Integrating TEI, NER/NEL, Textometry, and Linked Data for a Semantically Enriched Interview Corpus

Ranka Stanković*, Tamara Vučenović†, Biljana Rujević*, Milica Ikonić Nešić‡, Mihailo Škorić*

*University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka.stankovic, biljana.rujevic, mihailo.skoric}@rgf.bg.ac.rs
†University Metropolitan, Faculty of Management, Serbia
tamara.vucenovic@metropolitan.ac.rs
‡University of Belgrade, Faculty of Philology, Serbia
milica.ikonic.nesic@fil.bg.ac.rs

Abstract

This paper presents a pipeline that converts unstructured interview transcripts into a semantically enriched, queryable knowledge resource. The texts from the *Digitalne Ikone 20+* interview collection were first encoded in TEI XML (Text Encoding Initiative), marking interview boundaries, paragraph breaks, speaker turns with identifiers, dates, and topics. This structural encoding underpins downstream NLP and enables structured querying (e.g., by speaker). We then applied Named Entity Recognition to identify persons, places, organizations, and events, and embedded the results directly in TEI. In the third stage, Named Entity Linking mapped entity mentions to canonical Wikidata identifiers via context-aware disambiguation; missing entries were added to Wikidata when necessary. The resulting TEI+NER/NEL corpus, serialized as linked data, follows the NIF (NLP Interchange Framework). The pipeline also supports retrieval-augmented summarization that retrieves evidence passages and prompts LLMs (implemented with DSPy) to produce faithful interview summaries. We discuss design choices (TXM for textometry with JeRTeh resources; TESLA models for NER/NEL), report qualitative gains in interpretability through semantic links, and outline future work on domain-adapted NER/NEL, graph-based completion, and more expressive RAG architectures. The approach is replicable for other oral-history or media corpora and advances practical, evidence-grounded access to cultural archives and beyond.

Keywords: TEI, NER, NEL, summarization, linked data, textometry

1. Introduction

Large collections of interviews and broadcast transcripts remain underexploited in NLP because they are unstructured, heterogeneous, and costly to query reliably. This is especially true for low-resourced languages, where reusable pipelines and grounded questions and answers (QA) are scarce. Our motivation is twofold: (i) to turn a long-running Serbian interview series *Digitalne Ikone 20+* (Digital Icons 20+) (Vučenović, 2024) into a resource that supports reproducible research, semantic exploration, and evidence-based question answering; and (ii) to demonstrate a standards-driven workflow that other oral-history or media corpora can adopt.

Radio broacast *Digital Icons* curated by Tamara Vučenović has aired on *Radio Belgrade 2* since 2002, with more than 1,000 episodes. As an educational program dedicated to digital technologies, it is unique in concept and content within the Serbian media sphere. With knowledgeable experts as guests, the program covers topics such as computer science, the development of the information society, the digital divide and protection of the most vulnerable, the benefits and the misuse of social networks. Guests represent a wide range of social actors: senior government officials and specialized advisers, researchers and professional associations, foreign and domestic IT organizations, civil society groups, and media outlets. In addition to domestic guests, the show hosted international interlocutors on multiple occasions. The book *Digital Icons 20+*, derived from the interview transcripts of this broadcast, served as the basis for the development of *DigIko* corpus.

The motivation for combining TEI (Text Encoding Initiative) with NER (Named Entity Recognition) and NEL (Named Entity Linking) is to provide semantic enrichment as a foundation for RAG (Retrieval-Augmented Generation), where verifiable passages, linked to external resources, help curb hallucinations. NEL with Wikidata identifiers (QIDs) (Wikimedia, 2023) enables precise, multilingual entity control in both retrieval and prompting. TEI IDs and QIDs attached to each claim support replication, peer review, and integration with knowledge graphs.

The additional research utility is that entity and speaker-faceted retrieval yields stance profiles and temporal comparisons that plain summarization rarely affords. We implemented sequential transformation from raw text to semantically enriched corpus. The methodology for corpus preparation and exploitation is given in Section 3, while corpus editions are presented in Section 4. The analy-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1474–1484
11–16 May 2026.

sis and evaluation are presented in Section 5 and Section 7 brings limitations and future work.

The main contributions of this paper are: (i) a standards-based pipeline integrating various preprocessing tasks that turns interview transcripts into a semantically linked, structure-queryable resource, and (ii) grounded QA search over the corpus *DigIko* interviews collection, illustrating how evidence retrieval and LLM generation can support scholarly analysis. We argue that this workflow improves accessibility, supports semantic search and summarization, and offers a replicable path for building knowledge-centric interview corpora in Serbian and comparable languages. The repository with code and sample data will be freely accessible, while full corpus will be accessible for search.

2. Related Work

2.1. Interview-specific Models

Puren and Cafiero (2024) introduce *InTEIrviews*, an XML-TEI–conformant ODD[1] for qualitative interviews that promotes openness, reusability, and standardized annotation. An XML-TEI–conformant ODD selects TEI P5 modules and declares project-specific constraints while remaining compatible with TEI's element/class system. From a single ODD one can generate both validation schemas and human-readable documentation. Typical contents include a `<schemaSpec>` (module selection), `<elementSpec>` / `<classSpec>` (content models, attributes), and prose explaining encoding decisions and metadata. It builds entirely on existing TEI modules, chiefly Transcription of Speech and Language Corpora, and distinguishes macrostructure (headered metadata on participants/conditions; turns encoded with `<u who="">`; researcher notes in `<standOff>`) from microstructure (fine-grained linguistic and discourse features, named entities). The model also foregrounds ethics and privacy, recommending mechanisms such as `<gap>` (with `@resp, @unit, @quantity`) for redactions, and encourages controlled vocabularies (e.g., @ana on `<u>`) to make interviewer practices comparable. Overall, the ODD standardizes interview transcripts while preserving project-specific variation, enabling analysis, sharing, and FAIR-aligned reuse.

Davis (2015) surveys three practical approaches for aligning interview audio/video with transcripts: OHMS (Oral History Metadata Synchronizer), TEI-based workflows, and YouTube. It illustrates how each supports searchable, time-coded access to oral-history collections. It highlights using TEI structured markup to anchor utterances and timestamps, enables interoperable links between transcript segments and media.

2.2. TEI-Encoded Spoken Corpora

The CLARIN-conformant web services implement an end-to-end workflow for *TEI/ISO 24624* transcripts of spoken language, especially legacy interview corpora (Fisseni and Schmidt, 2019). The authors show how many tasks can operate directly on *ISO/TEI*, thereby keeping interoperability with common spoken-language tools while unlocking written-language NLP services. A central use case is the IDS Archive for Spoken German, where deposited interview data are curated to meet FAIR principles: full digitization, conversion to structured/standard formats (*ISO/TEI* for transcripts, CMDI for metadata), alignment between audio and text, linguistic enrichment, and integration into dissemination platforms.

The service pipeline has several building blocks: 1) *text2iso* that parse plain text transcripts into *ISO/TEI* with `<u>`, `<timeline>`, overlap handling, participants in `teiHeader`; 2) tokenization to `<w>` / `<pc>`, lift pauses to `<pause>`, preserve anchors; 3) per-utterance language detection (OpenNLP), annotate `xml:lang`; 4) OrthoNormal, like normalization (German), write `@norm` on `<w>`; 5) POS-tagging and lemmatization (TreeTagger/TT4J), language-appropriate models; 6) pseudo-alignment transcript-audio; 7) add/remove `xml:id` across elements.

The Slovene reference spoken corpus *GOS 2* is expansion of the original *GOS* from 2011 by integrating *GOS VideoLectures* (public academic speech) and the *Artur ASR* (Automatic Speech Recognition) database. The authentic, manually checked speech was selected and the divergent designs were reconciled via a unified metadata taxonomy. *TEI XML schema* (*CLARIN.SI* parameterization) was updated and transcription conventions were harmonized and annotated with CLASSLA-Stanza. A web concordancer offers simple/advanced search, metadata filtering, and links to other Slovene resources (Verdonik et al., 2024).

Parla-CLARIN is a *TEI* customization for modeling and exchanging parliamentary debate corpora, designed for interoperability and usability in multilingual research settings. Its recommendations define a consistent document structure (sittings, debates, speeches/utterances with `who` identifiers), rich headers with metadata (speakers, parties, roles, time and place), and facilities for linking to audio/video recordings and related legislative acts. The package includes schemas, validation guidance, and examples, and underpins projects

[1] ODD ("One Document Does-it-all") is the TEI XML-based specification language for defining, constraining, and documenting TEI customizations.

such as ParlaMint, thereby supporting FAIR access, comparability, and reuse.[2]

The British National Corpus (BNC) XML Edition is a classic, large-scale example of TEI–encoded spoken interaction. It demonstrates mature structural conventions for turn taking with `<u>` elements, speaker identification via `who` attributes (pointing to speaker registries), and consistent token/-markup practices at scale (e.g., `<w>` for words, `<pc>` for punctuation), all embedded in a rich `teiHeader` with participant and recording metadata. Because of its size and stability, the Spoken BNC offers robust precedents for utterance segmentation, speaker normalization, and metadata design that are readily transferable to interview corpora.[3]

3. Methodology

3.1. TEI for Spoken Interactions

Chapter 8 of the TEI P5 Guidelines: "Transcriptions of Speech" (TEI Consortium, 2025c) provides the primary specification for encoding interviews and other spoken data. It models discourse as utterances (`<u>`) organized within a standard TEI text structure, with speaker identification via `who` pointers to participant registries (e.g., `<listPerson>`). The chapter details how to encode pauses, nonlexical vocal events, kinesic incidents, overlaps, and subdivisions of discourse, and how to document recordings and transcription conventions in the `teiHeader` (e.g., `<transcriptionDesc>`, `<scriptStmt>`). For complex phenomena such as alignment and non-hierarchical structures, it points to TEI mechanisms for linking, segmentation, and timelines, enabling interoperable transcripts suitable for analysis and reuse.

Chapter 2 The TEI Header specifies how to document interview metadata in the `teiHeader`: bibliographic description (`fileDesc`), encoding and editorial practices (`encodingDesc`), participant/setting profiles (`profileDesc`: `listPerson`, roles, place/time), and versioning (`revisionDesc`). For interviews, this enables consistent recording of participants, recording circumstances, and provenance for citation and reuse. The guidelines Header/Profile pointers for transcribed speech point to what interview-specific facts to capture in the header/profile: participant identities and roles, interaction setting, recording and transcription conditions, and sampling, so that spoken data are discoverable, etc. (TEI Consortium, 2025b).

This structural layer segments discourse and provides context windows for downstream NLP as well as structure-aware querying (e.g., per speaker). TXM textometry, coupled with *JeRTeh*[4] resources (Stanković et al., 2020), supplies basic preprocessing, including sentence and other supplementary annotations.

3.2. Analysis with the TXM Tool

Beyond its role in enriching corpora with morphosyntactic layers, the TXM platform (Serge, 2020) serves as a versatile workbench for corpus exploration and quantitative reporting. It supports the computation of a wide range of text statistics (Krstev et al., 2019) and implements textometry, statistically grounded analysis of large text collections, providing methods to relate lexical distributions to metadata and structure (Heiden, 2010). The TXM tool combines search, inspection, and measurement within a single interface so that analysts can move seamlessly from evidence (concordances) to counts (frequencies) and onward to models (multivariate analyses), while keeping links back to the original passages.

Inspired by work (Stanković et al., 2022, 2024) on SrpELTeC corpus (Krstev et al., 2021), qualitative exploration of the corpus is centered on: (1) concordances powered by the CQP engine and its CQL syntax; (2) frequency lists computed over tokens, lemmas, POS, or structural tiers (including NER); (3) progression plots that visualize the distribution of word patterns across a text or corpus segment; and (4) an HTML-based reading view that is deeply cross-linked from other modules. Quantitative functionality builds on R backends to offer factorial correspondence analysis, clustering, specificities analysis, and collocation measures. Finally, TXM streamlines contrastive designs by letting users define subcorpora/partitions (by text type, time period, speaker, etc.) and then compare profiles across these slices using a consistent workflow (Heiden et al., 2015).

3.3. Entity Recognition and Linking

The NER and NEL models used are based on *TeslaXLM* (Škorić, 2025), Serbian and Serbo-Croatian language model built upon xlm-roberta-large. Respective classifiers were trained on top of the model using 73K manually curated sentences from an annotated dataset under continuous development. The annotation covers eight entity types: persons (`<PERS>`), places (`<LOC>`), organizations (`<ORG>`), and events (`<EVENT>`), roles and professions (`<ROLE>`), work of art (`<WORK>`), demonyms

[2] Parla-CLARIN: TEI schema and guidelines for parliamentary corpora, https://clarin-eric.github.io/parla-clarin/

[3] BNC XML Edition User Reference Guide: http://www.natcorp.ox.ac.uk/docs/URG.xml?ID=cdifsp.

[4] JeRTeh is Serbian Society for Language Resources and Technologies, https://jerteh.rs/

(<DEMO>), and products (<PRODUCT>).

After applying NER model, recognized spans were inserted into TEI corpus version. In the next step NEL model was used to link recognized entities with appropriate WikiData items, using context for disambiguation; where necessary, missing items are added to the knowledge base (KB). This step enables the normalization, meaning that different inflective forms referring to the same entity are linked to the same identifier (e.g., a WikiData QID). As a result, the corpus becomes anchored in a broader knowledge-graph ecosystem, facilitating data integration and reasoning.

We index *DigIko* corpus (Ranka Stanković and Tamara Vučenović, 2026) of TEI encoded interviews enriched with lemmatization, NER, and NEL (Wikidata QIDs). Each utterance (<u>), speaker turn, paragraph, and TEI division ( ) is treated as a retrievable unit with metadata: speaker ID, lemmas, entity spans, QIDs, and interview provenance. Indexing relies on tokenized TEI and canonical (lemmatized) text, TEI XPath/IDs, character offsets for entity spans, speaker/QID facets, and lemma and *n*-gram features. CQL queries that combine structural and lexical annotations are supported in TXM (desktop application) and on web instances of the NoSketch Engine and INCEpTION (Klie et al., 2018) on instance maintained by JeRTeh[5].

From the annotated corpus, we also generated a linked-data representation in NIF (NLP Interchange Format), following the approach used for the SrpELTeC corpus (Nešić et al., 2022), with additional speaker-level metadata. A build has been deployed to a local Apache Jena Fuseki (Apache Software Foundation, 2023) endpoint to enable retrieval and analysis via SPARQL [6].

Prompts are composed and optimized with *DSPy* (Declarative Self-improving Python) framework (Stanford NLP, 2025), which supports declarative task signatures, teleprompting, and modular policies for grounding and citation (Khattab et al., 2024). We experimented with *GPT-4.1* (OpenAI, 2024), *GPT-5* (OpenAI, 2025), and SM1 – *pilot5sumarizacija* (Škorić and Team, 2025) Serbian summarization model based on the Qwen3 architecture (800M parameters), trained on SD1 – *tesla/sumarizacija* dataset containing 30,000 pairs: (text, summary) (Škorić and Stanković, 2025).

In our DSPy pipeline, we define the signature to standardize faithful, style-controlled summarization of extracted question–answer pairs. The signature exposes two textual inputs: `question_in` and `answer_in`, which receive the raw question and answer, and two outputs: `question_out` and `answer_out`, which return concise Serbian summaries aligned with the source. A built–in instruction block enforces generation constraints crucial for evaluation and downstream use: neutral professional tone, avoidance of redundancy, omission of extraneous detail, and strict faithfulness. By declaring these fields and constraints at the signature level, DSPy enables reproducible prompting, clearer telemetry (inputs–outputs logged per call), and straightforward composition with retrieval modules, while ensuring that summarization quality is governed by explicit, testable criteria.

[5] http://inception3.jerteh.rs

[6] http://fuseki.jerteh.rs/#/dataset/digico

4. Many Faces of DigIko Corpus

The corpus DigIko - Digital Icons 20+ Interview Corpus is derived from the book *Digital Icons 20+*. The book organizes selected interviews from the radio broadcast into thematic sections that trace the evolution of digital society over the past 20 years. It serves as a substantive documentary record of how the digital revolution has reconfigured individual life, transformed contemporary culture, and influenced the policy environment for technology adoption in Serbia and beyond. Its aim is, above all, to underscore the importance of addressing Serbia's digitization/informatization, to map, through an authorial, "bird's-eye" perspective, the key moments, organizations, and individuals who have shaped the country's digital history, and to stimulate new, shared discussions with experts in IT and many other fields.

The eponymous book compiles interviews from 57 episodes featuring 67 guests. Featuring interlocutors who have actively shaped this trajectory, the book maps the shifting conditions in the Republic of Serbia while situating them within regional and global developments. The corpus *DigIko* offers curated interviews documenting the evolution of Serbia's digital ecosystem and can serve as a reference dataset for grounded, evidence-based NLP in a low-resourced language context.

4.1. DigIko TEI

The book was prepared in Microsoft Word using different custom styles, where sections of the text are precisely labeled: interview metadata, the introductory section presenting the discussion topic and interlocutors, the interviewer's questions and the interviewees' responses. Based on these styles, custom python script was prepared to to encode text transcripts with TEI XML tags, marking interview/session boundaries with , paragraph breaks (<p>), speaker turns with identifiers.

The folowing listing presents a snippet from one interview, rooted in `<div type="interview">` with administrative attributes (e.g., `@gender`, `@interviewid`, `@iyear`), followed by

a <DATE> block (<month>, <year>), a title (<title>), and participant labeling via <interviewee>. The discourse is partitioned into and . Utterances are encoded as <u> elements keyed to speakers with @who (and optionally addressees via @toWhom), with content organized as paragraphs <p> and fine-grained segments <seg> to support alignment, citation, and textometric analysis.

The tags related to the interview structure are shown in red, while NER tags are annotated in blue (<PRODUCT>, <PERS>, <WORK>, <DEMO>, <ORG>, <LOC>, <ROLE>).

```
<div type="interview" gender="M" interviewid="11"
    iyear="2021">
 <DATE><month>JUN</month><year>2021.</year></DATE>
 <title>Kada ljudima kažem čime se bavim, moram još
    pola sata da im objašnjavam šta je to</title>
 <interviewee>Sagovornik: <PERS>Voja Antonić</PERS>
    </interviewee>
  <div type="introduction">
   <p><seg>Kod prvih kompjutera to nije kao kod dana
    šnjih, oni su bili mnogo drugačiji od današnjih.
    </seg> [...]</p><p><seg>Sada već davne 1983.
    godine, u tadašnjoj <LOC>Jugoslaviji</LOC>,
    <PERS>Voja Antonić</PERS> je napravio i razvio
    <PRODUCT>Galaksiju</PRODUCT>, prvi domaći
    mikroračunar!</seg><seg> „Uradi sam" verzija
    ovog računara opisana je u specijalnom izdanju
    časopisa <WORK>Galaksija</WORK>, pod naslovom „
    <WORK>Računari u vašoj kući</WORK>".</seg><seg>
    <PERS>Voja</PERS> danas živi i radi u <LOC>
    Americi</LOC>, a u ekskluzivnom intervjuu za <
    ORG>Radio Beograd</ORG> [...]</seg></p>
  </div>
  <div type="conversation">
   <u who="#TaVu" toWhom="#VoAn"><iquestion><p><seg>
    Prvo moram da Vas pitam o <PRODUCT>Galaksiji</
    PRODUCT>.</seg></p></iquestion></u>
   <u who="#VoAn" toWhom="#TaVu"><ianswer>
    <p><seg>Jedna grupa mladih entuzijasta iz <LOC>
    Novog Sada</LOC> počela je jedan pravi
    dokumentarni film o tome, gde će to biti malo
    lepše ispričano nego što ću ja sada ispričati.<
    /seg><seg>To je bilo 1983. godine, ja sam bio
    tek oženjen, mlad.</seg><seg>Bio sam sa
    suprugom na letovanju u <LOC>Crnoj Gori</LOC>.<
    /seg><seg>Bili smo u <LOC>Risnu</LOC>, u <ORG>
    hotelu Teuta</ORG>.</seg>[...]</p></ianswer></u
    >
  <u who="#TaVu"><iquestion>
   <p><seg>Uskoro ste upoznali <PERS>Dejana
    Ristanovića</PERS> i, da se malo našalim,
    ostalo je istorija.</seg></p></iquestion></u>
  <u who="#VoAn" toWhom="#TaVu"><ianswer>
   <p><seg>Da, pošto je sve to radilo, pomislio sam
    da to negde objavim.</seg><seg>Prva ideja mi je
     bio zagrebački časopis <WORK>Sam svoj majstor<
    /WORK>, tada se zvao jednostavno <WORK>Sam</
    WORK>.</seg><seg>Ali s nekim ranijim projektima
     koje sam im slao nisam bio zadovoljan.</seg><
    seg>Oni su to priređivali i prevodili na <DEMO>
    hrvatski</DEMO> jezik.</seg><seg>Znate, ja vrlo
     poštujem <ROLE>novinare</ROLE></seg>[...]</p><
    /ianswer></u>
  </div>
```

For pragmatic marking, we wrap question/answer spans with project-specific <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 , (ii) required @who on <u>, and (iii) content models u → (p+) and p → (seg*), ensuring interoperable structure for NER/NEL offsets, speaker-level provenance, and downstream linked-data serialization.

DigIko-TEI was annotated with POS and lemma layers in TXM tool (Serge, 2020). The corpus contains 57 interviews, 1,334 utterances (665 questions and 669 answers, since few questions were addressed to many speakers), 8,887 sentences (segments) with 142,490 words. Distribution in number of words over different corpus partitions is given in Figure 1, showing size per year in histogram.

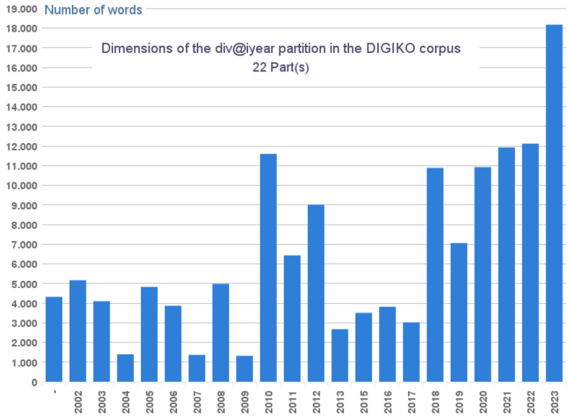

Figure 1: Corpus partition by years

The specificity graph for six technology-related terms across corpus 5-years periods is presented in Figure 2. Bars show term specificity (positive=above-expected usage; negative=below), with a banality threshold at -2. Early periods show moderate prominence of *internet* and *computer*, a mid-period rise for *data/information*, and a pronounced surge for *artificial intelligence* in 2021–2023 that far exceeds other terms.

Figure 3 presents specificity over time for frequent organisations: *Fejsbuk* [Facebook] (red), *Gugl* [Google] (blue) and *Tviter* [Twitter] (green), 2002–2023. Scores above the banality line ($+2$) mark salient periods. We observe an early surge for *Gugl* around 2008, a sharp peak for *Tviter* in 2012 (the highest value in the series), and elevated *Fejsbuk* around 2011–2013. Later years show a brief *Gugl* resurgence (2021) and mostly near-zero or negative values for *Tviter* and *Fejsbuk*, indicating reduced specificity outside the noted spikes.

4.2. DigIko Entity Linking

Figure 4 presents the local instance of INCEpTION used for manual correction of automatically annotated NER layer and NEL to Wikidata. Annotators label spans in running text and assign coarse types (listed in Subsection 3.3), then link each mention to a Wikidata item (e.g., *Srbija* Q403, *Jev-*

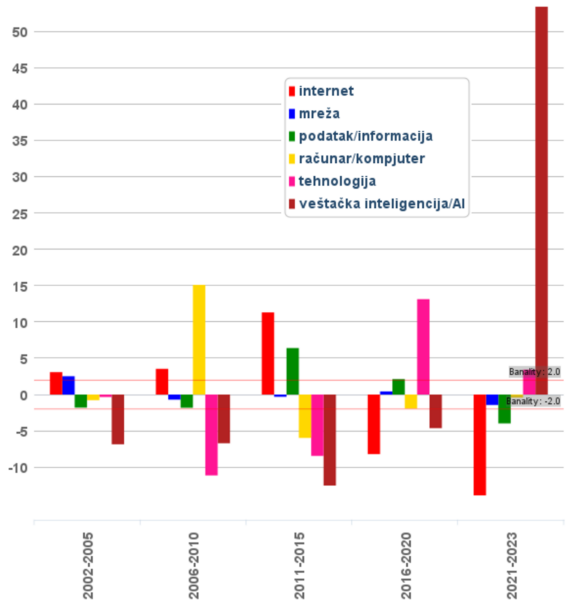

Figure 2: Specificity over time for selected key terms

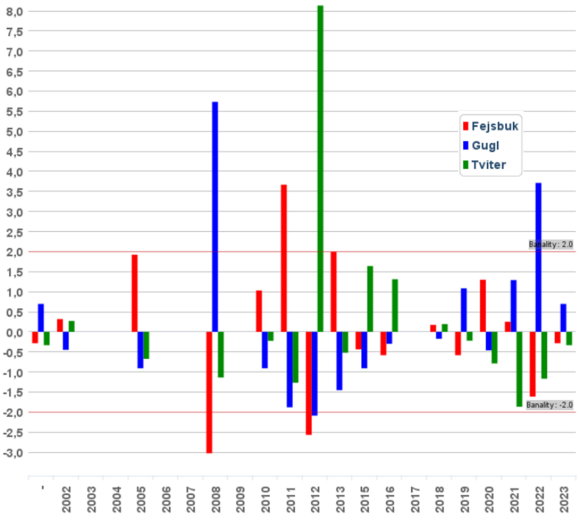

Figure 3: Specificity over time for frequent ORG

genij Kasperski `Q4864`). The left panel lists annotations grouped by position, the central panel shows color-coded spans, and KB pop-ups provide multilingual labels, aliases, and descriptions to aid disambiguation. The workflow stores QIDs with surface forms and roles, producing reproducible, knowledge-graph-ready annotations for downstream NER/NEL evaluation and knowledge graph (KG) construction.

To extract all mentioning of annotated locations query would be: `<Named_entity.value="LOC"/>`. To extract all mentioning of annotated exact item (for example `Q403` *Serbia*) from Wikidata query would be: `<Named_entity.identifier="http://www.wikidata.org/entity/Q403"/>`

Figure 5 presents distribution of annotated entity mentions by type. The inventory is dominated by roles and organizations, while events are comparatively rare. The most frequent locations are: Srbija [Serbia] (127), Beograd [Belgrade] (79), Amerika+SAD [USA] (34+18), Evropa [Europe] (32). The most frequent persons are: Stiv Džobs [Steve Jobs] (29), Nikola Tesla (27). Most frequent organisations are Fejsbuk [Facebook] (72) and Gugl [Google] (25). The most frequent roles (professions) are: umetnik [artist] (81), profesor [professor] (65), novinar [journalist] (48).

Top Wikidata items (QIDs) mentioned in the annotated corpus can be seen in Figure 6. The x–axis lists QIDs; the y–axis shows frequency (*Broj pojavljivanja*). The distribution is dominated by `Q403` *Serbia* (67). Next are `Q30` *USA* (29), `Q3711` *Belgrade* (25), and `Q458` *European Union / EU* (24). Mid–range items include `Q355` *Facebook* (17), `Q36704` *Yugoslavia/SFRJ* (16), and `Q46` *Europe* (13). Less frequent are `Q1709362` *UN* (7), `Q918` *X (Twitter)* (6), and `Q83286` *Socialist Federal Republic of Yugoslavia* (5). Overall, domestic geopolitical entities dominate, followed by major international actors and platforms.

For illustration of *DigIko* NIF edition, we will present a small part of Turtle (ttl) file. The main class `nif:String` represents strings of Unicode characters. The subclass of `nif:String` is `nif:Context`, that represents a text in its entirety and holds the characters of this text in the `nif:isString` property. In this particular scenario, it is evident that `itsrdf:taClassRef` is employed to connect with the relevant category of named entities, such as individuals, places, or organizations. When dealing with individuals (person), we included `wdt:Q5` from Wikidata, and `dbo:Person` from DBpedia with exact QID to the entity: `wd:Q1252236`.

```
<http://url/digiko.txt#char=5107,5119> a
 nif:RFC5147String, nif:String, nif:Word;
 nif:anchorOf "Voja Antonić";
 nif:beginIndex "5107";
 nif:endIndex "5119";
 ...
 itsrdf:taClassRef wd:Q5, dbo:Person ;
 itsrdf:taIdentRef wd:Q1252236 .
```

The summarization using DSPy on 661 QA pair with average length 189 characters for questions and 892 for answers produced with gpt-4.1 average 68 characters for answers and 214 for question (24% and 66% reduction). Reduction by gpt-5 was to 76 and 309 characters (36% and 53%) and for SM1 the reduction was to 61 and 94 characters (41% and 81%).

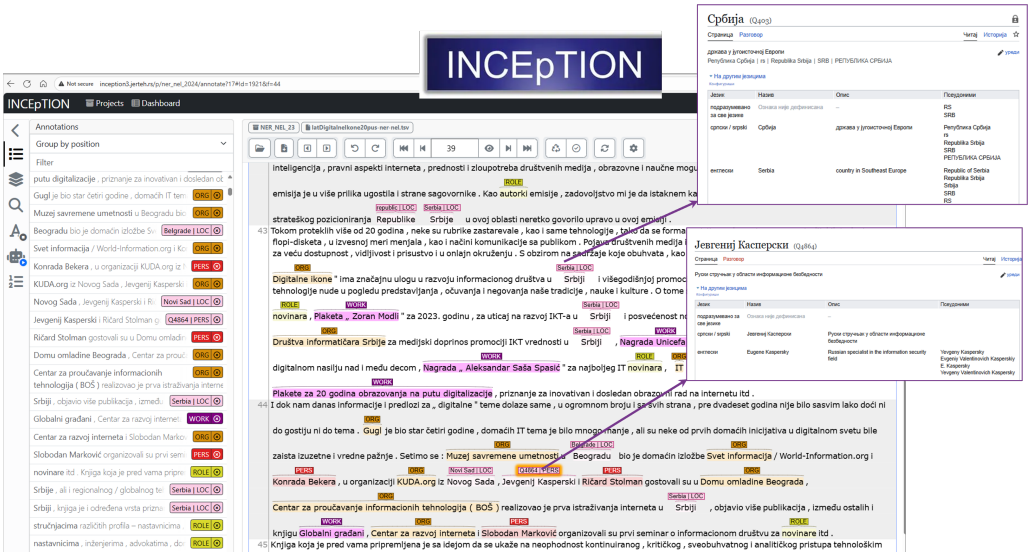

Figure 4: NER and NEL within INCEpTION

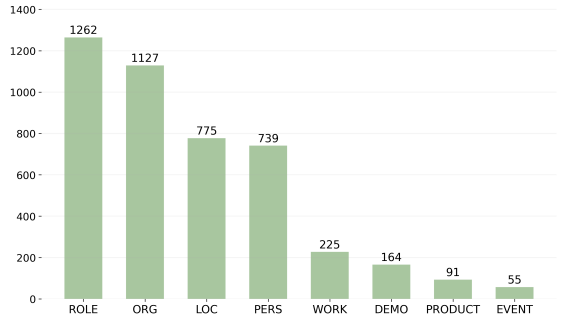

Figure 5: NER distribution over classes

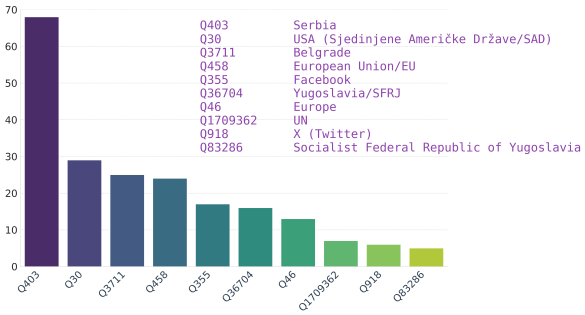

Figure 6: Top Wikidata items annotated in corpus

5. Evaluation

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

The evaluation for NER and NEL was performed on a subset of 1,000 sentences. In the class-wise evaluation, the model achieved the highest performance for PERS ($F_1 = 0.97$) and ROLE ($F_1 = 0.90$), followed by LOC ($F_1 = 0.88$), indicating strong recognition of persons, roles, and locations. Moderate scores were obtained for ORG ($F_1 = 0.66$) and WORK ($F_1 = 0.65$), reflecting confusion with semantically related categories and limited training diversity. The weakest results appeared for PRODUCT ($F_1 = 0.39$), driven by low recall (0.25) and a large number of missed instances, likely due to the rarity and variability of product names in the corpus. In total, NER achieved Precision: 0.84, Recall: 0.82 and F1 Score: 0.83. Overall, the system performs reliably for core entity types but would benefit from dataset balancing, domain-specific lexicons, refined annotation guidelines, and augmentation techniques to strengthen recognition of less frequent classes (Table 1). The evaluation

Class	TP	FP	FN	P	R	F1
PERS	155	6	3	.96	.98	.97
LOC	67	13	5	.84	.93	.88
ORG	66	49	18	.57	.79	.66
ROLE	167	12	26	.93	.87	.90
WORK	20	13	9	.61	.69	.65
PRODUCT	15	2	45	.88	.25	.39

Table 1: Evaluation results by entity class

for NEL was limited to locations, where system achieved a precision of 0.83, recall of 0.72, and an overall F_1-score of 0.77, based on 52 true positives (TP), 11 false positives (FP), and 20 false negatives (FN). These results indicate that the linker produces mostly correct associations (high precision), but misses a notable portion of true location men-

tions (lower recall). 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.

For both questions and answers longer than 200 characters, we evaluated summarization quality across three models: *GPT-4.1*, *GPT-5*, and *SM1* using ROUGE-L and BERTScore-F1 metrics (Table 2). ROUGE-L measures lexical overlap via the longest common subsequence between hypothesis and reference (Lin, 2004), while BERTScore-F1 captures semantic similarity using contextual embeddings (Zhang et al., 2019).

For question summarization, the GPT-5 model achieved the highest mean scores (ROUGE-L=0.23, BERTScore-F1=0.72), indicating improved lexical and semantic alignment with reference summaries. In comparison, *GPT-4.1* obtained a mean ROUGE-L of 0.16 and BERTScore-F1 of 0.69, suggesting more extractive behavior with slightly reduced semantic fidelity. The *SM1* model performed noticeably weaker (ROUGE-L=0.07, BERTScore-F1=0.63), generating shorter and more abstract summaries with limited overlap and coherence.

For answer-level summaries, the relative ranking remained consistent. *GPT-5* again outperformed other models with mean scores of ROUGE-L=0.27, BERTScore-F1=0.74, followed by *GPT-4.1* (ROUGE-L=0.16, BERTScore-F1=0.70) and *SM1* (ROUGE-L=0.05, BERTScore-F1=0.62). The narrower standard deviations for *GPT-5* (0.12 and 0.05) indicate higher consistency across samples, while its maximum BERTScore-F1 of 0.89 highlights strong semantic preservation. In contrast, *SM1* results show low variance but overall underperformance, reflecting limited training for long-form summarization and reduced factual retention.

Across both questions and answers, *GPT-5* consistently achieved the best balance between conciseness and semantic faithfulness, surpassing *GPT-4.1* in contextual abstraction and coherence. *GPT-4.1* remained reliable for extractive summaries with moderate overlap, whereas *SM1* tended to generate brief, semantically shallow outputs. These findings confirm that *GPT-5* is better suited for faithful abstractive summarization in Serbian, particularly for conversational and interview-style corpora. This also aligns with the impression gathered from the reading of generated summaries, where the superiority of *GPT-5* model was noticeable.

Although *SM1* underperformed, our analysis provided clear guidance for further improvements and precisely pinpointed where changes are most needed. Our goal is to build an open summarizer for Serbian, and these pilot evaluations offer actionable directions for iterative enhancements, perhaps via distillation of knowledge from larger, better-performing models through development of synthetic datasets.

Model	ROUGE-L	BERTScore-F1
GPT-5	0.23	0.72
GPT-4.1	0.16	0.69
SM1	0.07	0.63

Table 2: Summarization quality for questions and answers longer than 200 characters, evaluated with ROUGE-L and BERTScore-F1.

The proposed structure and multi-level annotation of the TEI corpus enable advanced functionalities such as semantic search, per-speaker retrieval, KG queries via SPARQL. Each TEI element is indexed through persistent pointers to allow direct cross-referencing within the corpus and external resources. During retrieval, speaker filters and Wikidata QID facets are applied to provide targeted contextual results. The retrieval component can be supported by multiple evidence views: speaker-centric digests that restrict evidence to a single speaker's utterances, entity-centric synopses that aggregate all passages mentioning a target Wikidata entity, and thematic overviews.

For factoid queries (e.g., *Who mentioned X and when?*), the retriever returns speaker– and turn–level evidence that the LLM cites explicitly. In addition to that, answers can be rewritten for non-experts while preserving evidence links to source utterances. The same mechanism enables concise bullet briefings or executive abstracts with controllable length and reading level, ensuring faithful simplification rather than paraphrase drift. The views can contrast how different speakers discuss the same entity or topic, aggregate cited spans into attribution maps over interview structure, and surface low-confidence claims or missing links (NEL gaps) for curator review and dataset improvement.

Figure 7 presents a graph where speakers (interviewees) of mentions linking locally extracted entities (Serbian labels) 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., *Voja Antonić* → *Jugoslavia*, *SAD*, *Beograd*).

6. Conclusion

We presented a pipeline that turns unstructured interview transcripts into a semantically enriched, queryable resource by integrating TEI encoding with NER/NEL to Wikidata, *textometry* (TXM), *linked–data* serialization (NIF), and *retrieval–augmented* summarization (DSPy+LLMs).

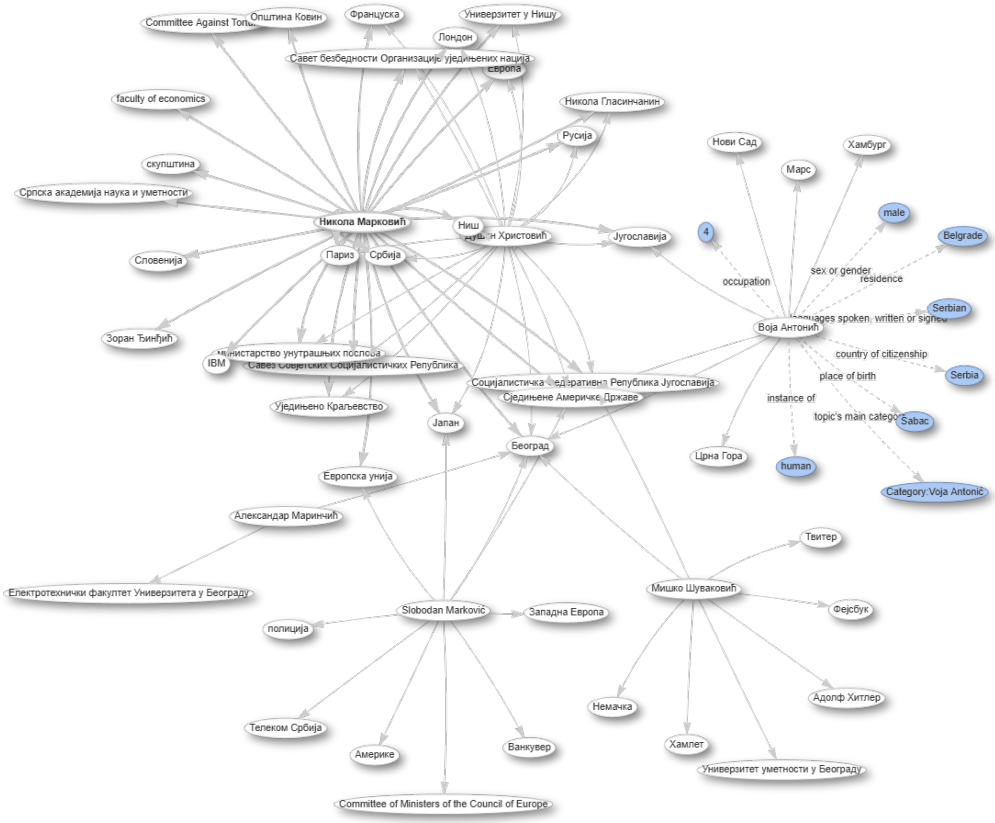

Figure 7: A graph: Who was mentioning who/what?

The TEI structure provides stable anchors for per–speaker retrieval, faceted search over entity QIDs, and reproducible citation. NEL normalizes mentions to canonical identifiers, enabling knowledge–graph queries (SPARQL) and interpretable, evidence–grounded exploration of the NIF version of *DigIko* corpus[7]. The NER model performs reliably on core types (PERS/ROLE/LOC), while ORG/WORK/PRODUCT remain challenging. NEL for locations achieves strong precision with room to improve recall. In summarization of questions and answers longer than 200 characters, *GPT–5* balances lexical faithfulness and semantic adequacy better than *GPT–4.1* and the Serbian baseline, confirming its suitability for faithful, concise abstracts in a low–resource setting. These pilot evaluations pinpoint concrete levers for improvement (dataset balancing, alias expansion for NEL), and they validate the end–to–end design for evidence–based access to cultural media.

Practically, the pipeline is *replicable*: TEI and ODD–driven constraints keep the format interoperable; TXM/NoSketch/INCEpTION support curation and analysis; NIF + Fuseki expose linked data; and DSPy enforces auditable prompting with explicit evidence fields. In future work, we will (i) extend the annotated dataset and release the NER+NEL models and TEI/NIF samples under permissive licenses, (ii) broaden NEL beyond locations with domain–adapted candidate generation and context constraints, (iii) refine RAG with per–speaker and entity–aware retrieval, attribution maps, and NIL/backoff strategies, and (iv) train and release an open Serbian summarizer improved by findings. We hope this resource and workflow will catalyze reproducible research on Serbian interviews and offer a transferable blueprint for other oral–history and media corpora.

7. Limitations and Future Work

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. The DigIko corpus is not big, but its expansion is envisaged. Summarization improvement is required, *SD1* will be expanded and *SM1* retrained.

Future activities will go towards RAG implementation offers a principled way to ground answers in large spoken corpora by retrieving evidence from TEI-encoded utterances and speaker metadata before generation (Gao et al., 2023; Izacard

[7] https://llod.jerteh.rs/digiko/

and Grave, 2021). In a TEI setting, each turn `<u>` is already anchored to `@who` (linked via `listPerson`), and rich setting descriptions in the `teiHeader` (TEI Consortium, 2025a). These structures enable granular chunking (per-utterance, per-speaker, per-session/interview) and dense indexing of both ASR-corrected transcripts and contextual facets (speaker role, topic, date, location), so that retrieval can return minimally sufficient evidence spans with stable pointers back to TEI IDs. When paired with citation-style prompting, the generator can surface inline references (e.g., `xml:id` of the `<u>` nodes) and short evidence snippets, improving verifiability while keeping the model's output concise and auditable (Lewis et al., 2020; Izacard and Grave, 2021).

8. Acknowledgements

This research was supported by the Science Fund of the Republic of Serbia: Text Embeddings – Serbian Language Applications–TESLA #7276, Ministry of Science, Technological Development and Innovation #451-03-34/2026-03/, and COST Action GOBLIN (CA23147) "Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs".

9. References

Robin Camille Davis. 2015. Synchronizing Oral History Text and Speech: a Tools Overview. *Behavioral & Social Sciences Librarian*, 34(4):234–238.

Bernhard Fisseni and Thomas Schmidt. 2019. 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*, pages 12–22. Linköping University Electronic Press.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Serge Heiden. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation*, volume 2, pages 389–-398, Japan.

Serge Heiden, Bénédicte Pincemin, and Matthieu Decorde. 2015. Manuel de TXM. Manuel d'utilisation du logiciel TXM.

Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the ACL: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, and et al. 2024. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Cvetana Krstev, Jelena Jaćimović, Branislava Šandrih, and Ranka Stanković. 2019. Analysis of the First Serbian Literature Corpus of the Late 19th and Early 20th Century with the TXM Platform. In *DH_BUDAPEST_2019*, pages 36–37. Centre for Digital Humanities - Eötvös Loránd University.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Karpukhin, et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81.

Milica Ikonić Nešić, Ranka Stanković, Christof Schöch, and Mihailo Škorić. 2022. 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 Language Resources and Evaluation Conference*, pages 7–16.

Marie Puren and Florian Cafiero. 2024. InTEIrviews: An ODD for Qualitative Interviews in the Humanities. *Journal of the Text Encoding Initiative*, 2024(15).

Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, Mihailo Škoric, and Milica Ikonić Nešić. 2022. Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3337–3345, Marseille, France. ELRA.

Ranka Stanković, Cvetana Krstev, and Duško Vitas. 2024. Srpeltec: A serbian literary corpus for distant reading. *Primerjalna književnost*, 47(2):45–63.

Ranka Stanković, Branislava Sandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In *LREC 2020-12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pages 3954–3962. ELRA (ELRA).

TEI Consortium. 2025a. Linking, segmentation, and alignment. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, version 4.10.2 edition. TEI Consortium. See §17.2.4 TEI XPointer Schemes (profile/pointer mechanisms relevant to transcribed speech).

TEI Consortium. 2025b. The TEI Header. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, version 4.10.2 edition. TEI Consortium. Last updated 4 Sept 2025.

TEI Consortium. 2025c. Transcriptions of Speech. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, version 4.10.2 edition. TEI Consortium. Last updated 4 Sept 2025.

Darinka Verdonik, Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić. 2024. Gos 2: A New Reference Corpus of Spoken Slovenian. In *Proceedings of LREC-COLING 2024*, pages 7825–7830.

Tamara Vučenović. 2024. *Digitalne ikone 20+ [Digital Icons 20+]*. Akademska knjiga; Radio Televizija Beograd - RTS.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*, page 43.

10. Language Resource References

Apache Software Foundation, Apache. 2023. *Apache Jena Fuseki*. The Apache Software Foundation, https://jena.apache.org/documentation/fuseki2/.

Klie, Jan-Christoph and Bugert, Michael and Boullosa, Beto and Eckart de Castilho, Richard and Gurevych, Iryna. 2018. *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation*. Association for Computational Linguistics, https://inception-project.github.io.

Cvetana Krstev and Branislava Šandrih Todorović and Ranka Stanković and Milica Ikonić Nešić. 2021. *SrpELTeC-gold - Named Entity Recognition Training Corpus for Serbian*. ELG, https://live.european-language-grid.eu/catalogue/corpus/9485, 1.0.

OpenAI. 2024. *GPT-4.1: OpenAI Large Language Model*. OpenAI. OpenAI. Accessed via OpenAI API on 2025-10-15. Model version: gpt-4.1.

OpenAI. 2025. *GPT-5: OpenAI Large Language Model*. OpenAI. OpenAI. Accessed via OpenAI API on 2025-10-15. Model version: gpt-5.

Ranka Stanković and Tamara Vučenović. 2026. *DigIko: Digital Icons 20+ – TEI Interview Corpus Annotated with Named Entities and Linked to Wikidata*. TESLA Project, Science Fund of the Republic of Serbia #7276. Hugging Face.

Heiden Serge. 2020. *The TXM Platform*. Equipe TXM, https://txm.gitpages.huma-num.fr/textometrie/.

Mihailo Škorić. 2025. *TeslaXLM: Multilingual Transformer for Serbian (te-sla/TeslaXLM)*. TESLA Project, Science Fund of the Republic of Serbia #7276. HugginFace. Hugging Face model card; base-model: XLM-R; license: CC-BY-SA-4.0.

Stanford NLP. 2025. *DSPy: The framework for programming—not prompting—language models*. Stanford NLP Group. Version 3.0.3, MIT License.

Wikimedia. 2023. *Wikidata*. Wikimedia, https://www.wikidata.org/.

Škorić, Mihailo and Stanković, Ranka. 2025. *TESLA/Sumarizacija: A Serbian summarization dataset*. Dataset hosted on Hugging Face Hub.

Škorić, Mihailo and TESLA Team. 2025. *PiloT5 – Serbian Summarization Model (te-sla/pilot5-sumarizacija)*. TESLA Project, Science Fund of the Republic of Serbia (#7276). Hugging Face Model Card. License: CC-BY-4.0.