

Nilo Pedrazzini

✉ npedrazzini@turing.ac.uk 🌐 npedrazzini.github.io 🌐 nilopedrazzini 🌐 npedrazzini

Skills

- Areas of specialization** 📖 Semantics and discourse coherence; language typology; computational linguistics (word embeddings and Large Language Models). Corpus building and linguistic annotation (low-resource languages and noisy data).
- Programming languages** 📖 Advanced Python, R and Bash. Intermediate Ruby.
- My go-to NLP stack** 📖 MacOS, Jupyter (*local proof-of-concept*) · Ubuntu Azure VMs/HPC, Azure Blob Storage, Bash/SLURM (*cloud development*) · Pandas, BeautifulSoup (*data wrangling*) · spaCy, NLTK (*preprocessing*) · Transformers/PyTorch, Gensim (*language modeling*) · SciPy, PyKrigge, gstat (R) (*interpolation*) · scikit-learn (*clustering, dimensionality reduction*) · Matplotlib/Seaborn, Plotly/Dash, Kepler (*visualization*) · MyBinder (*short-term dissemination*) · HuggingFace (*language model release*) · Weights & Biases (*experiment tracking*).
- Natural languages** 📖 Italian (native), English (C2), Russian (C1), Spanish (Dele B2), Serbian (B1-B2), German (Goethe-Zertifikat B1). Good reading knowledge of Bulgarian, French, and Norwegian (Bokmål). Working knowledge of Ancient Greek, Hittite, Latin, Old Church Slavonic, Arabic (Egyptian), and te reo Māori.

Employment

Research & Academia

- August 2023– 📖 **Turing Research Fellow**. The Alan Turing Institute (London). 🔗
- Feb 2023–July 2023 📖 **Research Software Engineer**. University of Oxford (*Digital Scholarship @ Oxford*) 🔗
- Jan 2022–July 2023 📖 **Research Associate in Corpus-Based Digital Humanities**. The Alan Turing Institute (London) (*Living with Machines*) 🔗👤
- Sept 2021–Mar 2022 📖 **Hourly-Paid Lecturer** in Digital Humanities, King's College London (Department of Digital Humanities).
- Sept 2020–June 2021 📖 **Tutor** in General Linguistics (Oriel College & Lady Margaret Hall, University of Oxford).
- Dec 2020–Aug 2021 📖 **Research Assistant**. Oxford School of Global and Area Studies, International Multimodal Communication Centre (IMCC) (*Depictions of Post-COVID-19 Futures in Russian International Media: Multimodal Viewpoint Analysis*) 🔗.
- Feb–Jun 2020 📖 **Research Assistant**. Department of Experimental Psychology, University of Oxford (*Read-Oxford*) 🔗

Industry & Service

- 2018 – 2020 📖 **AV/IT Technician**. Somerville College, University of Oxford.
- 2016 – 2019 📖 **Assistant Manager** (full-time). Le Pain Quotidien, London & Oxford.
- 2014 – 2016 📖 **Front of House** (full-time). Le Pain Quotidien, London.
- Apr–Jun 2015 📖 **Trainee Assistant Curator** (full-time). Slavonic Collections, The British Library.

Education

- 2023 📖 **DPhil Linguistics, Philology & Phonetics**, University of Oxford (St Hugh's College). Thesis: *A quantitative and typological study of Early Slavic participle clauses and their competition*. ESRC full scholarship (Grand Union DTP). Supervisor: Dr. Hanne M. Eckhoff.

Education (continued)

- Autumn 2020  **Visiting PhD student**, University of Oslo, Dept. of Linguistics and Scandinavian Studies. Supervisor: Prof. Dag Haug.
- 2019  **MSt Slavonic Studies (Serbian)**, University of Oxford (Somerville College). Bosnian-Croatian-Serbian language; history of Church Slavonic; Cyrillic palaeography. Supervisor: Dr. Hanne Eckhoff. Grade: Distinction.
- 2015 – 2017  **MPhil Theoretical and Applied Linguistics (70% complete)**, Università di Pavia. Thesis: *Secondary uses of the Russian second-genitive*. Supervisor: Prof. Silvia Luraghi. Exams: computational linguistics, first-order logic, advanced phonology, language typology, Celtic linguistics, Hittite, advanced Russian, Slavic philology. Grade: 29.2/30.
- 2015  **BA Modern Languages and Cultures (Russian and English)**, Università di Pavia. Thesis: *Old Believers' syllabic poetry and the Russian Baroque: Anonymous lamentations from the first half of the 18th century*. Supervisor: Prof. Maria C. Bragone. Grade: 110/110 Summa cum Laude.

Academic Service

Affiliations

- 2021–now  **Fellow**. Reproducibility Research Oxford (RROx) - UK Reproducibility Network (UKRN). I am specifically involved in discussing and promoting best Open Humanities practices.

Editorial

- 2021 – 2023  **Editorial Assistant**, Journal of Open Humanities Data (JOHD), Ubiquity Press.
- 2022  **Section Editor**, Journal of Open Humanities Data (JOHD), Ubiquity Press.
- 2020 – 2021  **Copyeditor**, Journal of Open Humanities Data (JOHD), Ubiquity Press.

Events

- 2024  Co-organizer of the workshop *Working with big historical data from maps and newspapers* @ The Alan Turing Institute. 
-  Co-organizer of the workshop *Working with big historical data from maps and newspapers* @ Lancaster University.
- 2022  Main organizer and chair of the *Open Humanities Seminar Series* (OHSS). 
-  Co-organizer of session on *Open Data @ Humanities & Data Science discussion group*, The Alan Turing Institute. 

Supervision

- 2024  Miriam Eisenbruch (Oxford, DPhil Linguistics) (technical supervision)
- 2021  Malithi Alahapperuma (UCL/Turing, MSc Digital Humanities) (work placement supervision)

Admin

- 2018 – 2023  **Undergraduate admissions interviewer** for candidates to BA in Modern Languages & Linguistics and BA in Psychology, Philosophy & Linguistics (Lady Margaret Hall, University of Oxford).
- 2019 – 2023  **Corpus building & annotation**. Tromsø Old Russian and Old Church Slavonic Treebank (TOROT)

Awards

- 2024  *Publications 2022 Best Paper Award for 10th Anniversary* 

Awards (continued)

- 2022  Circle U. Prize Award (PI: Barbara McGillivray), as a contributor to the project *The Language of Mechanization*, part of the *Living with Machines* project (The Alan Turing Institute), for innovative and novel approaches to inter- and trans-disciplinary research 
- 2020  Overseas Institutional Visit Award (Economic and Social Research Council), for a two-month visiting research period at the University of Oslo, Norway
 -  Digital Humanities Summer School (DHOx2020) merit-based bursary
- 2019-2023  Economic and Social Research Council (Grand Union DTP) Full PhD Scholarship
- 2019  Somerville College Archibald Jackson Prize

Teaching

University lectures

- Oct-Nov 2023  *Introduction to Text Mining with Python* (MSc in Digital Scholarship, University of Oxford).
- Sept-Dec 2021  *The Cultural Web: Building a Humanities Website*. (HTML/CSS, Jekyll, Git) (King's College London).

University tutorials

- April-June 2021  *Historical Linguistics* (University of Oxford).
- Jan-Mar 2021  *Semantics & Pragmatics* (University of Oxford).
 -  *Sociolinguistics* (University of Oxford).
- Sept-Dec 2020  *General Linguistics Foundations* (University of Oxford).
 -  *Psycholinguistics* (University of Oxford).

Summer schools

- 1-7 Jul 2023  *Applied Data Analysis*, Digital Humanities at Oxford Summer School (DHOxSS).
- 15 Jul 2022  *An introduction to Word2Vec*, Digital Humanities at Oxford Summer School (DHOxSS).

Workshops

- 30 Sept-2 Oct 2024  *Working with big historical data from maps and newspapers* workshop (London).
- 5-7 June 2024  *Working with big historical data from maps and newspapers* workshop (Lancaster).

Presentations

Invited lectures and talks

- 1 “T*xt as data: A (very small) sampler of (very cool) research projects in computational text analysis” (May 2024). Digital and Information Skills Community of Practice, Bodleian Libraries (University of Oxford).
- 2 “Language typologies using parallel text data and geo-statistics” (Apr. 2024). Office for Statistics Regulation (UK Statistics Authority; UKSA, London).
- 3 “T*xt as data: A sneak peek into the possibilities unleashed by computational methods” (May 2023). Creative & Heritage Innovation Challenge Workshop (Oxford, United Kingdom).
- 4 “Text mining per le Humanities: un’introduzione pratica con Python” (Mar. 2023). Hands-on introduction to text mining with Python for doctoral students at the University of Pavia (Italy).  URL: <https://github.com/npedrazzini/pavia-text-mining>.
- 5 “The Language of Mechanisation” (Nov. 2022). Presenting the Circle-U award-winning project at a round table at the Conference on interdisciplinary and transdisciplinary research for sustainable development (UCLouvain, Louvain-la-neuve, Belgium).  URL: <https://youtu.be/zx8HXmw0q4>.

Updated on September 19, 2024

- 6 “An introduction to Word2Vec” (July 2022). Digital Humanities at Oxford Summer School (“Text to Tech” strand), Oxford (United Kingdom). [URL: https://github.com/Living-with-machines/dhoxss-text2tech](https://github.com/Living-with-machines/dhoxss-text2tech).

Conference presentations

- 1 “GramTypix: a Python pipeline for subtoken-based language typology and probabilistic semantic mapping from parallel corpora” (Aug. 2024). Digital Humanities Conference (Arlington, Virginia).
- 2 “Mapping ‘when’-clauses in Latin American and Caribbean languages: an experiment in subtoken-based typology” (June 2024). Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), co-located with the North American Chapter of the Association for Computational Linguistics (NAACL) conference (Mexico City).
- 3 “Temporal clauses in Old Church Slavonic and Ancient Greek: typological differences emerged from massively parallel corpora” (Aug. 2023). 56th Annual Meeting of the Societas Linguistica Europaea (SLE, Athens, Greece), joint work with Dag Haug. [URL: https://societaslinguistica.eu/sle2023/](https://societaslinguistica.eu/sle2023/).
- 4 “Metadata Enrichment in the Living with Machines Project: User-focused Collaborative Database Development in a Digital Humanities Context” (July 2023). ADHO Digital Humanities Conference (DH2023, Graz, Austria), joint work with Kalle Westerling, David Beavan, Kaspar Beelen, Mariona Coll Ardanuy, Timothy Hobson, Christina Last, Griffith Reese, and Like Hare. [URL: https://dh2023.adho.org](https://dh2023.adho.org).
- 5 “Diachronic and diatopic word embeddings from British historical newspapers” (June 2023). 12th AIUCD (Convegno dell’Associazione per l’Informatica Umanistica e la Cultura Digitale), joint work with Barbara McGillivray. [URL: http://www.aiucd2023.unisi.it](http://www.aiucd2023.unisi.it).
- 6 “Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers” (Nov. 2022). The 2nd International Workshop on Natural Language Processing for Digital Humanities – NLP4DH 2022 (AACL-colocated, Taipei, Taiwan), with Barbara McGillivray. [URL: https://rootroo.com/en/nlp4dh-2022/](https://rootroo.com/en/nlp4dh-2022/).
- 7 “Evaluating Language Models for Ancient Greek: Design, Challenges, and Future Directions” (Aug. 2022). International Conference on Historical Linguistics 2022 (ICHL25: Oxford), with Barbara McGillivray, Malvina Nissim, Saskia Peels-Matthey & Silvia Stopponi. [URL: https://ichl.ling-phil.ox.ac.uk/abstracts/168](https://ichl.ling-phil.ox.ac.uk/abstracts/168).
- 8 “Historic machines from ‘prams’ to ‘Parliament’: new avenues for collaborative linguistic research” (June 2022). DH Benelux 2022: RE-MIX. Creation and alteration in DH, with Mia Ridge, Giorgia Tolfo, Kalle Westerling and Barbara McGillivray. [URL: https://zenodo.org/record/6578021#.YtjqFy8w1Yh](https://zenodo.org/record/6578021#.YtjqFy8w1Yh).
- 9 “Deep impact: A study on the impact of data papers and datasets in the humanities and social sciences” (June 2022). SciDataCon (International Data Week, Seoul, 20-23 June 2022), with Marton Ribary, Mandy Wigdorowitz, Eleonora Zordan, Barbara McGillivray. [URL: https://www.scidatacon.org/IDW-2022/sessions/400/paper/975/#](https://www.scidatacon.org/IDW-2022/sessions/400/paper/975/#).
- 10 “Le Journal of Open Humanities Data (JOHD): enjeux et défis dans la publication de data papers pour les sciences humaines” (Nov. 2021). DHNord2021 round table “Le data paper: une nouvelle forme de publication scientifique en SHS”, with Paola Marongiu, Marton Ribary & Barbara McGillivray. [URL: https://meshs.fr/page/enjeux_et_defis_dans_la_publication_de_data_papers_pour_les_shs](https://meshs.fr/page/enjeux_et_defis_dans_la_publication_de_data_papers_pour_les_shs).
- 11 “Tackling lack of linguistic data with HTR: A specialized model for the transcription of Serbian Church Slavonic manuscripts” (Apr. 2021). El’ Manuscript Conference 2020. [URL: http://www.elmanuscript2020.uni-freiburg.de/wordpress/wp-content/uploads/2021/04/ElManuscript-2021-Booklet-of-Abstracts_09.04.2021.pdf](http://www.elmanuscript2020.uni-freiburg.de/wordpress/wp-content/uploads/2021/04/ElManuscript-2021-Booklet-of-Abstracts_09.04.2021.pdf).
- 12 “Same question, different annotation depths: early Slavonic dative absolute in deeply versus shallowly annotated treebanks” (Nov. 2020). Grammar & Corpora International Conference, Kraków (Poland) [held as online event]. [URL: https://gac2020.ijp.pan.pl/GaC_2020_paper_25.pdf](https://gac2020.ijp.pan.pl/GaC_2020_paper_25.pdf).
- 13 “Exploiting Cross-Dialectal Gold Syntax for Low-Resource Historical Languages: Towards a Generic Parser for Pre-Modern Slavic” (Nov. 2020). Computational Humanities Research Workshop, Amsterdam (The Netherlands).

- 14 “Participial free adjuncts in early Slavic treebanks: The discourse properties of the dative absolute” (Sept. 2020). Slavic Linguistics Society (SLS) Conference, Bloomington (IN, USA) [held as online event].  URL: <https://slavicgf.sitohost.iu.edu/SLS/abstracts/Pedrazzini.PDF>.
- 15 “Challenges in historical corpus analyses of discourse relations: a case study in Old Church Slavonic” (June 2020). Corpora & Discourse International Conference, University of Sussex (UK) [held as online event].

Publications

Peer-reviewed articles

- 1 **Pedrazzini, Nilo** (2024). “Mapping ‘when’-clauses in Latin American and Caribbean languages: an experiment in subtoken-based typology”. In: *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*. Ed. by Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense. Mexico City, Mexico: Association for Computational Linguistics, pp. 24–33.  URL: <https://aclanthology.org/2024.americasnlp-1.4>.
- 2 Ridge, Mia, **Nilo Pedrazzini**, Miguel Vieira, Arianna Ciula, and Barbara McGillivray (2024). “Language of Mechanisation crowdsourcing datasets from the Living with Machines project”. In: *Journal of Open Humanities Data* 10 (1), pp. 1–9.  DOI: 10.5334/johd.195.
- 3 Stopponi, Silvia, **Nilo Pedrazzini**, Saskia Peels-Matthey, Barbara McGillivray, and Malvina Nissim (2024). “Natural language processing for Ancient Greek: Design, advantages and challenges of language models”. In: *Diachronica*.  DOI: 10.1075/dia.23013.sto.
- 4 Haug, Dag and **Nilo Pedrazzini** (2023). “The semantic map of *when* and its typological parallels”. In: *Frontiers in Communication* 8.  DOI: 10.3389/fcomm.2023.1163431.
- 5 Stopponi, Silvia, **Nilo Pedrazzini**, Saskia Peels, Barbara McGillivray, and Malvina Nissim (2023). “Evaluation of Distributional Semantic Models of Ancient Greek: Preliminary Results and a Road Map for Future Work”. In: *Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023*, pp. 49–58.  URL: <https://aclanthology.org/2023.alp-1.6>.
- 6 **Pedrazzini, Nilo** and Barbara McGillivray (2022). “Machines in the media: semantic change in the lexical field of mechanization in 19th-century British newspapers”. In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*. Association for Computational Linguistics (ACL), pp. 85–95.  URL: <https://aclanthology.org/2022.nlp4dh-1.12>.
- 7 McGillivray, Barbara, Paola Marongiu, **Nilo Pedrazzini**, Mandy Wigdorowitz, and Eleonora Zordan (2022). “Deep Impact: A study on the impact of data papers and datasets in the Humanities and Social Sciences”. In: *Publications* 10.4 (39).  DOI: 10.3390/publications10040039.
- 8 **Pedrazzini, Nilo** (2022). “One question, different annotation depths: A case study in Early Slavic”. In: *Journal of Historical Syntax* (Special issue: “Annotating Historical Corpora”) 6.7, pp. 1–40.  DOI: 10.18148/hs/2022.v6i4-11.96.
- 9 **Pedrazzini, Nilo** and Hanne Martine Eckhoff (2021). “OldSlavNet: A scalable Early Slavic dependency parser trained on modern language data”. In: *Software Impacts*, p. 100063.  DOI: 10.1016/j.simpa.2021.100063.
- 10 **Pedrazzini, Nilo** (2020). “Exploiting Cross-Dialectal Gold Syntax for Low-Resource Historical Languages: Towards a Generic Parser for Pre-Modern Slavic”. en. In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*. Ed. by Folgert Karsdorp, Barbara McGillivray, Adina Nerghees, and Melvin Wevers. Vol. 2723. Amsterdam, Netherlands: CEUR Workshop Proceedings, pp. 237–247.  URL: <http://ceur-ws.org/Vol-2723/short48.pdf>.

Chapters in books

- 1 McGillivray, Barbara, **Nilo Pedrazzini**, Arianna Ciula, Jon Lawrence, Tiffany Ong, Mia Ridge, and Miguel Vieira (Forthcoming). “Analysing the language of mechanisation in 19th-century British newspapers”. In: *Living with Machines*. Ed. by Ruth Ahnert, Emma Griffin, Jon Lawrence, and the Living with Machines Team. London: University of London Press.

Theses

- 1 **Pedrazzini, Nilo** (2023). "A quantitative and typological study of Early Slavic participle clauses and their competition". PhD thesis. University of Oxford.  URL: <https://ora.ox.ac.uk/objects/uuid:27636bb4-efd0-47bf-8e28-dce5a2b2ca4d>.

Other contributions

Blog posts

- 1 *Fairytales on trial: the Good and the Beautiful in early-Soviet children's literature* (Mar. 2019). British Library, European Studies Blog.  URL: <https://blogs.bl.uk/european/2019/03/fairytales-on-trial.html>.
- 2 *A reluctantly modern voice from the 17th-century Russian storm: Archpriest Avvakum and the Life written by himself* (June 2015). British Library, European Studies Blog.  URL: <https://blogs.bl.uk/european/2015/06/a-reluctantly-modern-voice-from-the-17th-century-russian-storm.html>.

Media

- 1 *Living with Machines: On Collaboration*. Docuseries episode. (2023).  URL: https://www.youtube.com/watch?v=A__ZJgw4_00.
- 2 *Living with Machines: The Language of Mechanisation*. Docuseries episode. (2023).  URL: <https://www.youtube.com/watch?v=qSciC6ymLK0>.