Slavic Forest,

Rudolf Rosa, **Dan Zeman**, David Mareček and Zdeněk Žabokrtský

Charles University, Prague, Czechia

The Task

- Train on language 1, parse related language 2
  -  Czech → Slovak 
  -  Slovenian → Croatian 
  -  Danish +  Swedish → Norwegian 
- Parallel data available
  - Movie subtitles (Open Subtitles 2016 from OPUS)

Closely Related Languages

- Translation can be almost word-by-word

Closely Related Languages

- Translation can be almost word-by-word
- Not necessarily found in subtitles (independent translations from a third language)
- Word-by-word translation of a treebank?

Closely Related Languages

- Translation can be almost word-by-word
- Not necessarily found in subtitles (independent translations from a third language)
- Word-by-word translation of a treebank?

- cs: *Našemu lidu hrozí velké nebezpečí.*
- sk: *Nášmu ľudu hrozí veľké nebezpečenstvo.*

- sl: *Še nikoli nisem videl česa takšnega.*
- hr: *Nikad nisam vidio nešto takvo.*

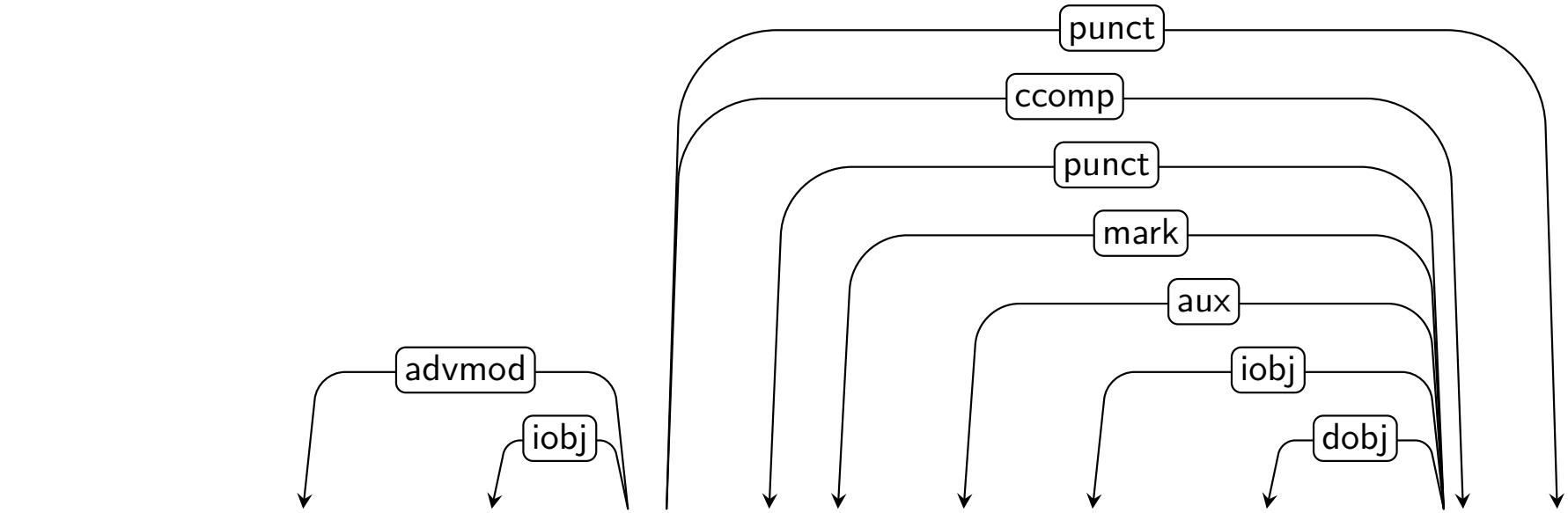
Closely Related Languages

- Translation can be almost word-by-word
- Not necessarily found in subtitles (independent translations from a third language)
- Word-by-word translation of a treebank?

- da: *London er verdens mest overvågede by.*
- no: *London er jordens mest overvåkede by.*

- sv: *Jag tänkte ducka och få ett bra spår in i nästa kurva.*
- no: *Jeg og få bare gå på mot neste sving.*

Word-by-Word Translation



cs:	Telefonicky	ho	požádal	,	aby	by	mu	dokument	poslal	.
sk:	Telefonicky	ho	požiadal	,	aby	by	mu	dokument	poslal	.
sk_{trans}:	telefón	ho	požiadal	,	aby	by	mu	dokument	poslal	.
en:	By-phone	him	asked	,	to	would	him	document	send	.

The Parser

- UDPipe
  - `http://ufal.mff.cuni.cz/udpipe`
  - baseline
  - “in-house” (by Milan Straka)
- Other parsers might perform better
  - we have not tested many parsers
  - our tweaks are mostly parser-independent

Training Data Sizes

Language	Sentences	Words
Czech	68,495	1,173,282
Slovak	8,483	80,575
Slovenian	6,471	112,334
Croatian	5,792	127,849
Danish	4,868	88,979
Swedish	4,303	66,645
Norwegian	15,696	243,887

- Target language training data
  - cannot be used by participants
  - except the tagger trained on them
  - but useful to compare supervised parsers

The Data

- Universal Dependencies 1.4
  - Same POS tags
  - Similar features
  - Same or similar dependency relations
  - Translation counterparts should have parallel trees
- Well, not quite...

Normalization

- Examples from Czech-Slovak
  - Dependency relations
    - ★ det, det:nummod, det:numgov → nmod
    - ★ nummod:gov → nummod
    - ★ remove sentences with compound, discourse, vocative (rare in Czech, not in Slovak, not easily replaceable)

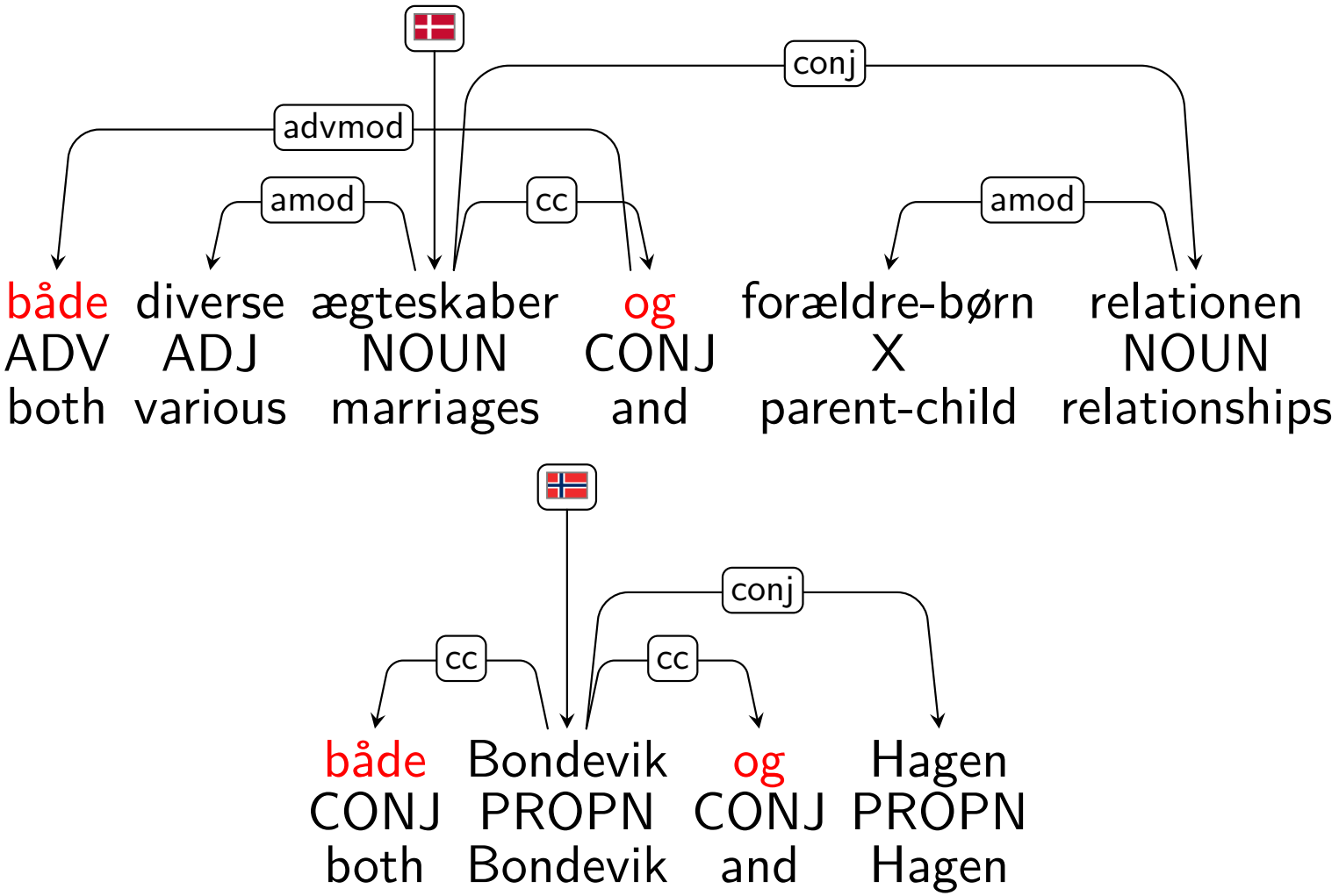
Normalization

- Examples from Czech-Slovak
  - Dependency relations
    - ★ det, det:nummod, det:numgov → nmod
    - ★ nummod:gov → nummod
    - ★ remove sentences with compound, discourse, vocative (rare in Czech, not in Slovak, not easily replaceable)
  - Part-of-speech tags
    - ★ SCONJ → CONJ
    - ★ (DET → PRON) ... not in final submission

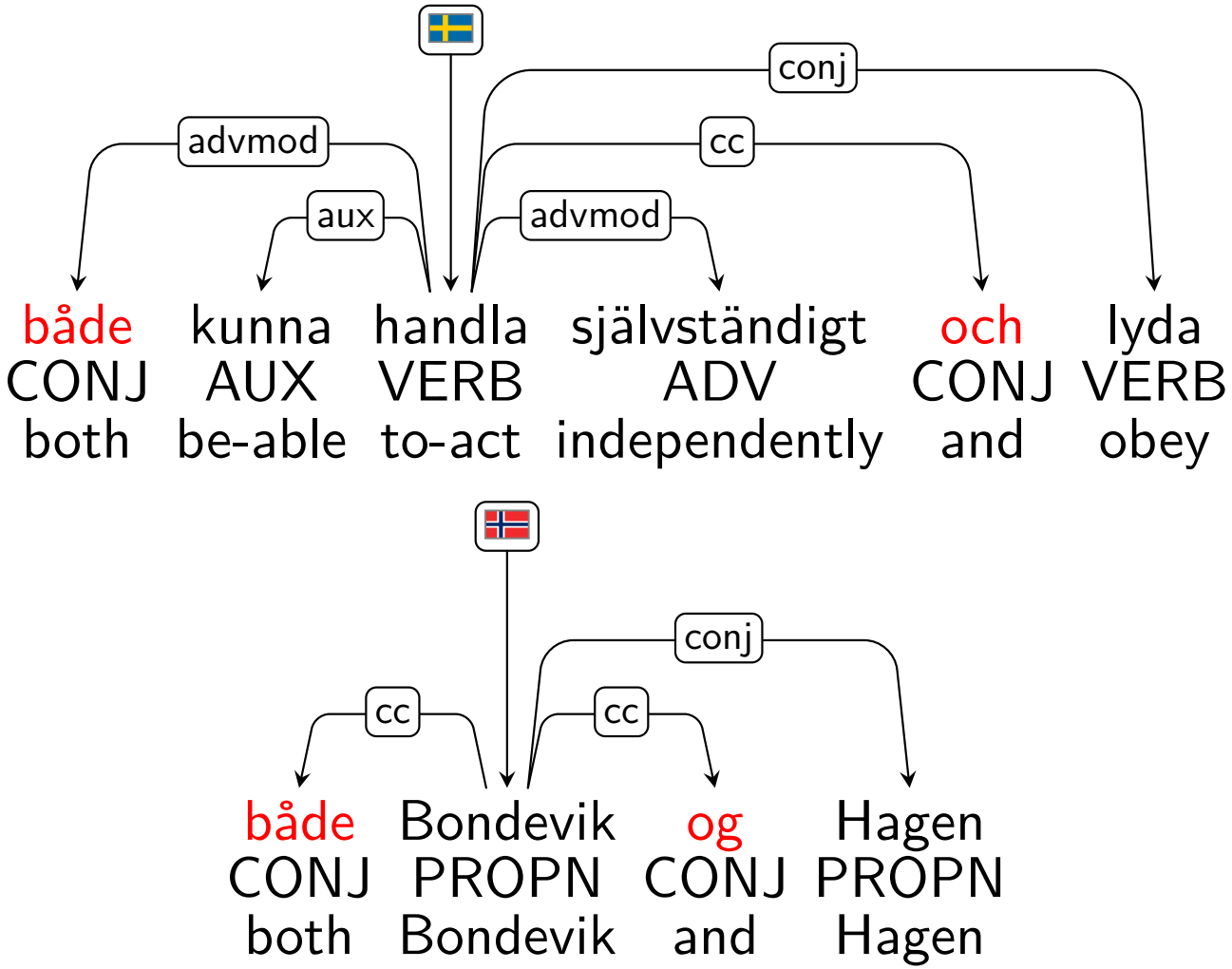
Normalization

- Examples from Czech-Slovak
  - Dependency relations
    - ★ det, det:nummod, det:numgov → nmod
    - ★ nummod:gov → nummod
    - ★ remove sentences with compound, discourse, vocative (rare in Czech, not in Slovak, not easily replaceable)
  - Part-of-speech tags
    - ★ SCONJ → CONJ
    - ★ (DET → PRON) ... not in final submission
  - Features
    - ★ Remove ConjType, Gender[psor], NameType, Number[psor], NumType, NumValue, Poss, PrepCase, Style, Variant

Normalization of *både*

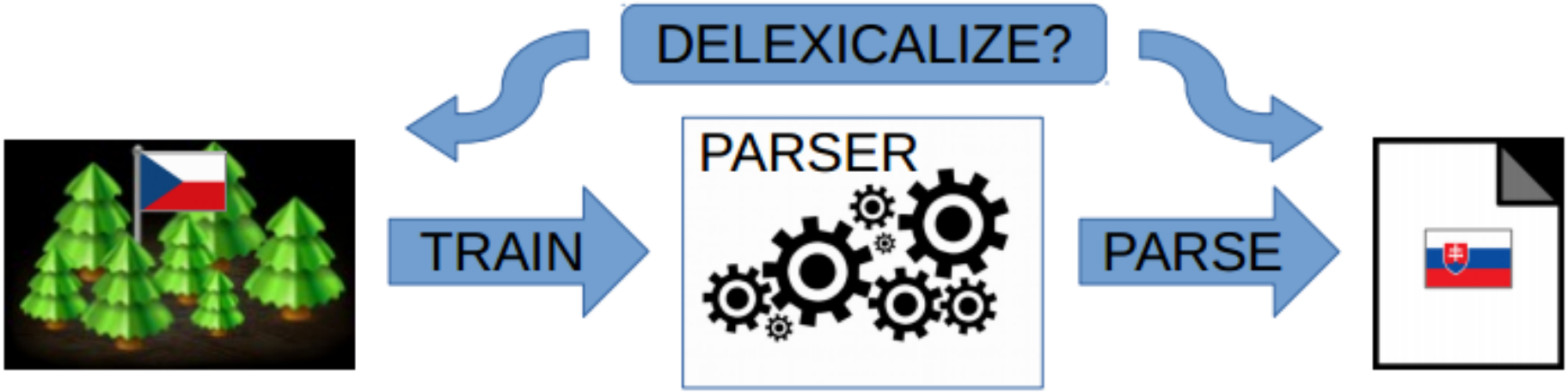


Normalization of *både*



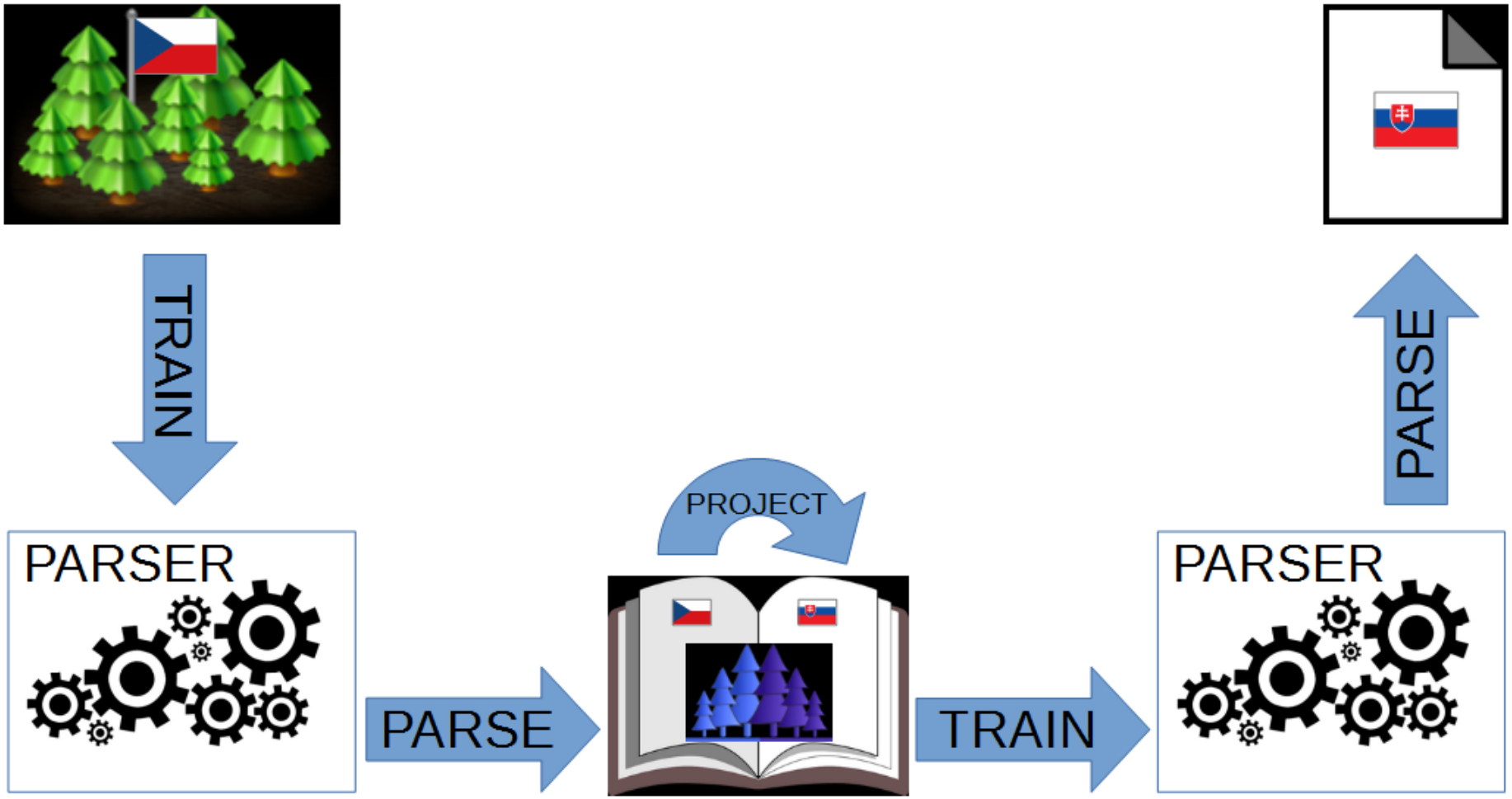
Possible Approaches

- **Direct:** train parser on source treebank; declare it a target parser.



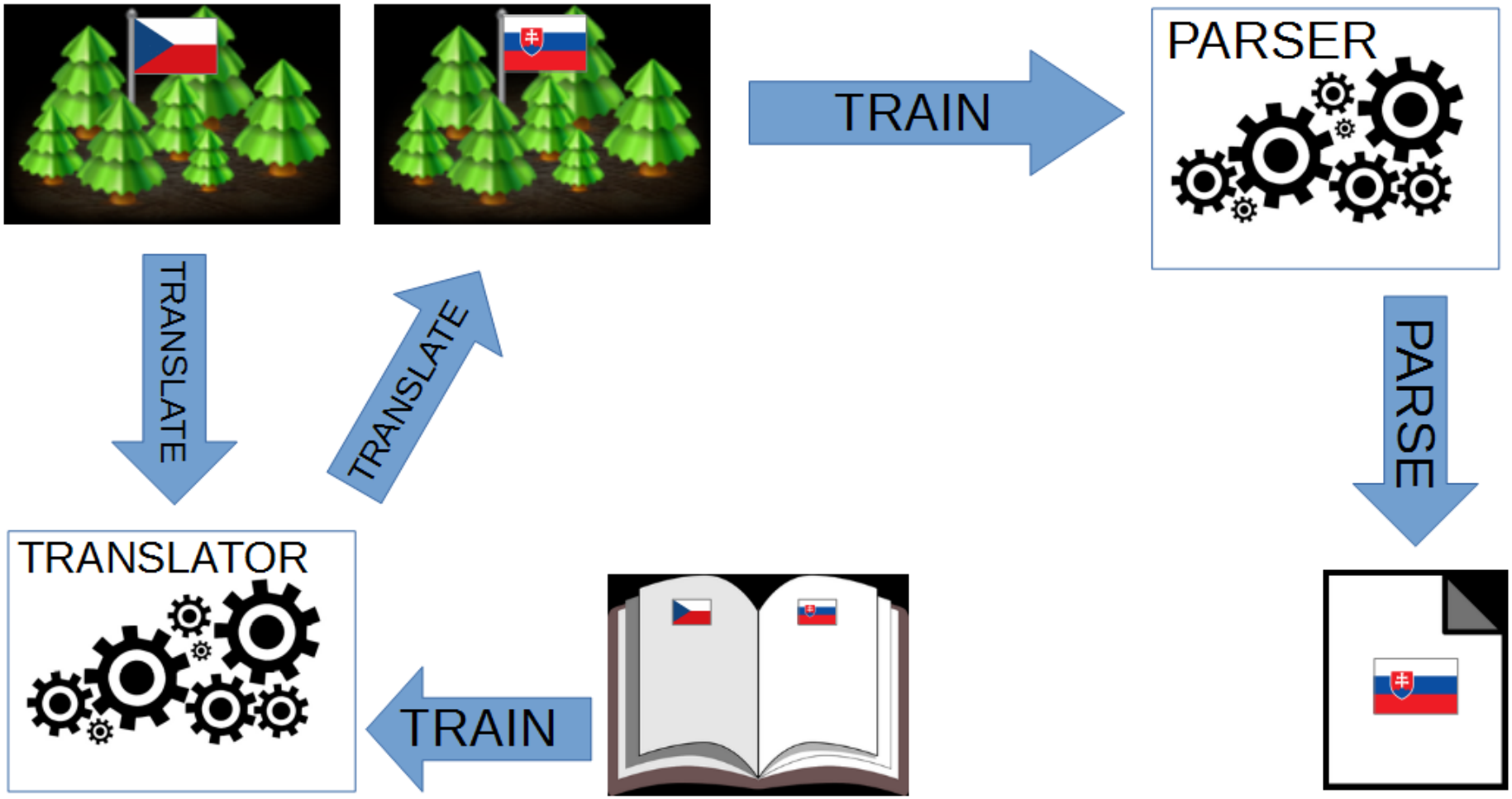
Possible Approaches

- **Project:** train parser on source treebank; parse source side of bitext; project trees to target side; train target parser on that.



Possible Approaches

- **Translate:** train translation on source-target bitext; translate source treebank to target language (only works because of 1-1 translation!); train target parser on that.



Cross-Tagging

- **Baseline:** tag target test data by a supervised target tagger (hopefully similar enough to the source tags the parser was trained on; used for sl-hr)

Cross-Tagging

- **Baseline:** tag target test data by a supervised target tagger (hopefully similar enough to the source tags the parser was trained on; used for sl-hr)
- **Translate + target tagger:** translate source treebank to target language, tag it by supervised target tagger, train the parser on it; tag the test data by the same tagger (called *source-xtag* in the paper; used for da/sv-no)

Cross-Tagging

- **Baseline:** tag target test data by a supervised target tagger (hopefully similar enough to the source tags the parser was trained on; used for sl-hr)
- **Translate + target tagger:** translate source treebank to target language, tag it by supervised target tagger, train the parser on it; tag the test data by the same tagger (called *source-xtag* in the paper; used for da/sv-no)
- **Translate + source tagger:** translate source treebank to target language, train both a tagger and a parser on it (tagger learns source tags); tag the test data by the same tagger (called *target-xtag* in the paper; used for cs-sk)

Feature Selection

- Only POS tags, or POS + features?
- UDPipe treats all features as one "tag"

- If features, which features?
  - Case (Collins et al., 1999)
  - Consistent across alignment
    - ★ At least 70% occurrences shared

Results

System	SK	HR	NO
LAS			
Ours	78.12	60.70	70.21
Helsinki	73.14	57.98	68.60
diff	*4.98*	*2.72*	*1.61*
UAS			
Ours	84.92	69.73	77.13
Helsinki	82.87	69.57	76.77
diff	*2.05*	*0.16*	*0.36*

Ablation Analysis

Component	SK	HR	NO
Normalize source annotations	2.50	3.11	1.67
Translate word forms	7.04	**5.02**	**6.66**
Pre-train form embeddings	2.83	3.88	5.28
Cross-tag	**11.36**	—	2.92
Add morphological features	2.09	1.70	1.43

Discussion

- Normalization helps, esp. for dependency labels
  - ... but it should be less important in the future

Discussion

- Normalization helps, esp. for dependency labels
  - ... but it should be less important in the future
- Very big data helps
  - cf. Czech-trained vs. Slovak-trained taggers

Discussion

- Normalization helps, esp. for dependency labels
  - ... but it should be less important in the future
- Very big data helps
  - cf. Czech-trained vs. Slovak-trained taggers
- Error analysis is important
  - we did best in our (near-)native language

Děkuji!	Thank You!	Ďakujem!
Otázky?	Questions?	Otázky?
Hvala!		Hvala!
Vprašanja?		Pitanja?
Tak!	Tack!	Takk!
Spørgsmål?	Frågor?	Spørsmål?

Scripts: http://hdl.handle.net/11234/1-1970
Models: http://hdl.handle.net/11234/1-1971