

# Milan Straka – Curriculum Vitae

## Personal Details

Name: RNDr. Milan Straka, Ph.D.  
Born: 11<sup>th</sup> April 1984, Plzeň  
Nationality: Czech  
Languages: Czech (native), English (FCE), German (basic)  
Home address: Liliová 1514  
Strakonice, 38601  
Czech Republic  
Email address: milan@strakovi.com  
Phone number: +420 605 435 586



## Education

- obtained Doctorate (Ph.D.) in Computer Science (thesis topic was Functional data structures and algorithms) in September 2013 at the Faculty of Mathematics and Physics of Charles University, all of university courses passed with grade A
- obtained Master (Mgr) degree in Information Security in June 2008 at the Faculty of Mathematics and Physics of Charles University, all of university courses passed with grade A
- obtained Bachelor (Bc) degree in Computer Science in June 2006 at the Faculty of Mathematics and Physics of Charles University, all of the university courses passed with grade A
- four times in a row received the university scholarship for the best students

## Work and Research Experience

- assistant professor (previously research assistant) at Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, since 2013
  - NLP deep learning research
  - implementations of NLP applications (POS tagging, lemmatization, parsing, named entity recognition, morphological derivation)
  - teaching deep learning and distributed programming courses
  - supervising master theses in NLP machine learning
  - member of organizing committee of CoNLL 2017 shared task in UD parsing
  - contract research for the Ministry of the Interior of the Czech Republic in 2017 (named entities recognition application)
- Hadoop cluster installation and tutorial, for ÚFAL, Charles University, January 2012
  - Hadoop updated to Apache Spark in 2014

## International Research Stays and Internships

- research assistant in Google, Zurich, Sep 2017–Mar 2018
  - deep learning in NLP – conversationalization of Google Home Assistant answers
- intern in Microsoft Research Labs, Cambridge, March–May 2010
  - working on functional data structures in Haskell and GHC Haskell compiler
- research stay at Simon Fraser University, Canada, February 2009
  - working on fully persistent data structures
- two times intern in Google (Czech Engineering Team), Jul–Sep 2009 and Jun–Sep 2008
  - distributed computations using MapReduce
  - Czech morphological analysis and Czech language model implementations

## Teaching

- Deep learning course (theoretical lectures and practical TensorFlow assignments)
  - followed by Deep learning seminar
- distributed programming course (using Spark and SGE)

- advanced functional programming techniques course (in Haskell)
- OCaml and F# course
- teaching assistant of various computer science courses
  - advanced programming
  - algorithms & data structures (both basic and advanced course)

## Papers

- Milan Straka, Jana Straková and Jan Hajič: **Prague at EPE 2017: The UDPipe System**. In Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation at the Fourth International Conference on Dependency Linguistics and the 15th International Conference on Parsing Technologies, Pisa, Italy, September 2017.
- Milan Straka and Jana Straková: **Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe**. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017.
- Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre et al.: **CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017.
- Natalia Klyueva and Antoine Doucet and Milan Straka: **Neural Networks for Multi-Word Expression Detection**. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017, Workshop on Multiword Expressions), Valencia, Spain, April 2017.
- Jana Straková, Milan Straka, Magda Ševčíková, and Zdeněk Žabokrtský: **Czech Named Entity Corpus**. In The Handbook of Linguistic Annotation, editors Nancy Ide and James Pustejovsky, ISBN 9402408797, Springer, Netherlands, 2017.
- Ševčíková Magda, Žabokrtský Zdeněk, Vidra Jonáš, Straka Milan. **Lexikální síť DeriNet: elektronický zdroj pro výzkum derivace v češtině**. In Časopis pro moderní filologii, Vol. 98, No. 1, Charles University, Prague, Czech Republic, ISSN 0008-7386, pp. 62-76, Sep 2016.
- Jana Straková, Milan Straka and Jan Hajič: **Neural Networks for Featureless Named Entity Recognition in Czech**. In Proceedings of the 19th International Conference on Text, Speech and Dialogue (TSD 2016), Brno, Czech Republic, September 2016.
- Milan Straka, Jan Hajič and Jana Straková: **UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing**. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016.
- Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra and Adéla Limburská: **Merging Data Resources for Inflectional and Derivational Morphology in Czech**. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016.
- Milan Straka, Jan Hajič, Jana Straková and Jan Hajič jr: **Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle**. In Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14), December 2015.
- Jana Straková, Milan Straka and Jan Hajič: **Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition**. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014): System Demonstrations, pages 13-18, Baltimore, Maryland, June 2014.
- Jana Straková and Milan Straka: **A New State-of-the-Art Czech Named Entity Recognizer**. In Text, Speech and Dialogue (TSD 2013), Pilsen, Czech Republic, Sep 2013.
- David Mareček and Milan Straka: **Stop-probability estimates computed on a large corpus improve Unsupervised Dependency Parsing**. In Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, August 2013.

- Milan Straka: ***Adams' Trees Revisited – Correct and Efficient Implementation.*** In Proceedings of Symposium on Trends in Functional Programming (TFP 2011), Madrid, Spain, May 2011.
- Milan Straka: ***The performance of the Haskell containers package.*** In Proceedings of Haskell 2010, 3rd ACM Haskell symposium on Haskell, Baltimore, Maryland, Sep 2010.
- Milan Straka: ***Optimal worst-case fully persistent arrays.*** In Symposium on Trends in Functional Programming (TFP 2009), Komarno, Slovakia, June 2009.
- Martin Mareš, Milan Straka: ***Linear-Time Ranking of Permutations.*** In Proceedings of ESA 2007, 15th Annual European Symposium, Eilat, Israel, October 2007.

## Theses

- Doctoral thesis: ***Functional Data Structures and Algorithms***
- Master thesis: ***Quadratic fields based cryptography***
- Bachelor thesis: ***Factoring polynomials over finite fields***

## Skills

- programming skills
  - excellent programming skills in C++, Python, Perl, Java, C#
  - advanced knowledge of functional languages, especially Haskell, F# and OCaml
  - everyday experience in Linux and Windows programming
  - thorough knowledge of MapReduce paradigm and Spark framework
  - fairly good knowledge of and experience in TCP/IP protocol suite
- machine learning
  - excellent knowledge of newest methods of NLP deep learning
  - several open-source implementations of NLP models (tokenizer, POS tagger, NE recognizer, parser) using deep artificial neural networks, implemented from scratch in C++
  - excellent knowledge of TensorFlow
  - bayesian nonparametrics modelling (unsupervised parsing)
- theoretical computer science
  - algorithmization and advanced graph algorithms (weighted perfect matching, micro/macrotree decomposition, LCA, planarity testing)
  - advanced data structures (interval, suffix, Sleator-Tarjan trees; Q-Heaps; persistent, self-adjusting and amortized data structures)
  - advanced knowledge of discrete mathematics and graph theory
  - probabilistic algorithms (maxIS, perfect hashing, random walks, derandomizing)
- computer graphics
  - experience in hardware accelerated 3D graphics, especially OpenGL and GLSL
- information security
  - symmetric ciphers, hash functions, public key cryptography (RSA, DSA)
  - algebraic number theory (primality testing, factorization, cryptanalysis)

## High Scores at Programming Competitions

- gold medal (14<sup>th</sup> out of 269) at the 15<sup>th</sup> International Olympiad in Informatics in 2003
- first place for four times at the national round of ACM International Collegiate Programming Contest in 2004–2007, second place in 2003
- fourth, fifth, fifth and sixth out of 50, 53, 62 and 67 at the Central European Programming Contest of the ACM International Collegiate Programming Contest in 2004–2007
- first or second place at six national programming competitions in 1999–2003
- first place out of 59 (nearly twice as many points than every one else) at the correspondence seminar in programming of the Faculty of Mathematics and Physics of Charles University in 2003
- participation at the 14<sup>th</sup> International Olympiad in Informatics in 2002
- silver medal (9<sup>th</sup> place) at the Central European Olympiad in Informatics in 2002

## Other Activities

- organization of programming competitions
  - the Czech Republic Championship in Programming, since 2004
  - the Czech National Olympiad in Informatics, 2004–2011
  - the main organizer of the correspondence seminar in programming of the Faculty of Mathematics and Physics of Charles University for 4 years. Each year I came up with 20 problems for the contestants, varying between logical ones, theoretical computer science tasks (no source needed) and practical tasks (the source is mandatory and points are assigned according to the results produced).
  - the Central European Olympiad in Informatics in Brno, 2007
- organization of programming camps
  - annual camp of the correspondence seminar in programming, 2003–2007
  - annual Programmer's Summer Camp, 2002–2007