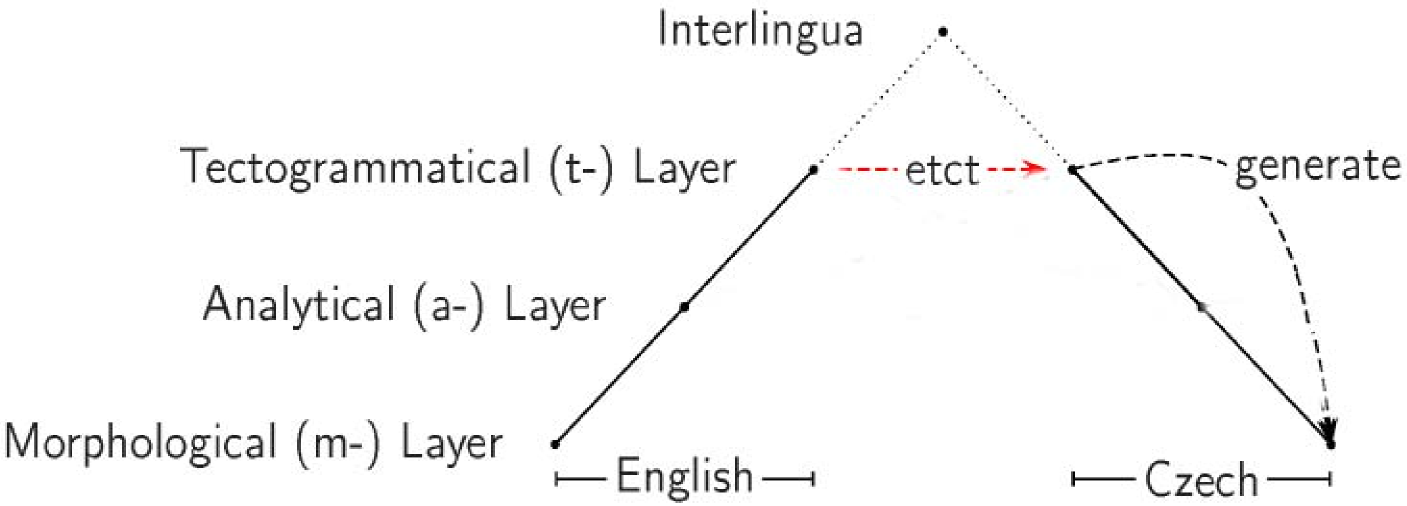
Towards English-to-Czech MT via Tectogrammatical Layer

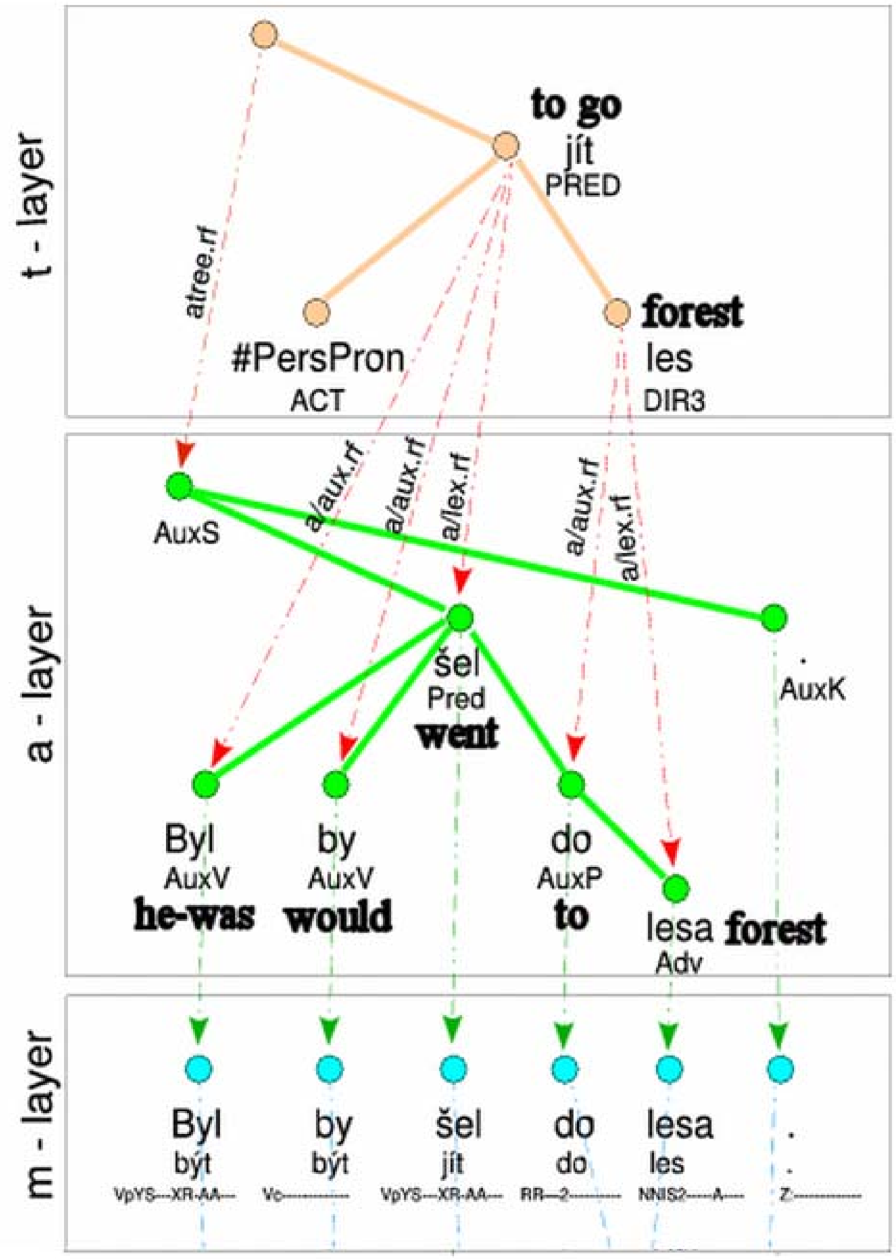
O.Bojar, S.Cinková, and J.Ptáček

Outline

- Overview
- Synchronous Tree Substitution Grammar
- Experimental Results
- Discussion

Overview: Deep Syntatic Machine Translation




t - layer
to go
jít
PRED
atree.rf
#PersPron
ACT
forest
les
DIR3
a - layer
a/aux.rf
a/aux.rf
a/lex.rf
a/aux.rf
a/lex.rf
AuxS
šel
Pred
went
.
AuxK
Byl
AuxV
he-was
by
AuxV
would
do
AuxP
to
lesa
Adv
forest
m - layer
Byl
být
VpYS---XR-AA---
by
být
Vc-------------
šel
jít
VpYS---XR-AA---
do
do
RR--2----------
lesa
les
NNIS2-----A----
.
.
Z:-------------

Under the Hood

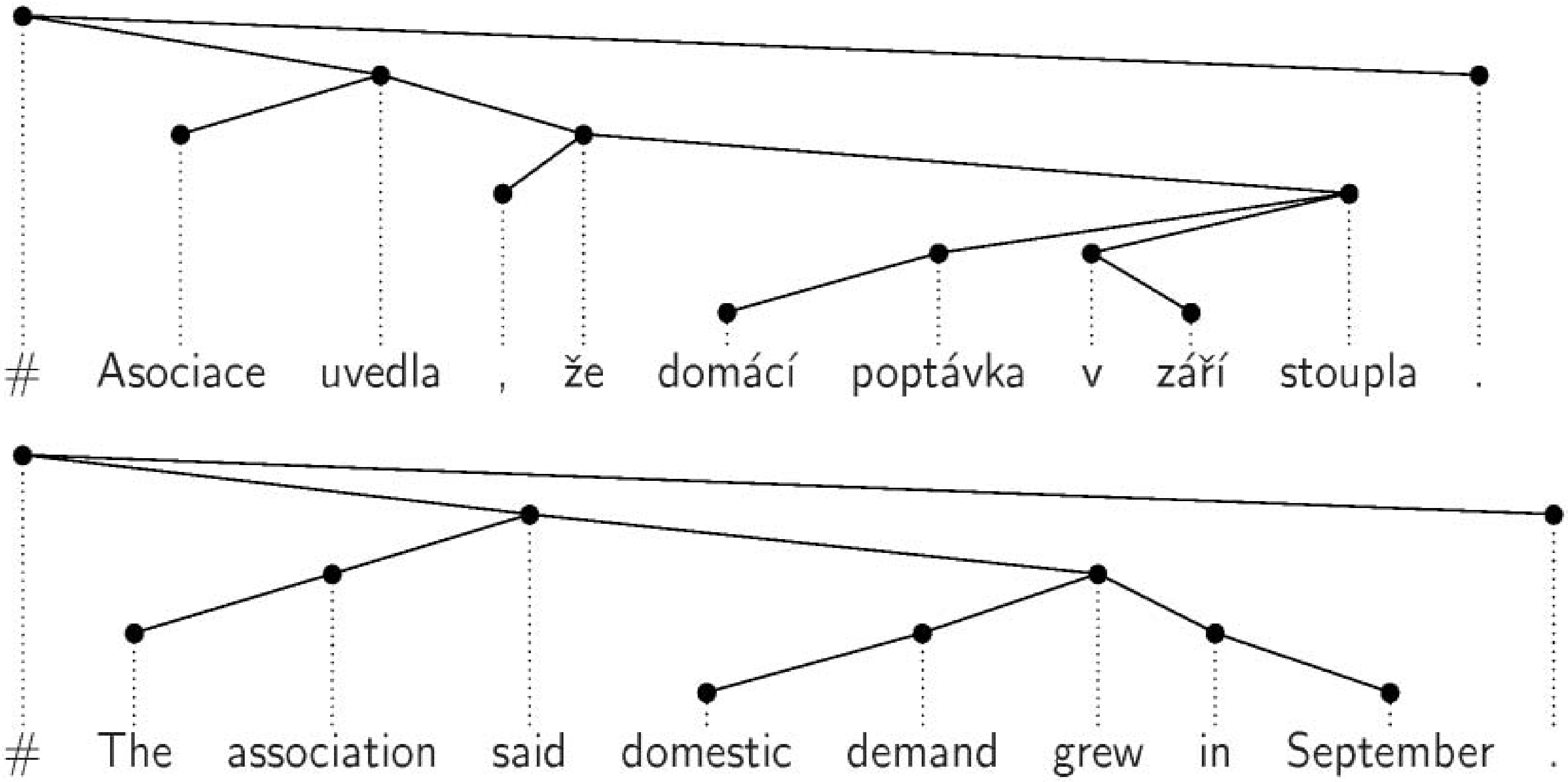
- Parallel treebank
- Little Trees Pairs Dictionary
- Synchronous Tree Substitution Grammar

- Parsing
- Covering the source tree with treelets with back-offs
- Reading the target tree
- Generation

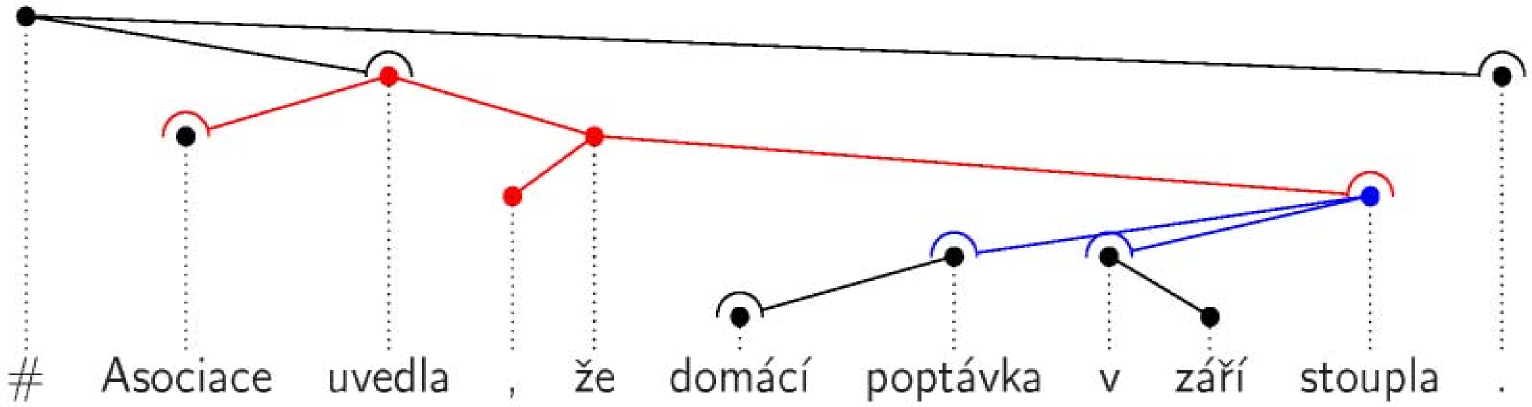
Synchronous Tree Substituon Grammar

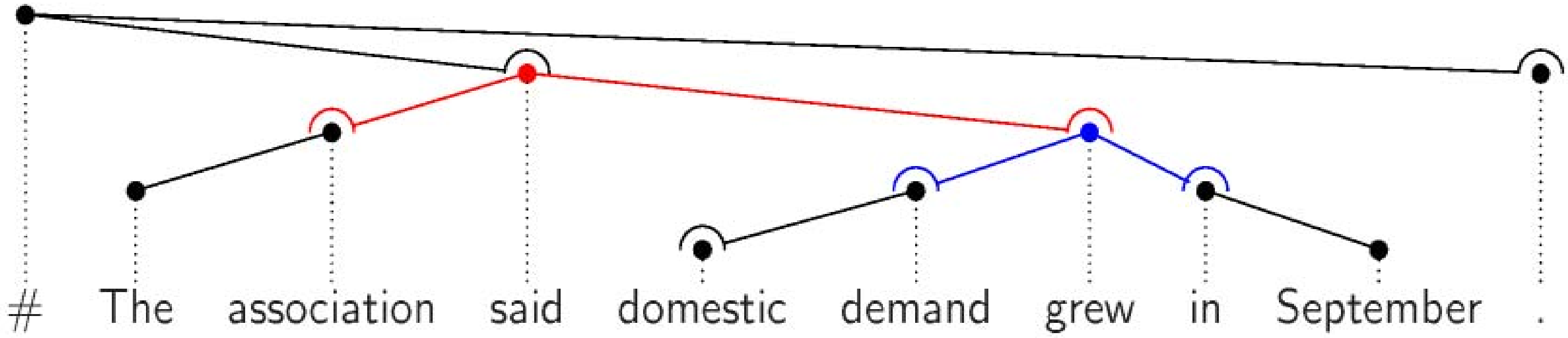
- (Hajič, 2002)
- (Eisner, 2003)
- (Čmejrek, 2006)
- Not bounded to t-trees

Idea: Observe a Pair of Dependency Trees

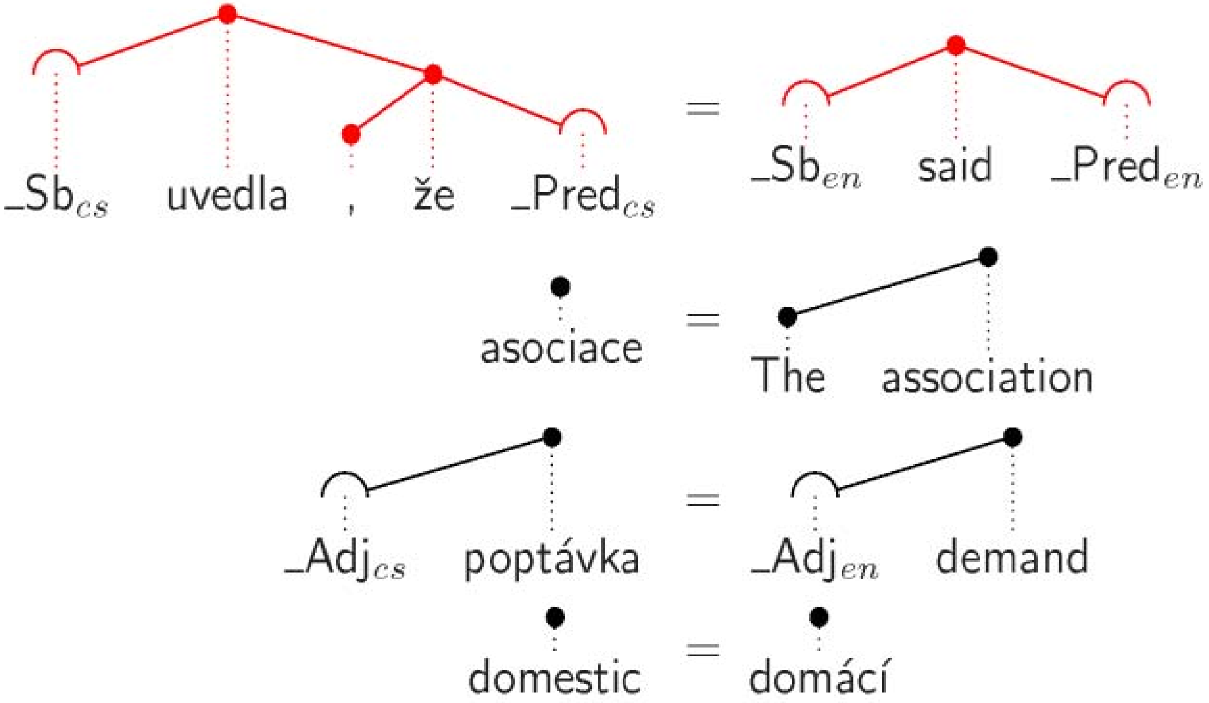


Idea: Decompose Trees into Treelets





Idea: Collect Dictionary of Treelet Pairs



_Sb$_{cs}$ uvedla , že _Pred$_{cs}$ = _Sb$_{en}$ said _Pred$_{en}$

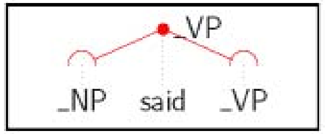
asociace = The association

_Adj$_{cs}$ poptávka = _Adj$_{en}$ demand

domestic = domácí

Little Trees Formally

Given a set of states Q and a set of word labels L, we define:



A LITTLE TREE or TREELET t is a tuple (V, V^i, E, q, l, s) where:

- V is a set of NODES,
- $V^i \subseteq V$ is a nonempty set of INTERNAL NODES. The complement $V^f = V \setminus V^i$ is called the set of FRONTIER NODES,
- $E \subseteq V^i \times V$ is a set of directed edges starting from internal nodes only and forming a directed acyclic graph,
- $q \in Q$ is the ROOT STATE,
- $l : V^i \rightarrow L$ is a function assigning labels to internal nodes,
- $s : V^f \rightarrow Q$ is a function assigning states to frontier nodes.

Optionally, we can keep track of local or global ordering of nodes in treelets.

I depart from Čmejrek (2006) in a few details, most notably I require at least one internal node in each little tree.

Treelet Alignments: Heuristics

- Similar to common phrase-extraction techniques given word alignments.
- Basic units are little trees instead of word spans.

1. Obtain **node-to-node alignments** (GIZA++ on linearized trees).
2. Extract all treelet pairs satisfying these conditions:
   - no more than i internal nodes and f frontier nodes,
   - **compatible with node alignment**,

     e.g. no node-alignment link leads outside the treelet pair and frontiers are linked.
   - satisfying **STSG property**.

     All children of an internal node have to be included in the treelet (as frontiers or internals), ie. assume no adjunction operation was necessary to construct the full tree.
3. Estimate probabilities, e.g. $p(t_1, t_2 | \text{root state}_1, \text{root state}_2)$

Back-off Schemes

Preserve all. Full-featured treelets are collected in training phase.
Required treelets often never seen in training data ⇒ back-off needed.

Drop frontiers. Observed treelets reduced to internal nodes only.
Given a source treelet, internals translated by the dictionary, frontiers generated on the fly, labelled and positioned probabilistically.

Word for word. Useful for single-internal treelets only: The label of the root internal translated independently, frontiers generated on the fly, labelled probabilistically, order unchanged.

Keep a word non-translated to handle unknown words.
Allowed only for single-internal treelets, frontiers mapped probabilistically.

Decoding STSG

- Find target tree such that the synchronous derivation δ is most likely.
- Implemented as two-step top-down beam-search similar to Moses:

1. Prepare **translation options table**:
   - For every source node consider every subtree rooted at that node.
   - If the subtree matches the source treelet in a treelet pair, we've got a translation option.
   - Keep only best τ translation options at a node.

2. Gradually **expand partial hypotheses**:
   - Starting at root use translation options to cover source tree.
   - Keep only best σ partial hypotheses of a given size (input nodes covered).

Implementation Details

- STSG model extended to log-linear combination of features:

$$\text{best derivation } \hat{\delta} = \underset{\delta \in \Delta(T_1)}{\operatorname{argmax}} \exp\Big(\sum_{m=1}^{M} \lambda_m h_m(\delta)\Big) \tag{1}$$

$$\text{instead of } \hat{\delta} = \underset{\delta \in \Delta(T_1)}{\operatorname{argmax}} p(t^0_{1:2}|Start_{1:2}) * \prod_{i=1}^{k} p(t^k_{1:2}|q^k_{1:2}) \tag{2}$$

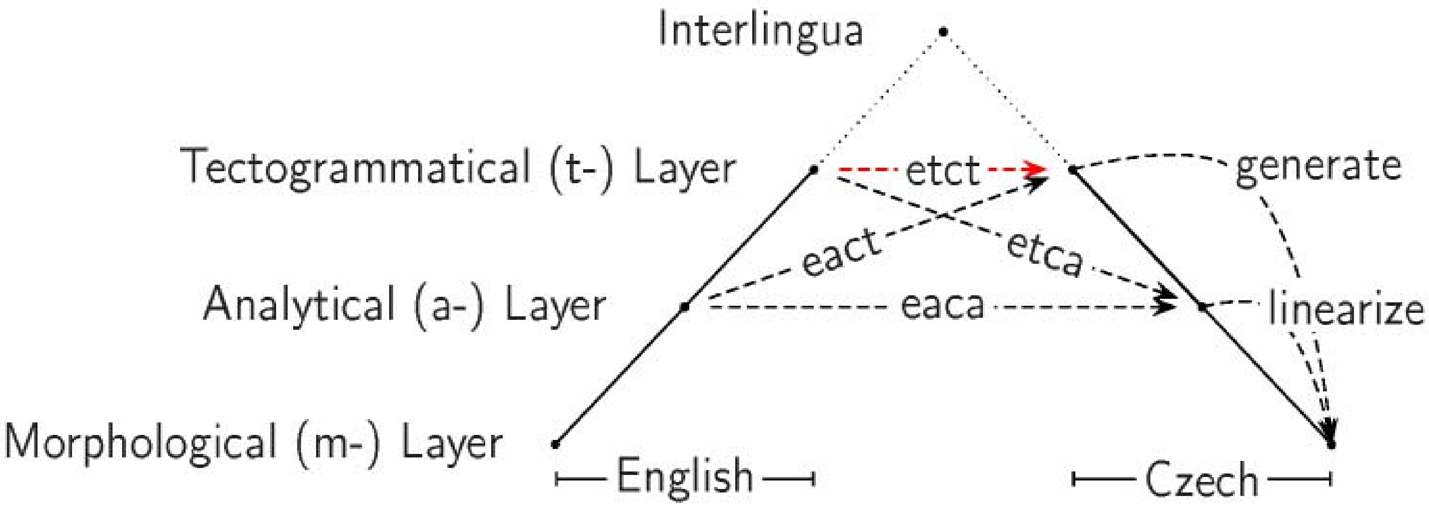
- Tinycdb (like GDBM) to store and access treelet dictionaries.
- Target tree structure can be disregarded (output linearized right away).
  - IrstLM to promote hypotheses containing frequent trigrams.

- Implemented in Mercury (Somogyi, Henderson, and Conway, 1995).
- Parallel computation on Sun Grid Engine cluster (160 CPUs in 40 machines).

Generator

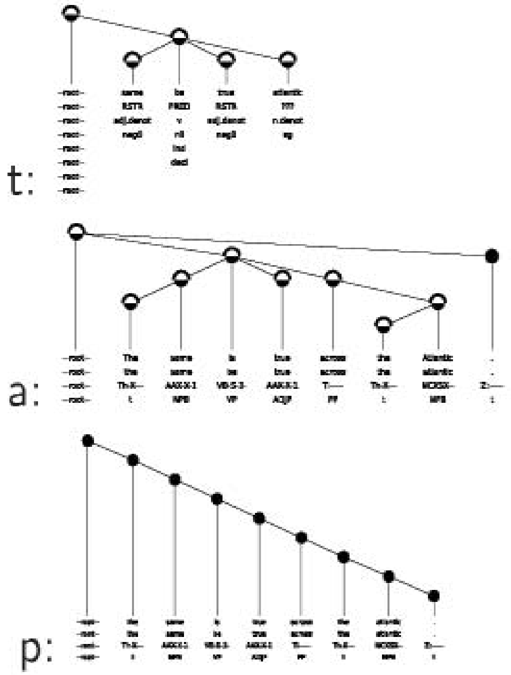
- Deterministic
- First plan - then fulfil
- Expects full featured t-trees on input

Overview: Deep Syntatic Machine Translation



Empirical Evaluation

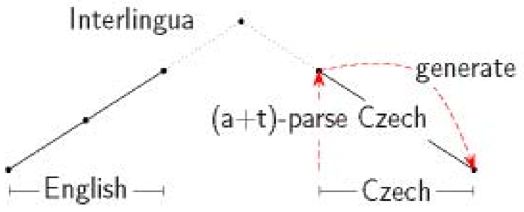
	BLEU
eact	3.0±0.3
etct	5.0±0.5
etca	6.3±0.6
eaca	8.6±0.6
epcp	10.3±0.7
epcp with no state labels	11.0±0.7
Phrase-based (Moses)	
Vanilla	12.9±0.6
Factored	14.2±0.7



ACL 2007 WMT shared task data, 55k training sentences, 964 test sentences.

Upper Bound on MT Quality via t-layer

- Analyse Czech sentences to t-layer.
- Optionally ignore some node attributes.
- Generate Czech surface.
- Evaluate BLEU against input Czech sentences.



	BLEU
Full automatic t-layer, no attributes ignored	36.6±1.2
Ignore sentence mood (assume indicative)	36.6±1.2
Ignore verbal fine-grained info (resultativeness, . . .)	36.6±1.2
Ignore verbal tense, aspect, . . .	24.9±1.1
Ignore all grammatemes	5.3±0.5

⇒ Node attributes obviously very important.