CzEng 0.7: Parallel Corpus with Community-Supplied Translations

Ondřej Bojar, Miroslav Janíček,
Zdeněk Žabokrtský, Pavel Češka and Peter Beňa
{bojar, zabokrtsky, ceska}@ufal.mff.cuni.cz,
mira.janicek@gmail.com, pitkinpb@yahoo.com

Charles University in Prague
Institute of Formal and Applied Linguistics (ÚFAL)

Outline of the Talk

- Processing pipeline.
- Data sources and licences.
- Community-supplied data:
  - Analysis of quality,
  - Utility for machine translation.
- Future plans.

Text Processing in CzEng 0,7

- Converted data from multitude of input formats,
  . . . HTML, PDF, Palm text, SGML, PO etc.
- Fixed some systematic errors due to the origin of the documents,
  . . . e.g. end-of-line hyphens, spaces in words.
- Removed some parts without counterpart in the corresponding translated/original document,
- Tokenized both Czech and English sides,
- Segmented documents into sentences,
- Aligned Czech and English sentence.
- All text processing done automatically, no manual annotation.

CzEng 0.7 now available in a uniform XML format (or a plaintext dump).

Tokenization & Segmentation: TextSeg

Both Czech and English handled by `TextSeg` by Pavel Češka:

- Decision tree based on 17 context attributes,
- Trained on manually processed data.

Evaluation of sentence segmentation (for Czech):

- 1,000 occurrences of tokens that can indicate the end of sentence
  . . . full-stop, hyphen, question mark, closing bracket, newline symbol.
- Evaluated against human judgements:
  Does the end of sentence coincide with the given token?

→ accuracy 98.4% on this set.

Sentence Alignment

- `Hunalign` to align pairs of segmented and tokenized documents.
- Taking advantage of the documents' structure where possible.
  . . . esp. community-supplied data often in pre-aligned segments.

Distribution of alignment types, English-Czech:

1-1	2-1	0-1	1-2	1-0	Others
1,096,940	68,856	63,185	43,057	30,694	33,123
82.1%	5.2%	4.7%	3.2%	2.3%	2.5%

Texts in CzEng 0.7

Legal texts:

- Acquis Communautaire Parallel Corpus
- The European Constitution proposal from the OPUS corpus
- samples from the Official Journal of the European Union

Stories and Commentaries:

- Readers' Digest stories
- e-books: Project Gutenberg and Palmknihy.cz and a subset of the Kačenka parallel corpus
- articles from Project Syndicate

User-supplied data: . . . not always complete sentences

- Czech localization of KDE and GNOME open-source projects
- user-contributed translations from the Navajo project

Texts in CzEng 0.7 – Data Sizes

	Sentences	Tokens
Acquis Communautaire	64.1%	69.0%
Readers' Digest	8.6%	8.6%
Project Syndicate	6.5%	8.9%
KDE Messages	**6.2%**	**1.9%**
GNOME Messages	**5.7%**	**1.9%**
Kačenka	4.2%	4.9%
Navajo User Translations	**2.3%**	**2.1%**
E-Books	1.2%	1.6%
European Constitution	0.8%	0.7%
Samples from European Journal	0.4%	0.5%
Total	1.4 mil.	21 mil.

Community-supplied data in bold.

Licencing Issues

- Much more data are available on the Internet,
- Only a fraction can be repackaged and distributed.

	Tokens Available			
Source of Texts and Translation	cs	en	cs	en
Community Translation of Proprietary Texts	19.5M	25.3M	37.8%	41.1%
Professional	21.3M	23.9M	41.2%	38.9%
Proprietary	9.6M	10.9M	18.6%	17.7%
Community	1.2M	1.4M	2.4%	2.3%
Total	51.6M	61.5M	100.0%	100.0%

CzEng 0.7 $\approx$ Professional + Community sources; in bold

Community-Supplied Translations (1/2)

The Navajo Project

- An effort of a Czech MT system vendor to improve the quality of their product by means of using a community-supplied corrections of their translation of English Wikipedia to Czech.
- Both sides freely available. ("free" as in "free speech")
- The users (translators) are anonymous.
- About 2,000 segments generated each month.
- CzEng 0.7 contains 30,208 translations of various parts of 3,724 Wikipedia articles.

Quality of Navajo User Translations

The contributors' anonymity rises doubts about the quality of their work.

Manual evaluation of 1,000 randomly selected segments:

Translation Quality	Proportion in the Sample
precise, flawless	**69.0%**
not translated	6.8%
incomplete	6.6%
imprecise	5.8%
precise, almost flawless	**4.5%**
machine-generated	4.4%
vandalism	2.7%
other	0.2%

Community-Supplied Translations (2/2)

KDE and GNOME Localizations

- Two major open-source software projects,
- Contributors **not** anonymous $\Rightarrow$ the quality considerably higher (almost professional)
- Only rarely full sentences, mostly short system messages and user interface elements e.g. "OK", "Yes" or "Delete file"

Achievable MT Quality Using CzEng

Evaluation in-domain vs. out-of-domain.

Disjoint sections of training data:

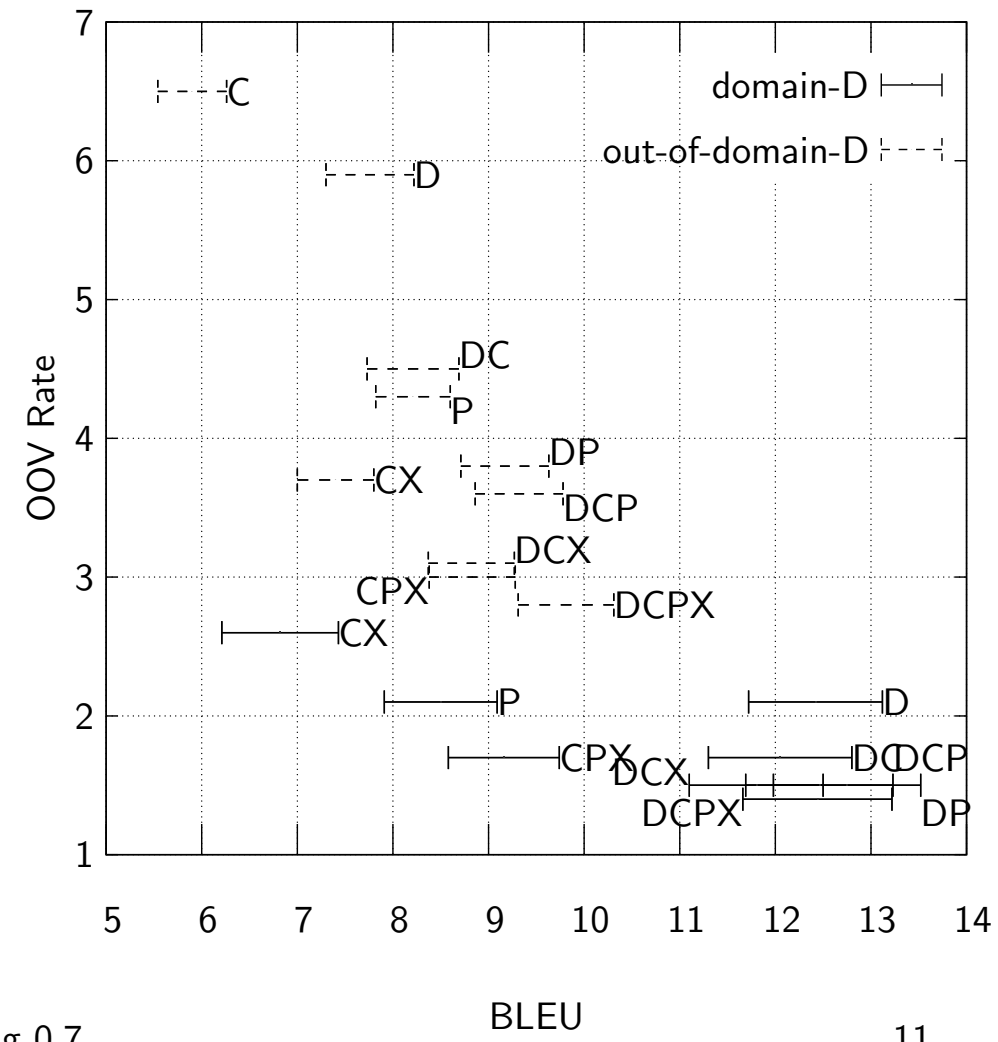
- **D** professional, in-domain
- **P** professional, out-of-domain
- **C** community
- **X** community translation, proprietary source

Observations within domain of **D**:

- **D** sufficient for BLEU score
- additional data reduce OOV

Observations outside domain of **D**:

- **CX** < **DP**
- **DCPX** best



Summary and Future Plans

- CzEng 0.7 (1.5M sents) available for research and educational purposes.
- Anonymous community generates huge data.
  The main issue is unresolved copyright issues.
  Quality seems reasonable (70% ok), given the zero cost.

Future plans with CzEng:

- Satisfied users.
  e.g. ACL Workshop on Machine Translation 2007 and 2008 shared tasks.
- Include more texts.
  Considering the idea of releasing everything (shuffled).
- Automatic annotation up to deep syntactic level.
- Designate subsections as development and evaluation data sets for MT.
  . . . properly cleaned-up.

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