

Introduction

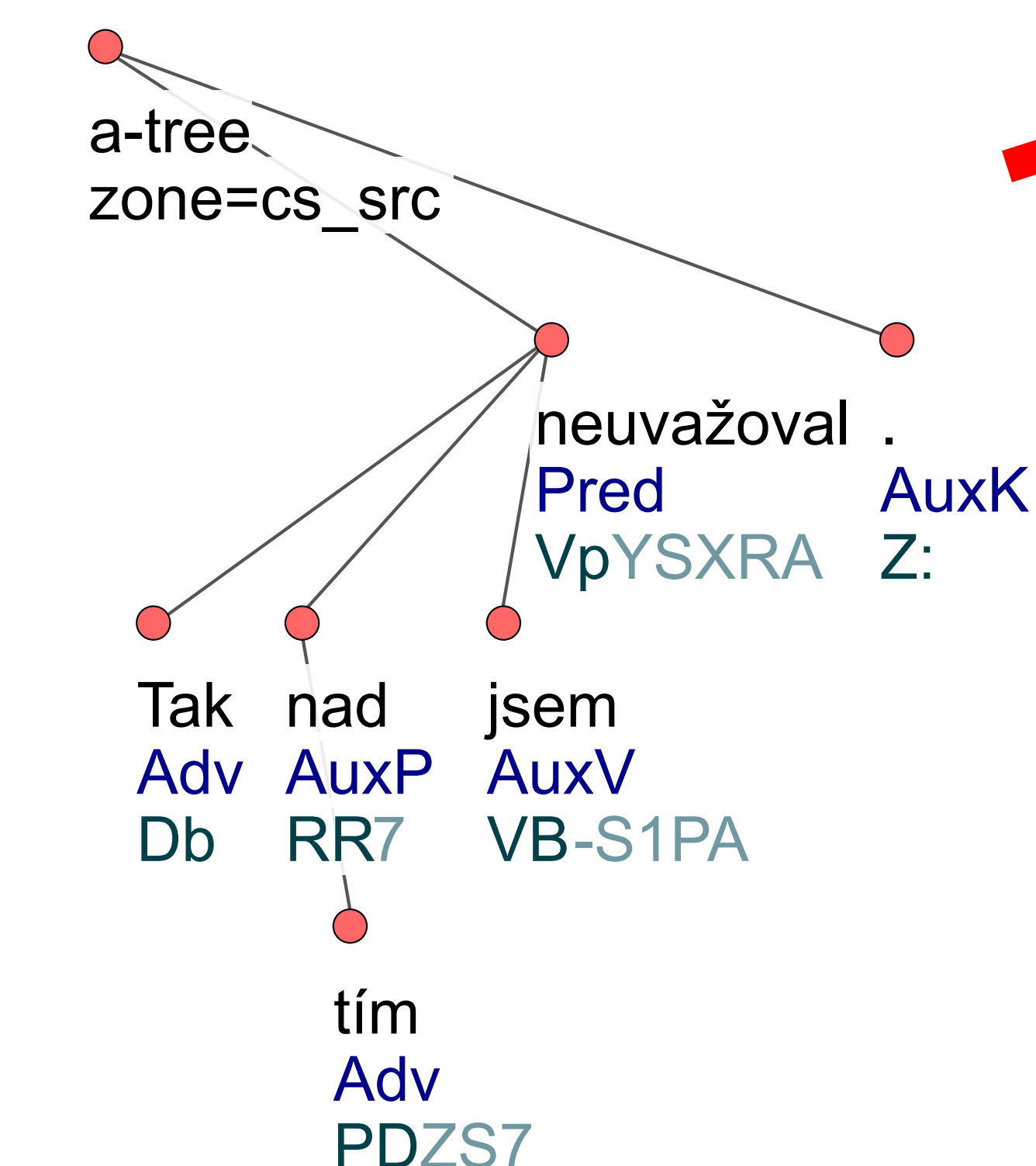
- TectoMT – competing in WMT since 2008, just English-Czech
- New language pairs** added in the QTLeap project:
 - Dutch, Spanish, Basque, Portuguese (to and from English)
 - Czech-English** – new language pair in WMT'15

New Languages: Standards & Training

- Treex blocks for new languages made easier:
 - Common morphology – Intersect
 - Common syntactic style – Hamlet 1.5 (3.0 / UD planned)
 - Base language-independent blocks
- Makefiles for easy translation model training

TectoMT System Operation

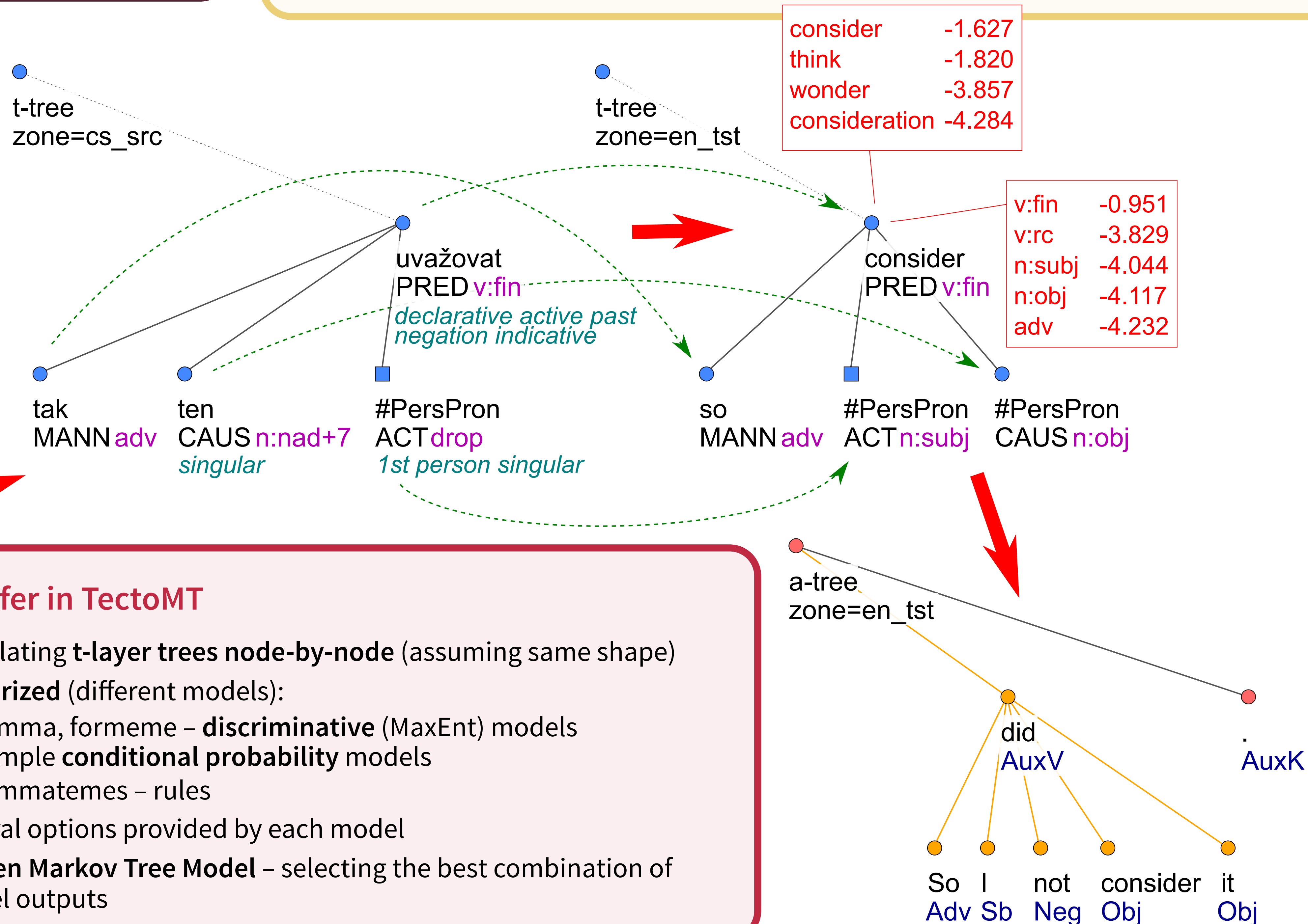
- Analysis – transfer – synthesis
- a-layer** – dependency trees
 - one node per token
- t-layer** – deep syntactic trees
 - only content words have nodes
 - t-lemma**: deep lemma
 - functor**: semantic/syntactic function label
 - formeme**: morpho-syntactic function label
 - grammatemes**: grammatical meaning



New Languages: Simple Model Combination

- Combining t-lemma + formeme predictions
- No training required**, as opposed to Hidden Markov Tree Model
- Five **non-parametric functions combining the models' outputs**:
 - AM-P – arithmetic mean of model-predicted probabilities
 - GM-P – harmonic mean of probs
 - HM-P – harmonic mean of probs
 - GM-Log-P – geometric mean of log-probs
 - HM-Log-P – harmonic mean of log-probs
- Tested on QTLeap corpus for 3 language pairs
- Performance **better than baseline** (using 1st variant of lemma and formeme)
- HMTM is better, if available**

Func/BLEU	EN-CS	EN-ES	EN-PT
Baseline	27.85	16.70	16.77
HMTM	28.76	-	-
AM-P	28.11	18.08	17.19
GM-P	28.18	18.06	17.11
HM-P	28.20	18.06	16.20
GM-Log-P	28.17	18.09	17.19
HM-Log-P	28.10	18.08	17.19



Transfer in TectoMT

- Translating **t-layer trees node-by-node** (assuming same shape)
- Factorized** (different models):
 - t-lemma, formeme – **discriminative** (MaxEnt) models + simple **conditional probability** models
 - grammatemes – rules
- Several options provided by each model
- Hidden Markov Tree Model** – selecting the best combination of model outputs

New in WMT'15: Czech-English Translation

Analysis

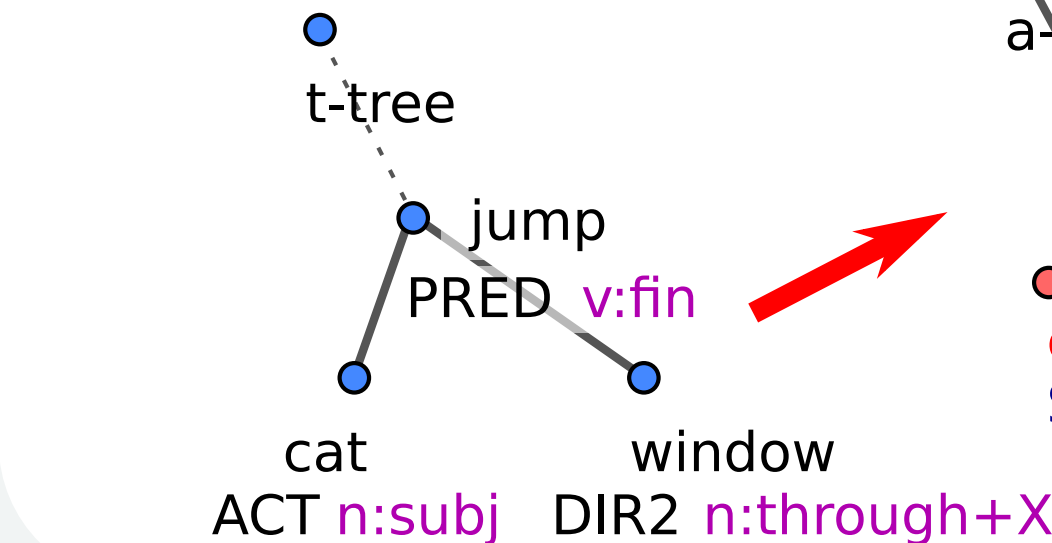
- Modified from CzEng, English-Czech TectoMT training
- MorphoDiTa tagger
- MST parser adapted for Czech

Transfer

- Basic t-lemma + formeme models with rule-based overrides
- Czech gender removed
- Double negatives removed
- Name translation fixes
- Grammateme fixes, e.g.: *těstoviny* (pl) -> *pasta* (sg)

Synthesis

- Gradually transform a copy of the translated t-tree



- Fill in morphological attributes (based on grammatemes)
- Mark subject (for agreement)
- Enforce basic word order
- Enforce subject-predicate agreement
- Add auxiliary words from formeme (prepositions, conjunctions)
- Add articles
- Add auxiliary verbs
- Remove imperative subjects
- Add negation particles
- Add punctuation
- Inflect words (Morphodita + Flect)
- Transform *a* -> *an*
- Delete repeated prepositions and conjunctions in coordinations
- Capitalize first word in sentence

WMT'15 Results

- English-Czech: 13.9% BLEU** – among the last, human eval. same
- used in Chimera, overall winner** in automatic+human ranking
 - 8.0% of reference tokens are only in TectoMT (and not Moses)
 - more than half of these tokens were used in Chimera
 - TectoMT is essential for Chimera's success
- Czech-English: 12.8% BLEU** – last, human eval. 2nd-to-last
 - pruning was too eager (bug)

Conclusion and Future Work

- Improvements and bugfixes required**
 - Hidden Markov Tree Model for Czech-English
 - word order fixes, article assignment (English)
- Further development plans:
 - Intersect** instead of grammatemes on t-layer, **Universal Dependencies**
 - Vowpal Wabbit** and **word embedding** features in transfer models
 - possibilities of non-isomorphic transfer