Automatic Translation Error Analysis

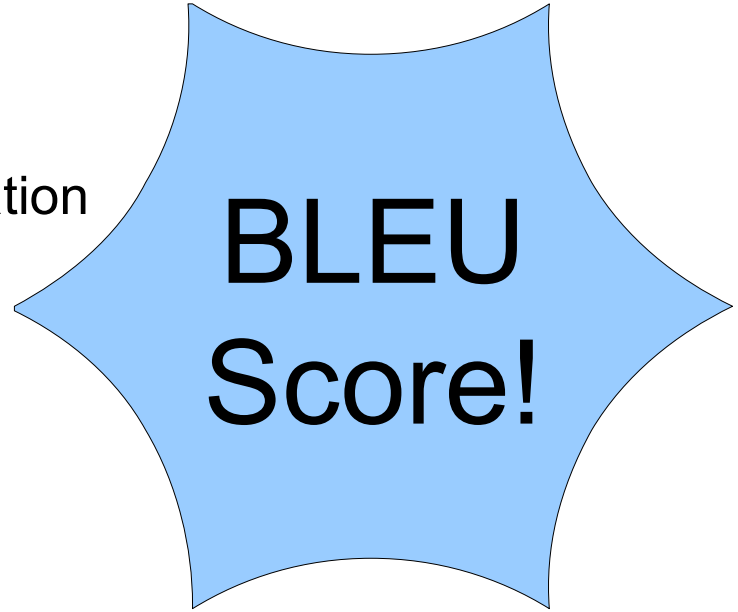
Mark Fishel, Ondřej Bojar,
Daniel Zeman, Jan Berka

University of Tartu
&
Charles University in Prague

The research has been supported by the grants P406/11/1499, P406/10/P259, SF0180078s08.

Machine Translation Evaluation?

- Most-widely used
- Automatic
  - Need 1 or more reference translation
- Easily computable
- Suitable for tuning weights

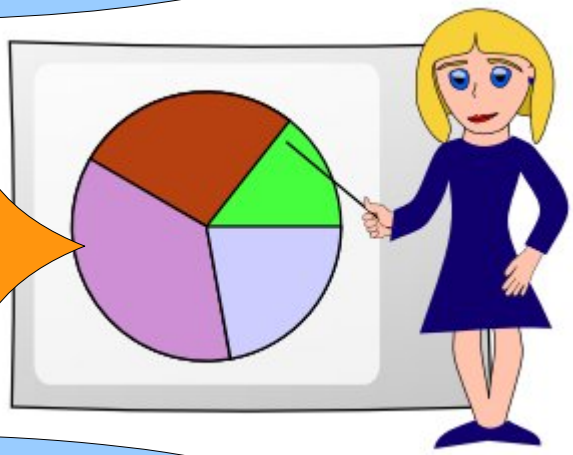


- Correlation with human judgments?

BLEU

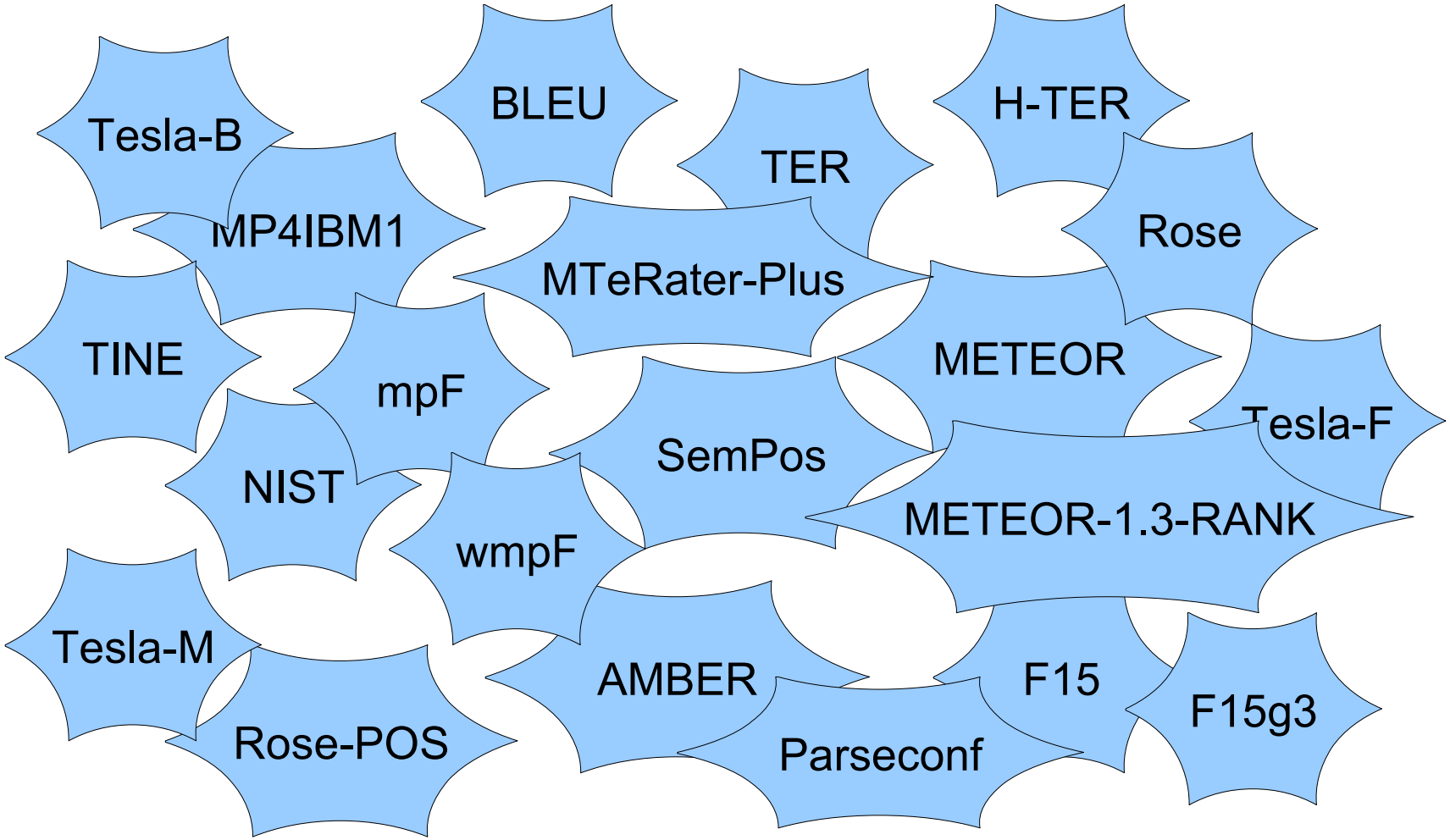
How good is the translation?

0.29!



Awesome!
And, actually, how good is it?

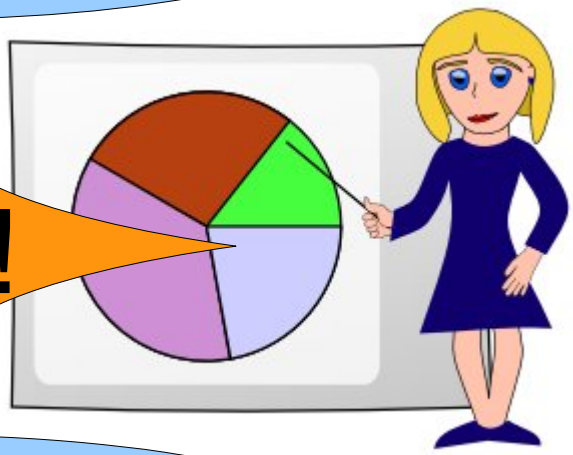
Don’t Like BLEU?



Don’t Like BLEU?

How good is the translation?

327.51 ± 6.021!



Awesome!
And, actually, how good is it?

Any Single-Number Metric…

- May be good for…
  - comparing two systems **on given dataset**
  - tuning model weights (if easily computable)
- Rarely, if at all…
  - does the absolute value tell anything

- **BUT NEVER…**
  - points directly to the particular **weaknesses** of the system

Except for OOV

- Out-of-Vocabulary rate:
  - How many test input words are unknown
    i.e. never seen in training data?

- But this is just one aspect.
- Can we perform a more detailed **error analysis?**
- Can we do it systematically?
- And (semi)automatically?

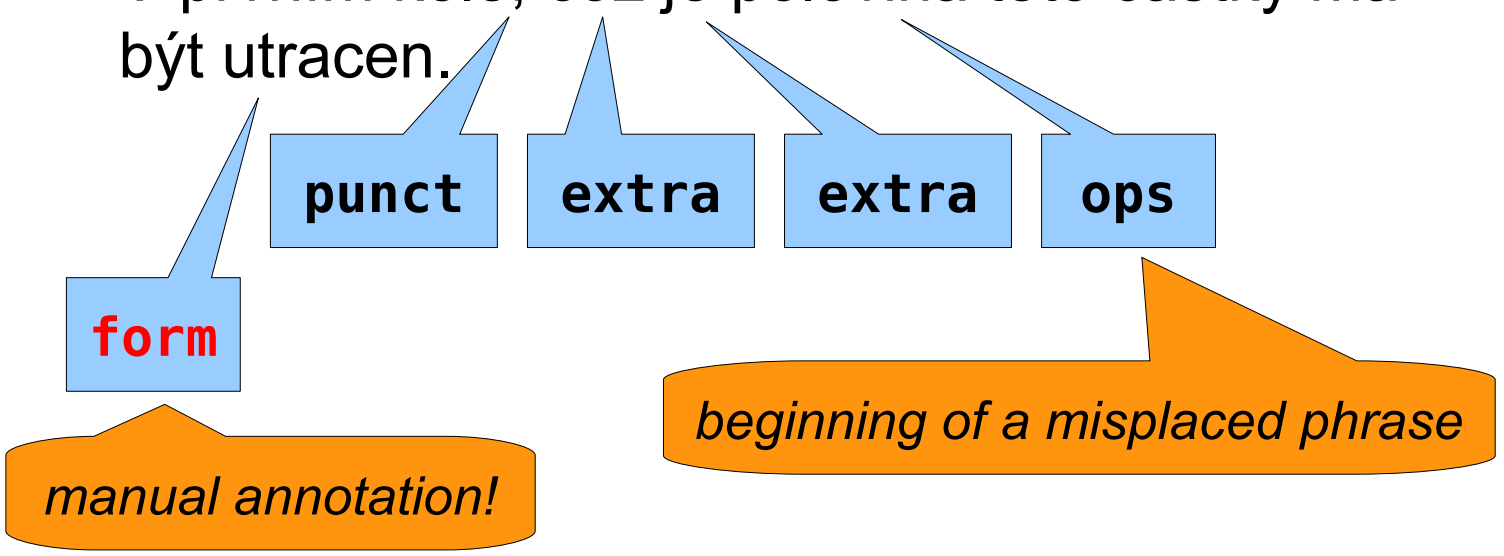
Vilar's Classification

Vilar et al., LREC 2006 [here simplified]

- Missing words
  - Content
  - Filler
- Word order
  - Word misplaced
    - Local
    - Long-distance
  - Phrase misplaced
    - Local
    - Long-distance
- Incorrect words
  - Sense
    - Incorrect disambiguation
    - Wrong lexical choice
  - Form (inflection)
  - Extra word
  - Style
  - Idiom
- Unknown words (OOV)
  - Unknown stem
  - Unseen form

Example (en → cs)

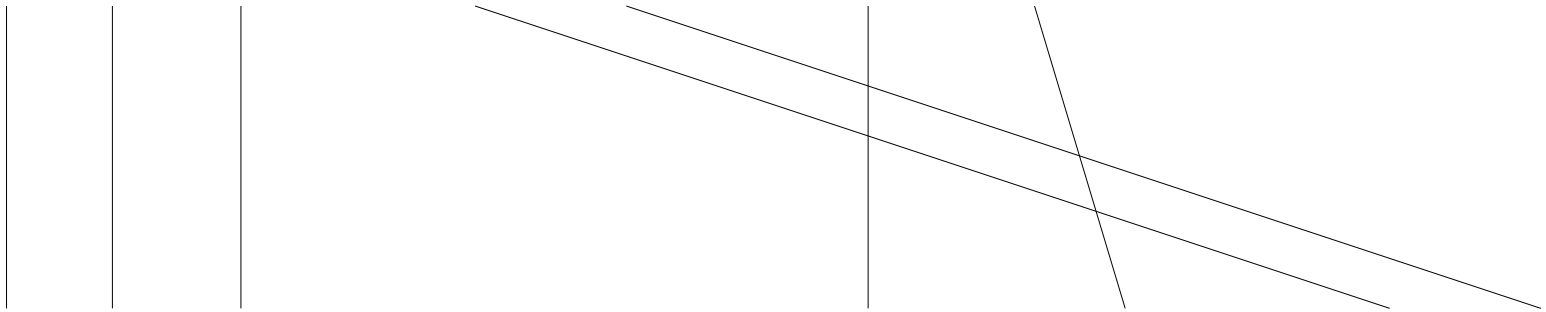
- IN: In the first round, half of the amount is planned to be spent.
- REF: V prvním kole bude použita polovina částky.
- GLOSS: *In the-first round will-be used half of-amount.*
- SYS: V prvním kole, což je polovina této částky má být utracen.



Automatic Error Analysis

- **Monolingual word alignment between:**
  - The reference translation
  - The hypothesis output by the system

V prvním kole	bude použita	polovina částky.	

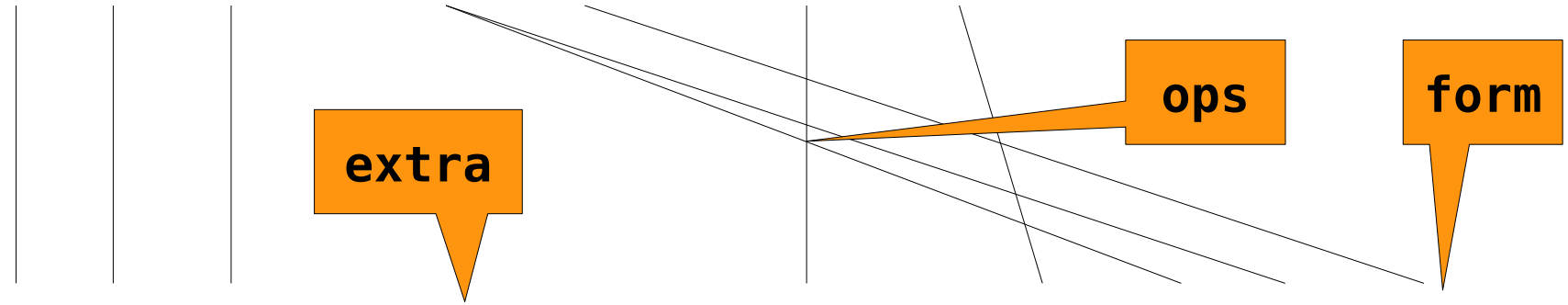


V prvním kole	což je	polovina této částky	má být utracen.

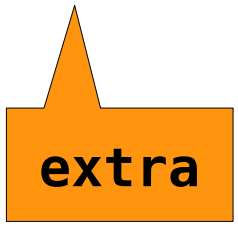
Automatic Error Analysis

- **Monolingual word alignment between:**
  - The reference translation
  - The hypothesis output by the system

V prvním kole	bude použita	polovina částky.	



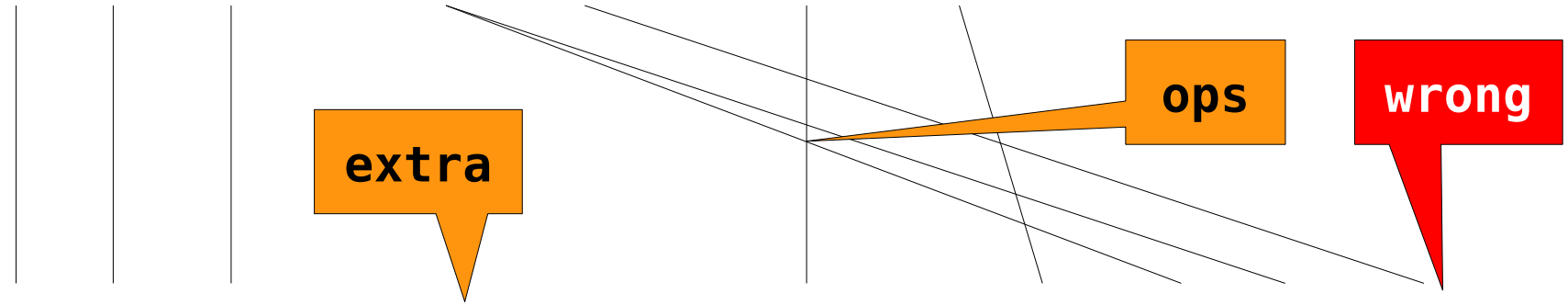
V prvním kole	což je	polovina této částky	má být utracen.



Automatic Error Analysis

- **Monolingual word alignment between:**
  - The reference translation
  - The hypothesis output by the system

V prvním kole	bude použita	polovina částky.	



V prvním kole	což je	polovina této částky	má být utracen.

extra

How to Align Words?

- There are numerous approaches to **bilingual** word alignment (GIZA++, Berkeley aligner, heuristics…)
- **Monolingual** alignment is easier
- Our lightweight approach:
  - Injective (any word linked max once)
  - Key idea: **align identical words**
    - **or lemmas**
  - Ambiguities: repeated tokens (punctuation, function words…)
    - Solution: first-order Markov dependency, i.e. reward adjacent words aligning to adjacent words

Alternative: Bilingual Alignment

- Via source language
- Use full training data (= large vocabulary)
- Run GIZA++ or other existing aligner
- Align hypothesis to source
- Align reference to source
- Assume transitivity, project links: hypothesis to reference

- No more injective, typically much slower
- But more robust: higher recall, lower precision

Error Labeler

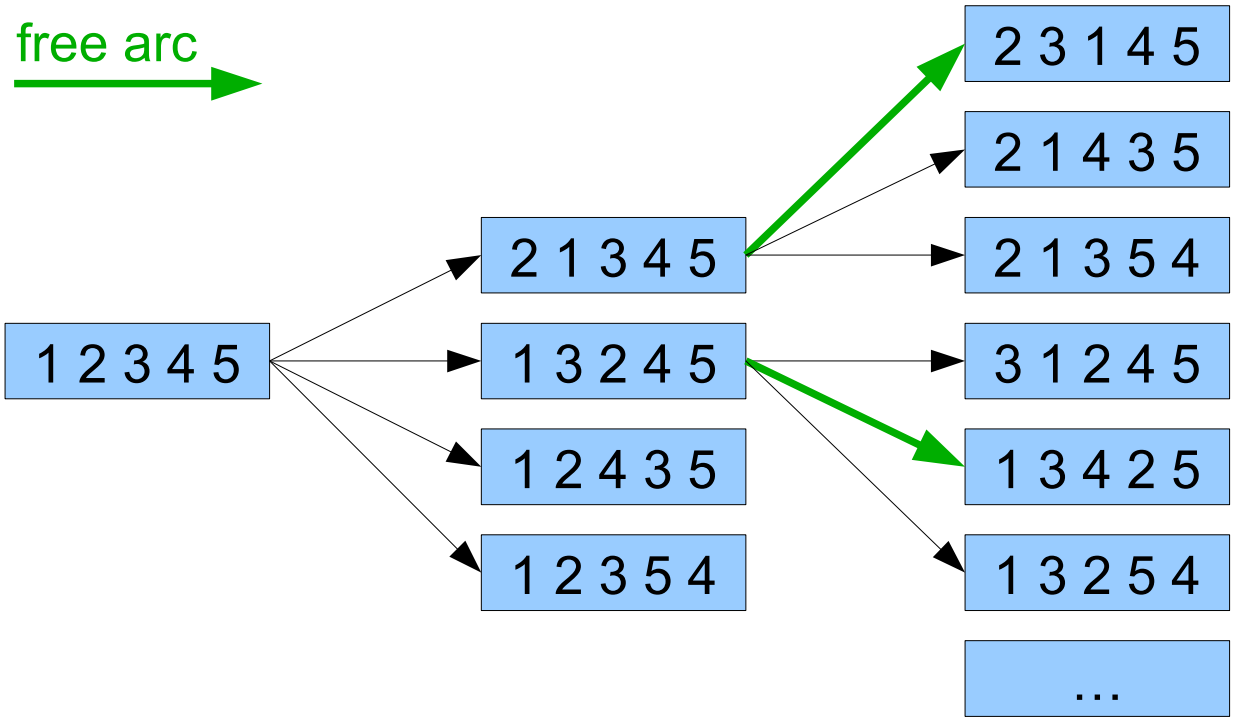
- Unaligned ref words → missing
  - Use POS tag, distinguish content words vs. auxiliaries
- Unaligned hyp words
  - Present in source → unknown
  - Otherwise → extra
- Aligned, same lemma, different form → wrong form
- Aligned, different lemma: synonym or wrong sense?
  - Future work (WordNet?)
- Aligned, differing in punctuation → punct
- Higher level errors (idioms etc.) currently not covered

Order Errors

- Weighted directed graph of hypothesis permutations
  - Nodes = permutations
  - Arc P1 → P2 iff P2 differs from P1 by 2 adjacent symbols
    - Default arc weight is 1
    - Weight is 0 if the arc continues shifting a token in the same direction
      - Helps with block shifts

- Breadth-first search
  - Ignore unaligned words
  - Find the cheapest path to a permutation that corresponds to the reference (i.e. all alignment links are perpendicular)

Breadth-First Search



Order Errors

- Permutation found? So we know:
  - Swapped adjacent words → local reordering
  - Other misplaced words → long-distance reordering

- The approach is word-based. No phrase reorderings.

Evaluation

- Manually annotated English-Czech data
  - See Bojar (PBML 95, 2011)
- Error labeler tested with different alignment methods:
  - Our lightweight monoling (“Addicter”)
  - Meteor monoling adapted to Czech
  - Via source (using Giza++)
  - Existing bilingual aligners used monolingually, enforced injectivity (Addicter search space = proposed alignment):
    - Giza++
    - Berkeley

Numbers numbers numbers…

Alignment Method	Alignment			Translation Errors		
	Prec	Rec	AER	Prec	Rec	F-score
addicter & source	86.39	85.89	13.86	15.27	54.06	23.82
addicter	98.89	72.18	16.55	10.36	43.76	16.75
addicter & meteor	97.90	71.54	17.33	10.38	43.78	16.78
addicter & giza++$_{\text{intersection}}$	85.99	77.78	18.32	13.47	49.61	21.18
addicter & berkeley & source	73.67	83.50	21.72	16.91	54.39	25.80
addicter & berkeley	71.23	78.31	25.40	15.38	52.02	23.74
addicter & giza++$_{\text{grow-diag}}$	65.93	74.58	30.01	14.71	48.56	22.58
via source	85.00	74.60	20.54	13.80	54.90	22.06
giza++$_{\text{intersection}}$	81.65	64.09	28.19	11.82	48.11	18.97
berkeley*	68.12	74.38	28.89	15.16	51.56	23.43
meteor	90.37	55.04	31.59	6.08	28.68	10.04
giza++$_{\text{grow-diag}}$*	61.54	69.95	34.52	14.50	47.99	22.27

Numbers numbers numbers…

Alignment Method	Alignment			Translation Errors		
	Prec	Rec	AER	Prec	Rec	F-score
addicter & source			13.86			23.82
addicter			16.55			16.75
addicter & meteor			17.33			16.78
addicter & giza++ $_{\text{intersection}}$						
addicter & berkeley & source			21.72			25.80
addicter & berkeley			25.40			23.74
addicter & giza++ $_{\text{grow-diag}}$						
via source						
giza++ $_{\text{intersection}}$						
berkeley*						
meteor						
giza++ $_{\text{grow-diag}}$ *						

Discussion

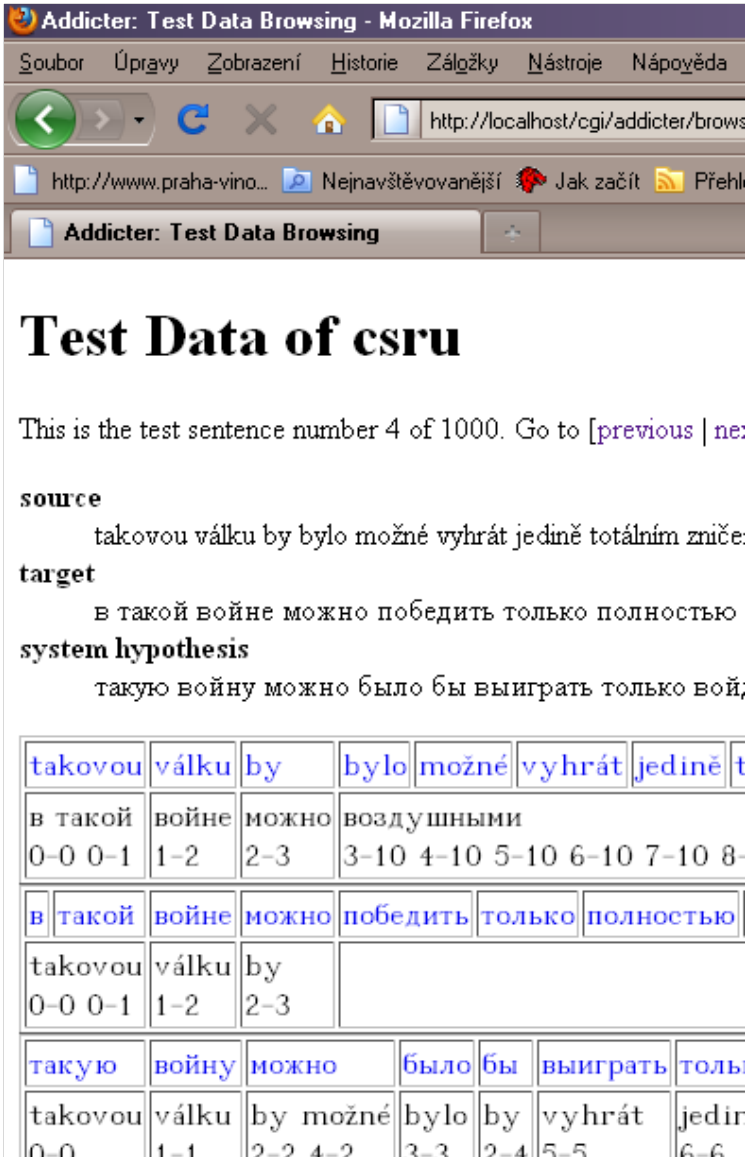
- Error detector tends to overkill
  - High recall, low precision
  - Holds for most error types
- Alignment error rate does not correlate with our F-score
  - Injective alignments are bad?
  - Word-level view is bad?

  - Future work: more focus on the phrase level!

Bonus: Addicter Corpus Browser

https://wiki.ufal.ms.mff.cuni.cz/user:zeman:addicter

- ADDICTER = Automatic Detection and DIsplay of Common Translation Errors
- A tool for
  - Browsing test corpus (source + reference + hypothesis)
  - Finding examples of aligned word pairs
    - In test and training corpus
  - Finding occurrences in Moses/Joshua phrase table
  - Alignment viewing
  - Alignment summarizing



Thank you

Děkuji

The research has been supported by the grants P406/11/1499, P406/10/P259, SF0180078s08.