Curriculum Vitae

VIDAS DAUDARAVIČIUS

PERSONAL DATA

Work address European Commission Joint Research Centre, Via Enrico Fermi 2749, I-21027 Ispra, Italy
E-mail vidas.daudaravicius@ec.europa.eu
Linkedin https://lt.linkedin.com/in/vidas-daudaravicius-28371159

EDUCATION

2008 - 2012 PhD student, Informatics
Faculty of Informatics, Vytautas Magnus University.
Dissertation title: *Collocation Segmentation for Text Chunking*, Defended: 31 Jan 2013,
Supervisor: prof. Minija Tamosiunaite.

2000–2002 Master's degree of Applied Informatics
Thesis Title: *Syntactical Analysis of Lithuanian Language.* Supervisor: prof. Jan Kuper.
Faculty of Informatics, Vytautas Magnus University

1995–2000 Bachelor's degree of the Informatics
Faculty of Informatics, Vytautas Magnus University

1992–1995 Religious studies
St. Anthony Religious Studies Institute at the Faculty of Catholic Theology
Vytautas Magnus University

WORK EXPERIENCE

2020-present ***The Joint Research Center (Ispra), The European Commission***
Senior Research Fellow
2020-present

- Conducting research in the area of knowledge extraction and engineering machine learning, deep learning and other natural language processing methods and apply these techniques for advancing the search capabilities of documents and data enrichment
- Administrative, advisory, linguistic and equivalent technical tasks

2011-2020 ***UAB VTeX***
Research manager
2015 - 2020
Researcher
2011-2015

- Conducting research projects in grammar error correction, language fluency evaluation of scientific writing, mathematical image to LaTeX translation, PDF to LaTeX conversion.
- Lead the development of research infrastructure within the company.
- Reporting trends and advising company's executives in automation of publishing services.
- Leading the R&D of Grammar Error Correction for English in LaTeX documents to create production line tool.
- Leading the R&D of Automated Subject Index compilation for science books and journals to create production line tool.
- Leading the R&D of PDF to LaTeX conversion of scholar documents.
- Lead the international group of researchers to organize a Shared Task for establishing the state-of-the-art key performance in automated evaluation of scientific writing (AESW 2016).
- Compiled a data-set for the AESW 2016 Shared task.
- Participated in the Build-It-Brake-It NLP 2017 Shared Task and generated sentences to fool sentiment analysis systems.
- Participated in Native Language Identification 2013 Shared Task and developed a system to identify native language of writers in English based on character n-grams.
- Participated in grammar error correction Shared task Helping Our Own 2012 and developed GEC system using HMM and word transformations.
- C++, LibSVM, word2vec, Haskell, LaTeX, Python, PyTorch, k-NN, HMM, scikit-learn, scikit-image, MongoDB, Stanford CoreNLP, Stanford Parser.

2000-2014 **Vytautas Magnus University**

FP7 *Learning and Execution of Action Categories ACAT* (Grant Agr. No. 600578)
Researcher
2013 (4 months)

- Introduced innovative design ideas on how robots could read manuals and perform actions without specific programming in assembling water-pumps (Grundfos).

Faculty of Informatics
Lecturer
2004-2014

- Gave *Natural Language Processing* courses to BA (4 times) and MA (5 times) students in computer science faculty, and MA students in linguistics (2 times).

Centre of Computational Linguistics
Senior engineer-programmer
2000-2008

- Linux server administration: http://donelaitis.vdu.lt (now http://tekstynas.vdu.lt).
- Leading research projects for development of the Contemporary Lithuanian Language Corpus infrastructure and web-service, Parallel English-Lithuanian corpus infrastructure and web-service, tagger for Lithuanian.
- Created Corpus analytics web tool for linguists, which still recent is the main and widely used by linguist researchers in Lithuania.
- Supervised computer science BA students in writing thesis.
- Developed methods for terminology extraction from corpora.
- Lithuanian online news analytics: advertisement trends, terminology.
- **Honors:** *The Corpus of Contemporary Lithuanian Language on the Internet*: **The best scientific work** and the best ICT product 2002 in Lithuania, Issuer - Infobalt Association.
- C++, Java, PHP, MySQL, SOAP services, Haskell, Perl, JavaScript, MPI.

Machine translation expert
2005-2007

- Writing feasibility study and project proposal for English->Lithuanian online translation facility.
- Evaluating project progress.
- Lead the development of technical (servers) and linguistic (parallel corpora and aligners) infrastructures.
- Cooperating with ProMT machine translation company.
- C++, Windows Server 2003, .NET, MS SQL Server, HPC.
- **http://vertimas.vdu.lt**

Education Studies Institute
Administrator
2000

- Created a database with *MS Access* to manage information about students, classrooms and study programs.

2006-2007 ***Lithuanian State Commission of Lithuanian Language***
Expert

- Expertise of Research projects related to Lithuanian language.
- Wrote 2 reviews.

1997 ***"Work&Travel USA" student exchange program:***

- "Merriam Park Painting", St. Paul, USA. **House painter**.
- "Metro Produce Dist.". Minneapolis, USA. **Driver, deliverer**. I passed exams for all driving license categories (ABCDE).

1995–1996 ***St. Anthony Religious Studies Institute at the Faculty of Catholic Theology***
Vytautas Magnus University
Technician of a computer classroom.

- Administrated computer classroom, installing and repairing :) programs.
- Gave lectures to bachelor students on how to use computers with Windows 3.1 and Windows 95 OS, introduced Word, Excel, CorelDraw.

1990–1991 ***Personal company "Nuotrauka"***
Owner, photography.

1989 **My first ever written program** '*Solar system*' to visualize Planetary movements over time with Focal programming language (It had Assembler and Pascal features. The only programing language on the computer ***DVK2*** at that time).

REASEARCH INTERESTS

Natural Language Processing
Machine Learning
Information Extraction and Information Retrieval
Text Classification and Clusterization
Text Mining
Collocation Extraction and Collocation Processing
Grammar Error correction
Natural Language Writing Evaluation
Functional programming
Lexicon extraction

LANGUAGE SKILLS

Lithuanian – native English – proficient Russian – moderate French – background

OTHER SKILLS

Programming languages:	C++, Haskell, Java, Python, JavaScript, Perl, PHP.
Mark-up languages:	HTML, XML, SGML.
Distributed computing:	MapReduce, Hadoop.
Parallel programming:	MPI.
Databases:	MySQL, MongoDB.
Other:	LaTeX, OpenOffice, MS Word, MS Access, MS Excel, Corel Draw.
Machine Learning:	PyTorch, DyNet, word2vec, LibSVM.

HONORS & AWARDS

2002 *The Corpus of Contemporary Lithuanian Language on the Internet*: The best scientific work and the best ICT product 2002 in Lithuania, Issuer - Infobalt Association.

SERTIFICATES

2005 *High Performance Computing: Application Tuning for Clusters*. Intel Software College
2015 *Machine Learning.* Coursera Verified Certificates License GDUK92FKBLA7
2016 *Hadoop Platform and Application Framework*. Coursera Course Certificates License GNSUXPNFQTGG
2017 *DeepLearn 2017 International Summer School on Deep Learning*. Univesity of Deusto

PARTICIPATION IN PROJECTS

2018-2020 *R&D for Development of Services at UAB VTEX* (Grant No J05-LVPA-K-03-0016). Applying Deep Learning for PDF layout recognition.
2017-2020 MC member of COST Action *European Network for Combining Language Learning with Crowdsourcing Techniques*. CA16105.
2013-2015 *Preparation of scientific publishing data and application of intelligent technologies at VTeX*. EU SF VP2-1.3-ŪM-02-K-04-019.
2013 FP7 *Learning and Execution of Action Categories ACAT* (Grant Agr. No. 600578)
2011-2012 *Application of machine learning in the enterprise activities – feasibility study*, EU SF VP2-1.3-ŪM-01-K-02-232.
2009 – 2012 HEalth teXt Analysis network in the Nordic and Baltic countries (HEXAnord). NordForsk foundation. Researcher networks 2009
2007 – 2008 *Internet resources: Annotated corpus of the Lithuanian language and tools of annotation (ALKA 2).* Lithuanian State Science and Studies Foundation
2006 - 2006 6FP IST, Network of excellence project *Resilience for Survivability in IST (ReSIST)* – (IST-4-026764-NOE)
2005 – 2007 *Webpage Information Translation Facility.* BPD2004-ERPF-3.3.0-02-04/0005. (http://vertimas.vdu.lt)
2003 – 2006 *Parallel Lithuanian Language Corpus*. Lithuanian State Commission of the Lithuanian language

2004	*WITFA (Webpage Information Translation Facility).* PHARE - 2002/000-620.05.01-03.07.
2005 – 2006	*The survivability of Lithuanian language under the globalisation circumstances: annotated Lithuanian language corpus (ALKA)*. Lithuanian State Science and Studies Foundation
2001	EU INCO. TELRI-II (Trans European Language Resources Infrastructure – II). Project E*lectronic dictionary of Czech-Lithuanian, Lithuanian-Czech*.

ONLINE TOOLS AND CORPORA

2018-2019	Online Subject Indexing of Books in LaTeX.
2015-2017	LaTeX to Text converter. It expands macros. A must tool when working with LaTeX files. Implemented in Haskell.
2016-2017	Keyphrase extraction from a set of scientific articles.
2016	Automated Evaluation of Scientific Writing Data Set.
2005-2008	Tagger for Lithuanian

SOCIAL ACTIVITIES

2019

1. AAAI-20, a member of the Program Committee.
2. **Program Committee Member** of *the 14th Workshop on Innovative Use of NLP for Building Educational Applications*. ACL workshop. 2019.
3. EMNLP-IJCNLP 2019 reviewer
4. ACL2019 reviewer

2018

5. NAACL2019 Reviewer
6. **Information Extraction Committee Member** of *the 2019 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*. Minnesota, USA.

2017

7. **Program Committee Member** of *the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. EMNLP workshop. 2017. Copenhagen. Denmark.
8. **Program Committee Member** of *the 4th Workshop on NLP Techniques for Educational Applications* (NLPTEA 2017). December 1, 2017, Taipei, Taiwan. IJCNLP 2017 Workshop. (https://sites.google.com/view/nlptea2017/)

2016

9. **Main Organizer** of Automated Evaluation of Scientific Writing (AESW) 2016 Shared Task. Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, Courtney Napoles. In connection with the 11th Workshop on Innovative Use of NLP for Building Educational Applications, ACL, San Diego, CA, USA, June 16, 2016. (http://textmining.lt/aesw/index.html)
10. **Program Committee Member** of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA-3). December 12 at Osaka, Japan in conjunction with COLING 2016. (http://nlptea2016.weebly.com/)

2012

11. ACL-2012, Jeju Island, **volunteer**.

PUBLICATIONS

2019

1. Vidas Daudaravičius. *Textual and Visual Characteristics of Mathematical Expressions in Scholar Documents*. In Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications (ESSD), ACL, Minneapolis, Minnesota, USA, June 06, 2019

2016

2. **Vidas Daudaravicius**, Rafael E. Banchs, Elena Volodina, Courtney Napoles. *A report on the AESW 2016 Shared Task*. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, ACL, San Diego, CA, USA, June 16, 2016
3. **Vidas Daudaravicius**. *A framework for keyphrase extraction from scientific journals*. Semantics,

Analytics, Visualization. Enhancing Scholarly Data: Second International Workshop, SAVE-SD 2016, Montreal, QC, Canada, April 11, 2016, Revised Selected Papers. LNCS 9792. Springer International Publishing

2015

4. **Vidas Daudaravicius**. *Automated Evaluation of Scientific Writing: AESW Shared Task Proposal*. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA'10), Denver, Colorado, June 2015, p. 56-63

2014

5. **Vidas Daudaravicius**. *Language Editing Dataset of Academic Texts*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland. May 2014, p. 1738-1742.
6. Henriksson, H. Moen, M. Skeppstedt, **V. Daudaravicius**, M. Duneld. *Synonym extraction and abbreviation expansion with ensembles of semantic spaces*. Journal of biomedical semantics. London: BioMed Central. 2014, 5:6, p. 1-25.

2013

7. **Vidas Daudaravicius**. *VTEX System Description for the NLI 2013 Shared Task*. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. Atlanta, Georgia. Jun 2013, p. 89-95.
8. **Vidas Daudaravicius**. *Collocation segmentation for text chunking*. PhD Thesis. 2013.

2012

9. Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, **Vidas Daudaravičius** and Martin Hassel. *Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models*. Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM 2012). Edited by: Ananiadou, Sophia; Pyysalo, Sampo; Rebholz-Schuhmann, Dietrich; Rinaldi, Fabio; Salakoski, Tapio. Zurich, 2012. ISBN 978-3-033-03823-3.
10. **Vidas Daudaravicius**. *Applying collocation segmentation to the ACL Anthology Reference Corpus*. In Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 66–75, Jeju Island, Korea, July 2012. Association for Computational Linguistics. ISBN 978-1-937284-29-9
11. **Daudaravicius, Vidas**. *VTEX Determiner and Preposition Correction System for the HOO 2012 Shared Task*. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, June 2012, Montreal, Canada, Association for Computational Linguistics, p. 225-232.
12. **Daudaravičius, Vidas**. *Automatic multilingual annotation of EU legislation with Eurovoc descriptors* // LREC 2012: 8th international conference on Language resources and evaluation, 21-27 May 2012, Istanbul, Turkey: proceedings. Paris: European Language Resources Association. ISBN 9782951740877. p. 14-20.

2011

13. Allvin, Helen; Carlsson, Elin; Dalianis, Hercules; Danielsson-Ojala, Riitta; **Daudaravičius, Vidas**; Martin, Hassel; Kokkinakis, Dimitrios; Lundgren-Laine, Heljä; Nilsson, Gunnar; Nytrø, Øystein; Salanterä, Sanna; Skeppstedt, Maria; Suominen, Hanna; Velupillai, Sumithra. *Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies* // Journal of biomedical semantics. London: BioMed Central. ISSN 2041-1480. vol. 2, suppl. 3, p. 1-11.

2010

14. Allvin, Helen; Carlsson, Elin; Dalianis, Hercules; Danielsson-Ojala, Riitta; **Daudaravičius, Vidas**; Martin, Hassel; Kokkinakis, Dimitrios; Lundgren-Laine, Heljä; Nilsson, Gunnar; Nytrø, Øystein; Salanterä, Sanna; Skeppstedt, Maria; Suominen, Hanna; Velupillai, Sumithra. *Characteristics and Analysis of Finnish and Swedish Clinical Intensive Care Nursing Narratives* // Louhi'10: Proceedings of the NAACL HLT 2010, Second Louhi Workshop on Text and Data Mining of Health Documents, June 5, 2010, Los Angeles, California. Stroudsburg, USA: The Association for Computational Linguistics, 2010. p. 53-60.
15. Costa-jussà, Marta R., **Daudaravičius, Vidas**, Banchs, Rafael E., *Integration of statistical collocation segmentations in a phrase-based statistical machine translation system*. EAMT 2010: Proceedings of the 14th Annual Conference of the European Association for Machine Translation, 27-28 May 2010, Saint-Raphaël, France
16. C. Henríquez, M.R. Costa-jussà, **V. Daudaravicius**, R.E. Banchs, J.B. Mariño (2010) *"UPC-BMIC-VDU system description for the IWSLT 2010: testing several collocation segmentations in a phrase-based SMT system"*, in Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT), pages 189-195, Paris
17. Henríquez, Carlos A. Q.; Costa-jussà, Marta R.; **Daudaravičius, Vidas**; Banchs, Rafael E.; Mariño, B. José. *Using collocation segmentation to augment the phrase table* // WMT'10: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, 15-16 July 2010. Morristown, USA: Association for Computational Linguistics. ISBN 9781932432718. p. 98-102.
18. Costa-jussà, Marta R.; **Daudaravičius, Vidas**; Banchs, Rafael E. Using *collocation segmentation to*

extract translation units in a phrase-based statistical machine translation system // Procesamiento del Lenguaje Natural. Barcelona: Sociedad Espanola para el Procesamiento del Lenguaje Natural. ISSN 1135-5948. 2010, no. 45, p. 215-220.

19. **Daudaravičius, Vidas**. *The influence of collocation segmentation and top 10 items to keyword assignment performance* // Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010: Proceedings. Berlin: Springer. ISBN 9783642121159. p. 648-660.
20. **Daudaravičius, Vidas**. *Automatic identification of lexical units* // Informatica: An International Journal of Computing and Informatics. Ljubljana: Slovensko društvo Informatika. ISSN 0350-5596. Vol. 34, no. 1 (2010), p. 85-91.

2007

21. Utka A., Kovalevskaitė J., Rimkutė E., **Daudaravičius V.** *Bilingual Parallel Corpora for English, Czech and Lithuanian* – The Third Baltic Conference on Human Language Technologies proceedings. Kaunas, 2007, 319–326
22. Gindiyeh M., Čulo O., Grigonytė G., Haller J., Avižienis A., Marcinkevičienė R., **Daudaravičius V.** *A document classification tool for the resilience knowledge base of the resist noe project* – The Third Baltic Conference on Human Language Technologies
23. E. Rimkutė, **V. Daudaravičius**, A. Utka *Morphological Annotation of the Lithuanian Corpus*. – In Proceedings of the Workshop on Balto-Slavonic Natural Language Precessing 2007, Special Theme: Information Extraction and Enabling Technologies, P. 94-99. Association of Computational Linguistics.

2006

24. **Daudaravičius V.** *Opening to infinity: Machine Translation and Lithuanian Language*. (In Lithuanian). Deeds and Days, 2006, Nr. 45, P. 7–18.
25. Rimkutė E., Kovalevskaitė J. **Daudaravičius V.** *The usage and application of multilingual corpora*. (In Lithuanian). Deeds and Days, 2006, Nr. 45, P. 41–62.

2005

26. Zinkevičius V., **Daudaravičius V.**, Rimkutė E. *The Morphologically annotated Lithuanian Corpus – The Second Baltic Conference on Human Language Technologies* proceedings. Tallinn, 2005, 365–370

2004

27. **V. Daudaravičius**, R. Marcinkevičienė Gravity Counts for the Boundaries of Collocations. In *International Journal of Corpus Linguistics*, 2004, Nr 9-2. Amsterdam: John Benjamins Publishing Company
28. Marcinkevičienė R., Bielinskienė A., **Daudaravičius V.**, Rimkutė E. *Corpora for Lithuanian Language Technologies* – In proceedings of *The First Baltic Conference Human Language Technologies. The Baltic Perspective*, 2004, P. 21–24.