Handling Overlapping Parallel Corpora

Mark Fishel, Heiki-Jaan Kaalep

University of Tartu, Estonia

Overview

- Overlapping parallel corpora?
- Handling them?
- Implementation
- Experiments
  - Corpora analysis
  - MT

Overlapping Parallel Corpora?

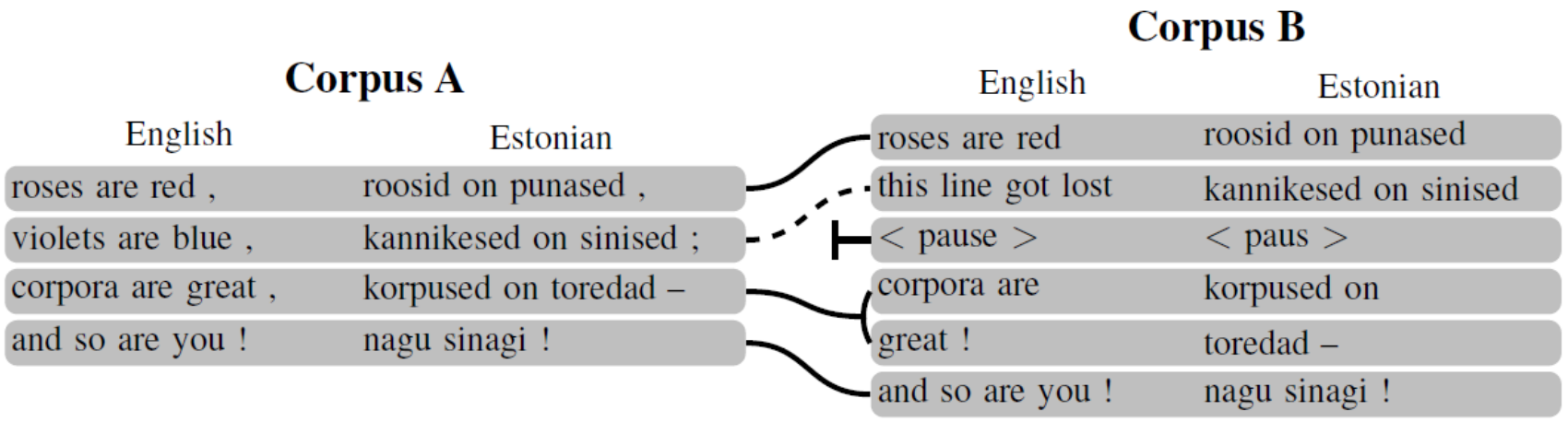


Figure 2.1: An example of overlapping parallel corpora with the correspondence of the two corpora shown. Second sentence pair of corpus B is an erroneous alignment.

Source

- Same source documents, corpora created independently
- Same corpus aligned independently

Problems

- Minor text differences
  - Typos corrected/added
  - Special symbols handled differently
- Different sentence alignment depths
- Added/omitted sentence pairs

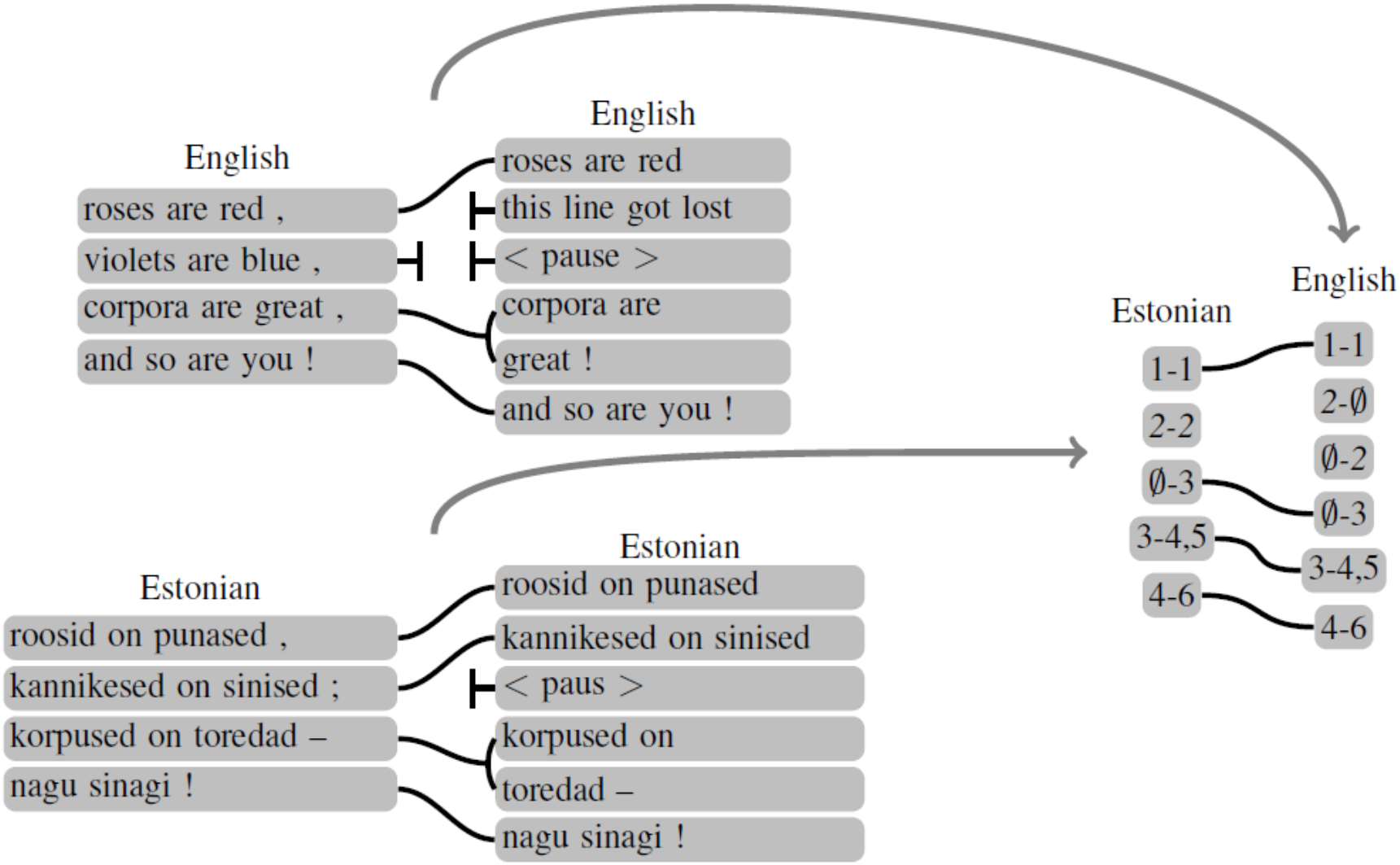
Benefits

- Increase segmentation depth
- Find potential sentence alignment error spots
- Combine corpora
- Check/improve one corpus by comparison to the other

Some examples from real life

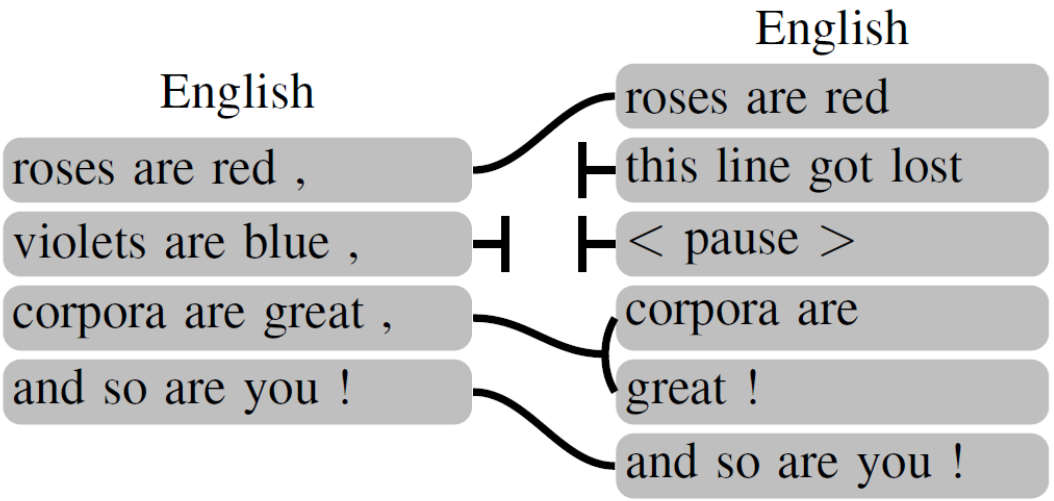
- JRC-Acquis corpus
  - Aligned with Vanilla and HunAlign alternatives
- Hunglish and JRC-Acquis
- CzEng and JRC-Acquis
- SUBTITLES
  - CzEng, Hunglish, OPUS
  - Kind of a special case

Method of processing

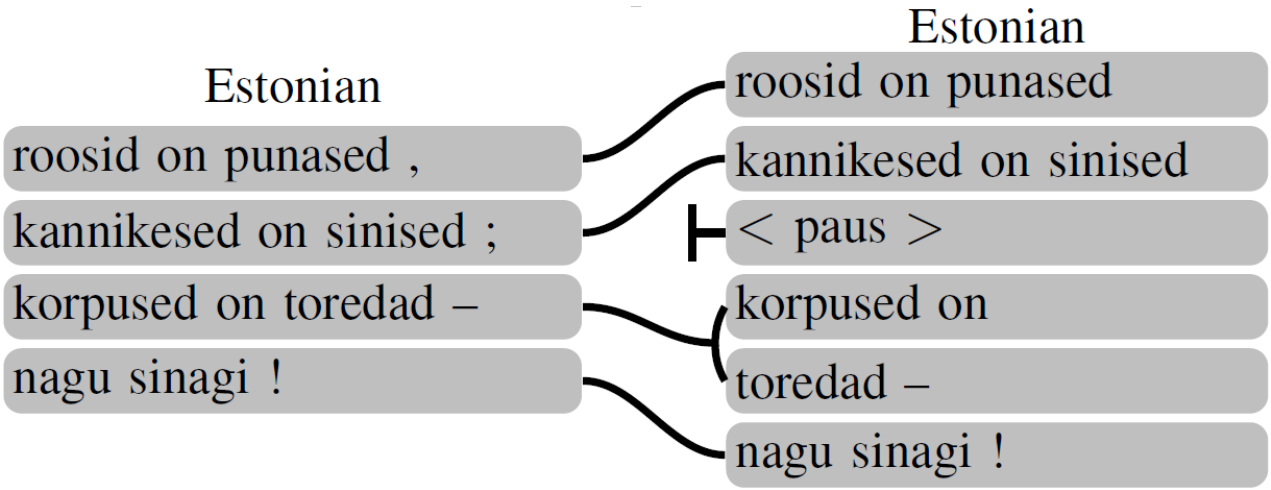


Method of processing

- Align language-parts independently
  - N-to-M edit distance sentence matching
    - Adding/omitting weight=1
    - Replacement weight = sentence pair distance



- Sentence distance = approximate matching with general edit distance
  - Weight(“,” -> “.”) = small
  - Weight(“D” -> “d”) = small
  - Weight(“3”->”6”) = really big

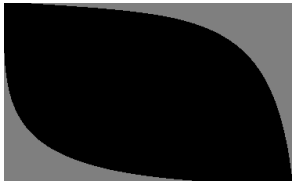


Optimization

- Head & tail

- Anchor-points

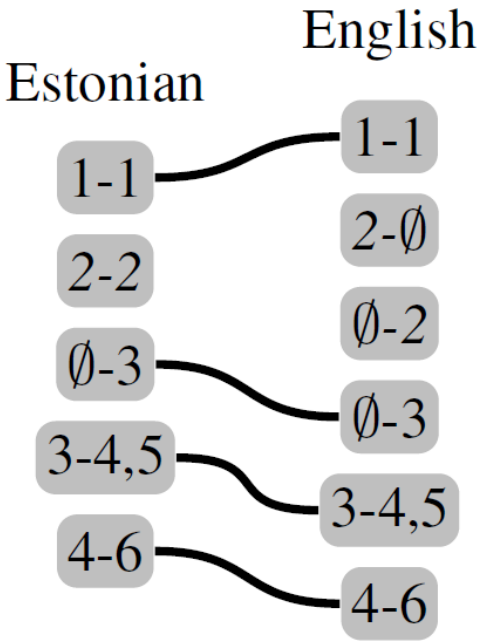
- Trimming the corners



- Traverse with a “front”,
  quit if threshold exceeded

Method of processing

- Align the alignments
  - Simple Levenstein distance



Implementation: CorporAl

- Aligns parallel corpora to each other
- Having an alignment
  - Outputs it or
  - Uses it to generate a combined corpus

Combining corpora

- Requires exact behavior specified
  - Include snt. pairs from just one corpus?
  - Include snt. pairs that match?
    - 3 sentences matched 2 – include what?
  - Include snt. pairs that did not match and how?
    - mismatch consists of 3 vs 2 sentences – include what?

Combining corpora

- Requires exact behavior specified
  - Include snt. pairs from just one corpus?
  - Include snt. pairs that match?
    - 3 sentences matched 2 – include what?
  - Include snt. pairs that did not match and how?
    - mismatch consists of 3 vs 2 sentences – include what?
- Max-size vs Max-accuracy

Corpora analysis

- UT vs JRC corpus (Est-Eng)

UT+JRC2, et-en	#docs	#snt pairs	#en words	#et words
Just UT	2048	134684	$3.12 \cdot 10^6$	$2.17 \cdot 10^6$
Just JRC2	5807	205025	$4.86 \cdot 10^6$	$3.25 \cdot 10^6$
Common UT	2009	93152	$1.88 \cdot 10^6$	$1.27 \cdot 10^6$
Common JRC2	2009	68165	$1.67 \cdot 10^6$	$1.09 \cdot 10^6$
Max-size	2009	98946	$2.03 \cdot 10^6$	$1.36 \cdot 10^6$
Max-acc	2009	56234	$1.35 \cdot 10^6$	$0.88 \cdot 10^6$

	UT	JRC2
$\emptyset$	7.12%	9.89%
0-1	0.00%	8.25%
1-0	32.57%	0.00%
1-1	59.30%	81.04%
1-2	0.06%	0.17%
2-1	0.91%	0.62%
2-2	0.00%	0.00%
3-1	0.01%	0.00%

Corpora analysis

- JRC HunAline vs Vanilla (Est-Eng-Lat, Ger-Eng)

JRC3, de-en	#docs	#snt pairs	#de words	#en words
Just Hun	4	66148	$0.84 \cdot 10^6$	$0.80 \cdot 10^6$
Just Van	83	3716	$0.11 \cdot 10^6$	$0.08 \cdot 10^6$
Identical	14733	614199	$13.79 \cdot 10^6$	$15.03 \cdot 10^6$
Common Hun	8598	658532	$15.75 \cdot 10^6$	$16.97 \cdot 10^6$
Common Van	8598	621816	$15.65 \cdot 10^6$	$16.94 \cdot 10^6$
Max-size	8598	658583	$15.75 \cdot 10^6$	$16.97 \cdot 10^6$
Max-acc	8072	575749	$14.19 \cdot 10^6$	$15.67 \cdot 10^6$

	JRC3 de-en	
	Hun	Van
$\emptyset$	11.9%	7.8%
0-1	0.0%	0.0%
1-0	0.6%	0.0%
1-1	86.7%	91.8%
1-2	0.0%	0.0%
2-1	0.8%	0.4%
2-2	0.0%	0.0%

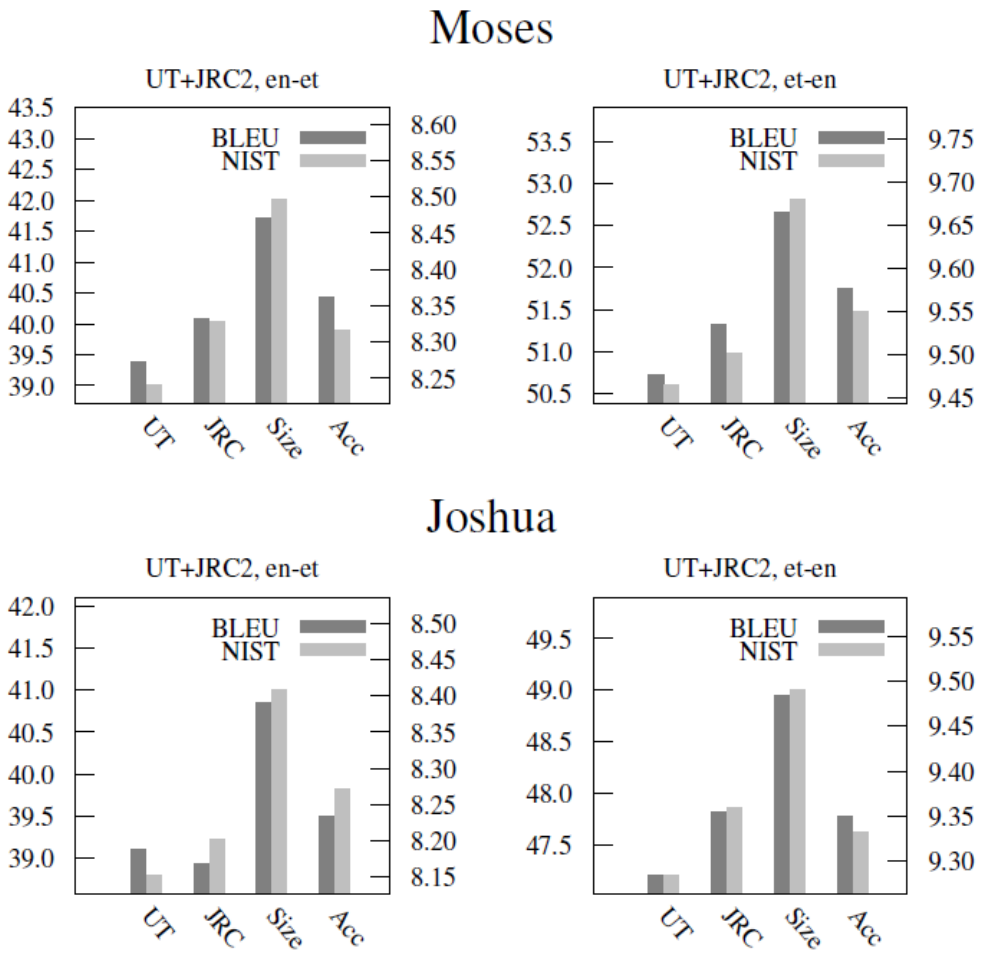
Influence on MT

- Overlapping corpora cannot be concatenated
  - data distribution gets skewed
  - freq. of the samples present in both parts increased relative to everyone else
- Baseline
  - snt. pairs from just corpus A +
    snt. pairs from just corpus B +
    snt. pairs from the overlapping part of
    either corpus B or corpus A

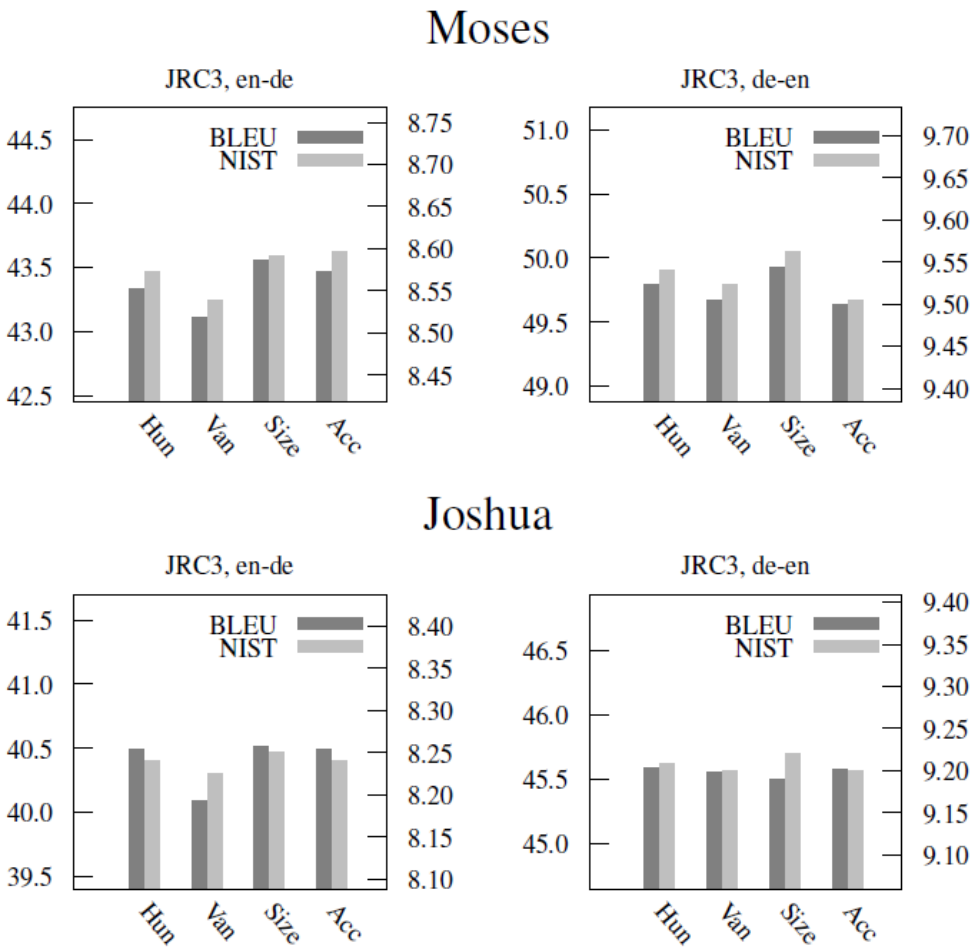
Experiment setup

- Baseline-1 and baseline-2 (from both corpora)
- vs max-accuracy and max-size
- Moses and Joshua default
- MERT
- GIZA++ default
- SRI LM 5-gram Kneser-Ney discounting
- 2500 snt. pairs in dev and test sets

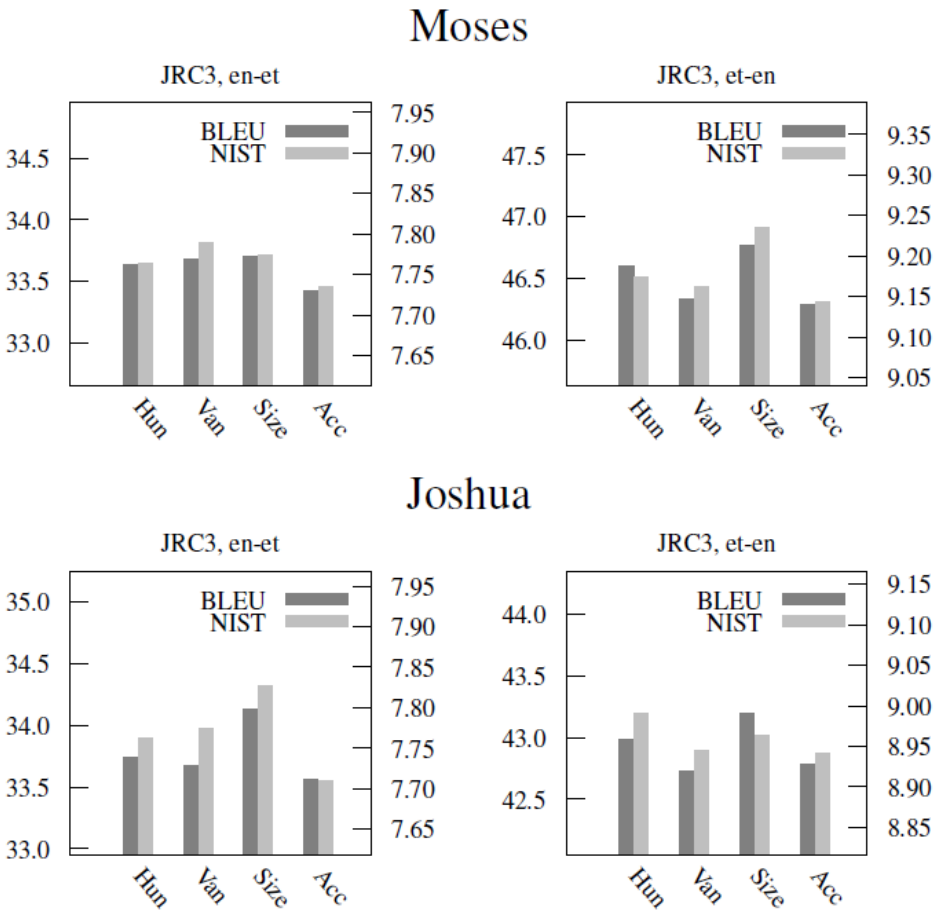
Influence on MT



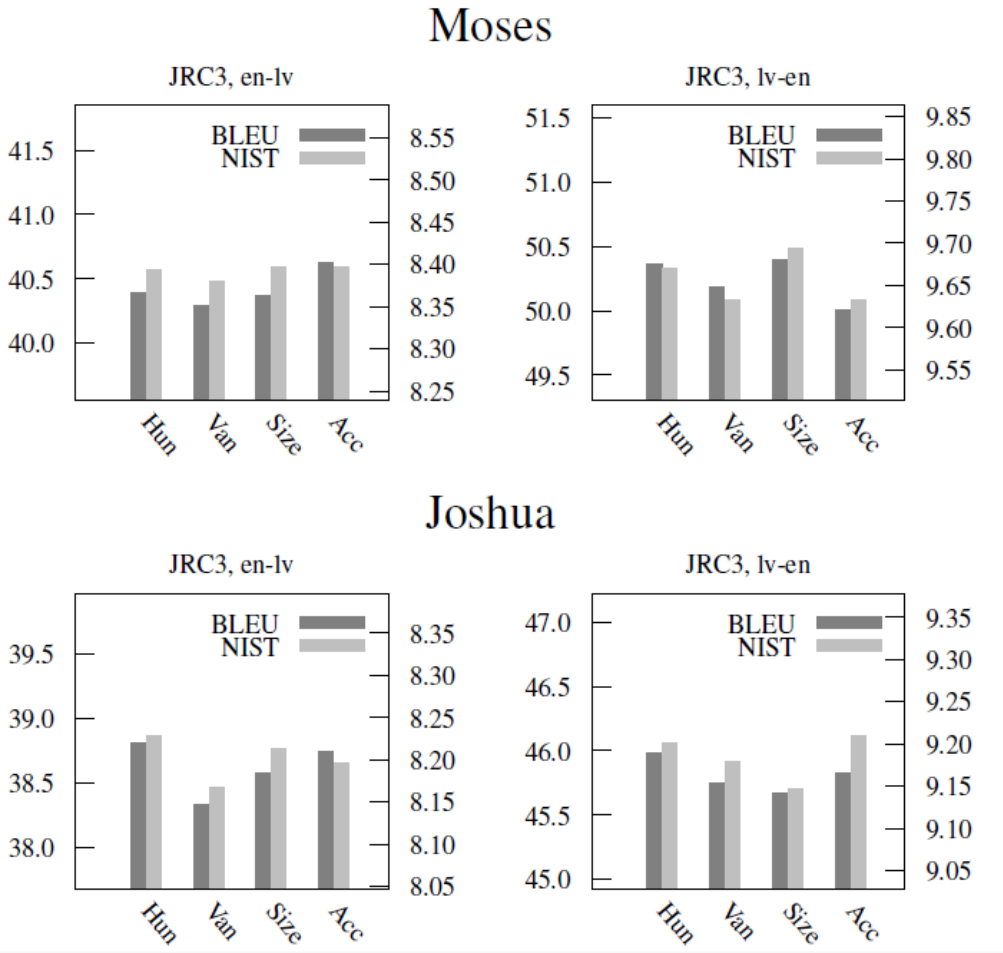
Influence on MT



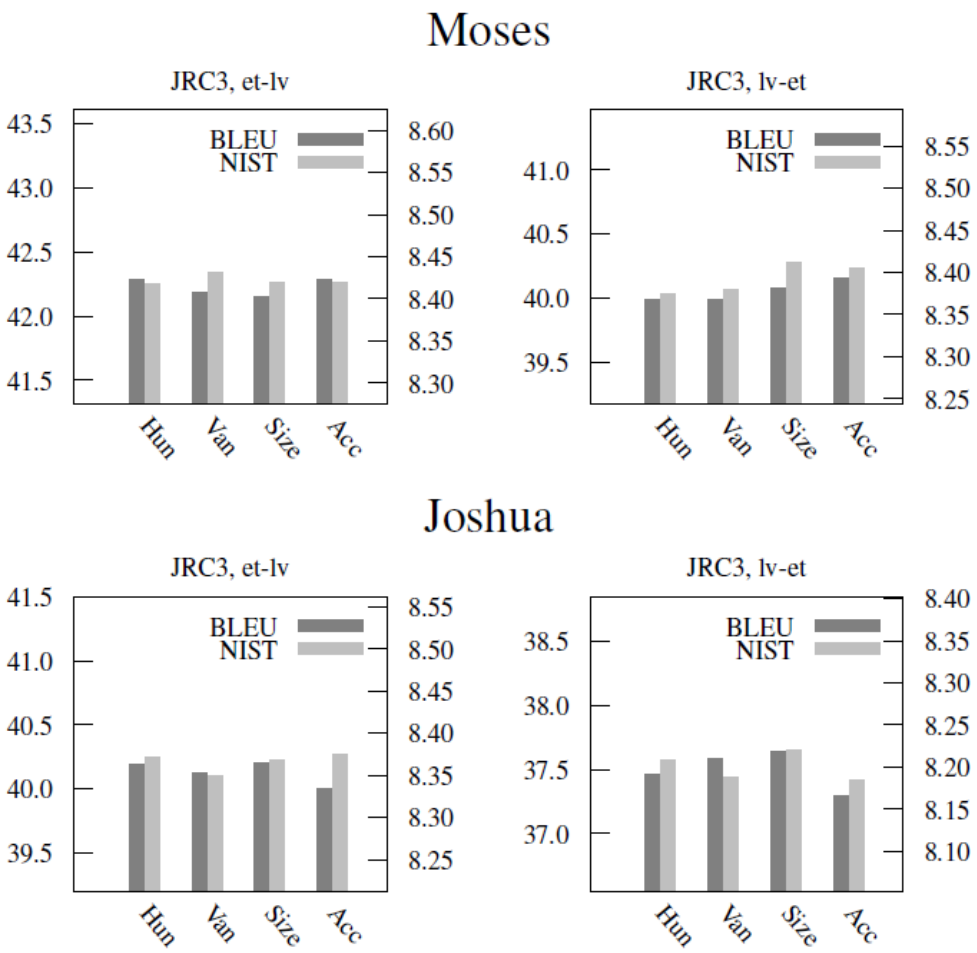
Influence on MT



Influence on MT



Influence on MT



Implementation

- PERL script
- corporal.sf.net

Future work

- Currently matches both language parts and looks for matches/mismatches
- Could be used to generate a Greek-German Europarl
- Extend to non-parallel corpora
  - treat text as language-1 and markup as language-2
  - combine OR
  - generate e.g. corpus, annotated morphologically AND syntactically

Thank you!