
**Aktuelnosti sa konferencije *LREC2022***

(*Language Resources and Evaluation Conference*)

Ranka Stanković

7.7.2022. JeRTeh seminar

LREC 2022 - LREC 2022 (lrec-conf.org)

# Palais du Pharo

LREC 2022 - Palais du Pharo (lrec-conf.org)

The Palais du Pharo was built in1858 by Napoleon III for his wife Eugénie .
Overlooking the entrance of Marseille Old Harbour, it is now a Conference Center owned by the City of Marseille.
http://palaisdupharo.marseille.fr/

# Keynote and Invited Speakers

We are glad to announce

**our two Keynote speakers at LREC 2022**

**Dr Julia Parish-Morris**
Dept of Psychiatry, Perelman School of Medicine
University of Pennsylvania
https://bit.ly/3MnVF5P

Language Resources for charting Linguistic Diversity in Neuroexpansive Populations

**June 21, 2022 at 14:30 (UTC+2)**

**Emmanuel Dupoux**
EHESS
https://bit.ly/3FQdd80

Textless NLP: towards language processing from raw audio

**June 22, 2022 at 14:30 (UTC+2)**

**our French Invited speakers**

**Dr Philippe Boula de Mareüil**
LISN CNRS
https://perso.limsi.fr/mareuil/

The languages of France illustrated by a Speaking Atlas

**June 22, 2022 at 12:30 (UTC+2)**

**Professor Henri-José Deulofeu**

Why we do not speak in Marseilles as they write in Paris?and consequently why NLP needs crucially more datafrom spontaneous speech ecological recordings

**June 23, 2022 at 12:30 (UTC+2)**

# ELRA Antonio Zampolli Prize

**Share this page!**

Steven Bird - Home

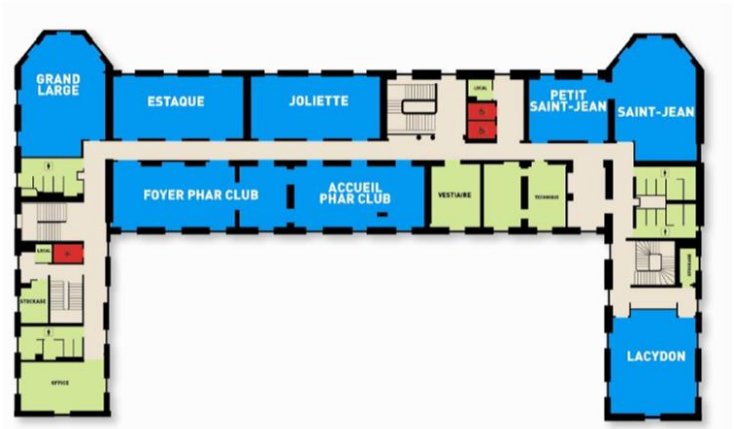
**Outstanding Contributions to the Advancement of Language Resources and Language Technology Evaluation within Human Language Technologies**

The ELRA Board has created a prize to honour the memory of its first president, professor Antonio Zampolli, a pioneer and visionary scientist who was internationally recognized in the field of computational linguistics and Human Language Technologies (HLT). He also contributed much through the establishment of ELRA and the LREC conference. To reflect Antonio Zampolli's specific interest in our field, the Prize will be awarded to individuals whose work lies within the areas of Language Resources and Language Technology Evaluation with acknowledged contributions to their advancements. The ELRA Antonio Zampolli Prize was awarded to:

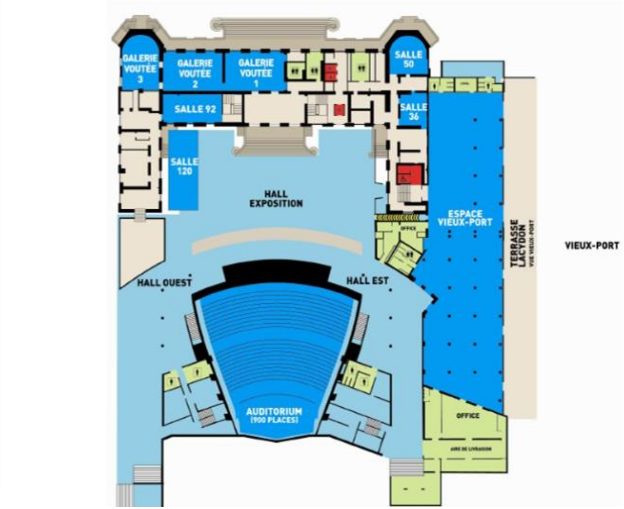
- **Fredrick Jelinek**, from John Hopkins University, Baltimore (USA), at **LREC 2004**, in Lisbon. The presentation can be viewed here.
- **Christiane Fellbaum** and **George A. Miller**, from Princeton University, Princeton (USA), at **LREC 2006**, in Genoa. The presentation can be viewed here.
- **Yorick Wilks**, from the Oxford Internet Institute and the Computer Science Department of the University of Sheffield (UK), at **LREC 2008**, in Marrakech. The presentation can be viewed here.
- **Mark Liberman**, from the University of Pennsylvania (USA), at **LREC 2010**, in Malta. The presentation can be viewed here.
- **Charles Fillmore** and **Collin Baker**, from the International Computer Science Institute (ICSI), University of California Berkeley (USA), and **Oriental Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (Oriental COCOSDA)** at **LREC 2012**, in Istanbul. The presentations are not available.
- **Alex Waibel** from Carnegie Mellon University (USA) and Karlsruhe Institute of Technology (Germany) at **LREC 2014**. His presentation can be seen in video on Youtube.
- **Roger K. Moore** from University of Sheffield (UK) at **LREC 2016**.The presentation can be viewed here.
- **Eva Hajičová** from Charles University, Prague, (Czech Republic) at **LREC 2018**.The presentation will be shortly available.


Pharo Old Palace Level 0


Pharo Old Palace Level 2




Pharo Old Palace Level 1




Pharo Extension Level -1


Proceedings of The 13th Language Resources and Evaluation Conference (lrec-conf.org)

**Full proceedings volume (PDF) | Programme | Author index | Bibliography (BibTeX) | Editors**

| pdf | bib | Papers | pages |
|---|---|---|---|
| pdf | bib | *Domain Adaptation in Neural Machine Translation using a Qualia-Enriched FrameNet*<br>Alexandre Diniz da Costa, Mateus Coutinho Marim, Ely Matos and Tiago Timponi Torrent | pp. 1-12 |
| pdf | bib | *HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation*<br>Serge Gladkoff and Lifeng Han | pp. 13-21 |
| pdf | bib | *Priming Ancient Korean Neural Machine Translation*<br>chanjun park, Seolhwa Lee, Jaehyung Seo, Hyeonseok Moon, Sugyeong Eo and Heuiseok Lim | pp. 22-28 |
| pdf | bib | *GECO-MT: The Ghent Eye-tracking Corpus of Machine Translation*<br>Toon Colman, Margot Fonteyne, Joke Daems, Nicolas Dirix and Lieve Macken | pp. 29-38 |
| pdf | bib | *Introducing Frege to Fillmore: A FrameNet Dataset that Captures both Sense and Reference*<br>Levi Remijnse, Piek Vossen, Antske Fokkens and Sam Titarsolej | pp. 39-50 |
| pdf | bib | *Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR Lexicon*<br>Bolette Pedersen, Nathalie Carmen Hau Sørensen, Sanni Nimb, Ida Flørke, Sussi Olsen and Thomas Troelsgård | pp. 51-60 |
| pdf | bib | *Sense and Sentiment* | pp. 61-69 |

# Stitistički gledano

- 38 radionica (od 50 prijavljenih)
- 9 tutorijela (od 12)
- 1302 submitovana rada 804 prihvaćeno (~62%):
  - 156 Orals,
  - 416 Posters,
  - 232 Remote.
- 1600 registracija
  - 942 prijavljenih uživo (verovatno više)
- 3001 autor
- Ukupno 804 rada na 7391 strana
  - Oral: O1-O39
  - Poster: P1-P40
  - Remote Papers, sesije R1-R21
- Po 4 sesije paralelno + poster sesija (sem predavanja
  - Šta odabrati?

**LREC 2022 - Programme - Jour 1**

**Mardi 21 juin 2022**

09:30 - 11:00 Cérémonie d'inauguration - Auditorium
11:00 - 11:20 ELRA : Les 25 prochaines années - Auditorium
*11:20 - 11:40 Pause café*
11:40 - 13:00 Sessions techniques : présentations orales et affichées
*13:00 - 14:30 Pause déjeuner*
14:30 - 15:10 Conférencière invitée : Julia Parish-Morris - Auditorium
15:15 - 16:35 Sessions Orales et Poster
*16:35 - 16:55 Pause café*
16:55 - 18:15 Sessions techniques : présentations orales et affichées
18:20 - 19:30 Réunion générale d'ELRA - Auditorium
20:00 LREC 2022 Cocktail de bienvenue - Palais du Pharo

**LREC 2022 - Programme - Jour 2**

**Mercredi 22 juin 2022**

09:30 - 10:50 Sessions techniques : présentations orales et affichées
09:30 - 10:50 Session Industrielle -Mucem
*10:50 - 11:10 Pause café*
11:10 - 12:30 Session Industrielle-Mucem
11:10 - 12:30 Sessions techniques : présentations orales et affichées
12:30 - 13:00 Conférencier invité Local : Philippe Boula de Mareüil - Auditorium
*13:00 - 14:30 Pause déjeuner*
14:30 - 15:10 Conférencier Invité Emmanuel Dupoux - Auditorium
*15:10 - 15:15 Courte pause (5mn)*
15:15 - 16:35 Sessions techniques : présentations orales et affichées
*16:35 - 16:55 Pause café*
16:55 - 18:35 Sessions techniques : présentations orales et affichées

**LREC 2022 - Programme - Jour 3**

**Jeudi 23 juin 2022**
09:30 - 10:50 Sessions techniques : présentations orales et affichées
*10:50 - 11:10 Pause café*
11:10 - 12:30 Sessions techniques : présentations orales et affichées
12:30 - 13:00 Conférencier local invité José Deulofeu - Auditorium
*13:00 - 14:30 Pause déjeuner*
14:30 - 15:10 Prix Antonio Zampolli : Conférence Invitée - Auditorium
*15:10 - 15:15 Courte pause (5mn)*
15:15 - 16:35 Sessions techniques : présentations orales et affichées
*16:35 - 16:55 Pause café*
16:55 - 18:00 LREC 2022 Cérémonie de Clôture- Auditorium
20:00 LREC 2022 GALA Dîner de Gala- RoofTop

# Čemu se teži

- **Multilingualism – and equal treatment of all languages –**
  - **is an essential feature of LREC, as it is the attempt of putting the text, speech and multimodal communities together as well as academics and industrials.**

- **LREC values topics – such as Less-resourced languages or Infrastructural issues, strategies and policies – that may not easily find proper venues in other big conferences. Research is strongly affected also by infrastructural (metaresearch) activities, really needed for our field to progress.**

- **LREC wants to be an "inclusive" conference: this is for us a very important feature.**

- **In 2021 Google Scholar Metrics h5-index, LREC ranks 6th of the big conferences/journals in Computational Linguistics.**

# The most popular areas

- (more than 100 submissions)
    - Corpora and Annotation (including Tools, Systems, Treebanks)
    - Information Extraction and Information Retrieval (including NER, QA,Text Mining, Document Classification, Text Categorisation)

- 50 or more submissions are:
    - Applications involving LRs and Evaluation (including applications in specific domains)
    - Less-Resourced/Endangered Languages
    - Dialogue, Conversational Systems, Chatbots, Human-Robot Interaction
    - Speech Resources and Processing (including Phonetic Databases, Phonology, Prosody)
    - Statistical Methods and Machine Learning for Language Technologies (including Language Models)
    - Multilinguality and Machine Translation (including Speech-to-Speech translation)
    - Evaluation and Validation Methodologies
    - Opinion Mining, Sentiment Analysis, Emotion Recognition/Generation
    - Multimodality and Cross-modality (including Sign Languages, Vision and other modalities) and Multimedia
    - Social Media Processing

## less attention (some are first time here)

- Natural Language Generation (including Summarization)
- Lexicons (also WordNet, FrameNet, Multimodal and Sign Language lexicons, etc.)
- Semantics (including Distributional Semantics, Word Sense Disambiguation, Coreference, etc.)
- Language Resources and Evaluation for Psycho-linguistics, Cognitive Linguistics and Linguistic Theories
- Digital Humanities and Cultural Heritage
- Discourse and Pragmatics
- Parsing, Tagging, Grammar, Syntax, Morphology
- Language Resource Infrastructures, Standards for LRs, Metadata, Policy issues, Ethics, Legal Issues
- Reproduction of Research Results in Science and Technology of Language

| | | La Major | Mucem | Notre-Dame | Petit Mucem | Estaque | Grand Large | Joliette | Lacydon | Saint-Jean | Salle 36 | Salle 50 | Salle 50 bis | Salle 92 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| June 20 | am | EURALI<br>Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia | OSACT<br>The 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Quran Q&A and Fine-Grained Hate Speech Detection | *Uniform Meaning Representation, a Cross-lingual Annotation Framework for Document-level Semantics*<br>Outline<br>Slides | *Lexical Semantic Change: Models, Data and Evaluation*<br>Outline<br>**Slides**<br>• Introduction<br>• Data<br>• Evaluation<br>• Models | ISA-18<br>Eighteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation | ParlaCLARIN III | CLTW<br>Celtic Language Technology Workshop | CMLC-10<br>Challenges in the management of large corpora 10 | *Building Reliable Datasets for Aggressive and Hateful Language Identification: Theory, Taxonomies and Approaches*<br>Outline | *Meta-Evaluation of Translation Evaluation Methods: a systematic up-to-date overview*<br>Webpage | *FAIRterm 4 ALL: Design and Implementation of FAIR (Findable, Accessible, Interoperable, Reusable) Terminological databases*<br>Outline | P-VLAM: People in Vision, Language And the Mind | 4th GLOBALEX Workshop @ LREC 2022 – Linked Lexicography |
| | pm | Perspectivist Approaches to NLP | | *Information extraction from social media: A hands-on tutorial on tasks, data, and open source tools*<br>Outline | WILDRE-6<br>6th Workshop on Indian Language Data Resource and Evaluation | | | Towards Digital Language Equality | *Cognitive and Linguistic BERTology: An Idiomatic Multiword Expression Perspective*<br>Outline | *Semantic Alignments across Languages*<br>Outline<br>Slides | Processing Language Variation: Digital Armenian | *Using LDC's Recording and Transcription Software*<br>Outline | Term21<br>Terminology in the 21st century: many faces, many places | |
| June 24 | am | Dataset Creation for Lower-Resourced Languages | LEGAL2022<br>Legal and Ethical Issues in Human Language Technologies | LDL-20228th Workshop on Linked Data in Linguistics : Revisiting a Decade of Linguistic Linked Open Data | ResT-UP 2<br>Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis | FNP2022<br>The 4th Financial Narrative Processing Workshop | PoliticalNLP<br>Workshop on Natural Language Processing for Political sciences | READI<br>Tools and Resources for REAding Difficulties | LAW XVI<br>The 16th Linguistic Annotation Workshop | SmiLa<br>Smiling and Laughter across contexts and the life-span | | | | |
| | pm | 7th Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives | | | | | SIGUL 2022 | | | The second workshop on Sentiment Analysis and Linguistic Linked Data (SALLD-2) @ LREC 2022 | | | | |
| June 25 | am | 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources | LT4HALA<br>Second Workshop on Language Technologies for Historical and Ancient Languages | 15th Workshop on Building and Using Comparable Corpora (with Shared Task on Multilingual Terminology Extraction from Comparable Corpora) | | RaPID-4: Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/ developmental impairments | | MWE 2022<br>18th Workshop on Multiword Expressions | LATERAISS<br>Language Technology and Resources for a Fair, Inclusive, and Safe Society | The 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results | | | | |
| | pm | | | | 1st Computing Social Responsibility Workshop-NLP Approaches to Corporate Social Responsibilities (CSR-NLP I) 2022 | | | | | Games and NLP | | | | |


*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3337–3345
Marseille, 20-25 June 2022


# Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection

**Ranka Stanković*, Cvetana Krstev†, Branislava Šandrih Todorović†, Duško Vitas‡, Mihailo Škorić*, Milica Ikonić Nešić†**
*University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka, mihailo.skoric}@rgf.bg.ac.rs
†University of Belgrade, Faculty of Philology, Serbia
cvetana@matf.bg.ac.rs, {branislava.sandrih, milica.ikonic.nesic}@fil.bg.ac.rs
‡University of Belgrade, Faculty of Mathematics, Serbia
vitas@matf.bg.ac.rs

**Abstract**

In this paper we present the Serbian part of the ELTeC multilingual corpus of novels written in the time period 1840–1920. The corpus is being built in order to test various distant reading methods and tools with the aim of re-thinking the European literary history. We present the various steps that led to the production of the Serbian sub-collection: the novel selection and retrieval, text preparation, structural annotation, POS-tagging, lemmatization and named entity recognition. The Serbian sub-collection was published on different platforms in order to make it freely available to various users. Several use examples show that this sub-collection is useful for both close and distant reading approaches.

**Keywords:** Corpus, Distant Reading, Digital Humanities, Linked Data, Named Entity Recognition, Text Analytics

## 1. Introduction

The term "distant reading" was first mentioned in (Moretti, 2000) for the use of quantitative text analysis methods in literary studies for the exploration of big text collections "at a distance". Observing particular features within the texts should help in discovery of new information and patterns in these collections "more objectively". The hypothesis of the current distant reading research is that useful (even if imperfect) formal and quantifiable textual features can be used as indicators or proxies for relevant literary phenomena (Schöch et al., 2020, p.1). Today, more than twenty years after its emergence, literary scholars are still discussing whether "distant reading" stands in opposition to the "close reading" ("patient, slow, word-by-word reading of literary texts that clarified the nuances and ambiguities of meaning...") and renders it obsolete (Glaubitz, 2018). For Underwood (2019) distant reading is simply a new scale of description that does not displace previous scales of literary description, but has the potential to expand the discipline, while Ciotti (2021) claims that the computational literary and cultural studies must find proper theoretical frameworks to take full advantage of the most advanced methods and analytical techniques, like text mining and machine learning.

In this paper Distant Reading (DR) refers to the currently ongoing COST action *Distant Reading for European Literary History* (CA16204) (2017–2022),[1] aiming at creating a network of scholars of different background that would produce resources and tools that can help in writing the European literary history from the new standpoint. Its main objective is the production of an unified, uniform, multilingual, digital novel collection dubbed ELTeC[2] (Odebrecht et al., 2021).

Our focus of this paper is the Serbian part of ELTeC, SrpELTeC sub-collection, and challenges that we had to overcome in order to produce it. This paper is organized as follows: Section 2 brings an overview of the ELTeC text collection and the production of its Serbian sub-collection, including the novel selection, text preparation, structural annotation, POS-tagging, lemmatization and named entity recognition (NER); Section 3 presents examples of the publication of SrpELTeC in digital libraries, corpus management systems and in Wikidata; some research tasks in digital humanities involving SrpELTeC in their solution are presented in Section 4; Section 5 presents some ongoing and future activities.

## 2. ELTeC and its Serbian Sub-Collection

In order to make ELTeC a solid basis for the implementation of distant reading methods it had to be meticulously prepared. First of all the eligibility criteria were defined that state that each language sub-collection should contain novels originally written in that language and first published, preferably as a book, in the period 1840–1920. For this purpose, a "novel" is defined as a fictional narrative text at least 10,000 words long. The choice of novels cannot be random, since some balancing criteria have also to be met:

- A sub-collection should contain 100 works that qualify as 'novels';

[1] Distant Reading COST action

[2] ELTeC: European Literary Text Collection

| | | | |
|---|---|---|---|
| pdf | bib | *Using Wiktionary to Create Specialized Lexical Resources and Datasets*<br>Lenka Bajčetić and Thierry Declerck | pp. 3457-3460 |

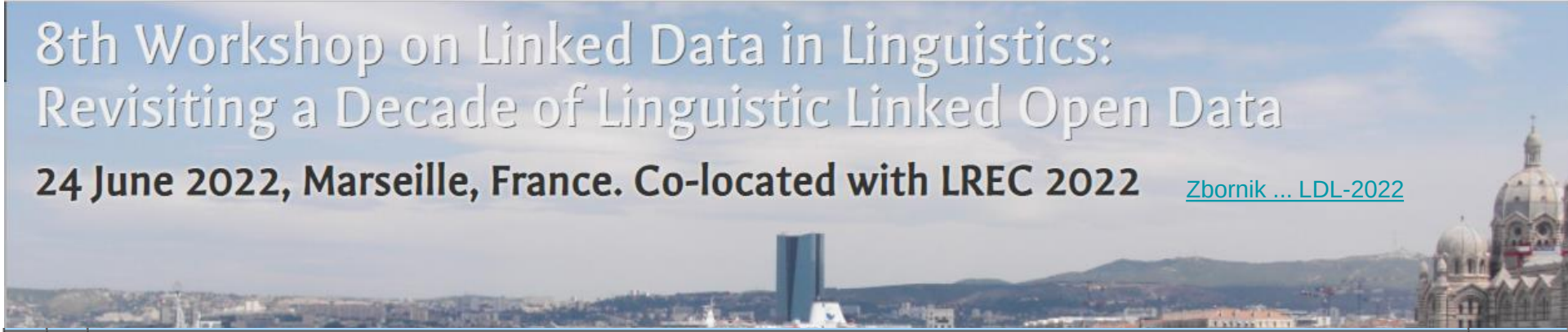
This paper describes an approach aiming at utilizing Wiktionary data for creating specialized lexical datasets which can be used for enriching other lexical (semantic) resources or for generating datasets that can be used for evaluating or improving NLP tasks, like Word Sense Disambiguation, Word-in-Context challenges, or Sense Linking across lexicons and dictionaries. We have focused on Wiktionary data about pronunciation information in English, and grammatical number and grammatical gender in German.

| | | | |
|---|---|---|---|
| pdf | bib | *GRhOOT: Ontology of Rhetorical Figures in German*<br>Ramona Kühn, Jelena Mitrović and Michael Granitzer | pp. 4001-4010 |

GRhOOT, the German RhetOrical OnTology, is a domain ontology of 110 rhetorical figures in the German language. The overall goal of building an ontology of rhetorical figures in German is not only the formal representation of different rhetorical figures, but also allowing for their easier detection, thus improving sentiment analysis, argument mining, detection of hate speech and fake news, machine translation, and many other tasks in which recognition of non-literal language plays an important role. The challenge of building such ontologies lies in classifying the figures and assigning adequate characteristics to group them, while considering their distinctive features. The ontology of rhetorical figures in the Serbian language was used as a basis for our work. Besides transferring and extending the concepts of the Serbian ontology, we ensured completeness and consistency by using description logic and SPARQL queries. Furthermore, we show a decision tree to identify figures and suggest a usage scenario on how the ontology can be utilized to collect and annotate data.

| | | | |
|---|---|---|---|
| pdf | bib | *Cross-Level Semantic Similarity for Serbian Newswire Texts*<br>Vuk Batanović and Maja Miličević Petrović | pp. 1691-1699 |

| | | | |
|---|---|---|---|
| pdf | bib | *DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations*<br>Ekaterina Lapshinova-Koltunski, Maja Popović and Maarit Koponen | pp. 1751-1760 |




Zbornik ... LDL-2022

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 7–16
Marseille, 24 June 2022


# From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)

**Milica Ikonić Nešić*, Ranka Stanković†, Christof Schöch‡, Mihailo Škorić†**
*University of Belgrade, Faculty of Philology, Serbia
milica.ikonic.nesic@fil.bg.ac.rs, †University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka.stankovic, mihailo.skoric}@rgf.bg.ac.rs
‡University of Trier, Germany; schoech@uni-trier.de

**Abstract**

In this paper we present the wikification of the ELTeC (European Literary Text Collection), developed within the COST Action "Distant Reading for European Literary History" (CA16204). ELTeC is a multilingual corpus of novels written in the time period 1840—1920, built to apply distant reading methods and tools to explore the European literary history. We present the pipeline that led to the production of the linked dataset, the novels' metadata retrieval and named entity recognition, transformation, mapping and Wikidata population, followed by named entity linking and export to NIF (NLP Interchange Format). The speeding up of the process of data preparation and import to Wikidata is presented on the use case of seven sub-collections of ELTeC (English, Portuguese, French, Slovenian, German, Hungarian and Serbian). Our goal was to automate the process of preparing and importing information, so OpenRefine and QuickStatements were chosen as the best options. The paper also includes examples of SPARQL queries for retrieval of authors, novel titles, publication places and other metadata with different visualisation options as well as statistical overviews.

**Keywords:** Wikidata, linked data, SPARQL, distant reading, literary corpus, named entity linking, ELTeC

## 1. Introduction

The COST Action "Distant Reading for European Literary History"[1] ran from 2017 to 2022 and aimed to use computational methods for the analysis of large collections of literary texts. The main goal of this networking project was to compile and analyse a multilingual open-source collection of novels, named European Literary Text Collection (ELTeC). ELTeC contains corpora of 100 novels per language written between 1840 and 1920 that are encoded in XML, are linguistically annotated and contain detailed metadata (Schöch et al., 2021).

The term *distant reading* (Moretti, 2000) describes an alternative or a complement to *close reading*: Instead of detailed, qualitative interpretations of selected literary texts, the idea is to analyse large collections of literary texts using quantitative methods of text analysis and machine learning. Formal and quantifiable textual features are used as indicators for relevant literary phenomena, with their patterns of occurrence then being related to categories such as authors, genres, or literary periods (Schöch et al., 2020).

This paper presents an approach for publishing the metadata and named entities (NE) from the sub-collections of ELTeC as linked open data. More precisely, the paper presents results for 700 novels from the first seven languages (English, Portuguese, French, Slovenian, German, Hungarian and Serbian) that are morpho–syntactically tagged (Stanković et al., 2022b) and partially annotated with named entities (Stanković et al., 2019; Frontini et al., 2020), as well as the case study on Named Entity Linking (NEL) for the Serbian ELTeC sub-collection.

Linked open data for literary texts is slowly gaining traction, as evidenced by resources such as BookSampo (Mäkelä et al., 2013) or projects like POSTDATA (Bermúdez-Sabel et al., 2021) and Mining and Modeling Text (Schöch et al., 2022). The motivation for the presented activity was to increase the visibility of the ELTeC collection, to connect it to open knowledge bases, as well as to allow searching and analyzing texts using linked open data. The incentive for the presented activity was the successful initial implementation for Serbian (Ikonić Nešić et al., 2021) that was further applied to other six languages with support of the sub-collection coordinators.

We use the term *wikification* not only for entity linking with Wikidata as the target Knowledge base, but also for creating and populating Wikidata items related to novels which will be further used for entity linking.

The crucial point for automation of wikification was the synergy of the powerful open source tools OpenRefine (Huynh, 2012) and QuickStatements (Manske, 2019). This enabled 700 novels from the core collections and 20 from extended sub-collections of ELTeC to be described in Wikidata, including associated items for their first editions, print editions, digital editions and the ELTeC (electronic) editions. This resulted in approximately 20,900 automatically added statements. To the best of our knowledge, this work is the first example of data about literary corpora for seven languages being automatically imported into Wikidata using different open source tools.

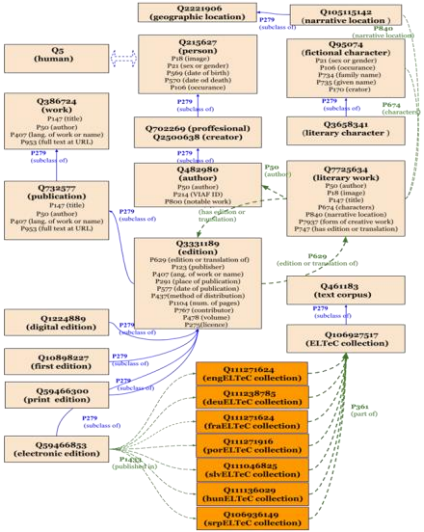
Section 2 is dedicated to the ELTeC: in Subsection 2.1 an overview of the text collection is given, in Subsec-

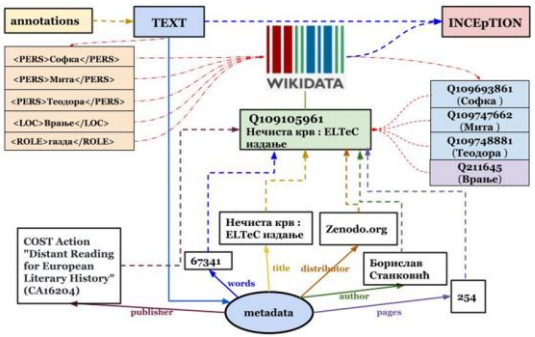
[1]Distant Reading for European Literary History (CA16204), https://www.distant-reading.net.

**The main objective** : wikification of old literary texts in ELTeC collection for seven languages (eng, deu, fra, hun, por, slo, srp)

## Main topics:

- Distant reading COST action
- ELTeC
  - Text Collection in XML/TEI
  - Linked Data model
  - Automation of Wikidata Population
- NER 4 WikiELTeC
  - Literary Characters and Narrative Locations in Novels
  - From NE Extraction to Wikidata across Inception to NIF
- The Overview of ELTeC@Wikidata by SPARQL Queries




Wikidata:WikiProject ELTeC — Wikidata

Nexus Linguarum

Workshops on Sentiment Analysis & Linguistic Linked Data

SALLD-2

Home → SALLD-2

24 June 2022, Marseille, France, Co-located with LREC 2022

# Sentiment Analysis of Sentences from Serbian ELTeC corpus

**Ranka Stanković**, Miloš Košprdić, Milica Ikonić Nešić, Tijana Radović
University of Belgrade, Serbia

Proceedings of The 18th Workshop on Multiword Expressions @LREC2022 (lrec-conf.org)

| **Papers** [pdf] [bib] | pages |
|---|---|
| *Figurative Language in Noun Compound Models across Target Properties, Domains and Time*<br>Sabine Schulte im Walde<br>[pdf] [bib] | pp. 1-1 |
| *Multiword Expressions and the Low-Resource Scenario from the Perspective of a Local Oral Culture*<br>Steven Bird<br>[pdf] [bib] | pp. 2-2 |
| *A General Framework for Detecting Metaphorical Collocations*<br>Marija Brkić Bakarić, Lucia Načinović Prskalo and Maja Popović<br>[pdf] [bib] | pp. 3-8 |
| *Improving Grammatical Error Correction for Multiword Expressions*<br>Shiva Taslimipoor, Christopher Bryant and Zheng Yuan<br>[pdf] [bib] | pp. 9-15 |
| *An Analysis of Attention in German Verbal Idiom Disambiguation*<br>Rafael Ehren, Laura Kallmeyer and Timm Lichte<br>[pdf] [bib] [optional] [supplementary] | pp. 16-25 |
| *Support Verb Constructions across the Ocean Sea*<br>Jorge Baptista, Nuno Mamede and Sónia Reis<br>[pdf] [bib] | pp. 26-36 |
| *A Matrix-Based Heuristic Algorithm for Extracting Multiword Expressions from a Corpus*<br>Orhan Bilgin<br>[pdf] [bib] [optional] [supplementary] | pp. 37-48 |
| *Multi-word Lexical Units Recognition in WordNet*<br>Marek Maziarz, Ewa Rudnicka and Łukasz Grabowski<br>[pdf] [bib] | pp. 49-54 |
| *Automatic Detection of Difficulty of French Medical Sequences in Context*<br>Anaïs KOPTIENT and Natalia Grabar<br>[pdf] [bib] | pp. 55-66 |
| *Annotating "Particles" in Multiword Expressions in te reo Māori for a Part-of-Speech Tagger*<br>Aoife Finn, Suzanne Duncan, Peter-Lucas Jones, Gianna Leoni and Keoni Mahelona<br>[pdf] [bib] | pp. 67-74 |
| *Metaphor Detection for Low Resource Languages: From Zero-Shot to Few-Shot Learning in Middle High German*<br>Felix Schneider, Sven Sickert, Phillip Brandes, Sophie Marshall and Joachim Denzler<br>[pdf] [bib] | pp. 75-80 |
| *Automatic Bilingual Phrase Dictionary Construction from GIZA++ Output*<br>Albina Khusainova, Vitaly Romanov and Adil Khan<br>[pdf] [bib] | pp. 81-88 |
| *A BERT's Eye View: Identification of Irish Multiword Expressions Using Pre-trained Language Models*<br>Abigail Walsh, Teresa Lynn and Jennifer Foster<br>[pdf] [bib] | pp. 89-99 |
| *Enhancing the PARSEME Turkish Corpus of Verbal Multiword Expressions*<br>Yagmur Ozturk, Najet Hadj Mohamed, Adam Lion-Bouton and Agata Savary<br>[pdf] [bib] | pp. 100-104 |
| *Sample Efficient Approaches for Idiomaticity Detection*<br>Dylan Phelps, Xuan-Rui Fan, Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton and Aline Villavicencio<br>[pdf] [bib] | pp. 105-111 |
| *mwetoolkit-lib: Adaptation of the mwetoolkit as a Python Library and an Application to MWE-based Document Clustering*<br>Fernando Zagatti, Paulo Augusto de Lima Medeiros, Esther da Cunha Soares, Lucas Nildaimon dos Santos Silva, Carlos Ramisch and Livy Real<br>[pdf] [bib] | pp. 112-117 |
| *Handling Idioms in Symbolic Multilingual Natural Language Generation*<br>Michaelle Dubé and François Lareau<br>[pdf] [bib] | pp. 118-126 |

# Za dalje čitanje …

http://www.lrec-conf.org/proceedings/lrec2022/index.html

http://www.lrec-conf.org/proceedings/lrec2022/workshops/index.html