# RNDr. Ondřej Bojar, Ph.D.


Date of Birth: 7th March 1979 in Prague
E-mail, Web: bojar@ufal.mff.cuni.cz; http://www.cuni.cz/~obo


**Education:**

| | |
|---|---|
| 2003-2008 | Ph.D. study at Institute of Formal and Applied Linguistics (ÚFAL), MFF UK<br>Doctoral thesis: Exploiting Linguistic Data in Machine Translation |
| 1997-2003 | Bachelor and Master degree (summa cum laude), Charles University in Prague, Faculty of Mathematics and Physics (MFF UK)<br>Master thesis: Automatic extraction of lexico-syntactic information from corpora |
| 1997-1999 | Parallel study of Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague; finished 4 semesters |
| 1993-1997 | High School Zborovská, Graduation Exam, best results |

**Other Experience:**

| | |
|---|---|
| since 2010 | assistant professor at ÚFAL, Charles University in Prague |
| 2010 | Two-month research stay at RWTH Aachen University |
| 2006-2007 | Twelve-month research visit at CSSE, University of Melbourne, tutor |
| Jul-Aug 2006 | Language Engineering Workshop (Machine Translation Team) at JHU, Baltimore (6 weeks) |
| since 2004 | Teaching assistant at both MFF UK and Czech Technical University |
| Oct-Nov 2005 | Two-month research stay at RWTH Aachen University |
| 2000-2005 | Programming and analysis, Internet protocols.<br>Internet Info, Ltd., http://www.iinfo.cz/ |
| 2003-2004 | Six-month study and research stay at University of Saarland, Saarbrücken |
| 1995-2003 | Teacher, Computer courses (grades 8 and 9), Primary School Fr. Plamínkové, Prague 7 |
| 1996-1999 | Created and maintained web pages at http://www.cestina.cz/ |
| 1996-1997 | Co-founder and vice-chairman of KPPM, the Czech Macintosh User Group |
| 1994-1997 | Publications in the Czech Macworld (not in the list of publications below) |

**Involved in Projects:**

| | |
|---|---|
| since 2012 | EU FP7 CSA MosesCore (Support for the Moses translation system.) |
| 2011-2013 | GAČR: Czech in the Machine Translation Era |
| 2010-2012 | GAČR PGS: Combining Phrase-Based and Deep Syntactic Machine Translation |
| 2009-2012 | EU FP7 Strep EuroMatrixPlus (Bringing Machine Translation for European Languages to the User) |
| 2006-2009 | EU FP6 Strep EuroMatrix (Machine translation between European languages) |
| since 2004 | PCEDT (The Prague Czech-English Dependency Treebank) |
| 2004-2005 | VALLEX (Valency Lexicon of Czech Verbs) |
| 2003-2005 | Machine translation project of economical texts from Czech to English |
| 2000-2002 | Software project The ENTs–A simulation of natural environment with human-like computer-driven agents, http://ufal.mff.cuni.cz/~bojar/enti |

**Other Skills:**

- English (fluent, Certificate in Advanced English), German (fluent, Zentrale Mittelstufenprüfung)
- programming languages Mercury, Perl, Prolog, PHP, SQL (excellent knowledge), C, C++, Java (good)
- extensive programming experience with Unix (and previously Mac OS and DOS)
- extensive programming experience with TCP/IP, Internet

## Publications:

### Books

Ondřej Bojar, Silvie Cinková, Jan Hajič, Barbora Hladká, Vladislav Kuboň, Jiří Mírovský, Jarmila Panevová, Nino Peterek, Johanka Spoustová, and Zdeněk Žabokrtský. 2012. *The Czech Language in the Digital Age*. Springer-Verlag Berlin Heidelberg, Berlin, Germany.

Ondřej Bojar. 2012. *Čeština a strojový překlad (Czech Language and Machine Translation)*, volume 11 of *Studies in Computational and Theoretical Linguistics*. ÚFAL, Praha, Czech Republic.

Ondřej Bojar. 2009. *Exploiting Linguistic Data in Machine Translation*, volume 3 of *Studies in Computational and Theoretical Linguistics*. ÚFAL, Prague, Czech Republic.

### Refereed

Jiří Maršík and Ondřej Bojar. 2012. TrTok: A Fast and Trainable Tokenizer for Natural Languages. *Prague Bulletin of Mathematical Linguistics*, 98:75–85, September.

Ondřej Bojar and Dekai Wu. 2012. Towards a Predicate-Argument Evaluation for MT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 30–38, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3921–3928, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Mark Fishel, Rico Sennrich, Maja Popović, and Ondřej Bojar. 2012. TerrorCat: a Translation Error Categorization-based MT Quality Metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 64–70, Montréal, Canada, June. Association for Computational Linguistics.

Mark Fishel, Ondřej Bojar, and Maja Popović. 2012. Terra: a Collection of Translation Error-Annotated Corpora. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 7–14, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Aleš Tamchyna, Petra Galuščáková, Amir Kamran, Miloš Stanojević, and Ondřej Bojar. 2012. Selecting Data for English-to-Czech Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 374–381, Montréal, Canada, June. Association for Computational Linguistics.

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. 2012. Probes in a Taxonomy of Factored Phrase-Based Models. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 253–260, Montréal, Canada, June. Association for Computational Linguistics.

Petra Galuščáková and Ondřej Bojar. 2012. Improving SMT by Using Parallel Data of a Closely Related Language. In *Human Language Technologies – The Baltic Perspective – Proceedings of the Fifth International Conference Baltic HLT 2012*, volume 247 of *Frontiers in AI and Applications*, pages 58–65, Amsterdam, Netherlands, October. IOS Press.

Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, and Daniel Zeman. 2012. Automatic MT Error Analysis: Hjerson Helping Addicter. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 2158–2763, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3153–3160, Istanbul, Turkey, May. ELRA, European Language Resources Association.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432, Edinburgh, Scotland, July. Association for Computational Linguistics.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-Based Evaluation of Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:77–86, March.

Ondřej Hálek, Rudolf Rosa, Aleš Tamchyna, and Ondřej Bojar. 2011. Named Entities from Wikipedia for Machine Translation. In Markéta Lopatková, editor, *ITAT 2011 Information Technologies – Applications and Theory*, volume 788, pages 23–30, September.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ondřej Bojar, Petra Galuščáková, and Miroslav Týnovský. 2011. Evaluating Quality of Machine Translation from Czech to Slovak. In Markéta Lopatková, editor, *Information Technologies – Applications and Theory*, pages 3–9, September.

Česlav Przywara and Ondřej Bojar. 2011. eppex: Epochal Phrase Table Extraction for Statistical Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 96:89–98, October.

Mark Fishel, Ondřej Bojar, Daniel Zeman, and Jan Berka. 2011. Automatic Translation Error Analysis. In *Text, Speech and Dialogue: 14th International Conference, TSD 2011*, volume LNAI 3658. Springer Verlag, September.

Matouš Macháček and Ondřej Bojar. 2011. Approximating a Deep-Syntactic Metric for MT Evaluation and Tuning. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 92–98, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ondřej Bojar. 2011. Analyzing Error Types in English-Czech Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:63–76, March.

Daniel Zeman, Mark Fishel, Jan Berka, and Ondřej Bojar. 2011. Addicter: What Is Wrong with My Translations? *Prague Bulletin of Mathematical Linguistics*, 96:79–88, October.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar, Adam Liška, and Zdeněk Žabokrtský. 2010. Evaluating Utility of Data Sources in a Large Parallel Czech-English Corpus CzEng 0.9. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 447–452, Valletta, Malta, May. ELRA, European Language Resources Association.

Ondřej Bojar, Pavel Straňák, and Daniel Zeman. 2010. Data Issues in English-to-Hindi Machine Translation. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 1771–1777, Valletta, Malta, May. ELRA, European Language Resources Association.

Jana Šindlerová and Ondřej Bojar. 2010. Building a Bilingual ValLex Using Treebank Token Alignment: First Observations. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 304–309, Valletta, Malta, May. ELRA, European Language Resources Association.

Aleš Tamchyna and Ondřej Bojar. 2010. Bohatá anotace ve frázovém strojovém překladu. In *ITAT 2010 Informačné technológie – Aplikácie a Teória, Zborník príspevkov prezentovaných na konferencii ITAT*, pages 99–106, September.

Jiří Diviš and Ondřej Bojar. 2010. Automatic Source Code Reduction. In *ITAT 2010 Information Technologies – Applications and Theory*, pages 9–16. PONT s. r. o., Seňa, Slovakia, September.

Ondřej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Hana Klempová, Michal Novák, Peter Fabian, Jan Ehrenberger, and Ondřej Bojar. 2009. Získávání paralelních textů z webu. In *ITAT 2009 Information Technologies – Applications and Theory*, September.

Jana Šindlerová and Ondřej Bojar. 2009. Towards English-Czech Parallel Valency Lexicon via Treebank Examples. In *Proceedings of 8th Treebanks and Linguistic Theories Workshop (TLT)*, pages 185–195, Milano, Italy, December.

David Kolovratník, Natalia Klyueva, and Ondřej Bojar. 2009. Statistical Machine Translation between Related and Unrelated Languages. In *ITAT 2009 Information Technologies – Applications and Theory*, September.

Kamil Kos and Ondřej Bojar. 2009. Evaluation of Machine Translation Metrics for Czech as the Target Language. *Prague Bulletin of Mathematical Linguistics*, 92:135–147.

Ondřej Bojar, Pavel Straňák, Daniel Zeman, Gaurav Jain, Michal Hrušecký, Michal Richter, and Jan Hajič. 2009. English-Hindi Translation—Obtaining Mediocre Results with Bad Data and Fancy Models. In *Proceedings of the 7th International Conference On Natural Language Processing (ICON-2009)*, Hyderabad, India, December. NLP Association of India.

Ondřej Bojar, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. 2009. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Ondřej Odcházel and Ondřej Bojar. 2009. Computer Aided Translation Backed by Machine Translation. In *Proceedings of the ASLIB International Conference Translating and the Computer 31*, London, UK, November.

Natalia Klyueva and Ondřej Bojar. 2008. UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proc. of International Conference Corpus Linguistics*, pages 188–195, October.

Ondřej Bojar and Jan Hajič. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, Ohio, June. Association for Computational Linguistics.

Ondřej Bojar, Pavel Straňák, and Daniel Zeman. 2008. English-Hindi Translation in 21 Days. In *Proceedings of the 6th International Conference On Natural Language Processing (ICON-2008) NLP Tools Contest*, Pune,

India, December. NLP Association of India.

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. 2008. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. ELRA.

Ondřej Bojar, Silvie Cinková, and Jan Ptáček. 2007. Towards English-to-Czech MT via Tectogrammatical Layer. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, Bergen, Norway, December.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, 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.

Ondřej Bojar. 2007. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague, Czech Republic, June. Association for Computational Linguistics.

Ondřej Bojar and Zdeněk Žabokrtský. 2006. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62.

Václava Benešová and Ondřej Bojar. 2006. Czech Verbs of Communication and the Extraction of their Frames. In *Text, Speech and Dialogue: 9th International Conference, TSD 2006*, volume LNAI 3658, pages 29–36. Springer Verlag, September.

Ondřej Bojar and Magdalena Prokopová. 2006. Czech-English Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1236–1239. ELRA, May.

Ondřej Bojar, Evgeny Matusov, and Hermann Ney. 2006. Czech-English Phrase-Based Machine Translation. In *FinTAL 2006*, volume LNAI 4139, pages 214–224, Turku, Finland, August. Springer.

Ondřej Bojar, Jiří Semecký, and Václava Benešová. 2005. VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83:5–17.

Markéta Lopatková, Ondřej Bojar, Jiří Semecký, Václava Benešová, and Zdeněk Žabokrtský. 2005. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings*, volume LNAI 3658, pages 99–106. Springer Verlag, September.

Ondřej Bojar, Cyril Brom, Milan Hladík, and Vojtěch Toman. 2005. The Project ENTs: Towards Modelling Human-like Artificial Agents. In Peter Vojtáš, Mária Bieliková, Bernadette Charron-Bost, and Ondrej Sýkora, editors, *SOFSEM 2005 Communications*, pages 111–122. Society for Computer Science, January.

Ondřej Bojar, Petr Homola, and Vladislav Kuboň. 2005. Problémy recyklování systému automatického překladu. In Peter Vojtáš, editor, *ITAT 2005 Information Technologies – Applications and Theory*, pages 335–344, Košice, Slovakia, September. University of P. J. Šafařík.

Ondřej Bojar, Petr Homola, and Vladislav Kuboň. 2005. Problems Of Reusing An Existing MT System. In *IJCNLP 2005 - Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts*, pages 181–186, October.

Ondřej Bojar and Jan Hajič. 2005. Extracting Translation Verb Frames. In Walther von Hahn, John Hutchins, and Christina Vertan, editors, *Proceedings of Modern Approaches in Translation Technologies, workshop in conjunction with Recent Advances in Natural Language Processing (RANLP 2005)*, pages 2–6. Bulgarian Academy of Sciencies, September.

Ondřej Bojar. 2005. Budování česko-anglického slovníku pro strojový překlad. In Peter Vojtáš, editor, *ITAT 2005 Information Technologies – Applications and Theory*, pages 201–211, Košice, Slovakia, September. University of P. J. Šafařík.

Ondřej Bojar, Petr Homola, and Vladislav Kuboň. 2005. An MT System Recycled. In *Proceedings of MT Summit X*, pages 380–387, September.

Ondřej Bojar. 2004. Problems of Inducing Large Coverage Constraint-Based Dependency Grammar for Czech. In *Constraint Solving and Language Processing, CSLP 2004*, volume LNAI 3438, pages 90–103, Roskilde University, September. Springer.

Ondřej Bojar. 2004. Czech Syntactic Analysis Constraint-Based, XDG: One Possible Start. *Prague Bulletin of Mathematical Linguistics*, 81:43–54.

Ondřej Bojar. 2004. Automated Extraction of Lexico-Syntactic Information. In Jana Šafránková, editor, *WDS'04 Proceedings of Contributed Papers: Part I - Mathematics and Computer Sciences*, pages 211–217, Prague, June 15–18. Charles University, Matfyzpress.

Ondřej Bojar. 2003. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120.

Ondřej Bojar. 2003. Building Subcorpora Suitable for Extraction of Lexico-Syntactic Information. In *Proceedings of the Student Session, ESSLLI*, August.

### Other

Ondřej Bojar, Aleš Tamchyna, and Jan Berka. 2012. Wild Experimenting in Machine Translation. CLARA Winter School in Prague, January.

Ondřej Bojar. 2012. Strojový překlad. *Vesmír*, 91:488–490, September.

Ondřej Bojar, Mauro Cettolo, Silvie Cinková, Philipp Koehn, Miroslav Týnovský, and Zdeněk Žabokrtský. 2012. Scientific Report on Rich Tree-Based SMT. Project EuromatrixPlus - Deliverable 1.4, ÚFAL, Charles University, March.

Ondřej Bojar and Aleš Tamchyna. 2011. Forms Wanted: Training SMT on Monolingual Data. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel, January.

Petra Galuščáková and Ondřej Bojar. 2011. Czech-Slovak Parallel Corpora. In *Proc. of Slovko 2011*, October.

Ondřej Bojar, Chris Callison-Burch, Jan Hajič, and Philipp Koehn, editors. 2009. *Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools*. Number 91 in Prague Bulletin of Mathematical Linguistics. Charles University, January.

Ondřej Bojar and Miroslav Týnovský. 2009. Evaluation of Tree Transfer System. Project Euromatrix - Deliverable 3.4, ÚFAL, Charles University, March.

Ondřej Bojar and Adam Lopez. 2008. Tree-based Translation. Handout for MT Marathon Tutorial, May.

Zdeněk Žabokrtský and Ondřej Bojar. 2008. TectoMT, Developer's Guide. Technical Report TR-2008-39, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, December.

Ondřej Bojar, Miroslav Janíček, and Miroslav Týnovský. 2008. Implementation of Tree Transfer System. Project Euromatrix - Deliverable 3.3, ÚFAL, Charles University, September.

Ondřej Bojar and Magdalena Prokopová. 2007. Czech-English Machine Translation Dictionary. Technical report, ÚFAL MFF UK, Prague, Czech Republic, April.

Ondřej Bojar. 2006. Strojový překlad: zamyšlení nad účelností hloubkových jazykových analýz. In *MIS 2006*, pages 3–13, Josefův Důl, Czech Republic, January. MATFYZPRESS.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondřej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, Alexandra Constantin, Christine Moran, and Evan Herbst. 2006. Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding. Technical report, Johns Hopkins University, Center for Speech and Language Processing, August.

Ondřej Bojar, Jiří Semecký, Shravan Vasishth, and Ivana Kruijff-Korbayová. 2004. Processing noncanonical word order in Czech. In *Proceedings of Architectures and Mechanisms for Language Processing, AMLaP 2004*, pages 91–91, Université de Provence, September 16-18.

Ondřej Bojar, Cyril Brom, Milan Hladík, Mikuláš Vejlupek, Vojtěch Toman, and David Voňka. 2003. ENTI – Simulátor přirozeného prostředí lidského světa. In *MIS 2003*, pages 3–14. MATFYZPRESS, January 18–25, 2003.

Ondřej Bojar. 2003. AX - Systém pro automatizovanou extrakci lexikálně-syntaktických údajů. In *MIS 2003*, pages 15–24. MATFYZPRESS, January 18–25.

Prague, February 13, 2013.

Ondřej Bojar