CzEng 0.7: Parallel Corpus with Community-Supplied Translations

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, Peter Beňa

Institute of Formal and Applied Linguistics (ÚFAL)
Faculty of Mathematics and Physics, Charles University, Prague
{bojar,zabokrtsky,ceska}@ufal.mff.cuni.cz, pitkinpb@yahoo.com, mira.janicek@gmail.com

Abstract

This paper describes CzEng 0.7, a new release of Czech-English parallel corpus freely available for research and educational purposes. We provide basic statistics of the corpus and focus on data produced by a community of volunteers. Anonymous contributors manually correct the output of a machine translation (MT) system, generating on average 2000 sentences a month, 70% of which are indeed correct translations. We compare the utility of community-supplied and of professionally translated training data for a baseline English-to-Czech MT system.

1. Introduction

CzEng 0.7[1] is a new release of a Czech-English parallel corpus compiled at the Institute of Formal and Applied Linguistics, Charles University, Prague in 2005-2007. The corpus contains no manual annotation. It is limited only to texts which have been already available in an electronic form and which are not protected by authors' rights in the Czech Republic. The main purpose of the corpus is to support Czech-English and English-Czech machine translation (MT) research with the necessary data. CzEng 0.7 is available free of charge for educational and research purposes, however, the users should become acquainted with the license agreement.[2]

In this paper we describe the current state and size of CzEng and we specifically focus on data coming from open-source and volunteer community. Section 2. gives full details on CzEng 0.7 data: data sizes, preprocessing and alignment steps. In Section 3., we evaluate the quality and growth of community-supplied translations. Final Section 4. evaluates the English-to-Czech MT quality thanks to the new data.

2. Czech-English Parallel Data

Table 1 summarizes the magnitude of Czech-English parallel texts we could easily obtain via the Internet. About 18% of tokens (running words and punctuation) in the collection are fully proprietary and despite being available, they cannot be repackaged for the purposes of NLP research. About 40% of tokens come from professionally translated texts with permissive copyright restrictions (e.g. European law texts, fiction in public domain). Community-supplied data is found at both ends of the magnitude scale. Data where both the original text and the translation have a permissive license constitute only about 2.5% (1.3 million) of tokens. The other end is represented by more than 20 million (40%) of tokens in texts with unresolved copyright issues: the original text was typically not released for public use but volunteers equipped it with a translation.

The striking contrast illustrates that while a community of random volunteers produces significant amounts of valuable data, it does not pay enough principled attention to intellectual property rights. Thus only a tiny fragment of the community output is usable without breaching any copyrights. Hopefully we will see a convergence of international law regulations and general awareness of the community in coming years so that the proportion of "locked" data will decrease. Ready-made licences such as the GNU FDL[3] or Creative Commons licenses[4] are already available.

2.1. CzEng 0.7 Data

Full details on data sections included in CzEng 0.7 are given in Table 2. We used texts from the following publicly available sources:

- Acquis Communautaire Parallel Corpus (Ralf et al., 2006),
- Readers' Digest texts, partially already made available in (Čmejrek et al., 2004),
- Localization of KDE and GNOME, the major open-source software projects, into Czech.
- Articles from Project Syndicate.[5]
- The Kačenka parallel corpus, previously released as (Rambousek et al., 1997); because of the authors' rights, CzEng 0.7 can include only its subset, namely the following books:
  - D. H. Lawrence: Sons and Lovers / Synové a milenci,
  - Charles Dickens: The Pickwick Papers / Pickwickovci,
  - Charles Dickens: Oliver Twist,
  - Thomas Hardy: Jude the Obscure / Neblahý Juda,
  - Thomas Hardy: Tess of the d'Urbervilles / Tess z d'Urbervillu,
- Other E-books were obtained from various Internet sources; the English side comes mainly from

[1] http://ufal.mff.cuni.cz/czeng/

[2] http://ufal.mff.cuni.cz/czeng/license.html

[3] http://www.gnu.org/licenses/fdl.html

[4] http://creativecommons.org/

[5] http://www.project-syndicate.org/, Copyright: Project Syndicate, 2007. Permission granted to use the data for educational and non-commercial purposes only. Reprinting the material without written consent from Project Syndicate is a violation of international copyright law.

Source of Texts and Translation	Sentences Czech	Sentences English	Words+Punctuation Czech	Words+Punctuation English
Community Translation of Proprietary Texts	3,762,116	3,950,173	19,489,458	25,278,025
	60.3%	61.4%	37.8%	41.1%
Professional	1,271,319	1,271,000	21,254,025	23,948,106
	20.4%	19.8%	41.2%	38.9%
Proprietary	1,012,124	1,012,124	9,611,711	10,877,297
	16.2%	15.7%	18.6%	17.7%
Community	196,900	195,950	1,230,416	1,388,949
	3.2%	3.0%	2.4%	2.3%
Total	6,242,459	6,429,247	51,585,610	61,492,377
	100.0%	100.0%	100.0%	100.0%

Table 1: Available Czech-English parallel texts.

Project Gutenberg,[6] the Czech from Palmknihy.cz.[7] CzEng 0.7 includes these books:

  - Jack London: The Star Rover / Tulák po hvězdách,
  - Franz Kafka: Trial / Proces,
  - E. A. Poe: The Narrative of Arthur Gordon Pym of Nantucket / Dobrodružství A. G. Pyma,
  - E. A. Poe: A Descent into the Maelström / Pád do Malströmu,
  - Jerome K. Jerome: Three Men in a Boat / Tři muži ve člunu.
- User-contributed translations from the Navajo project.
- The European Constitution proposal from the OPUS corpus (Tiedemann and Nygaard, 2004),
- Samples from the Official Journal of the European Union, which is a tiny collection of some rather randomly chosen issues of the the Official Journal of the European Union.

2.2. Preprocessing

Since the individual sources of parallel texts differ in many aspects, a lot of effort was required to integrate them into a common framework. Depending on the type of the input resource, (some of) the following steps were applied on the Czech and English documents:

- conversion from PDF, Palm text (PDB DOC), SGML, HTML and other formats,
- encoding conversion (everything converted into UTF-8 character encoding), sometimes manual correction of mis-interpreted character codes,
- removing scanning errors, removing end-of-line hyphens,
- file renaming, directory restructuring,
- sentence segmentation,
- tokenization,
- removing long text segments having no counterpart in the corresponding document,
- adding sentence and token identifiers,
- conversion to a common XML format.

[6] http://www.gutenberg.org/

[7] http://www.palmknihy.cz/

2.2.1. TextSeg: Tokenization and Segmentation

In comparison to the previous release of CzEng, we improved the tokenization and sentence segmentation algorithm. Both sentence segmentation and tokenization are now handled by TextSeg, a newly implemented tool described in (Češka, 2006).

Input text is first broken into tokens corresponding to words (whitespace delimits tokens and furthermore each non-letter non-digit character constitutes a token on its own, with the exception of floating point numbers). For the purposes of segmentation, special auxiliary tokens $\langle\backslash\mathsf{n}\rangle$ (forced newline) and $\langle\mathrm{D}\rangle$ (no whitespace between the two neighbouring tokens) are left in the stream of tokens.

Sentence boundaries are identified using a decision tree based on 17 context attributes:

- current token t_0 is a full stop
- current token t_0 is a hyphen
- current token t_0 is a question mark
- current token t_0 is a quotation mark or closing bracket
- current token t_0 is $\langle\backslash\mathsf{n}\rangle$
- current token t_0 is the last one within the paragraph
- token t_{+1} starts with a capital letter
- token t_{+1} is in uppercase
- token t_{+1} is $\langle\mathrm{D}\rangle$ (meaning that there was no whitespace after the current token)
- token t_{+1} is a full stop, hyphen, question mark, quotation mark or closing bracket
- token t_{+2} is a number
- token t_{-1} is $\langle\mathrm{D}\rangle$
- token t_{-1} is a full stop, hyphen, question mark, quotation mark or closing bracket
- token t_{-2} is a number
- token t_{-2} is an abbreviation which cannot indicate the end of the sentence (e.g. “adj.”)
- token t_{-2} is an abbreviation which can indicate the end of the sentence (e.g. “etc.”)
- token t_{-2} is an abbreviation which is homonymic with another word (e.g. “no.”)

We collected extensive lists of three different types of Czech and English abbreviations for the last three context attributes. Our decision tree was trained on a set of manually processed data and its evaluation was based on independent human judgements. The evaluation set consists of

	Sentences		Words+Punctuation	
	Czech	English	Czech	English
Acquis Communautaire	881,348	882,965	14,465,145	15,820,486
	64.1%	63.8%	69.0%	67.6%
Readers' Digest	118,972	126,975	1,794,045	2,233,022
	8.6%	9.2%	8.6%	9.5%
Project Syndicate	89,460	88,675	1,869,292	2,076,702
	6.5%	6.4%	8.9%	8.9%
KDE Messages	85,591	85,582	396,542	440,921
	6.2%	6.2%	1.9%	1.9%
GNOME Messages	79,021	79,083	399,933	434,039
	5.7%	5.7%	1.9%	1.9%
Kačenka	57,157	57,580	1,034,638	1,188,023
	4.2%	4.2%	4.9%	5.1%
Navajo User Translations	32,288	31,578	433,941	513,989
	2.3%	2.3%	2.1%	2.2%
E-Books	15,966	16,308	330,112	399,595
	1.2%	1.2%	1.6%	1.7%
European Constitution	11,101	9,500	138,990	176,032
	0.8%	0.7%	0.7%	0.8%
Samples from European Journal	5,004	4,957	104,392	133,136
	0.4%	0.4%	0.5%	0.6%
Total	1,375,908	1,383,203	20,967,030	23,415,945
	100.0%	100.0%	100.0%	100.0%

Table 2: CzEng 0.7 sections and data sizes.

1,000 occurrences of tokens which can indicate the end of sentence (full stop, hyphen, question mark, quotation mark, closing bracket or newline symbol). The task is to decide whether the end of sentence indeed coincides with the token. Our method achieves the accuracy of 98.4% on this set.

2.3. Sentence Alignment

We used the Hunalign tool[8] (Varga et al., 2005) to align sentences for all documents; the settings were all kept default and we did not use any dictionary to bootstrap from. Hunalign collected its own temporary dictionary to improve sentence-level alignments.

For some sources, however, a more refined approach was taken. The localization of software projects GNOME and KDE consists of pairs of original (English) messages and their translations—the translations are thus strictly delimited, which allowed us to align only sentences within the concerned messages and run Hunalign on each pair separately. Moreover, if both the original message and its translation were segmented to exactly one sentence, we could safely treat the resulting pair of sentences as aligned.

The situation is similar with the data from the Navajo project, which also do not constitute continuous texts but are formed by pairs of snippets of English and their translations. However, due to technical difficulties, we did not apply the above procedure to the Navajo data.

The number of alignment pairs in CzEng 0.7 according to the number of sentences on the English and Czech side is given in Table 3.

[8] http://mokk.bme.hu/resources/hunalign

3. Translation Supplied by Open-Source Community

In CzEng 0.7, we used a significant amount of input from the Internet community in the form of localization of two major open-source projects and data from a commercial machine translation project. With Web 2.0 and open-source software becoming popular, we expect that such data sources will grow even faster in the future.

3.1. Fixing Machine Translation

Project Navajo[9] is an effort of a Czech commercial machine translation system producer to improve the quality of their English-to-Czech machine translation by means of community-supplied data. Similar to Wikipedia,[10] the Navajo project is an on-line encyclopedia that allows users to modify its contents. This time though, the emphasis is solely on the improvement of the translation, i.e. the users are not allowed to add new facts to the encyclopedia.

Navajo provides selected articles from the English edition of Wikipedia, machine-translated using a proprietary system to Czech. Using a web interface, any user can (anonymously) correct the translations of selected parts of the machine-translated articles, ranging from phrases to multiple sentences; the data they provide is then used to enhance the training set of the translation system, which is then retrained and used to translate the articles in the encyclopedia again.

Users are not allowed to change the entire articles at once though; throughout the site, it is made clear that the focus of the contributions should be on the improvement of the

[9] http://www.navajo.cz/

[10] http://www.wikipedia.org/

English-Czech	1-1	2-1	0-1	1-2	1-0	3-1	1-3	0-2	Others
Alignment pairs	1,096,940	68,856	63,185	43,057	30,694	11,003	4,786	3,855	13,479
	82.1%	5.2%	4.7%	3.2%	2.3%	0.8%	0.4%	0.3%	1.0%

Table 3: Sentence alignment pairs according to number of sentences.

	# of Phrase Translations
August 2006	1,575
September 2006	1,874
October 2006	3,142
November 2006	2,707
December 2006	1,418
January 2007	1,526
February 2007	1,989
March 2007	1,771
April 2007	2,065
May 2007	2,469
June 2007	2,438
July 2007	1,933
August 2007	2,251
September 2007	1,626
October 2007	1,424

Table 4: Number of translations contributed to Navajo per month.

Czech translation, not the factual accuracy or richness. This diverges from the aim of the Czech edition of Wikipedia, whose relation to the English version is at most that of an inspiration and which does not follow any linguistic aims; in fact, the Czech Wikipedia is completely independent from the English one (in terms of topics, contributors and target audience) and as such is unsuitable for the task undertaken by Navajo, which can be described as providing the "English edition of Wikipedia in Czech".

The contributed translations are stored in a database that is browsable and that can be modified: each entry can be either deleted or changed, causing its immediate removal from the database and, in case of modification, the introduction of an updated entry. This again differs from the concept of Wikipedia, where all previous revisions of the article are stored and accessible.

It is the database of translations that we include in CzEng 0.7: in all, the release contains 30,208 translations of various parts of 3,724 different Wikipedia articles; at least 614 more translations were deleted by the community. Of the 3,724 articles, the 300 most-edited ones comprise a half of the entire number of translations. (See Table 5 for ten most-edited articles.)

Table 4 shows the numbers of contributed phrase translations included in CzEng 0.7 per month. The numbers oscillate around 2,000 translated phrases per month with no apparent tendency of overall growth.

3.1.1. Quality of the Supplied Translations

In order to measure the quality of the user-contributed data, we randomly selected 1,000 aligned sentence pairs and manually evaluated the translation.

We found that of the thousand pairs, 690 were flawless

	# of Phrase Translations
Black Death	370
Ancient Egypt	232
Programming language	212
Rammstein	211
Seoul	202
Eva Perón	193
System of a Down	184
Goth	183
Soviet war in Afghanistan	176
Agatha Christie	170

Table 5: Ten most-translated Wikipedia articles in Navajo.

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%

Table 6: Quality of Navajo user translations in CzEng.

translations (both grammatically correct and precise), 45 had slight grammatical, punctuation or diacritic errors, but besides these errors they conveyed the correct meaning, 58 were imprecise translations that did not keep the information value of the English part, but were more or less grammatically correct. In 68 pairs, the English and Czech parts were identical; most of them were names of people or artifacts, and as such may be considered correct translations.

66 pairs seemed to be incomplete, possibly due to the process of sentence alignment, as either the English or the Czech part was empty (i.e. the alignment was of type 0-x or x-0), or parts of the translated message seemed to be missing on one of the sides (which may happen in cases where the alignment is of type 1-2 or 2-1 etc.), or the Czech text was grammatically correct, but not related to the English original, probably because of a "shifted" alignment.

44 pairs were identified as just slight improvements of the machine-translated text with the machine-generated Czech easily recognizable. 27 pairs were identified as vandalism (i.e. the Czech part contains obscene language, nonsensical strings of characters or messages in languages other than Czech). Finally, the remaining 2 pairs consisted of punctuation tokens of no linguistic value of their own.

The results are summarized in Table 6.

Section	Training Data			Test Data	
				Domain-D	Out-of-domain-D
	Sentences	Tokens	Vocabulary	Tokens Out of Vocabulary	
D	84,141	1,952,352	35,736	504 (2.1 %)	3,009 (5.9 %)
DC	276,695	3,271,020	66,318	423 (1.7 %)	2,314 (4.5 %)
P	909,871	16,520,134	112,646	505 (2.1 %)	2,202 (4.3 %)
CX	2,844,950	18,207,559	164,690	637 (2.6 %)	1,873 (3.7 %)
DP	994,012	18,472,486	120,428	363 (1.5 %)	1,948 (3.8 %)
DCP	1,186,566	19,791,154	140,168	352 (1.5 %)	1,850 (3.6 %)
DCX	2,929,091	20,159,911	172,634	371 (1.5 %)	1,607 (3.1 %)
CPX	3,754,821	34,727,693	224,878	424 (1.7 %)	1,546 (3.0 %)
DCPX	3,838,962	36,680,045	230,136	347 (1.4 %)	1,416 (2.8 %)
Test Tokens	-	-	-	24,229 (100.0 %)	51,297 (100.0 %)
Test Sentences	-	-	-	964	2,051

Table 7: Training data sizes (English side) and out-of-vocabulary rates for various sections of CzEng.

3.2. Open-Source Software Localization

KDE[11] and GNOME[12] are open-source projects that deliver the two major free desktop environments for Unix-like operating systems. In CzEng 0.7, we used the data from their localization into Czech (the development language is English).

Translators contribute their translations either using specialized programs or via a web interface[13] that allows for manipulation with the strings in a wiki-like manner.

Due to the specialized technical nature of the localized strings and the fact that the translations cannot be submitted anonymously, the community of translators is much smaller in comparison to that of the Navajo project and the overall quality of the translations is higher, albeit limited to the domain of system messages. However, we did not undertake any evaluation procedure similar to that described in Section 3.1.1. for Navajo.

4. Achievable MT Quality Using CzEng 0.7

In order to evaluate the utility of CzEng 0.7 data for machine translation (MT), we trained the phrase-based decoder Moses (Koehn et al., 2007) on various sections of CzEng 0.7 for English-to-Czech translation.

We used the original tokenization of CzEng and automatically evaluated MT quality using BLEU (Papineni et al., 2002) (all lowercased), estimating empirical 95% confidence bounds using method by Koehn (2004). Moses was run in the most basic single-factored setup.

We chose Project Syndicate as our "domain corpus" because there are standard development and evaluation sets for Project Syndicate data available as part of shared task of ACL 2007 Workshop on Machine Translation (WMT07[14]). (WMT organizers refer to Project Syndicate data as "News Commentary" corpus.) In addition to in-domain test set of 964 sentences, we also use a contrastive out-of-domain set of 2,051 sentences, as made available for WMT08[15].

Table 7 summarizes the statistics about the source (English) side of sentences. For the purposes of this experiment, we use only sentences aligned 1-to-1 and delimit CzEng data into the following disjoint sections: "D"–in-domain Project Syndicate data, translated professionally, "P"–other professionally translated texts (e.g. Acquis, books, Readers' Digest; excluding Project Syndicate), "C"–community-supplied texts and translations (i.e. Navajo, KDE, GNOME), "X"–community translation of proprietary texts (not part of CzEng 0.7 release). We experiment with various combinations of the sections. The combination "DCP" corresponds to the official CzEng 0.7 release apart from a few books that cannot be distributed yet.

Figure 1 plots the out-of-vocabulary rate against the BLEU score achieved. We see that for in-domain translation, neither much OOV reduction, nor much increase in BLEU is achieved by any data added to "D". In fact, the slightly noisy community data can cause an insignificant loss in BLEU while high-quality professional translations can help a tiny bit.[16]

For out-of-domain evaluation, the picture is clearer: any reduction in OOV rate counts and helps to achieve a higher BLEU score. Adding the professional "P" section to the baseline "D" brings in more data of a higher quality, thus increasing BLEU more. However already the community sections (the relatively small "C" and especially the bigger "X") make a difference, see "D" vs. "DCX". Unsurprisingly, the best score is achieved when all data are employed ("DCPX").

The difference in training data suitability can be best seen when comparing "DP" and "DCX": while comparable in number of tokens (18 and 20 million, resp.), the community data contain much shorter sentences (6.9 vs. 18.6 tokens per sentence on average) due to a different nature of

[11] http://www.kde.org/
[12] http://www.gnome.org/
[13] http://translations.launchpad.net/
[14] http://www.statmt.org/wmt07/
[15] http://www.statmt.org/wmt08/

[16] For in-domain translation trained on "D" with some additional out-of-domain data, we employ two separately weighted language models (LM): one covering all training data and one just for the in-domain "D" section. Omitting the separate "D" LM significantly reduces BLEU scores. For out-of-domain translation, the separate "D" language model does not bring any improvement, so we use a single LM based on the target side of the respective training data.

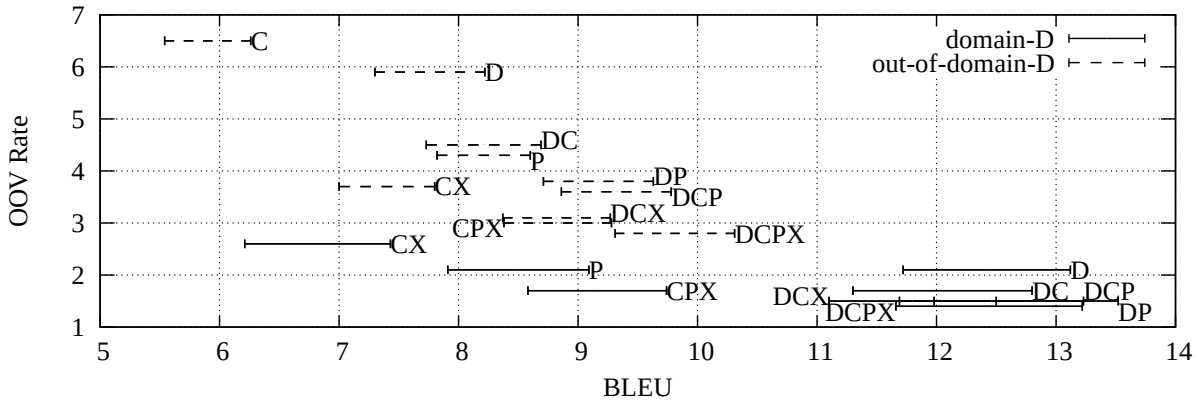

Figure 1: BLEU scores (empirical confidence intervals at 95%) compared to OOV rate.

the texts. While the average 6.9 tokens in a sentence are still well below the 3-gram language model context span, we are not surprised to see that phrase and language model probabilities get skewed: When translating the out-of-domain test set (25.0 tokens per sentence), "DCX" has lower predictive power for within-sentence translation and coherence compared to "DP". The same observation can be made by comparing BLEU scores for "D" (only about 2 million tokens) and "CX" (about 18 million tokens): while relevant to OOV reduction, "CX" is bad for longer phrases.

5. Summary And Further Plans

We have presented CzEng 0.7, a collection of Czech-English parallel texts. The corpus of about 20 million tokens is automatically sentence aligned. CzEng 0.7 is available free of charge for educational and research purposes, the licence allows collecting statistical data and making short citations. To our knowledge, it is the biggest and the most diverse publicly available parallel corpus for the Czech-English pair.

We observed that while copyright issues allow us to include only about 1.3 million tokens produced by a community of volunteers, the community has produced more than 15 times bigger collection of translations with unresolved status of the source texts. Hopefully the contributors will pay more attention to property rights and produce data outside of restrictive circumstances.

We evaluated the quality of user-supplied corrections of machine-translated sentences, showing that about 70% of the sentences are reasonably accurate. This data source grows for free at about 2,000 sentences every month.

While not as useful as professionally translated texts, community-supplied data can still lead to significant improvement in machine translation quality, especially when translating texts outside of the original domain.

Future versions of CzEng will contain (machine) annotation of the data on various levels up to deep syntactic layer. We also plan to designate subsections of CzEng as standard development and evaluation data sets for machine translation, paying proper attention to cleaning up of these sets.

6. Acknowledgement

The work on this project was partially supported by the grants FP6-IST-5-034291-STP (EuroMatrix), MSM 0021620838, and 1ET101120503.

7. References

Pavel Češka. 2006. Segmentace textu. Bachelor's Thesis, Faculty of Mathematics and Physics, Charles University in Prague.

Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of LREC 2004*, Lisbon, May 26–28.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, Barcelona, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2142–2147. ELRA.

Jiří Rambousek, Jana Chamonikolasová, Daniel Mikšík, Dana Šlancarová, and Martin Kalivoda. 1997. KAČENKA (Korpus anglicko-český - elektronický nástroj Katedry anglistiky).

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel & free. In *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*, Lisbon, May 26–28.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing RANLP 2005*, pages 590–596, Borovets, Bulgaria.