More on Syntax in MT

Ondřej Bojar
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

Thu Sep 12, 2013

Outline

- Refresher: Motivation to go beyond phrases.
- Constituency vs. dependency trees.
- Tree vs. linear context.
- Non-projectivity and why it matters in MT.

Refresher: Prove Google is Phrase-Based

Natáhnout bačkory.	Kick the bucket.	√

Refresher: Prove Google is Phrase-Based

Natáhnout bačkory.	Kick the bucket.	√
Proč musel natáhnout bačkory?	Why did he kick the bucket?	√

Refresher: Prove Google is Phrase-Based

Word form variations:

Natáhnout bačkory.	Kick the bucket.	✓
Proč musel natáhnout bačkory?	Why did he kick the bucket?	✓
Proč natáhl bačkory?	Why stretched slippers?	✗

Refresher: Prove Google is Phrase-Based

Word form variations:

Natáhnout bačkory.	Kick the bucket.	✓
Proč musel natáhnout bačkory?	Why did he kick the bucket?	✓
Proč natáhl bačkory?	Why stretched slippers?	✗

Pumping words into phrases:

Jan s Marií se vzali.	John and Mary were married.	✓

Refresher: Prove Google is Phrase-Based

Word form variations:

Natáhnout bačkory.	Kick the bucket.	✓
Proč musel natáhnout bačkory?	Why did he kick the bucket?	✓
Proč natáhl bačkory?	Why stretched slippers?	×

Pumping words into phrases:

Jan s Marií se vzali.	John and Mary were married.	✓
Jan s Marií se včera vzali.	John and Mary married yesterday.	✓

Refresher: Prove Google is Phrase-Based

Word form variations:

Natáhnout bačkory.	Kick the bucket.	✓
Proč musel natáhnout bačkory?	Why did he kick the bucket?	✓
Proč natáhl bačkory?	Why stretched slippers?	×

Pumping words into phrases:

Jan s Marií se vzali.

John and Mary were married. ✓

Jan s Marií se včera vzali.

John and Mary married yesterday. ✓

Jan s Marií se včera v kostele vzali.

John and Mary are married in church yesterday. ∼

Refresher: Prove Google is Phrase-Based

Word form variations:

Natáhnout bačkory.	Kick the bucket.	√
Proč musel natáhnout bačkory?	Why did he kick the bucket?	√
Proč natáhl bačkory?	Why stretched slippers?	×

Pumping words into phrases:

Jan s Marií se vzali.

John and Mary were married. √

Jan s Marií se včera vzali.

John and Mary married yesterday. √

Jan s Marií se včera v kostele vzali.

John and Mary are married in church yesterday. ∼

Jan s Marií se včera v kostele svatého Ducha vzali.

John and Mary yesterday in the Church of the Holy Spirit took. ×

PBMT vs. RBMT

(Prove Systran is not phrase-based.)

PBMT vs. RBMT

(Prove Systran is not phrase-based.)

	Stell dir das vor.	
Google	Imagine that.	✓
Systran	Imagine.	✓

PBMT vs. RBMT

(Prove Systran is not phrase-based.)

	Stell dir das vor.	
Google	Imagine that.	✓
Systran	Imagine.	✓
	Stell dir ein Haus vor.	
Google	Imagine a house before.	×
Systran	Imagine a house.	✓

PBMT vs. RBMT

(Prove Systran is not phrase-based.)

	Stell dir das vor.	
Google	Imagine that.	✓
Systran	Imagine.	✓
	Stell dir ein Haus vor.	
Google	Imagine a house before.	×
Systran	Imagine a house.	✓
	Stell dir ein kleines Haus vor.	
Google	Imagine a small house in front.	×
Systran	Imagine a small house.	✓

PBMT vs. RBMT

(Prove Systran is not phrase-based.)

	Stell dir das vor.	
Google	Imagine that.	✓
Systran	Imagine.	✓
	Stell dir ein Haus vor.	
Google	Imagine a house before.	×
Systran	Imagine a house.	✓
	Stell dir ein kleines Haus vor.	
Google	Imagine a small house in front.	×
Systran	Imagine a small house.	✓
	Stell dir ein kleines Haus mit vierzehn Fenster vor.	
Google	Imagine a small house with fourteen windows in front.	×
Systran	Imagine a small house with fourteen windows.	✓

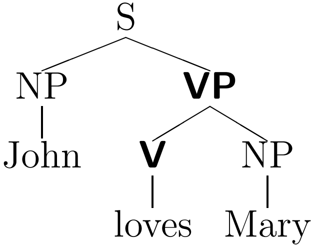
Constituency vs. Dependency

Constituency trees (CFG) represent only bracketing:
= which adjacent constituents are glued together.
Dependency trees represent which words depend on which.
+ usually, some agreement/conditioning along the edge.

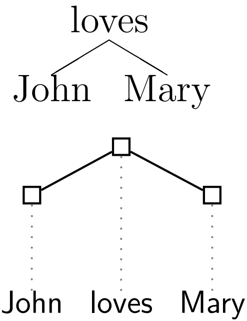
Constituency

John (loves Mary)

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



Dependency



What Dependency Trees Tell Us

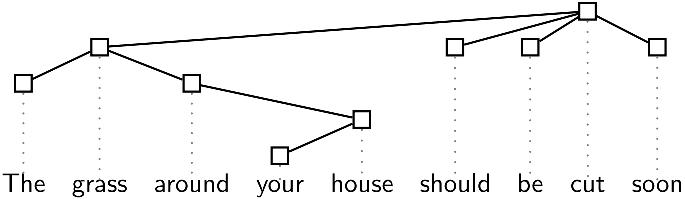
Input: The **grass** around your house should be **cut** soon.
Google: **Trávu** kolem vašeho domu by se měl **snížit** brzy.

- Bad lexical choice for *cut* = *sekat/snížit/krájet/řezat/. .*
  - 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



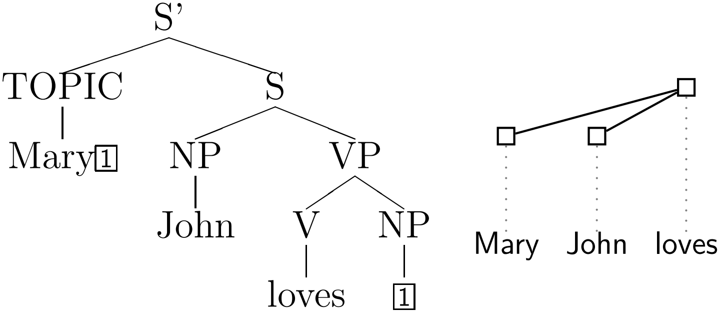
The grass around your house should be cut soon

- 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:

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

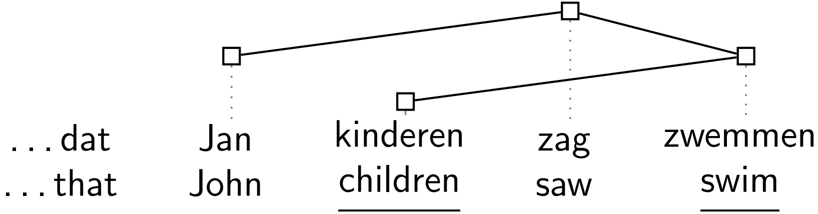
“Crossing Brackets”

- Constituent outside its father’s span causes “crossing brackets.”
  - Linguists use “traces” ($\boxed{1}$) 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.



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 = projective tree ⇒ representable in CFG.
- ≤ 1 gap & “well-nested” ⇒ mildly context sensitive (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:

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

  - Agreement in gender:

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

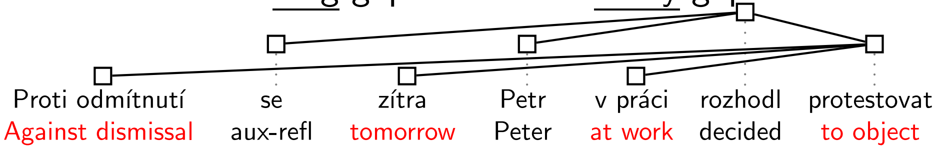
- Phrases can memorize <u>fixed</u> sequences containing:
  - the non-projective construction
  - and all the words in between! (⇒ extreme sparseness)

Is Non-Projectivity Severe?

Depends on the language.

In principle unlimited:

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



Proti odmítnutí	se	zítra	Petr	v práci	rozhodl	protestovat
Against dismissal	aux-refl	tomorrow	Peter	at work	decided	to object

Peter decided to object against the dismissal at work tomorrow.

In treebank data:

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

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

In parallel data:

- ˜3–15% English-Czech sents beyond ITG reordering.

Summary

- Limitations of phrase-based MT:
  - Little or no dependencies across phrases.
  - Practice: dependencies are often local enough.
- Limitations of hierarchical/constituency-based MT:
  - Non-projective constructions are bound to fail.

$\leadsto$ deep-syntactic (dependency) translation as a solution.

References

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.

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.