

Depfix

a Tool for Automatic Rule-based Post-editing of SMT

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



Input

source English sentence

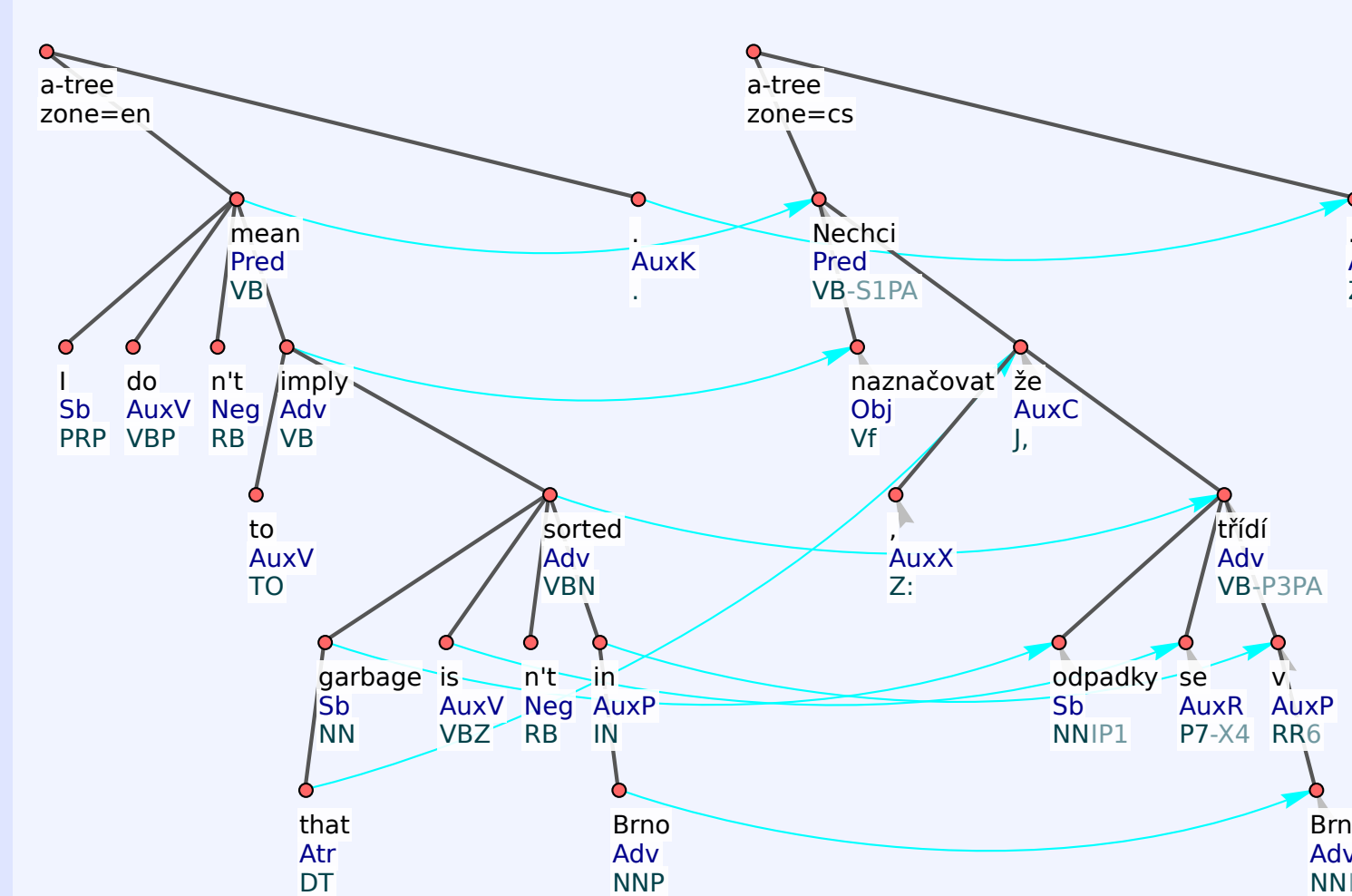
English → Czech
Machine Translation
Moses/Google Translate/...

Czech sentence from SMT

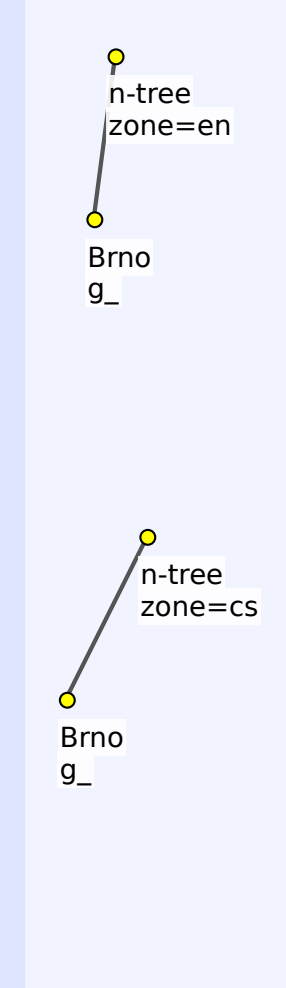
Treex Analysis

tokenizer (Treex)
lemmatizing tagger (MorphoDiTa)
word aligner (GIZA++)
dependency parser (MST)
dependency relations labeller (Treex)
named entity recognizer (Stanford)
conversion to tectogrammar (Treex)

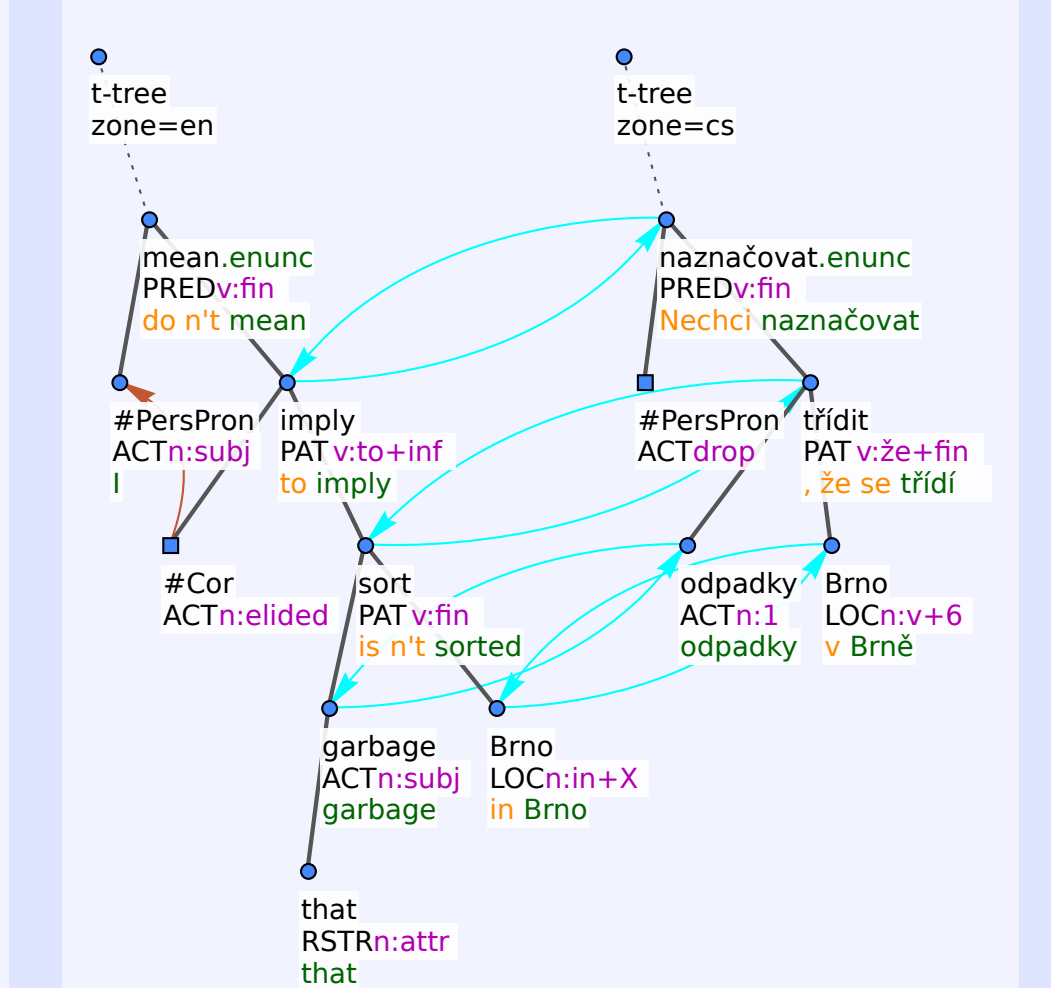
Analytical (surface syntax)
dependency trees



Named
entities



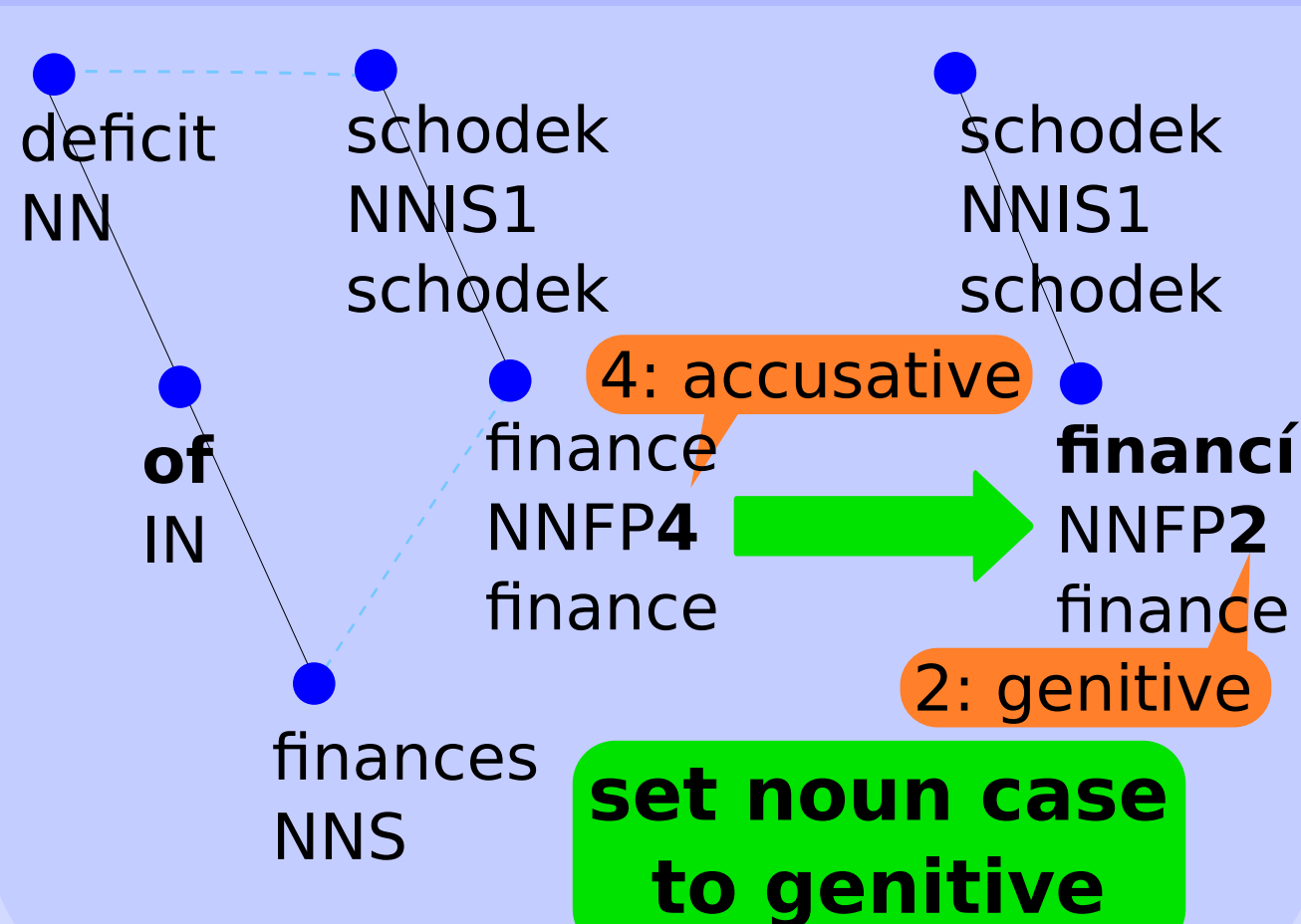
Tectogrammatical
(deep syntax) trees



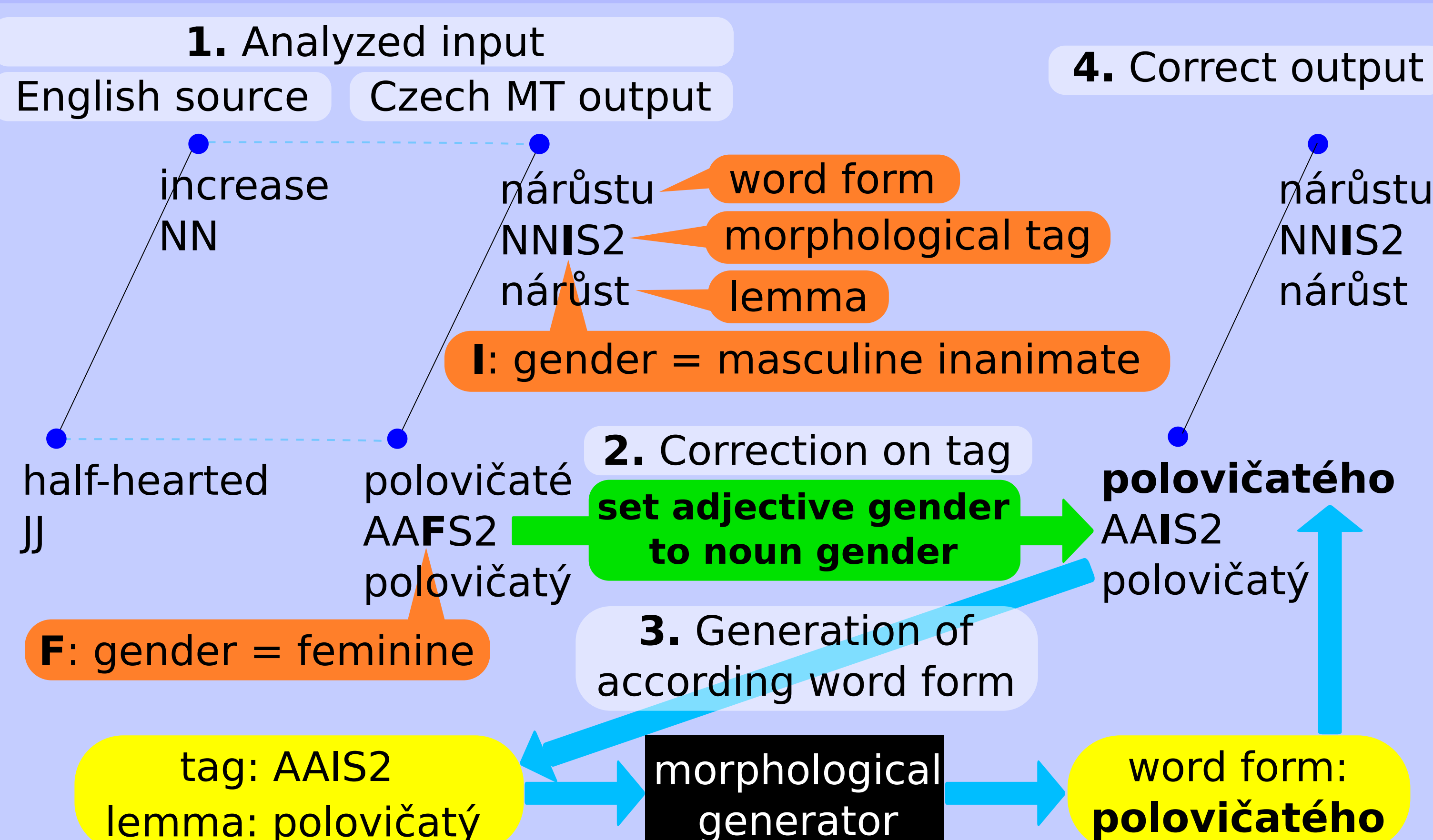
Depfix Correction Rules

current version: 28 rules

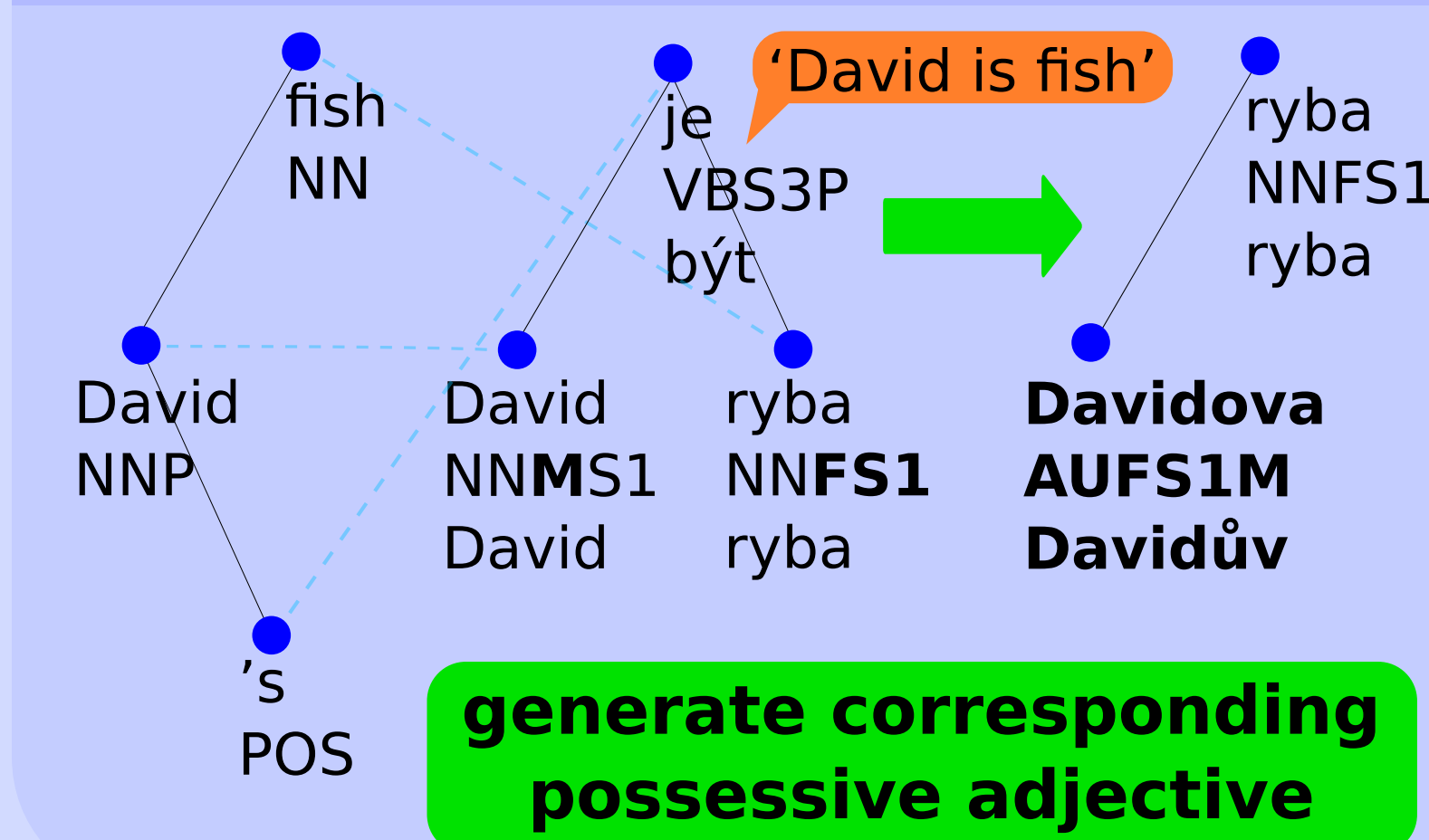
Translation of 'of'



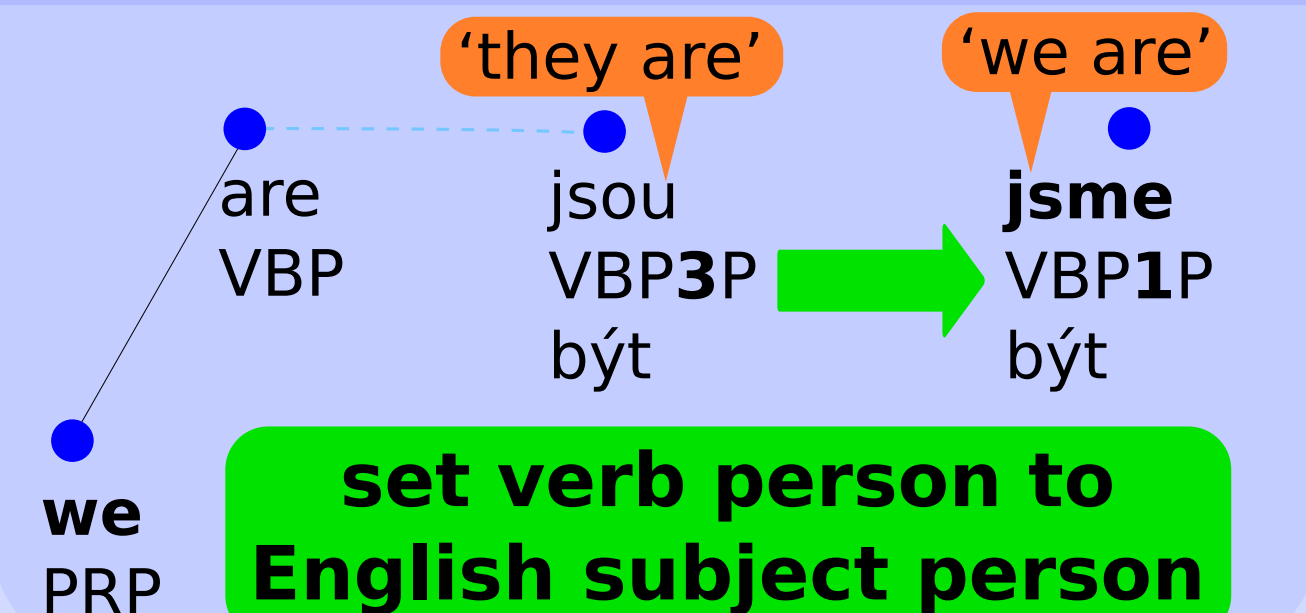
Noun - Adjective Agreement



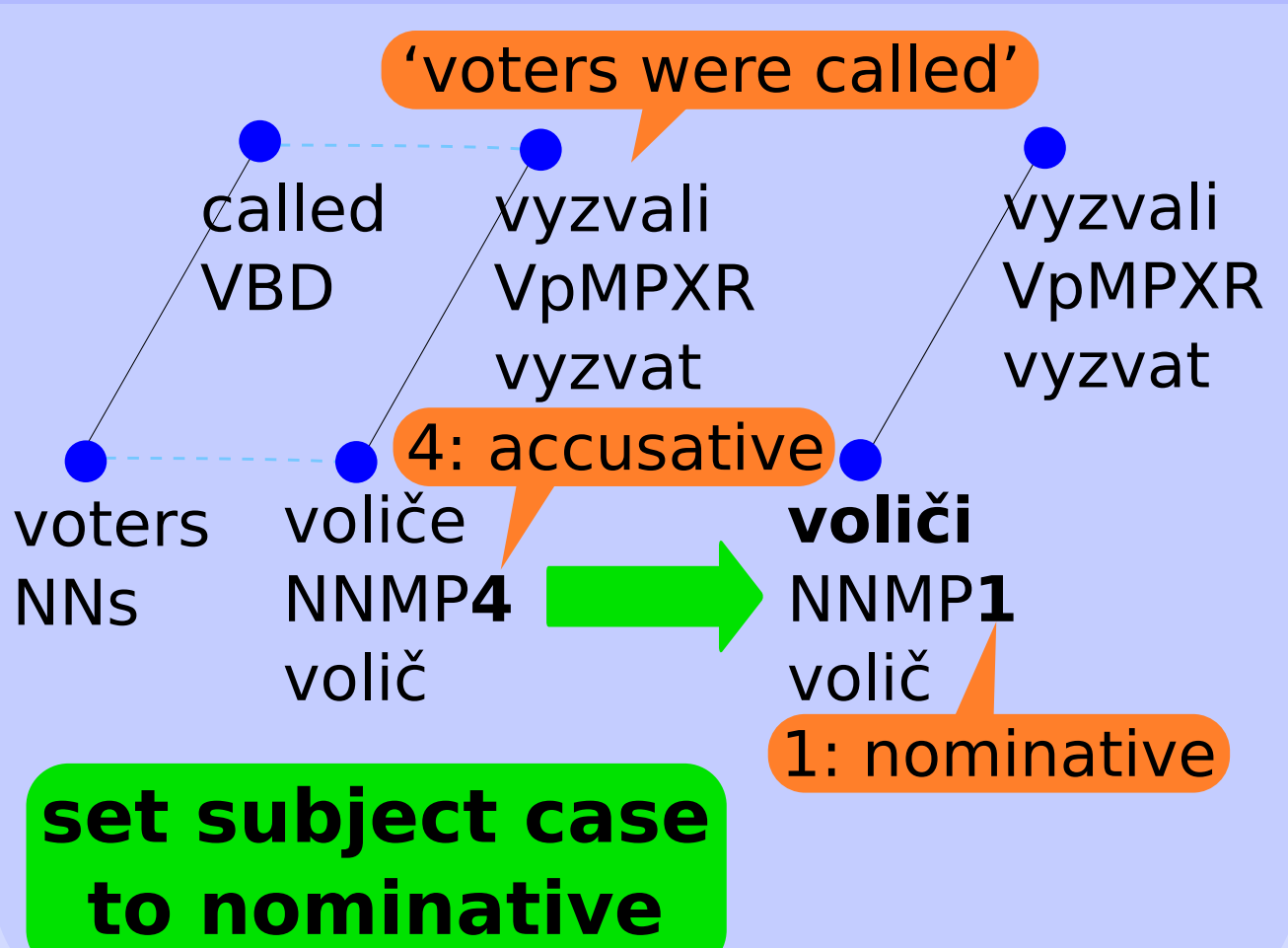
Translation of Possessive Nouns



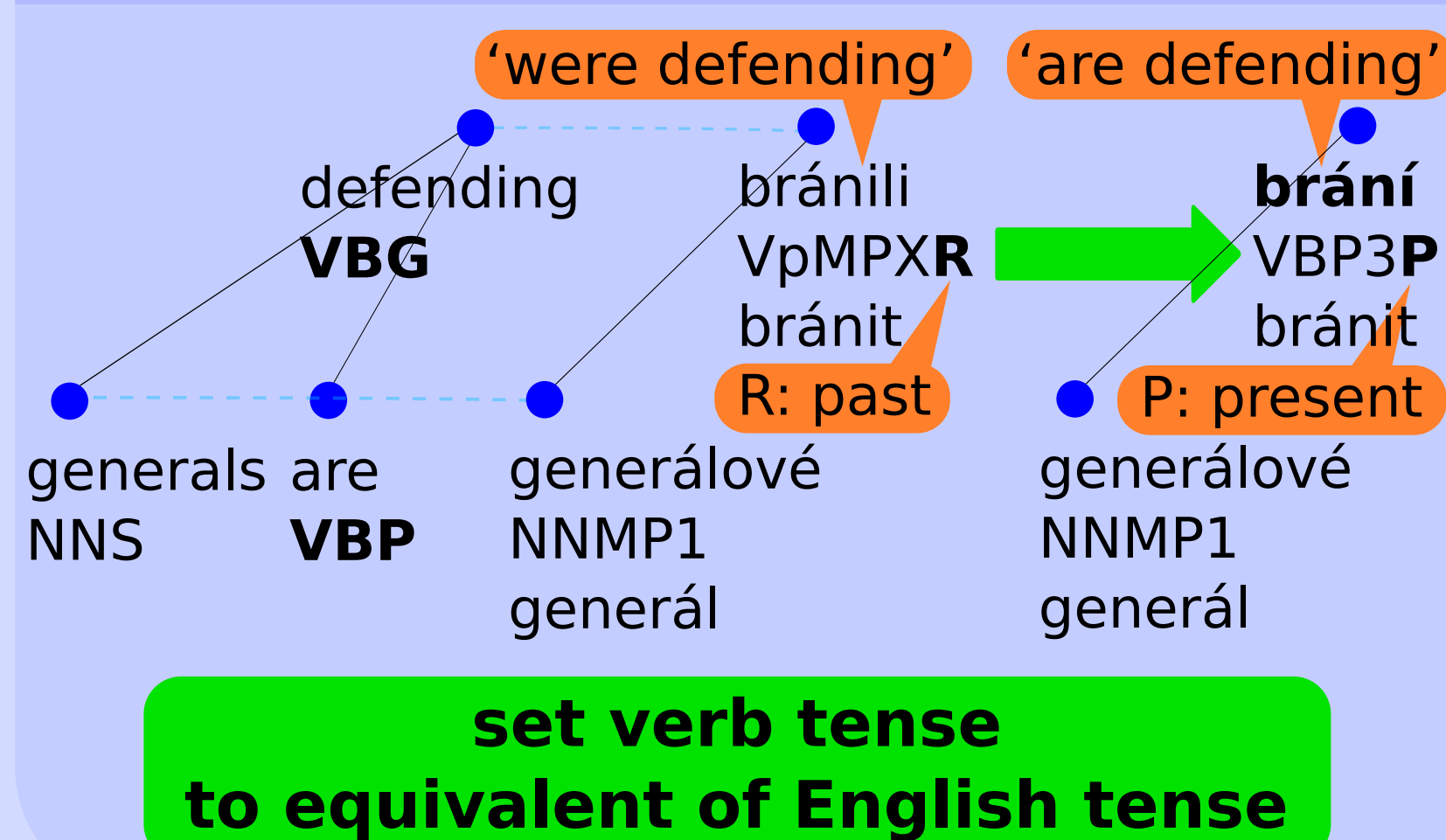
Subject Person Projection



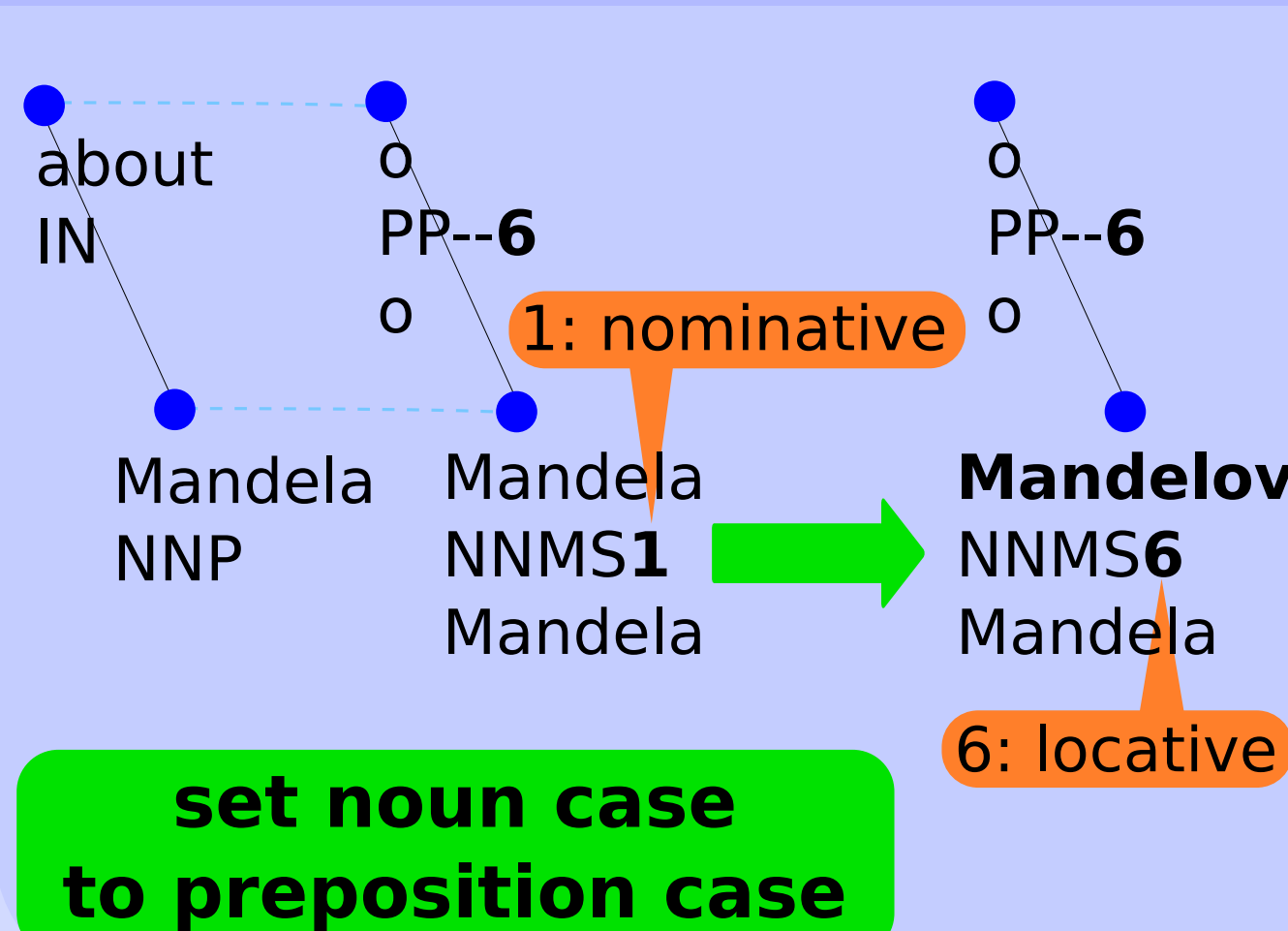
Subject Morphological Case



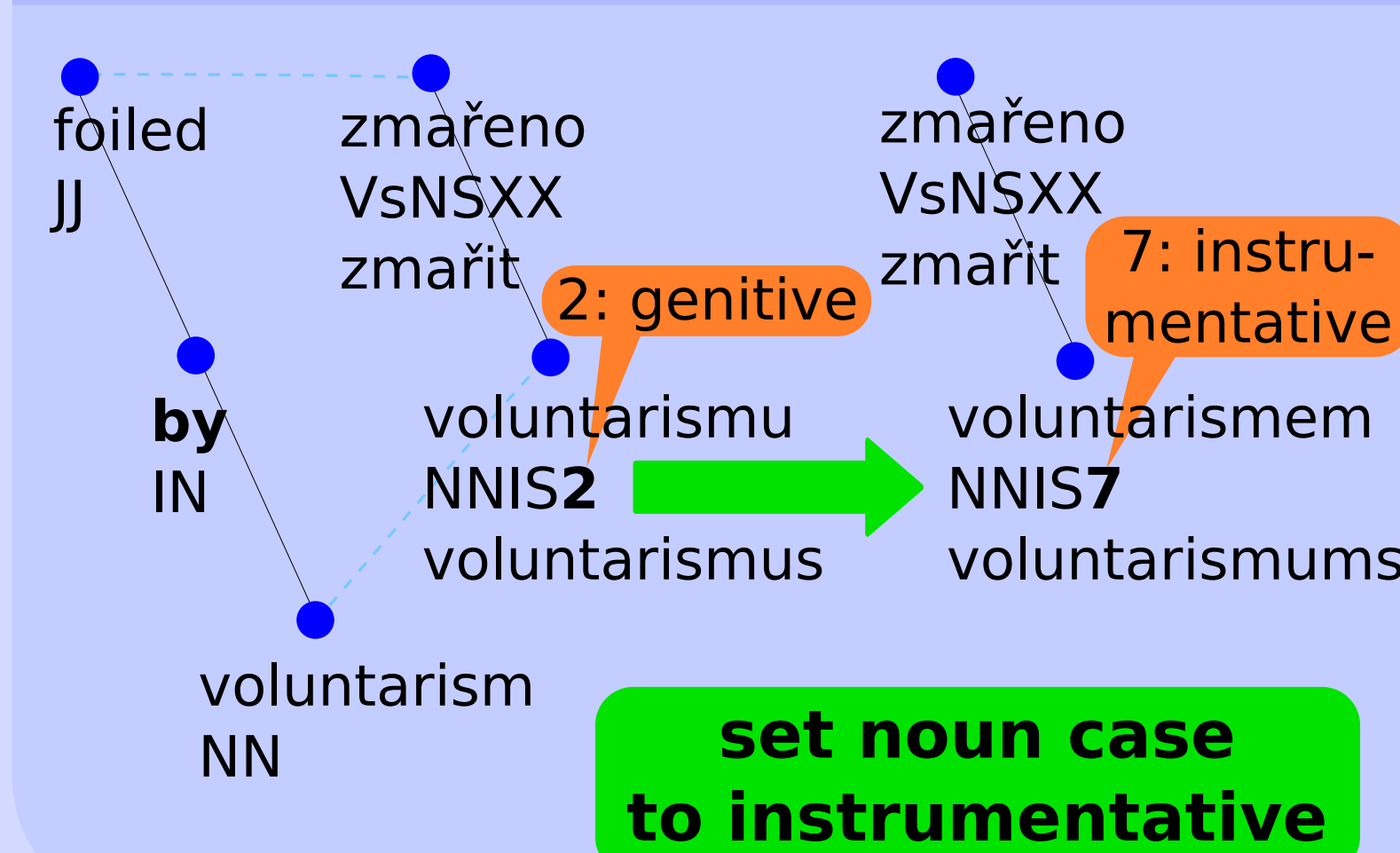
Verb Tense Translation



Preposition - Noun Agreement



Translation of 'by'



Depfix Evaluation (Δ BLEU)

Depfix improves most MT systems!

	WMT 2011	WMT 2012
CU Bojar	+0.47	+0.07
CU TectoMT	-0.10	-0.02
UEDIN	+0.64	+0.23
JHU	+0.42	+0.32
Eurotran	+0.21	+0.15
Microsoft Bing	?	+0.37
Google Translate	+0.23	0.00

Get Depfix Today!



- implemented in Perl in Treex NLP framework
- needs Linux + 4GB RAM
- 3 seconds per sentence
- released under GNU/GPL
- source is commented

<http://ufal.mff.cuni.cz/depfix>

CU Chimera English → Czech MT System

Input The source is available to all components to minimize error propagation.



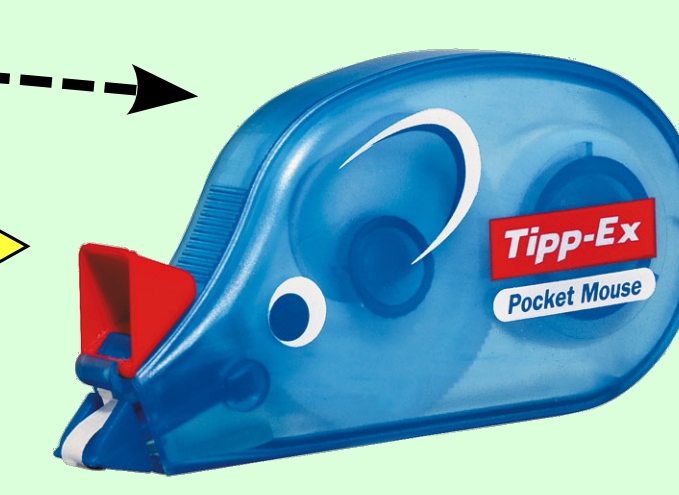
TectoMT (Popel)

- hybrid (rule-based/statistical) MT system
- transfer at tectogrammatical (deep syntactical) layer
- our combination: get an extra phrase table for Moses from TectoMT output



Moses (Bojar)

- phrase-based SMT
- large-scale data
- morphological tags as factors for a better grammatical coherence



Depfix (Rosa)

- automatic post-editing
- generates the final output

Evaluation of CU Chimera

CU Chimera is the state-of-the-art for English → Czech MT!

	WMT 2013			WMT 2014		
	cBLEU	BLEU	Manual	cBLEU	BLEU	Manual
CU Chimera	19.2	20.0	0.578	21.1	22.0	0.371
UEDIN	18.6	19.1	0.525	21.6	22.1	0.356
CU Bojar	18.9	20.1	0.580	20.9	22.1	0.333
Google Translate	?	?	0.562	20.2	?	0.169
CU TectoMT	14.2	14.7	0.476	15.4	15.8	-0.175