

# Kaja Dobrovoljc

Ljubljana, Slovenia | [Website](#) | [ORCID](#) | [Google Scholar](#) | [LinkedIn](#)

Corpus linguist with 15+ years of experience in language resource development, linguistic data analysis, and language technology evaluation. Specialized in annotation workflows, lexicogrammatical data analysis, and cross-linguistic research. Experienced in independent research, interdisciplinary collaboration, and knowledge dissemination.

## CURRENT EMPLOYMENT

---

- |              |  |
|--------------|--|
| 2019–present | Research Fellow, University of Ljubljana, Faculty of Arts, Dept. of Slovene Studies <ul style="list-style-type: none"><li>Research and dissemination in national and international projects of the <i>Language Resources and Technologies for Slovene</i> Research Group (UL FF, UL FRI)</li><li>Development and maintenance of corpora, dictionaries, online services, data processing tools at the <i>Centre for Language Resources and Technologies</i> (CJVT UL)</li><li>Student supervision</li></ul> |
| 2018–present | Research Associate, Jožef Stefan Institute, Artificial Intelligence Laboratory <ul style="list-style-type: none"><li>Support for NLP research, incl. linguistic data management and evaluations</li><li>Development and dissemination of the CLARIN.SI infrastructure</li></ul>  |

## PAST EMPLOYMENTS

---

- |           |   |
|-----------|---|
| 2018-2022 | Post-doctoral researcher, Faculty of Comp. and Information Science, Uni. of Ljubljana   |
| 2018–2019 | Post-doctoral researcher, Faculty of El. Engineering and Comp. Science, Uni. of Maribor |
| 2013–2018 | Young researcher, Trojina Institute for Applied Slovene Studies                         |
| 2011–2013 | Freelance Lexicographer and Translator  |

## EDUCATION

---

- |      |  |
|------|--|
| 2018 | PhD in Slovene Studies, University of Ljubljana, Faculty of Arts<br>Dissertation: <i>Lexical Features of Spoken Language in User-Generated Content</i>       |
| 2011 | University diploma in Translation Studies, University of Ljubljana, Faculty of Arts<br>University graduate translator for English and French (MA equivalent) |
| 2005 | Gimnazija Bežigrad, Ljubljana  |

## ADDITIONAL TRAINING

---

- |      |   |
|------|---|
| 2024 | Leveraging Science with AI Tools, Crowdhelix webinar  |
| 2022 | COST Academy Leadership workshop, Brussels, Belgium   |
| 2021 | Corpus Linguistics Summer School, University of Birmingham                                  |
| 2020 | 'Introduction to Statistics' Workshop, Faculty for Social Sciences, University of Ljubljana |
| 2016 | TextLink Training School, University of Valencia, Spain                                     |
| 2015 | Parseme Training School, Charles University, Prague, Czech Republic                         |
| 2015 | Introduction to programming, Faculty of Computer and Information Science                    |
| 2013 | Learn to Program: The Fundamentals, Coursera online course                                  |
| 2011 | European Summer School in Logic Language and Information, Ljubljana, Slovenia               |
| 2010 | CLARA Training School on Treebank Annotation, Charles University, Prague                    |

## RESEARCH VISITS

---

- |      |  |
|------|--|
| 2015 | Prof. Joakim Nivre, Department of Linguistics and Philology, Uppsala University;<br>PARSEME COST Action Short-Term Scientific Mission Grant (2 months) |
|------|--|

## RESEARCH PROJECTS

---

- |              |  |
|--------------|--|
| 2024–present | LLM4DH: Large Language Models for Digital Humanities (Task leader)                   |
| 2022–present | SPOT: Treebank-Driven Approach to the Study of Spoken Slovenian (Project Leader)     |
| 2024         | STARK 2: Enhancement of the STARK tool for analysing parsed corpora (Project Leader) |
| 2022–2023    | SLOKIT: CLARIN.SI service for corpus data analysis and summarization (WP Leader)     |
| 2022         | CLARIN.SI: Drevesnik online service for Slovenian treebank querying (Project Leader) |

|           |  |
|-----------|--|
| 2021–2023 | RSDO: Development of Slovene in a Digital Environment (Task Leader)                    |
| 2021–2022 | SLED: Monitor Corpus for Slovene and Related Language Resources                        |
| 2019–2022 | ELEXIS: European Lexicographic Infrastructure  |
| 2017–2020 | New grammar of contemporary standard Slovene: sources and methods (WP leader).         |
| 2019–2020 | KOLOS: Collocation as a basis for lang. description: semantic and temporal perspective |
| 2019      | STARK: dependency tree extraction tool (Project Leader)                                |
| 2018–2019 | Slovene in the palm of your hand: interactive e-environment for teaching Slovene       |
| 2014–2015 | Language Resources Portal (Co-developer)   |
| 2012–2013 | Language Technology Seminars for Teachers (Coordinator)                                |
| 2012      | xLike - Cross Lingual Knowledge Extraction   |
| 2012      | PISA Dependency Treebank for Slovene Student Assessment Texts                          |
| 2010–2013 | SSJ: Communication in Slovene  |

## RESEARCH OUTPUTS

---

|              |  |
|--------------|--|
| 2012–present | 10+ journal papers, among these 6 with first authorship<br>10+ book chapters, among these 7 with first authorship<br>40+ conference papers/presentations<br>60+ language resources deposited at the CLARIN.SI repository<br>10+ computational tools and services<br>3 invited talks<br>800+ citations (Google Scholar) |
|--------------|--|

## ORGANIZATION OF SCIENTIFIC MEETINGS

---

|           |   |
|-----------|---|
| 2025      | SyntaxFest 2025, Ljubljana, Slovenia: Head of Organizing Committee  |
| 2023      | UniDive 2 <sup>nd</sup> General Meeting, Naples, Italy: Co-chair of the WG1 Day   |
| 2016–2024 | Language Technologies and Digital Humanities Conference Series, Ljubljana, Slovenia:<br>OC member in 2016, 2018, 2022, 2024 |
| 2021      | EACL 2021 Language Diversity Games: OC member   |
| 2019      | Text, Speech and Dialogue Conference (TSD 2019), Ljubljana, Slovenia: OC Member   |
| 2018      | EURALEX International Congress, Ljubljana, Slovenia: OC Member  |
| 2016      | Language Technologies and Digital Humanities Conference 2016, Ljubljana, Slovenia:<br>Chair of Student Research Track       |

## EXPERT COMMITTEES

---

|              |   |
|--------------|---|
| 2024         | LREC-COLING 2024 Area Chair (Parsing, Tagging, Chunking, Grammar, Morphosyntax)   |
| 2023–present | Member of the EUTOPIA Connected Community: Grounding Human-Centred AI on Embodied Multimodal Interaction                  |
| 2023         | ESSLLI 2023 Programme Committee Member (Local PC Co-Chair)  |
| 2022–present | CA21167 COST Action UniDive: Universality, Diversity and Idiosyncrasy in Language Technology: MC Member and WG1 Co-leader |
| 2015–present | CLARIN.SI Consortium Management Committee Member  |
| 2014–present | Member of the ACL SIGLEX-MWE and SIGANN Sections  |
| 2013–present | Member of the Slovenian Language Technologies Society (MC member since 2015)  |
| 2013–2015    | Member of Young Academy Society   |
| 2015–2018    | MC Substitute at COST Action TextLink: Structuring Discourse in Multilingual Europe                                       |
| 2014–2017    | MC Substitute at COST Action Parseme: Parsing and multi-word expressions  |

## EDITORIAL BOARD MEMBER

---

|              |  |
|--------------|--|
| 2023–present | GOS 2: Corpus of Spoken Slovene Concordancer     |
| 2022–present | Oznacevalnik Online Text Annotation Services     |
| 2020–present | Jezikovni sledilnik Language Tracker             |
| 2019–present | Gigafida 2.0: Corpus of Written Standard Slovene |
| 2019–present | Sloleks 2.0: Slovene Morphological Lexicon       |
| 2018–present | Collocations Dictionary of Modern Slovene        |
| 2017–present | Thesaurus of Modern Slovene                      |

## **REVIEWING ACTIVITIES**

---

- 2018–present Journal paper reviewer: Language Resources and Evaluation (Springer Nature), Journal of Pragmatics (Elsevier), Humanities and Social Sciences Communications (Nature), Slovenščina 2.0 (Scopus), Slavia Centralis, Rasprave
- 2016–present Program committee member: International Conference on Language Resources and Evaluation (LREC), International Conference on Computational Linguistics (COLING), Text, Speech and Dialogue (TSD), Language Technologies and Digital Humanities Conference (JTDH), Universal Dependencies Workshop, SyntaxFest (Quasy, Depling, TLT, UDW), Workshop on the Semantics and Pragmatics of Dialogue (SemDial)
- 2021 Book reviewer for 'Slovene in the Palm of Your Hand 4' scientific monograph
- 2021 Book chapter reviewer for 'New Grammar of Contemporary Standard Slovene: Sources and Methods' scientific monograph

## **TEACHING**

---

- 2022 Workshop on CLARIN.SI language resources and services for UL PhD students
- 2019 Workshop on Q-CAT corpus annotation tool
- 2015–2017 Workshops on SketchEngine corpus concordancing tool
- 2015–2016 Workshops on WebAnno CLARIN.SI annotation platform
- 2015 Workshop on Nooj corpus processing tool

## **AWARDS AND SCHOLARSHIPS**

---

- 2020 E-Pub Special Mention Award at the Slovenian Book Fair (Gigafida 2.0 Reference Corpus)
- 2018 University of Ljubljana Best Research Achievement (Thesaurus of Modern Slovene)
- 2013–2015 Škrabec Association Scholarship for promising students of linguistics
- 2006 High distinction high-school graduate (Zlata matura)
- 2001–2010 Zois National Scholarship for gifted students