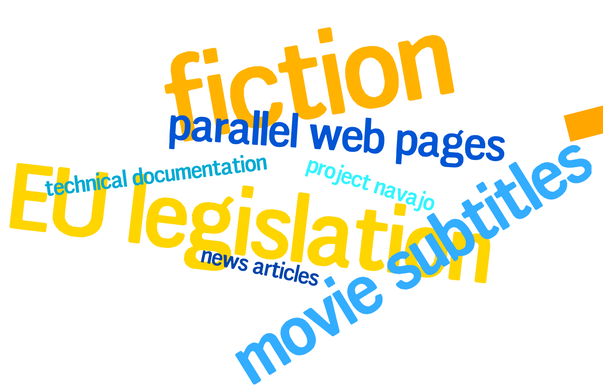
The Joy of Parallelism with CzEng 1.0

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, Aleš Tamchyna
surname@ufal.mff.cuni.cz, *except* {mnovak,odusek}@ufal.mff.cuni.cz, jiri.marsik89@gmail.com
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

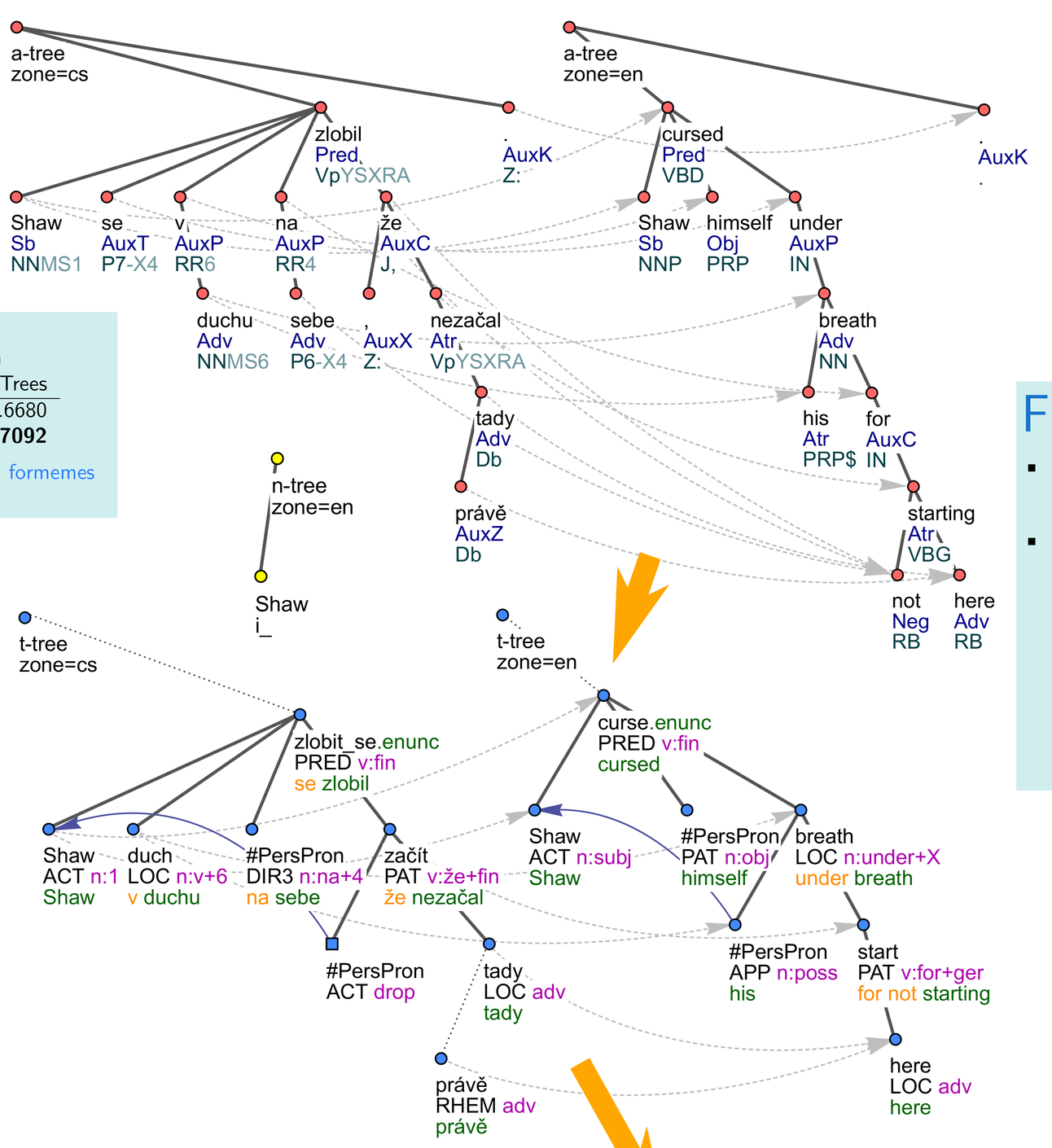
http://ufal.mff.cuni.cz/czeng

CzEng 1.0

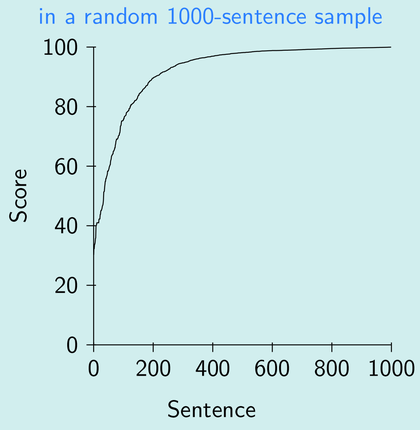
- updated release of our Czech-English parallel corpus
- freely available for non-commercial research and educational purposes
- 15 million sentence pairs (doubled since last release)
- ca. 200 million words per language
- improved automatic rich annotation
  - part-of-speech tags, morphology
  - surface syntax dependency trees (a-)
  - deep syntax dependency trees (t-)
  - co-reference links
- carefully filtered to reduce the amount of non-matching sentence pairs
- automatic sentence and word alignment



Document alignment → Sentence alignment → Rich annotation



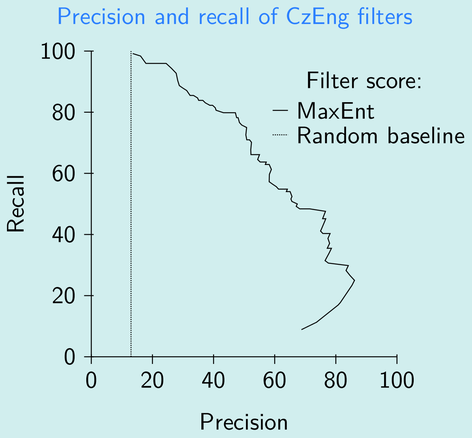
Sentence filter scores in a random 1000-sentence sample



Filtering

- maximum entropy classifier trained to distinguish matching sentence pairs
- features over all layers of annotation:
  - identical words and sentences
  - word and sentence lengths
  - alignment scores
  - dictionary coverage
  - matching grammatical categories
  - matching part-of-speech tags

Precision and recall of CzEng filters



	Formeme Detection on	
	Automatic Trees	Manual Trees
Baseline	1.5981	1.6680
Improved	**1.6873**	**1.7092**

Inter-lingual pointwise mutual information on formemes

Formemes

v:fin n:obj v:for+ger n:na+4

- description of morpho-syntactic form in t-tree nodes, including:
  - syntactic position and usage, morphological case
  - prepositions and conjunctions
- improved version aimed at greater cross-language consistency

Co-reference

- automatic co-reference links in t-trees for both languages
- using rules and perceptron ranking
- grammatical co-reference
  - relative and reflexive pronouns
  - verbs of control
- textual pronominal co-reference

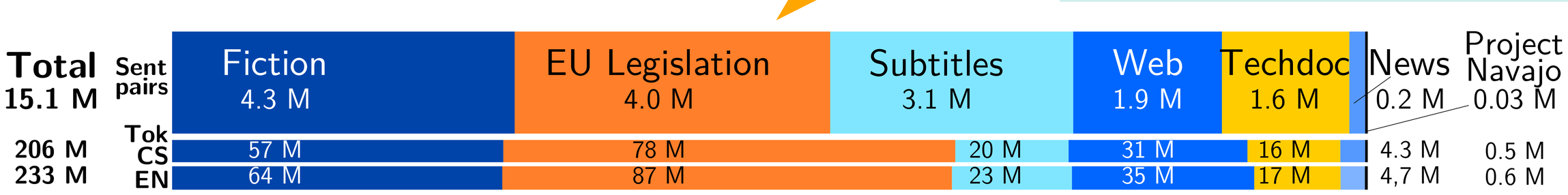
	Gold Standard Features			Automatic Features			Oracle Gender and Number		
Language	P	R	F	P	R	F	P	R	F
Czech	77.06	77.58	77.32	55.23	46.14	50.28	65.70	54.89	59.81
English	45.52	58.69	51.27	44.53	57.32	50.12			

Co-reference resolution evaluation

Filtering

Corpus and Domain		Sents	BLEU	Vocab. [k] En	Vocab. [k] Cs
CzEng 0.9	all	1M	14.77±0.12	187	360
CzEng 1.0	all	1M	**15.23±0.18**	221	396
CzEng 0.9	news	100k	**14.34±0.05**	53	125
CzEng 1.0	news	100k	14.01±0.13	47	113

Results of MT evaluation



Plaintext

```
A  a
fiction-b1-00train-f00001-s1
    0.93220559
    Shaw se v duchu na sebe zlobil, že nezačal právě tady.
    Shaw cursed himself under his breath for not starting
fiction-b1-00train-f00001-s2
    0.93224159
    Nebyli nuceni se přepychu vzdát ani tady, dumal.
    They certainly don't make them like they used to  he m
```

Factored

```
\t  |
fiction-b1-00train-f00001-s1
    0.93220559
    Shaw|Shaw_;S|NNMS1-----A----|1|7|Sb se|se_^(zvr._zájme
    Shaw|ACT|1|4|complex|n:1|n.denot|sg|neg0|-|-|-|-|-|-|-
    0-0 3-1 5-2 6-3 9-5 10-6 11-7 2-1 4-2 1-3 8-5
    Shaw|Shaw|NNP|1|2|Sb cursed|curse|VBD|2|0|Pred himself
    Shaw|ACT|1|2|complex|n:subj|n.denot|sg|-|-|-|-|-|-|-
```

Treex

```
<XML>
<?xml version= "1.0" encoding="UTF-8"?>
<treex_document xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
<head><schema href="treex_schema.xml"/></head><meta/>
<bundles><LM id="fiction-b1-00train-f00001-s1">
<czeng><blockid>b1</blockid><domain>fiction</domain><filte
<zones><zone language="cs"><sentence>Shaw se v duchu na sel
<trees><a_tree id="a_tree-cs-fiction-b1-00train-f00001-s1-
<giza_scores><counterpart.rf>a_tree-en-fiction-b1-00train-
```

This work was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Czech Republic), Czech Science Foundation grants P406/10/P259 and 201/09/H057, GAUK 4226/2011, and the FAUST project (FP7-ICT-2009-4-247762 of the EU and 7E11041 of the Czech Republic).
This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). This poster was presented at LREC 2012.