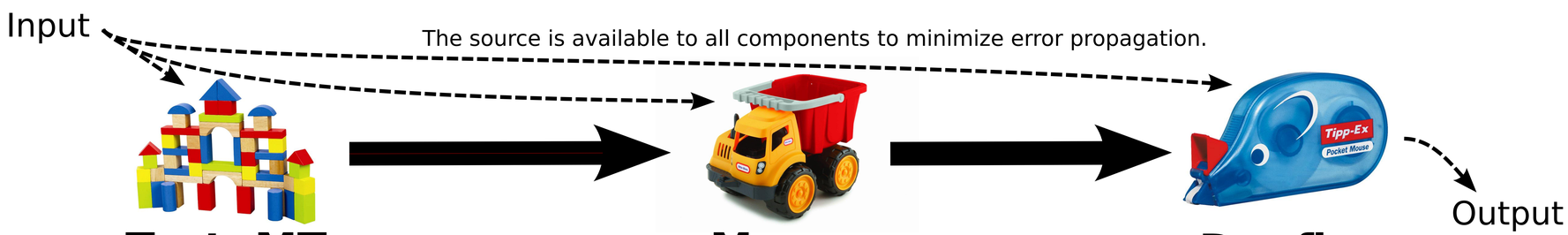
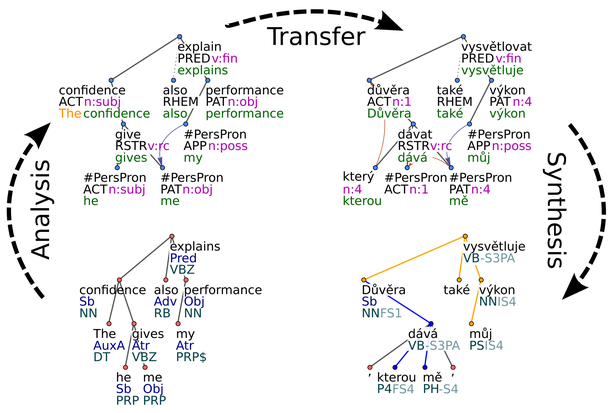
Chimera – Three Heads for English-to-Czech Translation

Ondřej Bojar, Rudolf Rosa, Aleš Tamchyna {bojar,rosa,tamchyna}@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague



TectoMT

Syntax-Based Translation improves **Sentence Structure** and reduces **OOV on Both Sides**



Moses

Main Search adds **0.2 GWord Parallel** **3.6 GWord Czech**

Parallel Data:

Corpus	Sents [M]	Tokens [M] English	Tokens [M] Czech
CzEng 1.0	14.83	235.67	205.17
Europarl	0.65	17.61	15.00
Common Crawl	0.16	4.08	3.63
+ testset translated by TectoMT and added as a separate phrase table	0.003	0.07	0.06

Three Separate Language Models:

Token	Order	Sents [M]	Tokens [M]	ARPA.gz [GB]	Trie [GB]
wordforms	4	201.31	3430.92	28.2	11.8
wordforms	7	24.91	444.84	13.1	8.1
tags	10	14.83	205.17	7.2	3.0

big *long* *morphological*

Depfix

Rule-Based Error Correction restores lost negation: **We're FC Barcelona!** *not*

Source: We're not FC Barcelona!
CU-Bojar: My **jsme** FC Barcelona!
*We **are** FC Barcelona!*
CU-Depfix: **Nejsme** FC Barcelona!
*We'**re not** FC Barcelona!*

✔ Depfix restores lost negation.

... and fixes other flaws:

Source: The biggest risk was for Tereshkova.
CU-Bojar: Největší riziko **je** pro Těreškovovou.
*The biggest risk **is** for Tereshkova.*
CU-Depfix: Největší riziko **bylo** pro Těreškovovou.
*The biggest risk **was** for Tereshkova.*

✔ Depfix provides the correct tense.

Depfix component	Precision
Subject - predicate agr.	68%
Pro-drop in subject	73%
Subject - past participle agr.	75%
Passive - aux 'be' agr.	77%
Possessive with 'of'	78%
Present continuous	78%
Missing reflexive verbs	80%
Subject categories projection	83%
Rehang children of aux verbs	83%
Lost negation recovery	90%

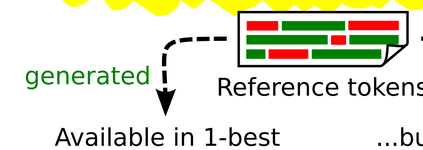
Source: Arizona was the first to introduce such a requirement.
Plain Moses: Arizona byla nejprve na zavedení takového požadavku.
Arizona was at first on introducing such a requirement.
TectoMT: Arizona byla první, zavede takový požadavek.
Arizona was the first, it will introduce such a requirement.
✔ TectoMT introduces a separate clause.
CU-Bojar: Arizona byla první, **kdo** zavedl takový požadavek.
Arizona was the first who introduced such a requirement.
✔ CU-Bojar improves on that.

Source: The main anti-Soviet war leaders returned to power in 2001.
Plain Moses: Hlavní **protisovětské** války vůdci vrátili k moci v roce 2001.
The main anti-Soviet$_{fem}$ war leaders returned to power in year 2001.
TectoMT: Hlavní **protisovětští** váleční vůdci se vrátili k moci v roce 2001.
The main anti-Soviet$_{masc}$ war leaders returned to power in year 2001.
✔ TectoMT provides the correct word form, never seen in the training data.

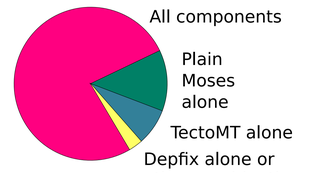
Our Chimera: Beat Google

System	BLEU	TER	WMT Ranking Appraise	WMT Ranking MTurk
CU-TECTOMT	14.7	0.741	0.455	0.491
CU-BOJAR	20.1	0.696	0.637	0.555
CU-DEPFIX	20.0	0.693	0.664	0.542
PLAIN Moses	19.5	0.713	–	–
GOOGLE TR.	–	–	0.618	0.526

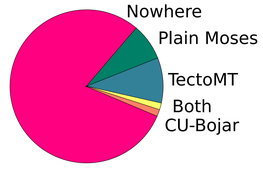
Credits & Opportunities



Available in 1-best output of:



...but could have been taken from:



Sample Errors

TectoMT suffers from errors in analysis and often mistranslates multiword expressions or idioms:

Source: ...turning a blind eye.
TectoMT: ...obrátí slepé oko.
✘ Literal translation.

Moses makes its many usual errors.

DepFix sometimes applies rules in inappropriate situations:

Source: But we're waiting in the sidelines.
CU-Bojar: Ale čekáme v ústraní$_{sg}$. ✔
CU-Depfix: Ale čekáme v ústraních$_{pl}$.
✘ Depfix too eager at recovering plural.

Presented at WMT 2013, Sofia, Bulgaria.

This work was partially supported by the grants P406/11/1499 of the Grant Agency of the Czech Republic,FP7-ICT-2011-7-288487 (MosesCore) and FP7-ICT-2010-6-257528 (Khresmoi) of the European Union and by SVV project number 267314.