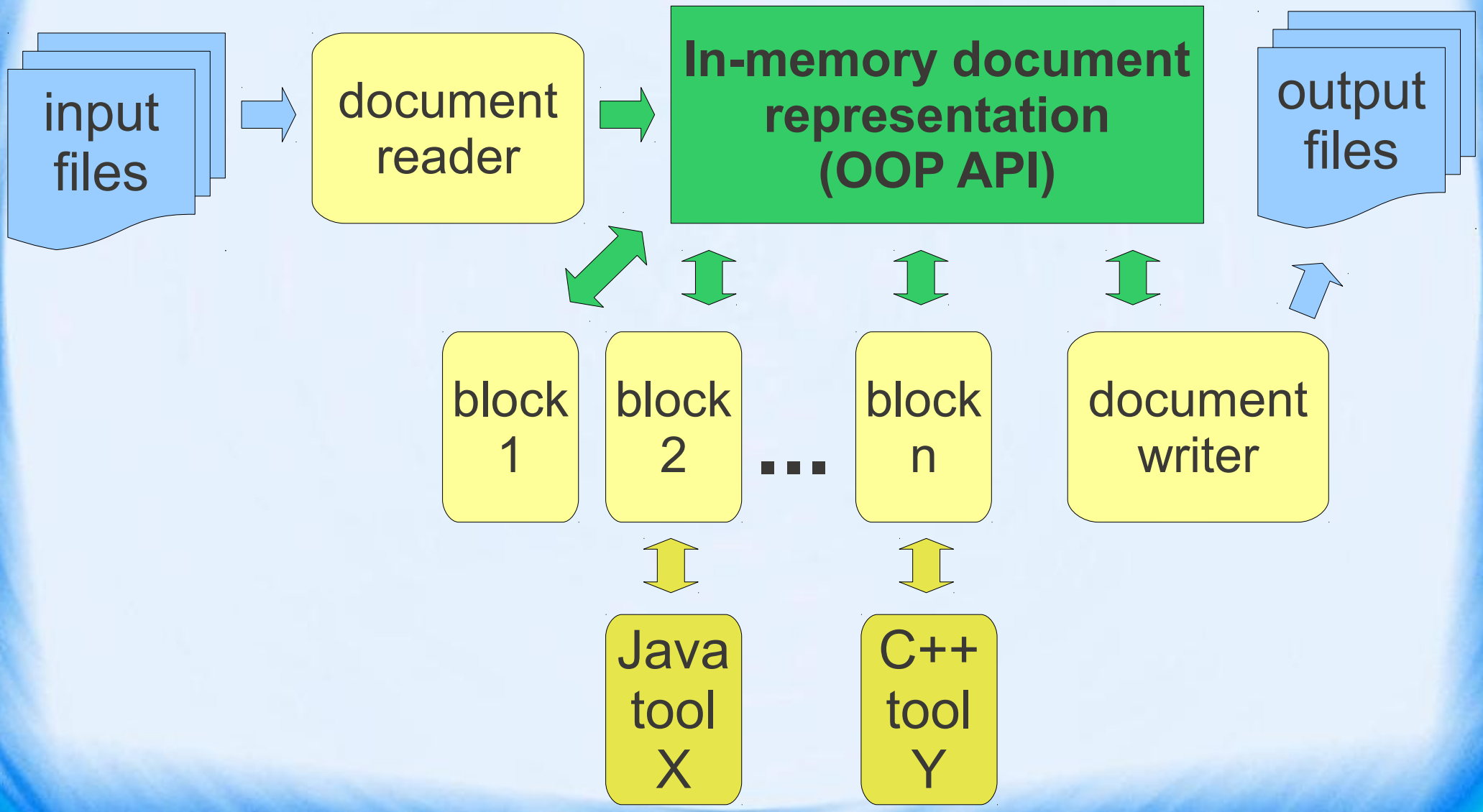
sentence #838 ID: ? Matching n-grams: GOOGLE - TECTOMT = -28

SRC:
The Obama administration is considering a package of sanctions that would target Iran's military and political elite, but Gates signaled that

Treex

Treex architecture

scenario



Treex

Treex architecture
block code example

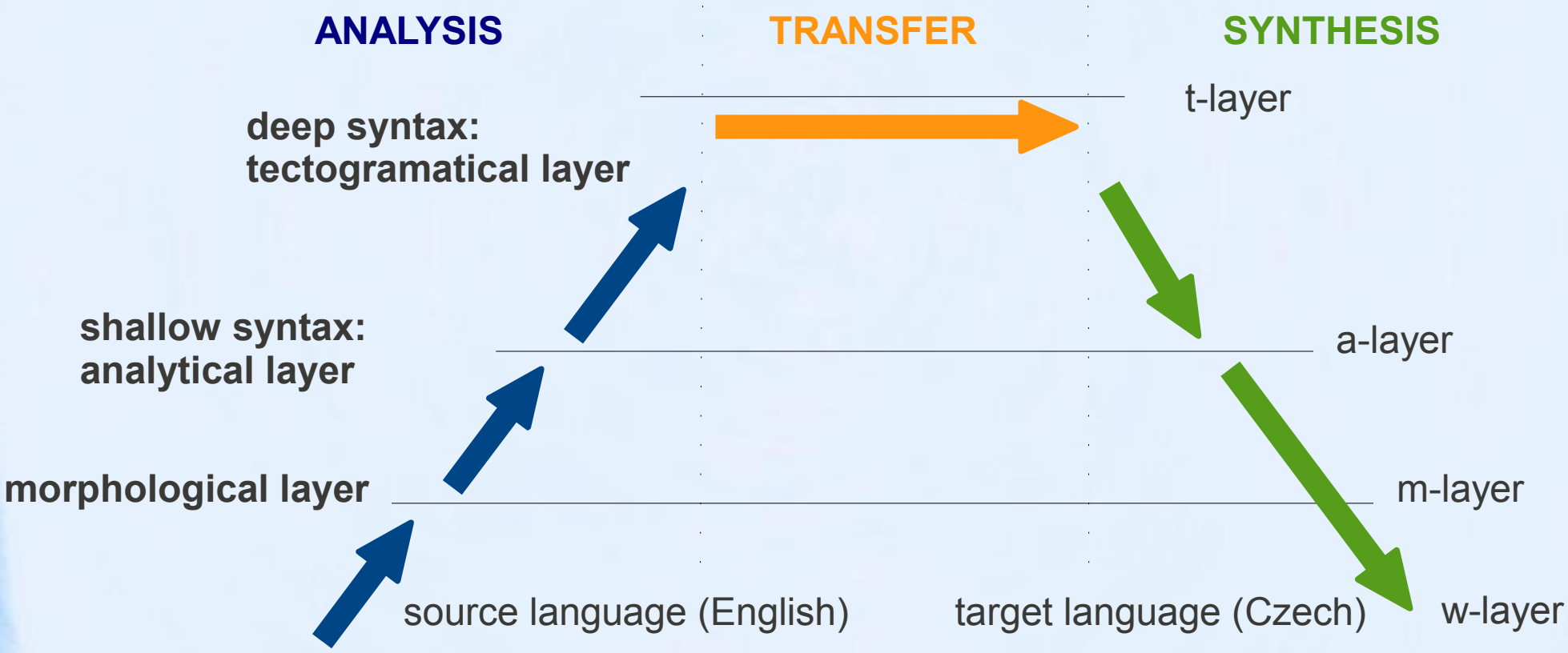
```
package Tutorial::Solution::Svo2Sov;
use Moose;
use Treex::Core::Common;
extends 'Treex::Core::Block';

sub process_anode {
  my ( $self, $a_node ) = @_;
  if ( $a_node->tag =~ /^V/ ) {          # verb found
    foreach my $child ( $a_node->get_echildren() ) {
      if ( $child->afun eq 'Obj' ) {     # object found
        # Move the object and its subtree so it precedes the verb
        $child->shift_before_node($a_node);
      }
    }
  }
  return;
}
1;
```

Treex core
Treex convention
Perl keyword/convention

TectoMT translation scheme

TectoMT

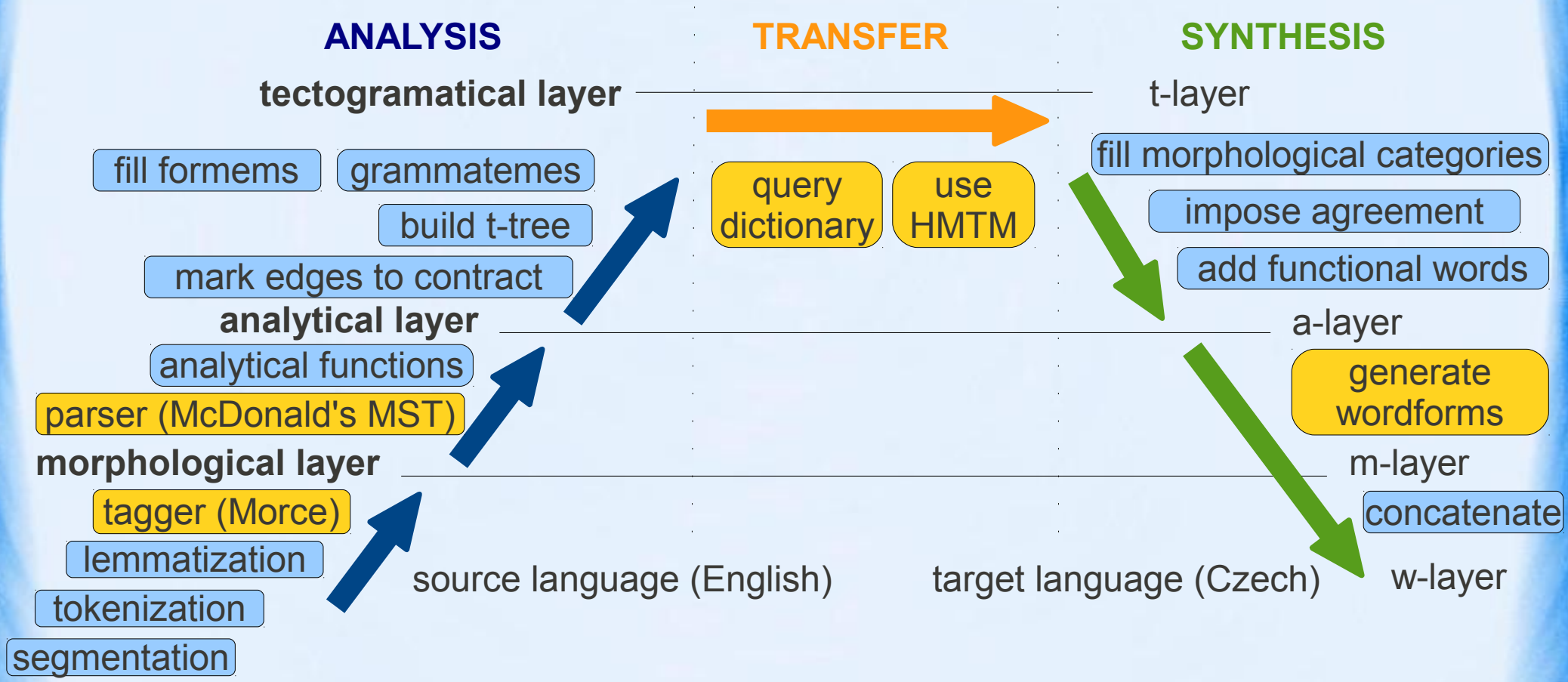
transfer over the tectogrammatical layer



TectoMT translation scheme

TectoMT

rule based & statistical blocks



The Real Scenario

TectoMT

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

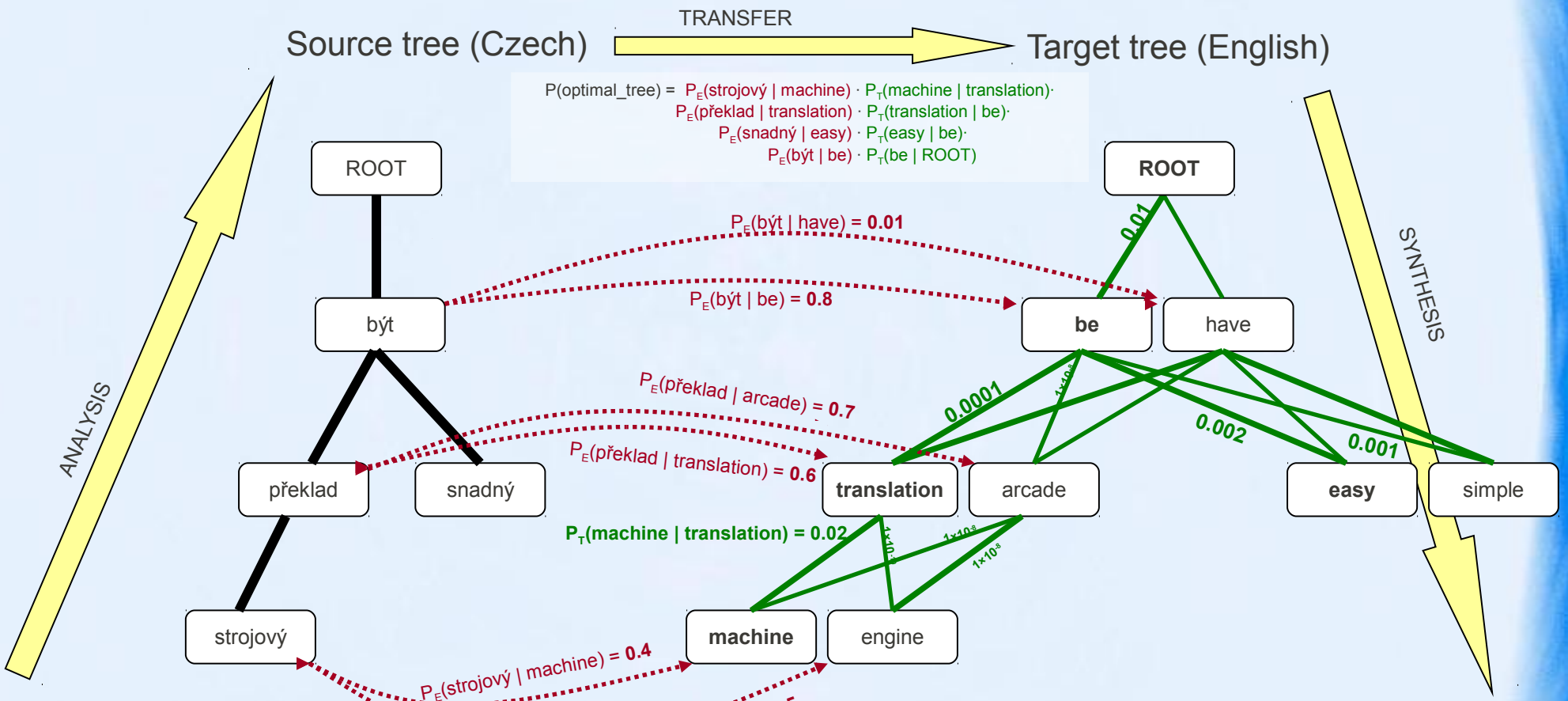
TRANSFER:
CopyTtree
TrLFPhrases
TrLFJointStatic
DeleteSuperfluousTnodes
TrFTryRules
TrFAddVariants
TrFRerank
TrLTryRules
TrLAddVariants
TrLFNumeralsByRules
TrLFilterAspect
TransformPassiveConstructions
PrunePersonalNameVariants
RemoveUnpassivizableVariants
TrLFCompounds
CutVariants
RehangToEffParents
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
FindGramCorefForReflPron
NeutPersPronGenderFromAntec
ValencyRelatedRules
SetClauseNumber
TurnTextCorefToGramCoref

SYNTHESIS TO A-LAYER:
CopyTtree
DistinguishHomonymous.
ReverseNumberNounDep.
InitMorphcat
FixPossessiveAdjs
MarkSubject
ImposePronZAgr
ImposeRelPronAgr
ImposeSubjpredAgr
ImposeAttrAgr
ImposeComplAgr
DropSubjPersProns
AddPrepos
AddSubconjs
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

HMTM Transfer

TectoMT



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**

Maximum Entropy Dictionary

TectoMT

Baseline Dictionary

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

- Maximum likelihood estimates (from the training sections of CzEng 0.9)
- Pruned by thresholds on *p(x|y)* and *p(y|x)*
- No context used
  x = source lemma
  y = target lemma

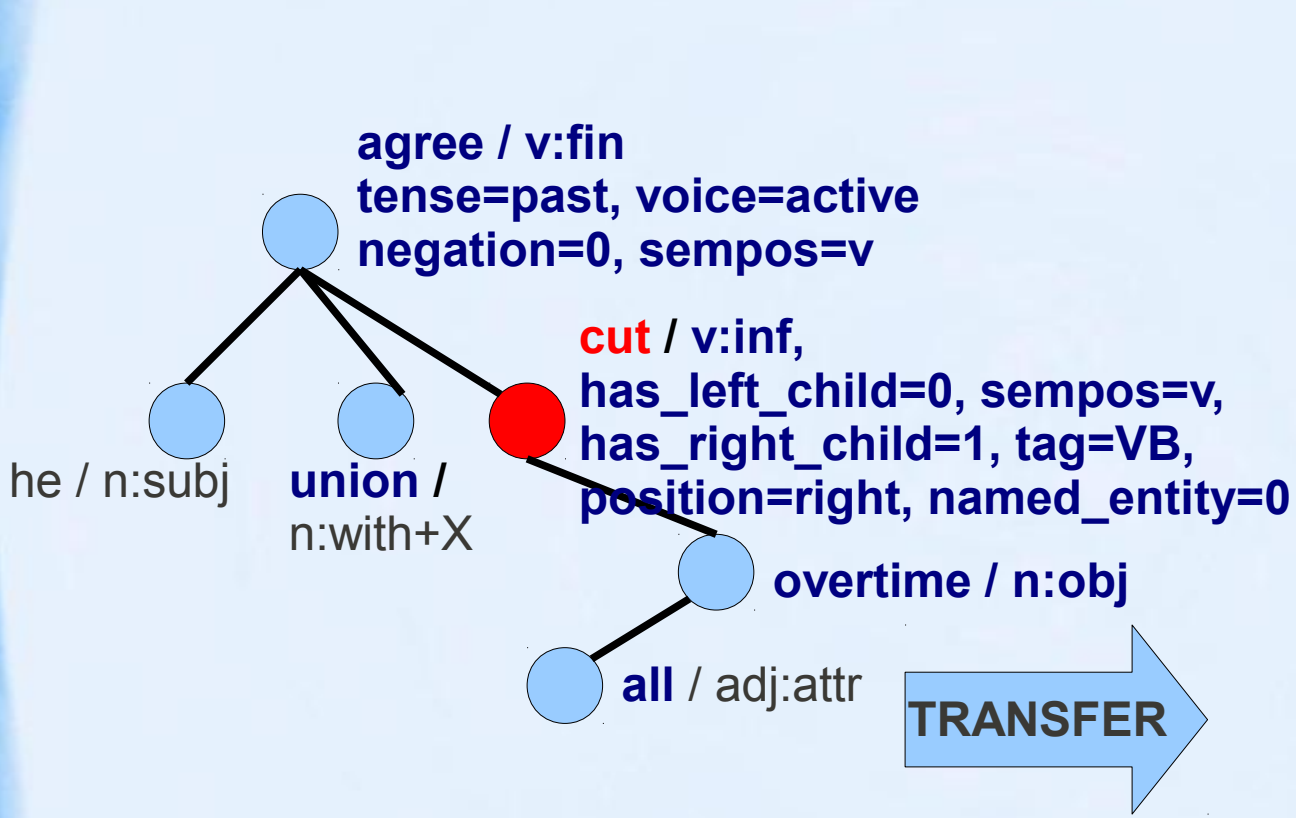
MaxEnt Dictionary

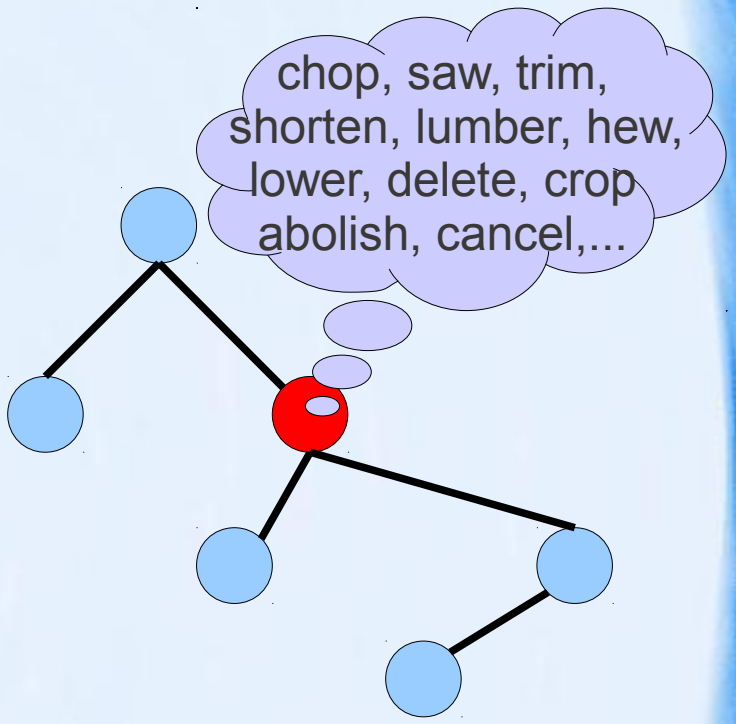
$$p(y|x) = \frac{1}{Z(x)} \exp \sum_i \lambda_i f_i(x, y)$$

- 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

TectoMT





ANALYSIS

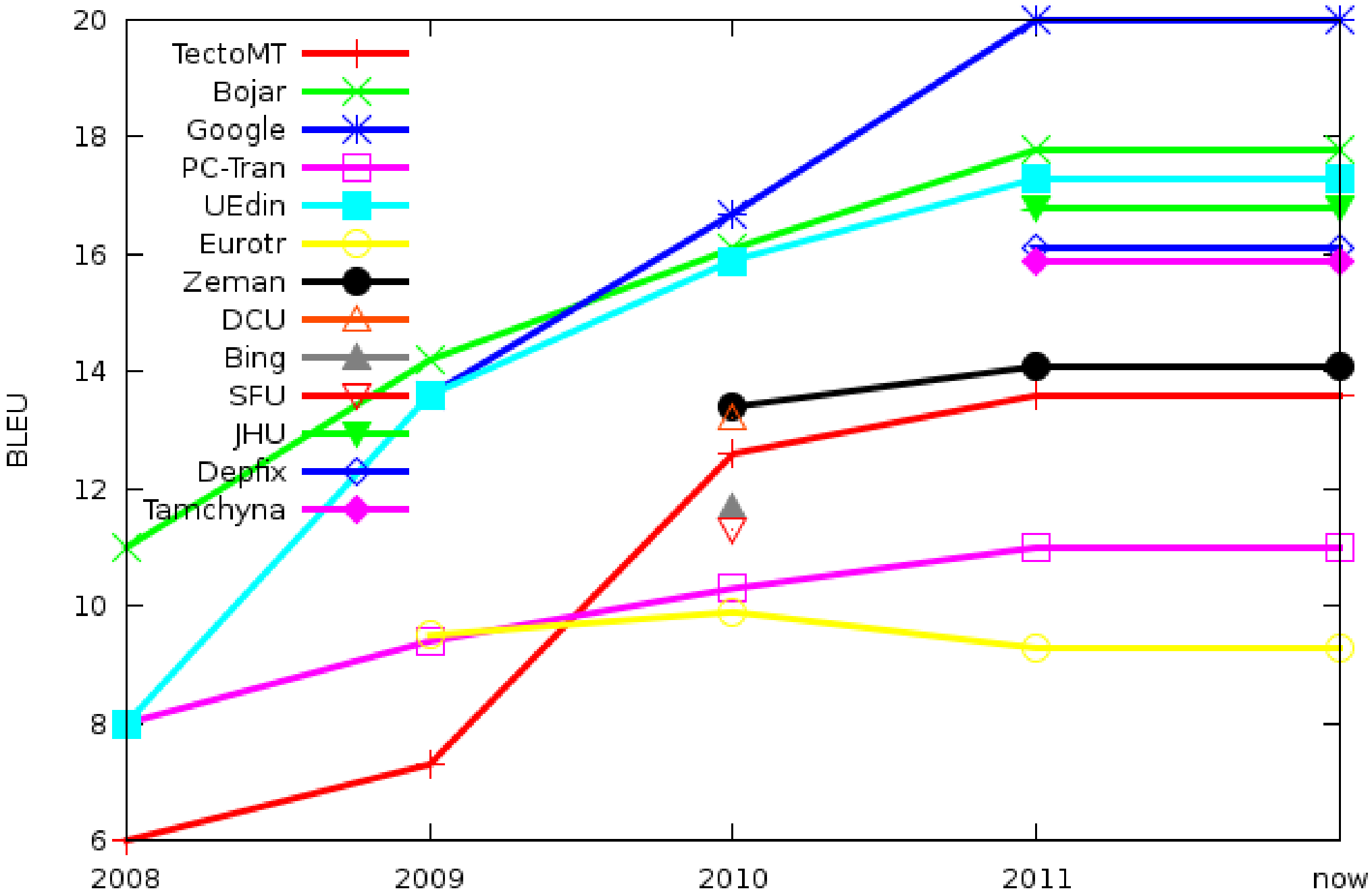
SYNTHESIS

He agreed with the unions to cut all overtime.

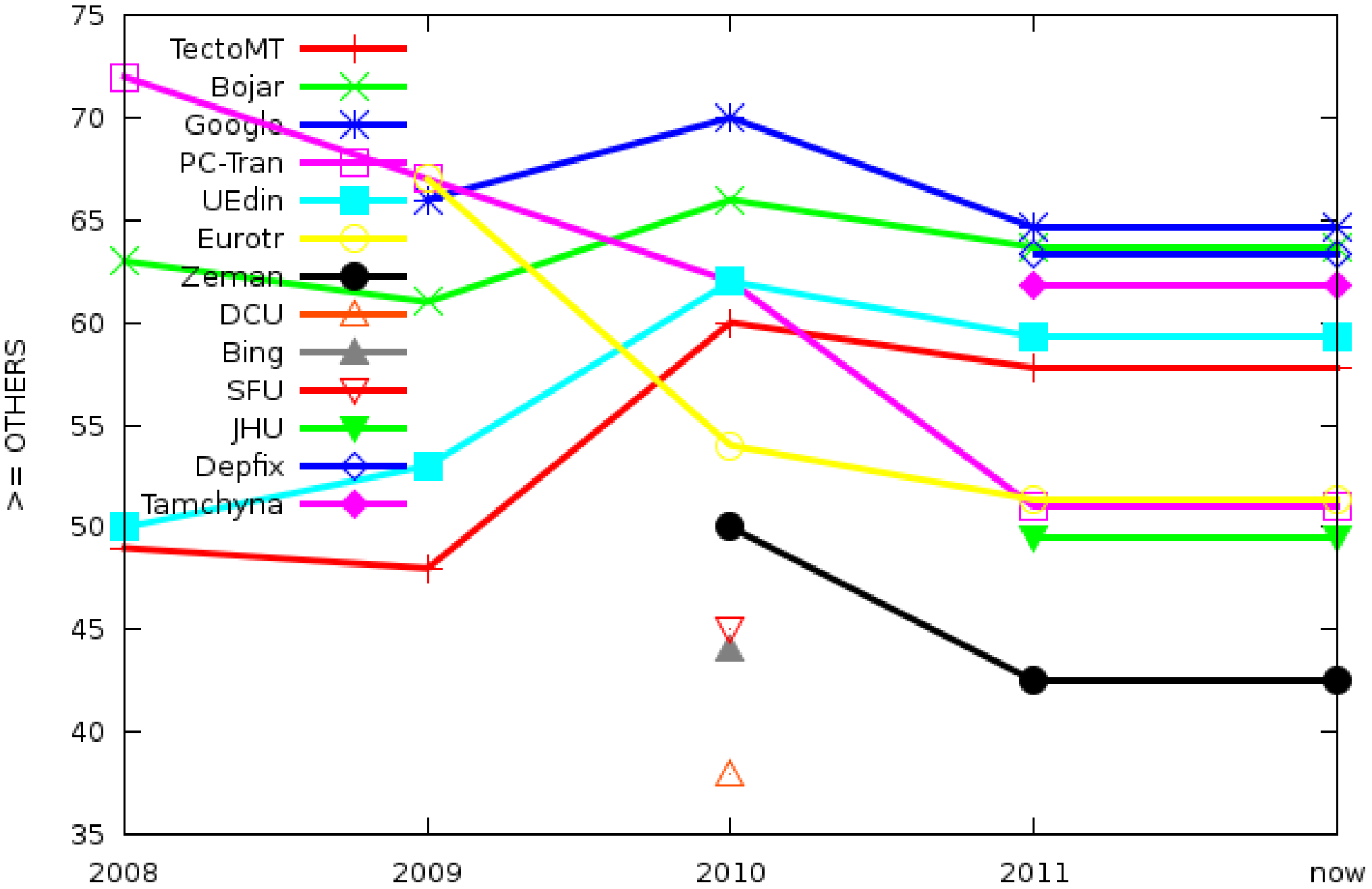
Dohodl se s odbory na zrušení všech přesčasů.

TectoMT

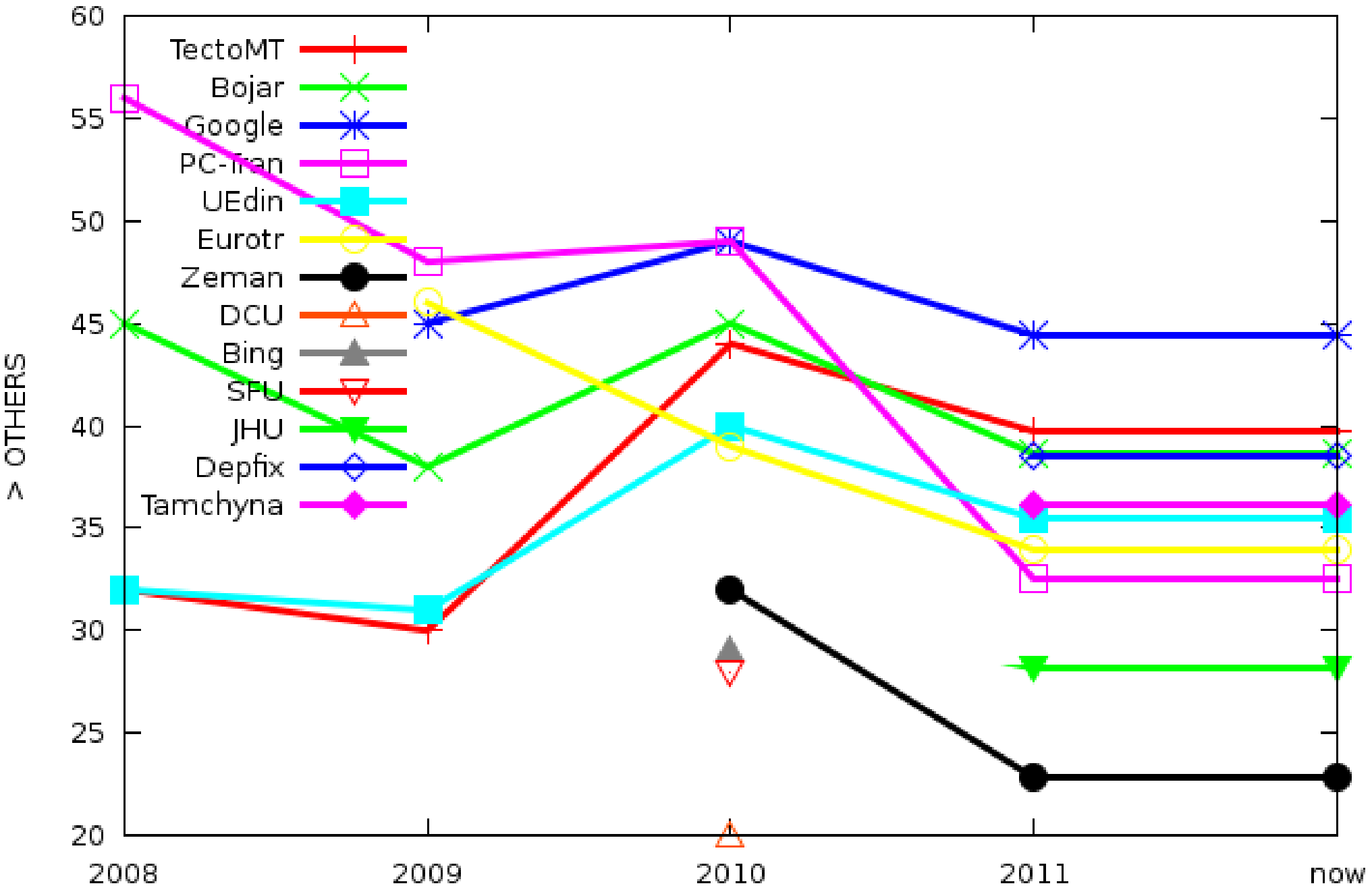
Results – BLEU



Results – manual ranking (≥ others)



Results – manual ranking (> others)



Examples of Translation (2009)

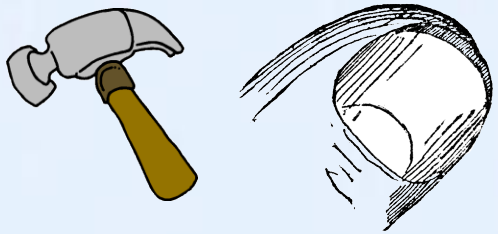
TectoMT

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.

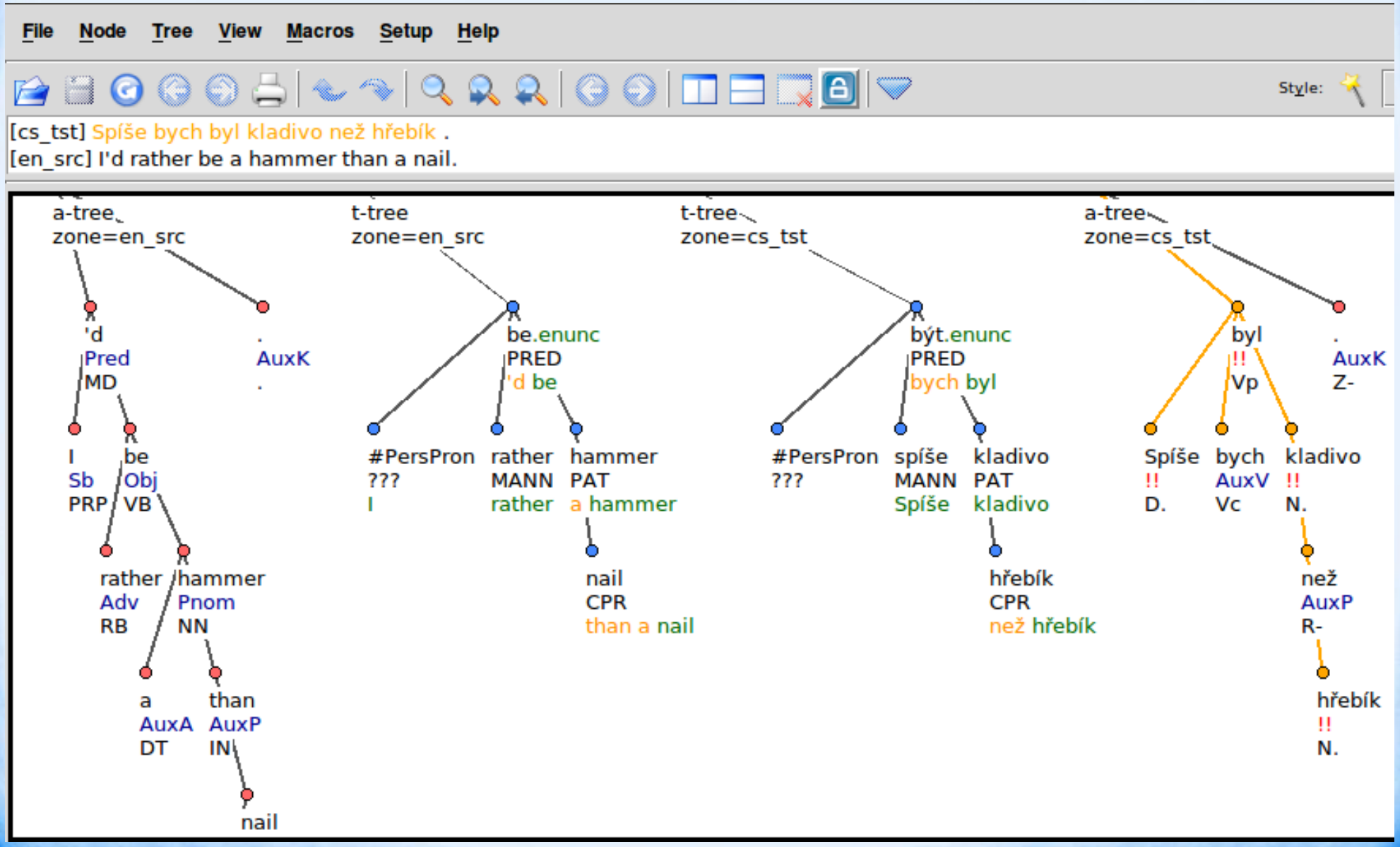


A bird in the hand is worth two in the bush.

Pták v ruce je cenný dvakrát v Bushovi.

Example of Translation (2011)

TectoMT



Sample of MaxEnt Features

TectoMT

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
...	

Treex

TectoMT

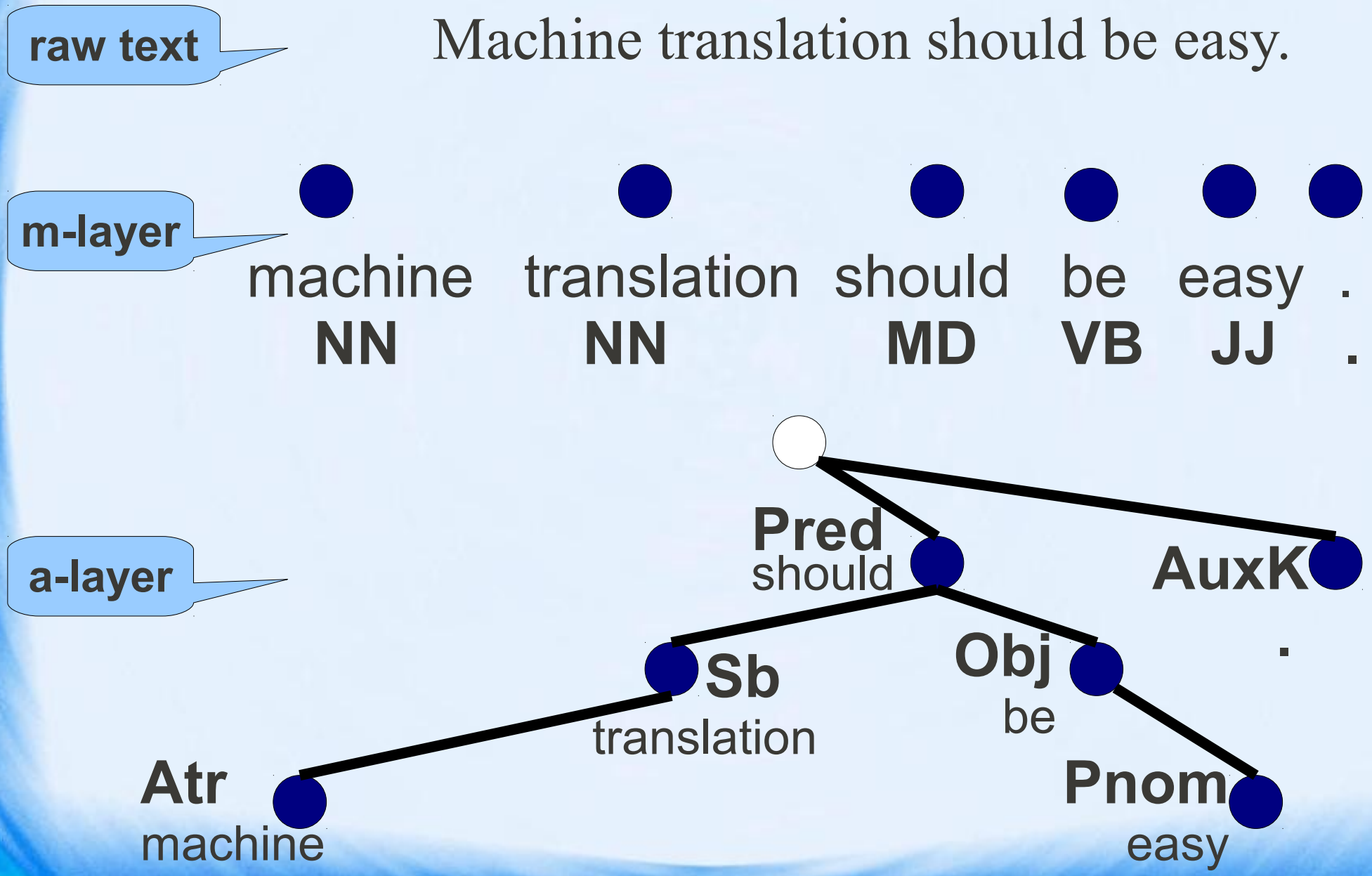
Thank you

Demo Translation – Analysis

TectoMT

raw text

Machine translation should be easy.

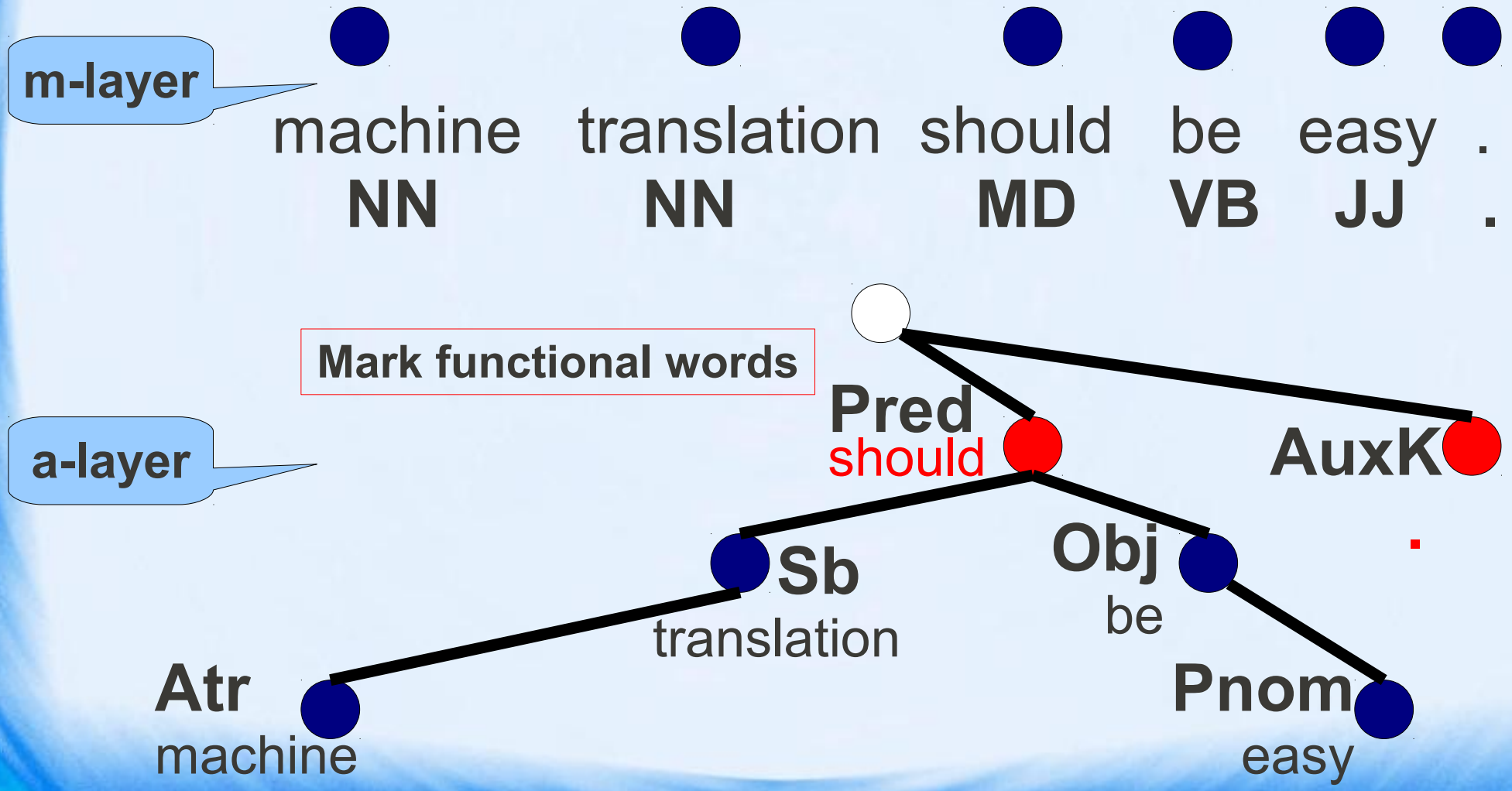


Demo Translation – Analysis

TectoMT

raw text

Machine translation should be easy.

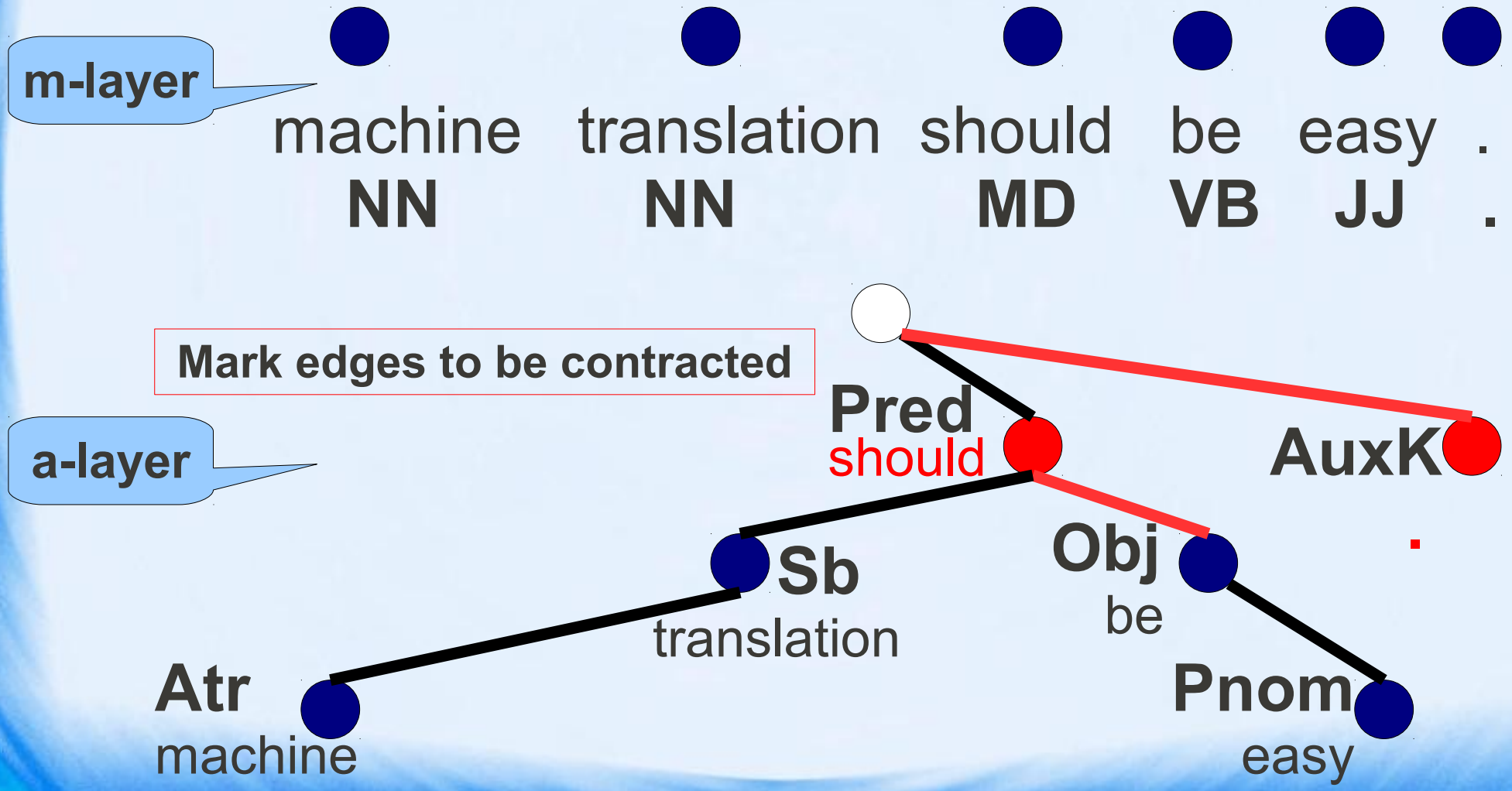


Demo Translation – Analysis

TectoMT

raw text

Machine translation should be easy.



Demo Translation – Analysis

TectoMT

raw text

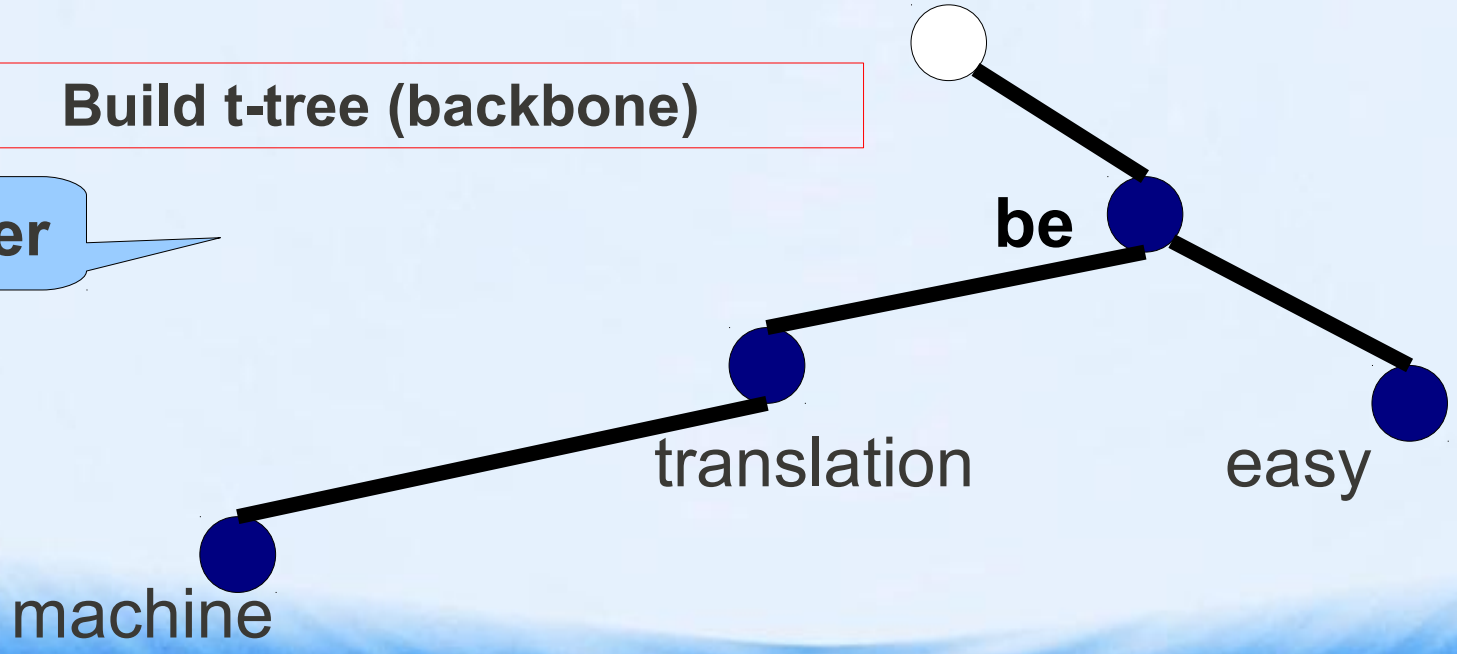
Machine translation should be easy.

m-layer

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

Build t-tree (backbone)

t-layer

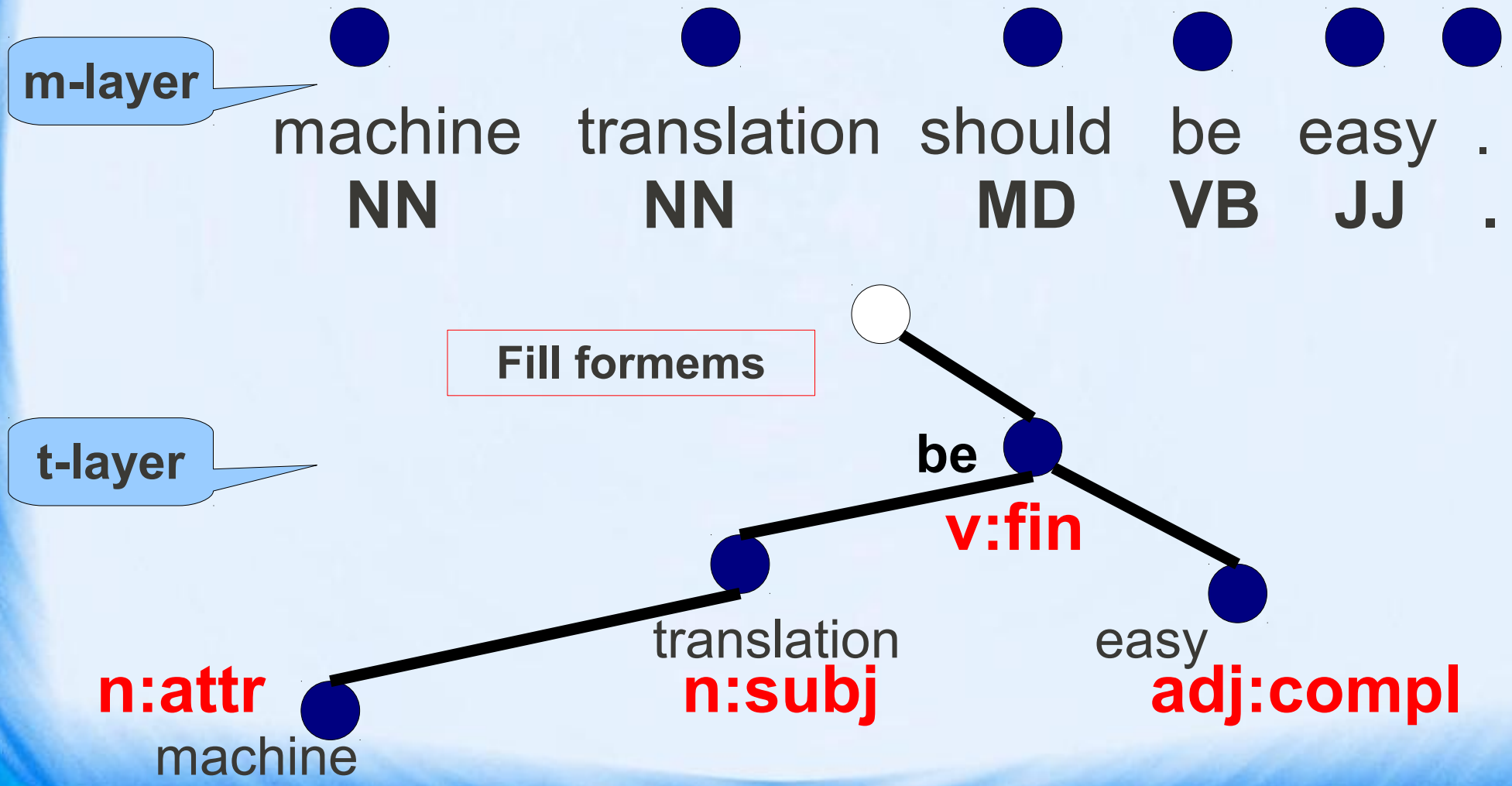


Demo Translation – Analysis

TectoMT

raw text

Machine translation should be easy.

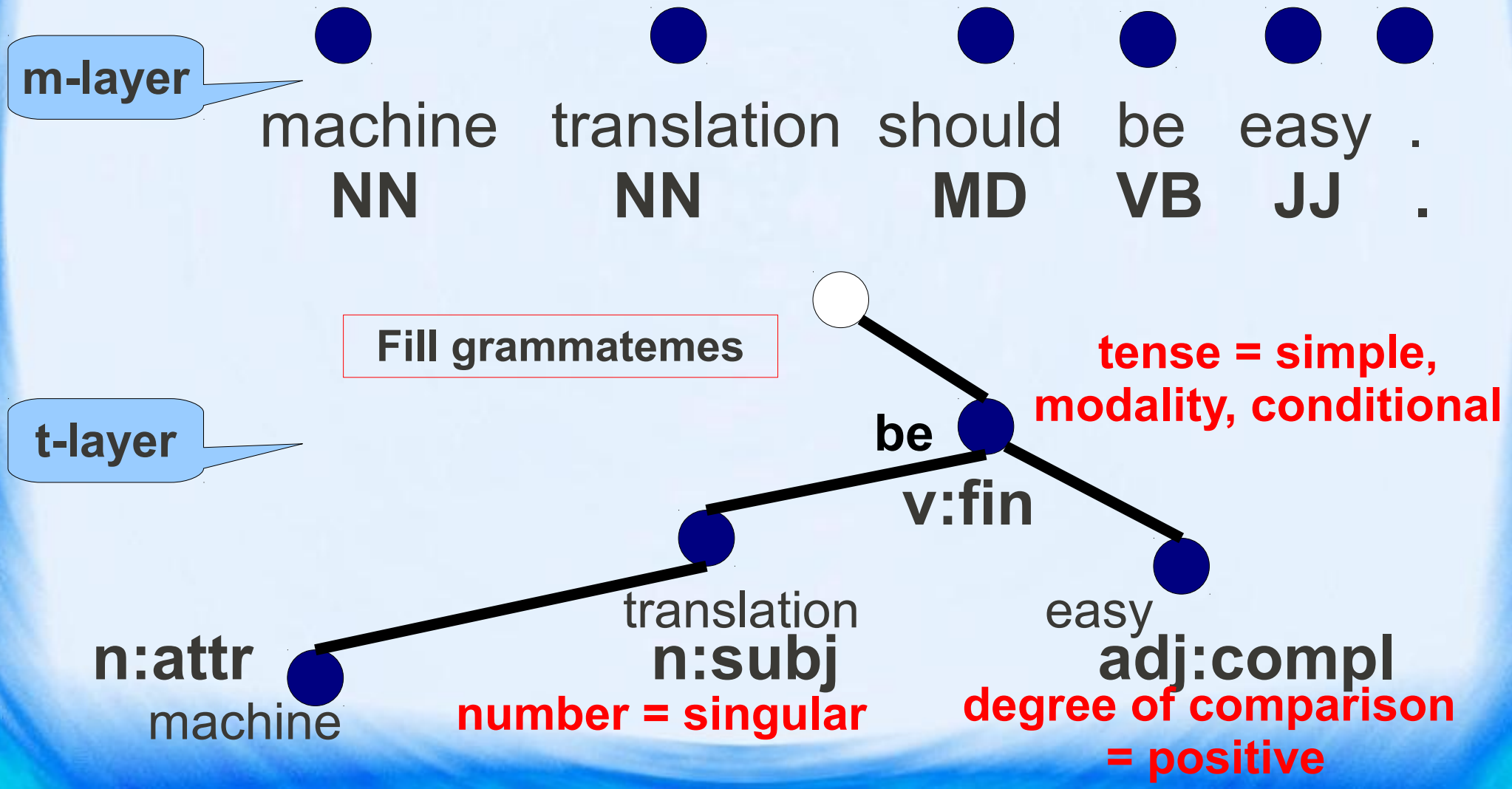


Demo Translation – Analysis

TectoMT

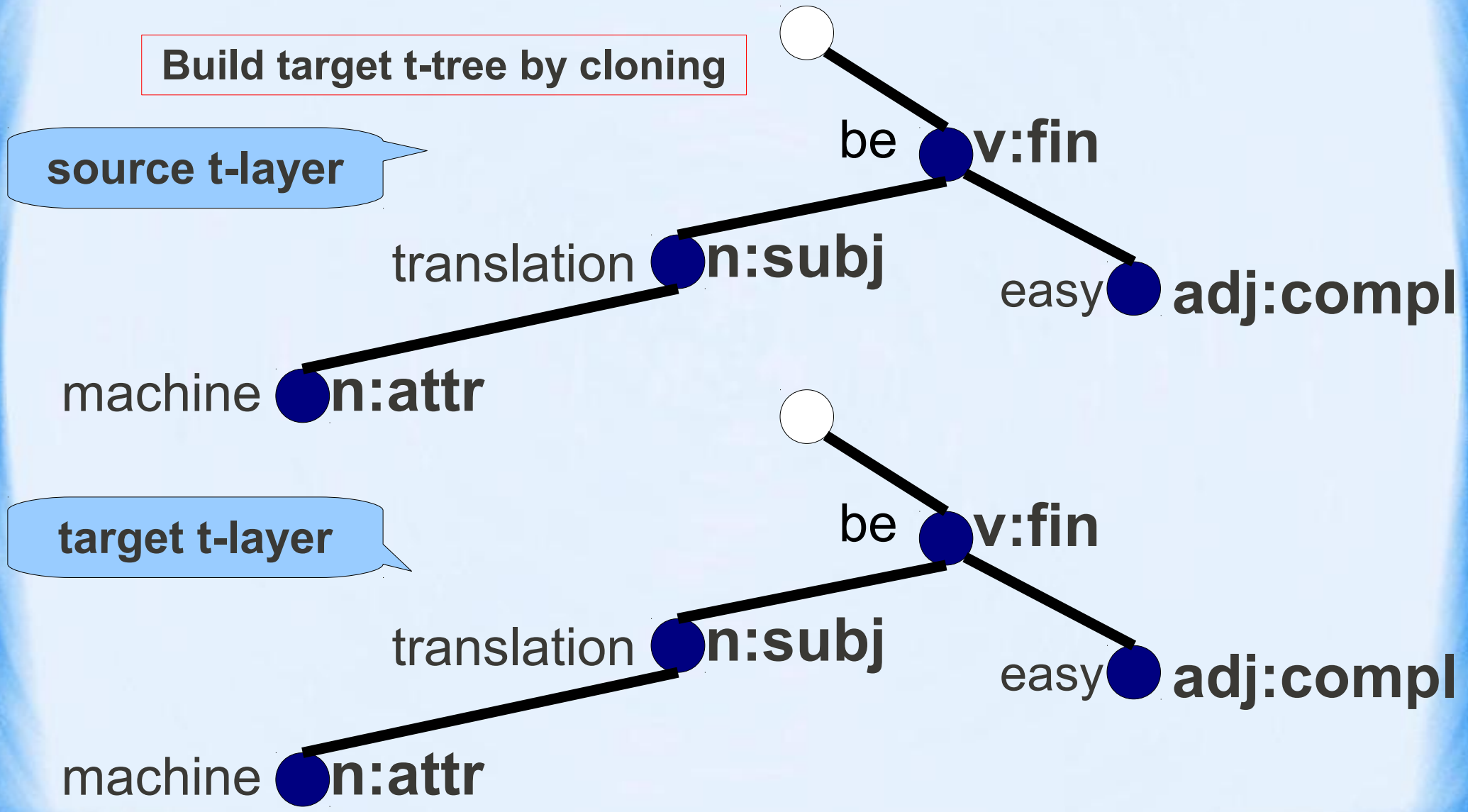
raw text

Machine translation should be easy.



Demo Translation – Transfer

TectoMT

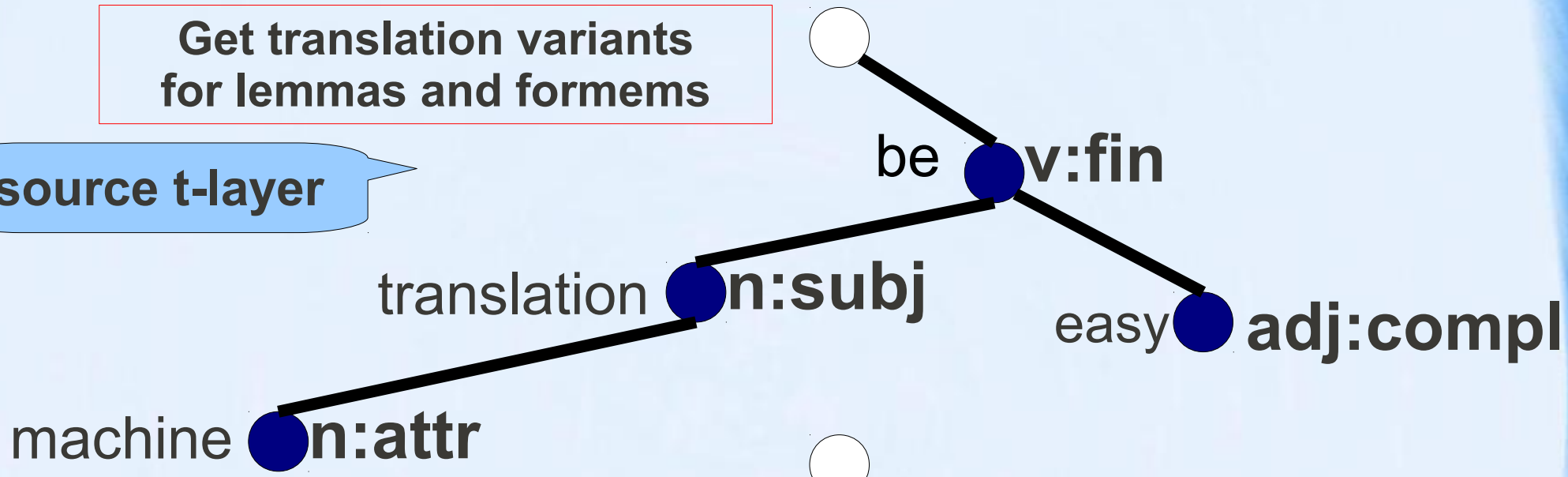


Demo Translation – Transfer

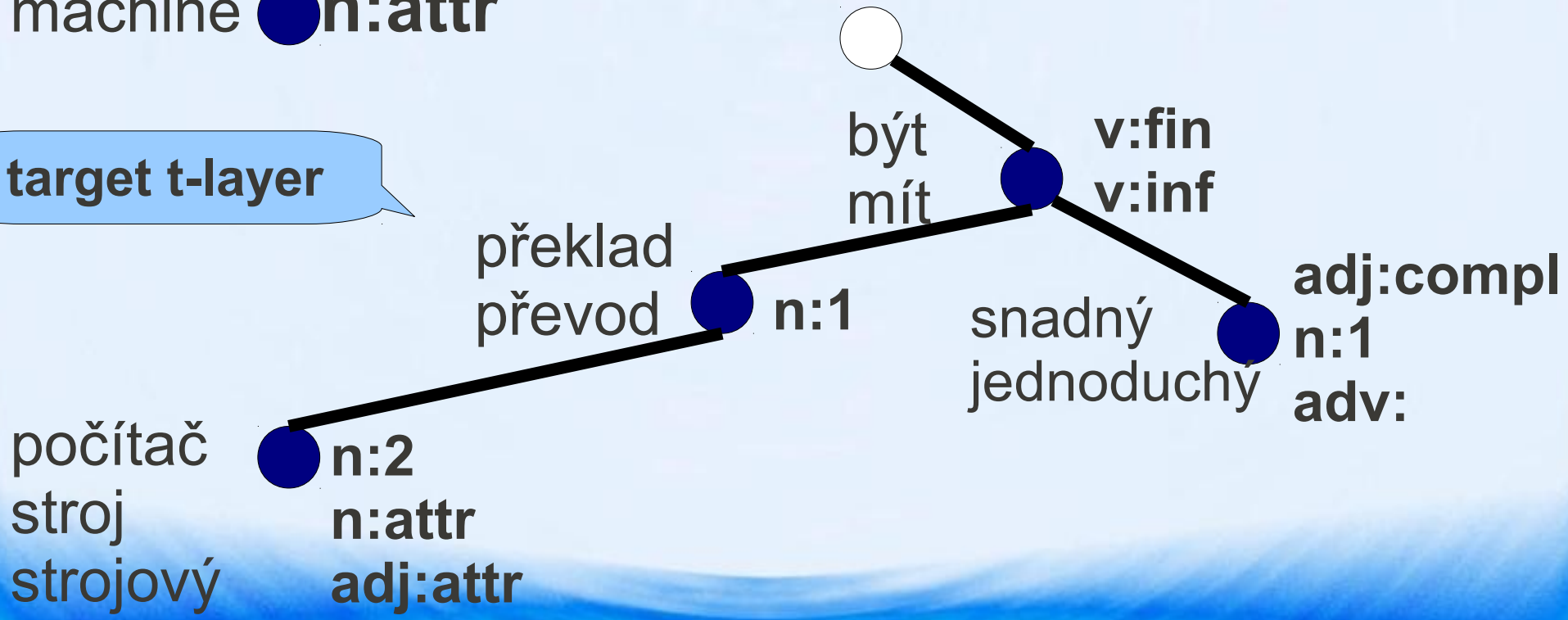
TectoMT

Get translation variants for lemmas and formems

source t-layer



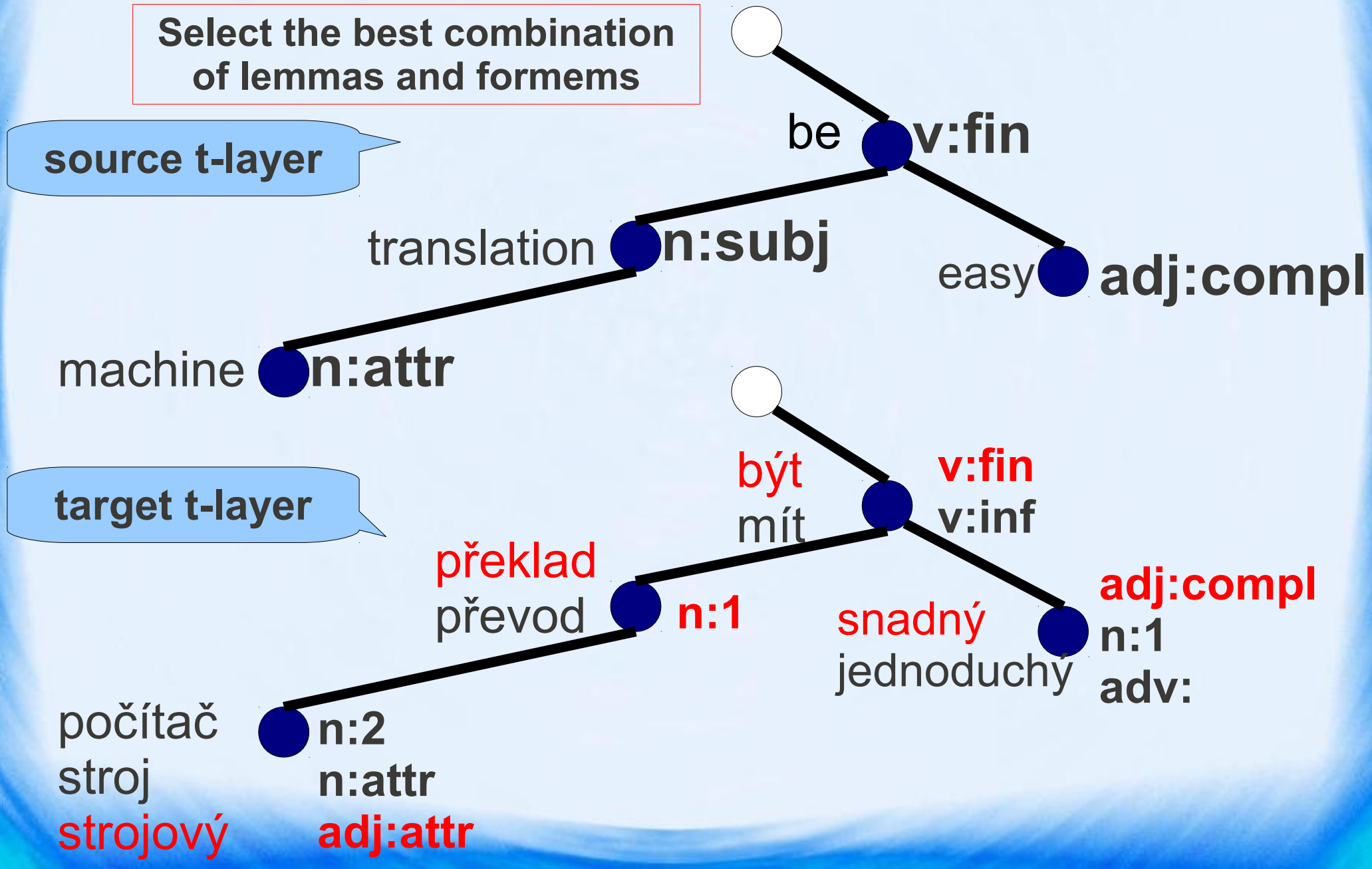
target t-layer



Demo Translation – Transfer

TectoMT

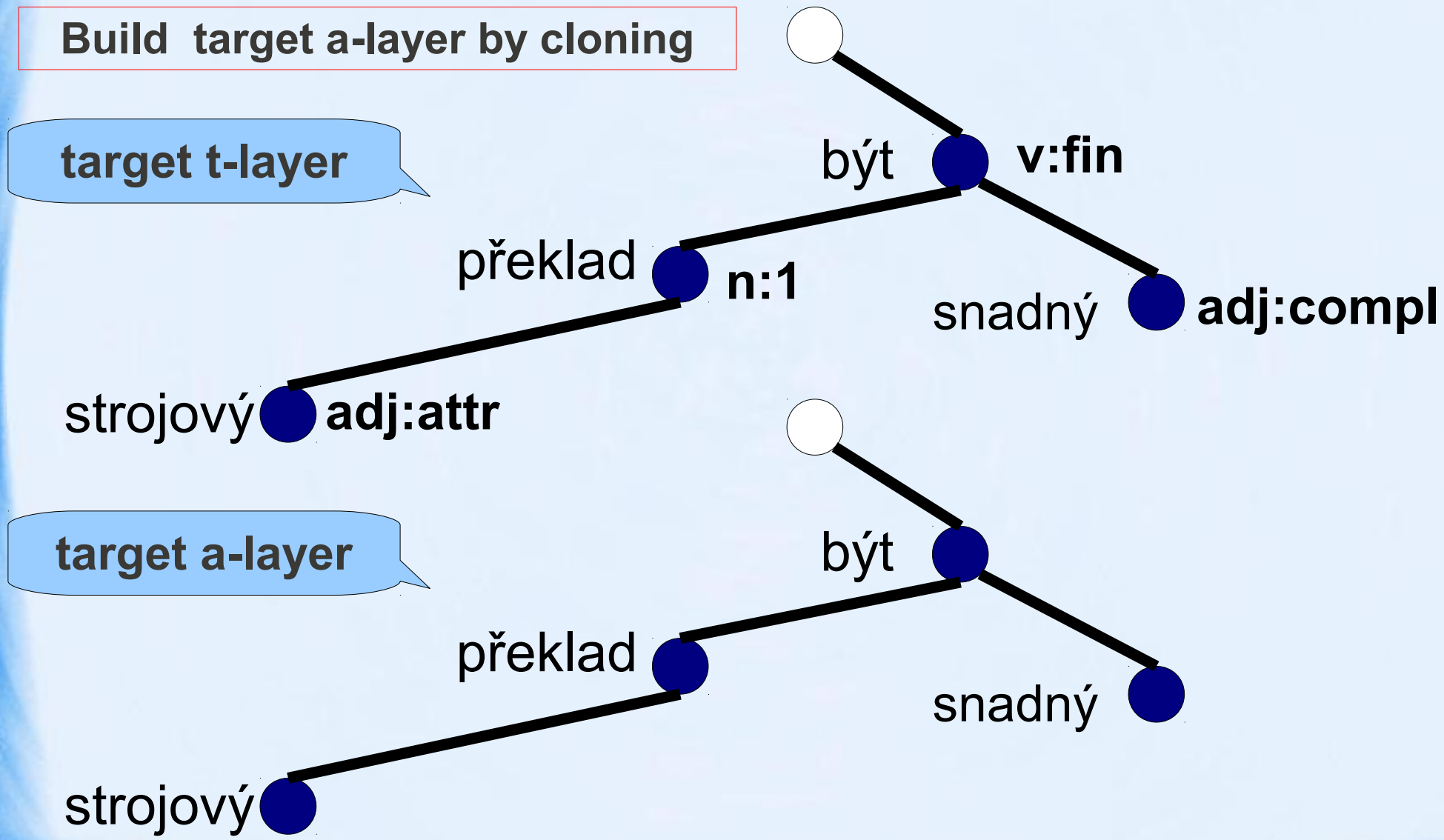
Select the best combination of lemmas and formems



Demo Translation – Synthesis

TectoMT

Build target a-layer by cloning

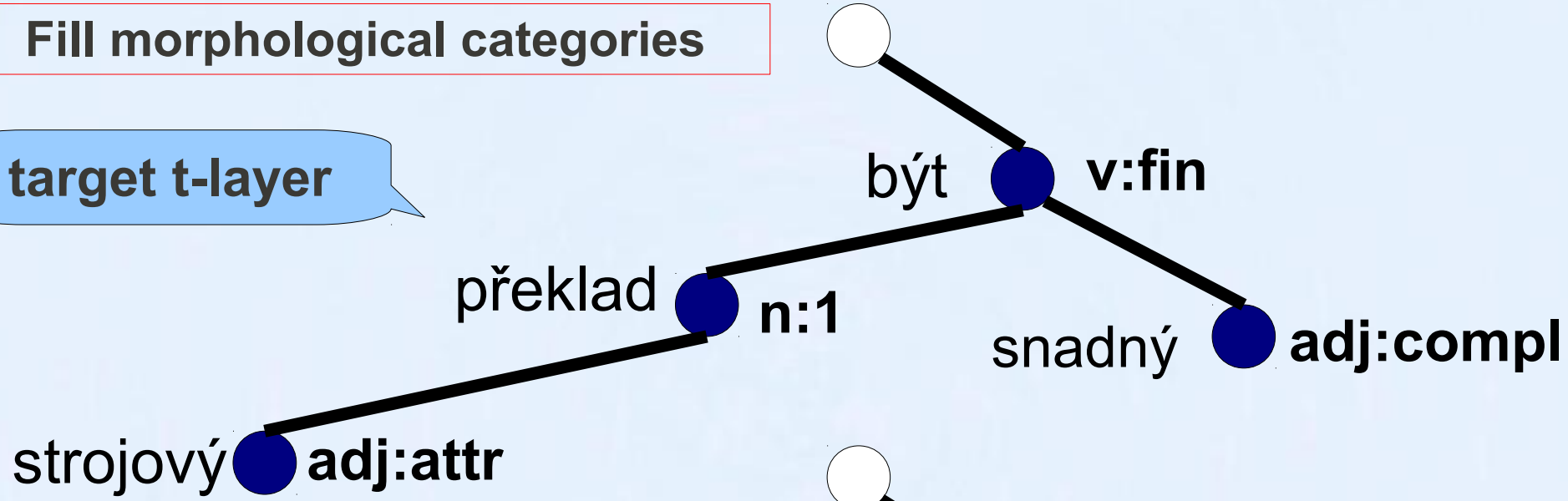


Demo Translation – Synthesis

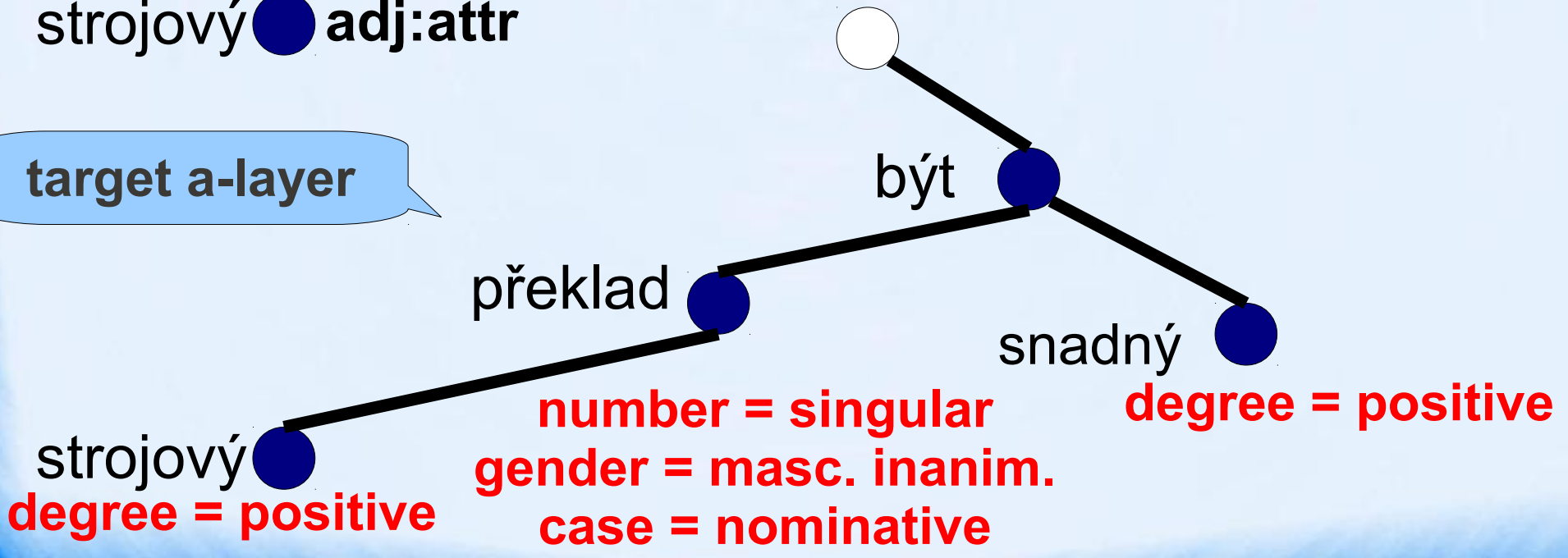
TectoMT

Fill morphological categories

target t-layer



target a-layer

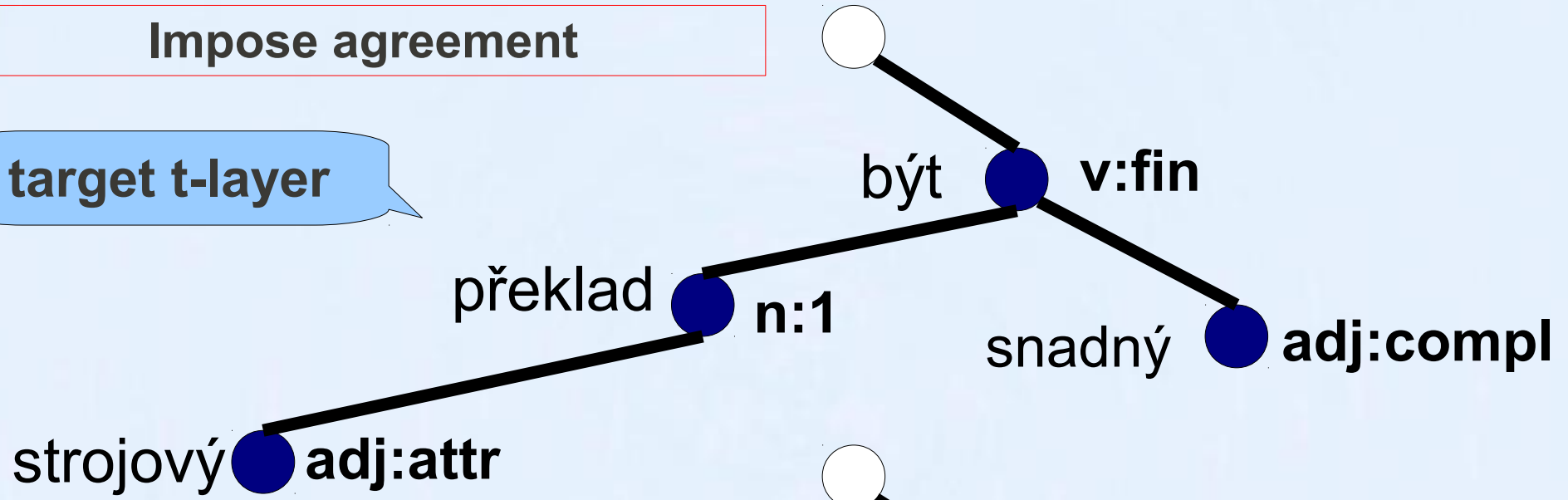


Demo Translation – Synthesis

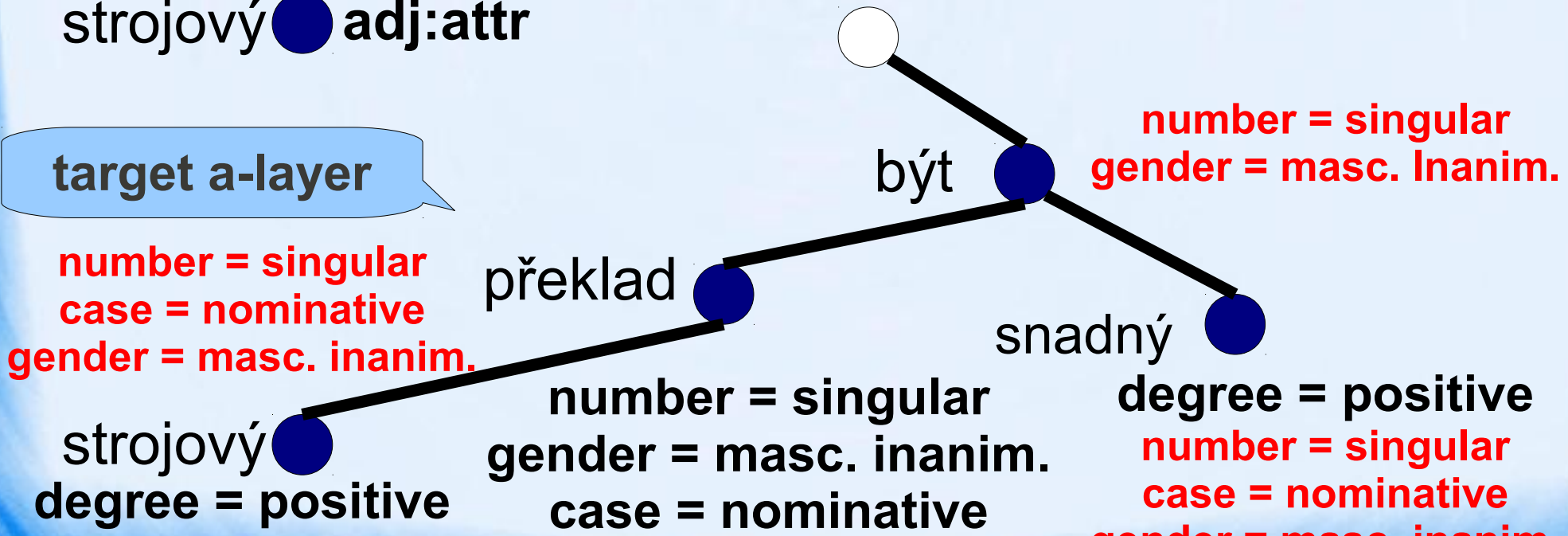
TectoMT

Impose agreement

target t-layer



target a-layer

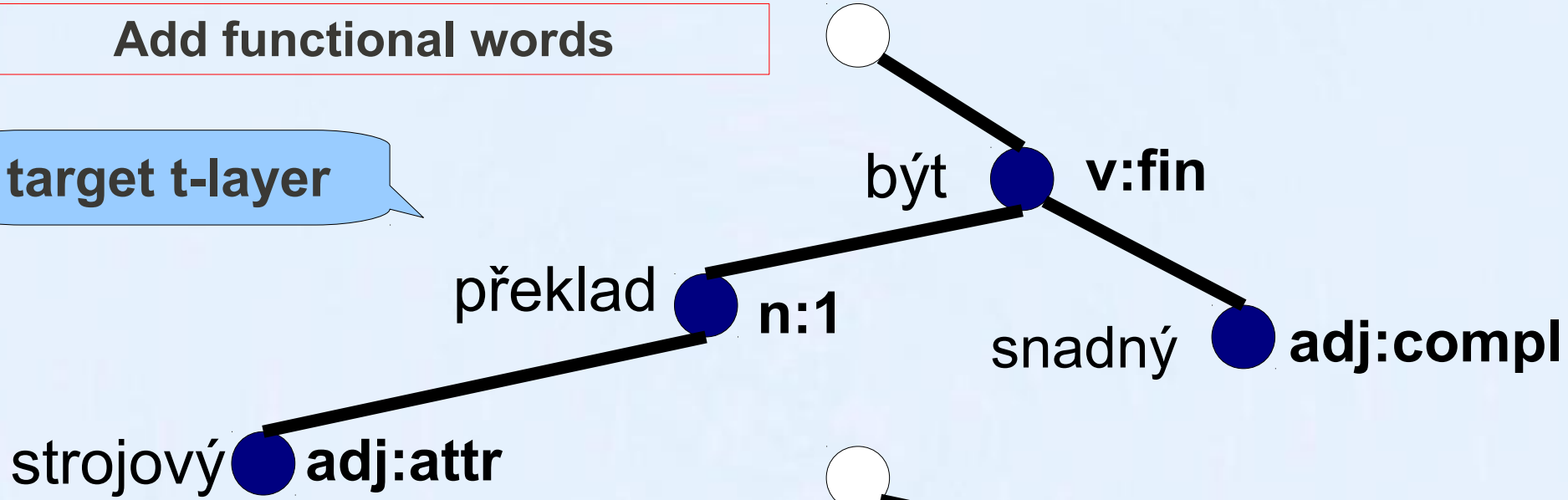


Demo Translation – Synthesis

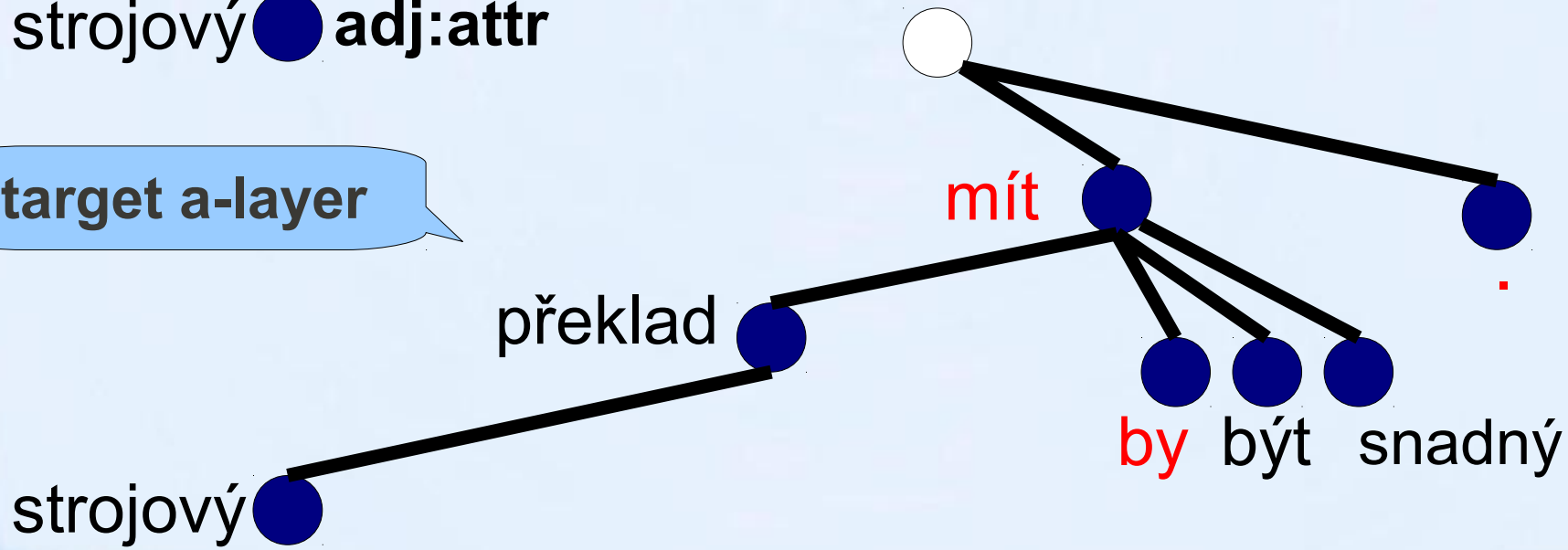
TectoMT

Add functional words

target t-layer



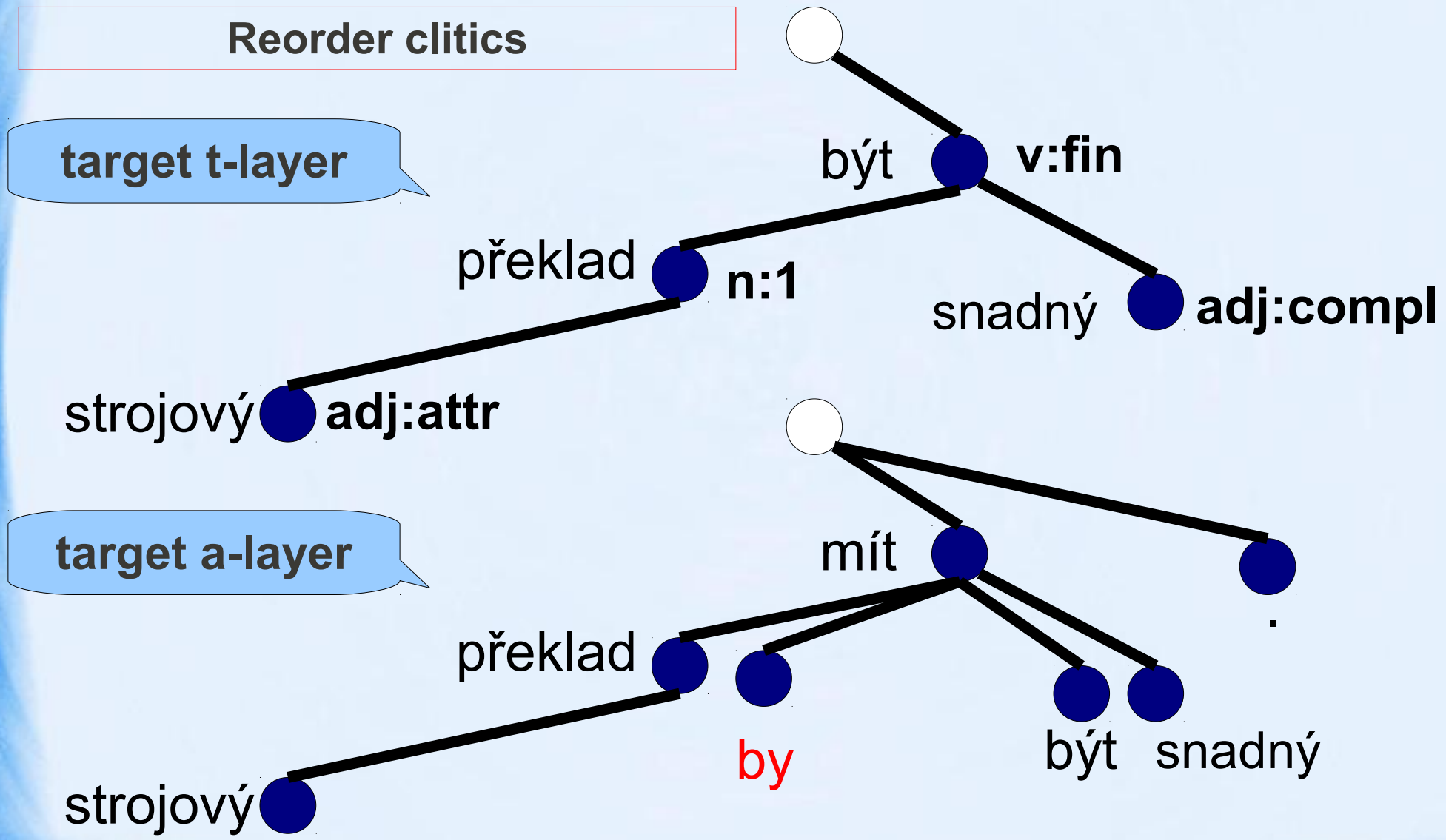
target a-layer



Demo Translation – Synthesis

TectoMT

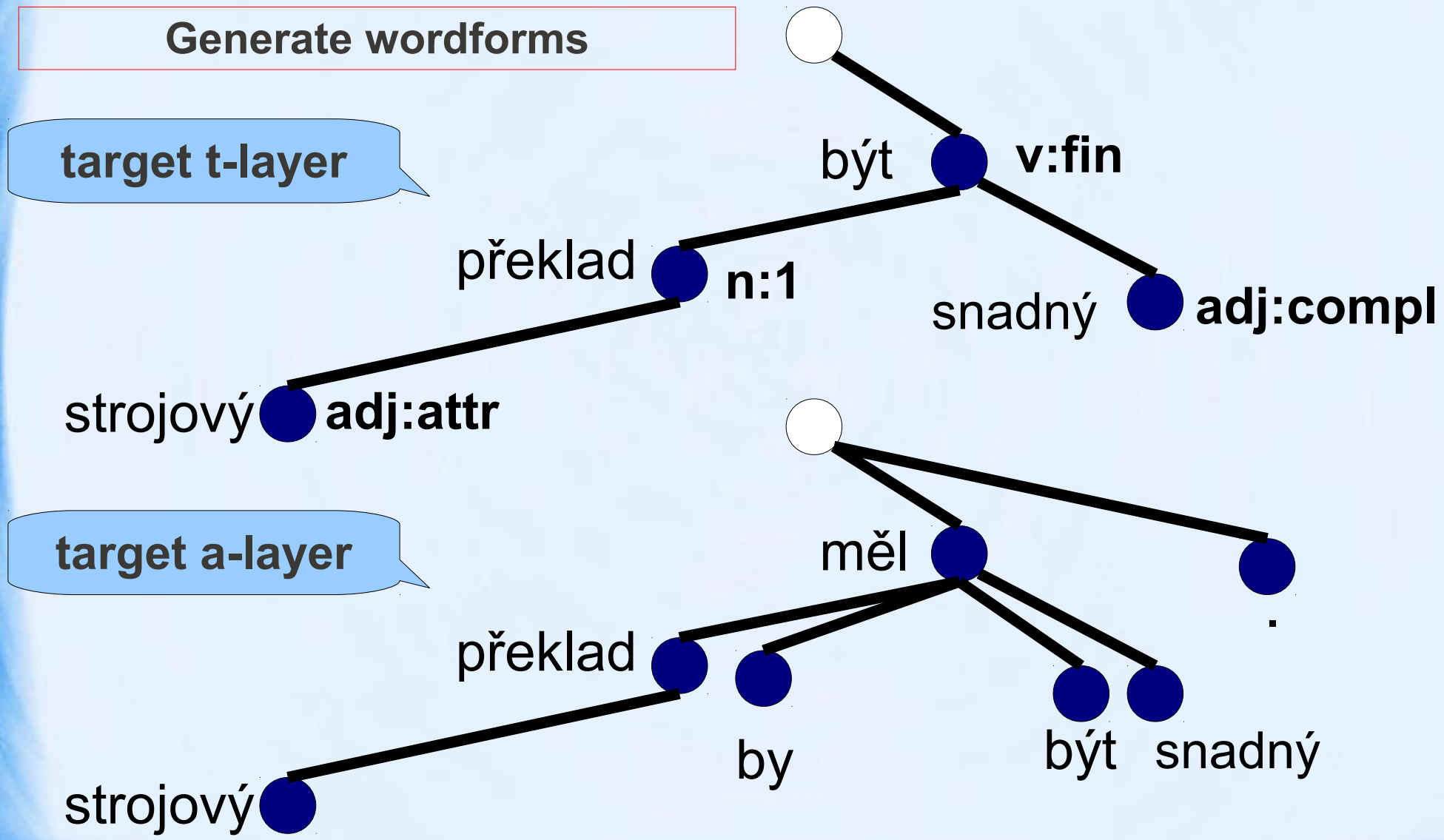
Reorder clitics



Demo Translation – Synthesis

TectoMT

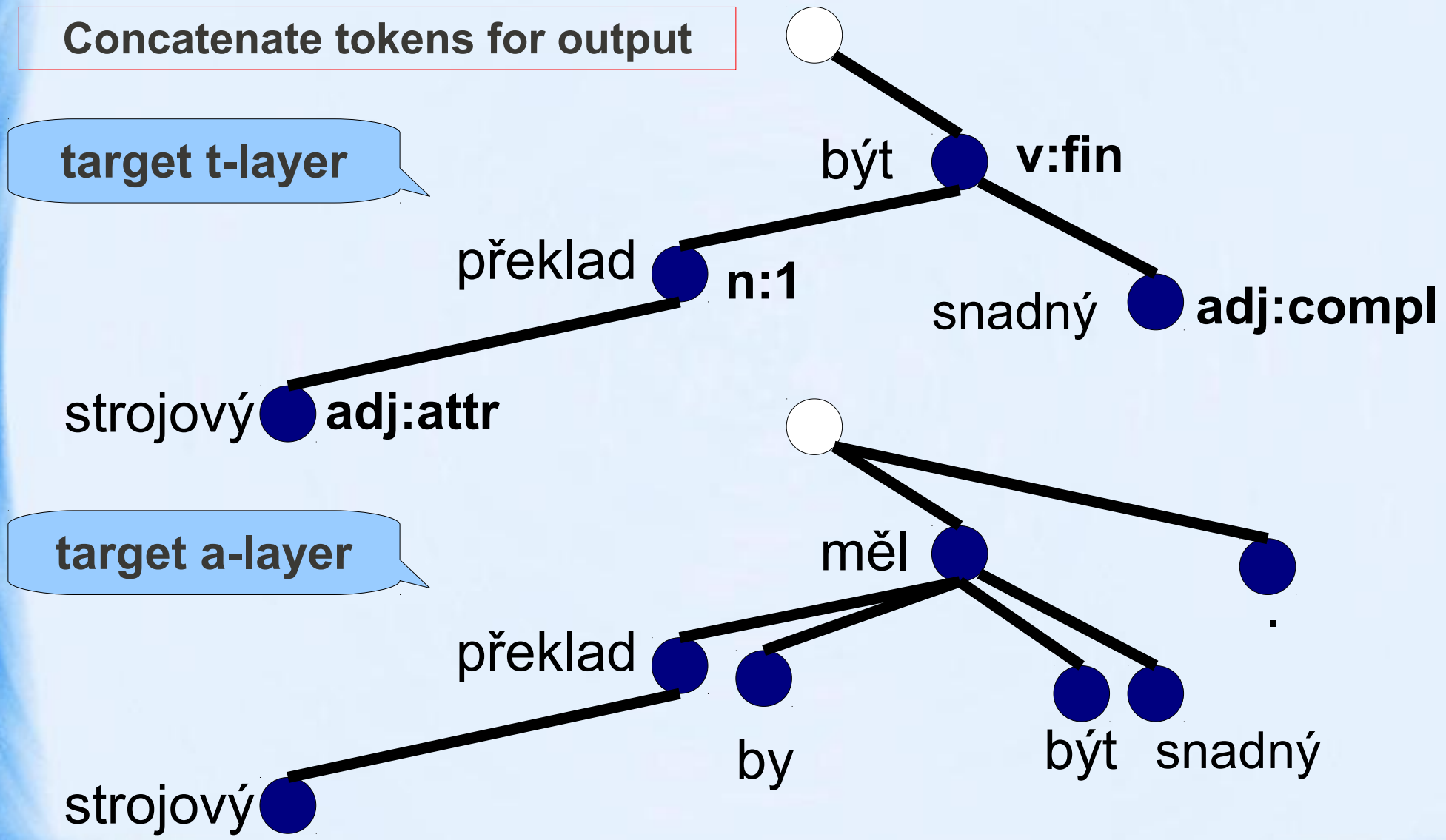
Generate wordforms



Demo Translation – Synthesis

TectoMT

Concatenate tokens for output



Strojový překlad by měl být snadný.

Demo Translation – Real Scenario

TectoMT

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
FindGramCorefForReflPron
NeutPersPronGenderFromAntec
ValencyRelatedRules
SetClauseNumber
TurnTextCorefToGramCoref

SYNTHESIS TO A-LAYER:
CopyTtree
DistinguishHomonymous.
ReverseNumberNounDep.
InitMorphcat
FixPossessiveAdjs
MarkSubject
ImposePronZAgr
ImposeRelPronAgr
ImposeSubjpredAgr
ImposeAttrAgr
ImposeComplAgr
DropSubjPersProns
AddPrepos
AddSubconjs
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

Combining Dictionaries

TectoMT

- new general interface (for lemmas and formems)

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

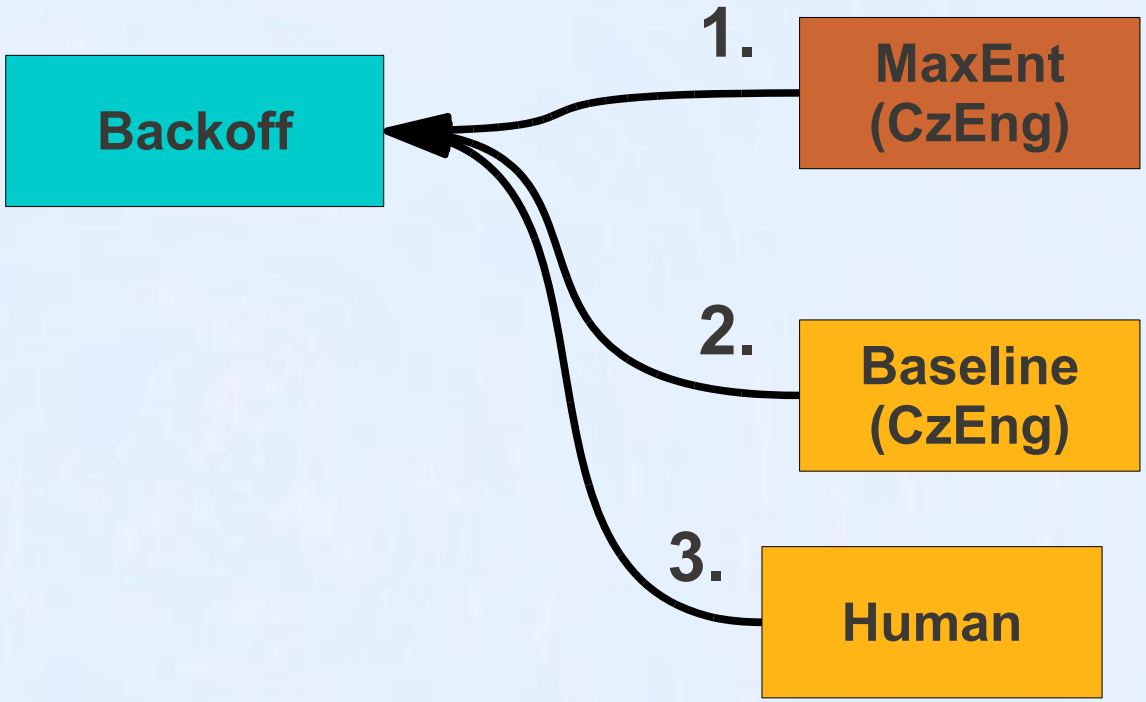
returns a list of translation variants including probabilities

- OOP style, dictionary constructor can take another dictionary (or more) as a parameter → hierachy
- Four basic types of dictionaries:

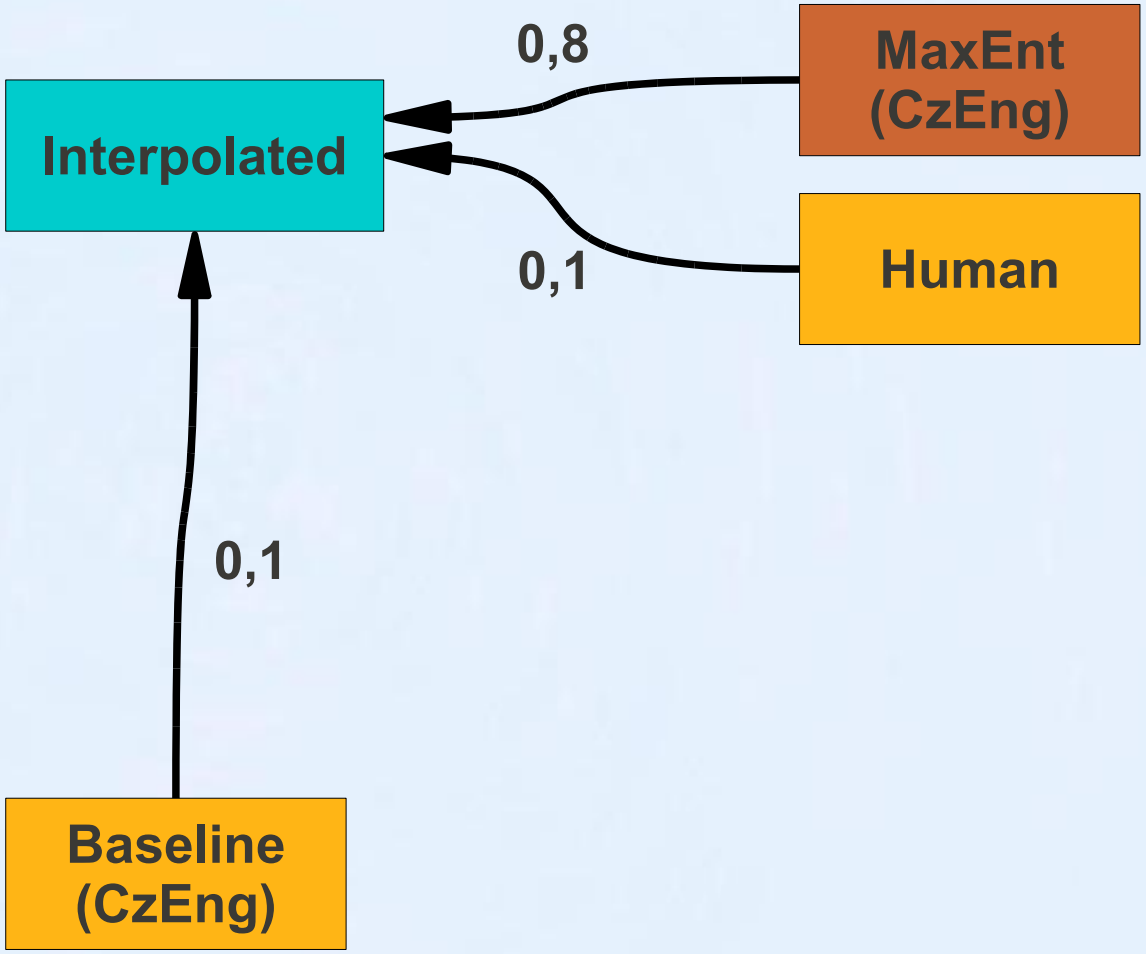
Static plain	loaded from a file „lemma → lemma“
Context	loaded from a file „lemma,features → lemma“
Derivational	translations derived dynamicaly, input dictionary
Combinaional	combination of more input dictionaries

Hierarchy of lemma dictionaries

TectoMT

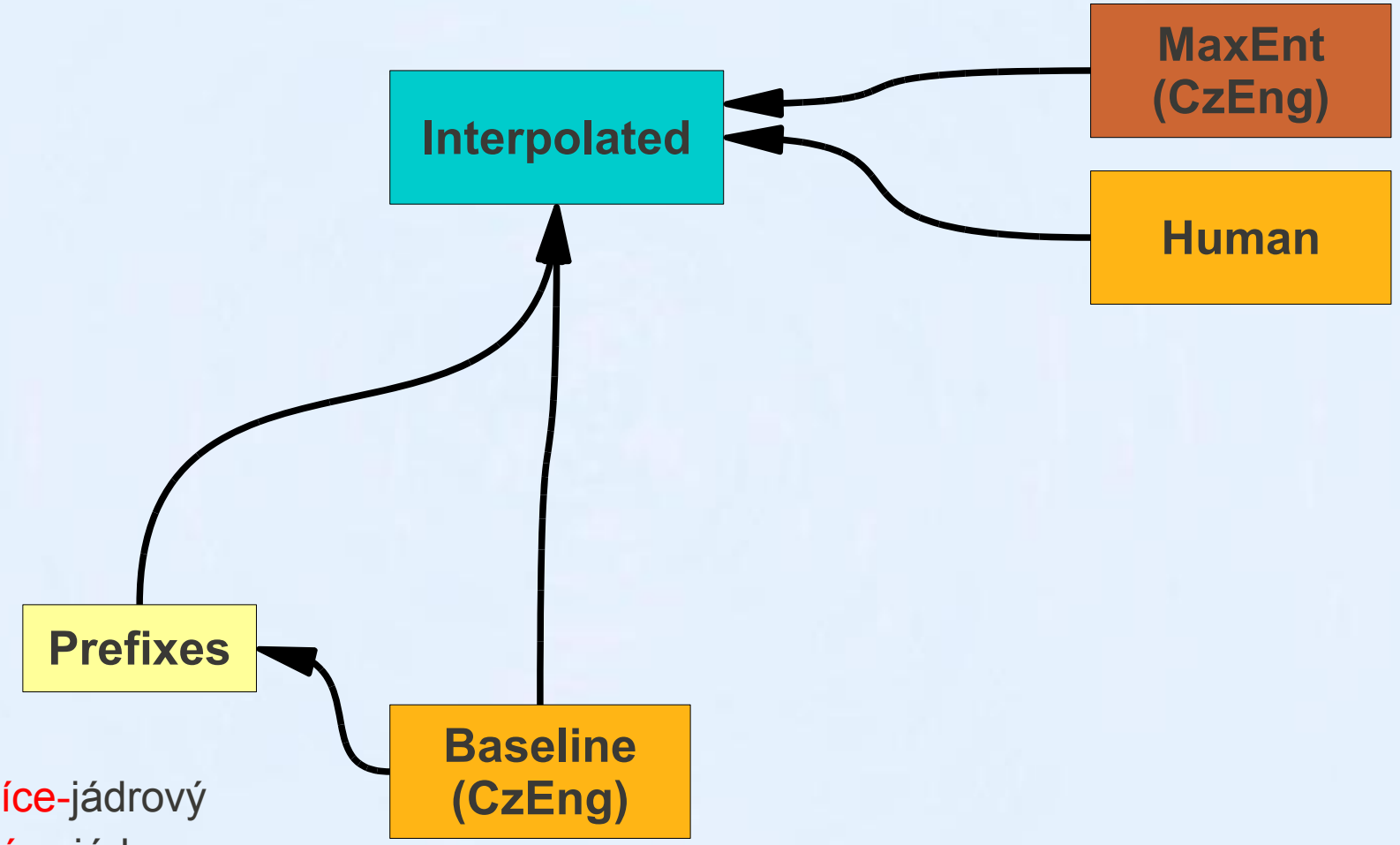


Hierarchy of lemma dictionaries



Hierarchy of lemma dictionaries

TectoMT



multi-core → více-jádrový

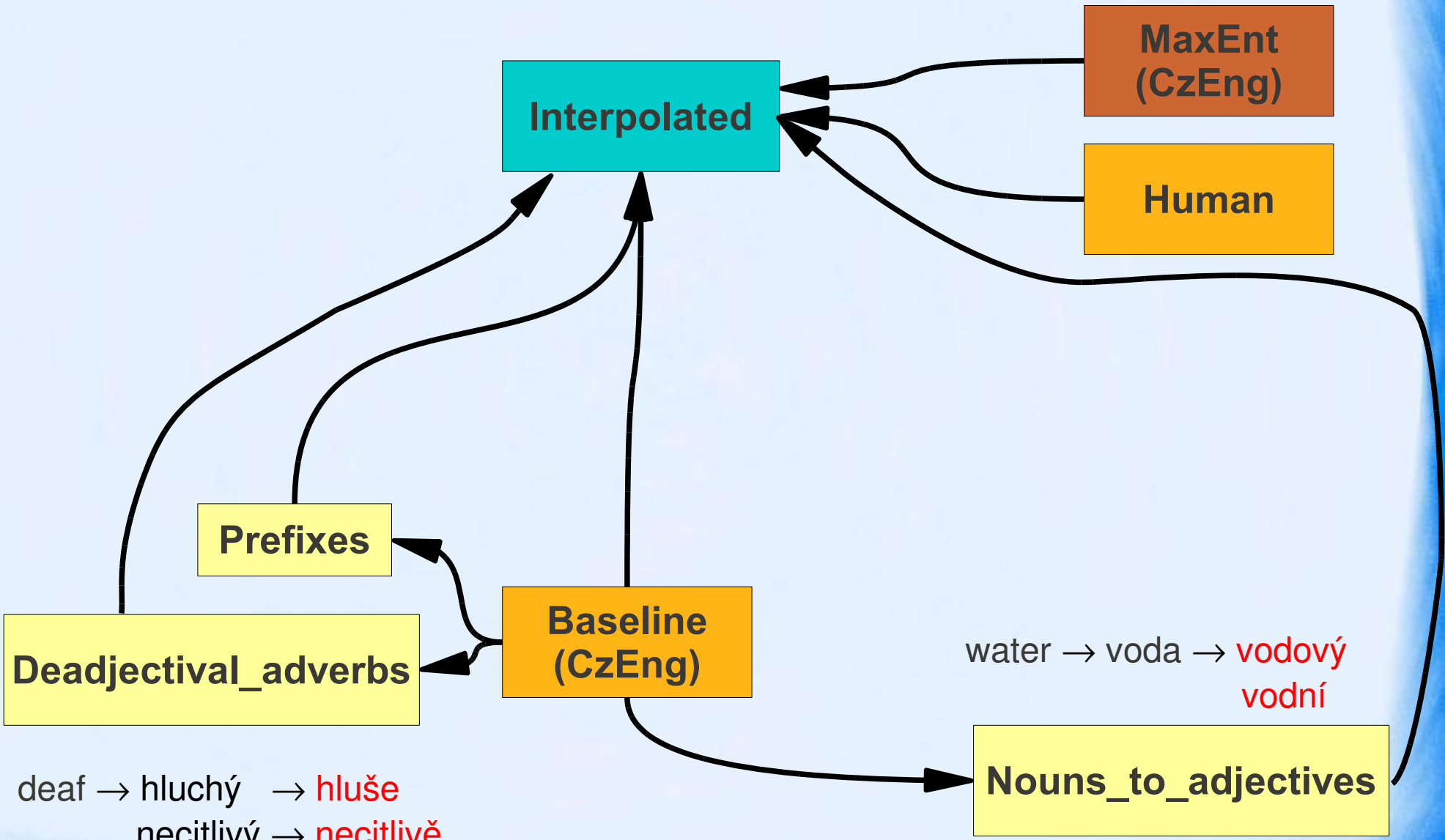
více-jádro

multi-jádrový

multi-jádro

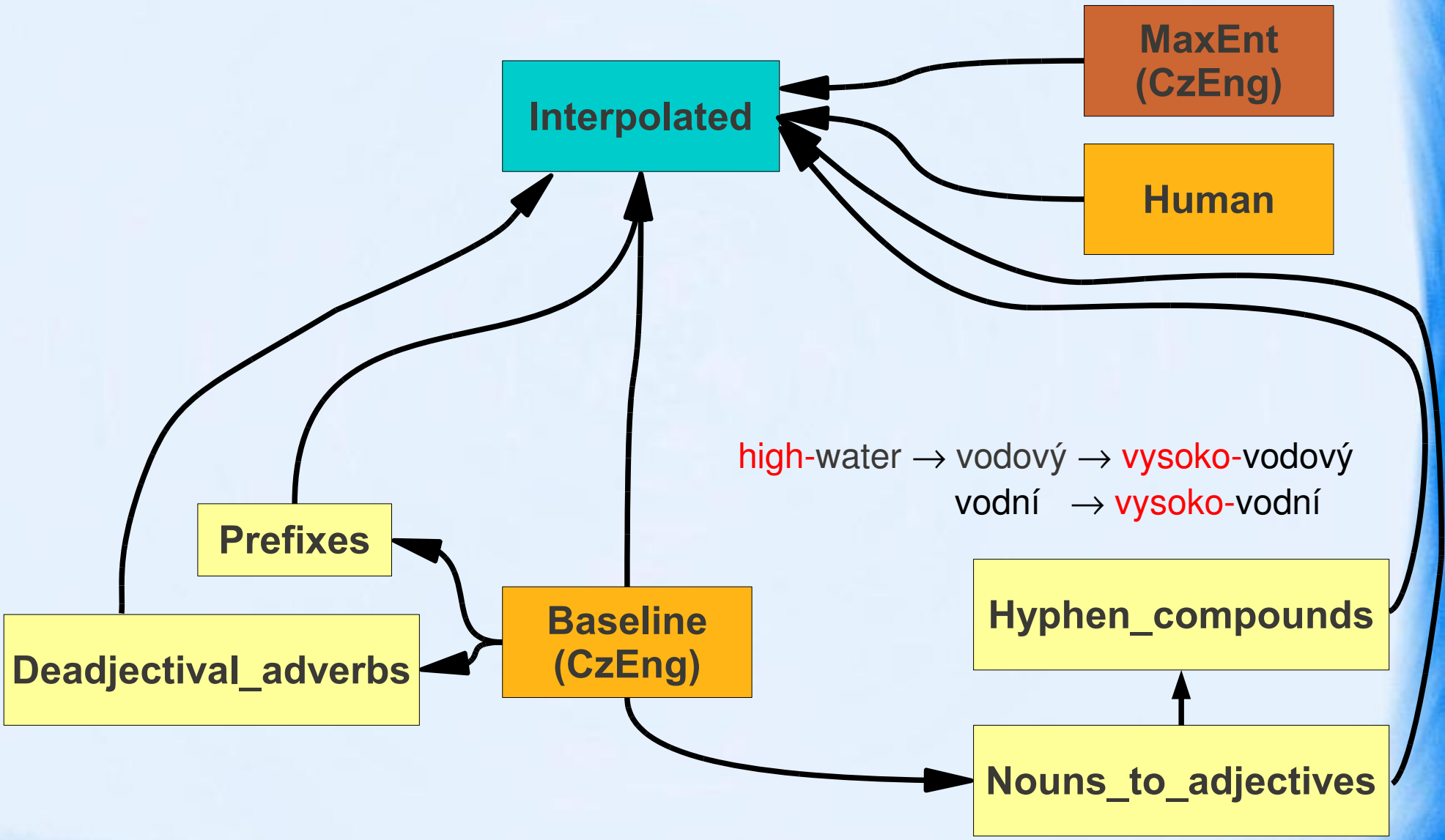
Hierarchy of lemma dictionaries

TectoMT



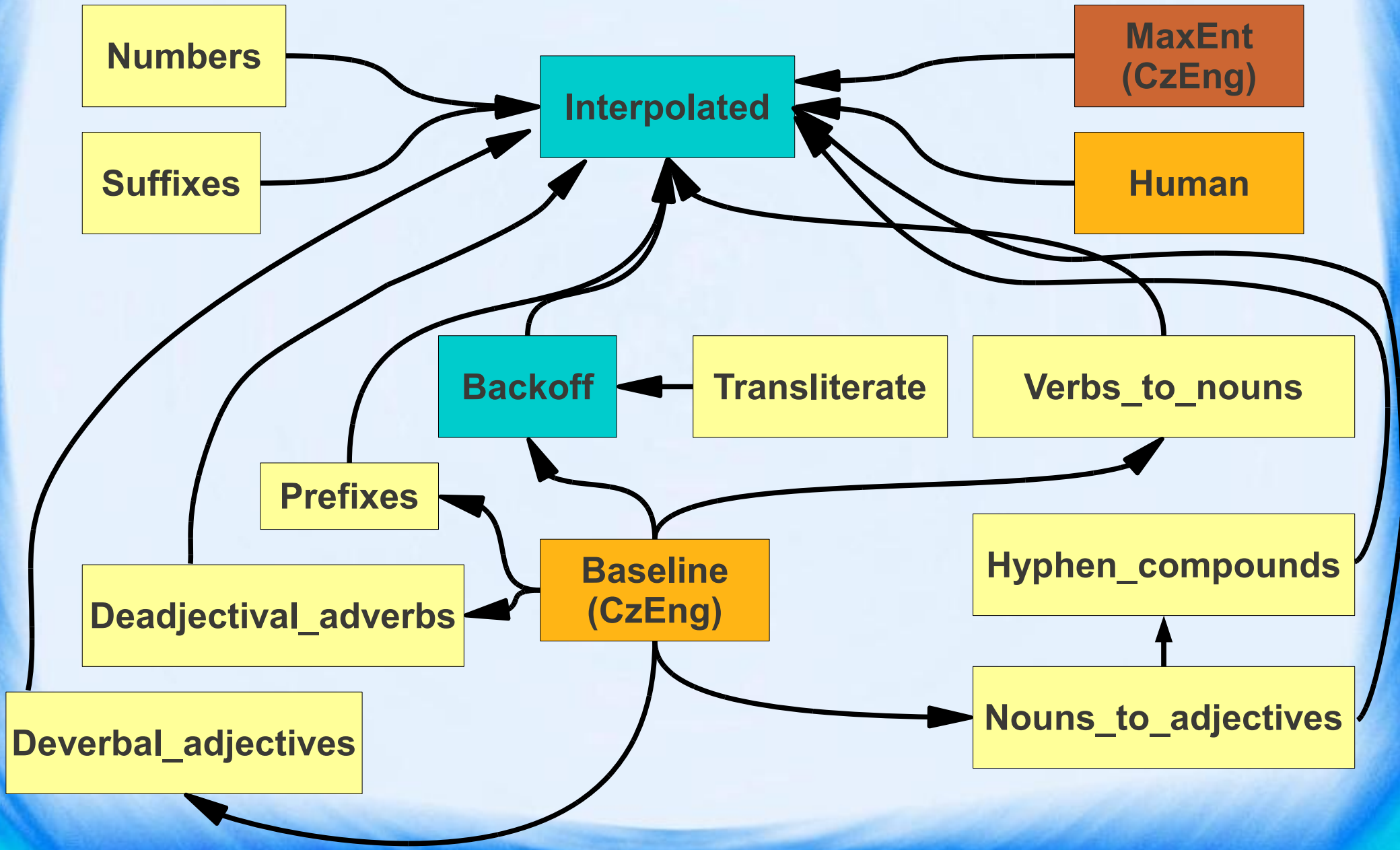
Hierarchy of lemma dictionaries

TectoMT



Hierarchy of lemma dictionaries

TectoMT

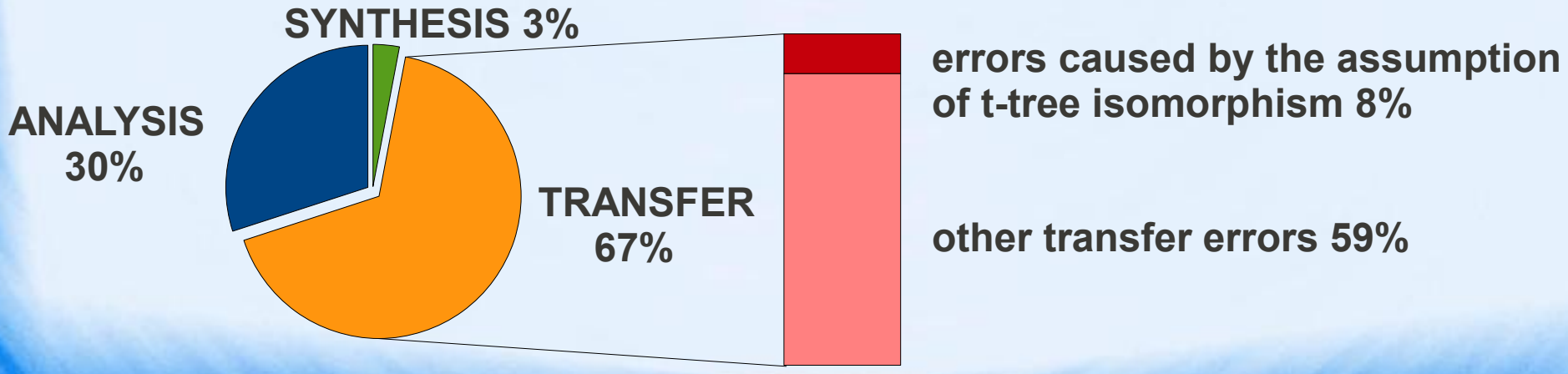


Annotation of Translation Errors

TectoMT

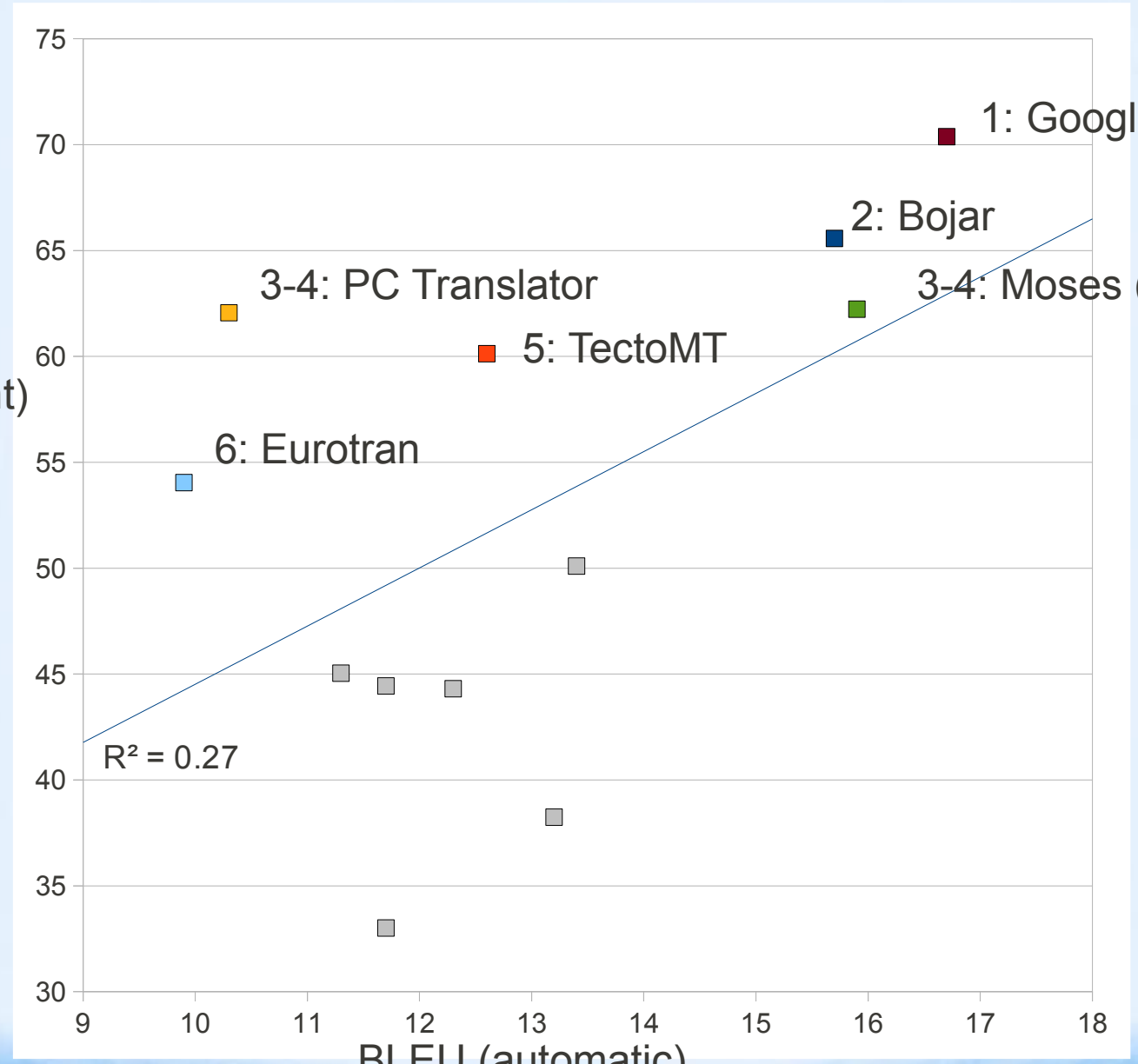
sample of 250 sentences, 1463 errors in total

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, ?



Results – BLEU vs. Ranks (WMT 2010)

TectoMT

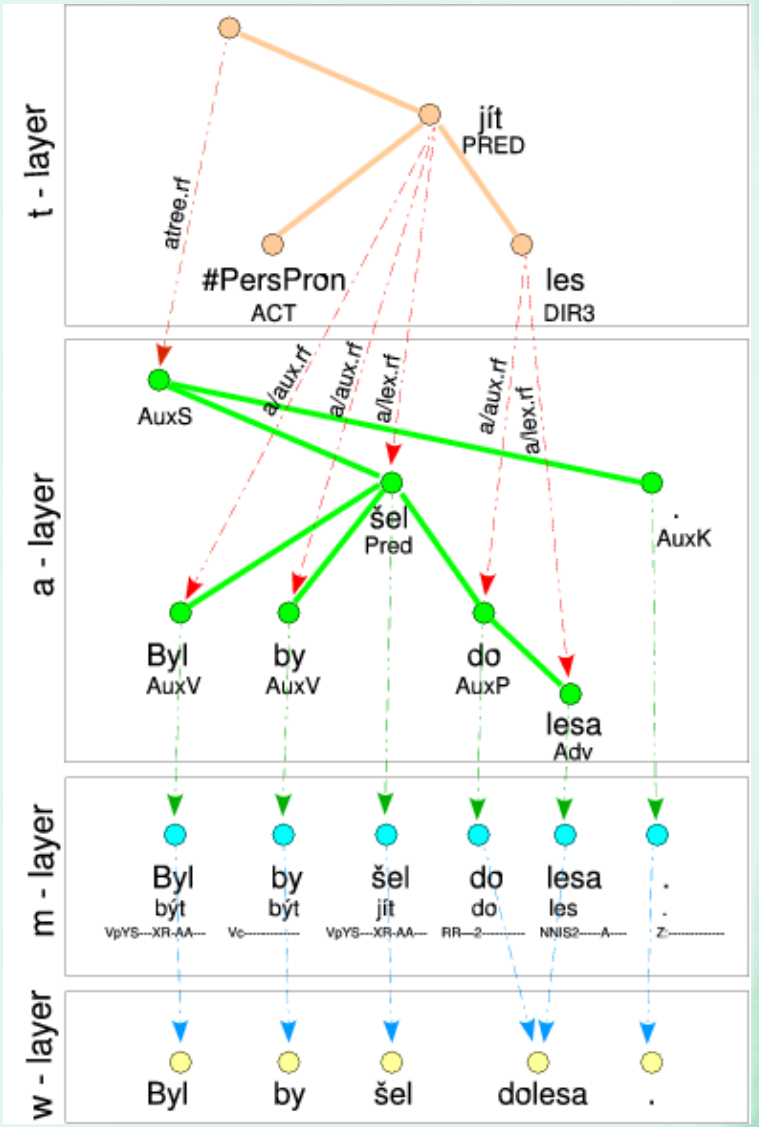


4 layers of language description

implemented in Prague Dependency Treebank (PDT)

Treex

- **tectogrammatical layer**
  deep-syntactic dependency trees
- **analytical layer**
  surface-syntactic dependency trees, labeled edges
- **morphological layer**
  lemma & POS tag for each word
- word layer
  raw (tokenized) text



4 layers of language description
implemented in Prague Dependency Treebank (PDT)

Treex

- **tectogrammatical layer**
  deep-syntactic dependency trees

- abstraction from many language-specific phenomena

- autosemantic (meaningful) words
  ~ nodes

- functional words (prepositions, auxiliaries)
  ~ attributes

- syntactic-semantic relations (dependecies)
  ~ edges

- added nodes (e.g. becasue of pro-drop)

- ...