Machine Translation via Deep Syntax

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.
- Delving deeper.
  - Motivation for deep syntax.
  - Approaches (being) tested in Prague.
  - New pitfalls.
- TectoMT, the platform.
- Summary.

Constituency vs. Dependency

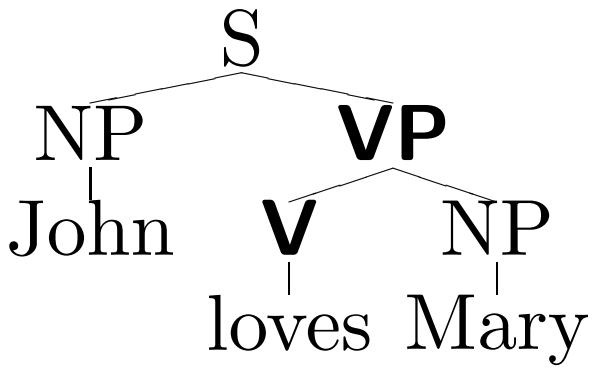
Constituency trees (CFG) represent only bracketing:
= which adjacent constituents are glued tighter to each other.

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

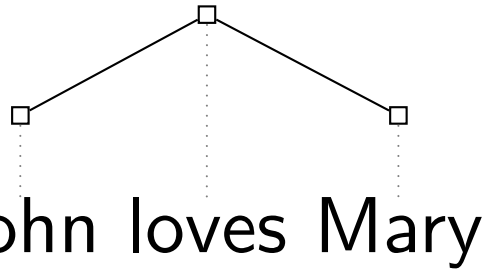
Constituency

John (loves Mary)

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



Dependency



What Dependency Trees Tell Us

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

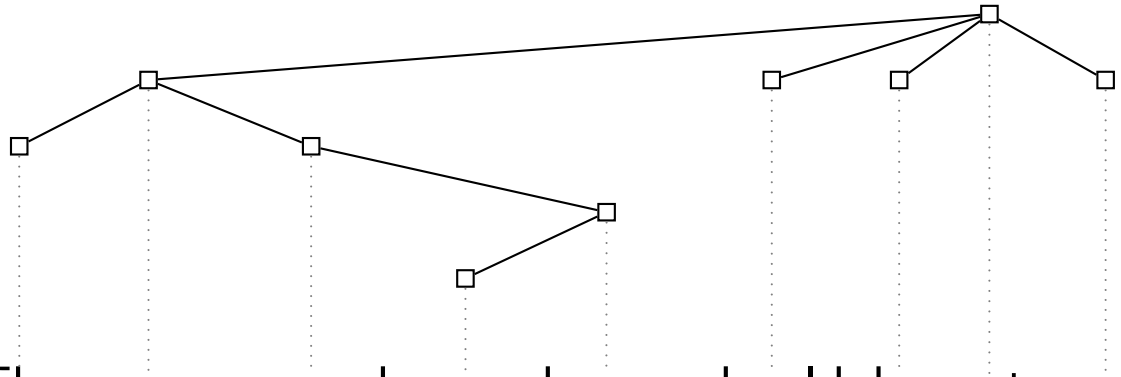
Google: **Trávu** kolem vašeho domu by se měl **snížit** brzy.

- Bad lexical choice for $cut = sekat/snížit/krájet/řezat/\ldots$
  - Due to long-distance 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

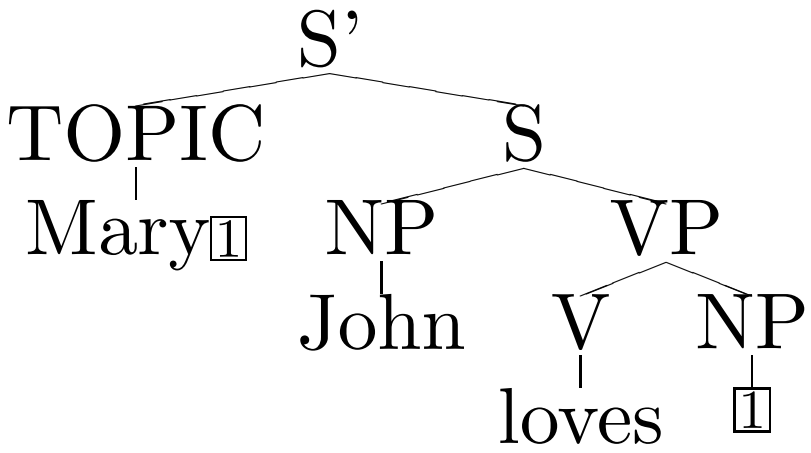


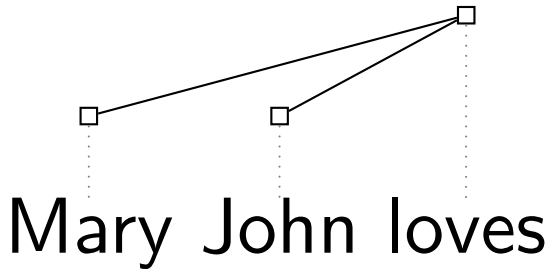
- Tree context (neighbours in the dependency tree):
  - is better at predicting lexical choice than n-grams.
  - often equals linear context:
    Czech manual trees: 50% of edges link neighbours,
    80% of edges fit in a 4-gram.
- Phrase-based MT is a very good approximation.
- Hierarchical MT can even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



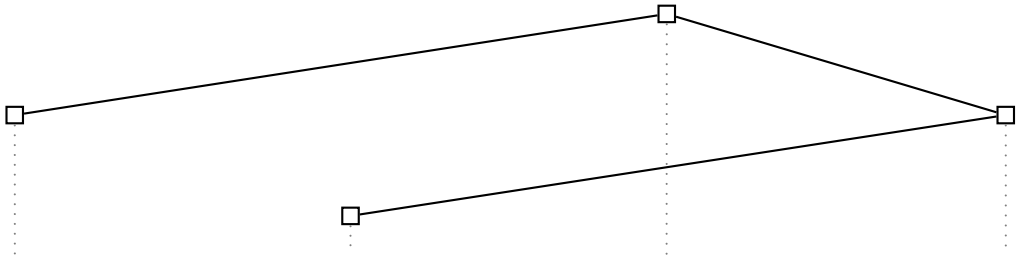


Despite this shortcoming, CFGs are popular and “the” formal grammar for many. Possibly due to the charm of the father of linguistics, or due to the abundance of dependency formalisms with no clear winner (Nivre, 2005).

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 **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:

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

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

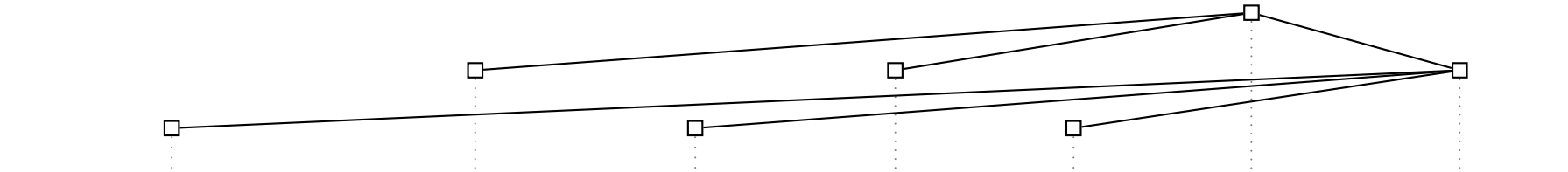
- Phrasal chunks can memorize fixed sequences containing:
  - the non-projective construction
  - and all the words in between! ($\Rightarrow$ extreme sparseness)

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.

This is a corrected and extended version of the slide I originally presented.

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!

Another 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).

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), . . .

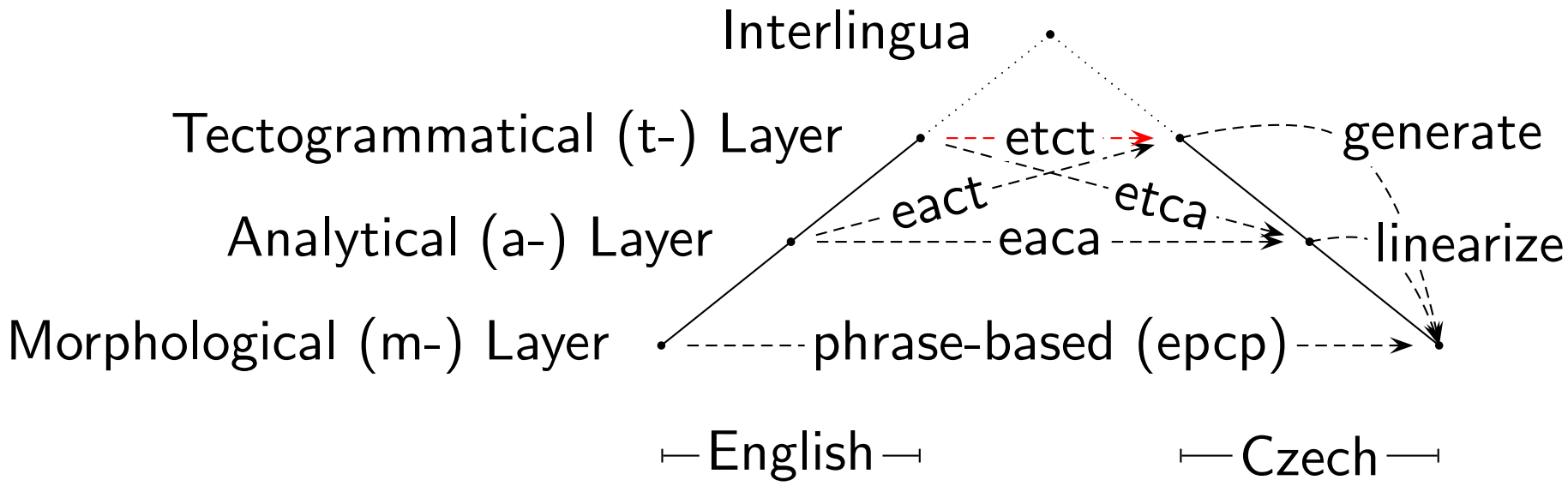
I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	kočky	.
	pily	dvě	zelená	pruhovaná	koček	
	. . .	dvou	zelené	pruhované	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným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

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

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).

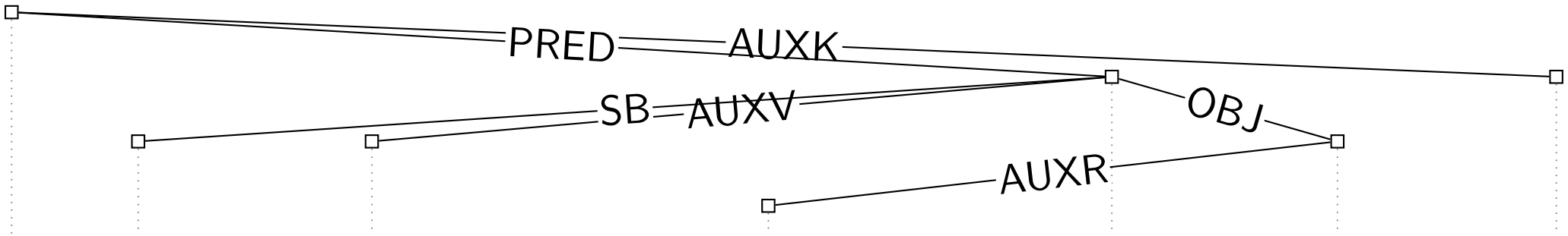
Materialized theory — Treebanks:

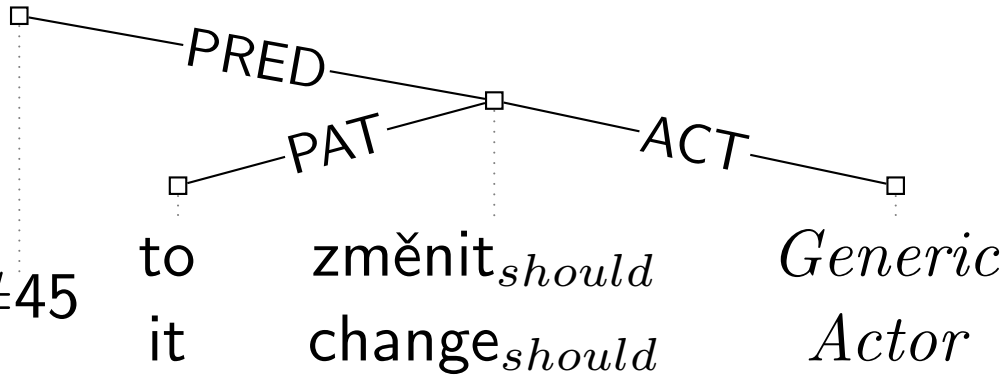
- Czech: PDT 1.0 (2001), PDT 2.0 (2006)
- Czech-English: PCEDT 1.0 (2004), PCEDT 2.0 (in progress)
- English: PEDT 1.0 (2009); Arabic: PADT (2004)

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)

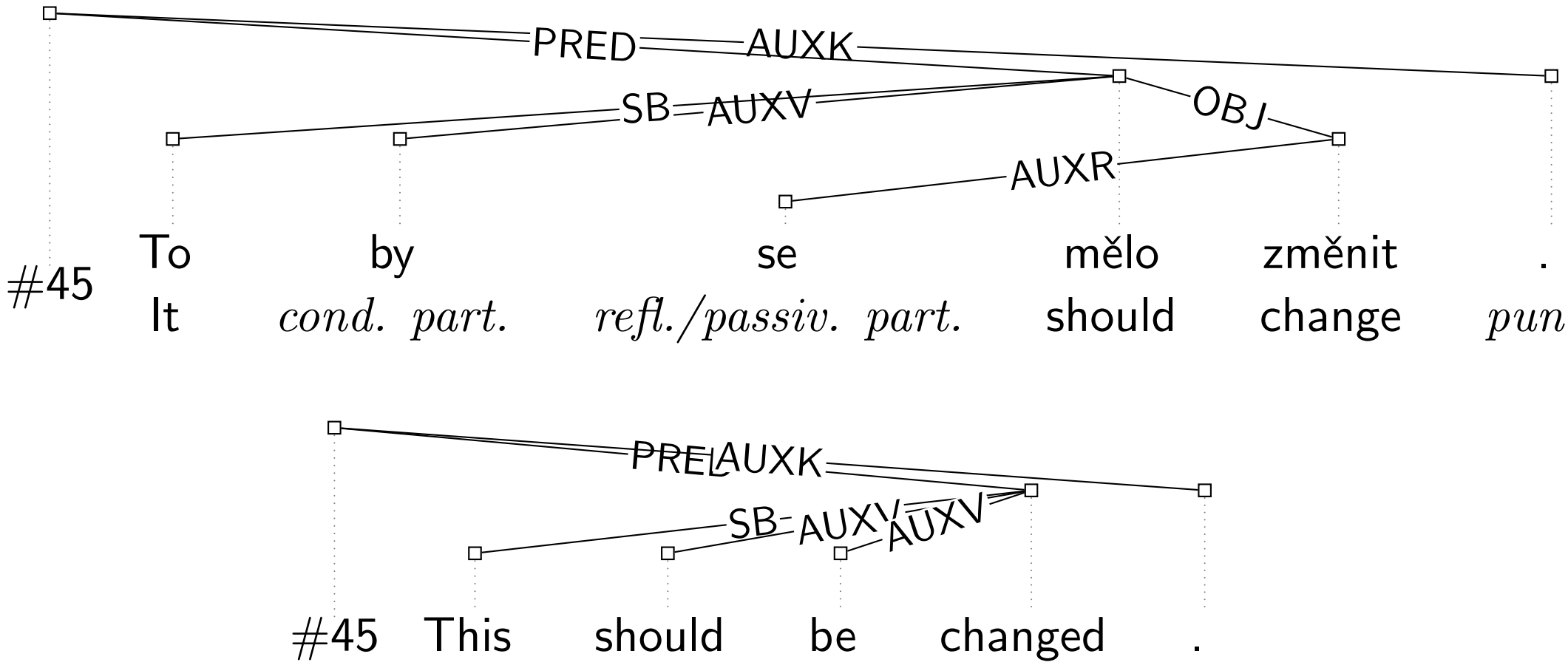
Analytical vs. Tectogrammatical



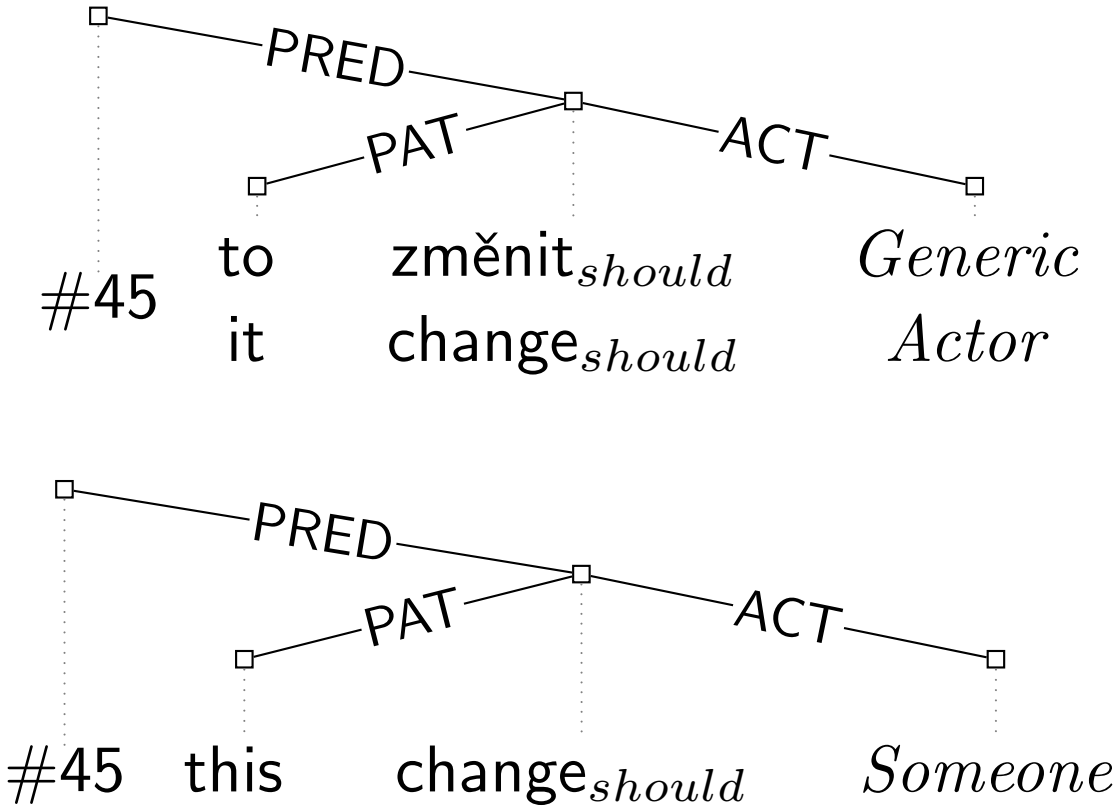


- hide auxiliary words, add nodes for "deleted" participants
- resolve e.g. active/passive voice, analytical verbs etc.
- "full" tecto resolves much more, e.g. topic-focus articulation or anaphora

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).
  $\Rightarrow$ Non-projectivities resolved at t-layer.
- Tree context used instead of linear context.
- 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.

Implementations of Deep MT

In Prague, using t-layer:

- TectoMT (Žabokrtský et al., 2008)
  - preserves t-tree structure
  - a maxent model to score choices of node and edge labels
  - a Viterbi-like alg. to pick the best combination of labels
- TreeDecode (Bojar et al., 2008)
  - based on Synchronous Tree Substitution Grammars
  - top-down stack-based decoder
  - applicable to any pair of dependency trees (a-/t-layer)

Others:

- Sulis (Graham, 2010) – LFG
- Richardson et al. (2001), Bond et al. (2005), Oepen et al. (2007).

WMT09 Scores for English→Czech

System	BLEU	NIST	Rank
Vanilla Moses (Prague)	**14.24**	**5.175**	-3.02 (4)
Google	13.59	4.964	-2.82 (3)
Vanilla Moses (Edinburgh)	13.55	5.039	-3.24 (5)
Clever Moses T+C+C&T+T+G 84k	10.01	4.360	–
Eurotran XP	09.51	4.381	-2.81 (2)
PC Translator	09.42	4.335	**-2.77 (1)**
TectoMT 2009	07.29	4.173	-3.35 (6)
TreeDecode **"phrase-based"** 84k	08.07	3.942	–
TreeDecode **via t-layer** 643k	05.53	3.660	–
TreeDecode **via t-layer** 43k	05.14	3.538	–
Vanilla Moses 84k, even weights	08.01	3.911	–
Vanilla Moses 84k, MERT	10.52	4.506	–

TectoMT 2009 had a very simple transfer, not the maxent+Viterbi.

Pitfalls Hit by TreeDecode

- **Cumulation of Errors**:
  - e.g. 93% tagging * 85% parsing * 93% tagging * 92% parsing = 67%
- **Data Loss** due to incompatible structures:
  - Any error in the parses and/or the word-alignment prevents treelet pair extraction.
- **Data Sparseness** when attributes or treelet structure atomic:
  - E.g. different tense requires a new treelet pair.
  - There is no adjunction in STSG, new modifier needs a new treelet pair.
- **Combinatorial Explosion** when generating attributes dynamically:
  - Target treelets 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 varied in unimportant attributes.
  - "Delaying" some attributes until the full tree is built does not help enough.

Details in project deliverables (http://ufal.mff.cuni.cz/euromatrix/) and lab session.

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/

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.
- "**Deep syntax**":
  - Aims at solving morphological richness, non-projectivity, . . .
  - T-layer is an example; (parallel) treebanks and tools ready.
  - ⊖ No win thus far.
- **TectoMT Platform** is a (great) tool for rich annotation.

Lab session for all the details.

References

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, Miroslav Janíček, and Miroslav Týnovský. 2008. Implementation of Tree Transfer System. Project Euromatrix - Deliverable 3.3, ÚFAL, Charles University, September.

Francis Bond, Stephan Oepen, Melanie Siegel, Ann Copestake, and Dan Flickinger. 2005. Open source machine translation with DELPH-IN. In Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit, pages 15–22, Phuket, Thailand, September.

Yvette Graham. 2010. Sulis: An Open Source Transfer Decoder for Deep Syntactic Statistical Machine Translation. In Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools, number 93 in Prague Bulletin of Mathematical Linguistics. Charles University, January.

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

References

Spanning Tree Algorithms. In Proceedings of HLT/EMNLP 2005, October.

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

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards Hybrid Quality-Oriented Machine Translation. On Linguistics and Probabilities in MT. In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07), Skövde, Sweden.

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.

Stephen D. Richardson, William B. Dolan, Arul Menezes, and Monica Corston-Oliver. 2001. Overcoming the Customization Bottleneck Using Example-Based MT. In Proceedings of the workshop on Data-driven methods in machine translation, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

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,

References

December.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In Proc. of the ACL Workshop on Statistical Machine Translation, pages 167–170, Columbus, Ohio, USA.