Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data

Ranka Stanković
University of Belgrade, Serbia
ranka.stankovic@rgf.bg.ac.rs

Christian Chiarcos
University of Augsburg, Germany
christian.chiarcos@uni-a.de

Miloš Utvić
University of Belgrade, Serbia
misko@matf.bg.ac.rs

Olivera Kitanović
University of Belgrade, Serbia
olivera.kitanovic@rgf.bg.ac.rs

Abstract

This paper describes a case study on the generation of Linked Data text corpora using the NLP Interchange Format (NIF). The ELTEC corpus subset, which consists of 900 novels from the period 1840-1920 for 9 European languages, served as the basis for this research. The annotated version of the novels, in the so-called TEI level-2 format, was transformed into NIF, an RDF/OWL-based format that aims to achieve interoperability between NLP tools, language resources, and annotations. In this paper, we present our approach for transformation, and the implemented pipeline, and offer the code and results for similar use cases.

1 Introduction

Linguistic data science is a specialized area within the broader field of data science. It concentrates on the structured analysis and investigation of extensive data sets, employing various techniques and methodologies to extract valuable insights.[1] A crucial aspect of this field is the development of use cases that facilitate the integration of different language data types into a standardized ecosystem. This process utilizes tools and open standards established by the W3C to enable intelligent access, integration, and distribution of language data that caters to various user requirements. (Bosque-Gil et al., 2021)

Here, we illustrate the application of this approach to a subset of the ELTEC corpus (Burnard et al., 2021; Schöch et al., 2021; Stanković et al., 2022), which consists of 900 novels from the period 1840-1920 for 9 European languages. While working on the development of the ELTeC text collection, which includes numerous novels in many under-resourced languages, the concept of transforming the collection into linked data and adding it to the Linguistic Linked Open Data (LLOD) cloud was conceived. This would have the advantage of enhancing the exposure of under-resourced language data by linking it with other language resources already present in the LLOD cloud, thereby increasing its visibility.

The ELTeC core collection[2] has 12 corpora of 100 novels comparable in their internal structure. The ELTeC plus corpora take the total number of available full-text novels to 338 and ELTEC extension 547, with the ELTeC extensions, more than 2000 full-text novels are included in ELTeC. This research is focused on transformation and publishing a set of novels from ELTEC text collection from period 1840-1920 as open linked data according to best practice and guidelines fostered by CA18209 - European network for Web-centred linguistic data science (NexusLinguarum)[3].

The ELTeC novels format was developed within the COST Action CA16204 Distant Reading for European Literary History (D-Reading) (Burnard et al., 2021) in the so-called XML/TEI level-2[4]. Given the current lack of comparable corpus data in the LLOD cloud, they represent a particularly valuable resource for LLOD, as this technology allows not only interlinking different language versions, but potentially, also integrates dictionaries of the respective languages, prosopographical networks, geographical information, and other knowledge bases. The contribution is especially important since several low-resourced languages have ELTeC sub-collections with 100 novels. An overview of part of the ELTeC collection that was used in this case study will be presented in Section 1.3.

This paper will present a data model in Section 2.2 and approach for transformation from XML/TEI (Text Encoding Initiative) into NIF[5]

[1] This is the definition adopted by the Cost Action CA18209, *Nexus Linguarum - European network for Web-centred linguistic data science* (2019-2023), https://nexuslinguarum.eu/ (Declerck et al., 2020)

[2] https://www.distant-reading.net/eltec/

[3] https://nexuslinguarum.eu/

[4] https://distantreading.github.io/Schema/eltec-2.html

[5] http://bpmlod.github.io/report/nif-corpus/index.html (Un-

(NLP Interchange Format) in Section 2.3. The description of the results of transformation in the form of RDF graphs will be discussed in Section 3.1 and the examples of SPARQL query in Section 3.2. Discussion with open issues, dilemmas, difficulties, and constraints in research will be given in Section 4, followed by current results and plans for further activities in Section 5.

1.1 Motivation

In our research, results of literary scholars and the digital humanities community developed within the Cost Action D-Reading, are brought together with technologies for web-centered linguistic data science semantic networks developed in the Cost Action NexusLinguarum, fostering interdisciplinary research in these two areas. In the digital humanities community, the XML-based standards of the Text Encoding Initiative (TEI)[6] represent the prototypical approach to publishing electronic text and data. Yet, they have been criticized for not establishing a sufficient degree of interoperability, and their synchronization with formal semantics and web standards such as RDF and OWL have been repeatedly suggested since the 2000s (Bański, 2010; Ciotti and Tomasi, 2016). With the development of the Linguistic Linked Open Data (LLOD) cloud (Chiarcos et al., 2012; Pareja-Lora et al., 2019; Cimiano et al., 2020b), interest in formalizing this bridge has been intensified, albeit, so far, with a focus on lexical data (Bellandi, 2023).

ELTeC as a carefully selected and balanced text collection for each language, when available in LLOD could become a playground for various types of research in different scientific disciplines. The main contribution is a complex project which includes the preparation and publishing of 900 novels in LLOD. The developed procedure could be used for other ELTEC sub-collections and other XML/TEI corpora, and thus serve as a point of orientation for future publication workflows of multilingual corpus data on the web. This activity is directly related to the activities of Nexus Linguarum Working Group 1 'Linked-Data based language resources' that include creation, interlinking, enrichment, and evolution of the linguistic resources, especially in the context of a designated task of the action regarding the Development of the LLOD cloud for under-resourced languages and domains.

official Draft)

[6] https://tei-c.org/guidelines/

Motivation for this research was found in several previous successful use cases of transformation and publication using the NLP Interchange Format (NIF) (Hellmann et al., 2013), a community standard for representing the linguistic annotations of textual data in RDF, as produced by conventional NLP tools available at the time. It has been primarily designed for NLP web services but is also applicable for linguistically annotated corpora if their annotations do not exceed a certain level of complexity. Its primary goal has been to provide interoperable web services connecting NLP services, data, and applications and to build modular, flexible workflows on that basis (Hellmann et al., 2012; Cimiano et al., 2020c). NIF supports the annotation of named entities, part-of-speech tags, dependency parses, sentiment analysis, and other types of linguistic information. By its use of string URIs, NIF also supports multilingual text resources, enabling the representation of text in multiple languages and the alignment of annotations and translations across languages by means of RDF properties.

1.2 Related Research

Examples of electronically edited *text* in TEI and linked data complementing include the recent application of the Web Annotation standard to annotate TEI editions (del Rio Riande and Vitale, 2020). While such standoff annotation with JSON-LD is appropriate for *completed* editions, digital editions that are being worked on at the time of Linked Data annotation require a representation in *inline* XML, as demonstrated, by the experimental edition of a Middle French medical treatise (Tittel et al., 2018), as well as the Diachronic Spanish Sonnet Corpus: TEI and linked open data encoding, data distribution, and metrical findings (Ruiz Fabo et al., 2021). Aside from JSON-LD standoff and XML inline annotation with RDF, a third line of research on electronically edited text as Linked Data includes the full conversion of individual texts, structured corpora, and annotations. This is what is being pursued here. Normally, this line of research is conducted on data that follows conventions in the NLP and corpus linguistics communities rather than the DH communities, and here, tabular formats or, more recently, JSON have fully superseded the XML formats of the early 2000s. Cimiano et al. (2020a) presented prototypical applications of Linguistic Linked Data in Digital Humanities technologies and LOD resources in Digital Human-

ities as well as frequently used vocabularies. We see a special contribution to our work in discussing how to establish bridges between Linked Data technologies developed for NLP and TEI data produced and consumed in digital humanities.

Hellmann et al. (2010) and Brümmer (2015) described early experiments on the application of NIF to corpus data, and Brümmer et al. (2016) introduces the DBpedia Abstract Corpus - an open, large-scale corpus of annotated Wikipedia texts in six languages. The corpus contains over 11 million texts and more than 97 million entity links. The paper discusses the characteristics of the Wikipedia texts, the process of creating the corpus, its format, and interesting use cases, such as training and evaluating Named Entity Linking. NIF (Hellmann et al., 2013) was used as the corpus format to provide DBpedia compatibility using Linked Data as well as NLP tool interoperability. NIF is featured as a format for corpus data in the Best Practice Recommendations of the W3C Community Group Best Practices for Multilingual Linked Open Data (BP-MLOD).[7] As an illustration of the capacities of NIF, FrameNet (FN), an extensive lexical database for the English language has been published into RDF Linked Open Data (LOD) format, along with a vast corpus of text that has been annotated using FN. Alexiev and Casamayor (2016) examined the FN-LOD representation, compares it with NIF, and proposes an approach for the integration of FN into NIF that does not require any custom classes or properties.

Another widely used standard for linguistic annotations in RDF is Web Annotation (Sanderson et al., 2013) (formerly known as Open Annotation), published as a W3C standard (recommendation) in 2017[8]. Unlike NIF, however, it does not provide specific data structures for linguistic annotation, but only formalizes markables ('annotation targets') and information they are annotated with ('annotation bodies') in a reified annotation property. As Web Annotation does not provide specifically linguistic annotation, we focus on NIF-based vocabularies, here.

Yet another RDF-based corpus format is POWLA (Chiarcos, 2012), a reconstruction of the Linguistic Annotation Framework (Ide and Suderman, 2014, LAF, ISO 24612:2012) in OWL2/DL. As a proprietary standard, however, LAF seems not to be used much in the field, so the current role of POWLA seems to be primarily that of a companion vocabulary that serves to augment shallow data models such as NIF or Web Annotation data with generic data structures for linguistic annotation (Cimiano et al., 2020d). We are not aware of any corpus or annotation projects using POWLA independently of either NIF or Web Annotation since de Araujo et al. (2017), and for the rather shallow annotations of the ELTeC data, core NIF data structures are sufficient so we decided to focus on NIF.

Other RDF-based corpus formalisms we are aware of are either limited to a specific technology or software, e.g., the NewsReader Annotation Format (Fokkens et al., 2014, NAF-RDF), or the LAPPS Interchange Format (Ide et al., 2016, LIF), or they are focusing on a particular user community and their specific needs, e.g., the compatibility with tabular ('CoNLL') formats as used in NLP (Chiarcos and Glaser, 2020, CoNLL-RDF)or on the representation of interlinear glossed text (IGT) as used in language documentation, language teaching, and linguistic typology (Ionov, 2021, Ligt). CoNLL-RDF is based on a reduced core vocabulary taken from NIF, but it introduces its own URI schema, based on the counting of tokens and sentences. Unlike NIF, CoNLL-RDF URIs thus do not directly refer to a document, but only to a unit of annotation. Furthermore, CoNLL-RDF is more specialized in the annotation of syntax and semantics, whose treatment in NIF requires NIF extensions, whereas here, we focus on matters well covered by NIF, morphosyntactic annotation and named entities. Nevertheless, a future direction of our research is to compare NIF and CoNLL-RDF editions of our data with respect to verbosity and scalability issues.

In a recent overview of these and related vocabularies, Cimiano et al. (2020b) described the principles for annotating text data using RDF-compliant formalism, that are providing the basis for making annotated corporate and text collections accessible from the LLOD ecosystem. Because web documents may change, to preserve interpretability, it is recommended to include the full text of the annotated document in the RDF data.

Based on our literature overview and the char-

[7] However, these have not progressed beyond the level of a draft, available under http://bpmlod.github.io/report/nif-corpus/index.html, cf. https://www.w3.org/community/bpmlod/.

[8] https://www.w3.org/TR/annotation-model/

acteristics of our data, we decided to follow the BPMLOD draft recommendation and apply NIF 2.0 to our data. In the light of the alternatives, this offers a number of advantages:

- NIF is widely used (about as much as Web Annotation or CoNLL-RDF, but much more than tool- or community-specific RDF vocabularies or than generic formats such as LAF/POWLA).
- NIF provides explicit, native data structures for linguistic annotation (unlike Web Annotation).
- For the current annotations of the ELTeC corpus (morphosyntax, named entities), the native NIF vocabulary is sufficient. Additional data structures that could also account for morphological segmentation (as in Ligt), dependency syntax, and semantic role labeling (as in CoNLL-RDF) or generic linguistic annotations (as in POWLA) are not required.
- NIF is designed for standoff annotation, i.e., it uses string URIs to point to documents provided in their native formats on the web. Web Annotation is similar in this regard, but both are different from designated data models for linguistically annotated corpora whose basic unit of analysis is not the (primary text in the) document, but units of annotations imposed over these (e.g., CoNLL-RDF, Ligt). As an example, NIF URIs directly resolve against an offset in the annotated document, whereas CoNLL-RDF URIs are generated from sentence ID and token number, i.e., they require pre-processed documents.

For this reason, we eventually went with the NIF vocabulary for data modeling. It is to be noted though, that NIF has a number of potential downsides, including a high degree of verbosity (in comparison to tool- or domain-specific formats as well as to tabular formats as currently used in NLP – but probably less than or comparable to traditional XML-based formats such as LAF), so that one of the research questions we aim to contribute to is the discussion of scalability issues for such kind of data. Also, we would like to contribute to an effort of comparing and harmonizing data models for linguistic annotations on the web that has been initiated in 2020 in the context of the W3C Community Group Linked Data for Language Technology (LD4LT).[9] To the best of our knowledge, progress in this working group is slow. On the one hand, this can be attributed to external factors such as the involvement of many contributors in the development of a lexical companion vocabulary for corpus data, OntoLex-FrAC (Chiarcos et al., 2022a), which is in the process of finalization and which is expected to provide important stimuli for the discussion of annotations in LD4LT. On the other hand – and probably, more importantly –, the LLOD cloud diagram[10] currently suffers from a lack of corpus data, to begin with, so only limited data is available that can serve as a basis for comparison and benchmarking to evaluate or demonstrate the potential of LLOD technologies for corpus data. With the data set produced as a result of our efforts, such a dataset becomes available for the first time. As this is a relatively large-scale, annotated parallel corpus, it allows to both explore the potential of RDF technology for cross-lingual linking, as well as for the linking of corpora with annotations or, prospectively, lexical resources – for which the application of LLOD technologies is by far more established, and for which tremendous amounts of data are available (Gracia et al., 2018).

The field of literature and the Semantic Web encompasses various research areas and applications where semantic technologies are applied to enhance the understanding, analysis, and organization of literary works. While the intersection of literature and the Semantic Web is relatively new, several notable works have explored this interdisciplinary domain. These works represent a fraction of the research carried out at the intersection of literature and the Semantic Web. The field continues to evolve, and ongoing studies explore novel ways to leverage semantic technologies for improved understanding, analysis, and accessibility of literary works.

The specific research questions that can be explored when transforming TEI literary corpus into a linked NIF corpus: RQ1) What are the challenges and potential improvements for named entities to be recognized and linked to external resources in the NIF corpus? RQ2) How annotations, such as part-of-speech tags and lemma, should be represented for the literary works in the linked NIF corpus? RQ3) How effectively does the linking of enti-

[9] https://www.w3.org/community/ld4lt/wiki/LD4LT_Annotaton_Workshop_Zaragoza_2021.

[10] http://linguistic-lod.org/.

ties in the NIF corpus contribute to the enrichment and integration of the literary works with other linked data sources, such as DBpedia, Wikidata, or other semantic web datasets?

1.3 ELTeC collection

ELTeC is a multilingual collection of roughly comparable corpora each containing 100 novels from a given national (or rather: language-based) literary tradition (Schöch et al., 2021). The multiple encoding levels are defined in the ELTeC scheme: at level zero, only the bare minimum of markup defined above is permitted, while at level 1 a slightly richer (though still minimalist) encoding is defined. At level 2, additional tags are introduced to support linguistic processing of various kinds, as discussed further below. (Burnard et al., 2021).

The current version comprises 10 languages: German (deu), English (eng), French (fra), Hungarian (hun), Polish (pol), Portuguese (por), Romanian (rom), Slovenian (slv), Spanish (spa), Serbian (srp), with level-2 annotations for 100 novels per language. Further in the paper ISO 639-2:1998 Codes for the representation of names of languages — Part 2: Alpha-3 code[11] will be used.

The obligatory annotations for ELTeC TEI level-2 are POS tags and lemma, but some of them have also NER (named entity recognition) layer and some of them have detailed grammatical descriptions for tokens. All annotated novels are publicly available and published as XML/TEI files under CC-BY license. Input data collection with novels in XML/TEI level-2 is available in the following repositories: https://github.com/COST-ELTeC/ELTeC-lng/tree/master/level2 where "lng" is substituted with 3-letter code for language.

All language sub-collections are annotated with Universal Dependencies POS tag set and lemmatized. All, except French, have sentence boundaries marked with <s> XML element. NER tag sets do not have the same number of categories for different languages: most frequently used are PERS (person), ORG (organization), and LOC (location), but few also have DEMO (demonym, name of kinds of people: national, regional, political e.g. Frenchwoman, German, Parisians), ROLE (names of the profession, but also titles, nobility, office, military), WORK (titles of books, songs, plays, newspaper, paintings, sculptures, and other creations), EVENT (important events e.g. Christmas, Victory Day). Some text collections (srp, slv, por) have unique IDs for paragraphs, sentences, and tokens, while others are without identifiers.

Metadata from 700 novels, named WikiELTeC is available in Wikidata. WikiELTeC was semi-automatically populated from TeiHeader using OpenRefine, QuickStatents, and custom-made procedures (Ikonić Nešić et al., 2022). Each item for a novel is connected with an appropriate item that is an instance of electronic edition (`Q59466853`), first edition (`Q10898227`), print edition (`Q59466300`), and digital edition (`Q1224889`) using property (`P747`) (has edition or translation), and every item of edition must be connected with a corresponding item for a novel with inverse property (`P629`) (edition or translation of). The list of all properties used for novels in Wikidata is documented in WikiProject_ELTeC[12].

2 Methods

2.1 Standards for linguistic annotation

There are two prominent RDF standards for linguistic annotation: NLP Interchange Format (NIF) and Web Annotation. Both standards use URIs (or IRIs) for addressing corpora, which coincides with the use of URIs in other formats such as TEI and XML standoff formats. However, these standards are relatively technical and not particularly user-friendly, and there is a need for clearer documentation that provides guidelines (GL's) and best practices (BP's) for implementation. Apart from NIF standards, two resources were used: 'Best Practices for Multilingual Linked Open Data' (BPMLOD) W3C community group, and the output of the LIDER project[13]

NIF is a community standard developed in a series of research projects at the AKSW Leipzig, Germany, and still maintained by that group. A typical UR/IRI consists of two main components, a base name that serves to locate the document, and an optional fragment identifier. For numerous media types and different file formats, different fragment identifiers have been defined, often as best practices (BPs; also referred to as Requests for Comments, RFCs) of the Internet Engineering Task Force (IETF).

Khan et al. (2022a) report that this is one area where there is a real necessity for documentation that provides clear GL's and BP's. The presented research could be a showcase for the use of NIF

[11] https://www.iso.org/standard/4767.html

[12] https://www.wikidata.org/wiki/Wikidata:WikiProject_ELTeC

[13] https://lider-project.eu

and the transformation of TEI-compliant corpora to NIF. This paper contributes to this effort by providing a case study on NIF as an RDF-based format for describing strings in the novel, relaying on the classes and properties that are formally defined within the NIF Core Ontology 2.0[14]. The reason not to use the latest version 2.1 of NIF Ontology is the lack of full documentation, but some features introduced in 2.1 version will be discussed.

2.2 ELTEC-NIF data model

An overview of the linguistic annotation of corpora by NLP tools in a way that integrates Semantic Web standards and technologies is given in (Khan et al., 2022b), focusing on NIF and Web annotation. For this case study we selected the NLP Interchange Format (NIF), designed to facilitate the integration of NLP tools in knowledge extraction pipelines, as part of the building of a Semantic Web toolchain and a technology stack for language technology on the web. NIF provides support for part-of-speech tagging, lemmatization, and entity annotation, enabling ELTeC level-2 layers transformation.

The first version of ELTeC novels excerpts in NIF format is produced using the INCEPTION tool (Klie et al., 2018). TTL files are available in JeRTeh (Serbian Society for Language Resources and Technologies) web portal[15]. Several changes were introduced, mostly related to named entities and metadata linking. Selected metadata from WikiELTeC (Ikonić Nešić et al., 2022) is linked with novel content triples. Figure 1 presents an outline of the model for ELTeC-NIF.

For named entities, several ontologies were consulted. From OLIA[16] were user equivalents: `olia:Person`, `olia:Space`, `olia:Organization`, `olia:Event`. To link with DBpedia, dbo[17] namespace is introduced, and for Wikidata wd[18]. To link the type of recognized named entities are used following classes: `dbo:Person` = `wd:Q5`, `dbo:Place` = `wd:Q7884789`, `dbo:Organisation` = `wd:Q43229`, `dbo:Event` = `wd:Q1656682`, `dbo:Profession` = `wd:Q28640`, DEMO = `dbo:demonym` = `wd:Q217438`, `dbo:Work`= `wd:Q386724`. The recognized named entities are not linked with Wikidata or DBpedia items, they are just marked and classified in one of seven predefined types.

The presented research connects the previous results from the fields of Digital Humanities (Burnard et al., 2021; Schöch et al., 2021; Ikonić Nešić et al., 2022; Krstev, 2021; Stanković et al., 2022) and Linked Data (Hellmann et al., 2012; Brümmer, 2015; Alexiev and Casamayor, 2016; Cimiano et al., 2020c) which are traditionally considered separate areas of research. TEI is a widely used standard for encoding and representing textual data, while Linked Data focuses on interlinking and integrating diverse datasets. By bridging these two areas, the paper contributes to the integration of TEI-encoded literary resources with the broader Linked Data ecosystem.

2.3 Transformation procedure

A collab notebook was prepared for the transformation of XML/TEI into NIF. For Wikidata management *mkwikidata*[19] library was used for working with RDF *rdflib*. The code is available as a Python notebook in the GitHub repository TEI2NIF[20]. Code comprises classes: *Novel, Sentence, Token, NamedEntity* for appropriate transformation and set of additional functions.

For each novel in selected language in the set: $Lngs = \{deu, eng, fra, hun, pol, por, rom, slv, spa, srp\}$ the graph is created. Main function *write_gnovel* instantiate Graph with the following namespaces: *itsrdf, nif, olia, dc, dct, ms, wd, wdt, dbo, eltec*. After the instantiation of *Novel*, initial triples for the novel are added.

The parsing through selected XML/TEI level-2 version of the novel comprises several parts for generating triples: 1) novels metadata 2) sentences 3) named entities, and 4) words/tokens.

3 Results

3.1 NIF Terse RDF Triple Language (ttl)

From ELTeC level–2 described in Section 1.3, 900 novels from 9 language sub-collections with 100 *ttl* files were published. The number of sentences is limited to 1000 per novel in this edition. For the Serbian additional option, the dataset was prepared without a sentence limit.

[14] https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html
[15] http://llod.jerteh.rs/ELTEC/srp/NIF-INCEPTION/
[16] http://purl.org/olia/discourse/olia_discourse.owl
[17] https://dbpedia.org/ontology/
[18] https://www.wikidata.org/wiki/
[19] https://pypi.org/project/mkwikidata/
[20] https://github.com/rankastankovic/TEI2NIF

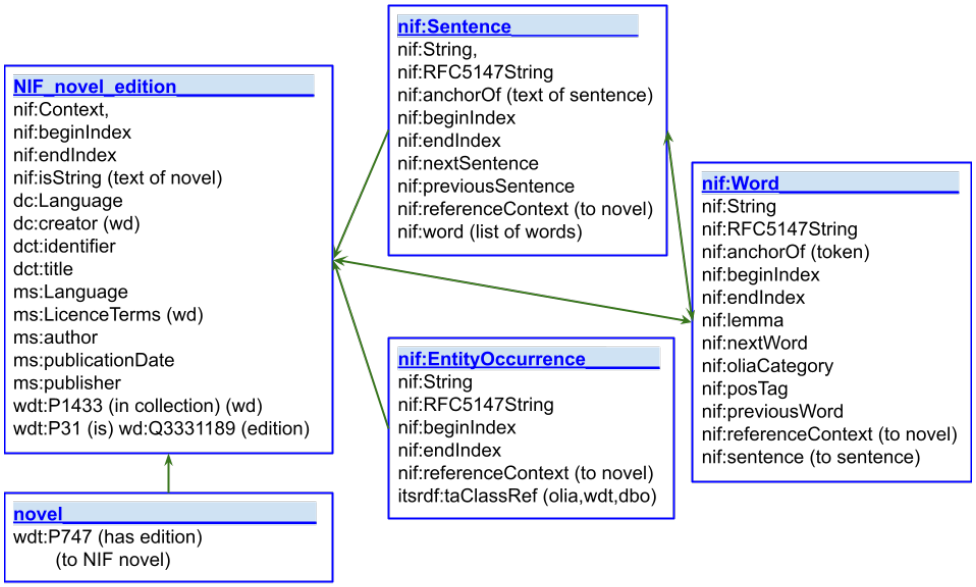

Figure 1: Data model for novel in ELTeC corpus .

Uncompressed files are accessible at: $http://llod.jerteh.rs/ELTEC/lng/NIF/$, where $lng \in Lngs$, with Creative Commons Attribution 4.0 International license. Zipped files are available also: http://llod.jerteh.rs/ELTEC/lng/NIF-lng-1000.zip, where $lng \in Lngs$ and they will be available on European Language Grid portal and other language repositories.

The core classes `nif:String` is used for the novel's content itself, described by `nif:beginIndex` and `nif:endIndex` . Dublin Core vocabulary is used for predicates related to the language, author, identifier, and title. The author is linked with Wikidata items for example Emili Bronthe is represented by (`wd:Q80137`). The novel "Wuthering Heights" (`wd:Q202975`) with the property *has edition or translation* (`wdt:P747`) is linked to digital version of the novel (`eltec:ENG18471.txt`) used as a source for NIF version. Further on, novel `eltec:ENG18471.txt` is linked by property *is published in* (`wdt:P1433`) with "engELTeC: English Literary Text Collection (ELTeC)" (wd:Q111271624). META–SHARE ontology[21] is used to describe language, licence terms, author, publisher, and publication year:

```
wd:Q202975 wdt:P747 eltec:ENG18471.txt.
eltec:ENG18471.txt a nif:Context,
    nif:String, nif:RFC5147String;
  nif:beginIndex "0";
  nif:endIndex "98583";
  nif:isString "Wuthering Heights A novel
  , By Ellis Bell , ... and Mr. Hindley
  will have to proceed to extremities
  , see if he wont .";
  dc:Language "en";
  dc:creator wd:Q80137 ;
  dct:identifier "ENG18471";
  dct:title "Wuthering Heights :
  ELTeC edition"^^xsd:string ;
  ms:Language "en"^^xsd:string ;
  ms:LicenceTerms wd:Q20007257 ;
  ms:author "Bronte, Emily (1818-1848)";
  ms:publisher "COST Action \"Distant
  Reading for European Literary
  History\" (CA16204)";
  ms:publicationDate "1847";
  wdt:P1433 wd:Q111271624;
  wdt:P31 wd:Q3331189.
```

For illustration, a short sentence "This is certainly , a beautiful country !" from "Wuthering Heights" Emily Brontë (1847) is presented and illustrative parts will be discussed. Substring of the `nif:Context` can be: a single word, sentence, or named entity that is linked to the relevant Context resource via `nif:referenceContext`. Beginning and end indices refer to the string content (sentence) represented by the context. The previous and next sentences are references as well as a list of words.

```
<http://llod.jerteh.rs/ELTEC/eng/NIF/
    ENG18471.txt#char=285,327>
  a nif:String, nif:RFC5147String,
    nif:Sentence ;
  nif:anchorOf "This is certainly ,
  a beautiful country ! " ;
  nif:beginIndex "285" ;
  nif:endIndex "327" ;
```

[21] http://w3id.org/meta-share/meta-share/2.0.0

```
  nif:nextSentence
  <http://llod.jerteh.rs/ELTEC/eng/NIF/
  ENG18471.txt#char=328,450>;
  nif:previousSentence
  <http://llod.jerteh.rs/ELTEC/eng/NIF/
  ENG18471.txt#char=169,284>;
  nif:referenceContext eltec:ENG18471.txt ;
  nif:word <http://llod.jerteh.rs/ELTEC/
    eng/NIF/ENG18471.txt#char=285,289>,
    <http://llod.jerteh.rs/ELTEC/eng/NIF/
    ENG18471.txt#char=290,292>,
        ...
    <http://llod.jerteh.rs/ELTEC/eng/NIF/
    ENG18471.txt#char=325,326> .
```

Following listing presents triplets for tokens (words). The property `nif:anchorOf` is used to explicate the annotated string. Apart from indices, `nif:lemma` and `nif:posTag` are included, `nif:previousWord` and `nif:nextWord`, `nif:sentence` and `nif:refeneceContext`.

```
<http://llod.jerteh.rs/ELTEC/eng/NIF/
    ENG18471.txt#char=307,316> a
  nif:String, nif:RFC5147String, nif:Word;
  nif:anchorOf "beautiful" ;
  nif:beginIndex "307" ;
  nif:endIndex "316" ;
  nif:lemma "beautiful" ;
  nif:nextWord
  <http://llod.jerteh.rs/ELTEC/eng/NIF/
  ENG18471.txt#char=317,324>;
  nif:posTag "ADJ" ;
  nif:oliaCategory  olia:Adjective ;
  nif:previousWord
  <http://llod.jerteh.rs/ELTEC/eng/NIF/
  ENG18471.txt#char=305,306>;
  nif:referenceContext eltec:ENG18471.txt;
  nif:sentence
  <http://llod.jerteh.rs/ELTEC/eng/NIF/
  ENG18471.txt#char=285,327>.
```

Since the English corpus has not NER layer annotated, example is taken from the Portuguese corpus. One can see that `itsrdf:taClassRef` is used to link to the appropriate type on NER, in this case for person: `olia:Person`, `wdt:Q5`, `dbo:Person`:

```
<http://llod.jerteh.rs/ELTEC/por/NIF/
POR0100.txt#char=78337,78365> a
  nif:Phrase,  nif:String,
  nif:RFC5147String ; nif:anchorOf
    "D. Diogo Furtado de Mendonça";
  nif:beginIndex "78337";
  nif:endIndex "78365";
  nif:referenceContext eltec:POR0100.txt;
  itsrdf:taClassRef
    olia:Person, wd:Q5, dbo:Person .
```

Total size of the repository for all nine languages is 12.87 GB, which includes 900 txt files, 900 ttl and 900 zip files. Table 1 gives an overview per language. The calculation in Fuseki database is calculated for the Serbian corpus. The database has

Language	zip (MB)	txt+ttl (GB)
deu	118	1.6
eng	106	1.42
hun	103	1.4
pol	90	1.12
por	96	1.23
rom	78	1.01
slv	100	1.32
spa	124	1.64
srp	95	1.22

Table 1: Size of corpus file repositories.

20.7 GB (17 times more than the files in the repository). There are 21,416,099 triples, 99012 sentences, 1.731.440 words, 32625 persons (`wd:Q5`), 5937 places (`wd:Q7884789`) etc.

3.2 SPARQL Endpoint

Apache Jena Fuseki[22] is used for uploading and testing Serbian ELTeC corpus (Krstev, 2021; Stanković et al., 2022) tranformed to NIF, as a SPARQL server web application at JeRTeh site[23]. Fuseki provides the SPARQL 1.1 protocols for query and update as well as the SPARQL Graph Store protocol. It is integrated with TDB (component of Jena for RDF storage) to provide a robust, transactional persistent storage layer, and incorporates Jena text query.

Six most frequent nouns in a novels of writer Jakov Ignjatović (`wd:Q570913`): kuća (house) 275, otac (father) 208, dan (day) 144, mati (mother) 140, godina (year) 127, ruka (hand) 123 can be found with following SPARQL query:

```
SELECT ?lemma (COUNT(?lemma) AS ?count)
   WHERE {
    ?subject nif:lemma ?lemma ;
       nif:posTag "NOUN"^^xsd:string;
       nif:referenceContext ?novelid.
    # Jakov Ignjatović
    ?novelid dc:creator wd:Q570913.
   }
GROUP BY   ?lemma
ORDER BY desc(?count)
```

List of recognised named entities linked with entity types in Wikidata can be retrieved with following query :

```
SELECT ?subject ?nentity  ?etype
WHERE {
    ?subject nif:anchorOf ?nentity ;
      itsrdf:taClassRef ?etype.
  FILTER (isURI(?etype) &&
  contains(str(?etype), ("wiki") ) )
}
```

[22] https://jena.apache.org/documentation/fuseki2/

[23] http://fuseki.jerteh.rs/#/dataset/SrpELTeC/query

The total numbers of recognised named entities grouped by type from Wikidata: person (Q5) 32625, name for a geographical entity or location (Q7884789) 5937, role - profession (Q28640) 24287, demonyms - name for a resident of a locality (Q217438) 5387, organization (Q43229) 451, events (Q1656682) 267, individual intellectual or artistic work (Q386724) 129, are retrieved with following query:

```
SELECT  ?etype (COUNT(?etype) AS ?count)
WHERE {
    ?subject nif:anchorOf ?nentity ;
      itsrdf:taClassRef ?etype.
  FILTER (isURI(?etype) &&
  contains(str(?etype), ("wiki") ) )
}
group by ?etype
```

4 Discussion

The primary issue at hand concerned which version of NIF to use - 2.0 or 2.1. Although version 2.1 offered some additional features that could have been advantageous for our case study, such as detecting information and subsequently linking entities, we opted for version 2.0. This was because, to the best of our knowledge, version 2.1 was only a release candidate and lacked comprehensive documentation. The service introduced two NIF substring resources that had the potential to be named entities. Each of these substring resources contained multiple pieces of annotation information:

- Indicating that a particular substring had been identified as a probable reference to a named entity. In NIF 2.1, this was achieved by assigning the `nif:EntityOccurrence` class to the substring resource.
- Providing potential references to Linked Data identifiers for the mentioned named entities, as well as classifying or referencing the entities into one or more categories. To reference these entities, we used the `itsrdf:taIdentRef` property from ITSRDF.

The dilemma related to NER was also mapping od NER types to appropriate ontology and choosing the best-fitting ontology class. We already mentioned that tagsets for NER classes are not the same for all languages and each language used specific tools and models. The general suggestion was to use 7 classes, that are mapped in our approach but some were used less and some more. For example, the Polish corpus is annotated with a very detailed tagset including *MISC, nam_adj_country, nam_fac_road, nam_fac_square, nam_liv_god, nam_liv_person, nam_loc_country_region, nam_loc_gpe_city, nam_loc_gpe_country, nam_loc, nam_org_nation, nam_org_organization, nam_pro_media_periodic, nam_pro_title,...* [24] In order to keep those detailed information, this is encoded as:

```
<...POL0004.txt#char=17646,17662>
        a nif:RFC5147String ;
 nif:anchorOf "Marya błogosławi"^^xsd:string ;
 nif:beginIndex "17646"^^xsd:nonNegativeInteger ;
 nif:endIndex "17662"^^xsd:nonNegativeInteger ;
 nif:referenceContext eltec:POL0004.txt ;
 itsrdf:taClassRef "<nam_liv_person>"^^xsd:string .
```

For syntactic quality we are using custom Python scripts and SPARQL queries, while RDFUnit tool (Kontokostas et al., 2016) is used as an RDF Unit-Testing suite for semantic quality to validate the RDF data against the NIF Ontology.

Named entities annotated with the proposed dataset with seven categories are properly linked, but some collections, like Polish, have different NER tagset, which should be handled in the next version. Ongoing efforts are being made to develop a solution based on NIF corpus for entity linking with Wikidata.

The interlinking of entities in the NIF corpus offers the potential for new discoveries and valuable insights into literary works, authors, historical figures, and cultural contexts. Moreover, the linked NIF corpus holds the promise of shedding light on language variation, including dialectal differences, historical language evolution, and specific geographic or temporal language usage. This, in turn, can reveal patterns of language change, borrowings, and semantic shifts within literary works. The findings presented in the corpus can facilitate comparative analysis of literary works, genres, and authors, uncovering shared linguistic features, stylistic trends, and thematic connections.

The ELTeC-NIF corpora benefit various users and stakeholders in NLP tasks. NIF's flexibility and interoperability make it valuable for sharing and utilizing NLP data across different domains. Researchers can analyze linguistic annotations and extract features, Tool Developers can use NIF corpora for training or testing, Linguists can study language phenomena, and Semantic Web Developers can integrate NLP data with linked sources for advanced analysis and knowledge discovery.

5 Conclusion and future directions

Future plans include several activities. We would like to generate a version of our corpus adhering to

[24] https://github.com/CLARIN-PL/Liner2

the CoNLL-RDF vocabulary (Chiarcos and Fäth, 2017), a direct rendering of the CoNLL format in RDF, that mimicks CoNLL's original TSV-style layout, and describe a novel extension of CoNLL-RDF, introducing a formal data model, formalized as an ontology. The transformation will rely on the ontology as a basis for linking RDF corpora with other Semantic Web resources. (Chiarcos et al., 2021) Since CoNLL-RDF is easy to read, easy to parse, close to conventional representations and facilitates LLOD integration by applying off-the-shelf Semantic Web technology to CoNLL corpora and annotations, we would like to compare it with NIF. As it doesn't use string URIs directly, CoNLL-RDF is probably less suitable for philological corpora than NIF or Web Annotation – these can directly be used to provide standoff annotations over a digitally edited text on the web, regardless of its format. At the same time, however, it is less verbose than NIF, but limited to a minimal core vocabulary from NIF, so it is possible that it has advantages in speed and scalability. Yet, with the limited amount of data published in both formats currently available, this suspicion cannot be directly evaluated, and such an evaluation would be a prospective goal of our efforts.

Next steps will be integration into the Linguistic Linked Open Data (LLOD) Cloud[25], coordinated effort of the he Open Linguistics Working Group (OWLG), its members and collaborating initiatives. The LLOD cloud is visualized by means of a cloud diagram that displays all the resources with their relative sizes and their connections. (Cimiano et al., 2020b) Finally, due to the available resources, the current version has limited the number of sentences to 1000, but the final version will be produced from the whole novels. Moreover, set of additional novels in extended edition and some novels for languages that do not have level-2 but have level-1 could be playground for testing web services for POS-tagging, morphosyntactic annotation, and named entities recognition and linking.

We also hope that soon an appropriate SPARQL endpoint with with the adequate capacity will become available, so that this valuable resource can be used in linguistic community working with linked data. Publishing RDF data on the web in a sustainable way has previously been proven challenging, and again, we would like to evaluate different approaches and the adequacy of existing hosting solutions for larger-scale data such as linguistic corpora. Also, in the context of European infrastructure initiatives for NLP services, the role of linked data remains somewhat underexplored,[26] and we expect our upcoming experiences in developing such a solution – both on a technological and a political level – to be of particular value for future initiatives on corpus data in RDF.

Also, last, but not least, publishing data is only the very first step in the process. The development of tools that allow their users to benefit from the advantages promised by the application of Linked Data technology to language resources (findability, federation, interoperability, ease of information integration, queriability) will be decisive for the future of LLOD technology. For lexical data, some of these effects can already be seen, as tools for lexicographers to become available, both with respect to automated support for lexicography (Gracia et al., 2021) and with respect to end-user tools for creating and maintaining dictionaries (Fiorelli et al., 2020). Although initial applications have been proposed for annotation engineering (Chiarcos et al., 2022b) and corpus querying (Ionov et al., 2020), the general progress on corpus data may be hampered by the limited amount of data previously available, as well as by the diversity of vocabularies applied for their publication.

Acknowledgements

This paper is based upon work from COST Action NexusLinguarum – "European network for Webcentered linguistic data science" (CA18209), supported by COST www.cost.eu, through Virtual mobility grant and other activities.

References

Vladimir Alexiev and Gerard Casamayor. 2016. Fn goes nif: integrating framenet in the nlp interchange format. In *Proceedings of the LDL 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, pages 1–10.

Piotr Bański. 2010. Why tei stand-off annotation doesn't quite work. In *Balisage: The markup conference*.

[25] http://linguistic-lod.org/llod-cloud

[26] LOD technology is usually seen as a key to achieve sustainable management of scientific data, and it has thus been integrated into the technology stack of initiatives such as SSHOC (Dumouchel et al., 2020). The wide usage of RDF and Linked Data for language resources substantially pre-dates the FAIR principles (Farrar and Lewis, 2007; Chiarcos et al., 2011), but has gained a lot of traction in the course of this development, more in Khan et al. (2022b).

Andrea Bellandi. 2023. Building linked lexicography applications with lexo-server. *Digital Scholarship in the Humanities*.

Julia Bosque-Gil, Verginica Barbu Mititelu, Hugo Gonçalo Oliveira, et al. 2021. Balancing the digital presence of languages in and for technological development. *A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud*.

Emily Brontë. 1847. *Wuthering Heights: A novel by Ellis Bell*. London: T. C. Newby, London.

Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. 2016. Dbpedia abstracts: a large-scale, open, multilingual nlp training corpus. In *Proceedings of the Tenth International Conference on LREC'16*, pages 3339–3343.

Martin Brümmer. 2015. Expanding the nif ecosystem. corpus conversion, parsing and processing using the nlp interchange format 2.0.

Lou Burnard, Christof Schöch, and Carolin Odebrecht. 2021. In search of comity: Tei for distant reading. *Journal of the Text Encoding Initiative*, (14).

Christian Chiarcos. 2012. Powla: Modeling linguistic corpora in owl/dl. In *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, 2012. Proceedings 9*, pages 225–239. Springer.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling frequency, attestation, and corpus-based information with ontolex-frac. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027.

Christian Chiarcos and Christian Fäth. 2017. Conll-rdf: Linked corpora done in an nlp-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 74–88. Springer.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022b. Querying a dozen corpora and a thousand years with fintan. In *Proceedings of the 13th LREC*, pages 4011–4021.

Christian Chiarcos and Luis Glaser. 2020. A tree extension for conll-rdf. In *Proceedings of the 12th LREC*, pages 7161–7169.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *Trait. Autom. des Langues*, 52(3):245–275.

Christian Chiarcos, Maxim Ionov, Luis Glaser, and Christian Fäth. 2021. An ontology for conll-rdf: Formal data structures for tsv formats in language technology. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data in Linguistics*. Springer.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020a. *Linguistic Linked Data in Digital Humanities*, pages 229–262. Springer International Publishing, Cham.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020b. *Linguistic Linked Open Data Cloud*, pages 29–41. Springer International Publishing, Cham.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020c. Linked data-based nlp workflows. *Linguistic Linked Data: Representation, Generation and Applications*, pages 197–211.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020d. Modelling linguistic annotations. In *Linguistic Linked Data: Representation, Generation and Applications*, pages 89–122. Springer.

Fabio Ciotti and Francesca Tomasi. 2016. Formal ontologies, linked data, and tei semantics. *Journal of the Text Encoding Initiative*, (9).

Denis Andrei de Araujo, Sandro José Rigo, and Jorge Luis Victória Barbosa. 2017. Ontology-based information extraction for juridical events with case studies in brazilian legal realm. *Artificial Intelligence and Law*, 25:379–396.

Thierry Declerck, Jorge Gracia, and John P McCrae. 2020. Cost action "european network for web-centred linguistic data science"(nexuslinguarum). *Procesamiento del Lenguaje Natural*, 65:93–96.

Gimena del Rio Riande and Valeria Vitale. 2020. Recogito-in-a-box: From annotation to digital edition. *Modern Languages Open*.

Suzanne Dumouchel, Emilie Blotière, Laure Barbot, et al. 2020. Triple project: building a discovery platform to enhance collaboration. In *ITM Web of Conferences*, volume 33, page 03005. EDP Sciences.

Scott Farrar and William D Lewis. 2007. The gold community of practice: An infrastructure for linguistic data on the web. *Language Resources and Evaluation*, 41:45–60.

Manuel Fiorelli, Armando Stellato, Tiziano Lorenzetti, et al. 2020. Editing ontolex-lemon in vocbench 3. In *Proceedings of the 12th LREC*, pages 7194–7203.

Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, et al. 2014. Naf and gaf: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–16.

Jorge Gracia, IIan Kernerman, and Besim Kabashi. 2021. Results of the translation inference across dictionaries 2021 shared task. In *CEUR workshop proc.*, ART-2021-131934.

Jorge Gracia, Marta Villegas, Asuncion Gomez-Perez, and Nuria Bel. 2018. The apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, 2013, Proceedings, Part II 12*, pages 98–113. Springer.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. 2012. Nif combinator: Combining nlp tool output. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, 2012. Proceedings 18*, pages 446–449. Springer.

Sebastian Hellmann, Jörg Unbehauen, Christian Chiarcos, and Ngonga Ngomo. 2010. The tiger corpus navigator. In *Ninth International Workshop on Treebanks and Linguistic Theories*, volume 91.

Nancy Ide and Keith Suderman. 2014. The linguistic annotation framework: a standard for annotation interchange and merging. *Language Resources and Evaluation*, 48:395–418.

Nancy Ide, Keith Suderman, Marc Verhagen, and James Pustejovsky. 2016. The language application grid web service exchange vocabulary. In *Worldwide Language Service Infrastructure: Second International Workshop, WLSI 2015, Kyoto, Japan, 2015.*, pages 18–32. Springer.

Milica Ikonić Nešić, Ranka Stanković, Christof Schöch, and Mihailo Skoric. 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 LREC*, pages 7–16, Marseille, France. ELRA.

Maxim Ionov. 2021. Apics-ligt: Towards semantic enrichment of interlinear glossed text. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Maxim Ionov, Florian Stein, Sagar Sehgal, and Christian Chiarcos. 2020. cqp4rdf: Towards a suite for rdf-based corpus linguistics. In *The Semantic Web: ESWC 2020 Satellite Events: ESWC 2020 Satellite Events, Heraklion, Crete, Greece, 2020*, pages 115–121. Springer.

Fahad Khan, Christian Chiarcos, Thierry Declerck, et al. 2022a. A survey of guidelines and best practices for the generation, interlinking, publication, and validation of linguistic linked data. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th LREC*, pages 69–77, Marseille, France. ELRA.

Fahad Khan, Christian Chiarcos, Thierry Declerck, et al. 2022b. When linguistics meets web technologies. recent advances in modelling linguistic linked data. *Semantic Web*, (vol. 13, no. 6):987–1050.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, et al. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

Dimitris Kontokostas, Christian Mader, Christian Dirschl, et al. 2016. Semantically enhanced quality assurance in the jurion business use case. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, 2016, Proceedings 13*, pages 661–676. Springer.

Cvetana Krstev. 2021. The serbian part of the eltec collection through the magnifying glass of metadata. *Infotheca - Journal for Digital Humanities*, 21(2):26–42.

Antonio Pareja-Lora, Barbara Lust, Maria Blume, and Christian Chiarcos. 2019. *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. The MIT Press.

Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Martínez Cantón, and Elena González-Blanco. 2021. The diachronic spanish sonnet corpus: Tei and linked open data encoding, data distribution, and metrical findings. *Digital Scholarship in the Humanities*, 36(Supplement_1):i68–i80.

Robert Sanderson, Paolo Ciccarese, and Herbert Van de Sompel. 2013. Designing the w3c open annotation data model. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 366–375.

Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. Creating the european literary text collection (eltec): Challenges and perspectives. *Modern Languages Open*.

Ranka Stanković, Cvetana Krstev, , Duško Vitas, et al. 2022. Distant reading in digital humanities: Case study on the serbian part of the eltec collection. In *Proceedings of the LREC*, pages 3337–3345, Marseille, France. ELRA.

Sabine Tittel, Helena Bermúdez-Sabel, and Christian Chiarcos. 2018. Using RDFa to link text and dictionary data for medieval french. In *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2016): Towards Linguistic Data Science. ELRA, Paris, France, Miyazaki, Japan*.