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

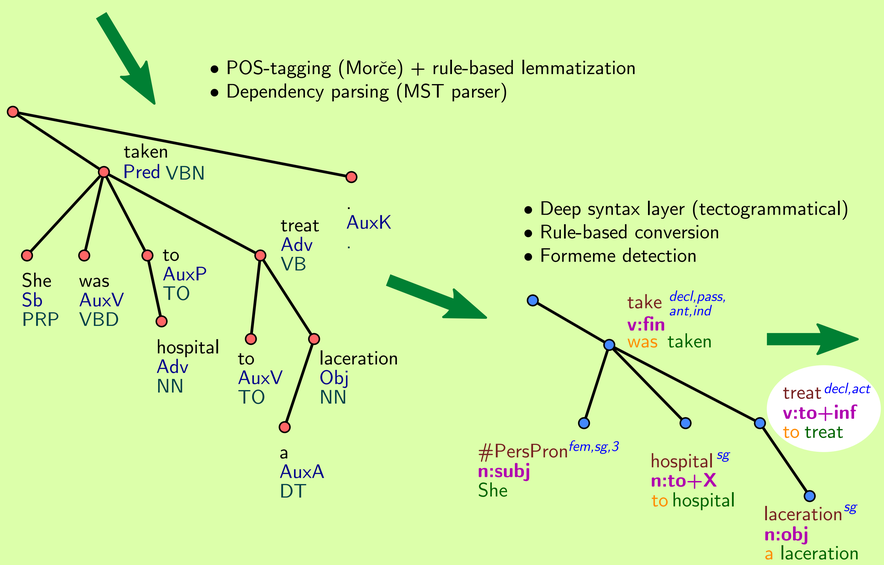
Formemes in English-Czech Deep Syntactic MT

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, David Mareček – {odusek,zabokrtsky,popel,majlis,mnovak,marecek}@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

The TectoMT System

Analysis

She was taken to hospital to treat a laceration.



Transfer

- On the deep (tectogrammatical) layer
- Separated into three virtually independent subtasks

Grammatemes
- Rule-based modules

decl,act + [noun lemma] **n:k+3** = *sg*

Lemma
- Maximum-Entropy model: *P(cs-lemma|en-context-features)*
- Simple model: *P(cs-lemma|en-lemma)*
- Derivation back-off models

treat →

léčba	-3.05
považovat	-3.45
léčit	-3.75
léčení	-3.77
zacházet	-3.91

Target-Language Tree Model
- Viterbi algorithm adapted for trees
- Selects best lemma+formeme

léčba **n:k+3**

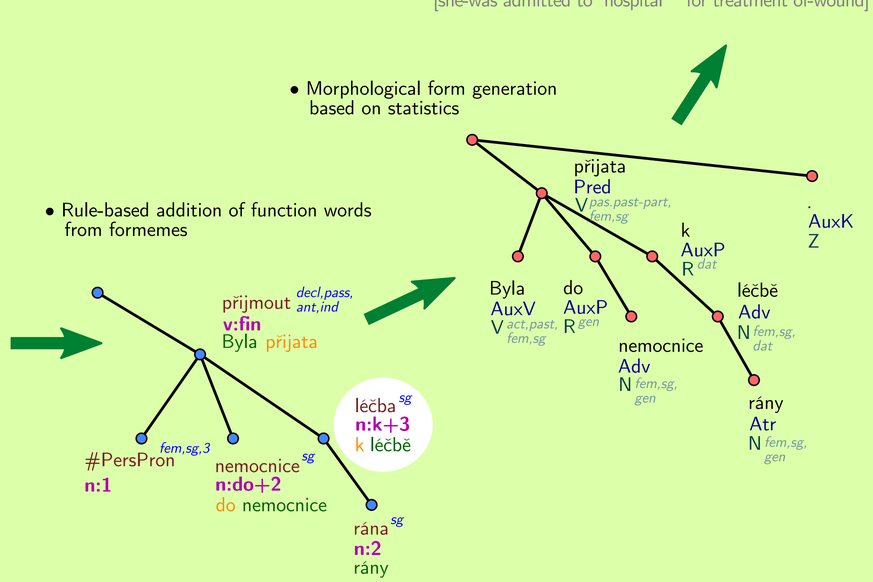
Formeme
- Maximum-Entropy model: *P(cs-formeme|en-context-features)*
- Simple model: *P(cs-formeme|en-formeme)*
- Valency reranker: *P(cs-formeme|en-formeme,en-parent-lemma)*

take →

v:to+inf →	**v:inf**	-1.48 → -2.80
	v:aby+fin	-2.72 → -2.32
	n:k+3	-2.85 → -2.01
	n:pro+4	-3.61 → -3.57
	v:že+fin	-3.83 → -5.84

Synthesis

Byla přijata do nemocnice k léčbě rány.
[she-was admitted to hospital for treatment of-wound]



What is a formeme

v:to+inf

- Syntactic part-of-speech (verb, noun, adjective, adverb)
- Preposition or subordinate conjunction
- Morpho-syntactic form

- Description of morpho-syntactic features in deep syntax trees:
  - Syntactic position and usage
  - Morphological case
  - Prepositions and subordinate conjunctions
- Simplifies transfer (no need for semantic labels)
- Enables a straightforward transition from deep syntax to the surface
- Reduces data sparsity (non-redundant information)
- Rule-based annotation modules for English and Czech
- Implemented within the **Treex NLP Framework**

Examples

So	*far,*	*everything*	*was*	*just*	*right.*				
So	far	everything	be	just	right				
adv	**adv**	**n:subj**	**v:fin**	**x**	**adj:compl**				

The tram	*was forced*	*to break*	*abruptly*	*on Herzbergallee*	*when a vehicle*	*stopped*	*suddenly*	*in front of it.*		
tram	force	break	abruptly	Herzbergallee	vehicle	stop	suddenly	#PersPron		
n:subj	**v:fin**	**v:to+inf**	**adv**	**n:on+X**	**n:subj**	**v:when+fin**	**adv**	**n:in_front_of+X**		

	Thank	*you*	*for staying*	*with us*	*through all*	*these*	*years.*	
#PersPron	thank	#PersPron	stay	#PersPron	all	this	year	
drop	**v:fin**	**n:obj**	**v:for+ger**	**n:with+X**	**adj:attr**	**adj:attr**	**n:through+X**	

I	*had*	*no*	*idea*	*what*	*was coming.*		
#PersPron	have	no	idea	what	come		
n:subj	**v:fin**	**n:attr**	**n:obj**	**n:subj**	**v:rc**		

K jeho schválení	*bylo zapotřebí*	*226*	*hlasů.*					
[for its	approval	was	needed	226	votes]			
#PersPron	schválení	být	zapotřebí	226	hlas			
adj:poss	**n:k+3**	**v:fin**	**adv**	**adj:attr**	**n:1**			

Skoro	*každý*	*– bez ohledu na*	*hmotnost –*	*pije*	*každý*	*den*	*sladké*	*limonády*	*(70*	*procent).*
[almost	everyone	– regardless	of	weight	– drinks	every	day	sweet	lemonades	(70 | percent)]
skoro	každý	hmotnost		pít	každý	den	sladký	limonáda	70	procento
adv	**n:1**	**n:bez_ohledu_na+4**		**v:fin**	**adj:attr**	**n:4**	**adj:attr**	**n:4**	**adj:attr**	**n:1**

Formeme Improvements

- Reducing redundancy
- Reducing data sparsity
- Increasing inter-language consistency
- Maintaining linguistic adequacy

He is one ***of the best*** *at school.* adj:of+X → **n:of+X**

Many *of them were late.* adj:compl → **n:subj**

Koupil pět ***banánů.*** (acc, gen) [He-bought five bananas] n:2 → **n:4**

five *potatoes* n:attr → **adj:attr**

Apollo ***11*** **n:attr**

Evaluation

Inter-language mutual information

Version	MI
Original formemes	1.5981
Revised formemes	1.6873

Czech-to-Czech round trip

Version	BLEU
Original formemes	0.6818
Revised formemes	0.7092

TectoMT translation of the WMT'12 test set (trained on 1/2 of CzEng 1.0)

Version	BLEU
Original formemes	0.1190
Revised formemes	0.1199

TectoMT Performance

WMT'12 manual ranking constrained system winner

WMT'12 Human Evaluation

	System	C?	>others
1.	CU-Depfix •	N	0.66
2.	Online B	N	0.63
3.-4.	U. Edinburgh ⋆	Y	0.56
3.-4.	CU-Tamchyna	N	0.56
5.	CU-Bojar ⋆	Y	0.54
6.-7.	**CU-TectoMT ⋆**	**Y**	**0.53**
6.-7.	Online A	N	0.53
8.	Commercial 1	N	0.48
9.	Commercial 2	N	0.46

(• = Overall winner, ⋆ = Constrained winner)

WMT'12 Automatic Evaluation (BLEU)

	System	BLEU
1.	CU-Depfix	0.163
2.	Online B	0.162
3.	U. Edinburgh	0.155
4.	CU-Bojar	0.142
5.	CU-Tamchyna	0.140
9.	**CU-TectoMT**	**0.120**

(Systems marked as "primary" listed here)

This research has been supported by the grants FP7-ICT-2009-4-247762 (FAUST), FP7-ICT-2009-4-249119 (Metanet), LH12093 (Kontakt II), DF12P01OVV022 (NAKI), 201/09/H057 (Czech Science Foundation), GAUK 116310, and SVV 265 314.This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

Presented at WMT'12, June 7 – 8, Montreal, Canada.