

Encoding, Linking, Retrieving: A Methodological Framework for Knowledge-Enriched Interview Corpora^{*}

Ranka Stanković¹, Tamara Vučenović², Milica Ikonić Nešić³, and Biljana Rujević¹

¹ University of Belgrade, Faculty of Mining and Geology,
Djusina 7, Belgrade, Serbia

ranka.stankovic@rgf.bg.ac.rs (corresponding author), biljana.rujevic@rgf.bg.ac.rs

² University Metropolitan, Faculty of Management
Tadeuša Košćuška 63, Belgrade, Serbia
tamara.vucenovic@metropolitan.ac.rs

³ University of Belgrade, Faculty of Philology
Studentski Trg 2, Belgrade, Serbia
milica.ikonik.nesic@fil.bg.ac.rs

Sažetak This paper presents an AI-driven pipeline for transforming the interviews from “Digitalne Ikone 20+” book from unstructured transcripts into a structured, semantically enriched, and queryable knowledge resource. The raw text was first converted into XML-TEI format, with explicit structural markup of interview boundaries, speaker turns, paragraphs, temporal metadata, and topics. This encoding established logical segmentation and enabled targeted queries, such as retrieving content by speaker or thematic segment. An NLP and textometric analysis was conducted using the TXM tool and JeRTeh resources, followed by automatic Named Entity Recognition (NER) using models from the TESLA project. Key entity types were identified and embedded into the TEI structure. In the subsequent Named Entity Linking (NEL) stage, entities were disambiguated and connected to Wikidata identifiers, enriching the corpus with external knowledge graph references. Missing entities were added to Wikidata, contributing new structured knowledge. The resulting resource allows researchers, students, and the public to explore cultural heritage interviews through intelligent querying, automated dataset generation, and knowledge graph integration. The pipeline offers a replicable methodology for converting oral archives into AI-accessible knowledge bases.

Keywords: Corpus, Named Entity Recognition, Named Entity Linking, Information Extraction, Textometry, Knowledge Graphs.

1. Introduction

This paper builds upon our preliminary findings first introduced in the AI Conference 2025 abstract, where the initial implementation of the “Digitalne Ikone” AI pipeline

^{*} This paper is an extended version of an abstract “From Unstructured Interviews to Queryable Knowledge: An AI Pipeline for the “Digitalne Ikone” Corpus using TEI, NER, NEL, and RAG”, published in Artificial Intelligence Conference Book of Abstracts. <https://www.mi.sanu.ac.rs/>

showed promising results, and now presents a fully extended study with deeper analysis, methodological enhancements, and a broader evaluation [32].

Large archives of interviews and broadcast transcripts remain significantly underutilized in Natural Language Processing (NLP), primarily due to their inherently unstructured nature, diverse formats, and the high cost of performing reliable semantic queries over them. The challenge is even more pronounced in low- and less-resourced languages, where reusable processing pipelines and datasets that enable grounded question answering (QA) are still scarce or entirely absent.

In this work, our motivation is twofold. First, we aim to transform the long-standing Serbian interview series *Digitalne Ikone 20+* (Digital Icons 20+) [38] into a structured, semantically enriched resource that facilitates reproducible research, exploratory analysis, and evidence-based question answering. Second, we seek to establish a standards-compliant methodological framework, based on widely accepted practices such as XML-TEI encoding (Text Encoding Initiative), lexical and semantic annotation, text analysis, knowledge management and information retrieval, that can be adapted and adopted by other oral-history archives, cultural heritage collections, or media corpora across different languages. By doing so, we aim not only to improve accessibility and interoperability, but also to highlight the broader potential of AI methodologies for unlocking knowledge embedded in complex narrative datasets.

The radio broadcast *Digital Icons*, curated by Dr. Tamara Vučenović, has been continuously aired on *Radio Belgrade 2* since 2002, with more than 1,000 episodes to date. As an educational program devoted to digital technologies and their societal impact, it holds a unique position within the Serbian media landscape, not only among radio programs but also across television formats. Its long-term significance stems from both its thematic coherence and its informative public role: the program invites domain experts to discuss topics spanning computer science, the evolution of the information society, digital inclusion and the protection of vulnerable groups, ethical issues of social media, and the benefits and risks associated with technological development.

The diversity of guests has been a defining feature of the broadcast. Participants include senior officials and policy advisers, researchers and academic representatives, professionals from domestic and international IT organizations, civil society groups, media outlets, and specialized associations. Moreover, the program has hosted numerous international interlocutors, contributing to the global exchange of ideas about digital culture and technology.

Importantly, the book *Digitalne Ikone 20+* (Digital Icons 20+), derived from the transcripts of selected broadcast interviews, provided the foundation for building the *DigIko* [32], as first version, and *DigIko2 – Digital Icons 20+ Interview Corpus* used in this research. It now represents a valuable bridge between journalistic content, public discourse, and computational linguistics, enabling further exploration of societal narratives through AI-based methods.

The integration of TEI with NER (Named Entity Recognition) and NEL (Named Entity Linking) is motivated by the need to provide layered semantic enrichment that serves as a reliable foundation for Retrieval-Augmented Generation (RAG). By grounding AI-generated answers in verifiable textual evidence, RAG allows retrieved passages, rather than free-form generative output, to guide the reasoning process and mitigate hallucinations.

ons. In this setting, NEL based on Wikidata⁴ identifiers (QIDs) [36] plays a crucial role, as it enables precise multilingual entity resolution and control during both retrieval and prompting stages. Each entity mention is not only classified but also linked to an external canonical reference, contributing to interoperability and long-term reusability.

Furthermore, the consistent use of TEI IDs and Wikidata QIDs attached to each extracted claim ensures transparency of provenance, supports reproducibility and peer-review practices, and facilitates integration with existing or emerging knowledge graphs. This standards-driven combination of TEI, NER, and NEL therefore establishes a structured backbone for explainable AI workflows, allowing interview-based corpora to be queried, validated, and semantically explored using modern NLP (Natural Language Processing) techniques.

The contributions of this paper can be summarized as follows. First, we present a standards-driven AI pipeline that integrates multiple preprocessing tasks, ranging from structural encoding to semantic annotation, thereby transforming raw interview transcripts into a semantically linked and structurally queryable resource. This approach ensures interoperability, transparency, and reproducibility across all stages of processing. Second, we demonstrate a grounded question–answering (QA) generation over the *Digiko2* interview corpus, combining evidence retrieval with LLM-based generation to support scholarly inquiry and exploratory analysis.

The workflow incorporates TEI structural annotation, TXM/Jerteh preprocessing, and NER/NEL models from the TESLA⁵ project, resulting in TEI-embedded entity tags resolvable to Wikidata identifiers. We argue that such integration substantially improves accessibility, enables semantic search and automated dataset generation, and provides a replicable pathway for constructing knowledge-centric interview corpora for Serbian and other less-resourced languages.

In addition, the repository with code and sample data will be made openly available, while the complete corpus will be searchable through CQL (Corpus Query Language) and SPARQL endpoints for the linked-data version. Together, these components establish a foundation for transparent peer review, empirical validation, and future extensions toward knowledge graph construction and more advanced RAG architectures.

Beyond conventional transformation, the proposed approach enables entity- and speaker-faceted retrieval, which opens possibilities for opinion tracking, stance profiling, and temporal comparison, analytical dimensions that are typically unavailable in plain text processing. By implementing a sequential transformation from raw transcripts to a semantically enriched and queryable corpus, we provide a foundation for evidence-based analysis that goes beyond narrative abstraction.

The discussion of prior studies and relevant methodological frameworks is provided in Section 2, which outlines existing approaches to TEI-based corpus construction, NER/NEL integration, textometry and retrieval-augmented methods in cultural and interview-based datasets. The complete methodology for corpus preparation and exploitation is detailed in Section 3, followed by the presentation of the corpus editions and data formats in Section 4, along with the analytical procedures, practical applications. The evaluation of results and the discussion of current limitations with directions for future development is provided in Section 5. Together, these sections outline a replicable

⁴ <https://www.wikidata.org/>

⁵ TESLA – Text Embeddings – Serbian Language Applications, <https://tesla.rgf.bg.ac.rs/>

workflow for building and systematically interrogating interview-based corpora through semantically grounded AI methods.

2. Related Work

In this section, we provide an overview of related research across several interconnected areas, all relevant to the construction and exploitation of the *DigIko2* corpus. First, we discuss prior studies that model interview transcripts and specifically those that use TEI as a structural framework for capturing interview transcripts and conversational data. Next, we review approaches to semantic enrichment through NER with particular emphasis on linking to external knowledge bases such as Wikidata. We conclude this section with work related to textometry, as a computational method for analyzing textual corpora that combines statistical modeling with linguistic structure to uncover patterns, distributions, and relationships within texts that are not easily visible through manual reading.

2.1. Modeling of Interview Corpora

The work presented in [26] introduces *InTEIviews*, an XML–TEI–conformant ODD⁶ specifically tailored for qualitative interviews, with the aim of promoting openness, reusability, and standardized annotation practices. An XML–TEI–conformant ODD enables the selection of relevant TEI P5 modules and the declaration of project-specific constraints while remaining compatible with the TEI element and class system. From a single ODD specification, it is possible to generate both validation schemas and human-readable documentation, ensuring both methodological transparency and reproducibility.

The ODD typically includes a `schemaSpec` (for module selection), `elementSpec` and `classSpec` (for content models and attributes), as well as prose guidance on encoding decisions and metadata requirements. The model builds upon existing TEI modules, particularly *Transcription of Speech* and *Language Corpora*, and draws a clear distinction between macrostructure (e.g., headered metadata on participants and recording conditions, turns encoded via `u who=`, and researcher notes within `standOff`) and microstructure (e.g., fine-grained linguistic and discourse features, including named entities). Both structures and the following TEI elements are presented in Sample 1.1.

Ethical and privacy considerations are also addressed by recommending the use of `gap` (in combination with `@resp`, `@unit`, and `@quantity`) for redacting sensitive content. Controlled vocabularies, such as annotating speakers with `@ana`⁷ on `u`⁸ are encouraged to support cross-interview comparability. Overall, the *InTEIviews* model demonstrates how interview transcripts can be standardized through TEI while still preserving project-specific variation, enabling semantic analysis, corpus sharing, and FAIR-aligned reuse.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt> <title>TEI interview</title> </titleStmt>
```

⁶ ODD (“One Document Does-it-all”) is the TEI XML-based specification language for defining, constraining, and documenting TEI customizations.

⁷ analysis refers to elements that contain interpretations of the annotated element.

⁸ utterance: contains a stretch of speech usually preceded and followed by silence or by a change of speaker.


```

    </fileDesc>
  </teiHeader>
  <!-- MACROSTRUCTURE -->
  <standOff>
    <listPerson>
      <person xml:id="SP1" gender="F" age="35" .../>
      <person xml:id="INT" role="interviewer" .../>
    </listPerson>
  </standOff>
  <text>
    <body>
      <!-- MICROSTRUCTURE -->
      <div type="interview" ...>
        <u who="#INT" xml:id="u1"> Can you describe life during that period? </u>
        <u who="#SP1" xml:id="u2" ...>It was very nice in <name type="LOC" ref="https://www.wikidata.org/wiki/Q55630">Novi Sad</name> for three months.</u>
        <u who="#SP1" xml:id="u3">After that... it is hard to explain.</u>
      </div>
    </body>
  </text>
</TEI>

```

Sample 1.1. XML schema *InTEIreviews*

The DDHI Encoding Guidelines [3] developed for the Dartmouth Digital History Initiative (DDHI) provide practical guidance for TEI-based oral history encoding, with a strong emphasis on interview metadata and TEI headers. These guidelines focus on capturing descriptive information about interviewees, interviewers, recording circumstances, equipment, rights, and ethical considerations, offering clear examples and recommended practices for structuring metadata within `teiHeader`. In addition, they demonstrate techniques for annotating speaker turns, editorial notes, and research-related contextual information, making them particularly valuable for projects involving oral narratives and historical accounts. What distinguishes the DDHI approach is its balance between TEI compliance and practical usability: the guidelines include actionable examples, recommended element usage, and documentation that supports reproducibility and interoperability across oral history projects. This makes DDHI a relevant reference point for corpus preparation methodologies such as *DigIko2*, where interview metadata, participant roles, and recording context are essential for downstream tasks such as entity linking, semantic retrieval, and RAG-based question answering.

A valuable overview of transcript–media alignment methods is provided in [4], which analyzes three practical approaches for synchronizing interview audio/video with their corresponding transcripts: the OHMS (Oral History Metadata Synchronizer) system, TEI-based encoding workflows, and YouTube-based captioning practices. These methods are examined in terms of their ability to provide searchable, time-coded access to oral-history materials, enabling users to navigate interviews based on speaker turns, topics, or temporal segments. Of particular relevance to TEI-based corpus construction is the demonstration of how structured markup, especially when combined with timestamp anchoring that supports interoperability and facilitates links between transcript segments and media files. This approach not only enhances accessibility and user engagement but also enables multimodal analysis, making oral-history collections more suitable for scholarly interrogation, pedagogical use, and AI-based retrieval.

2.2. Text Encoding Initiative for Spoken Corpora

CLARIN-compliant web services provide an end-to-end processing workflow for *TEI/ISO 24624* transcripts of spoken language, with particular emphasis on legacy interview corpora [7]. Their framework demonstrates how linguistic tools and processing components can operate directly on *ISO/TEI* formats, thus ensuring semantic interoperability with established spoken-language tools while also granting access to a broader range of written-language NLP functionalities. A key application is showcased through the IDS Archive for Spoken German, where interview materials are curated and processed according to FAIR principles. The workflow involves complete digitization of archival recordings, conversion into structured and standardized formats (including *ISO/TEI* for transcripts and CMDI for metadata), precise alignment between audio and transcript segments, and successive layers of linguistic enrichment. Furthermore, the resulting corpora are integrated into dissemination platforms that support search, visualization, and processing services, thereby enabling sustainable reuse and scholarly exploration of oral-history datasets. This approach offers a compelling model for similar initiatives, particularly for interview-based collections that seek to combine long-term preservation with modern AI-based processing methods.

The service pipeline consists of several modular components designed to convert raw interview transcripts into linguistically enriched *ISO/TEI* representations. The workflow begins with *text2iso*, which parses plain-text transcripts and generates *ISO/TEI*-compliant structures including *u* for utterances, *timeline* for time anchoring, participant metadata in `<teiHeader>`, and mechanisms for overlap resolution. Subsequently, tokenization transforms the transcript into *w* (word) and *pc* (punctuation) elements, while preserving original anchors and lifting pauses into dedicated *pause* tags. Per-utterance language identification is performed using OpenNLP, with the detected code assigned to `@xml:lang` attributes. Language-specific normalization is handled through OrthoNormal, which writes normalized values to `@norm` within the *w* elements, allowing both orthographic variants and surface forms to be retained. Further enrichment involves POS-tagging and lemmatization via TreeTagger/TT4J with language-appropriate models, followed by pseudo-alignment between transcript segments and the corresponding audio stream. Finally, consistent `@xml:id` management ensures traceability across elements and supports downstream referencing, validation, and corpus querying. Overall, the pipeline illustrates a structured and extensible approach to preparing spoken-language corpora for both linguistic research and computational processing.

The Slovene reference spoken corpus *GOS 2* represents a substantial expansion of the original *GOS* corpus from 2011, integrating two major additional sources: *GOS Video-Lectures* (public academic speech) and the *Artur ASR* database (Automatic Speech Recognition). The corpus includes authentic spoken data that have been manually verified, and heterogeneous designs from the original datasets were reconciled through a unified metadata.

A revised *TEI XML* schema-based on the *CLARIN.SI* parameterization was implemented to ensure formal alignment and long-term interoperability. Transcription conventions were harmonized across datasets; however, certain segmentation and punctuation variations were deliberately preserved in order to stimulate community feedback and support alternative annotation practices. Linguistic enrichment was conducted using CLAS-SLA–Stanza, enhancing the corpus with token-level annotation and morphosyntactic in-

formation. The resulting resource is openly released in *TEI* format, with transcriptions publicly available and audio data accessible for the newer subsets. A dedicated web concordancer enables simple and advanced search capabilities, metadata-based filtering, and integration with other Slovene linguistic resources [35].

The *Parla-CLARIN* initiative provides a specialized *TEI* customization for modeling and exchanging parliamentary debate corpora, with a strong emphasis on interoperability and usability across multilingual research contexts. Its guidelines propose a consistent document architecture that includes hierarchical structures for sittings, debates, and individual speeches or utterances, each annotated with `@who` attributes to identify speakers unambiguously. The framework further encourages rich `teiHeader` metadata, covering information about participants, political affiliations, institutional roles, temporal information, and geographic context. In addition to textual representation, *Parla-CLARIN* supports links to audio/video recordings and references to legislative acts. The package provides *TEI* schemas, validation instructions, and illustrative examples, making it suitable both for corpus construction and integration with downstream NLP workflows.⁹

The British National Corpus (BNC) XML Edition represents one of the most influential large-scale implementations of *TEI* for spoken interaction. As a well-established resource, it showcases mature structural conventions for modeling dialogue and turn-taking using `u` elements, with speaker identities assigned through `@who` attributes that reference a dedicated speaker registry. At the token level, the BNC employs consistent markup practices, including `w` for words and `pc` for punctuation, enabling corpus-wide uniformity and compatibility with a wide range of linguistic tools. A comprehensive `teiHeader` encodes rich participant and recording metadata, covering demographic information, situational context, and technical details of the audio capture. Owing to its scale, stability, and longstanding use, the Spoken BNC provides methodological precedents for utterance segmentation, speaker normalization, metadata design, and reproducibility, many of which can be effectively transferred to the development of interview-based corpora such as *DigIko2*. Its structural conventions remain a reference point for both *TEI*-based encoding practices and best practices in spoken-language corpus construction.¹⁰

2.3. Named Entities in Serbian Text

Named Entity Recognition and Linking (NER+NEL) inherently involve the task of entity disambiguation, which is particularly challenging for morphologically and less resourced languages [20,9]. Historically, early research in NEL relied on rule-based approaches and handcrafted heuristics, as seen in systems such as *SrpNER* [20] and the disambiguation algorithms of [39]. Over time, these were replaced by supervised machine learning models, including SVMs, CRFs, and neural architectures, that demonstrated substantially higher accuracy in complex linking scenarios [14,8]. The availability of structured multilingual resources such as Wikipedia, DBpedia, and especially Wikidata [2] has been crucial in providing canonical knowledge representations for entity resolution.

A major shift occurred with the emergence of transformer-based architectures, notably BERT and a suite of domain- or task-adapted models, that significantly improved NEL performance by leveraging context-aware embeddings, cross-lingual transfer, and

⁹ *Parla-CLARIN* <https://clarin-eric.github.io/parla-clarin/>

¹⁰ BNC XML Reference Guide: <http://www.natcorp.ox.ac.uk/docs/URG.xml?ID=cdfisp>.

domain-specific adaptation [40,6]. Recent studies have also highlighted limitations of LLMs in NEL tasks within specialized domains. Recent work presented in [23] emphasized the importance of geographic bias evaluation in LLM outputs, demonstrating that the diversity and specificity of geographic references vary across models and may be influenced by prompt phrasing.

For Serbian, NEL development was supported by resources produced in this research, including the multilingual silver dataset DaMuEL [21] and the `SrpCNNeL` model for named entity recognition and linking [15]. The inflectional complexity of Serbian, characterized by different forms for three gender, seven grammatical cases, numbers, requires models to handle morphological variation. Two pretrained transformer models optimized for Serbian give the best results: 1) `TeslaXLM`,¹¹ a fine-tuned variant of *FacebookAI/xlm-roberta-large* with 561M parameters, that was trained on a 20-billion-token Serbian (Latin + Cyrillic) corpus and achieves high robustness across dialectal and orthographic variation [42], and 2) `Jerteh-355`,¹² based on *RoBERTa-large* and includes 355M parameters, trained from scratch on a 4-billion-token Serbian corpus, it excels in monolingual contextual representations [41]. These models provide the linguistic robustness necessary for entity extraction and disambiguation in highly inflected Serbian interview data, forming the basis for the NER+NEL pipeline presented in this work.

2.4. Textometry

The origins of textometry can be traced to mid-20th century France, notably through the pioneering work of Pierre Guiraud [10] and Charles Muller [24], who established early methods for linguistic statistics. Subsequent developments by Jean-Paul Benzécri [1] expanded the statistical foundations and influenced later implementations of textometric analysis. The methodological basis of modern textometry also draws significantly from the work of Ludovic Lebart and André Salem [22], who formulated techniques for extracting meaningful linguistic patterns from textual data.

Building upon these foundations, textometry has introduced new statistical models for identifying specific textual phenomena such as contextual word salience, linear organization and internal structure of texts, intertextual contrast, and lexical evolution over time (e.g., scope of usage, periods of lexical change). The results of a textometric analysis typically yield visually organized outputs, including hierarchical lists, maps, clusters, and reorganized corpora, offering a synthetic yet flexible view of the underlying textual structures.

Unlike traditional summarization or keyword extraction, textometry provides a close reading of the text through numerical indicators, enabling both global corpus-level perspectives and local contextual insights. This duality highlights its hermeneutic potential and demonstrates how large-scale textual analysis can reveal latent patterns that are difficult, if not impossible, to detect manually. Interpretation, however, remains dependent on expert knowledge and the quality of the tools used, indicating that textometry is best conceived as a methodology that generates data-driven signals requiring expert validation.

Several adjacent disciplines intersect with textometry. Information Retrieval (IR) shares an interest in processing large text collections, though its primary goal is document

¹¹ <https://huggingface.co/te-sla/TeslaXLM>

¹² <https://huggingface.co/jerteh/Jerteh-355>

and information retrieval rather than internal textual analysis. Latent Semantic Analysis (LSA) similarly employs mathematical models but is generally oriented toward cognitive and semantic tasks outside the scope of corpus-based linguistic inquiry. Text mining (TM) applies quantitative techniques to extract valuable insights; however, its focus is on information extraction, whereas textometry concentrates on structural and linguistic exploration of texts. Meanwhile, NLP relies on statistics to build systems for linguistic unit recognition and tagging; indeed, corpora may serve as a foundation for both NLP development and subsequent textometric exploration. Thus, these two fields are complementary and mutually reinforcing.

A major advancement in applied textometry is the emergence of software environments that simplify corpus exploration through integrated tools and graphical workflows. Among these, the *TXM platform* [11] stands out as an open-source environment widely used in the social sciences and digital humanities (history, literature, linguistics, sociology, geography, political science). TXM combines qualitative and quantitative text analysis, built upon CQP (Corpus Query Processor) browsing and the R statistical package, thereby enabling rich corpus navigation and statistical modeling.

3. Methodology

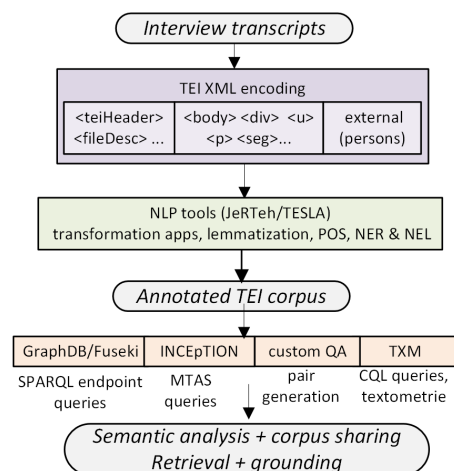
This section outlines the toolchain, starting with TEI structure encoding, followed by annotation and textometry, and finally linked-data serialization and text generation. The diagram in Figure 1 illustrates the end-to-end processing pipeline of the *DigIko2* corpus, from raw interview transcripts through TEI-compliant XML encoding to parallel NLP enrichment via TXM textometry and the JeRTeh/TESLA annotation stack (lemmatization, NER, NEL). The enriched corpus feeds three output channels: a linked-data serialization in NIF exported to an Apache Jena Fuseki SPARQL endpoint¹³ and GraphDB¹⁴, a CQL-queryable INCEpTION [18] annotation layer, and a QA-pair generation module powered by GPT-4.1/GPT-5.1 that produces a RAG-ready resource.

3.1. TEI for Spoken Interactions

Chapter 8 of the TEI P5 Guidelines, entitled “*Transcriptions of Speech*” [34], provides the core specification for representing spoken language and interviews in TEI-compliant XML. It conceptualises spoken discourse in terms of utterances (u) that are embedded within a standard TEI text hierarchy and anchored to speakers via the @who attribute, which may reference participants defined in structured registries such as listPerson. The chapter further outlines recommendations for encoding pauses, vocalisations, non-lexical phenomena, kinesic events, overlapping speech, and subdivisions of discourse. In addition, it specifies how alignment, segmentation and linking mechanisms can be employed to handle non-hierarchical annotations and time-based structures, thereby enabling interoperable and analysable transcripts suitable for digital preservation and reuse. Documentation of transcription conventions and recording conditions is expected to be maintained in the `teiHeader`, via elements such as `transcriptionDesc` and `scriptStmt`.

¹³ <http://fuseki.jerteh.rs/>

¹⁴ <http://graphdb.jerteh.rs/sparql?repositoryId=digico>



Slika 1. The components of the proposed workflow/toolchain

Complementing this, Chapter 2 of the TEI P5 Guidelines addresses the structure and function of the `teiHeader`, which provides the essential metadata framework for archiving interview material (see Figure 2). The header facilitates the systematic documentation of bibliographic and file-related information (`fileDesc`), editorial and encoding practices (`encodingDesc`), participant profiles and communication settings via `profileDesc` (e.g. `listPerson`, roles, place, time), as well as revision tracking (`revisionDesc`). In the context of interview data, these structures allow for the explicit recording of provenance, participant identities, recording circumstances, and methodological decisions—ensuring traceability, reproducibility and reliable citation. For an illustration, Sample 1.2 introduces an excerpt from *DigIko2* TEI header.

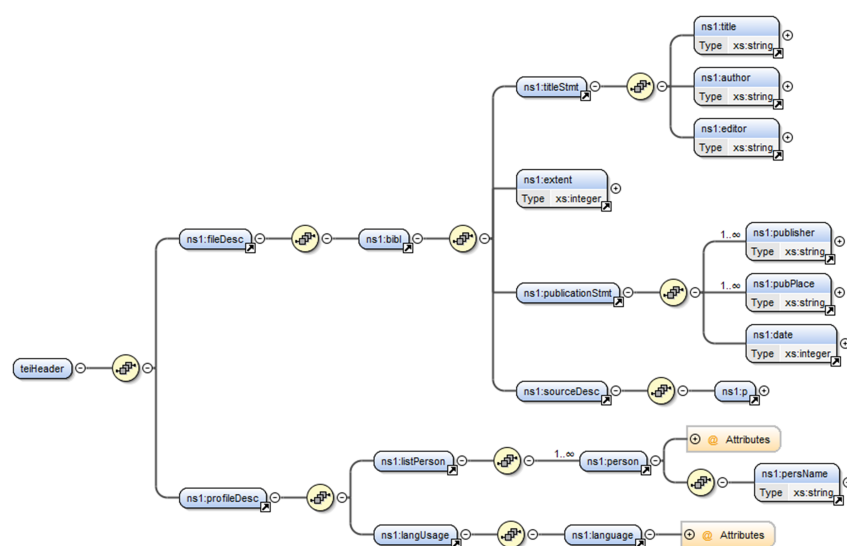
```
<teiHeader>
<fileDesc>
  <bibl>
    <titleStmt>
      <title>Digitalne ikone 20+</title>
      <author>Vučenović, Tamara</author>
      <editor>Ružić, Nevena</editor>
    </titleStmt>
    <extent>510</extent>
    <publicationStmt>
      <publisher>Akademska knjiga</publisher>
      <publisher>RTS</publisher>
      <pubPlace>Novi Sad</pubPlace>
      <pubPlace>Beograd</pubPlace>
      <date>2024</date>
    </publicationStmt>
    <sourceDesc>
      <p><seg>Born digital - Publikacija sadrži intervjue [...]</seg></p>
    </sourceDesc>
  </bibl>
</fileDesc>
<profileDesc>
  <listPerson>
    <person xml:id="TaVu" role="interviewer" qid="Q109916651">
      <persName>Tamara Vučenović</persName></person>[...]
    <person xml:id="SlMa" role="interviewee" qid="Q19559876">
      <persName>Slobodan Marković</persName></person>
    <person xml:id="VoAn" role="interviewee" qid="Q1252236">
      <persName>Voja Antonić</persName></person>[...]
```

```

<langUsage>
  <language ident="sr-Latn">Serbian (Latin script)</language>
</langUsage>
</profileDesc>
</teiHeader>

```

Sample 1.2. TEI header in DigIko2



Slika 2. Schema of <teiHeader> element

Additional guidance on how to represent participant information and interaction settings is provided in the sections of the TEI Header/Profile concerned with transcribed speech (cf. Section 5.2.9 of the Guidelines). These recommendations specify which interview-specific attributes should be recorded, such as roles, sampling criteria, recording conditions, transcription methods, and where they should be encoded within the header. This approach fosters the discoverability, interoperability and long-term sustainability of spoken language resources [33].

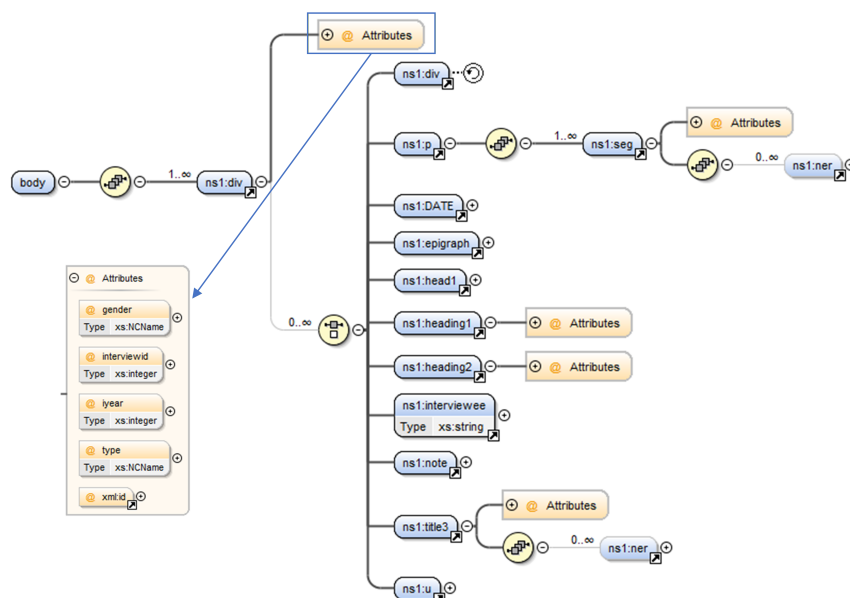
The diagram 3 illustrates a TEI-aware XML schema where each interview is encoded within a `div` element enriched with descriptive attributes (gender, interview ID, year, type). For example, in *DigIko2* we use this element for different types of text divisions and text type is defined by type attribute. The following Sample 1.3 presents an interview introduction that provides context for the reader (or listener). The internal structure follows TEI's Transcriptions of Speech recommendations, segmenting the discourse into `u` units, each containing paragraphs (`p`), segments (`seg`), and optional entity annotations (`name`). Additional structural metadata such as titles, epigraphs, notes, and participant roles are supported through labeled heading elements and a dedicated interviewee field. This schema enables precise modeling of conversational flow while supporting linguistic annotation, knowledge-graph linking, and textometric analysis.

```

<div type="introduction"><p><seg id="1382">Domen .yu je registrovan negde 1989. ili
1990. godine od strane <name type="ORG" ref="https://www.wikidata.org/wiki/
Q5918873"> Instituta „Jožef Stefan“</name> u <name type="LOC" ref="https://www.
wikidata.org/wiki/Q437">Ljubljani</name>.</seg></p>[...]</div>

```

Sample 1.3. Element `<div>` in *DigIko2*



Slika 3. Schema of div element

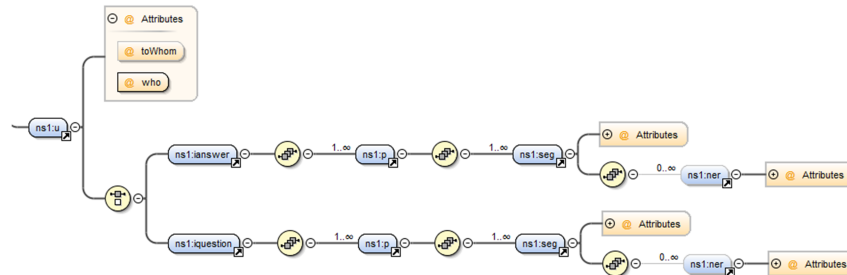
Figure 4 illustrates the structural model used for encoding question–answer pairs within the conversational layer of the corpus. Each utterance is represented using the `u` element, which can optionally contain two attributes: `@who`, identifying the speaker, and `@toWhom`, indicating the intended addressee. Within `u`, the model allows the occurrence of either `iquestion` or `ianswer` elements, reflecting the annotation of dialogic roles. The following example in Sample 1.4 illustrates an excerpt from *DigIko2* where the `u` element is used to group the speaker turn.

```
<u who="#TaVu"><question><p><seg id="1407">Šta je potrebno izmeniti od podataka  
nalazi se u dokumentu koji možemo da pronademo na sajtu <name type="ORG" ref="<br>https://www.wikidata.org/wiki/Q136527687">Registra nacionalnog domena Srbije</<br>name>.</seg><seg id="1408">Da li je sada već kasno?</seg></p></question></u><br><u who="#MiTa" toWhom="#TaVu"><answer><p><seg id="1409">Za neke manje korekcije<br>nije kasno.</seg>[...]</p></answer></u>
```

Sample 1.4. Element $\langle u \rangle$ in *DigIko2*

Both `iquestion` and `ianswer` share an identical internal structure: they consist of one or more paragraphs (`p`), which are further segmented into `seg` units. These segments serve as operational anchors for sentence-level alignment, citation, and downstream linguistic processing. The schema also allows optional `name` elements associated with specific segments, enabling the annotation of named entities with additional attribute-level metadata. This design ensures that conversational flow remains TEI-compliant while remaining compatible with NLP pipelines that require fine-grained, segment-level supervision. This

structural layer segments discourse and provides context windows for downstream NLP as well as structure-aware querying (e.g., per speaker).



Slika 4. Schema of u element

3.2. Textometry in TXM

TXM facilitates import of a variety of textual formats, including plain texts and XML documents encoded according to TEI guidelines and generates an intermediate XML-TXM representation optimized for analysis [11]. This representation identifies three core text units:

- Textual units (e.g., books, articles, interviews) with associated metadata such as author, title, date, or genre;
- Structural units (chapters, passages, segments) representing document organization;
- Lexical units (tokenized words) that include graphic form, lemma, and grammatical category.

Metadata are treated as structured units within TXM and can be used for corpus partitioning, filtering, and comparison. Through frequency lists, concordance views, and sub-corpus extraction, TXM enables qualitative inspection, while quantitative techniques, such as specificity analysis, correspondence analysis, co-occurrence analysis, and clustering, allow detection of lexical trends and contextual patterns across texts. Results may be exported in tabular or graphical formats for further processing or visualization.

Although TXM supports automatic tagging and lemmatization via TreeTagger [29], it also accommodates external NLP tools whose annotations are added to the XML-TEI representation. In addition, TXM supports CQL (Corpus Query Language) querying, enabling expressive regular-expression search within structured corpora. Overall, TXM serves as a methodological bridge between classical textual analysis and computational linguistics. In the context of this research, TXM aligns well with the goals of the *DigIko2* corpus, where structured representation, linguistic annotation, and flexible querying pave the way for subsequent semantic enrichment and AI-based exploration.

In addition to providing morphosyntactic enrichment, the TXM platform functions as a comprehensive workbench for corpus exploration and quantitative linguistic reporting. It enables the calculation of a broad spectrum of textual statistics [19] and supports textometric analysis [11]. The strength of TXM lies in its integration of search, inspection, and measurement within a single interface, allowing researchers to move fluidly across analytical levels: from qualitative evidence via concordances, to quantitative measurements

through frequency counts, and further to multivariate statistical models. Crucially, the interface preserves links back to the original text passages, ensuring interpretability and supporting a tight connection between empirical findings and source-based validation.

Building on earlier work with the *SrpELTeC* corpus [30], the exploratory use of TXM is organized around several complementary analytical functions. Qualitative investigation is supported by: (1) concordance retrieval driven by the CQP engine and CQL syntax; (2) frequency lists calculated over tokens, lemmas, POS categories, or structural layers (including named entities); (3) progression plots that visualize lexical patterns across individual texts or corpus segments; and (4) an HTML-based reading interface that is tightly cross-linked with other analytical modules. The contrastive corpus relies on subcorpora or partitions, based on criteria such as text type, period, speaker, or metadata, and subsequently compare their statistical profiles using a unified workflow [12].

3.3. Corpus Management and Queries

TXM textometry is combined with *JeRTeh* language resources [31] and TESLA models [41], which provide essential preprocessing components, including sentence segmentation and supplementary linguistic annotations, thereby forming the foundation for further semantic enrichment and corpus-based inquiry.

The NER and NEL components are based on *TeslaXLM_2025*¹⁵ trained as part of the TESLA-NER-NEL corpus [17]. The dataset includes 150,000 sentences, with entities linked to Wikidata identifiers. Annotation covers eight entity types: persons (PERS), places (LOC), organizations (ORG), events (EVENT), roles/professions (ROLE), works of art (WORK), demonyms (DEMO), and products (PRODUCT). Although joint NER+NEL architectures can enhance performance, they demand large annotated datasets and high computational capacity. To balance accuracy and efficiency, our approach utilizes spaCy's shared transformer backbone, enabling the joint training of both NER and NEL layers while preserving computational tractability. After applying the NER model, the recognized spans are inserted directly into the TEI-encoded corpus. The NEL step then links each entity to corresponding Wikidata items, resolving ambiguity using contextual cues; when necessary, new items are added to the knowledge base. This procedure normalizes references across different inflected forms by assigning them a unique identifier (e.g. a Wikidata QID), effectively anchoring the corpus within a broader knowledge-graph ecosystem and enabling semantic interoperability, data integration, and advanced inference.

CQL queries that combine structural and lexical annotations are supported in TXM and INCEpTION [18] instance maintained by JeRTeh Society.¹⁶ From the annotated corpus, we also generated a linked-data representation in NIF (NLP Interchange Format) [13], following the approach used for the *SrpELTeC* corpus [16], with additional speaker-level metadata. The NIF is a Semantic Web-based RDF/OWL format that enables interoperability between different NLP tools, annotations, and corpora. It provides a standardized way to represent text spans, offsets, linguistic annotations (e.g., NER tags, links to Wikidata) and supports multilingual annotation workflows. NIF makes it possible to integrate heterogeneous resources into Linked Open Data, allowing annotations to be queried with SPARQL, visualized, and reused in downstream applications such as

¹⁵ Serbian language model built on xlm-roberta-large

¹⁶ <http://inception.jerteh.rs/>

knowledge graphs and RAG pipelines. In our work, NIF was used to export annotated TEI segments, preserve character-level spans, and link entities to Wikidata, thereby enabling interaction with external knowledge bases and graph visualization. A build has been deployed to a local Apache Jena Fuseki¹⁷ endpoint to enable retrieval and analysis via SPARQL.

3.4. From Interview Transcripts to Short QA Pairs

The conversational structure of the *DigIko2* corpus, encoded in TEI with explicit speaker turns (*u*), paragraph units (*p*), and fine-grained segments (*seg*), provides an operational foundation for deriving question–answer (QA) items suitable for supervised QA training and RAG-oriented querying. Each interview turn contains both an interrogative component (often marked in *iquestion*) and an extended response (*ianswer*), which can be semantically decomposed into multiple atomic information units. Instead of treating a full turn as a single QA pair, we segment longer questions and answers into coherent discourse chunks, typically at the level of individual *seg* elements or small groups of contiguous segments. This enables the generation of multiple focused QA pairs grounded in a shared context, each targeting distinct facts or aspects.

To operationalise this process, each answer segment is analyzed in terms of factual content, temporal markers, named entity annotations, and narrative transitions. Questions are then reformulated in a shorter, direct manner (e.g. “Who was involved in the Galaksija computer project?”) while preserving alignment with the original TEI context. For model training and evaluation, answers are stored either as (i) short extractive spans inside the relevant *seg* or (ii) concise abstractive paraphrases of one to three consecutive *seg* elements.

Previous studies suggest that the typical length of a question ranges from 5–25 tokens, while extractive answer spans vary between 3–30 tokens [27]. In contrast, abstractive answers tend to be substantially longer, often spanning 10–80 tokens, especially in dialogue-oriented QA settings [5,28]. Each QA item records its provenance via TEI identifiers (e.g. `xml:id` of the *u* or *seg*) and includes references to Wikidata QIDs when linked entities were detected during NEL.

In this procedure, a single dialogue turn can yield multiple QA pairs based on separate facts, narrative transitions, or entity mentions. TEI structure therefore acts as a natural scaffold for question segmentation, while NER/NEL layers provide entity-aware anchors for retrieval. This design is intentionally compatible with RAG pipelines: TEI segments or turns can serve as retrievable passages, entity links enable knowledge-graph expansion, and QA pairs provide supervision for evaluating both retrieval accuracy and answer grounding. Consequently, the semantically enriched interview corpus evolves into a RAG-ready resource, supporting automated reasoning and evidence-based exploration of oral-history materials.

We employed two versions of the OpenAI language model: GPT-4.1 and GPT-5.1, accessed via the official OpenAI API in November 2025 [25]. Both models were prompted using a controlled experimental setup for paraphrasing and restructuring QA pairs. In pipeline, the prompt is designed to standardize faithful and style-controlled restructuring and paraphrasing of extracted question–answer pairs. It takes two textual inputs,

¹⁷ <http://fuseki.jerteh.rs/#/dataset/digico/query>

question_in and answer_in, corresponding to the original question and answer, and produces a JSON-formatted output consisting of a list of QA pairs. These outputs represent concise Serbian summaries that remain semantically aligned with the source content.

A predefined instruction block enforces key generation constraints required for evaluation and downstream processing, including a neutral professional tone, elimination of redundancy, omission of irrelevant details, and strict semantic fidelity. By explicitly defining constraints, the prompt enables seamless integration with the retrieval module while ensuring that the quality of generated outputs is governed by transparent and verifiable criteria.

In practice, the user provides the question_in and answer_in fields with raw inputs (i.e., the content of an <iquestion> and its corresponding <ianswer> within a <u> element). The GPT model, accessed via the OpenAI API, processes these inputs and generates JSON list of simplified QA pairs. Each pair captures distinct factual elements, entity mentions, or narrative transitions present in the original question–answer content.

4. Serialization and Exploitation Framework for the *DigIko2* Corpus

The *DigIko2* is derived from the book *Digital Icons 20+* [38], which compiles selected interviews originally broadcast on the radio program. The book arranges these interviews into thematic sections that illustrate how digital society has evolved over the past two decades. It functions as a documentary account of the digital transformation in Serbia, capturing its impact on everyday life, cultural change, and the policy environment for technology adoption, both nationally and in a broader regional context [37]. Its primary aim is to highlight the importance of Serbia’s digital transition and to identify, from a curated “bird’s-eye” perspective, the key events, institutions, and individuals that have influenced this process.

The collection is designed to stimulate further interdisciplinary dialogue, not only with specialists in IT but also with experts from various fields. In total, the book includes interviews from 57 broadcast episodes and features 67 distinguished guests, many of whom actively contributed to shaping Serbia’s digital history, offering an insightful narrative that links local developments to wider international trends. It includes, among others, conversations with researchers, engineers, and IT specialists who are, sadly, no longer with us, such as Vukašin Masnikosa, a member of the team that built Serbia’s first digital computer (CER-10), and Dejan B. Popović, a close collaborator of the creator of the “Belgrade Hand”, Academician Aleksandar Marinčić, and the author of numerous publications on Nikola Tesla. Notable international guests, Richard Stallman, Lev Manovich, Mike McGee, Eugene Kaspersky, and others—also appear in *Digital Icons 20+*, alongside many prominent figures in Serbian IT. In total, the volume features 67 interlocutors.

The *DigIko2* corpus provides a curated collection of interviews that chronicle the development of Serbia’s digital ecosystem and technological landscape. In addition to its documentary value, the corpus represents a valuable reference resource for grounded and evidence-based NLP research, particularly within the context of low-resourced languages.

4.1. TEI-Encoded Representation of the *Digiko2* Corpus

The source material was originally prepared in *Microsoft Word* using a set of carefully designed custom styles, ensuring a consistent structural segmentation of the text. Each stylistic category captured a specific functional layer of the interview: *metadata about the session*, the *introductory segment* presenting the discussion topic and interlocutors, as well as clearly distinguished *interviewer questions* and *interviewee responses*. This structured annotation in Word served as the basis for an automated conversion pipeline, implemented in Python, which transformed the transcripts into TEI-compliant XML. The script mapped each custom Word style to a corresponding TEI element, encoding interview/session boundaries via `div` elements, marking paragraph breaks with `p`, and assigning unique speaker identifiers to `u` tags.

The Sample 1.5 shown in listing below illustrates a representative fragment of a single interview. The structure is rooted in a `div type="interview"` element, enriched with administrative attributes such as `@gender`, `@interviewid`, and `@iyear`. This is followed by a `DATE` block (containing `month` and `year` elements), a descriptive interview title encoded using `title`, and the identification of the interlocutor via `interviewee`.

The discourse is subsequently divided into `div type="introduction"`, which is the thematic context of the conversation, and `div type="conversation"`, which contains the actual dialogic exchange. Speaker turns are encoded as `u` elements, each associated with a distinct speaker through the `@who` and `@toWhom` attribute to indicate addressed recipients). Within these `u` elements, the textual content is further structured into sentences and paragraphs using `p`, while fine-grained segmentation is achieved through `seg` tags. This level of granularity facilitates alignment, detailed citation practices, cross-referencing, and a wide range of textometric and computational analyses.

The tags related to the interview structure are shown in red, NER tags are annotated with two attributes: `type` and `ref`. Attribute `type` has one of the values from the list: `PERS`, `ORG`, `LOC`, `WORK`, `DEMO`, `ROLE`, `PRODUCT`. Attribute `ref` is the link of the entity in Wikidata knowledge base. For this fragment (Sample 1.5), many parts are omitted, but first QA pair is not, because it will be later used for demonstration of set of shorter QA pairs (in Subsection 4.5).

```
<div type="interview" gender="F" interviewid="08" iyear="2010">
  <DATE> <month>MART</month> <year>2010.</year> </DATE>
  <title id="1381">Opraštanje od .yu domena</title>
  <interviewee>Sagovornica: Mira Tasić</interviewee>
  <div type="introduction"><p><seg id="1382">Domen .yu je registrovan negde 1989. ili
    1990. godine od strane <name type="ORG" ref="https://www.wikidata.org/wiki/
    Q5918873">Instituta „Jožef Stefan“</name> u <name type="LOC" ref="https://www.
    wikidata.org/wiki/Q437">Ljubljani</name>.</seg><seg id="1383">Ta institucija
    je bila vođa projekta akademske mreže tadašnje <name type="LOC" ref="https://
    www.wikidata.org/wiki/Q83286">Jugoslavije</name>.</seg>[...]</p><p><seg id="
    1386">Danas u podne, u skladu sa odlukom <name type="ORG" ref="https://www.
    wikidata.org/wiki/Q242540">Internet korporacije za dodeljena imena i brojeve</
    name> (<name type="ORG" ref="https://www.wikidata.org/wiki/Q242540">ICANN</
    name>), prestaje da funkcioniše .yu domen.</seg><seg id="1389"><name type="ORG
    " ref="https://www.wikidata.org/wiki/Q136527687">Registar nacionalnog domena
    Srbije</name> podseća građane da će sve internet adrese koje koriste a završ
    avaju se sa .yu postati nevažeće, a sve internet prezentacije sa istim završ
    etkom biće nedostupne od danas u podne.</seg></p></div>
  <div type="conversation">
    <u who="#TaVu"><iquestion><p><seg id="1390">Danas je veoma važan datum za <name
      type="DEMO" ref="https://www.wikidata.org/wiki/None">srpski</name> internet.
      </seg><seg id="1391">O tome šta će se tačno desiti danas u 12 sati govoriće

```

```

nam <name type="PERS" ref="https://www.wikidata.org/wiki/None">Mira Tasić</
name>, jedna od ključnih ličnosti za razvoj interneta u <name type="LOC" ref
="https://www.wikidata.org/wiki/Q403">Srbiji</name>.</seg></p><p><seg id="
1392">Šta će se zaista desiti?</seg><seg id="1393">Na koji način ćemo se
konačno oprostiti od .yu domena?</seg></p></iquestion></u>
<u who="#MiTa" toWhom="#TaVu"><ianswer><p><seg id="1394">Za ovaj korak smo se
pripremali dve godine; od trenutka kada je počeo da se koristi .rs domen,
znali smo da će .yu domen morati da ode sa interneta.</seg><seg id="1395">
Trebalo je da se to dogodi u septembru prošle godine, u međuvremenu smo
postigli produžetak rada .yu domena još pola godine.</seg><seg id="1396">Svi
nazivi domena su zabeleženi na domen serverima.</seg><seg id="1397">Domen
server je mašina kojoj je posao da, kada dobije upit od bilo koga na
internetu za određeni domen, vrati mrežnu adresu računara na kome se nalazi
odgovarajući domen.</seg><seg id="1398">Naš domen server za domen .yu nalazi
se na <name type="ORG" ref="https://www.wikidata.org/wiki/Q6590075">
Elektrotehničkom fakultetu Univerziteta u Beogradu</name> (<name type="ORG"
ref="https://www.wikidata.org/wiki/Q6590075">ETF</name>).</seg><seg id="1399
">On u ovom trenutku ima oko 4000 .yu domena koji su produženi da važe do
kraja postojanja .yu domena na internetu.</seg><seg id="1400">Danas u podne
<name type="ROLE" ref="https://www.wikidata.org/wiki/Q255063">administrator<
/name> će sve podatke o tih 4000 domena skloniti sa servera koji je određen
da odgovara na pitanja gde se nalaze .yu domeni.</seg><seg id="1401">U slede
ća 24 sata ta informacija će se propagirati preko interneta.</seg><seg id="
1402"><name type="ROLE" ref="https://www.wikidata.org/wiki/Q278368">
Korisnici</name> koji budu pokušali da dođu do adresa koje se završavaju .yu
domenom dobiće poruku da ova adresa više ne postoji.</seg><seg id="1403">U
toku današnjeg dana, <name type="ORG" ref="https://www.wikidata.org/wiki/
Q242540">ICANN</name> će sa svojih osnovnih servera ukloniti zapis koji
ukazuje na naš server na <name type="ORG" ref="https://www.wikidata.org/wiki/
Q6590075">ETF-u</name> gde se nalaze podaci o .yu domenima.</seg><seg id="
1404">Od tog trenutka .yu domen definitivno nestaje sa interneta.</seg></p><
p><seg id="1405">To znači da će elektronske adrese koje koristimo, a koje su
formirane na .yu domenu, postati nevažeće, sve internet prezentacije sa
istim domenom će takođe biti nevažeće.</seg><seg id="1406">U pripremi za
nestanak .yu domena sa interneta bilo je potrebno izvršiti izmenu podataka u
protokolu svih servisa koji koriste nazive domena, od elektronske pošte,
preko podešavanja veb-stranica i linkova ka njima.</seg></p></ianswer></u>
<u who="#TaVu"><iquestion><p><seg id="1407">Šta je potrebno izmeniti od podataka
nalazi se u dokumentu koji možemo da pronademo na sajtu <name type="ORG" ref
="https://www.wikidata.org/wiki/Q136527687">Registra nacionalnog domena
Srbije</name>.</seg><seg id="1408">Da li je sada već kasno?</seg></p></
iquestion></u>
<u who="#MiTa" toWhom="#TaVu"><ianswer><p><seg id="1409">Za neke manje korekcije
nije kasno.</seg><seg id="1410">Ako se koriste adrese koje se završavaju na
.yu i neki korisnik pošalje imejl na takvu adresu, mejl-servis će pokušavati
24 sata da isporuči tu poštu.</seg><seg id="1411">Nakon toga korisnik će
dobiti poruku da taj mejl više ne postoji i moraće da se raspita na neki
drugi način, telefonom ili preko veb-adrese kako da dođe do nove
imejl-adrese.</seg>[...]</p></ianswer></u>
[...]
```

```

<u who="#TaVu"><iquestion><p><seg id="1487">Volela bih da spomenemo i ćirilički
domen na internetu, koji je u pripremi.</seg>[...]</p></iquestion></u>
<u who="#MiTa" toWhom="#TaVu"><ianswer><p><seg id="1489">U dogovoru sa <name type=
"ORG" ref="https://www.wikidata.org/wiki/Q38216395">Ministarstvom za
telekomunikacije</name>, inicijativu je pokrenuo <name type="ORG" ref="https:
//www.wikidata.org/wiki/Q136527687">Registar nacionalnih internet domena
Srbije</name>.</seg>[...]<seg id="1494">Procedura izbora nacionalnog ćirilič
kog domena mora da bude transparentna, svi detalji procesa moraju da budu
stavljani javnosti na uvid.</seg></p></ianswer></u></div>
</div>

```

Sample 1.5. XML representation of interview structure and NER+NEL tags

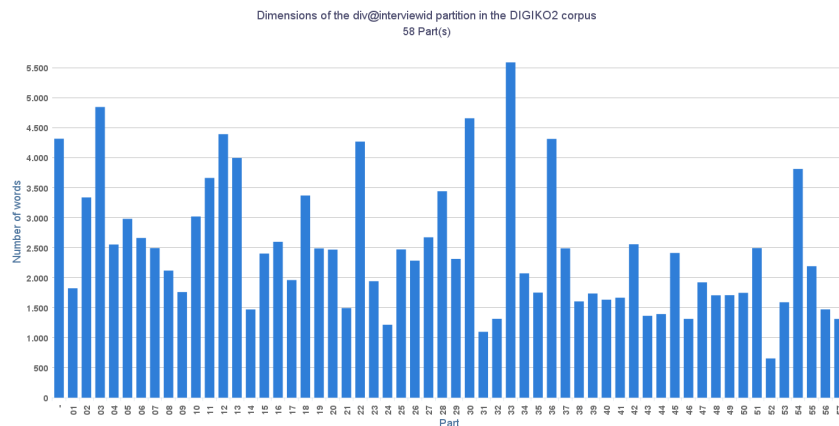
For pragmatic marking, we wrap question/answer spans with project-specific elements `iquestion` and `ianswer`; in a standards-conformant profile these may be modeled as `seg type="question"/seg type="answer"`. The schema enforces (i) controlled values for `@type` on `div`, (ii) required `@who` on `u`, and (iii) content models `u`

$\rightarrow (p+)$ and $p \rightarrow (seg*)$, ensuring interoperable structure for NER/NEL offsets, speaker-level provenance, and downstream linked-data serialization.

4.2. Textometric Methods Applied to *DigIko2* via TXM

The *DigIko2-TEI* corpus was enriched with part-of-speech (POS) and lemmatization layers using the TXM platform [12,11], enabling corpus-wide linguistic analysis and structured querying. The current version consists of 57 TEI-encoded interviews, amounting to 1,334 annotated utterances, of which 665 are questions and 669 are responses—the latter exceeding question counts because certain prompts were addressed to multiple interlocutors.

The corpus comprises a total of 8,868 sentence-level segments, aligned with the *seg* layer, and contains 142,665 running word tokens, yielding sufficient linguistic density for both qualitative inspection and quantitative modelling. Corpus size distribution across subpartitions that represents interviews is illustrated in Figure 5. The minimum length is 654 words, the maximum is 5,591, with an average of 2,423 words and a standard deviation of 1,061. The histogram visualizes word counts per interview, including corresponding introductory sections as first bar. Similar partitioning can be done by year, speakers, gender of speakers, etc., all, selected or grouped classes. This statistic provides an initial profile of corpus dimensions and are instrumental for downstream analyses such as topic analysis, diachronic comparison, and speaker-based profiling.

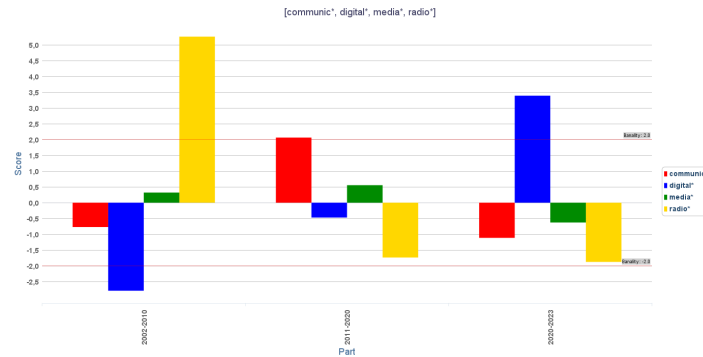


Slika 5. Corpus partition by interviews

Figure 6 presents the specificity analysis for four technology-related concepts across three temporal periods of the *DigIko2* corpus. The bars indicate the specificity score of each concept, where positive values reflect a frequency above what would be statistically expected under a uniform distribution, while negative values reflect underuse. A conventional “banality” threshold is set at -2 , below which a term is considered not characteristic for the given corpus segment. The concepts were operationalized using CQL search expressions; for instance, *communic** captures forms such as *komunikacija*, *komunikacij-*

ski, komunikacioni by expression: [srlemma="komunikac.*"], while *digital** retrieves *digitalan, digitalizacija, digitalizovati*, by expression: [srlemma=digital.*].

The resulting temporal patterns reveal meaningful conceptual shifts across the corpus. During the early interviews (2002–2010), the concept *radio* shows positive specificity, reflecting its well-established presence in the discourse of that period. In the middle period (2011–2020), communication-related terms are only marginally specific, indicating a more diffuse or transitional relevance. The most notable development concerns the concept *digital*, which exhibits negative specificity in the earliest period but becomes strongly positive in the most recent segment (2020–2023), clearly surpassing all other terms. This sharp rise points to a reorientation of public discourse towards digitally driven paradigms, aligned with broader technological shifts and growing societal awareness of digital transformation.



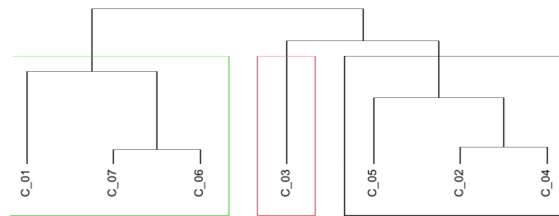
Slika 6. Specificity over time for selected key terms

The interviews in the *DigIko2* corpus were originally grouped into seven thematic chapters, each reflecting a specific perspective on the evolution of digital society in Serbia. These chapters range from biographical and narrative portrayals of key figures (e.g., *C_01* and *C_07*), to domain-oriented discussions on engineering, communication and art (*C_02*, *C_04* and *C_05*), and broader socio-cultural reflections in *C_03*. In order to assess whether the human editorial structure corresponds to linguistic patterns in the data, we applied a TXM-based hierarchical clustering procedure, relying on lexical distributions and structural features extracted from the TEI-encoded corpus. The resulting dendrogram (Figure 7) reveals three macro-clusters that offer meaningful insights into the corpus composition and thematic coherence.

The first cluster (*C_01*, *C_07*, *C_06*) groups interviews that share a predominantly narrative style and exhibit a biographical or reflective tone. These chapters tend to foreground personal trajectories, digital icons, and literary perspectives rather than technical discourse, which may explain their strong linguistic similarity. The second cluster emerges as a singleton, containing only *C_03* (*Society and the Individual*), which behaves as a linguistic outlier. This is consistent with chapter content: the interviews in *C_03* involve broader commentary on societal implications of digital transformations and include heterogeneous registers, ranging from philosophical to policy-oriented discourse. Its sepa-

ration in the dendrogram therefore underscores a distinctive conceptual mode that bridges personal narrative and analytical reflection.

The third cluster links *C_05*, *C_02* and *C_04*, all of which revolve around applied technological domains, including engineering, internet infrastructure, communication practices, and artistic uses of digital media. These chapters show stronger concentrations of domain-specific vocabulary and technical terminology, resulting in higher linguistic cohesion at the lexical level. This correspondence between human categorization and computational clustering suggests that thematic organization can be partially inferred from textual features alone, although certain hybrid or meta-level reflections (as seen in *C_03*) may require additional semantic and discourse-oriented analyses. Overall, the TXM clustering supports the structural logic of the original editorial design and highlights the potential of corpus-driven methods to inform future curatorial or annotation strategies.



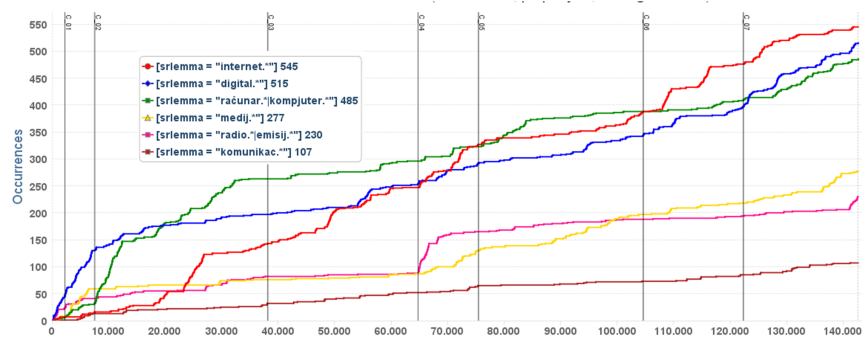
Slika 7. Hierarchical clustering, cluster dendrogram over chapters

Figure 8 presents the cumulative progression of six technology-related concepts across the *DigIko2* corpus chapter partitions. Each line represents the cumulative frequency of a CQL-based lemma expression (*internet**, *digital**, *računar*|kompjuter**, ...) tracked across the full linear order of interviews. The vertical grid lines correspond to transitions between chapters as corpus partitions. The progression curves reveal clear shifts in terminology and discourse focus over time. Terms related to *internet*, *digital*, and *računar/kompjuter* (computers) demonstrate strong and steadily increasing usage, with marked acceleration in the second half of the corpus. Notably, *internet.** becomes the dominant term in later segments.

Conversely, terms associated with *radio* and *emission* display a modest and relatively stable trajectory, reflecting the original media format of the interviews but decreasing thematic centrality. The lexical field *medij** shows a similar but stronger growth trend, particularly in later segments, likely aligned with discussions of digital journalism, media convergence, and online platforms. The term *komunikac** remains the least frequent overall but exhibits steady incremental growth, suggesting that communication-related discourse emerges more strongly as the digital paradigm matures. Taken together, the cumulative trajectories indicate a clear transition from traditional media terminology toward digitally centered concepts.

4.3. Semantic Annotation of the *DigIko2* Corpus in INCEpTION

Figure 9 illustrates the local deployment of the INCEpTION platform, which was used to manually refine the automatically generated NER layer and to verify or assign NEL links to Wikidata identifiers. In this environment, annotators directly select spans of running text and assign entity types, following the inventory described in Subsection 3.3.



Slika 8. Progression with selected queries

Each annotated mention is then linked to a corresponding Wikidata entry, such as *Mihajlo Pupin Institute* (Q6589283), thereby grounding the annotation in a canonical knowledge base. The left panel provides an ordered list of annotations grouped by different option, in this case text position; the central panel highlights entity spans with color-coded markers; and panel with NER class and Wikidata metadata, assisting annotators in resolving ambiguity. This explicit linking between surface forms and QIDs facilitates consistent disambiguation and ensures that all annotations are both human-interpretable and machine-readable.

By capturing Wikidata identifiers together with entity spans and their assigned roles, the INCEpTION workflow produces a reproducible and interoperable annotation layer suitable for downstream processing. The resulting corpus is therefore compatible with knowledge-graph construction, entity-based retrieval, and RAG-style question answering, offering a principled foundation for future AI-driven exploration of the *DigIko2* corpus.

The screenshot displays the INCEpTION interface for Named Entity Recognition (NER) and Named Entity Linking (NEL). On the left, a sidebar lists annotations grouped by position, showing various roles like 'computer engineer', 'consultant', 'coauthor', 'doctor', 'electrical engineer', 'pioneer', and 'physician'. The central panel shows a text snippet with color-coded markers for named entities and their roles. The right panel displays Wikidata metadata for the identified entities, including the Mihajlo Pupin Institute (Q6589283) and the Radio Beograd 2 (Q100000000).

Slika 9. NER and NEL within INCEpTION

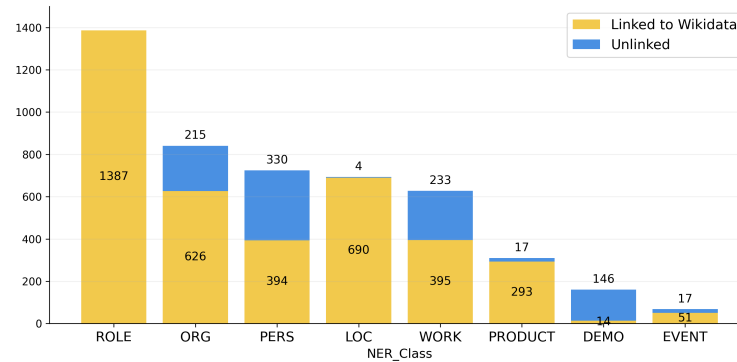
To extract annotated organisations in INCEpTION query would be: `<Named_entity.value = "ORG"/>`. For linked entities, for example Q6589283 *Mihajlo Pupin Institute*, query is: `<Named_entity.identifier = "http://www.wikidata.org/entity/Q6589283"/>`. To extract annotated organisations in TXM query would be: `<name_type="ORG» []+</name>`. The most frequent entity type is ROLE 1,387 (329), where in brackets are distinct occurrences. The most frequent roles are: umetnik (artist, 79), profesor (professor, 70), student (55), novinar (journalist, 44), korisnik (user, 41). The second most frequent class is ORG 841 (469), followed by: PERS 724 (384), LOC 694 (181), WORK 628 (307), PRODUCT 309 (140), DEMO 160 (45), and EVENT 68 (48).

4.4. Publishing *DigIko2* as Linked Data via NIF Serialization

In order to enable interoperable, Linked Data friendly processing of the *DigIko2* corpus, we developed a custom export pipeline from TEI to the NLP Interchange Format (NIF). Starting from TEI-encoded segments with manually corrected NER and NEL layers, each `<seg>` unit is converted into a `nif:Context/nif:String` with canonical character offsets, while every `<name>` element is mapped to a `nif:Phrase` anchored via `nif:anchorOf` and `nif:beginIndex/nif:endIndex`. Entity links are preserved by attaching Wikidata identifiers from the TEI `@qid` attribute as `itsrdf:taIdentRef`, and speaker information from the TEI header (e.g. `<person xml:id="...">` with `@role` and `@qid`) is propagated into the NIF graph using a custom `di:speaker` property pointing to `foaf:Person` nodes. In this way, *DigIko2* becomes a richly annotated NIF corpus that simultaneously encodes textual spans, entity linking, and discourse roles, and can be queried with SPARQL or visualized via Wikidata Query Service to analyse who mentions which entities in the interviews, in line with the general NIF design principles [13].

After automatic annotation, manual curation was performed on the entire *DigIko2* corpus. Figure 10 presents the absolute number of annotated entity spans, distinguishing between those linked to Wikidata and those not linked. Out of the total annotated spans, ROLE and LOC show almost complete coverage, with 100% and 99.4% successfully linked to Wikidata, respectively, indicating that occupational roles and geographical names are highly standardised and consistently recognised. PRODUCT also shows high alignment (94.5%), followed by EVENT (76.1%), ORG (74.3%), and WORK (63.0%), where partial Wikidata coverage still allows for meaningful knowledge-graph construction but leaves room for targeted enrichment. In contrast, PERS exhibits a more balanced distribution (54.1% linked vs. 45.9% unlinked), suggesting the need for additional manual disambiguation of personal names, especially in cases of local relevance or ambiguous referents. The DEMO class (demographic references) remains largely unlinked (only 8.8% coverage), signalling that this category requires the most attention in future annotation cycles, but this is related how this type of data is recorded in Wikidata. Taken together, these results provide actionable insight into which entity classes are mature enough for direct knowledge-graph ingestion and which require further curation to ensure full interoperability with external knowledge bases such as Wikidata.

In the following fragment from *DigIko2* NIF edition (Sample 1.6), the TEI segment containing the sentence “*Rodoljub Šabić je poverenik za informacije od javnog značaja i zaštitu podataka o ličnosti u Republici Srbiji.*” is represented as a `nif:Context/nif:String` resource `di:seg-2886`. The fragment’s character offset



Slika 10. Number of entities that are recognized and linked with Wikidata

is specified by `nif:beginIndex` and `nif:endIndex` and a local identifier stored in `dct:identifier`.

The mention of the person *Rodoljub Šabić* is modelled as a separate `nif:Phrase` node `di:seg-2886_char_0_14`, anchored to the surface form via `nif:anchorOf` and aligned with its span in the context using `nif:beginIndex=0` and `nif:endIndex=14`. The custom property `di:nerClass` encodes the NER label "PERS", while `itsrdf:taIdentRef` property links `wd:Q19057243` as the mention to the corresponding Wikidata entity for Rodoljub Šabić. In addition, both the context and the phrase are connected to the discourse participant `di:person-TaVu` through the `di:speaker` relation, enabling downstream analysis of which speakers introduce which entities in the *DigIko2* corpus.

```

di:seg-2886 a nif:Context, nif:String ;
  di:speaker di:person-TaVu ;
  nif:beginIndex "0"^^xsd:nonNegativeInteger ;
  nif:endIndex "110"^^xsd:nonNegativeInteger ;
  nif:isString "Rodoljub Šabić je poverenik za informacije od javnog
značaja i zaštitu podataka o ličnosti u Republici Srbiji." ;
  dct:identifier "2886" .
di:seg-2886_char_0_14 a nif:Phrase ;
  di:nerClass "PERS" ;
  di:speaker di:person-TaVu ;
  nif:anchorOf "Rodoljub Šabić" ;
  nif:beginIndex "0"^^xsd:nonNegativeInteger ;
  nif:endIndex "14"^^xsd:nonNegativeInteger ;
  nif:referenceContext di:seg-2886 ;
  itsrdf:taIdentRef wd:Q19057243 .

```

Sample 1.6. Fragment from *DigIko2* NIF edition

4.5. Generating QA dataset

The generation of the new QA dataset based on the original dialogue in the *DigIko2* corpus was performed in Python on a dataset of 661 question–answer (QA) pairs. It should be noted that several questions had multiple answers because more speakers were involved in the conversation. The original questions had an average length of 189 characters, whereas the answers averaged 892 characters. We applied GPT-4.1 and GPT-5.1 (November 2025 versions) to paraphrase and reformulate the QA pairs.

Table 1 compares the original dataset with the QA pairs generated by GPT-4 . 1 and GPT-5 . 1. While the original interviews contained 665 questions and 669 answers, both models produced more than three times the number of QA pairs. This indicates that the conversation was reconstructed so that longer QA segments were split into multiple QA units, where each question had exactly one answer, thus forming a dialogue structure closer to that of a chatbot.

Tabela 1. Statistics of Original and Generated QA Pairs

Type	Original		GPT-4 . 1		GPT-5 . 1	
	N°	Avg. Chars	N°	Avg. Chars	N°	Avg. Chars
Questions	665	271	2,256	57	1,987	68
Answers	669	1,091	2,256	126	1,987	232

GPT-4 . 1 yielded the largest number of QA pairs (2256), but with a substantial reduction in length (avg. 57 and 126 chars). This suggests a fragmentation-oriented strategy, prioritising quantity over contextual depth. GPT-5 . 1, however, generated fewer QA pairs (1987) but substantially longer answers (avg. 232 characters), indicating a shift towards richer and more semantically dense content. These findings illustrate an important trade-off between coverage and context preservation. GPT-5 . 1 appears better suited for downstream tasks such as NER, NEL, and RAG pipelines, where the quality and informativeness of each QA pair are more important than their sheer number. To illustrate, we give a Sample 1.7 of one long QA pair given in Section 4.1 that was refactored into four QA pairs.

```

<generated>
<qapair xml:id="qa-152">
  <q>Šta će se tačno desiti danas u 12 sati sa .yu domenom?</q>
  <a>Danas u podne administrator će obrisati sve podatke o oko 4000 preostalih .yu
    domena sa servera na Elektrotehničkom fakultetu u Beogradu, koji je zadužen da
    odgovara na upite za .yu domene. U naredna 24 sata ta promena će se proširiti
    internetom, a ICANN će ukloniti zapis koji upućuje na ovaj server, čime .yu
    domen definitivno nestaje sa interneta.</a>
</qapair>
<qapair xml:id="qa-153">
  <q>Zašto i kako je pripreman prelazak sa .yu na .rs domen?</q>
  <a>Od uvođenja .rs domena bilo je poznato da .yu mora biti ugašen. Gašenje je
    prvobitno bilo planirano za septembar prošle godine, ali je rad .yu domena
    produžen za još šest meseci. Tokom tog perioda korisnici su trebalo da pređu
    na .rs i prilagode svoje servise novom domenu.</a>
</qapair>
<qapair xml:id="qa-154">
  <q>Šta će se desiti sa sajtovima i mejl adresama na .yu domenu?</q>
  <a>Svi sajtovi i elektronske adrese koji koriste .yu domen postaće nevažeći.
    Korisnici koji pokušaju da pristupe adresama koje se završavaju na .yu dobiće
    poruku da ta adresa više ne postoji.</a>
</qapair>
<qapair xml:id="qa-155">
  <q>Koje tehničke promene su bile potrebne zbog gašenja .yu domena?</q>
  <a>Bilo je potrebno izmeniti podatke u konfiguraciji svih servisa koji koriste
    nazive domena: elektronske pošte, veb-stranica i linkova ka njima, kako bi
    umesto .yu koristili nove domene, pre svega .rs.</a>
</qapair>
</generated>

```

Sample 1.7. Example of context-aware QA pair refactoring in GPT-5 . 1

4.6. Extending dataset

To expand our dataset, we collected and analysed a set of 586 posts published on the RTS (Radio Television of Serbia) website. For each post, we extracted the title, the introductory lead segment, the full HTML body, and the corresponding URL. After removing some HTML markup (fonts, comments, colors), the textual content was manually annotated and compiled into a unified XML file containing 586 structured entries. The average length of the title was approximately 60 characters, the lead averaged 190 characters, and the body of the post around 620 characters. The following Sample 1.8 illustrates the structure of an annotated item, where named entities are consistently marked using the `ner` element with classes for persons, organizations, roles, creative works, and products, accompanied by Wikidata identifiers to support disambiguation and interoperability.

```
<item>
  <Title id="196"><name type="PERS" ref="https://www.wikidata.org/wiki/Q131271652">
    Anđelka Zečević</name></Title>
  <Lead id="197">U emisiji <name type="WORK" ref="https://www.wikidata.org/wiki/
    Q136525085">Digitalne ikone</name> danas je sa nama <name type="PERS" ref="
    https://www.wikidata.org/wiki/Q131271652">Anđelka Zečević</name>, <name type
    ="ROLE" ref="https://www.wikidata.org/wiki/Q11900058">istraživač</name> na <
    name type="ORG" ref="https://www.wikidata.org/wiki/Q123951838">Matematičkom
    institutu SANU</name>, u sektoru za računarstvo.</Lead>
  <Link>https://www.rts.rs/lat/radio/radio-beograd-2/5632717/andjelka-zecevic.html<
    /Link>
  <Body>
    <p id="198">Razgovaraćemo o Anđelkinim oblastima interesovanja, o obradi
      prirodnog jezika, digitalnoj humanistici i medicinskoj informatici, o
      projektu u oblasti mašinskog učenja i tome kako nam obrada velike kolekcije
      podataka može pomoći u otkrivanju ranih prediktora demencije, kao i o
      aktivnostima društva za jezičke resurse, <name type="ORG" ref="https://www.
      wikidata.org/wiki/Q97203619">JerTeh</name>, čiji je jedan od osnivača.
      Poseban deo razgovora posvetili smo primeni veštačke inteligencije u
      medicini, istraživanjima koja treba da doprinesu adekvatnoj bezbednosti <
      name type="ROLE" ref="https://www.wikidata.org/wiki/Q181600">pacijenata</
      name>, kao i tome zašto kaže da alatu <name type="PRODUCT" ref="https://www
      .wikidata.org/wiki/Q115564437">ChatGPT</name> dugujemo zahvalnost za
      globalnu i najbolju reklamu u proteklih 30-ak godina, ali ga grdimo i zbog
      mnogih turbulencija koje je uveo u ovu oblast.</p>
    <p id="199">Emisiju uređuje i vodi <name type="PERS" ref="https://www.wikidata.
      org/wiki/Q109916651">Tamara Vučenović</name>.</p>
  </Body>
</item>
```

Sample 1.8. An XML representation of a post from the RTS website

In total, 9,111 entity mentions were annotated across the corpus. The distribution of categories reflects the interview-driven structure typical of RTS content: roles and professions were the most frequent type with 2,849 mentions, followed by personal names with 2,266 mentions, organizations with 2,262 mentions, geographical entities with 853 occurrences, and a smaller number of creative works, events, demographic descriptors, and products. Despite the high number of annotated mentions, only 952 entities were linked to Wikidata, predominantly geographical entities (766) and organizations (167). This disparity highlights both the richness of the corpus in context-dependent or local expressions, many of which lack stable knowledge-base identifiers, and the challenges of grounding Serbian media content into global knowledge graphs. At the same time, the dataset provides a valuable resource for evaluating NER and NEL methods and for developing Serbian-language knowledge-graph applications.

5. Discussion

The automatic transformation from MS Word into TEI required several iterations of corrections in order to achieve a consistent structure. For TEI schema generation and corpus validation XMLSpy tool was used. It should also be noted that we tested several schema variations to produce structure appropriate for the analysis, corpus partitioning, search, extraction and transformation.

The evaluation of the NER and NEL components over the whole *DigIko2* interview corpus shows that the model performs strongly on several entity types, despite the considerable linguistic variability introduced by the fact that approximately 70% of the training corpus consists of literary text. The model is transformer-based NER+NEL model based on a pretrained model: a multilingual model with 561 million parameters (TeslaXLM) [41,42]. For NEL, the results for the LOC category demonstrate high precision (0.90) with recall of 0.78, yielding an F1-score of 0.84. This indicates that the linking module is generally effective in associating recognized locations with the correct Wiki-data identifiers, although a moderate number of false negatives suggests that recall could still be improved. Other classes for NEL are at early stage.

In the NER evaluation, several categories, especially PERS, LOC, ROLE, and DEMO, exhibit strong performance. The PERS class achieves an F1-score of 0.95, reflecting highly reliable recognition of person entities across both interview-based and literary segments of the corpus. The LOC class also performs very well (F1 = 0.91), and its high recall (0.96) aligns with the consistent behavior observed in the NEL evaluation. Strong results for ROLE (F1 = 0.84) and DEMO (F1 = 0.93) further confirm that categories with clearer semantic boundaries are captured effectively. More challenging categories include ORG, WORK, and PRODUCT. The ORG class shows moderate recall (0.76) but lower precision (0.57), indicating over-generation, likely due to contextual ambiguity or similarity to other nominal expressions. The model often made errors when recognizing ORG entities such as *Radio Beograd 2*, failing to include the number 2 as part of the entity name. It also misclassified examples like *Ljudmila iz Ljubljane*, labeling it as an PERS instead of correctly identifying *Ljudmila* as an ORG entity. The WORK (F1 = 0.41) and PRODUCT (F1 = 0.28) categories suffer from low recall values (0.28 and 0.18), suggesting that many relevant entities remain unrecognized, which is expected since there was not many labeled entities of this category in the training dataset.

Overall, the model robustly handles categories with more stable linguistic and semantic patterns, such as PERS, LOC, DEMO, and ROLE, even within a hybrid corpus combining interviews and narrative prose. However, categories characterized by greater variability or domain-specificity, such as EVENT, WORK and PRODUCT, require further improvement. Future enhancements may focus on domain-adaptive training, targeted augmentation of underrepresented entity types, and improved handling of culturally or stylistically marked references. The present results establish a strong baseline for further development of NER+NEL pipelines for complex, mixed-genre corpora such as *DigIko2*.

The main sources of error likely stem from ambiguous or morphologically inflected toponyms, incomplete gazetteers, and conservative confidence thresholds. Future improvements should focus on expanding alias lists, enhancing candidate generation for inflected forms, and leveraging contextual cues such as co-occurring regions or temporal information. The proposed TEI-based structure, together with its multi-layer annotation schemas, enables a range of advanced functionalities, including semantic search, per-

Tabela 2. NER and NEL Evaluation Results on the *DigIko2* Corpus

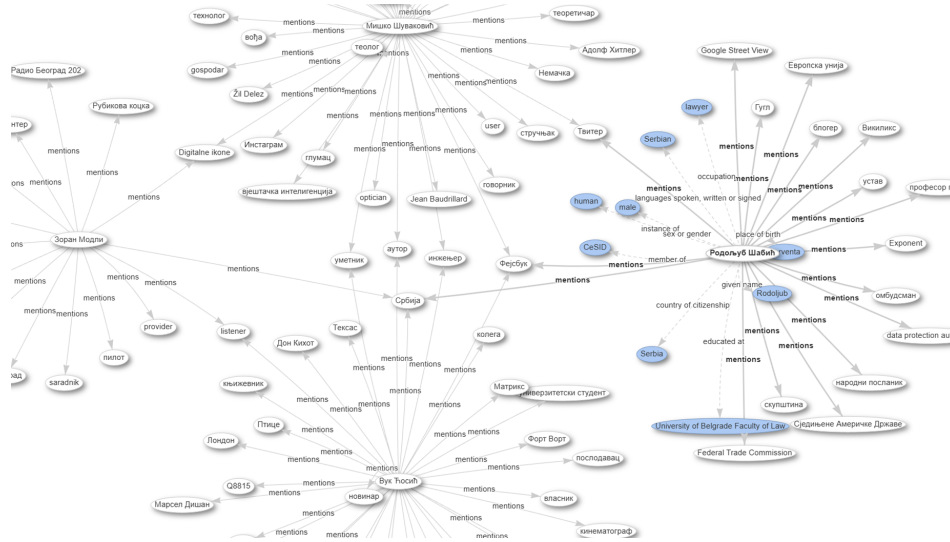
	PERS	LOC	ORG	ROLE	WORK	DEMO	EVENT	PRODUCT
TP	697	667	639	1,111	175	151	43	57
FP	42	108	488	151	51	13	12	34
FN	27	27	202	276	453	9	25	253
P	0.94	0.86	0.57	0.88	0.77	0.92	0.78	0.63
R	0.96	0.96	0.76	0.80	0.28	0.94	0.63	0.18
F1	0.95	0.91	0.65	0.84	0.41	0.93	0.70	0.28

speaker retrieval, and knowledge graph querying via SPARQL. Every TEI element is indexed with persistent pointers, allowing direct cross-referencing both within the corpus and across external resources. During retrieval, speaker-specific filters and Wikidata QID facets can be applied to return context-aware results.

The retrieval framework supports multiple evidence views: (i) *speaker-centric digests*, which restrict results to utterances of a single participant; (ii) *entity-centric synopses*, aggregating all segments linked to a given Wikidata entity; and (iii) *thematic overviews*, which provide high-level content grouping to support topic-driven exploration.

For factoid-style queries (e.g., *Who mentioned X, and in which turn?*), the system returns speaker-level and turn-level evidence, which the LLM may cite explicitly. Furthermore, the responses can be reformulated for non-expert audiences while keeping traceable references to the original source utterances. This mechanism also permits the generation of concise briefings or controlled-length executive summaries, ensuring faithful content simplification rather than paraphrastic drift. The evidence views can additionally highlight how different speakers discuss the same entity or theme, consolidate cited segments into attribution maps over the interview structure, and reveal low-confidence claims or missing links (NEL gaps). Such outputs are particularly valuable for curator inspection and iterative dataset refinement. Domain adaptation is required for NEL since many recognized entities are not linked with Wikidata items. The current knowledge graph should be completed, relations better explored. Figure 11 presents a graph where speakers (interviewees) of mentions linking locally extracted entities to Wikidata items. Each node is a person, organization, place, event, or concept; solid directed edges encode *who mentions who/what* in the corpus (e.g., *Rodoljub Šabić* and *Miško Šuvaković* → *Serbia, Facebook, Twitter*).

The obtained results for NER are highly encouraging and broadly comparable to those reported for large, well-resourced languages. In contrast, the evaluation of NEL demonstrates that substantial room for improvement remains. This should not be interpreted as a limitation of the approach, but rather as an indication of the inherent difficulty of grounding Serbian entities in global knowledge bases, where coverage is still uneven and context-dependent expressions are frequent. It is important to emphasise that this is currently the only model that explicitly addresses the NEL task for the Serbian language. The lower recall is partly a consequence of the rich and highly inflectional morphology of Serbian, but an even more important factor is the absence of many context-specific or culturally bound concepts in Wikidata, which prevents reliable linking even when the entities are correctly recognised. Intensive work on this problem is ongoing within the TESLA project, and significantly improved NEL performance can be expected in the near future as resources, lexicons, and training data continue to expand. The dataset is publicly



Slika 11. A graph: Who was mentioning who/what?

available through the DigIko open data page,¹⁸ and HuggingFace¹⁹ while a dedicated SPARQL endpoint has been established to enable programmatic access and querying of the data.²⁰

6. Conclusion

This paper introduced a comprehensive and standards-based pipeline for transforming unstructured interview transcripts into a semantically enriched, interoperable, and queryable resource. By integrating TEI-encoded structure with NER and NEL models linked to Wikidata, textometric analysis in TXM, NIF-based linked-data serialization, and retrieval augmented generation (RAG), the approach establishes a multi-layer framework for evidence-grounded exploration of the *DigIko2* corpus.

The TEI encoding provides stable anchors for segmentation, speaker attribution, metadata-driven retrieval, and reproducible citation practices, enabling granular analysis of interview discourse. The NER component achieves reliable performance for the core entity types (PERS, ROLE, LOC), reflecting the strength of Serbian transformer models trained within the TESLA project. More challenging categories—ORG, WORK, and PRODUCT—reveal the need for extended training data and richer domain-adapted lexicons. NEL performance remains promising but incomplete: high precision for locations is contrasted with moderate recall, hindered primarily by rich Serbian morphology and the limited coverage of culturally specific entities in Wikidata. These findings highlight the importance of continued knowledge-graph enrichment and improved candidate-generation strategies for low-resourced languages.

¹⁸ <https://l1lod.jerteh.rs/digiko/>

¹⁹ <https://huggingface.co/datasets/te-sla/DigIko>

²⁰ <http://graphdb.jerteh.rs/sparql?repositoryId=digiko>

The generation of derived QA datasets using GPT-4.1 and GPT-5.1 demonstrates the feasibility of producing concise, semantically faithful question-answer pairs directly from TEI-encoded transcripts. GPT-5.1, in particular, balances informativeness and lexical fidelity, making it suitable for downstream RAG applications. The resulting resources thus support multiple pathways of exploitation: semantic search, speaker- and entity-centric retrieval, SPARQL querying over NIF graphs, and evidence-based generation for digital-humanities scholarship.

Future work will expand the annotated dataset, further refine TESLA NER/NEL models, and release the TEI and NIF corpus under permissive open licenses. Planned extensions include (i) domain-sensitive NEL for non-canonical and culturally specific entities; (ii) enhanced RAG pipelines with per-speaker and per-entity evidence retrieval and citation-aware generation; (iii) construction of a richer knowledge graph integrating Wikidata with project-specific ontologies; and (iv) the generation of additional training sets to strengthen the ecosystem of Serbian NLP resources. With these developments, the *DigIko2* pipeline offers not only a replicable methodological blueprint but also a foundation for transparent, verifiable, and linguistically grounded AI applications in less-resourced language settings.

Acknowledgement. This research was supported by the Science Fund of the Republic of Serbia, #7276, *Text Embeddings – Serbian Language Applications – TESLA* and based upon work from COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, <https://www.cost.eu>).

Literatura

1. Benzécri, J.P.: L'analyse des données, vol. 2. Dunod, Paris (1973)
2. Cetoli, A., Bragaglia, S., O'Harney, A.D., Sloan, M., Akbari, M.: A Neural Approach to Entity Linking on Wikidata. In: Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, Proceedings, Part II 41. pp. 78–86. Springer (2019)
3. Dartmouth Digital History Initiative: DDHI Encoding Guidelines: Applying the Basic Layer of the DDHI Encoding Schema to Oral History Interview Transcripts (2022), <https://ddhi.dartmouth.edu/ddhi-encoding-guidelines>, accessed: 2025-11-23
4. Davis, R.C.: Synchronizing oral history text and speech: a tools overview. *Behavioral & Social Sciences Librarian* 34(4), 234–238 (2015)
5. Dinan, E., Roller, S., Shuster, K., et al.: Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In: Proceedings of ICLR (2019)
6. Dong, H., Chen, J., He, Y., Liu, Y., Horrocks, I.: Reveal the Unknown: Out-of-Knowledge-Base Mention Discovery with Entity Linking. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. p. 452–462. CIKM '23, ACM (2023)
7. Fisseni, B., Schmidt, T.: CLARIN web services for TEI-annotated transcripts of spoken language. In: Selected Papers from the CLARIN Annual Conference 2019. Leipzig, 30 September–2 October 2019. pp. 12–22. Linköping University Electronic Press (2019)
8. Francis-Landau, M., Durrett, G., Klein, D.: Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks. In: Knight et al. (ed.) Conference of North American Chapter of ACL: Human Language Technologies. pp. 1256–1261. ACL (2016)
9. Fu, X., Shi, W., Yu, X., Zhao, Z., Roth, D.: Design Challenges in Low-resource Cross-lingual Entity Linking (2020), <https://arxiv.org/abs/2005.00692>

10. Guiraud, P.L.: Problèmes et méthodes de la statistique linguistique. Synthese Library, Springer Dordrecht, 1 edn. (1960)
11. Heiden, S.: The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In: Otaguro et al. (ed.) 24th Pacific Asia Conference on Language, Information and Computation. vol. 2, pp. 389–398. Sendai, Japan (2010)
12. Heiden, S., Pincemin, B., Decorde, M.: Manuel de TXM (Jul 2015), <https://halshs.archives-ouvertes.fr/halshs-01223949>, manuel d'utilisation du logiciel TXM
13. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: International semantic web conference. pp. 98–113. Springer (2013)
14. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust Disambiguation of Named Entities in Text. In: Barzilay et al. (ed.) 2011 Conference on Empirical Methods in NLP. pp. 782–792. ACL, Edinburgh, UK. (2011)
15. Ikonić Nešić, M., Petalinkar, S., Stanković, R., Utvić, M., Kitanović, O.: SrpCNNeL: Serbian model for named entity linking. In: 2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS). pp. 465–473. IEEE (2024)
16. Ikonić Nešić, M., Stanković, R., Schöch, C., Škorić, M.: From ELTeC text collection metadata and named entities to linked-data (and back). In: Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th LREC Conference. pp. 7–16 (2022)
17. Ikonić Nešić, M., Utvić, M.: Overview of the Tesla-Ner-Nel-Gold Dataset: Showcase on Serbian-English Parallel Corpus. Technical editors p. 57 (2024)
18. Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In: Proc. of 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9. ACL (2018)
19. Krstev, C., Jaćimović, J., Šandrih, B., Stanković, R.: Analysis of the First Serbian Literature Corpus of the Late 19th and Early 20th Century with the TXM Platform. In: DH_BUDAPEST_2019. pp. 36–37. Centre for Digital Humanities - Eötvös Loránd University (2019)
20. Krstev, C., Obradović, I., Utvić, M., Vitas, D.: A system for named entity recognition based on local grammars. Journal of Logic and Computation 24(2), 473–489 (2014)
21. Kubeša, D., Straka, M.: DaMuEL 1.0: A large multilingual dataset for entity linking (2023), <http://hdl.handle.net/11234/1-5047>, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL)
22. Lebart, L., Salem, A.: Statistique textuelle. Dunod, Paris (1994)
23. Liu, Z., Janowicz, K., Majic, I., Shi, M., Fortacz, A., Karimi, M., Mai, G., Currier, K.: Operationalizing Geographic Diversity for the Evaluation of AI-Generated Content. Transactions in GIS 29(3), e70057 (2025)
24. Muller, C.: Initiation aux méthodes de la statistique linguistique. Honoré Champion, 1 edn. (1992), ISBN-10: 2852032708. Available in DJVU format
25. OpenAI: OpenAI API: ChatGPT-4.1 and ChatGPT-5.1 Models. <https://platform.openai.com/docs> (2025), models used: gpt-4.1, gpt-5.1
26. Puren, M., Cafiero, F.: InTElviews: An ODD for Qualitative Interviews in the Humanities. Journal of the Text Encoding Initiative 2024(15) (2024)
27. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of EMNLP (2016)
28. Reddy, S., Chen, D., Manning, C.D.: CoQA: A Conversational Question Answering Challenge. In: Proceedings of NAACL (2019)
29. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proc. of International Conference on New Methods in Language Processing. pp. 44–49. Manchester, UK (1994)
30. Stanković, R., Krstev, C., Šandrih Todorović, B., Vitas, D., Škorić, M., Ikonić Nešić, M.: Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection. In: Calzolari et al. (ed.) Proc. of 13th LREC. pp. 3337–3345. ELRA, Marseille, France (2022)

31. Stanković, R., Sandrih, B., Krstev, C., Utvić, M., Škorić, M.: Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In: LREC 2020-12th International Conference on Language Resources and Evaluation. pp. 3954–3962. ELRA (2020)
32. Stanković, R., Vučenović, T., Ikonić Nešić, M.: From Unstructured Interviews to Queryable Knowledge: An AI Pipeline for the Digitalne Ikone Corpus using TEI, NER, NEL, and RAG. In: Book of Abstracts of AI Conference 2025. pp. 129–130. SASA, Belgrade, Serbia (2025)
33. TEI Consortium: The TEI Header. In: TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium, version 4.10.2 edn. (2025), <https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html>, last updated 4 Sept 2025
34. TEI Consortium: Transcriptions of Speech. In: TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium, version 4.10.2 edn. (2025), <https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html>, last updated 4 Sept 2025
35. Verdonik, D., Dobrovoljc, K., Erjavec, T., Ljubešić, N.: Gos 2: A new reference corpus of spoken slovenian. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024). pp. 7825–7830 (2024)
36. Vrandečić, D., Krötzsch, M.: Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM* 57(10), 78–85 (2014)
37. Vučenović, T.: Communication Through Social Media: Blogs and Wikis in Libraries. *Qualitative and Quantitative Methods in Libraries* 10(3), 315–328 (2021)
38. Vučenović, T.: Digitalne ikone 20+ [Digital Icons 20+]. kademaska knjiga; Radio Televizija Beograd - RTS (2024)
39. Wacholder, N., Ravin, Y., Choi, M.: Disambiguation of Proper Names in Text. In: Fifth Conference on Applied Natural Language Processing. pp. 202–208. ACL, Washington, USA (1997)
40. Zhou, S., Rijhwani, S., Wieting, J., Carbonell, J., Neubig, G.: Improving Candidate Generation for Low-resource Cross-lingual Entity Linking. *Transactions of the Association for Computational Linguistics* 8, 109–124 (01 2020), https://doi.org/10.1162/tacl_a_00303
41. Škoric, M.: New Language Models for Serbian. *Infotheca - Journal for Digital Humanities* 24(1), 7–28 (2024)
42. Škorić, M., Petalinkar, S.: New XLM-R-based Language Models for Serbian and Serbo-Croatian. In: Artificial Intelligence Conference. pp. 59–60. SASA, Belgrade (2024)

Ranka Stanković is a Full Professor at the University of Belgrade, Faculty of Mining and Geology, Head of Chair for Mathematics and Informatics, vice-president of Language Resources and Technologies Society and . She specializes in informatics, statistics, and natural language processing (NLP). Her research interests span lexical resources, information extraction, semantic web technologies, geoinformatics, and artificial intelligence. She teaches informatics, geoinformatics, GIS, statistics, and programming for linguists.

Tamara Vučenović is an Associate Professor with the Faculty of Management, Metropolitan University, Belgrade, Serbia, and the Editor-in-Chief of the Educational Programme at the Radio Television of Serbia (RTS). She received her degrees from the University of Belgrade. She serves as an international expert and Vice-Chair for the European Commission (EACEA) and the OSCE. She has published numerous scientific papers, including respected venues such as the Language Resources and Evaluation Conference (LREC). Her current research interests include Digital Media, Natural Language Processing, and Artificial Intelligence Applications.

Milica Ikonić Nešić is an Assistant at the Faculty of Philology, University of Belgrade, Department of Librarianship and Information Science, and a PhD student in the Multidisciplinary PhD program, Intelligent Systems module, at the University of Belgrade. Her research interests include mathematics, natural language processing, computational linguistics, the semantic web, machine learning, and deep learning, applied to the development of tools, resources, and models for the Serbian language.

Biljana Rujevic is a Research Associate and librarian at the University of Belgrade, Faculty of Mining and Geology. She received her Ph.D. degree from the University of Belgrade, Faculty of Philology in 2022. Her doctoral research focused on electronic lexicography. Her research interests include electronic lexicography, digital humanities, and language resources.

Received: December 12, 2025; Accepted: May 23, 2026.

