Rich Morphology and Hybrid Approaches to MT

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

Thu Nov 18, 2011

Outline

- MT pipeline and problems caused by rich morphology.
- Targeting Czech.
  - Source of the morphological explosion.
  - OOV rates.
- Caveat: Evaluating at Low BLEU.
- Individual Systems:
  - Tuning towards SemPOS.
  - Two-Step Translation.
  - TectoMT.
- Hybridization:
  - Rover System Combination.
  - DEPFIX Grammatical Post-Processing.
- Summary.

(Phrase-Based) MT Pipeline

- Word Alignment.
- Extraction of Translation Units.
- Translation of New Text.
- Reordering.
- Language Modelling.
- MT Evaluation.
- Model Optimization.

(Phrase-Based) MT Pipeline

- Word Alignment.
- Extraction of Translation Units.
- Translation of New Text.
  - New forms of known words.
  - Unknown words.
- Reordering.
- Language Modelling.
  - Sparser unigrams and higher-grams (reordering).
- MT Evaluation.
  - Fewer matches with the reference.
- Model Optimization.

... rich morphology makes everything harder.

(Phrase-Based) MT Pipeline

- Word Alignment.
- Extraction of Translation Units.
- Translation of New Text. ← Chance for RBMT.
  - New forms of known words.
  - Unknown words.
- Reordering. ← Chance for RBMT.
- Language Modelling.
  - Sparser unigrams and higher-grams (reordering).
- MT Evaluation.
  - Fewer matches with the reference.
- Model Optimization.

... rich morphology makes everything harder.

(Phrase-Based) MT Pipeline

- Word Alignment. ⇒ (Lemmatize, chop or LEAF.)
- Extraction of Translation Units.
- Translation of New Text.
  - New forms of known words. ⇒ Two-Step Translation.
  - Unknown words. ⇒ TectoMT.
- Reordering.
- Language Modelling.
  - Sparser unigrams and higher-grams (reordering).
- MT Evaluation. ⇒ SemPOS.
  - Fewer matches with the reference.
- Model Optimization. ⇒ SemPOS+BLEU.

... rich morphology makes everything harder.

Morphological Explosion in Czech

(In)flective lang.: suffix encodes many categories:

- Czech nouns and adjs: 7 cases, 4 genders, 3 nums, ...
- Czech verbs: gender, num, 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			...	...		

Out-of-Vocabulary 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 lemmas	3.1%	35.7%	5.1%	38.1%

Out-of-Vocabulary 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 lemmas	3.1%	35.7%	5.1%	38.1%

- Significant vocabulary loss during phrase extraction:
  - e.g. 2.2%→3.9% for 7.5M Czech.

Out-of-Vocabulary 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 lemmas	3.1%	35.7%	5.1%	38.1%

- Significant vocabulary loss during phrase extraction:
  - e.g. 2.2%→3.9% for 7.5M Czech.
- OOV of Czech forms ∼twice as bad as in English.

Out-of-Vocabulary 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 lemmas	3.1%	35.7%	5.1%	38.1%

- Significant vocabulary loss during phrase extraction:
  - e.g. 2.2%→3.9% for 7.5M Czech.
- OOV of Czech forms ∼twice as bad as in English.
- OOV of Czech lemmas lower than in English.

Out-of-Vocabulary Rates

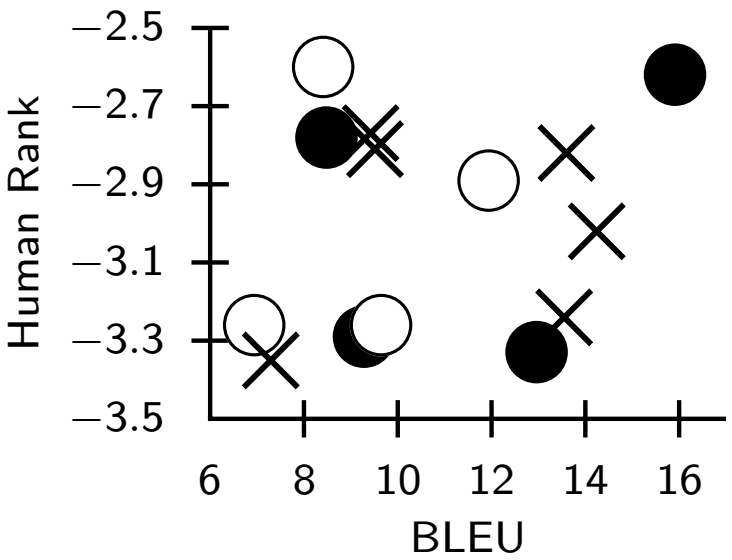
Dataset (# Sents)	Language	*n*-grams Out of: Corpus Voc. 1	Corpus Voc. 2	Phrase-Table Voc. 1	Phrase-Table Voc. 2
7.5M	Czech	2.2%	30.5%	3.9%	44.1%
	English	1.5%	13.7%	2.1%	22.4%
	Czech + English input sent	1.5%	29.4%	3.1%	42.8%
126k	Czech	6.7%	48.1%	12.5%	65.4%
	English	3.6%	28.1%	6.3%	45.4%
	Czech +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

- Significant vocabulary loss during phrase extraction:
  - e.g. 2.2%→3.9% for 7.5M Czech.
- OOV of Czech forms ∼twice as bad as in English.
- OOV of Czech lemmas lower than in English.
- Free word order of Czech apparent.

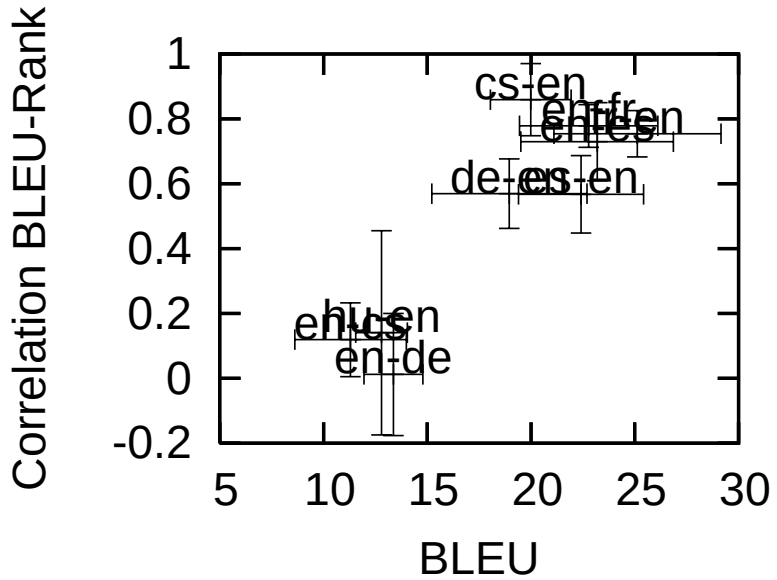
Side Note: BLEU vs. Human Rank

- 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

$\Rightarrow$ Bojar et al. (2010) use SemPOS, a coarse metric that correlates better with humans for Czech and English.

Overview of MT Systems Discussed

- Phrase-Based:
  - Vanilla Moses.
  - Optimized to SemPOS+BLEU. (Macháček and Bojar, 2011)
  - Two-Step Translation.
- TectoMT (deep, not quite RBMT).
- Rover System Combination.
- DEPFIX Grammatical Post-Editing. (Mareček et al., 2011)

Overview of MT Systems Discussed

- Phrase-Based:
  - Vanilla Moses.
  - Optimized to SemPOS+BLEU. . . . Deep Evaluation.
  - Two-Step Translation.
- TectoMT (deep, not quite RBMT).
- Rover System Combination. . . . PBMT over *.
- DEPFIX Grammatical Post-Editing. . . . RBMT over *.

Optimizing Towards SemPOS

SemPOS compares bags of lemmas, not sequences of forms.

- Sequences not overvalued
  $\Rightarrow$ better correlation with human ranking.
- Not fit for selecting best output from n-best list.
  $\Rightarrow$ Need to combine with e.g. BLEU.

WMT11 Tunable Metrics Task, manual ranking:

System	$\geq$others	$>$others
bleu•	0.79	0.28
bleu-single•	0.77	0.27
cmu-meteor•	0.76	0.27
rwth-cder	0.76	0.26
cu-sempos-bleu•	0.74	**0.29**
stanford-dcp•	0.73	0.27
nus-tesla-f	0.68	0.28
sheffield-rose	0.05	0.00

- Among the many "winners" (•).
- Best in "$>$others", i.e. when ties are not rewarded.

Optimizing Towards SemPOS

SemPOS compares bags of lemmas, not sequences of forms.

- Sequences not overvalued
  $\Rightarrow$ better correlation with human ranking.
- Not fit for selecting best output from n-best list.
  $\Rightarrow$ Need to combine with e.g. BLEU.

WMT11 Tunable Metrics Task, manual ranking:

System	$\geq$others	$>$others
bleu•	0.79	0.28
bleu-single•	0.77	0.27
cmu-meteor•	0.76	0.27
rwth-cder	0.76	0.26
cu-sempos-bleu•	0.74	**0.29**
stanford-dcp•	0.73	0.27
nus-tesla-f	0.68	0.28
sheffield-rose	0.05	0.00

- Among the many "winners" (•).
- Best in "$>$others", i.e. when ties are not rewarded.
- Generally hard to interpret the ranking.

Two-Step Moses 1/2

- English → lemmatized Czech
  - meaning-bearing morphology preserved
  - max phrase len 10, distortion limit 6
  - large target-side (lemmatized LM)
- 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, lattices on our todo list.
- See also works by Alex Fraser for targetting German.

Two-Step Moses 2/2

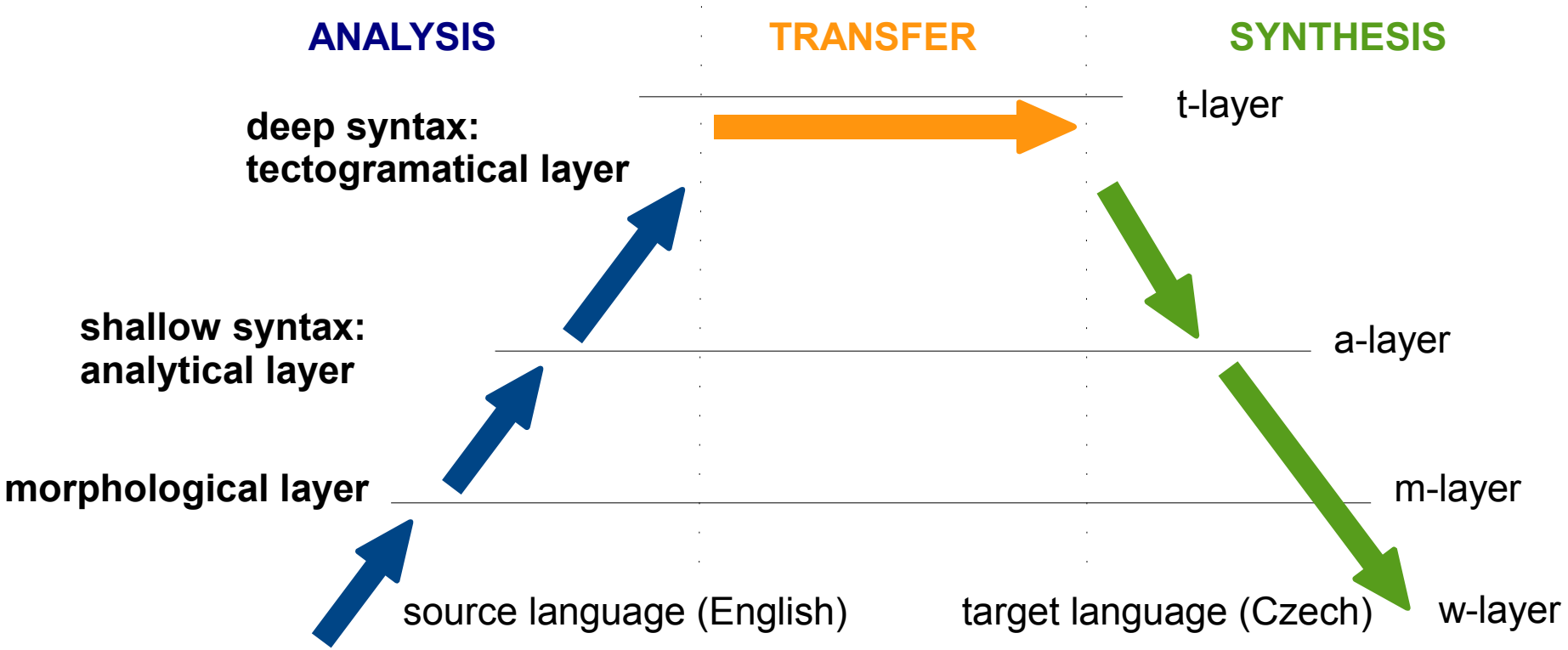
Data Size		Simple		Two-Step		Diff
Parallel	Mono	BLEU	SemPOS	BLEU	SemPOS	B. S.
126k	126k	10.28±0.40	29.92	10.38±0.38	30.01	↗ ↗
126k	13M	12.50±0.44	31.01	12.29±0.47	31.40	↘ ↗
7.5M	13M	14.17±0.51	33.07	14.06±0.49	32.57	↘ ↘

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

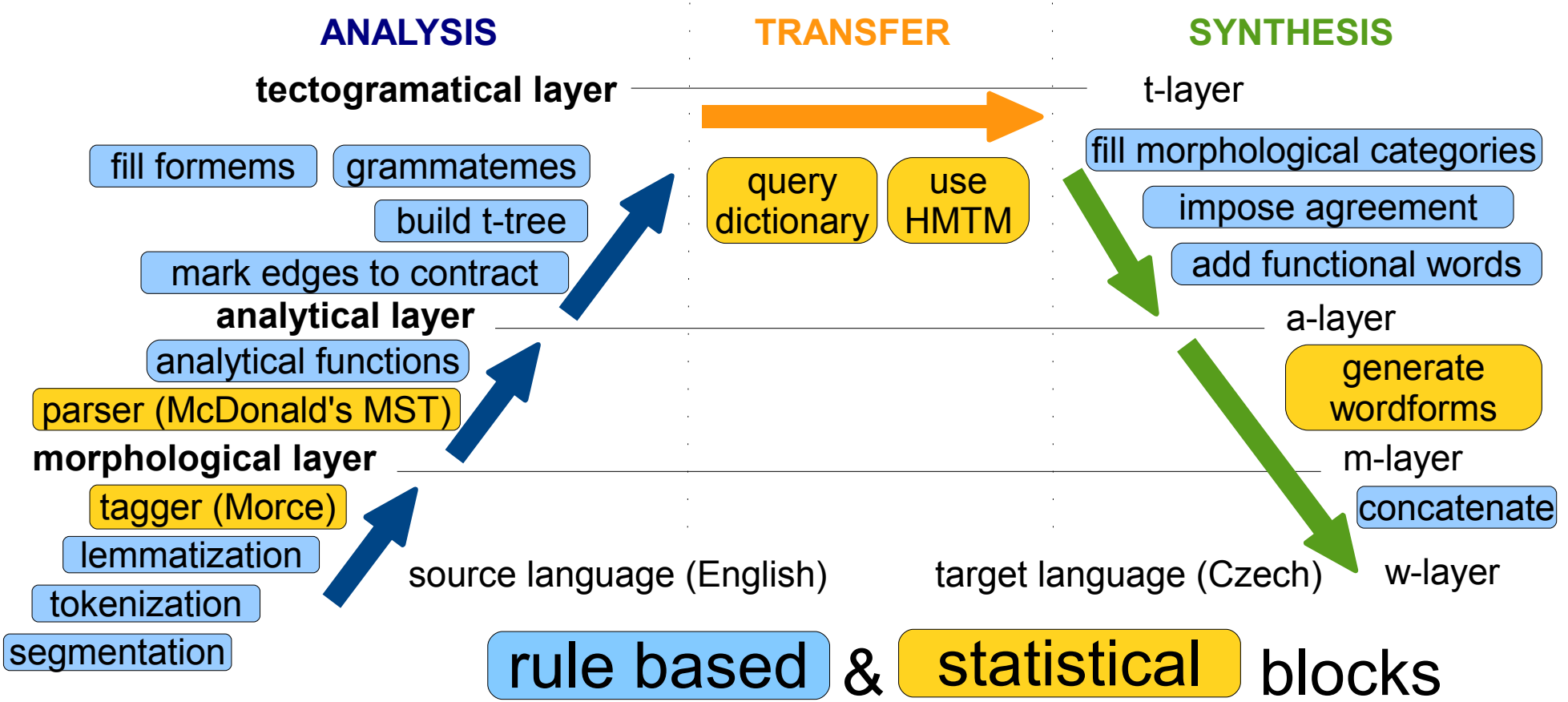
- Each annotator weakly prefers Two-step
  - but they don't agree on individual sentences.

TectoMT



- TectoMT = MT system built using blocks in Treex platform (on its way to CPAN).
- See Prague Dependency Tbk for tectogrammatical layer.

TectoMT



- TectoMT = MT system built using blocks in Treex platform (on its way to CPAN).
- See Prague Dependency Tbk for tectogrammatical layer.

TectoMT vs. PBMT: Can't Quite Tell

Metric	**CU-Bojar**	**TectoMT**
WMT10 $\geq$ others (official)	**66**	60
WMT10 $>$ others	**45**	44
WMT10 Edits acceptable [%]	**40**	34
WMT11 $\geq$ others (official)	**64**	58
WMT11 $>$ others	39	**40**
Quiz-based evaluation [%]	76	**82**

Even 6 annotators annotating the same 100 sentences pairwise don't agree:

	Better		Both	
Annotator	CU-BOJAR	TECTOMT	fine	wrong
A	**24**	23	5	11
C	10	**12**	5	36
D	**32**	20	2	9
M	11	**18**	7	27
O	**23**	18	4	18
Z	25	**27**	2	9
Total	**125**	118	25	110

Rover System Combination (1/2)

Following Fiscus (1997) and Matusov et al. (2008):

Systems vote which individual words to put in the output.

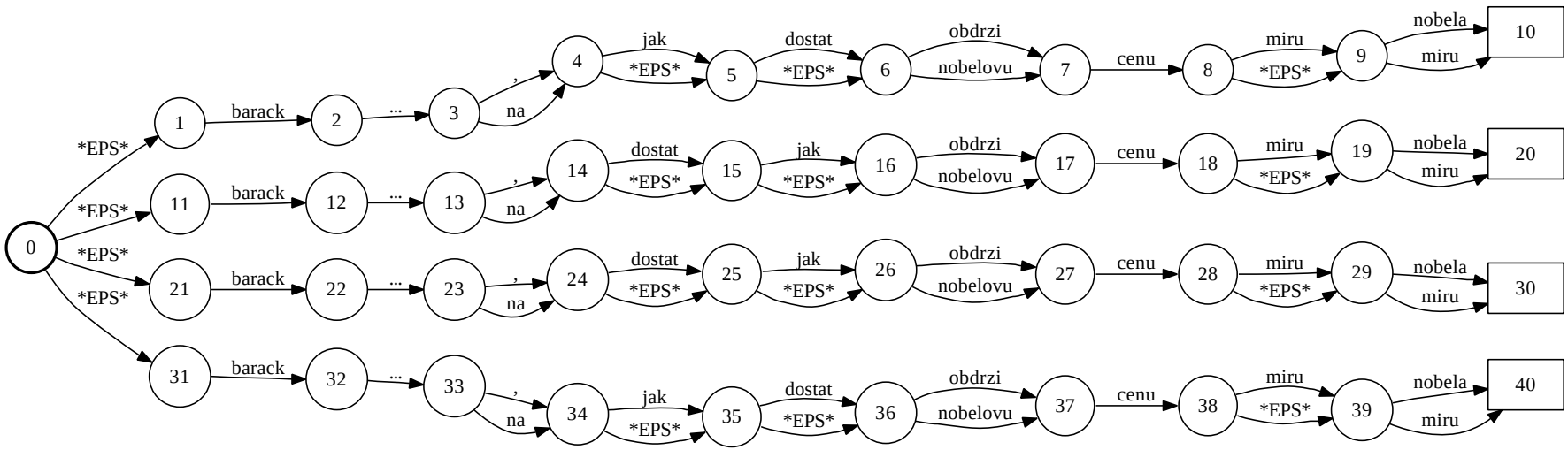
Procedure:

1. Given a "primary system" / "skeleton";
   - Align each system to one skeleton (bold), producing "bitexts":
     barack|**barack** ... ,|**na** dostat|ϵ jak|ϵ nobelovu|**nobelovu** cenu|**cenu** m
     barack|**barack** ... na|**na** nobelovu|**nobelovu** cenu|**cenu** míru|**míru**
     barack|**barack** ... ,|**na** obdrží|**nobelovu** cenu|**cenu** míru|ϵ nobela|**míru**
   - Combine all bitexts to confusion network:

barack	...	na	ϵ	ϵ	nobelovu	cenu	ϵ	míru
barack	...	,	dostat	jak	nobelovu	cenu	ϵ	míru
barack	...	na	ϵ	ϵ	nobelovu	cenu	ϵ	míru
barack	...	,	ϵ	ϵ	obdrží	cenu	míru	nobela

Rover System Combination (2/2)

2. Combine confusion networks of various skeletons to one lattice:



3. Add language model scores.
4. Optimize weights (word penalty, LM, skeleton choice, number of votes, did primary system vote for this, ...).
5. Select best path.

Combined Systems

In the following, we:

- Combine only ÚFAL's systems built for the WMT10 shared task.
- Tune and test on WMT10 combination task datasets.

	Dev Set		Test Set	WMT10 Manual Rank
BOJAR-PRIMARY	16.00±1.15	↗	16.90±0.61	65.5
BOJAR-SEMPOS	15.76±1.12	↗	16.61±0.59	-
BOJAR-2STEP	13.59±1.12	↗	14.38±0.58	-
TECTOMT	11.48±1.04	↗	13.19±0.58	60.1
GOOGLE	17.32±1.25	↘	16.76±0.60	70.4
EUROTRAN	9.64±0.92	↗	11.04±0.48	54.0
PCTRANS2010	10.24±0.92	↗	10.84±0.46	62.1

Note Google discrepancy between Dev and Test ⇒ overfitting would be very likely.

Even "Bad" Systems Offer Words

Analyzing 44193 toks in the ref of WMT10 syscomb testset.

What is the % tokens produced by:

- ... the primary system only BOJAR-PRIMARY?
- ... one of the secondary systems only?

	BOJAR-PRIMARY (16.90±0.61) vs.			
	BOJAR-SEMPOS	BOJAR-2STEPSL	TECTOMT	the 3 other
	16.61±0.59	14.38±0.58	13.19±0.58	-
In Both	48.3	43.8	41.2	50.8
Nowhere	45.4	42.8	41.0	37.0
Primary Only	3.5	8.0	10.6	1.0
Secondary Only	**2.8**	**5.4**	**7.1**	**11.2**

- TectoMT may bring up to 7.1% tokens, Two-Step 5.4%.
- 37% tokens of the ref not available in 1-best outputs.

Manual System Combination

To check the plausibility of "voting assumption" we manually do the task:

- Myself:
  - English→Czech, WMT10, 4 systems, 52 sents.
  - Reference translation available.
  - Attempted to stick to the original word order.
- Matusov (2009):
  - probably Chinese(?)→English, IWSLT 2006.
    (Matusov (2009) p. 140 talks about TC-STAR07 es→en.)
  - 4 systems, 489 sents.
  - Without looking at source or reference.
  - Allowed any reordering.

Plausibility of Voting Assumption

How many produced tokens actually had the majority support?

	Matusov (2009) Manual		My en→cs WMT10 Manual		My en→cs WMT10 Auto	
Supported by	Toks	%	Toks	%	Toks	%
1	978	15.8	160	19.4	30	3.6
2	1117	18.1	110	13.3	183	21.9
≤ 2	2095	**33.9**	270	**32.7**	213	**25.5**
3	1279	20.7	137	16.6	188	22.5
4	2806	45.4	417	50.6	435	52.0
Total	6180	100.0	824	100.0	836	100.0

... about $\frac{1}{3}$ of manually and $\frac{1}{4}$ of automatically combined tokens have no majority support (weights influence this).

Directions Examined in System Comb.

No Rover, just Moses, simply "add to training":

- Add the 3 other outputs to training data of BOJAR-PRIMARY.

Within RWTH Rover implementation (minor modifications):

- Improving monolingual word alignments.
  - Various automatic synonym dictionaries, ...

RWTH alignment + Moses path selection and MERT:

- More detailed lattice arc weights.
- Larger LMs.
- LMs for morphological tags.

Baseline Combinations

Combination Method	Weights Default	Optimized
RWTH Rover	17.50±0.64	17.42±0.63
Moses Add-to-training	-	17.25±0.62
Moses Rover	-	17.19±0.61
bojar-primary	-	16.90±0.61
google	-	16.76±0.60

- RWTH marginally better unoptimized (sys. weights equal).
- MERT optimizer in Moses worse than JaneOpt in RWTH setup.
- Add-to-training works but very inefficient implementation:
  - Need to re-align, re-extract phrases, re-tune in MERT.

Larger LMs

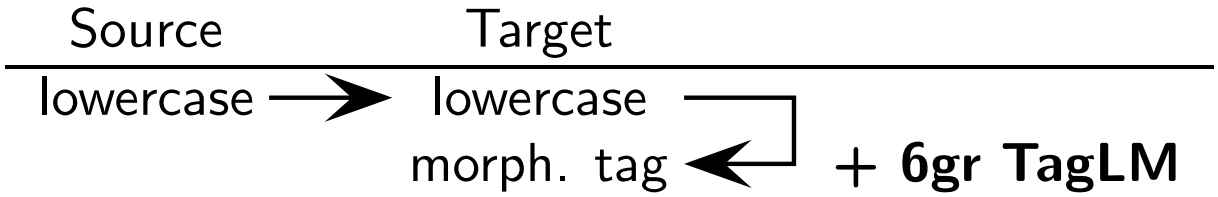
- Default: only 3gr LM based on input hypotheses used.
- G. Leusch (RWTH) saw no gains from additional LM.
  - en→cs and Moses MERT do make use of that.
- Additional data: WMT10mono, 13M sents, 211M toks.

		Underlying Alignment	
Combination Method	Baseline	Eqvoc+Lemmas	$\oslash \pm \sigma$ Across All
RWTH Unoptimized	17.50±0.64	17.53±0.63	17.52±0.01
Moses +5grLM	17.36±0.61	17.49±0.61	17.48±0.06
Moses +4grLM	17.63±0.59	17.45±0.62	17.46±0.08
RWTH Optimized	17.42±0.63	17.47±0.61	17.45±0.05
Moses +3grLM	17.46±0.61	17.44±0.63	17.41±0.07
Moses (default LM)	17.32±0.63	17.34±0.61	17.32±0.06

- With the additional LM, Moses can reach RWTH optimizer.
- Higher *n*-grams marginally better.

LMs for Morphological Tags

- Additional LM over morphological tags can help in factored translation (Bojar, 2007).



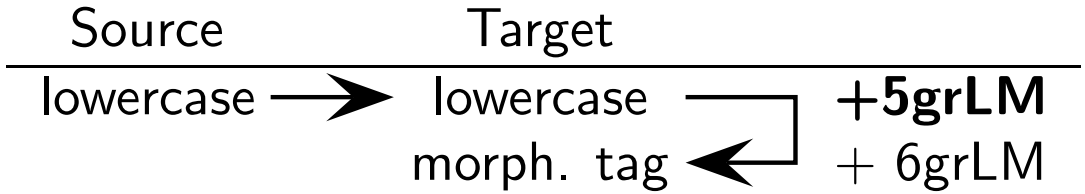
- Hypotheses "tagged with unigram tagger" on the fly.

	Underlying Alignment		
	Baseline	Eqvoc+Lemmas	$\varnothing \pm \sigma$ Across All
Moses +tagLM	**17.88±0.62**	**17.95±0.59**	**17.90±0.12**
RWTH Unoptimized	17.50±0.64	17.53±0.63	17.52±0.01
RWTH Optimized	17.42±0.63	17.47±0.61	17.45±0.05
Moses (default LM)	17.32±0.63	17.34±0.61	17.32±0.06

- Beating RWTH Rover (no support for factors) at last.

TagLM and Large LM

- We can combine TagLM and regular LM.
- This makes 15 weights in MERT optimization:
  - 9 arc weights, 3 LM weights, 2 tagger weights, word penalty.



	Underlying Alignment		
	Baseline	Eqvoc+Lemmas	$\oslash \pm \sigma$ Across All
Moses +tagLM +5grLM	**18.01±0.66**	**17.80±0.59**	**17.97±0.09**
Moses +tagLM	17.88±0.62	17.95±0.59	17.90±0.12
RWTH Unoptimized	17.50±0.64	17.53±0.63	17.52±0.01
Moses +5grLM	17.36±0.61	17.49±0.61	17.48±0.06
RWTH Optimized	17.42±0.63	17.47±0.61	17.45±0.05
Moses (default LM)	17.32±0.63	17.34±0.61	17.32±0.06
RWTH Optimized 7 Systems	18.02±0.65	18.07±0.67	-

- In terms of BLEU score, this approaches the combination of all 7 systems.

Manual Evaluation of System Comb.

- Manually ranked 65 sentences.
  - All hyps get equally-poor/equally-ok, or
  - All hyps get a rank (at least one of them gets rank 1).

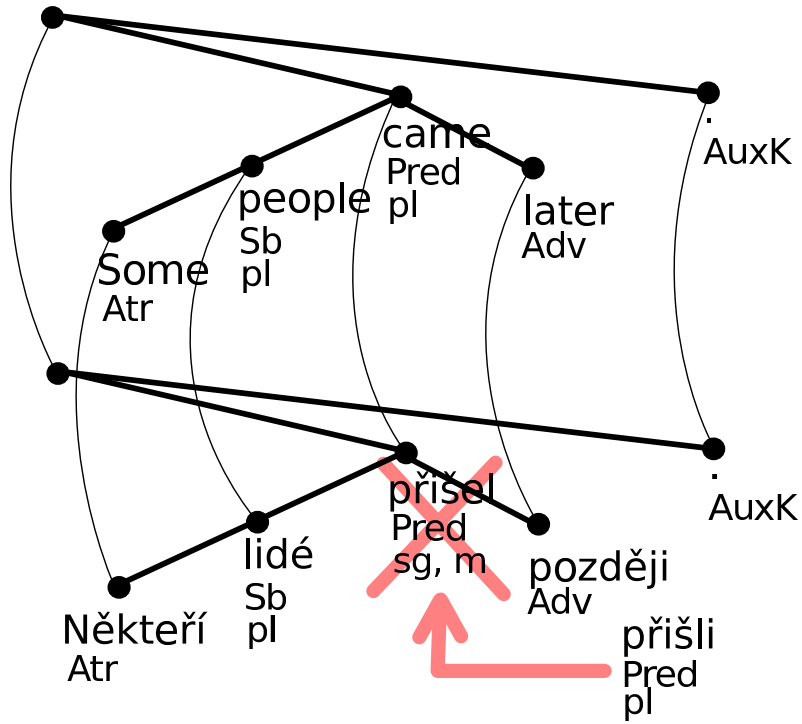
		Equally		Ranked as			
		Poor	Ok	1	2	3	4
Moses +tagLM +5grLM	**18.01±0.66**	11	7	18	16	10	3
RWTH Optimized	17.42±0.63	11	7	**22**	17	7	1
Moses (default LM)	17.32±0.63	11	7	17	14	14	2
bojar-primary	16.90±0.61	11	7	14	20	9	4
google	16.76±0.60						

- Improvement over individual systems confirmed.
- Other differences not clear.

  . . . which is in line with confidence bounds.

Grammatical Post-Processing (DEPFIX)

1. Align source and MT output (or use detailed output).
2. Parse MT output.
3. Apply rule-based corrections:
   - Hack wrong tagging/parse based on English.
   - Agreement (subject case, noun-adj, subject-predicate, subject-past participle).
   - Case implied by preposition.
   - Remove extra reflexive particles.



. . . implemented in Treex platform.

DEPFIX Impact on WMT11 Systems

System	Before	After	Delta (Confidence)
CU-ZEMAN	14.61	14.80	0.19 (0.09–0.29)
UEDIN	17.80	17.88	0.08 (-0.02–0.17)
CMU-HEAFIELD	20.24	20.32	0.08 (-0.03–0.19)
JHU	17.36	17.42	0.06 (-0.03–0.16)
ONLINE-B	20.26	20.31	0.05 (-0.06–0.16)
CU-TWOSTEP	16.57	16.60	0.03 (-0.07–0.13)
UPV-PRHLT.	20.68	20.69	0.01 (-0.08–0.11)
COMMERC2	09.32	09.32	0.00 (-0.04–0.04)
CU-POPEL	14.12	14.11	-0.01 (-0.06–0.03)
CU-BOJAR	16.88	16.85	-0.03 (-0.12–0.07)
CU-TAMCHYNA	16.32	16.28	-0.04 (-0.14–0.06)

- No gain for TectoMT (CU-POPEL).
- Surpringly no gain for our tuned Moseses (CU-BOJAR, CU-TAMCHYNA).

Manual Evaluation of DEPFIX

System	Annotator	Sents	Percent Better	Worse	Equal
cu-bojar-twostep	A	269	56.5	14.5	29.0
cu-bojar-twostep	B	269	64.3	18.6	17.1
online-B	A	247	63.1	15.9	21.1
online-B	B	247	66.8	25.9	7.3

- Around 60% sentences that change are actually improved.

TestSet	Sents.	Percent Better	Worse	Equal	BLEU Before	After	Diff
2010	104	50.0	19.2	30.8	16.99	17.38	0.39
2011	101	65.3	18.8	15.8	13.99	13.87	-0.12

- Confirmed improvement around 50–60% sentences.
- BLEU does not indicate that.

Summary

- Rich morphology makes everything in MT harder.
- Improvements impossible to check with BLEU.
- Manual ranking of hypotheses:
  - Clear when comparing 2 similar outputs (DEPFIX).
  - May fail for already 2 dissimilar outputs (TECTOMT vs. CU-BOJAR).
- Rover System Combination (PBMT over *):
  - 1/3 of tokens should be chosen despite no majority.
  - Helps over individual systems.
  - Tuning in Moses Poor.
  - LMs (over factors) help.
- DEPFIX Grammatical Post-Processing (RBMT over *):
  - Treex platform proves versatile.
  - Got minor improvements of most systems.

References

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.

Ondřej Bojar. 2007. English-to-Czech Factored Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 232–239, Prague, Czech Republic, June. Association for Computational Linguistics.

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, pages 347–354. IEEE.

Matouš Macháček and Ondřej Bojar. 2011. Approximating a Deep-Syntactic Metric for MT Evaluation and Tuning. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 92–98, Edinburgh, Scotland, July. Association for Computational Linguistics.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 426–432, Edinburgh, Scotland, July. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. IEEE Transactions on Audio, Speech and Language Processing, 16(7):1222–1237, September.

Evgeny Matusov. 2009.
Combining Natural Language Processing Systems to Improve Machine Translation of Speech. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, December.