Wrestling with Deep Syntactic Translation from English to Czech

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

Outline

- Properties of Czech
- Motivation for deep syntax.
- Tree-to-tree transfer using Synchronous Tree-Substitution Grammars.
- Methods of back-off.
- Empirical evaluation.
- Sources of errors.

Properties of Czech language

	Czech	English
Rich morphology	$\geq$ 4,000 tags possible, $\geq$ 2,300 seen	50 used
Word order	free	rigid

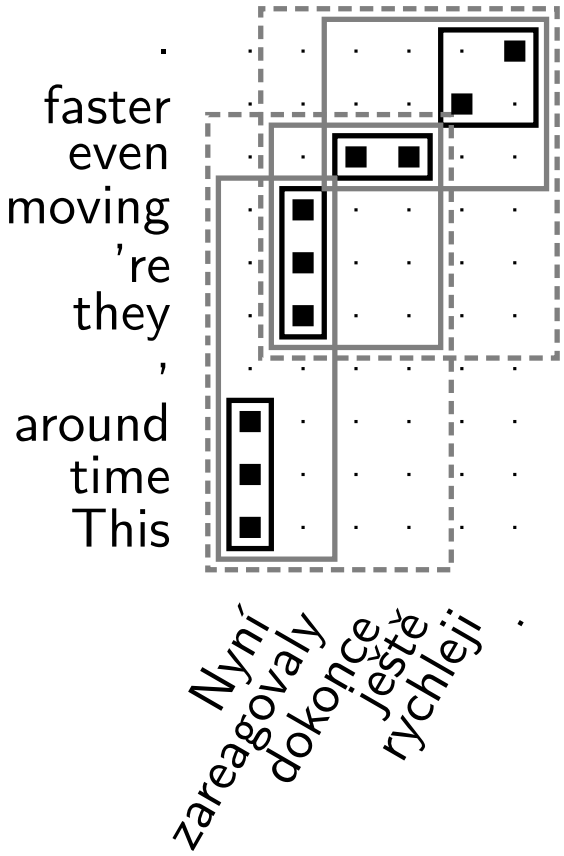
- rigid global word order phenomena: clitics
- rigid local word order phenomena: coordination, clitics mutual order

Nonprojective sentences	16,920	23.3%
Nonprojective edges	23,691	1.9%

Known parsing results	Czech	English
Labelled edge accuracy	80.19%	89.61%
Unlabelled edge accuracy	86.28%	90.63%

Data by Nivre et al. (2007), Zeman (personal web page), Holan (2003), and Bojar (2003). Consult Kruijff (2003) for empirical measurements of word order freeness.

Baseline: Phrase-Based MT

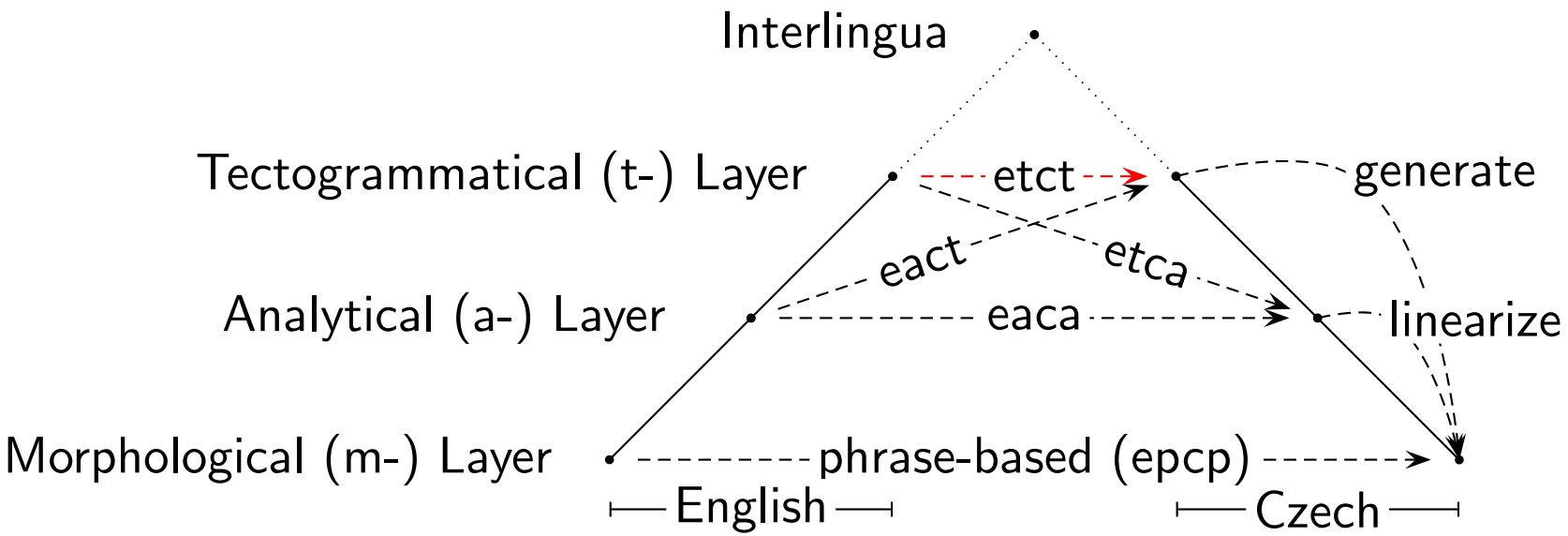


This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Phrase-based MT: choose such segmentation of input string and such phrase "replacements" to make the output sequence "coherent" (3-grams most probable).

Open-source implementation: www.statmt.org/moses

Overview: Deep Syntatic Machine Translation



Tectogrammatics: Deep Syntax Culminating

Background: Prague Linguistic Circle (since 1926).

Theory: Sgall (1967), Panevová (1980), Sgall et al. (1986).

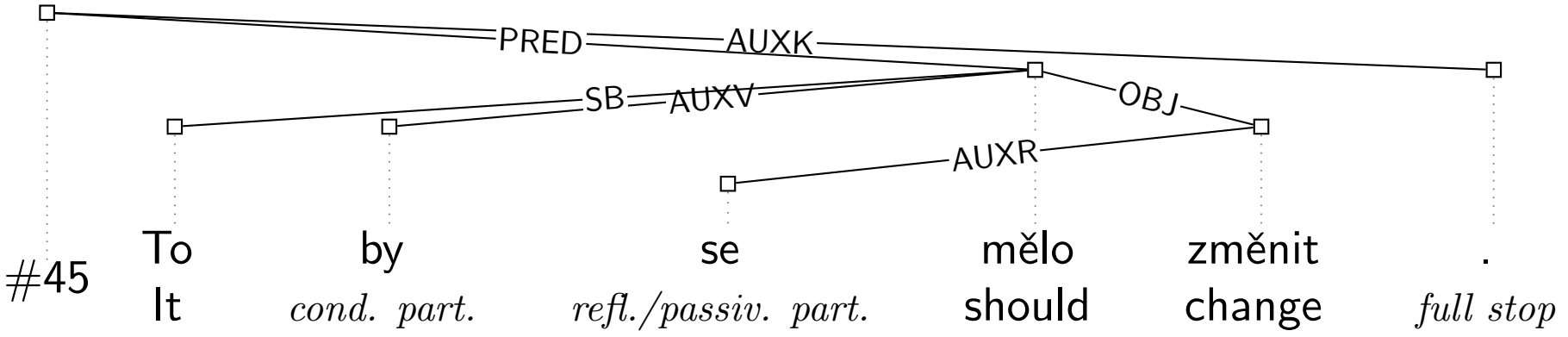
Materialized theory — Treebanks:

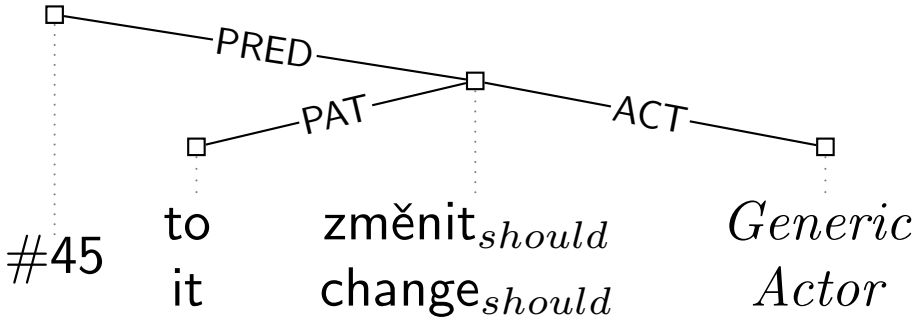
- Czech: PDT 1.0 (2001), PDT 2.0 (2006)
- Czech-English: PCEDT 1.0 (2004), PCEDT 2.0 (in progress)
- Arabic: PADT (2004)

Practice — Tools:

- parsing Czech to surface: McDonald et al. (2005)
- parsing Czech to deep: Klimeš (2006)
- parsing English to surface: well studied (+rules convert to dependency trees)
- parsing English to deep: heuristic rules (manual annotation in progress)
- generating Czech surface from t-layer: Ptáček and Žabokrtský (2006)

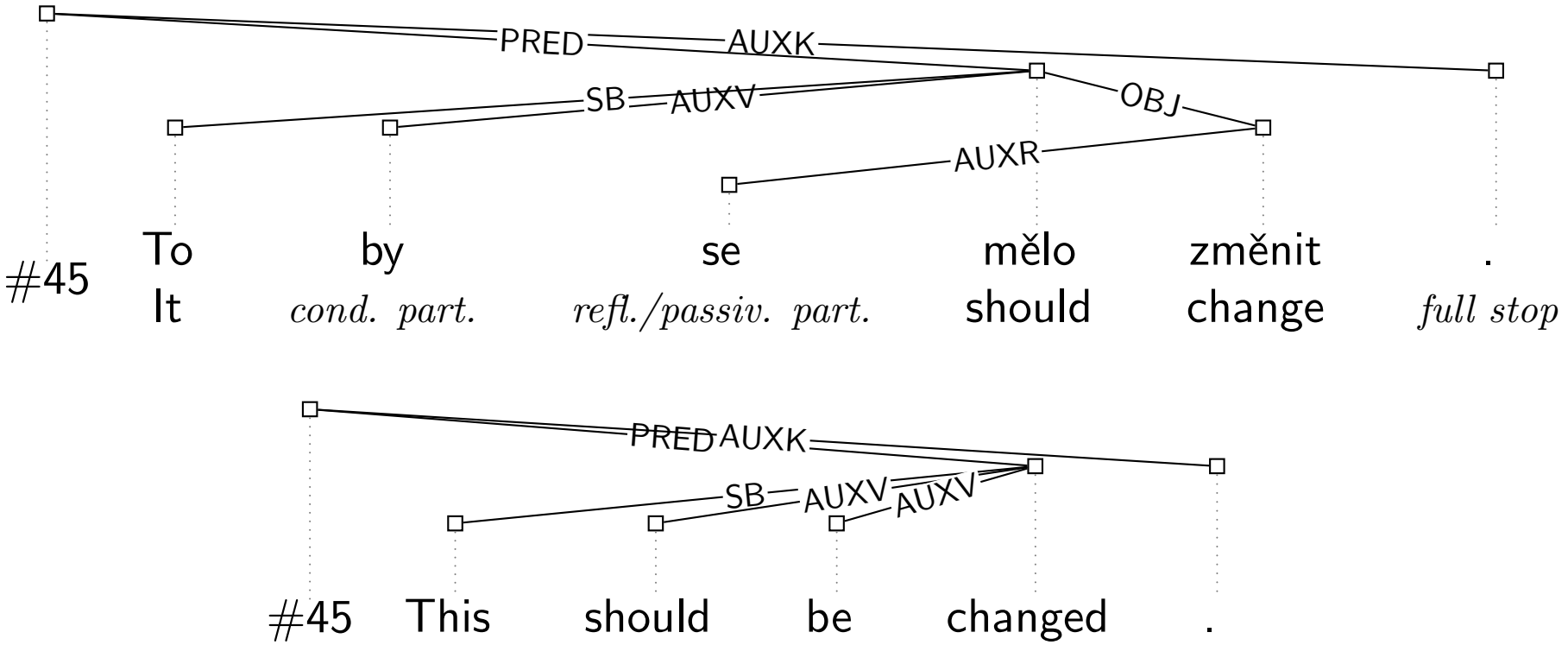
Analytical vs. Tectogrammatical



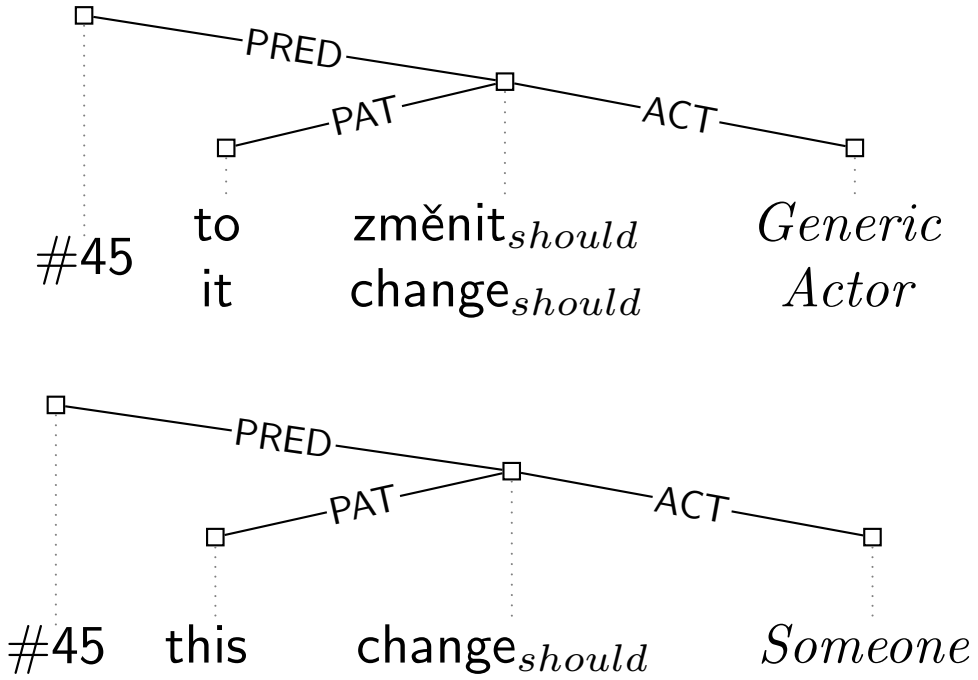


- hide auxiliary words, add nodes for "deleted" participants
- resolve e.g. active/passive voice, analytical verbs etc.
- "full" tecto resolves much more, e.g. topic-focus articulation or anaphora

Czech and English Analytical



Czech and English Tectogrammatical



Predicate-argument structure: change$_{should}$(ACT: someone, PAT: it)

The Tectogrammatical Hope

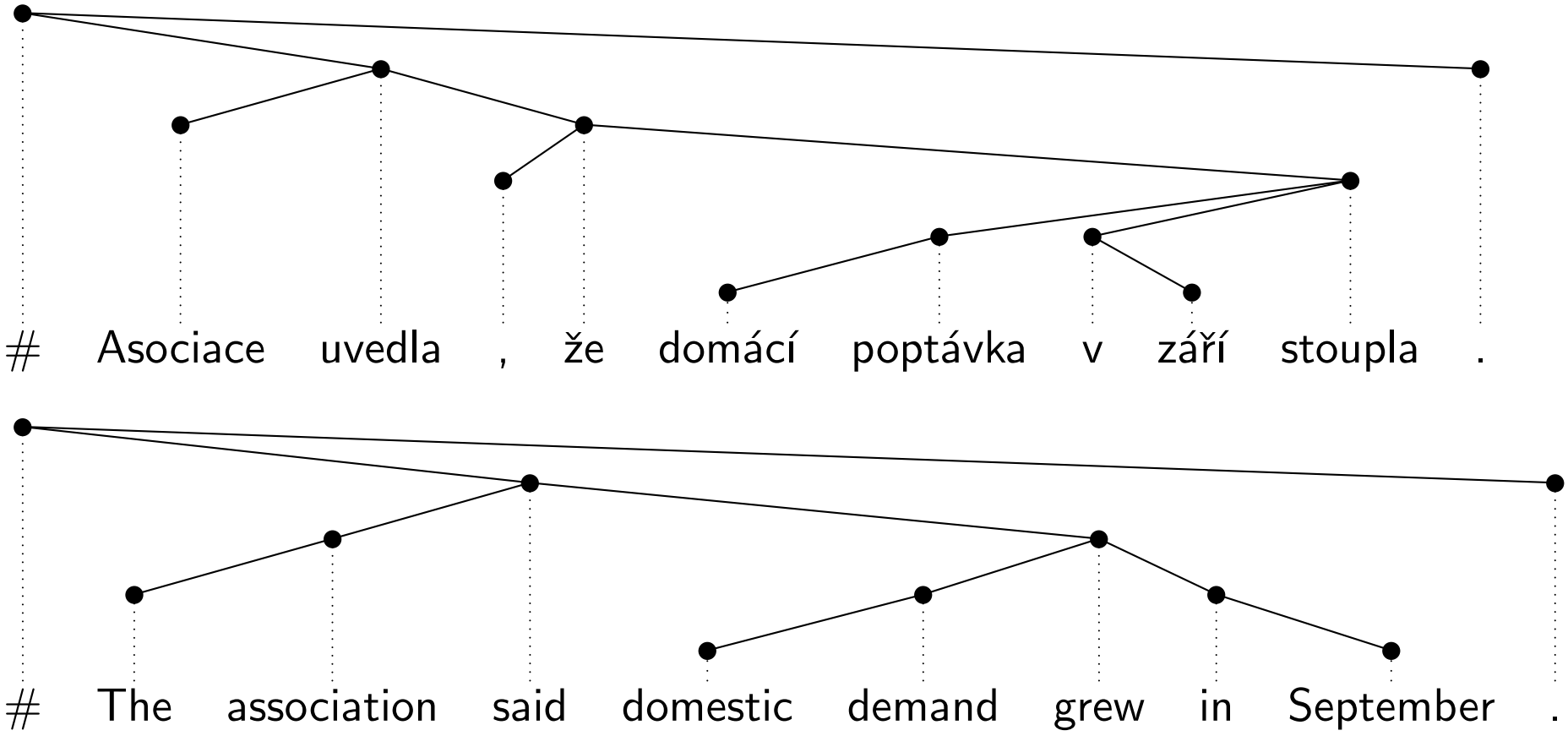
Transfer at t-layer should be easier than direct translation:

- Reduced structure size (auxiliary words disappear).
- Long-distance dependencies (non-projectivites) solved at t-layer.
- Word order ignored / interpreted as information structure (given/new).
- Reduced vocabulary size (Czech morphological complexity).
- Czech and English t-trees structurally more similar
  $\Rightarrow$less parallel data might be sufficient (but more monolingual).
- Ready for fancy t-layer features: co-reference.

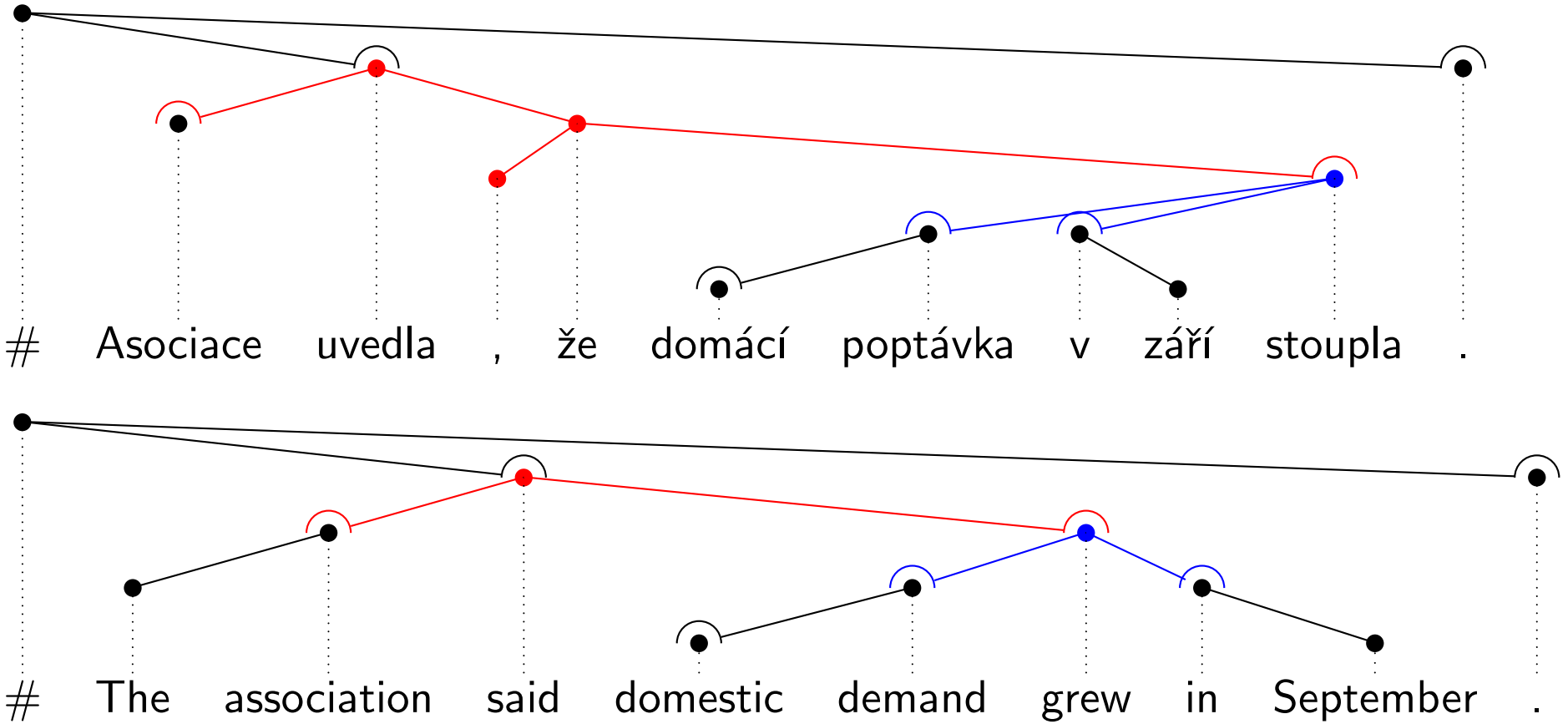
The complications:

- 47 pages documenting data format (PML, XML-based, sort of typed)
- 1200 pages documenting Czech t-structures
  "Not necessary" once you have a t-tree but useful understand or to blame the right people.

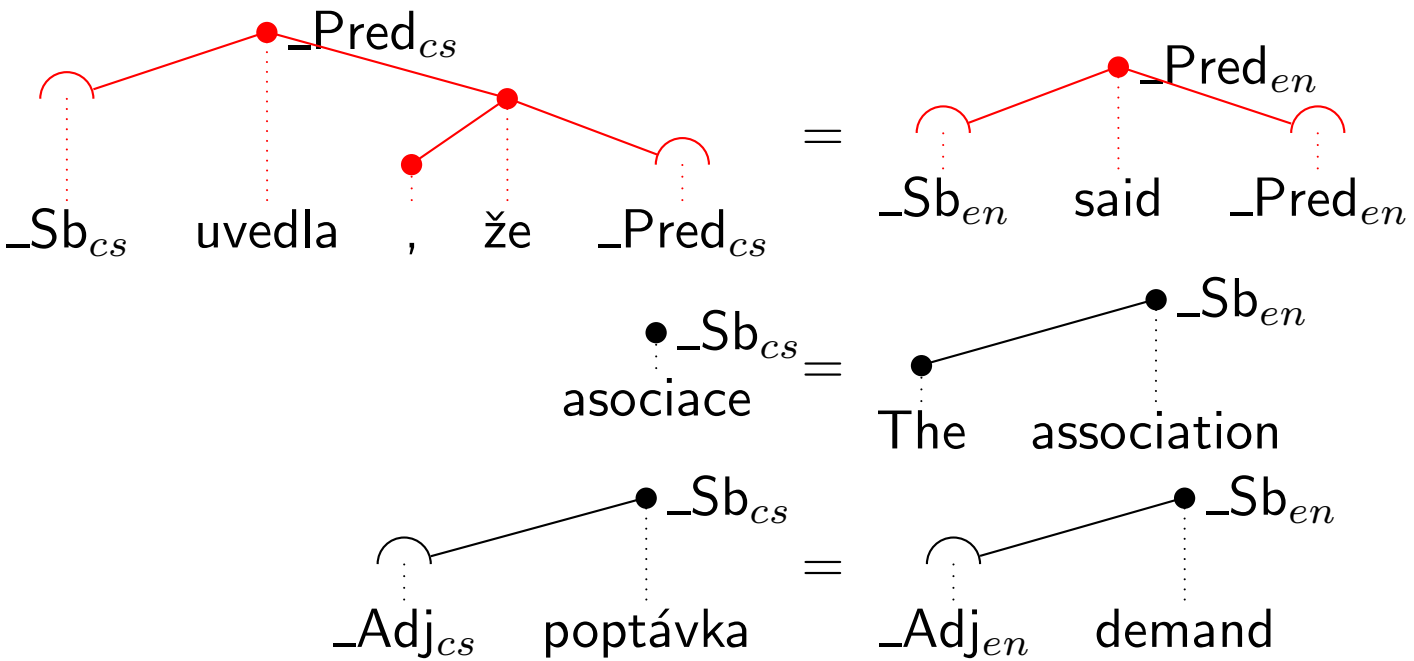
Idea: Observe a Pair of Dependency Trees



Idea: Decompose Trees into Treelets



Idea: Collect Dictionary of Treelet Pairs



. . . **Synchronous Tree Substitution Grammar**, e.g. Čmejrek (2006).

Decoding STSG

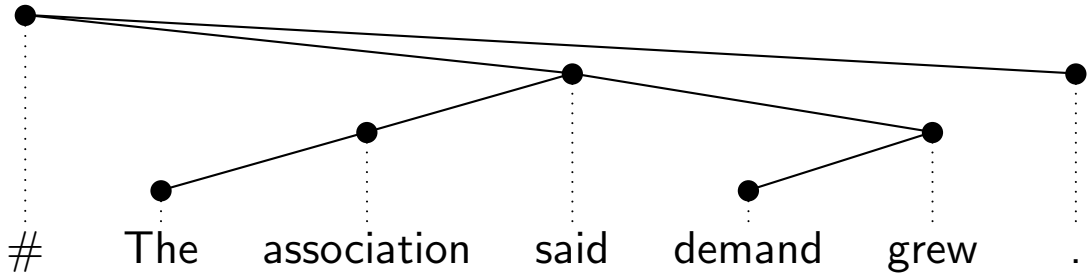
Given an input dependency tree:

- decompose it into known treelets,
- replace treelets by their treelet translations,
- join output treelets and produce output final tree; linearize or generate plaintext.

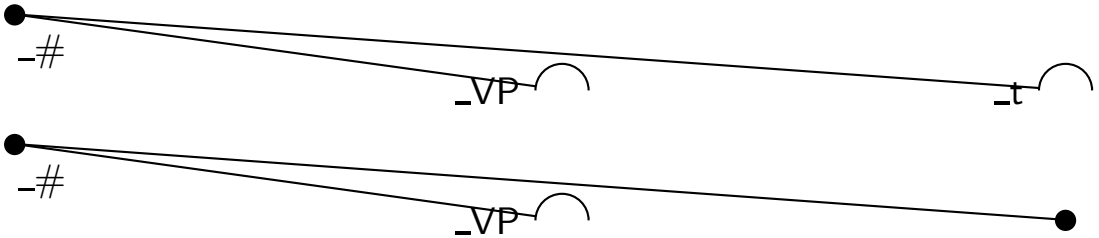
Implemented as two-step top-down beam-search similar to Moses:

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).

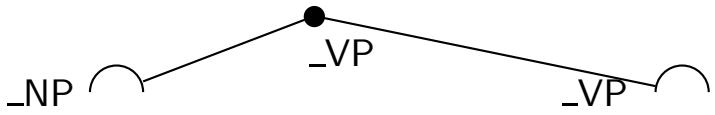
Translation Options Example



Sample translation options at root:



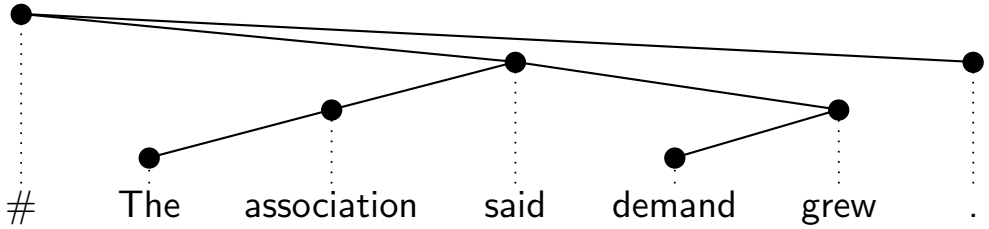
Sample translation options at 'said':



Sample translation options at '.':


⇒ .

Hypothesis Expansion Example



Sample Derivation:

Linearized output:

h_0 _# ⇒ _#

h_1 _# _VP ⇒ # _Pred .

h_2 _# _NP _VP ⇒ # _Sb uvedla , že _Pred .

h_3 _# _NP _NP ⇒ # _Sb uvedla , že _Sb stoupla .

Treelet Alignments: Heuristics

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

1. Parse both sides of the parallel corpus.
2. Obtain **node-to-node alignments** (GIZA++ on linearized trees).
3. 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.
4. Estimate probabilities, e.g. $p(t_1, t_2 | rootstate_1, rootstate_2)$

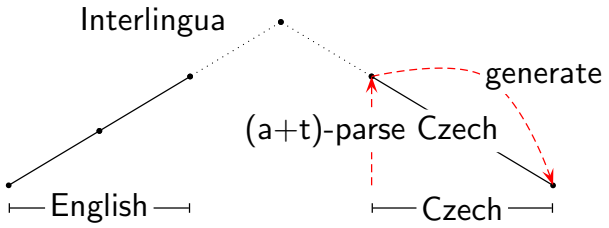
Another option is an EM-loop by Eisner (2003).

In Reality, t-nodes are not Atomic!

t-nodes have about 25 attributes: t-lemma, functor, gender, person, tense, iterativeness, dispositional modality, . . .

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.

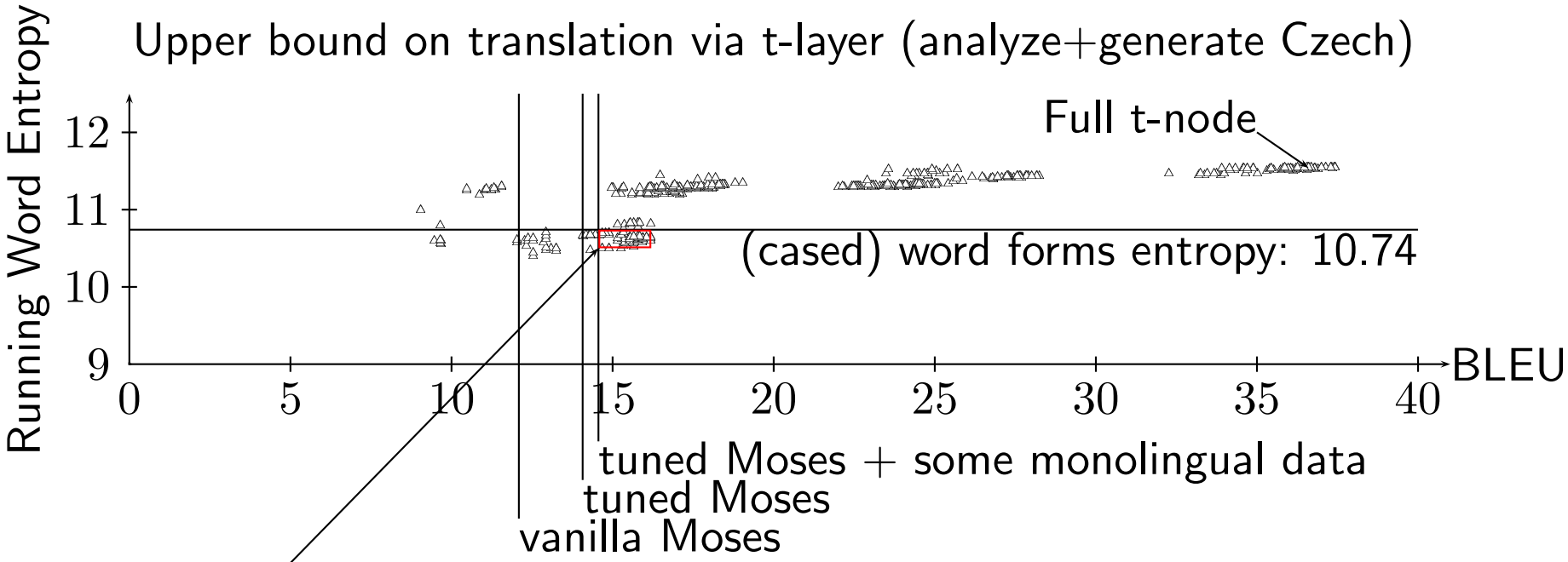


	BLEU
Full automatic t-layer, no attributes ignored	36.6±1.2
Ignore sentence mood (assume indicative)	36.6±1.2
Ignore verbal fine-grained info (resultativeness, . . .)	36.6±1.2
Ignore verbal tense, aspect, . . .	24.9±1.1
Ignore all grammatemes	5.3±0.5

⇒ Node attributes obviously very important.

No More Fairy Tales on Vocabulary Reduction

Can we find a balance of small vocabulary and high achievable BLEU?



Space for improvement assuming:

- t-nodes atomic (with a restricted set of attributes)
- we wish to stay below the entropy of plain text

⇒ Very limited achievable BLEU even if tranfer were absolutely perfect.

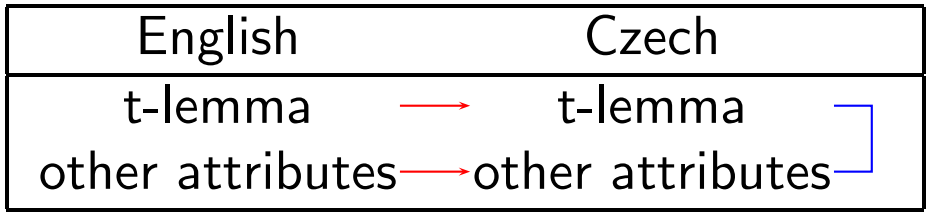
Consequence: Need for Factored Translation

- t-layer by itself *increases* complexity of node label choice.

⇒ cannot treat output nodes labels as atomic: `go.V.past.third.sg...`

The bare minimum:

- two factors to translate lexical item and grammatical features separately
- check for internal output compatibility (using more monolingual data)

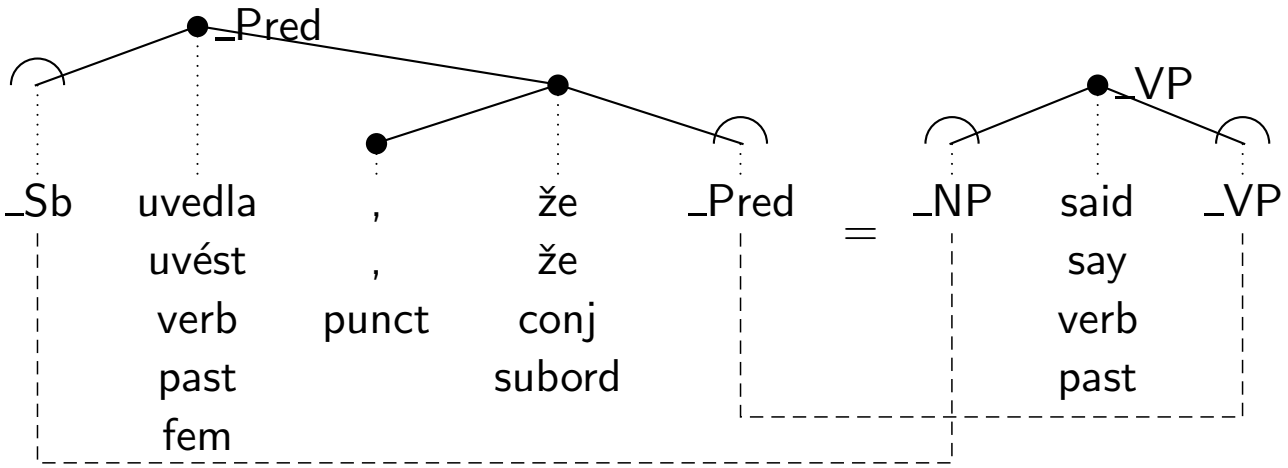


Conflicting with the key concept of STSG: treelet shape (and size) alternations.

Current implementation is "synchronous", as in Moses:
– translation options fully specify all output factors.

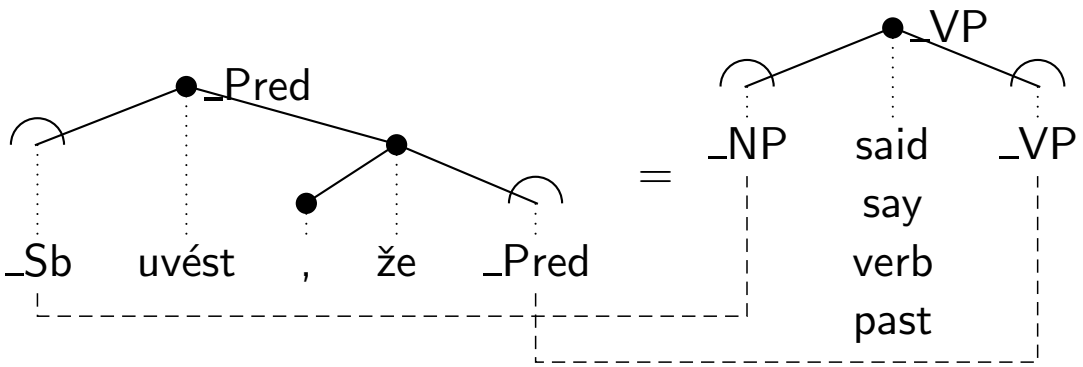
Treelet Construction 1: Preserve Everything

- Most basic method (no back-off).
- Preserves:
  - shapes of treelets, ordering of nodes,
  - all factors (attributes) of internal nodes,
  - position and states of frontier nodes.



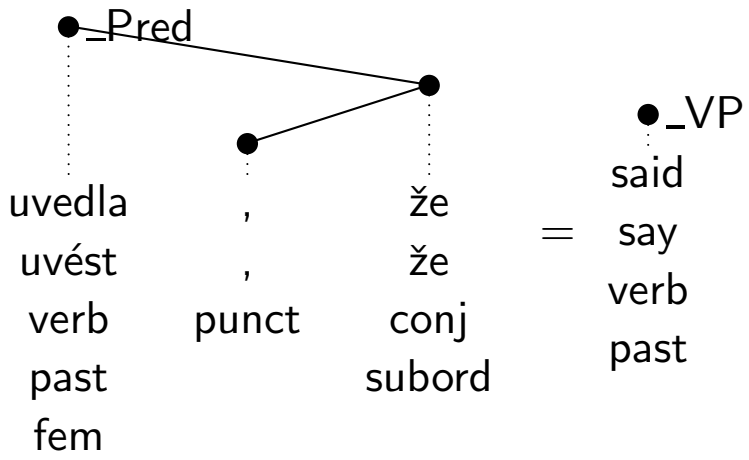
Construction 2: Ignore Input Factors

- Back-off by ignoring some of input factors ⇒ reduced source data sparseness.
- Output factors fully specified (i.e. their values guessed).



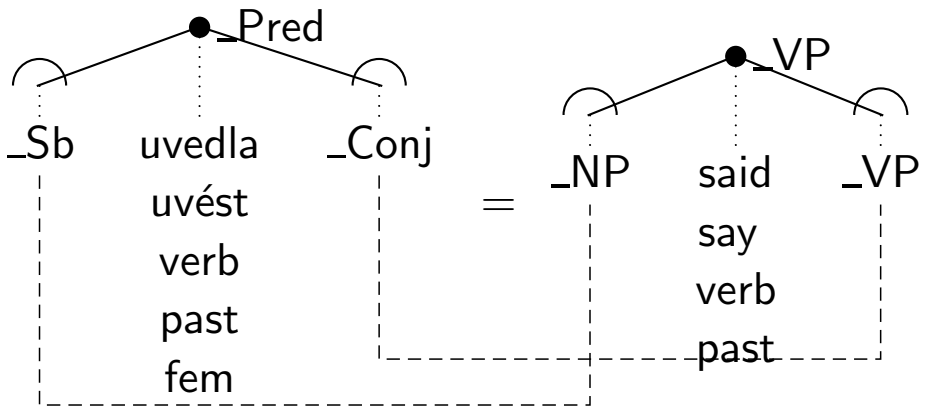
3: Translate Internals, Generate Frontiers

- We preserve: treelet shape, all factors of internal nodes.
- In training, frontier nodes are ignored (dropped).
- In translation, frontier nodes are translated and positioned one-by-one.
  - Available only when producing linearized output (no need to reconstruct the structure).
  - Frontiers are placed before or after internal nodes, not in between.



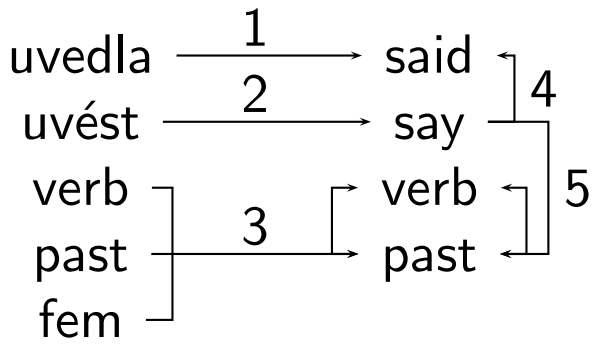
4: Translate Node-by-node

- Like (3), but treelets limited to one internal only.
  $\Rightarrow$ trivial to reconstruct treelet structure.
- Frontiers ignored in training, and translated one-by-one.
  - Ordering preserved for the sake of simplicity.
- If used alone, the source and target trees will have equal number of nodes:
  $\Rightarrow$ Not suitable for transfer at a-layer.



5: Factored Translation (node-by-node)

- Analogous to factored phrase-based translation (Koehn and Hoang, 2007).
- The configuration specifies a sequence of steps:
  - **Mapping steps** convert input factors to output factors.
  - **Generation steps** bind values of output factors.
  - The order of steps is important due to the limited stack of partial hypotheses.
- Currently limited to node-to-node translation to avoid conflicting structures.



Combining Models

- The original STSG (Eisner, 2003) extended to a log-linear model:

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

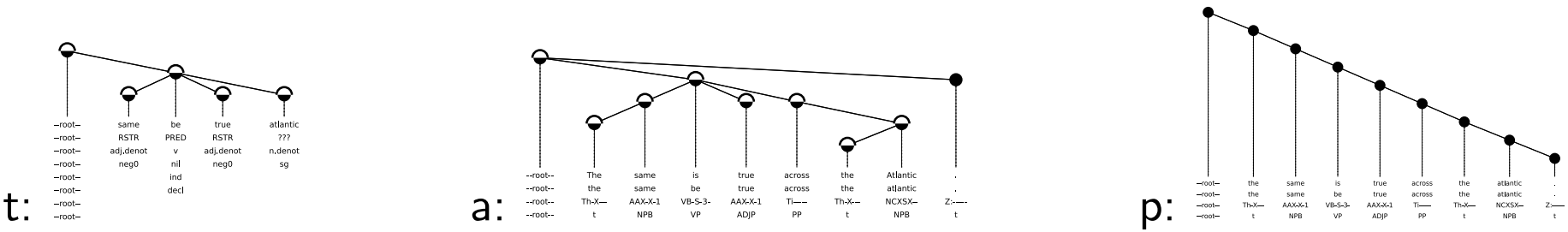
$$\text{instead of } \hat{\delta} = \underset{\delta \in \Delta(T_1)}{\operatorname{argmax}} p(t^0_{1:2} | Start_{1:2}) * \prod_{i=1}^{k} p(t^k_{1:2} | q^k_{1:2}) \tag{2}$$

- The configuration specifies treelet-construction methods to use:
  - E.g. prefer "Preserve everything" but back-off to factored node-by-node.
- Weights λ_m of simultaneously used models chosen to achieve high BLEU.
  - Implemented binding to two MERT methods: (Och, 2003) a (Smith and Eisner, 2006)
  - Fails to converge (too many weights) $\Rightarrow$ manually pick some values.

Current Results (BLEU)

Layers \ Language Models	no LM	n-gram/*binode*
epcp, no factors	8.65±0.55	10.90±0.63
eaca, no factors	6.59±0.52	8.75±0.61
etca, no factors	-	6.30±0.57
etct factored, preserving structure	5.31±0.53	*5.61±0.50*
eact, source factored, output atomic	-	*3.03±0.32*
etct, no factors, all attributes	1.61±0.33	*2.56±0.35*
etct, no factors, just t-lemmas	0.67±0.19	-

t: a: p:



Why the t-layer Performs So Poorly?

- **Cumulation of Errors**:
  - e.g. 93% tagging * 85% parsing * 93% tagging * 92% parsing = 67%
  - Still using rather ancient tools: (Ratnaparkhi, 1996), (Collins, 1996), . . .
- **Data Loss** due to incompatible structures:
  - Any error in either of the parses and/or the word-alignment prevents treelet pair extraction.
- **Combinatorial Explosion** of factored output:
  - Translation options are first fully built, before combination is attempted.
  - Abundance of t-node attribute combinations
    $\Rightarrow$ e.g. lexically different translation options pushed off the stack
    $\Rightarrow$ n-bestlist varies in unimportant attributes.
- **Deterministic Sentence Generation**:
  - Does not use an n-gram language model.
  - Tuned on manual t-trees, not on automatic trees coming from English.

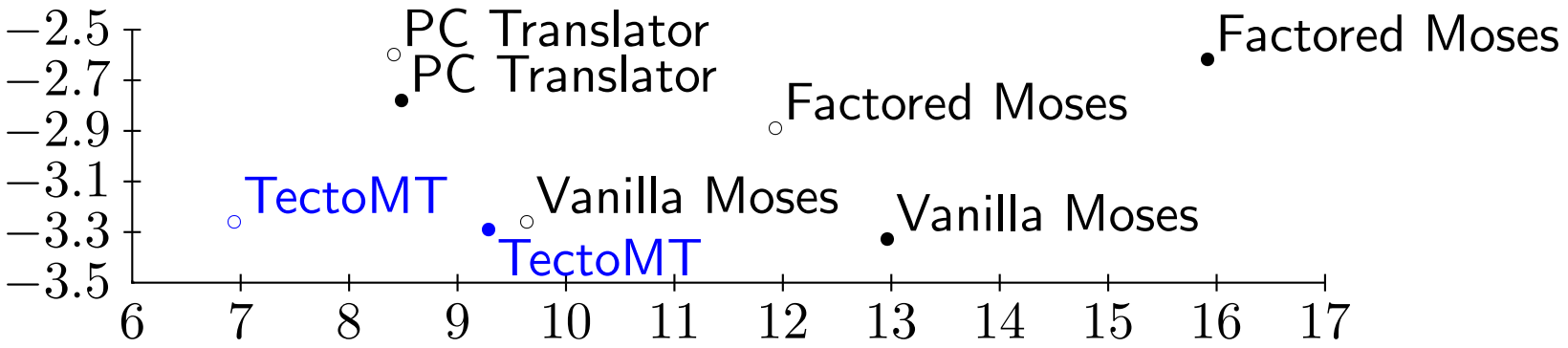
Comparison with Phrase-Based MT

Method	Language Model	BLEU
factored Moses, more data	wordform 4-grams + tag 7-grams, big corpus	15.3±0.9
factored Moses, more data	wordform 3-grams + tag 7-grams	14.2±0.7
basic Moses (no factors)	wordform 3-grams	12.9±0.6
epcp, no factors	3-grams	10.90±0.63
epcp, no factors	none	8.65±0.55
best of etct	binode	5.61±0.50

- Moses beats "epcp" because:
  - "epcp" does not allow any phrase reordering.
  - MERT works well in Moses.
- n-gram LM clearly helpful but even "epcp" without a LM $>$ "etct".

Don't Dump Deep Syntax Yet

TectoMT (Žabokrtský, 2008): mostly deterministic heuristics for t-transfer.



WMT08 Results	In-domain ●		Out-of-domain ○	
	BLEU	Rank	BLEU	Rank
Factored Moses	15.91	-2.62	11.93	-2.89
PC Translator	8.48	-2.78	8.41	-2.60
TectoMT	9.28	-3.29	6.94	-3.26
Vanilla Moses	12.96	-3.33	9.64	-3.26
etct	4.98	-	3.36	-

System Combination

Bleeding edge results due to Teresa Herrmann: (Rosti et al., 2007)

- Outputs of various systems combined to a confusion network.
  One of the systems is always taken as the backbone.
- Moses used to choose the best path according to edge probabilities and LM.

		System Used as Backbone	
	System Alone	Baseline	Word Length
etct	4.92±0.34	10.27±0.50	10.27±0.50
pct	8.63±0.40	10.09±0.49	10.02±0.50
tectomt	9.49±0.45	10.78±0.46	12.22±0.45
moses	15.22±0.61	10.27±0.48	10.28±0.49
mosesBIG	16.45±0.60	10.65±0.47	10.64±0.48
google-2008-05-15	21.21±0.74	10.90±0.51	10.90±0.51

BLEU known to correlate badly with human judgements, cf. the previous slide.

Summary

- Deep syntactic transfer provides a hope for grammatical MT.
  - including a vision of fancy features (co-reference, topic-focus).
- More complicated setup ⇒ many more parameters to tune.
- Linguistic features can easily explode the search space.
- Errors cumulate, conflicts cause data loss (two parsers worse than one).
- Vanilla STSG not usable:
  - Strong assumptions (structural compatibility, atomic labels) ⇒ sparse data.
  - Need to improve back-off methods.
  - Need to carefully explore the search space.

(Never use "clever" methods where copy-paste works.)

References

Ondřej Bojar, Jiří Semecký, Shravan Vasishth, and Ivana Kruijff-Korbayová. 2004. Processing noncanonical word order in Czech. In *Proceedings of Architectures and Mechanisms for Language Processing, AMLaP 2004*, pages 91–91, Université de Provence, September 16-18.

Ondřej Bojar. 2003. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120.

Martin Čmejrek. 2006. *Using Dependency Tree Structure for Czech-English Machine Translation*. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Michael Collins. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

Jason Eisner. 2003. Learning Non-Isomorphic Tree Mappings for Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume*, pages 205–208, Sapporo, July.

Tomáš Holan. 2003. K syntaktické analýze českých(!) vět. In *MIS 2003*. MATFYZPRESS, January 18–25, 2003.

Václav Klimeš. 2006. *Analytical and Tectogrammatical Analysis of a Natural Language*. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proc. of EMNLP*.

Geert-Jan M. Kruijff. 2003. 3-Phase Grammar Learning. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT/EMNLP 2005*, October.

References

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.

Jarmila Panevová. 1980. *Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech sentence]*. Academia, Prague, Czech Republic.

Jan Ptáček and Zdeněk Žabokrtský. 2006. Synthesis of Czech Sentences from Tectogrammatical Trees. In *Proc. of TSD*, pages 221–228.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania, May.

Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007. Combining Outputs from Multiple Machine Translation Systems. In *HLT-NAACL*, pages 228–235.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

David A. Smith and Jason Eisner. 2006. Minimum-Risk Annealing for Training Log-Linear Models. In *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL), Companion Volume*, pages 787–794, Sydney, July.

Kateřina Veselá, Jiří Havelka, and Eva Hajičová. 2004. Condition of Projectivity in the Underlying Dependency Structures. In *Proceedings of Coling 2004*, pages 289–295, Geneva, Switzerland, August. COLING.

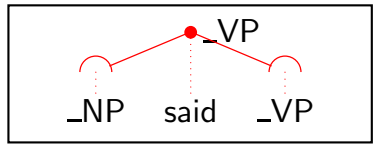
References

Zdeněk Žabokrtský. 2008. TectoMT: Highly Modular Hybrid MT System
with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, page In print, Columbus, Ohio, USA.

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.

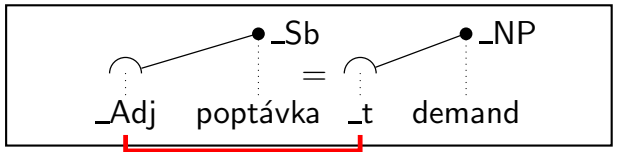
Optionally, we can keep track of local or global ordering of nodes in treelets.

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

Treelet Pair Formally, Synchronous Derivation

A TREELET PAIR $t_{1:2}$ is a tuple (t_1, t_2, m) where:

- t_1 and t_2 are little trees for source and target languages (L_1 and L_2) and states (Q_1 and Q_2),
- m is a 1-1 MAPPING of frontier nodes in t_1 and t_2.



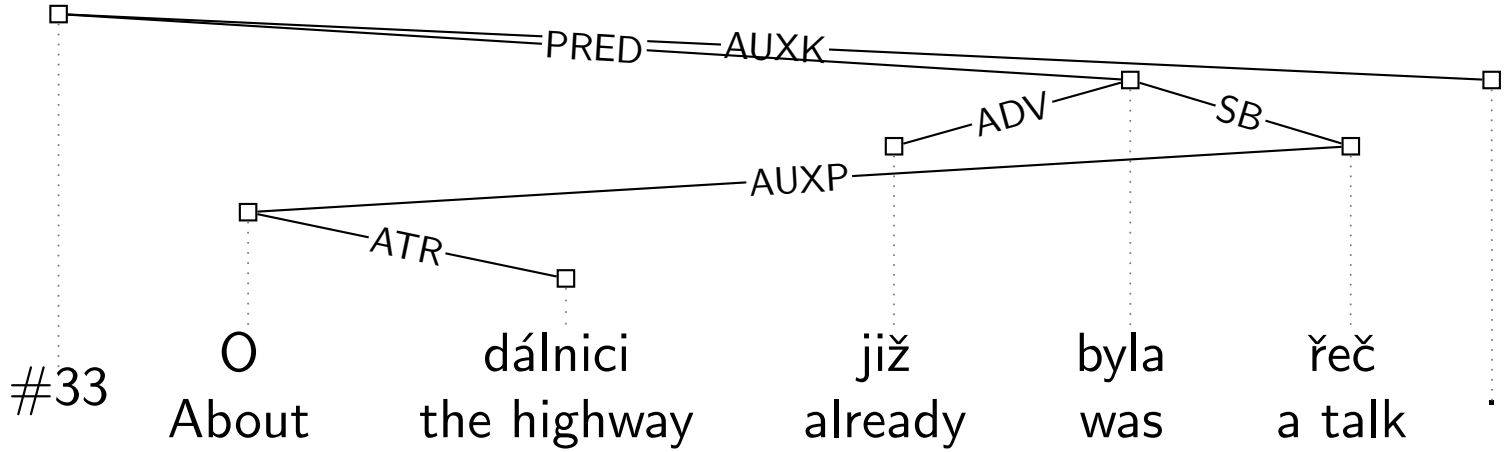
Unlike Čmejrek (2006), I require all frontier nodes mapped, i.e. equal number of left and right frontier nodes.

From a starting SYNCHRONOUS STATE $Start_{1:2} \in Q_1 \times Q_2$,
a SYNCHRONOUS DERIVATION δ constructs a pair of dependency trees by:

- attaching treelet pairs $t^0_{1:2}, \ldots, t^k_{1:2}$ at corresponding frontier nodes, and
- ensuring that the root states $q^0_{1:2}, \ldots, q^k_{1:2}$ of the attached treelets pairs $t^0_{1:2}, \ldots, t^k_{1:2}$ match the frontier states of the corresponding frontier nodes.

Can define probability of a derivation: $p(\delta) = p(t^0_{1:2}|Start_{1:2}) * \prod_{i=1}^{k} p(t^k_{1:2}|q^k_{1:2})$

Nonprojectivity



Non-projectivity:

- does not seem to cause delays in reading experiments (Bojar et al., 2004)
- disappears at the deep syntactic level (Veselá et al., 2004)
- parsing ($O(n^2)$) solved only recently (McDonald et al., 2005)