EatTalk: Syntax and Rich Morphology in MT

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

Outline

- Syntax is more than bracketing:
  - Dependency vs. constituency trees.
  - Non-projectivity and why it matters.
- Rich morphology.
  - Vocabulary sizes, OOV.
  - Factored and Two-step attempts in PBT.
  - Impact on MT evaluation.
- What we call deep syntax.
  - Motivation for deep syntax.
  - Tectogrammatical layer, TectoMT.
- Summary.

Constituency vs. Dependency

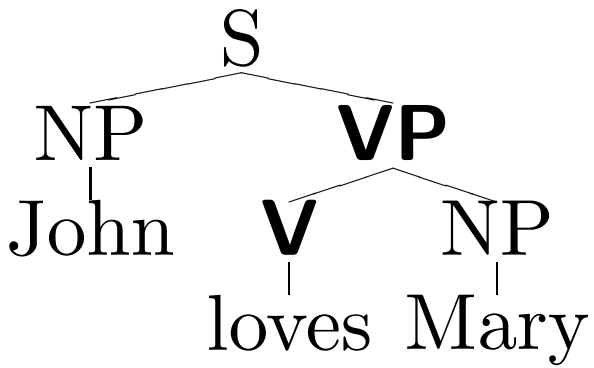
Constituency trees (CFG) represent only bracketing:
= which adjacent constituents are glued tighter to each other.

Dependency trees represent which words depend on which.
+ usually, some agreement/conditioning happens along the edge.

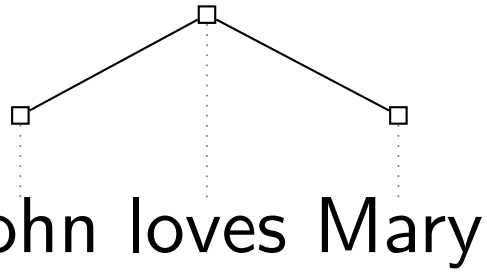
Constituency

John (loves Mary)

John $_{\mathrm{VP}}$(loves Mary)



Dependency



What Dependency Trees Tell Us

Input: The **grass** around your house should be **cut** soon.

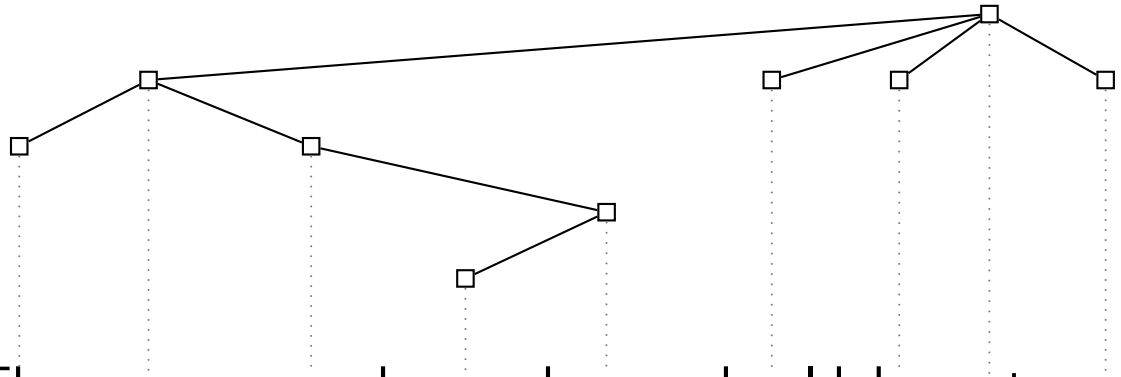
Google: **Trávu** kolem vašeho domu by se měl **snížit** brzy.

- Bad lexical choice for $cut = sekat/snížit/krájet/řezat/\ldots$
  - Due to long-distance lexical dependency with $grass$.
  - One can "pump" many words in between.
  - Could be handled by full source-context (e.g. maxent) model.
- Bad case of $tráva$.
  - Depends on the chosen active/passive form:

active⇒accusative	passive⇒nominative
trávu . . . by**ste** ~~se~~ měl posekat	tráva . . . by **se** měl**a** posekat
	tráva . . . by měl**a být** posek**ána**

Examples by Zdeněk Žabokrtský, Karel Oliva and others.

Tree vs. Linear Context

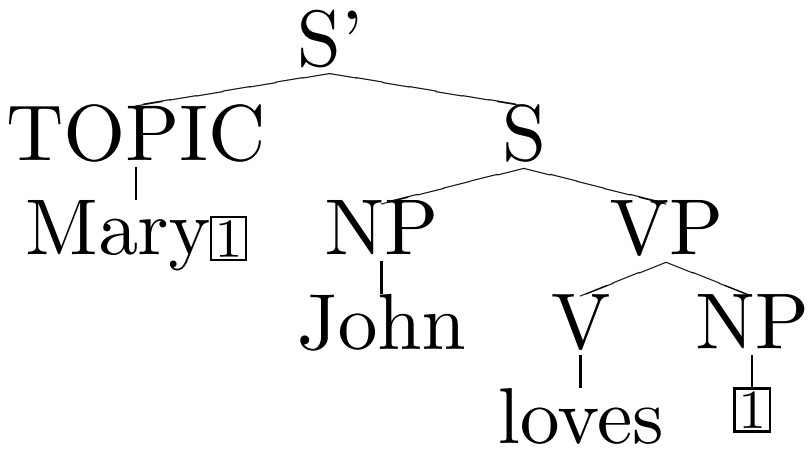


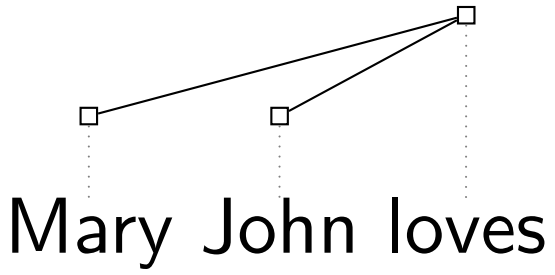
- Tree context (neighbours in the dependency tree):
  - is better at predicting lexical choice than n-grams.
  - often equals linear context:
    Czech manual trees: 50% of edges link neighbours,
    80% of edges fit in a 4-gram.
- Phrase-based MT is a very good approximation.
- Hierarchical MT can even capture the dependency in one phrase:

$$X \rightarrow <\text{ the grass } X \text{ should be cut, trávu } X \text{ byste měl posekat } >$$

"Crossing Brackets"

- Constituent outside its father's span causes "crossing brackets."
  - Linguists use "traces" (1) to represent this.
- Sometimes, this is not visible in the dependency tree:
  - There is no "history of bracketing".
  - See Holan et al. (1998) for dependency trees including derivation history.



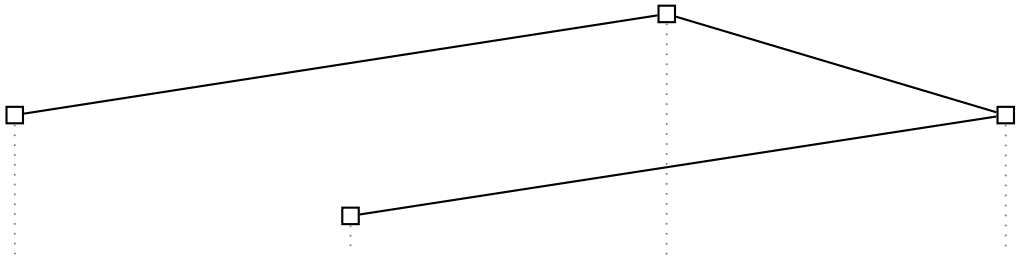


Despite this shortcoming, CFGs are popular and "the" formal grammar for many. Possibly due to the charm of the father of linguistics, or due to the abundance of dependency formalisms with no clear winner (Nivre, 2005).

Non-Projectivity

= a gap in a subtree span, filled by a node higher in the tree.

Ex. Dutch "cross-serial" dependencies, a non-projective tree with one gap caused by saw within the span of $swim$.



. . . dat	Jan	kinderen	zag	zwemmen
. . . that	John	children	saw	swim

. . . that John saw children swim.

- 0 gaps $\Rightarrow$ projective tree $\Rightarrow$ can be represented in a CFG.
- ≤ 1 gap & "well-nested" $\Rightarrow$ mildly context sentitive (TAG).

See Kuhlmann and Möhl (2007) and Holan et al. (1998).

Why Non-Projectivity Matters?

- CFGs cannot handle non-projective constructions:

  Imagine John **grass** saw **being cut**!

- No way to glue these crossing dependencies together:
  - Lexical choice:

    $$X \rightarrow <\text{grass } X \text{ cut}, \text{trávu } X \text{ sekat} >$$

  - Agreement in gender:

    $$X \rightarrow <\text{John } X \text{ saw}, \text{Jan } X \text{ viděl} >$$

    $$X \rightarrow <\text{Mary } X \text{ saw}, \text{Marie } X \text{ viděl}\mathbf{a} >$$

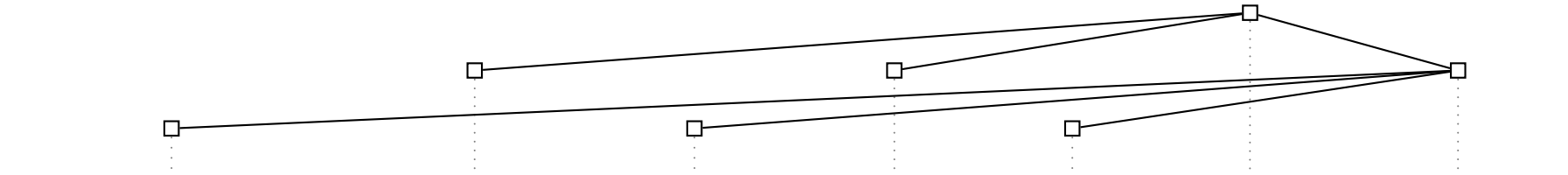
- Phrasal chunks can memorize fixed sequences containing:
  - the non-projective construction
  - and all the words in between! ($\Rightarrow$ extreme sparseness)

Is Non-Projectivity Severe?

Depends on the language.

In principle:

- Czech allows long gaps as well as many gaps in a subtree.



In treebank data:

⊖ 23% of Czech sentences contain a non-projectivity.

⊕ 99.5% of Czech sentences are well nested with ≤ 1 gap.

Parallel View

Ignoring formal linguistic grammar, do we have to reorder beyond swapping constituents (ITG/Hiero with ≤ 2 nonterminals)?

		English-Czech Parallel Sents	
Domain	Alignment	Total	Beyond ITG
WSJ	manual Sure	515	2.9%
WSJ	manual S+P	515	15.9%
News	GIZA++, gdfa	126k	10.6%
Mixed	GIZA++, gdfa	6.1M	3.5%

- searched for (discontinuous) 4-tuples of alignment points in the forbidden shapes (3142 and 2413).
- additional alignment links were allowed to intervene (and could force different segmentation to phrases) $\Rightarrow$ we overestimate.
- no larger sequences of tokens were considered as a unit $\Rightarrow$ we underestimate.

Don't Care Approach (cs→en)

Input: Zítra **se** v kostele Sv. Trojice budou **brát** Marie a Honza.
Google: Tomorrow **is** the Holy Trinity church will **take** Mary and John.

- Bad lexical choice:
  $brát = take$ vs. $brát\ se = get\ married$
- Superfluous is:
  - se is very often mis-aligned with the auxiliary is.

The straightforward bag-of-source-words model would fail here:

- se is very frequent and it often means just $with$.
- An informed model would use the source parse tree.
  - Remember to use a non-projective parser!

Complementary Issue: Morphology

News Commentary Corpus (2007)	Czech	English
Sentences	55,676	
Tokens	1.1M	1.2M
Vocabulary (word forms)	91k	40k
Vocabulary (lemmas)	34k	28k

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

Czech tagging and lemmatization: Hajič and Hladká (1998)
English tagging (Ratnaparkhi, 1996) and lemmatization (Minnen et al., 2001).

OOV Rates

Dataset (# Sents)	Language	n-grams Out of: Corpus Voc. 1	Corpus Voc. 2	Phrase-Table Voc. 1	Phrase-Table Voc. 2
7.5M	Czech	2.2%	30.5%	3.9%	44.1%
	English	1.5%	13.7%	2.1%	22.4%
	Czech + English input sent	1.5%	29.4%	3.1%	42.8%
126k	Czech	6.7%	48.1%	12.5%	65.4%
	English	3.6%	28.1%	6.3%	45.4%
	Czech + English input sent	5.2%	46.6%	10.6%	63.7%
126k	Czech lemmas	4.1%	36.3%	5.8%	52.6%
	English lemmas	3.4%	24.6%	6.9%	53.2%
	Czech + English input sent lemmas	3.1%	35.7%	5.1%	38.1%

- OOV of Czech forms ˜twice as bad as in English.
- OOV of Czech lemmas lower than in English.
- Significant vocabulary in extraction.

WMT 2010 test set; more details in Bojar and Kos (2010).

Morphological Explosion in Czech

MT to Czech has to choose the word including its form:

- Czech nouns and adjectives: 7 cases, 4 genders, 3 numbers, . . .
- Czech verbs: gender, number, aspect (im/perfective), . . .

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	kočky	.
	pily	dvě	zelená	pruhovaná	koček	
	. . .	dvou	zelené	pruhované	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
	viděl jsem		zelenými	pruhovanými		
	viděla jsem		. . .	. . .		

Margin for improvement: Standard BLEU ∼12% vs. lemmatized BLEU ∼21%

Factored Attempts (WMT09)

Data	System	BLEU	NIST	Sent/min
2.2M	Vanilla	**14.24**	**5.175**	12.0
2.2M	T+C	13.86	5.110	2.6
84k	T+C+C&T+T+G	10.01	4.360	4.0
84k	Vanilla MERT	10.52	4.506	–
84k	Vanilla even weights	08.01	3.911	–

T+C = form→form (i.e. vanilla), generate tag, use extra tag LM
T+C+C = form→form, generate lemma and tag, use extra lemma LM and tag LM
T+T+G = lemma→lemma, tag→tag, generate form

- T+T+G explodes the search space
  - too many translation options ⇒ stacks overflown
    ⇒ important options pruned before LM context can pick them

Two-Step Attempts (WMT10) 1/2

1. English → lemmatized Czech
   - meaning-bearing morphology preserved
   - max phrase len 10, distortion limit 6
   - large target-side (lemmatized LM)
2. Lemmatized Czech → Czech
   - max phrase len 1, monotone

Src	after a sharp drop		
Mid	po+6	ASA1.prudký	NSA-.pokles
Gloss	after+voc	adj+sg...sharp	noun+sg...drop
Out	po	prudkém	poklesu

- Only 1-best output passed, will try lattice.

Two-Step Attempts (WMT10) 2/2

Data Size		Simple		Two-Step		Diff
Parallel	Mono	BLEU	SemPOS	BLEU	SemPOS	B.S.
126k	126k	10.28±0.40	29.92	10.38±0.38	30.01	↗ ↗
126k	13M	12.50±0.44	31.01	12.29±0.47	31.40	↘ ↗
7.5M	13M	14.17±0.51	33.07	14.06±0.49	32.57	↘ ↘

Manual micro-evaluation of ↘↗, i.e. 12.50±0.44 vs. 12.29±0.47:

	Two--Step	Both Fine	Both Wrong	Simple	Total
Two-Step	**23**	4	8	-	**35**
Both Fine	7	14	17	5	43
Both Wrong	8	1	28	2	39
Simple	-	3	7	**23**	33
Total	**38**	22	60	30	150

- Each annotator weakly prefers Two-step
  - but they don't agree on individual sentences.

Two-Step Has Words to Offer

Analyzing 52889 tokens in the Czech reference of WMT10:

- # tokens produced by cu-bojar-primary?
- # tokens among translation options of cu-bojar-primary?
- # tokens in two-step single-best output only?

	In Primary we Consider	
	1-Best Hyp	Tr. Opts
In Both	41.8 %	45.5 %
Nowhere	44.8 %	**17.7 %**
Primary Only	8.1 %	35.1 %
Two-step Only	**5.4 %**	1.7 %

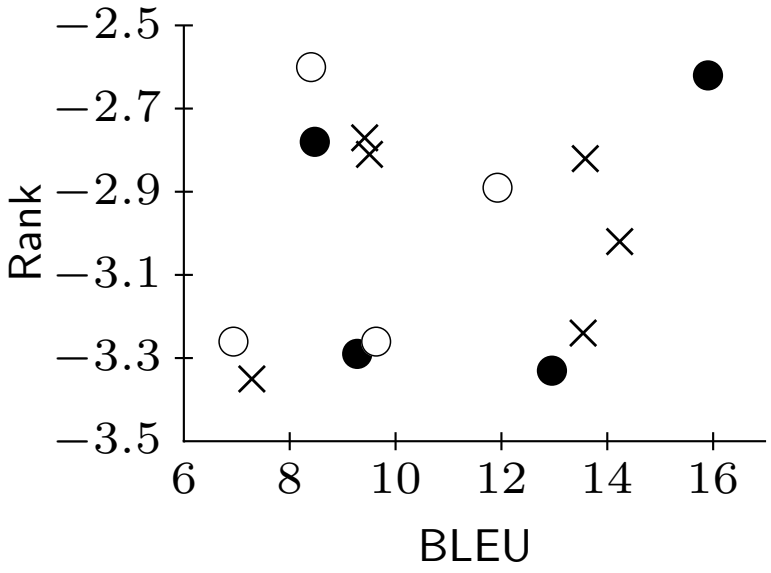
- ~50% of ref toks not produced by Primary.
- ~20% of ref toks not available among Primary tropts.
- ~2–5% of ref toks only in Two-Step 1-Best.

BLEU vs. Human Rank

- Large vocabulary impedes the performance of BLEU.

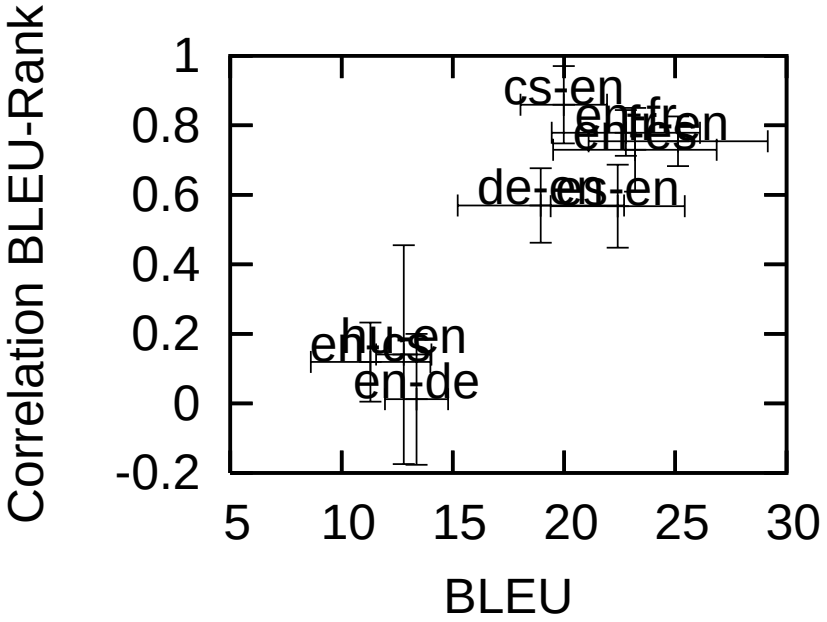
En→Cs Systems

WMT08, WMT09



Various Language Pairs

WMT08, WMT09, MetricsMATR



⇒ BLEU does not correlate with human rank if below ˜20.

Reason 1: Focus on Forms

SRC	Prague Stock Market falls to minus by the end of the trading day
REF	pražská burza se ke konci obchodování propadla do minusu
cu-bojar	praha stock market klesne k minus na konci obchodního dne
pctrans	praha trh cenných papírů padá minus do konce obchodního dne

- Only a single unigram in each hyp. confirmed by the reference.
- Large chunks of hypotheses are not compared at all.

Confirmed by Reference	Yes	Yes	No	No
Contains Errors	Yes	No	Yes	No
Running words	6.34%	36.93%	22.33%	**34.40%**

Reason 2: Sequences Overvalued

BLEU overly sensitive to sequences:

- Gives credit for 1, 3, 5 and 8 four-, three-, bi- and unigrams,
- Two of three serious errors not noticed,

  $\Rightarrow$ Quality of cu-bojar overestimated.

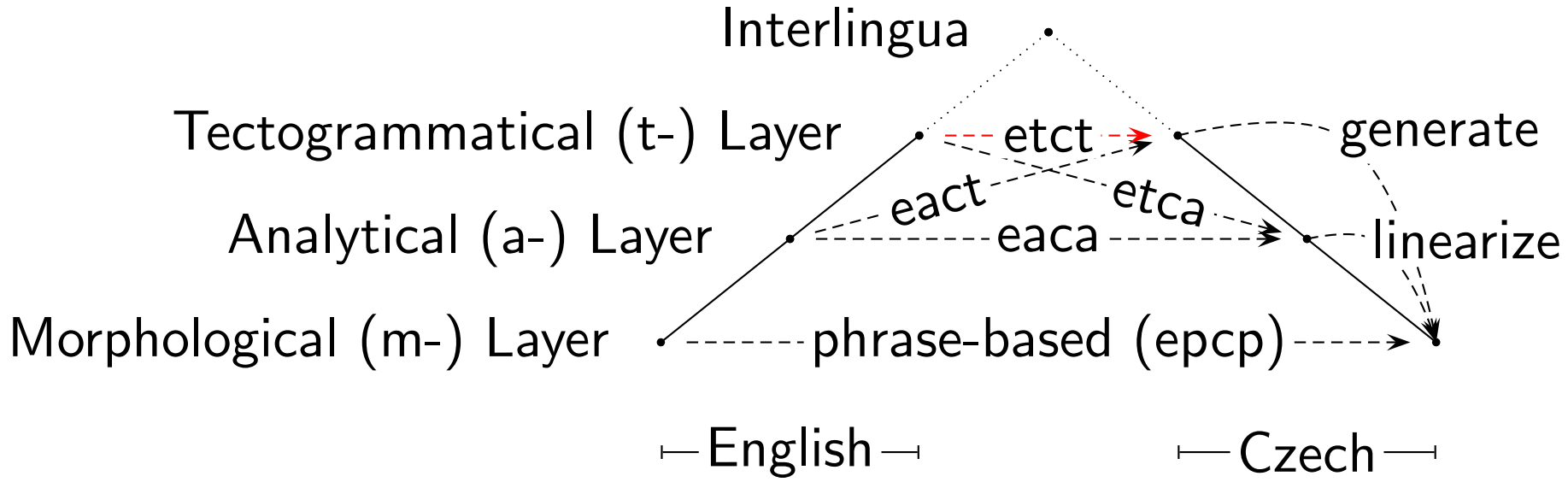
SRC	Congress yields: US government can pump 700 billion dollars into banks
REF	kongres ustoupil : vláda usa může do bank napumpovat 700 miliard dolarů
cu-bojar	kongres výnosy : vláda usa může čerpadlo 700 miliard dolarů v bankách
pctrans	kongres vynáší : us vláda může čerpat 700 miliardu dolarů do bank

More details in Bojar et al. (2010).

Motivation for Deep Syntax

Let's introduce (an) intermediate language(s) that handle:

- auxiliary words,
- morphological richness,
- non-projectivity,
- ~~meanings of words~~.



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)
- English: PEDT 1.0 (2009); Arabic: PADT (2004)

Practice — Tools:

- parsing Czech to a-layer: McDonald et al. (2005)
- parsing Czech to t-layer: Klimeš (2006)
- parsing English to a-layer: well studied (+rules convert to dependency trees)
- parsing English to t-layer: heuristic rules (manual annotation in progress)
- generating Czech surface from t-layer: Ptáček and Žabokrtský (2006)
- **all-in-one TectoMT platform**: Žabokrtský and Bojar (2008)

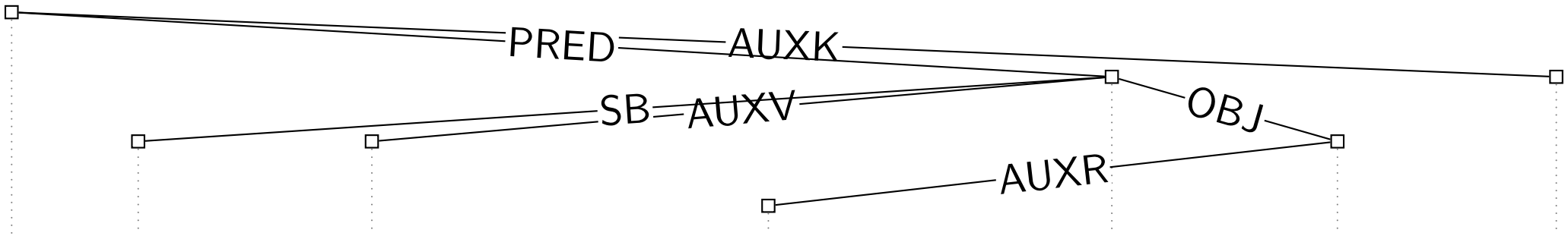
TectoMT Platform

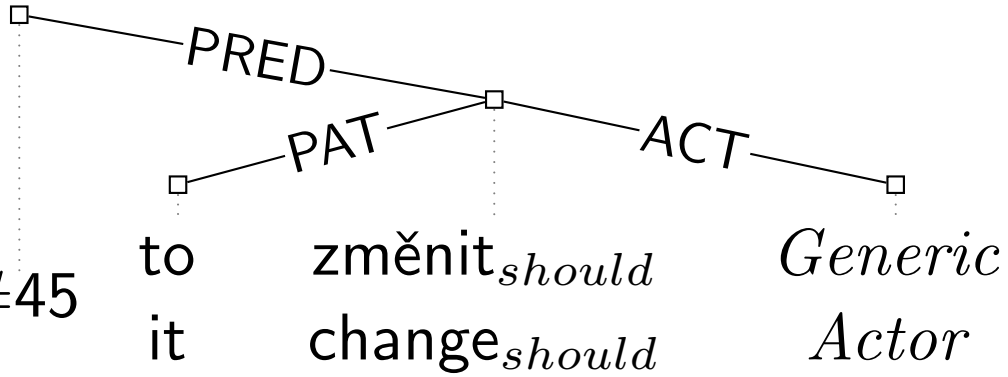
- TectoMT is not just an MT system.
- TectoMT is a highly modular environment for NLP tasks:
  - Provides a unified rich file format and (Perl) API.
  - Wraps many tools: taggers, parsers, deep parsers, NERs, . . .
  - Sun Grid Engine integration for large datasets:

    e.g. CzEng (Bojar and Žabokrtský, 2009), 8.0M parallel sents. at t-layer.
- Implemented applications:
  - MT, preprocessing for other MT systems (SVO→SOV in 12 lines of code),
  - dialogue system, corpus annotation, paraphrasing, . . .
- Languages covered: Czech, English, German; and going generic

http://ufal.mff.cuni.cz/tectomt/

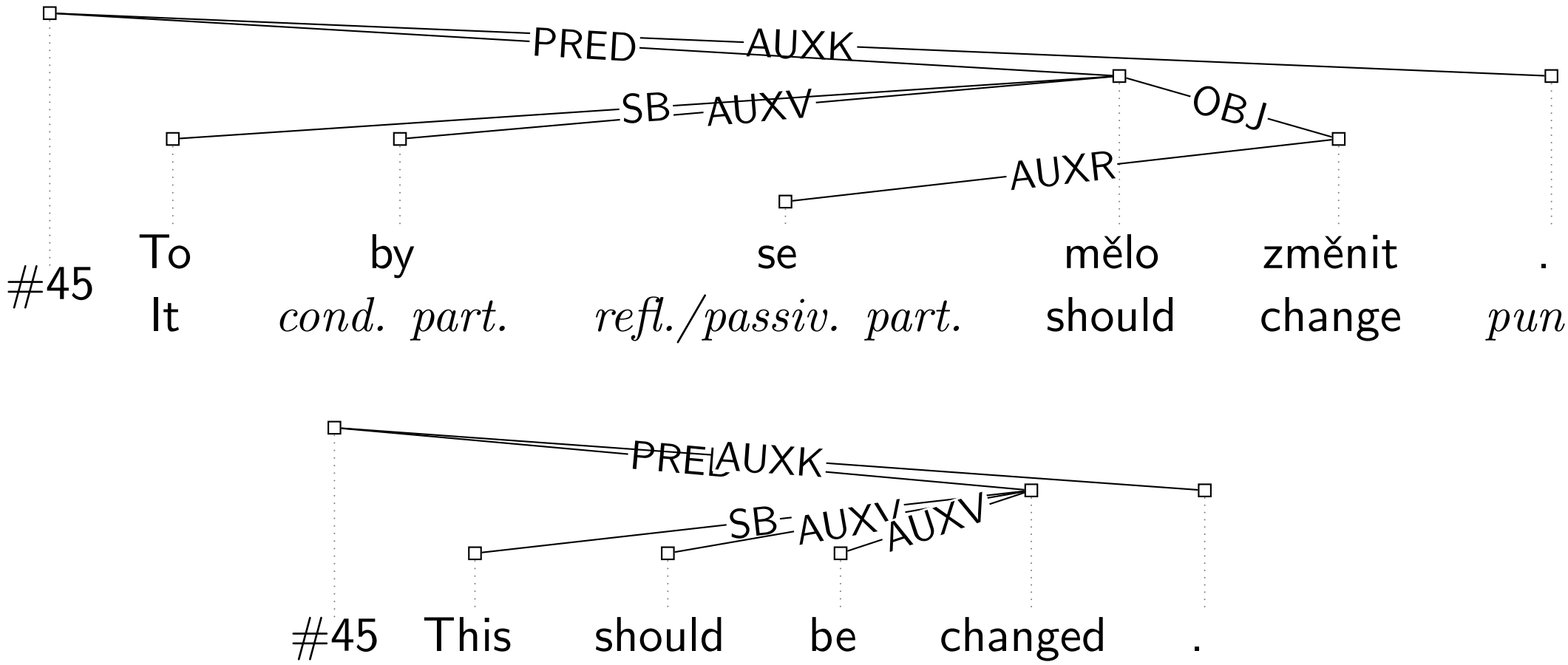
Analytical vs. Tectogrammatical



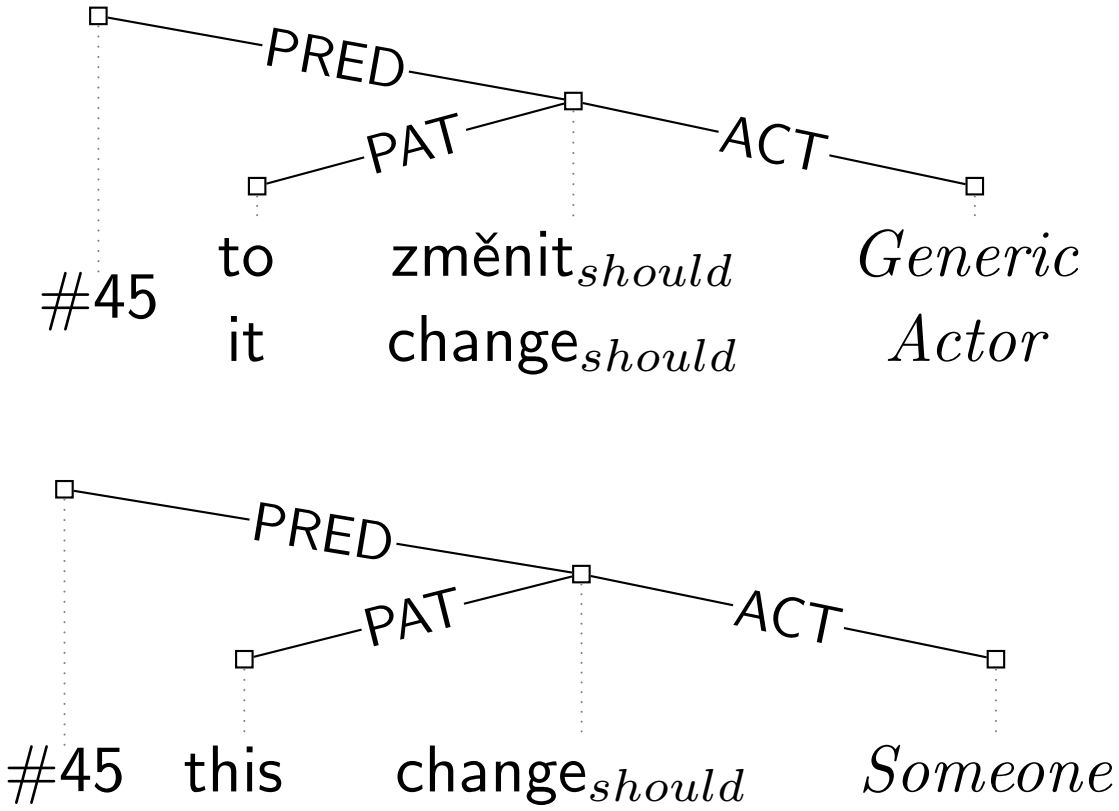


- hide auxiliary words, add nodes for "deleted" participants
- resolve e.g. active/passive voice, analytical verbs etc.
- "full" tecto resolves much more, e.g. topic-focus articulation or anaphora

Czech and English A-Layer



Czech and English T-Layer



Represents predicate-argument structure:

`change`$_{should}$`(ACT: someone, PAT: it)`

The Tectogrammatical Hope

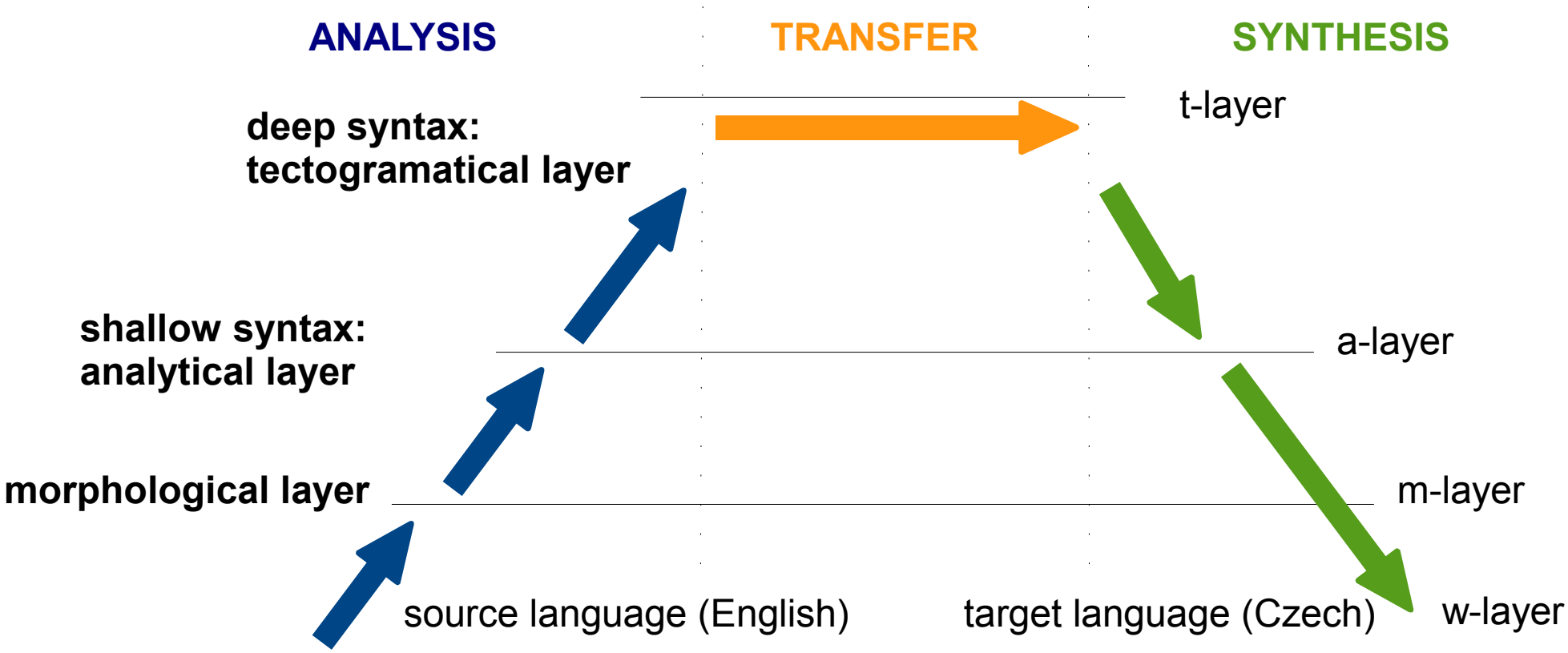
Transfer at t-layer should be easier than direct translation:

- Reduced vocabulary size (Czech morphological complexity).
- Reduced structure size (auxiliary words disappear).
- Word order ignored / interpreted as information structure (given/new).
  ⇒ Non-projectivities resolved at t-layer.
- Tree context used instead of linear context.
- Czech and English t-trees structurally more similar
  ⇒ Less parallel data might be sufficient (but more monolingual).
- Ready for fancy t-layer features: co-reference.

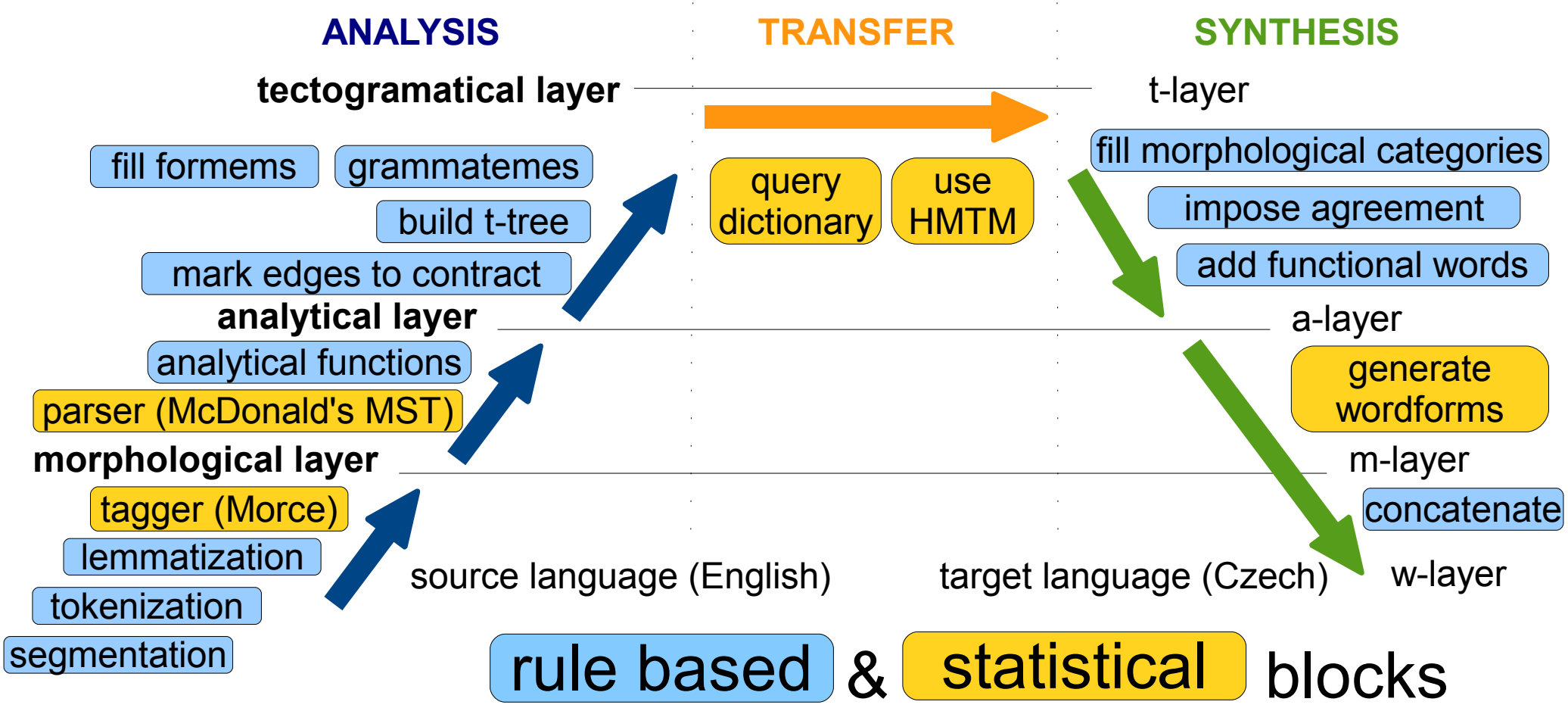
Anyone welcome to try!

`http://ufal.mff.cuni.cz/czeng/` **= 8.0M parallel sents at t-layer**

“TectoMT Transfer” (1/2)



“TectoMT Transfer” (2/2)



WMT10 Evaluation

	REF	CU-BOJAR	CU-TECTO	EUROTRANS	ONLINEB	PC-TRANS	UEDIN
REF	-	4.3	4.3	5.1	3.8	3.6	2.3
CU-BOJAR	**87.1**	-	**45.7**	28.3	**44.4**	39.5	**41.1**
CU-TECTO	**88.2**	35.8	-	38.0	**55.8**	**44.0**	36.0
EUROTRANS	**88.5**	**60.9**	**46.8**	-	**50.7**	**53.8**	**48.6**
ONLINEB	**91.2**	31.1	29.1	32.8	-	43.8	**39.3**
PC-TRANS	**88.0**	**45.3**	42.9	28.6	**49.3**	-	36.6
UEDIN	**94.3**	39.3	**44.2**	31.9	32.1	**49.5**	-
$>$ others	90.5	45.0	44.1	39.3	49.1	**49.4**	39.6
$>=$ others	95.9	65.6	60.1	54.0	**70.4**	62.1	62.2
Official rank	-	2	5	6	**1**	4	3
# pairwise wins	6	2	3	0	**4**	3	3
BLEU		.16	.13	.10	**.17**	.10	.16
TER	-	**74.5**	76.9	81.9	74.6	82.4	75.2

- TectoMT 5th, between two traditional commercial systems.
- Pairwise comparisons more favourable (beated the 2nd and the 3rd system).

TectoMT Has Words to Offer

Analyzing 52889 tokens in the Czech reference of WMT10:

	In Primary we Consider	
	1-Best Hyp	Tr. Opts
In Both	39.3 %	45.6 %
Nowhere	41.8 %	**17.4 %**
Primary Only	10.6 %	35.0 %
TectoMT Only	**8.4 %**	2.0 %

- ˜2–8% of ref toks only in TectoMT.
- Primary and TectoMT less similar than Primary and Two-Step.
  - Here, 10.6% of toks exclusively by Primary,
  - On slide 17, 8.1% exclusively from Primary.
- Still ˜17% of ref toks not available at all.

Summary

- There is some **dependency syntax**.
  - Dependency reveals, well, dependencies between words.
  - Non-projective constructions cannot be handled by CFGs.
- **Morphological richness** is a challenge for MT.
  - Factored setup explodes the search space.
  - Two-step setup not convincing but promising.
  - BLEU correlates worse.
- "**Deep syntax**":
  - Aims at solving morphological richness, non-projectivity, . . .
  - T-layer is an example; (parallel) treebanks and tools ready.
  - No win thus far, but clearly different type of errors.
  - TectoMT as a platform for NLP (pre-)processing.

. . . so I am here to combine the outputs.

References

Ondrej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng 0.9: Large Parallel Treebank with Rich Annotation. Prague Bulletin of Mathematical Linguistics, 92:63–83.

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling Sparse Data Issue in Machine Translation Evaluation. In Proceedings of the ACL 2010 Conference Short Papers, pages 86–91, Uppsala, Sweden, July. Association for Computational Linguistics.

Jan Hajič and Barbora Hladká. 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In Proceedings of COLING-ACL Conference, pages 483–490, Montreal, Canada.

Tomáš Holan, Vladislav Kuboň, Karel Oliva, and Martin Plátek. 1998. Two Useful Measures of Word Order Complexity. In A. Polguere and S. Kahane, editors, Proceedings of the Coling '98 Workshop: Processing of Dependency-Based Grammars, Montreal. University of Montreal.

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

Marco Kuhlmann and Mathias Möhl. 2007. Mildly context-sensitive dependency languages. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 160–167, Prague, Czech Republic, June. Association for Computational Linguistics.

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.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English.

References

Natural Language Engineering, 7(3):207–223.

Joakim Nivre. 2005. Dependency Grammar and Dependency Parsing. Technical Report MSI report 05133, Växjö University: School of Mathematics and Systems Engineering.

Jarmila Panevová. 1980. Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech s 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.

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

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

Zdeněk Žabokrtský and Ondřej Bojar. 2008. TectoMT, Developer's Guide. Technical Report TR-2008-39, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, December.