Fine-grained Human Evaluation of Neural versus Phrase-based Machine Translation

EAMT, Praha, 31st May 2017

Filip Klubička
University of Zagreb

Antonio Toral
University of Groningen

Víctor M. Sánchez-Cartagena
Prompsit Language Engineering

Introduction

Introduction

In many setups, NMT has surpassed the performance of the mainstream MT approach to date: PBMT

Introduction

In many setups, NMT has surpassed the performance of the mainstream MT approach to date: PBMT

E.g. news translation shared task at WMT'16

- 10 language directions: EN $\leftrightarrow$ CS, DE, FI, RO, RU
- Automatic evaluation: BLEU, TER
- Human evaluation: ranking translations

Overall Evaluation (Automatic)

System	CS	DE	FI	RO	RU
			From EN		
PBMT	23.7	30.6	15.3	27.4	24.3
NMT	**25.9**	**34.2**	**18.0**	**28.9**	**26.0**
			Into EN		
PBMT	30.4	35.2	23.7	**35.4**	**29.3**
NMT	**31.4**	**38.7**	-	34.1	28.2

Table 1: BLEU scores of the best NMT and PBMT systems

Bold: statistical significance

Overall Evaluation (human)

System	CS	DE	FI	RO	RU
	From EN				
PBMT	23.7	30.6	15.3	27.4	24.3
NMT	**25.9**	**34.2**	**18.0**	**28.9**	**26.0**
	Into EN				
PBMT	30.4	35.2	23.7	**35.4**	**29.3**
NMT	**31.4**	**38.7**	-	34.1	28.2

Table 2: BLEU scores of the best NMT and PBMT systems

Bold: statistical significance (BLEU)

Green: statistical significance (human evaluation)

Background

Overall, NMT outperforms PBMT, but... which are its strengths? And what are its weaknesses?

Background

Paper	**Direction**	**Findings: NMT...**
Bentivogli et al., 2016	EN→DE	1. Improves on reordering and inflection 2. Decreases PE effort 3. Degrades with sentence length

Background

Paper	Direction	Findings: NMT...
Bentivogli et al., 2016	EN→DE	1. Improves on reordering and inflection 2. Decreases PE effort 3. Degrades with sentence length
Toral and Sánchez-Cartagena, 2017	EN→CS, DE, FI, RO, RU	1. Corroborated findings 1 and 2 from Bentivogli 2. Higher inter-system variability
	CS, DE, RO, RU→EN	3. More reordering than PBMT but less than hierarchical PBMT

Background

Limitations of these analyses

- Performed automatically. E.g. inflection errors detected with a PoS tagger
- Coarse-grained. 3 error types: inflection, reordering and lexical

Background

Limitations of these analyses

- Performed automatically. E.g. inflection errors detected with a PoS tagger
- Coarse-grained. 3 error types: inflection, reordering and lexical

This work

This work: fine-grained human analysis of NMT vs PBMT and factored PBMT

- Fine-grained. Errors annotated following a detailed error taxonomy (>20 error types)
- Human. Errors annotated manually
- Factored PBMT. Not compared to NMT to date[1]
- Direction. English-to-Croatian, i.e. MT into a morphologically-rich target language, challenge for phenomena such as agreement (case, gender, number)

[1] To the best of our knowledge

Data sets and MT systems

Data sets

- Dev. First 1k sentences from English test set at WMT'12, translated into Croatian
- Test. Same but from WMT'13
- Train
  - Parallel. 4.8M sentence pairs selected according to cross-entropy from different sources: EU/legal, news, web, subtitles
  - Monolingual. Web + target side of parallel data

MT systems

All systems trained on the same data set. NMT does not use monolingual data.

- Pure PBMT. Standard Moses + hierarchical reordering, bilingual neural LM, OSM
- Factored PBMT. Maps 1 factor in the source (surface form) to 2 in the target (surface form and morphosyntactic description)
- NMT
  - Sequence-to-sequence with attention
  - Unsupervised word segmentation (byte pair encoding)
  - Trained for 10 days, models saved every 4.5h. Ensemble of 4 best models on dev set

MT systems

Results with automatic metrics

System	BLEU	TER
PBMT	0.2544	0.6081
Factored PBMT	0.2700	0.5963
NMT	0.3085	0.5552

Human Evaluation

Error taxonomy

Multidimensional Quality Metrics (MQM)

- Framework for defining custom quality metrics
- Provides a flexible vocabulary of quality issue types

Error taxonomy

Multidimensional Quality Metrics (MQM)

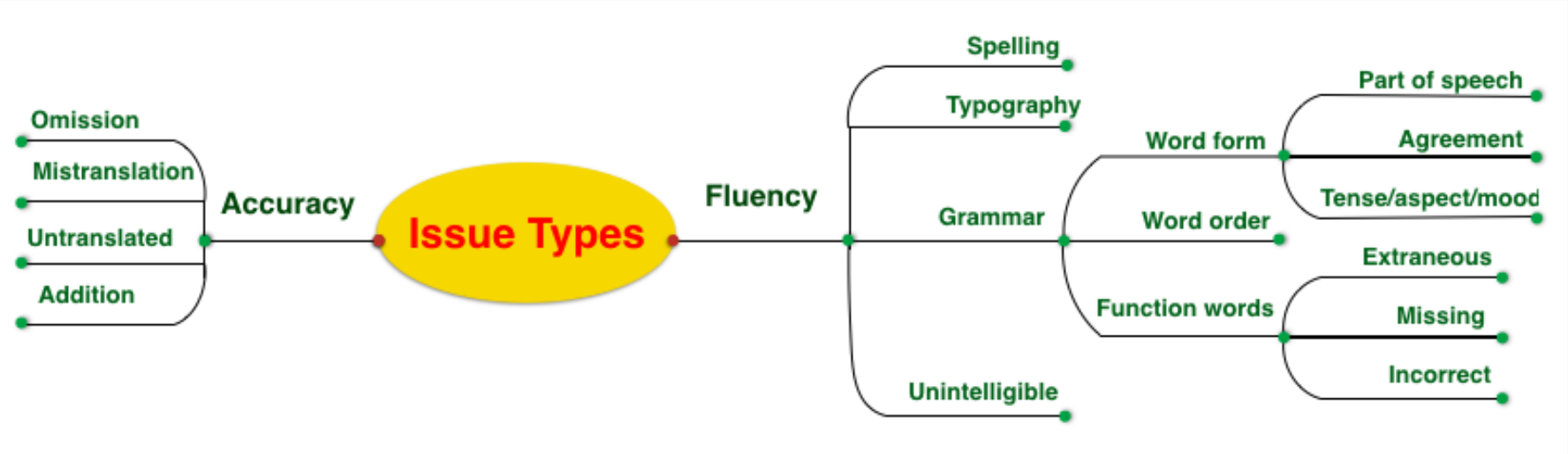
- Framework for defining custom quality metrics
- Provides a flexible vocabulary of quality issue types

We devised an MQM-compliant taxonomy with these aims

- Right level of granularity: trade-off between having a detailed taxonomy and the annotation process being viable
- Error types relevant for the translation direction

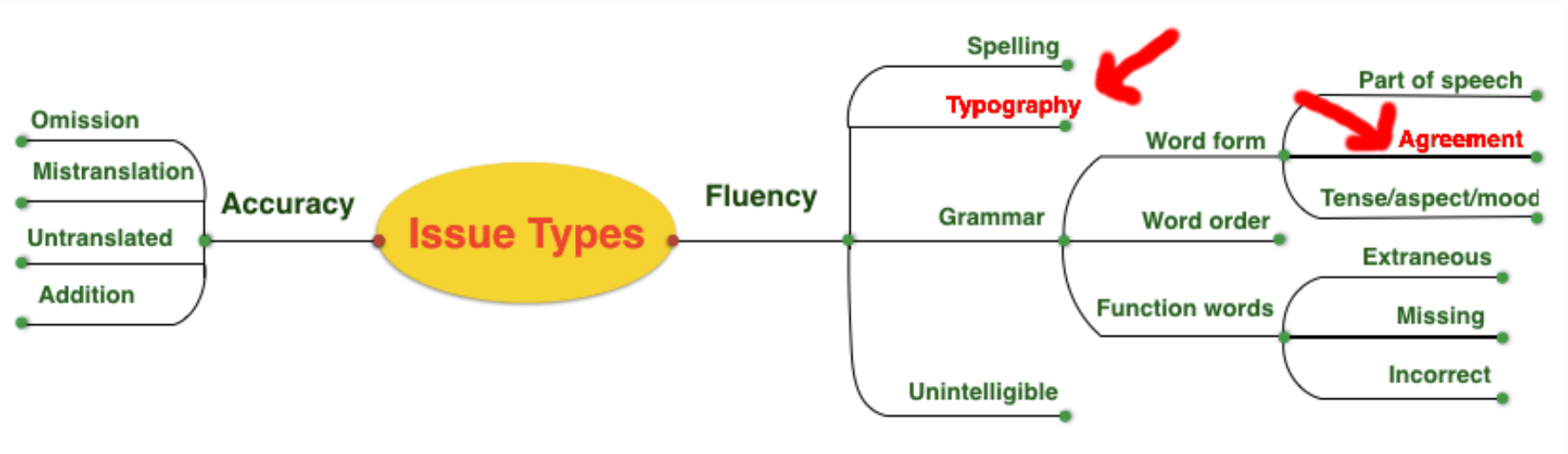
Error taxonomy

MQM core taxonomy



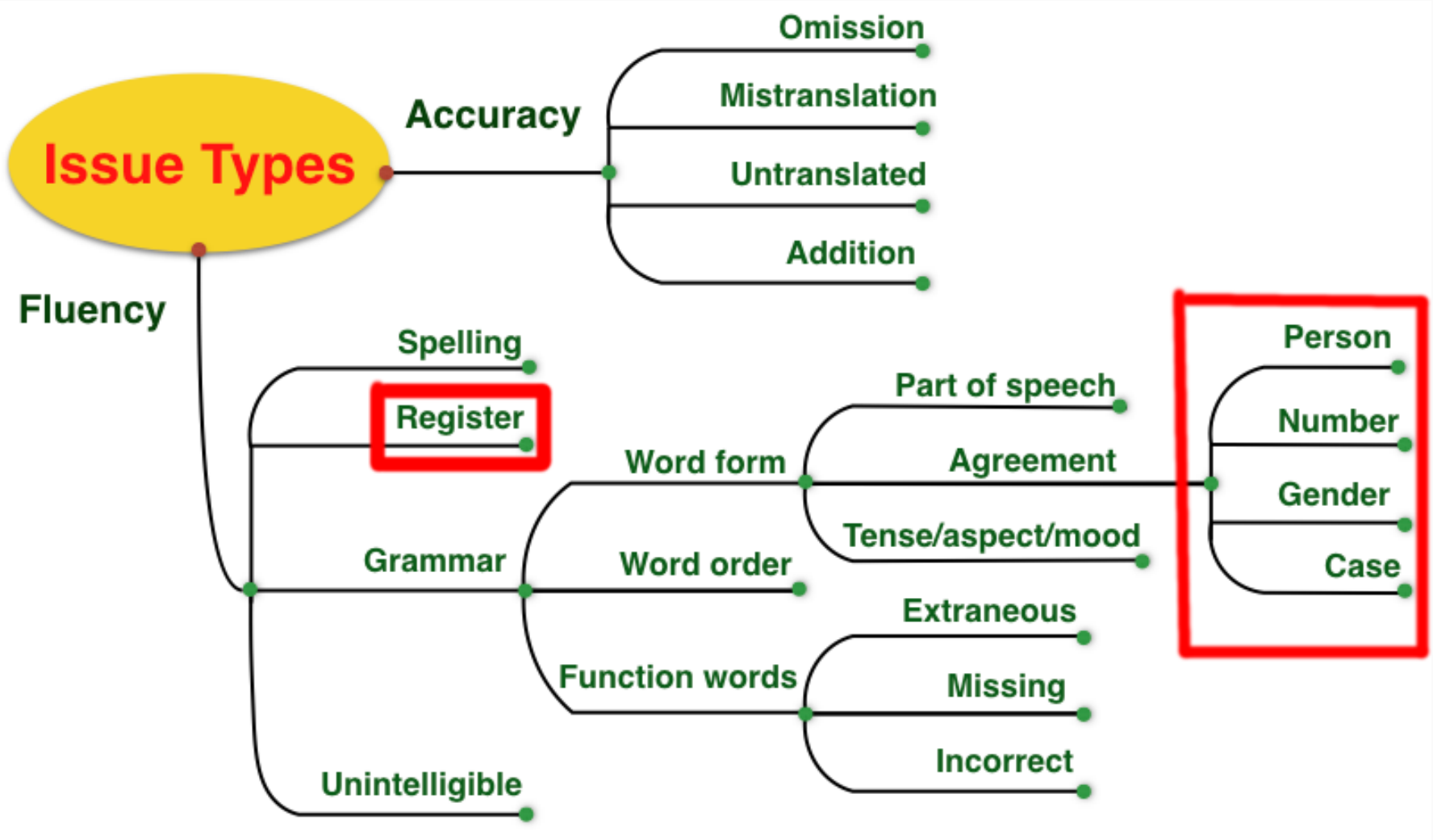
Error taxonomy

MQM core taxonomy



Error taxonomy

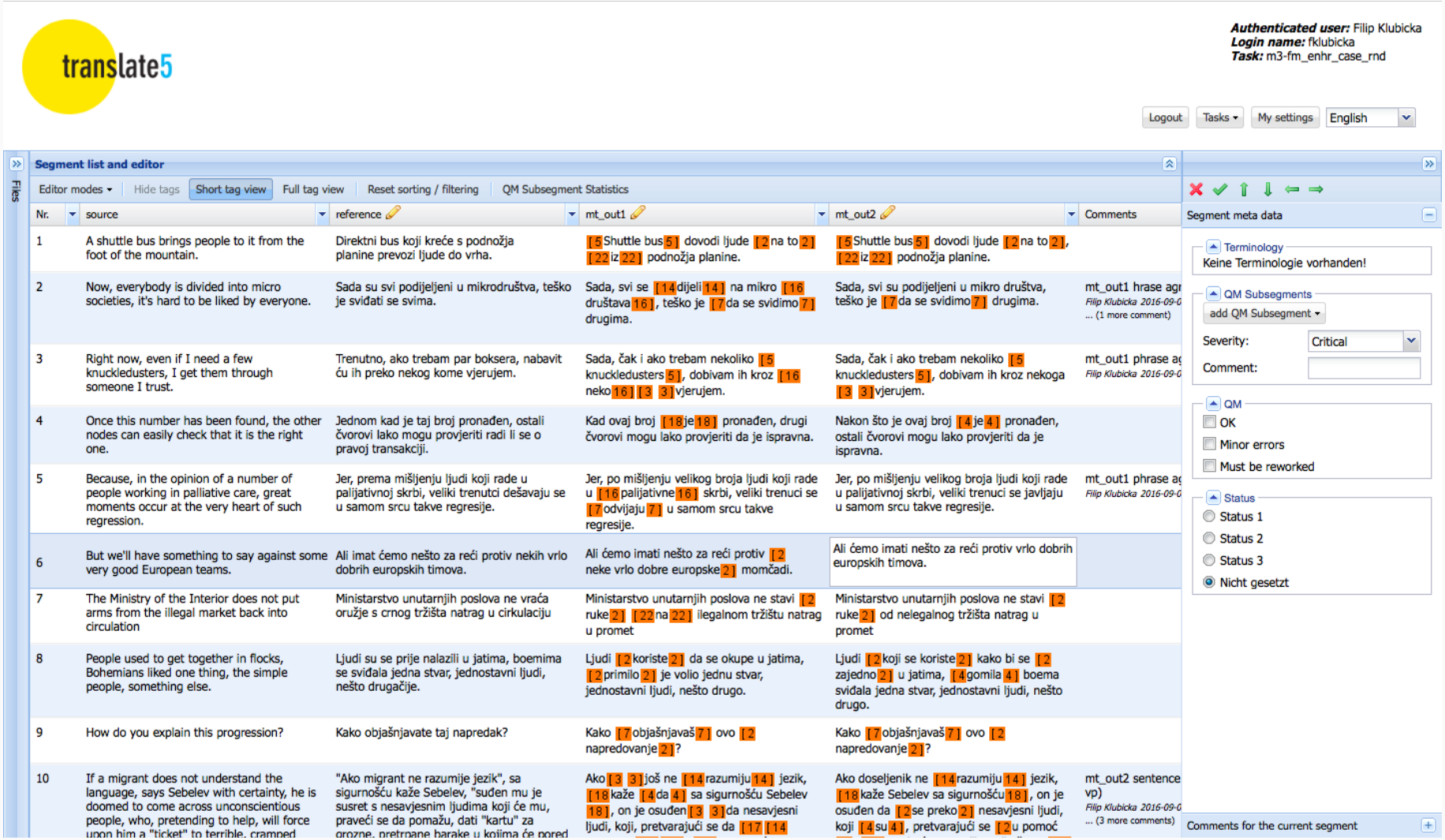
MQM Slavic taxonomy



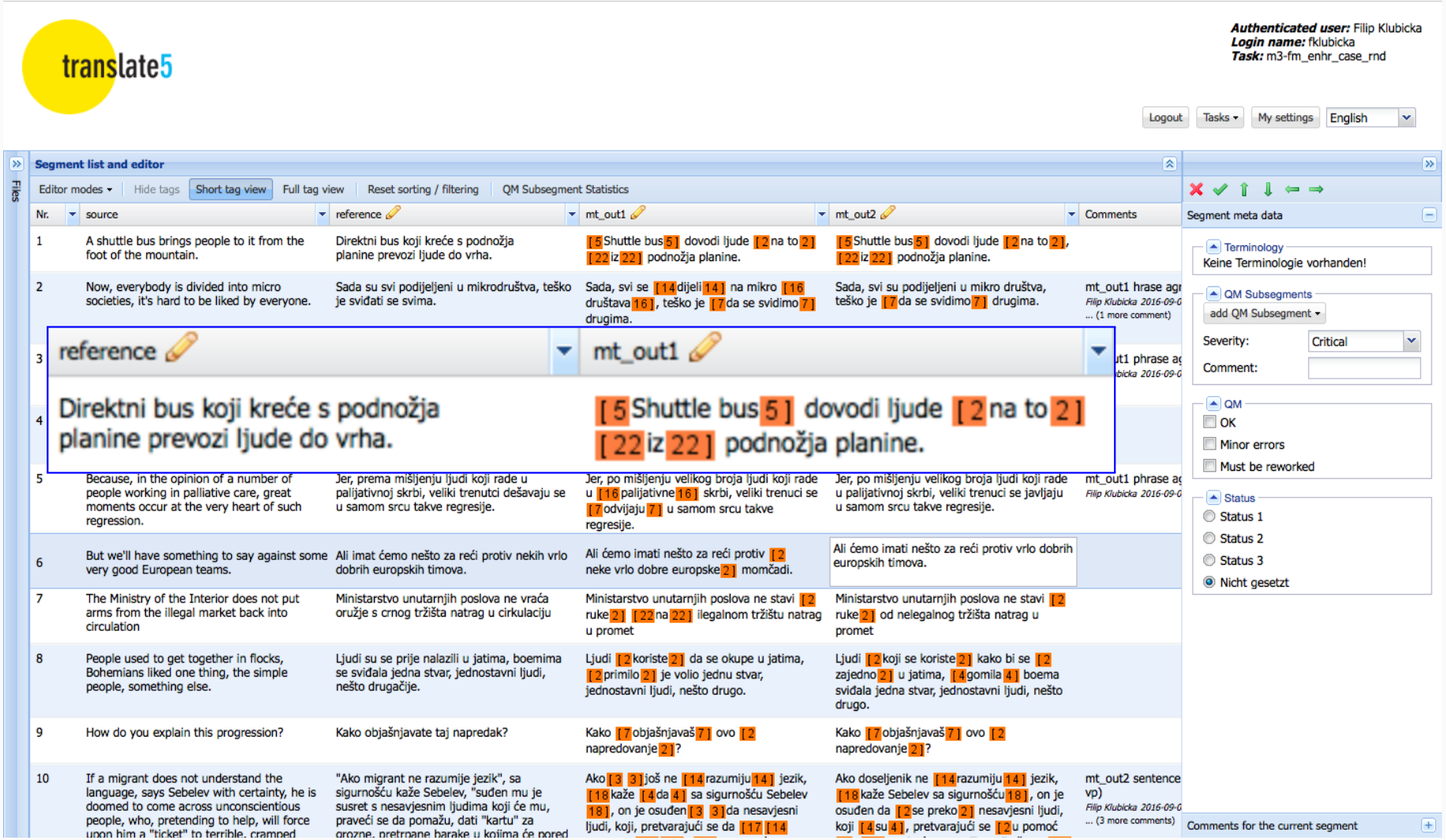
Annotation Setup

- Tool: `translate5`
- 2 annotators (native Croatian, C1 English)
- 100 randomly selected sentences from the test set annotated
  - Total: 600 annotated sentences (100 sentences * 3 systems * 2 annotators)

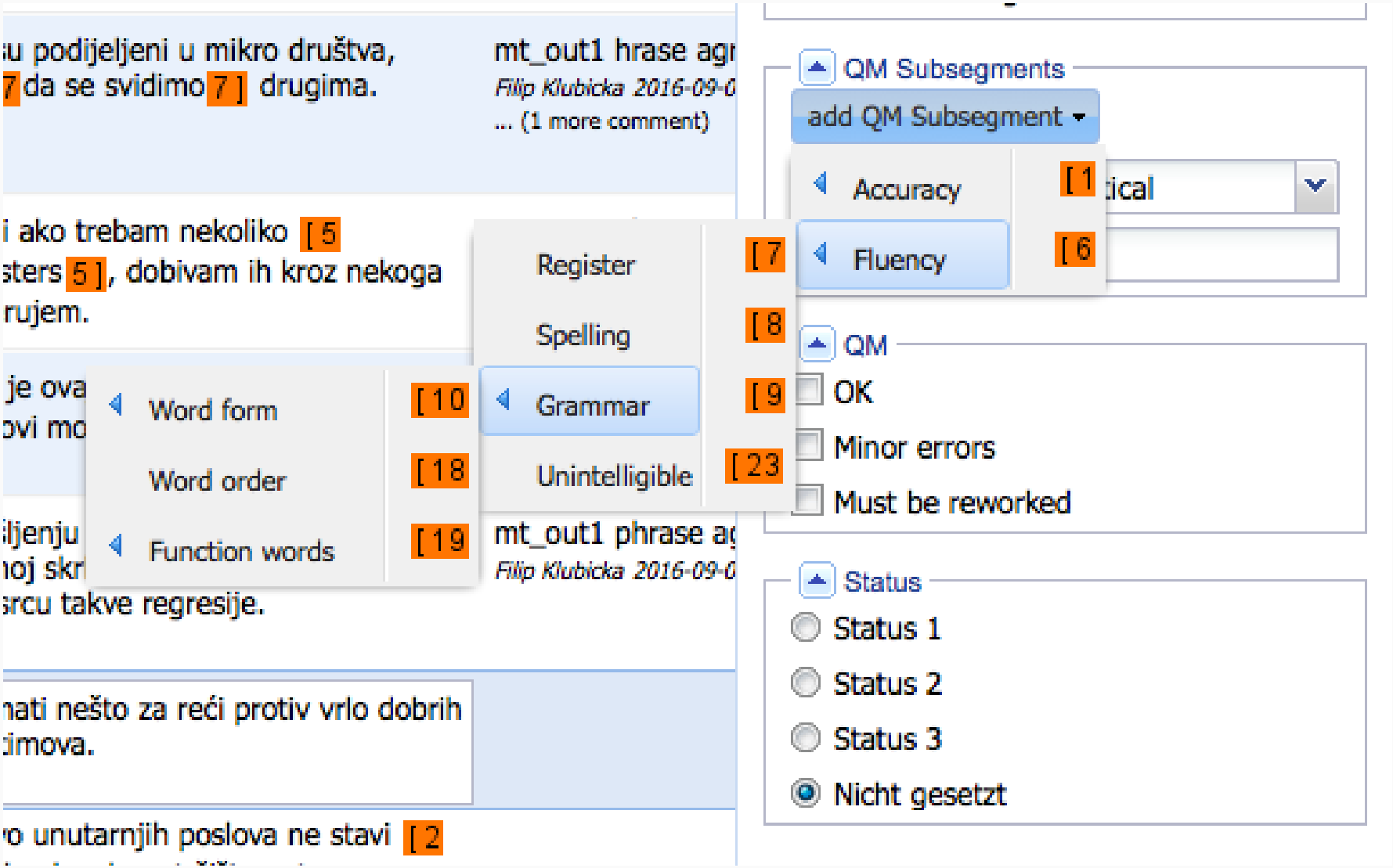
Annotation Process



Annotation Process



Annotation Process



Inter Annotator Agreement

Calculated at sentence level with Cohen's κ

Inter Annotator Agreement

Calculated at sentence level with Cohen's κ

Inter annotator agreement for each MT system

PBMT	Factored	NMT	Concatenated
0.56	0.49	0.44	0.51

Inter Annotator Agreement

Calculated at sentence level with Cohen's κ

Inter annotator agreement for each MT system

PBMT	Factored	NMT	Concatenated
0.56	0.49	0.44	0.51

Inter annotator agreement for each error type (min: 0.27, max: 0.72)

Results

Notes

- Outputs have different length
  - normalise errors by number of tokens: ratio of tokens with and without errors
- Statistical significance with χ^2
  - 2x2 contingency tables for each pair of systems: (PBMT, factored), (PBMT, NMT), (factored, NMT)
- Error types: concatenated and separately

Results

Overall: considering all error types

	PBMT		Factored		NMT	
	No error	**Error**	**No error**	**Error**	**No error**	**Error**
Overall	2826	1010	3007	**809	3199	**469

** $p<0.01$ (compared to the system on its left)

Results

Overall: considering all error types

	PBMT		Factored		NMT	
	No error	**Error**	**No error**	**Error**	**No error**	**Error**
Overall	2826	1010	3007	**809	3199	**469

** $p<0.01$ (compared to the system on its left)

Relative reduction of errors:

- Factored: 20%
- NMT: 42% (wrt factored), 54% (wrt PBMT)

Results (by error type, accuracy branch)

	PBMT		Factored		NMT	
Error type	**No error**	**Error**	**No error**	**Error**	**No error**	**Error**
Accuracy	3467	369	3525	*291	3402	266
Mistranslation	3547	289	3586	*230	3471	197
Omission	3801	35	3793	23	3619	*49
Addition	3814	22	3797	19	3655	13
Untranslated	3813	23	3797	19	3662	*6

* $p<0.05$ (compared to the system on its left)

Results (by error type, accuracy branch)

	PBMT		Factored		NMT	
Error type	**No error**	**Error**	**No error**	**Error**	**No error**	**Error**
Accuracy	3467	369	3525	*291	3402	266
Mistranslation	3547	289	3586	*230	3471	197
Omission	3801	35	3793	23	3619	*49
Addition	3814	22	3797	19	3655	13
Untranslated	3813	23	3797	19	3662	*6

* $p<0.05$ (compared to the system on its left)

- Factored and NMT have less accuracy errors than PBMT
- NMT reduces untranslated (better coverage due to sub-word segmentation?)
- NMT leads to more omission errors than factored

Results (by error type, fluency branch)

	PBMT		Factored		NMT	
Error type	**No error**	**Error**	**No error**	**Error**	**No error**	**Error**
Fluency	3195	641	3298	*518	3465	**188
Unintelligible	3790	46	3769	47	3668	**0
Grammar	3270	566	3371	**445	3497	**156
Word order	3752	84	3752	64	3646	**22
Word form	3389	447	3471	*345	3538	**102
Tense...	3775	61	3765	51	3648	*20
Agreement	3466	370	3540	*276	3566	**102
Number	3778	58	3772	44	3646	*22
Gender	3788	48	3756	60	3644	*24
Case	3614	222	3694	*122	3622	**46
Person	3836	0	3816	0	3664	4

** $p<0.01$, * $p<0.05$ (compared to the system on its left)

Conclusions

Conclusions: contributions

1. Human fine-grained error analysis of NMT
2. NMT compared not only to pure and hierarchical PBMT but also to factored models
3. Devised an MQM-compliant taxonomy for Slavic languages
4. Approach to analyse statistically MQM results

Conclusions: findings

- Overall errors. NMT reduces them by 54% (wrt PBMT) and by 42% (wrt factored PBMT)
- Agreement errors (number, gender and case). NMT is specially effective, 72% reduction (wrt PBMT), and 63% (wrt factored PBMT)
- Omission. The only error type for which NMT underperformed factored PBMT (40% increase)

Future work

- Compare to PBMT with morph segmentation
- NMT-focused MQM evaluation: add fine-grained tags under the Accuracy branch
- NMT vs PBMT analysis for novels

Thank you! Děkuji!

Questions?

Inter Annotator Agreement

Inter annotator agreement for each error type (min: 0.27, max: 0.72)

Error type	**Cohen's κ**
Accuracy	
Mistranslation	0.53
Omission	0.37
Addition	0.47
Untranslated	0.72
Fluency	
Unintelligible	0.35
Register	0.27
Word order	0.4
Function words	
Extraneous	0.46
Incorrect	0.29
Missing	0.33
Tense...	0.38
Agreement	0.33
Number	0.54
Gender	0.53
Case	0.56