Towards Semantic Interoperability: Parallel Corpora as Linked Data Incorporating Named Entity Linking

Ranka Stanković*, **Milica Ikonić Nešić**†, **Olja Perišić**‡, **Mihailo Škorić***, **Olivera Kitanović***

*University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka.stankovic,mihailo.skoric,olivera.kitanovic}@rgf.bg.ac.rs
†University of Belgrade, Faculty of Philology, milica.ikonic.nesic@fil.bg.ac.rs,
‡University of Turin, Italy, olja.perisic@unito.it

Abstract

The paper presents the results of the research related to the preparation of parallel corpora, focusing on transformation into RDF graphs using NLP Interchange Format (NIF) for linguistic annotation. We give an overview of the parallel corpus that was used in this case study, as well as the process of POS tagging, lemmatization, and named entity recognition (NER). Next, we describe the named entity linking (NEL), data conversion to RDF, and incorporation of NIF annotations. Produced NIF files were evaluated through the exploration of triplestore using SPARQL queries. Finally, the bridging of Linked Data and Digital Humanities research is discussed, as well as some drawbacks related to the verbosity of transformation. Semantic interoperability concept in the context of linked data and parallel corpora ensures that data exchanged between systems carries shared and well-defined meanings, enabling effective communication and understanding.

Keywords: parallel corpora, named entity linking, named entity recognition, NER, NEL, linked data, NIF, Wikidata

1. Introduction

The motivation for publishing parallel corpora as linked data lies in the benefits of increased accessibility, interoperability, semantic enrichment, community collaboration, and the promotion of open science. These motivations collectively contribute to advancing linguistic research, language technology, and cross-disciplinary insights.

Parallel corpora are essential for multilingual studies, and publishing them as linked data simplifies cross-lingual research. Researchers can efficiently compare and analyze texts in multiple languages, enabling more comprehensive linguistic and cultural studies. Linked data enables semantic enrichment through the integration of annotations, linguistic metadata, and cross-lingual alignments. This enrichment provides deeper context and insights for linguistic research, machine translation, and language technology development.

Previous successful use cases of representation of the linguistic annotations of textual data in RDF (Stanković et al., 2023; Stanković et al., 2024) using NLP Interchange Format (NIF) (Hellmann et al., 2013) inspired this research. NIF facilitates the annotations of various types of linguistic data, e.g. part-of-speech, lemmas, and named entities. By using string-based URIs (Uniform Resource Identifier), NIF additionally accommodates multilingual text materials, allowing the annotations of translation equivalents across different languages via RDF properties. This is directly aligned with the activities of Nexus Linguarum COST Action (Declerck et al., 2020), devoted to the creation, interlinking, enrichment, and evolution of linguistic resources, especially in the context of under-resourced languages and domains. In this paper, the showcase of the Italian-Serbian parallel corpus will be used to illustrate previously mentioned possibilities for annotation and linking.

The Serbian language boasts a rich and intricate morphology, allowing for the declension of toponyms and other proper nouns, which foreign students may not always find easy to identify and derive to their basic form (lemma) searchable in dictionaries and encyclopedias. Some of the difficulties are the transcription of proper names e.g. *Džon* (John), *Đovani* (Giovanni), their declension (*Đovaniju*, loc./dat., *Džona*, acc.), the formation of possessive adjectives from personal names such as *Đovanijev* (m.sg., Đovani's) and *Džonove* (f.pl., John's) all subject to declension. Conversely, Italian, lacking grammatical cases, conveys numerous syntactic relationships through the use of prepositions. For instance, "di Giovanni" can be rendered in Serbian as a possessive adjective, such as "*Đovanijev(a/o/i/e/a)*", or as a genitival phrase, "*od Đovanija*" with its precise semantic interpretation heavily contingent on the context (of Giovanni, by Giovanni...).

To overcome these and similar problems the project "It-Sr-NER: Web services for named entity recognition, linking, and mapping" was implemented as part of CLARIN's "Bridging Gaps" call in 2022 (Perisic et al., 2023). Within this project, web services were developed for annotating named

LDL: Linked Data in Linguistics @ LREC-COLING 2024, pages 115–125
20-25 May, 2024.

entities in text, namely personal names, places, organizations, ethnicities, events, and works of art.

The project participants were experts from Serbian and Italian academic institutions: University of Turin and the Society for Language Resources and Technologies JeRTeh. The result was the creation and publication of web applications and services for annotating named entities (NE) in monolingual and bilingual parallel texts for 24 languages, with a case study focused on Italian and Serbian parallel texts. Furthermore, an Italian-Serbian parallel corpus comprising 10,000 segments of extracted and aligned sentences, chosen from classic works of Italian and Serbian literature, was also created and made publicly available (Perišić et al., 2022b). [1]

The main research objective and contribution of this paper was to provide an existing parallel corpus as linked data that adheres to standardized formats and structures, ensuring interoperability with other datasets and systems. This interoperability will allow researchers to integrate parallel corpora into larger linguistic databases or use them in conjunction with other linked data resources for more comprehensive analyses. The developed procedure can be used for other monolingual or parallel corpora, and thus serve as a point of orientation for future publication workflows of multilingual corpus data publication on the web.

Several aligned corpora exist in which Serbian is one of the languages. In most cases, the second language is English or French, while corpora including the Serbian-Italian combination are rare. Additionally, we see a special contribution to our work in discussing how to establish bridges between Linked Data technologies developed for NLP and data produced and consumed in digital humanities.

In Section 2 we give a short overview of related work concerning the preparation and annotation of parallel corpora, named entity recognition and their linking, linked data standards for corpora, and NLP Interchange Format (NIF) for linguistic annotation. Section 3 brings an overview of the parallel corpus that was used in this case study, the process of POS-tagging and lemmatization, as well as named entity recognition (NER). Section 4 describes integration results: NEL, data conversion to RDF, incorporation of NIF annotations, while in Section 5 validation of produced NIF through the exploration of triplestore using SPARQL is described and the NER and linking is presented. Section 6 is dedicated to the bridging of Linked Data and Digital Humanities research. The concluding remarks and plans for future research are given in Section 7.

[1] It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian

2. Related Work

2.1. Parallel corpora

More than ten years ago Zanettin (2012) emphasized the limited availability of readily accessible sources of parallel corpora across various domains and text genres. The availability of parallel corpora remains limited even for languages with a large number of speakers and a wide range of digital resources despite the ever-increasing demand for them. These available parallel corpora often serve as examples for testing new tools and methods for the less spoken languages with limited resources and for which translations of literary works and other texts are primarily in print, going slowly through digital conversion (Jenn and Fraisse, 2022).

Although the significance of parallel corpora in literary and translation studies has been confirmed (Moratto and Li, 2022), literary parallel corpora are particularly challenging to create due to the increased resources required for their development and concerns related to copyright issues (Dimitroulia, 2023). If recent research has shown that the potential of parallel corpora remains invisible and unknown to most literary translators, the introduction of these technologies into the education of future translators could bring about a change in this trend. At the same time, the exploration of parallel corpora can improve reciprocal language learning from a contrastive perspective that enables the observation of different cross-cultural and linguistic asymmetries (Hunston, 2002).

2.2. Corpus Linked Data Standards

NIF and Web Annotation are two well-known RDF standards for linguistic annotation. Both specifications use URIs (or IRIs) to address corpora, which is similar to how URIs are used in other formats. The 'Best Practices for Multilingual Linked Open Data' (BPMLOD) W3C community group and the LIDER project's[2] results were used in addition to NIF standards, since standards themselves are somewhat technical and not very user-friendly. This document describes NIF as a format for corpus data.[3]

In (Hellmann et al., 2013) NIF was employed as the corpus format to ensure compatibility with DBpedia through Linked Data and to facilitate interoperability with NLP tools. DBpedia abstracts were one of the first implementations of NIF (Brümmer, 2015; Brümmer et al., 2016) on an open, large-scale corpus of annotated Wikipedia texts in six languages, with over 11 million texts and more than 97 million entity links.

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

[3] BPMLOD-NIF, http://bpmlod.github.io/report/nif-corpus/index.html

FrameNet (FN) lexical database for English has been published as RDF Linked Open Data (LOD), including the corpus of text that has been annotated using FN. Alexiev and Casamayor (2016) compared FN-LOD with NIF, and proposed to integrate FN into NIF. The widely used standards for linguistic annotations in RDF are: 1) Annotation (Sanderson et al., 2013), published as a W3C standard (recommendation) in 2017;[4], 2) POWLA (Chiarcos, 2012), a reconstruction of the Linguistic Annotation Framework (Ide and Suderman, 2014) in OWL2/DL; 3) CoNLL-RDF focusing on the compatibility with tabular ('CoNLL') formats as used in NLP (Chiarcos and Glaser, 2020).

While describing principles for annotating text data using RDF-compliant formalism to be accessible from the LLOD ecosystem, Cimiano et al. (2020a) recommended including the full text of the annotated document in the RDF data, to preserve interoperability.

After studying the relevant literature and taking into consideration the characteristics of our data, we decided to follow the BPMLOD draft recommendation and apply NIF (version 2.0) to our data, similar to our approach in the previous project (Stanković et al., 2023; Stanković et al., 2024). We are working with an annotated parallel corpus, which opens up opportunities to explore the potential of RDF technology for cross-lingual linking, as well as for the linking of corpora with annotations and lexical resources.

2.3. NLP Interchange Format (NIF) for Linguistic Annotation

NIF is a community standard developed through a series of research projects at the AKSW Leipzig, Germany, and maintained by the same group. A typical URI/IRI consists of two main components, a base name that serves to locate the document, and an optional fragment identifier. For different media types and 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. (2022) report that this is one area where there is a real necessity for documentation that provides clear guidelines (GLs) and BPs. The research we present could be a showcase for the use of NIF and the transformation of parallel corpora to NIF. This paper contributes to this effort by providing a case study on the use of NIF as an RDF-based format for describing strings in the novel, relying on the classes and properties that are formally defined within the NIF Core Ontology 2.0.[5] The reason not to use the latest version 2.1 of NIF Ontology is the lack of full documentation.

3. Data Preparation and Preprocessing

3.1. Description of the Parallel Corpus

The Italian-Serbian corpus It-Sr-NER (Perišić et al., 2022b) consists of 10,000 aligned segments (sentences) taken from Italian and Serbian translations of ten different novels. For the presented work, 1000 aligned sentences from various novels were selected. Table 1 presents an overview of the novels in It-Sr-NER, where the last column designates the novels whose sentences belong to the 1000-sentence corpus.

The novels were aligned and converted into the TMX (Translation Memory eXchange) (Serge, 2020) format using the ACIDE program for creating parallel corpora (Obradović et al., 2008; Krstev and Vitas, 2011). Each segment in Italian and Serbian is numbered and paired with the corresponding language segment(s) indicated by the `"xml:lang"` attribute.[6] The It-Sr-NER corpus is available on the CLARIN Center and can be accessed through the VLO (Virtual Language Observatory) and Language Resource Switchboard. The corpus includes the aligned bilingual version, as well as individual monolingual versions, and named entities that have been automatically annotated (Perišić et al., 2022a). Additional information can be found in the *GitHub* [7] and it is searchable in the Biblisha digital library (Stanković et al., 2018) (Stanković et al., 2017). [8]

The resources developed in this project are open and accessible to researchers, teachers, and students, but the biggest benefit will be for those interested in the Italian language in Serbia and the Serbian language in Italy. Given the polycentrism of the Serbo-Croatian language, the students and teachers in Croatian, Montenegrin, and Bosnian universities and schools could also benefit from this corpus and web services.

3.2. POS tagging and lemmatization

The complete parallel corpus was annotated with part-of-speech (POS) tags using *Universal POS* tagset, and lemmas.

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

[5] https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html

[6] TXM file of the novel "Around the World in Eighty Days"

[7] It-Sr-NER GitHub repository

[8] Bibliša digital library

Name of Novel	Novel Name Translation	Samples in NIF
Il nome della rosa	The Name of the Rose	✓
Le avventure di Pinocchio	The Adventures of Pinocchio	✓
Storia di chi fugge e di chi resta, L'amica geniale	Those Who Leave and Those Who Stay	✓
Uno, nessuno e centomilar	One, None and a Hundred Thousand	✓
Anikina vremena	Legends of Anika	✓
Na drini ćuprija	The Bridge on the Drina	✓
Nečista krv	Impure Blood	
Opštinsko dete	Municipal child	
Bašta, pepeo	Garden, Ashes	
Le Tour du monde en quatre-vingts jours	Around the World in Eighty Days	

Table 1: An overview of the novel samples included in the corpus.

The Serbian part of the corpus was annotated using a multi-model tagger for the Serbian language, *BEaST* (Stanković et al., 2022) which uses both *TreeTagger* (Schmid, 2013) and *spaCy*[9] models trained on part-of-speech tagging task using the manually annotated, publicly available corpus *SrpKor4Tagging* (Vitas et al., 2021). The lemmatization is performed after the POS-tagging step, using electronic morphological dictionaries for Serbian (Krstev and Vitas, 2006) (Krstev, 2008; Vitas and Krstev, 2012), incorporated through the aforementioned TreeTagger model.

The Italian part of the corpus was annotated using *spaCy* model for Italian (Explosion, 2022), using the UD annotation scheme obtained by conversion from the Italian Stanford Dependency Treebank, released for the dependency parsing shared task of Evalita-2014 (Bosco et al., 2014).

3.3. Named Entity Recognition

For NER in Serbian texts Jerteh-355-tesla (Ikonić Nešić et al., 2024), a version of Jerteh-355 (Škorić, 2024) language model was used. Jerteh-355, based on the RoBERTa-large architecture (Liu et al., 2019), was pre-trained for Serbian on a diverse corpus of approx 4 billion tokens. Jerteh-355-tesla was fine-tuned specifically for NER task, using spacy framework on the corpus of Serbian novels published between 1840 and 1920, named SrpELTeC-gold (Krstev et al., 2021), newspaper articles and sentences generated from the Wikidata (Wikimedia, 2023) and Leximirka lexical database (Stanković et al., 2021). It achieves an F_1 score of approx 96% on the test dataset.

For the Italian language texts, spacy model `it_core_news_sm-3.4.0` (Explosion, 2022) was used, which was trained on a synthetic NER corpus WikiNER, based on the text and structure of Wikipedia (Nothman et al., 2013). The model achieved F_1 score of 86% on the test set.

After automatic annotation, the INCEpTION (Klie et al., 2018) was used for manual correction and linking of named entities. In this paper, the focus was on the three most frequent types of named entities across language-specific models: persons (`<PERS>`), locations (`<LOC>`), and organizations (`<ORG>`), as explained in Subsection 2.3.

Table 2 presents statistics of several named entities per class in Serbian (sr) and Italian (it) datasets, with explanations of entity types.

4. Integration

4.1. Named Entity Linking

After annotating the parallel corpus as described in the previous section, the next step was to link entities belonging to one of the NE classes with Wikidata (Wikimedia, 2023). Extracted PERS entities refer mostly to the characters of novels, LOC entities designate places where the action of a novel takes place (geopolitical locations) while ORG represents organizations mentioned in novels. Entries in Wikidata didn't exist for characters of some novels; thus, similar to the approach in (Ikonić Nešić et al., 2021), the OpenRefine (David Huynh, 2022) and QuickStatements (Manske, 2019) were used to create 111 appropriate items for 5 novels (56 characters of the novel *"Storia di chi fugge e di chi resta, L'amica geniale"* (Q55517451) by Elena Ferante). For novel *"Le avventure di Pinocchio"* (Q8065468) all characters were already in Wikidata.

The named entities for both languages were linked with Wikidata in additional layer of annotation a Wikidata identifier is assigned to each entity. For example, *Jakša*, a character from the novel "Legends of Anika" (`wd:Q61133860`), is recognized as a person, assigned NE tag `<PERS>` and linked with URL `http://www.wikidata.org/entity/Q122730462`. The annotation and linking with Wikidata using the INCEpTION tool is presented in Figure 1. Two more entities are recognized in the text presented in this figure: PERS *Anika* (wd:Q122730455) and LOC *Višegrad* (wd:Q239266).

The full process of linking entities with knowledge bases using the INCEpTION annotation platform is described in (Klie et al., 2020).

For annotating named entities (NE), several ontologies were consulted. The following NE type equivalents were used

[9] SpaCy

Entity	Explanation	sr	it
PERS Personal names	First names, surnames, nicknames and their combinations (of real people and fictional characters, gods and saints).	901	1036
LOC Locations	Continents, countries, regions, populated places, oronyms, water surfaces, names of celestial bodies, city locations.	257	310
ORG Occupations and titles	Names of companies, political parties, educational institutions, sports teams, hospitals, museums, libraries, hotels, cafes, and places of worship.	31	30

Table 2: Number of named entities per class

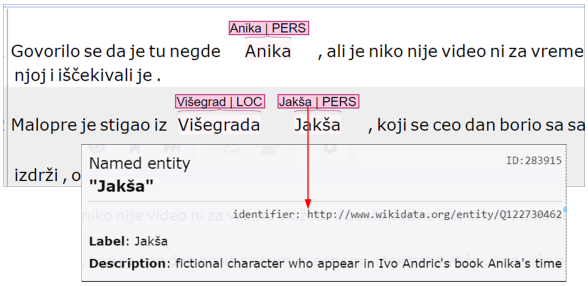

Figure 1: An annotated example

from OLIA:[10] `olia:Person`, `olia:Space`, `olia:Organization`. dbo[11] namespace was introduced to link NEs with DBpedia, and *wd* namespace for Wikidata. The following classes were used to link types of recognized NEs: `dbo:Person` = `wd:Q5`, `dbo:Place` = `wd:Q7884789`, `dbo:Organisation` = `wd:Q43229`.

4.2. Data Conversion to RDF

A collab notebook was prepared for the transformation of the parallel corpus into NIF. The library *rdflib*[12] was used for RDF management.[13] Code comprises classes *Corpus_mono*, *Sentence*, *Word*, *NamedEntity*, *Corpus_bili* for necessary transformations and a set of additional functions. *Corpus_mono* takes as input *TSV* file with annotations and produces a RDF graph (a *ttl* file) for one language, instantiating further for each sentence an object of a *Sentence* class, that produces RDF triples related to the object of `nif:Sentence` type. Further, class *Word* manages tokens from the file and generates RDF triples for objects of the `nif:Word` type, while *NamedEntity* finds the words (and tokens) that belong to one name entity, specify its type, and link it to Wikidata, if exists.

For interlinking sentences that are translation units, class *Corpus_bili* is used.

Two monolingual corpora consist of the same number of segments, that are aligned as translation equivalents. Since NIF does not support translation units and translation unit variants (as TMX standard), the sentence class `nif:Sentence` is used, as the most similar NIF concept.

The main function *write_gcorpus_mono* instantiate RDF Graph with the following namespaces: *itsrdf, nif, olia, dc, dct, ms, wd, wdt, dbo, eltec*. After *Corpus_mono* is created, the first set of triples is introduced to the monolingual corpus.

Figure 2 presents an outline of the model for a parallel corpus in NIF.

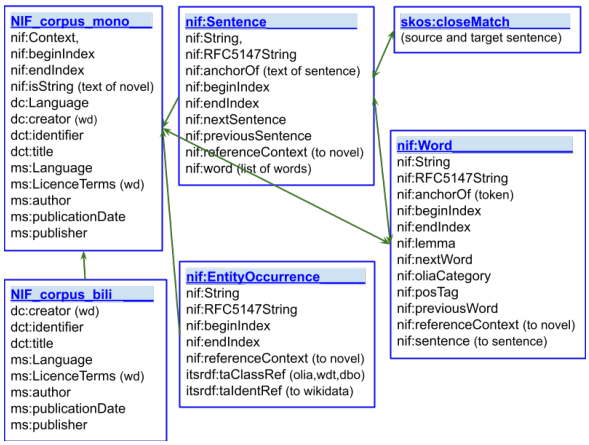

Figure 2: A data model for a parallel corpus in NIF

For establishing links between translation equivalents in different languages we used `skos:closeMatch` from *SKOS* (Simple Knowledge Organization System).[14] The `skos:closeMatch` property indicates that the two objects are sufficiently similar that they can be used alternately in applications.

4.3. Incorporating NIF Annotations

NIF Terse RDF Triple Language (ttl) was used as a serialization for transformation into linked data.

[10] http://purl.org/olia/discourse/olia_discourse.owl

[11] https://dbpedia.org/ontology/

[12] https://rdflib.readthedocs.io/en/stable/

[13] The code is available in the GitHub repository.

[14] https://www.w3.org/TR/skos-reference/

Dataset with 1000 aligned sentences within six *ttl* files derived from the corpus described in Subsection 3.1, is published and included in LRE map.[15]

The core class `nif:String` is used for the monolingual corpus content itself (a text in Italian and a corresponding text in Serbian), described by `nif:beginIndex` and `nif:endIndex`. Dublin Core vocabulary is used for predicates related to the language, author, identifier, and title. META–SHARE ontology[16] is used to describe language, license terms, author, publisher, and publication year.

For illustration, we will present a part of an Italian sentence *"- Dunque, compar Geppetto, - disse il falegname in segno di pace fatta, - qual è il piacere che volete da me ?"*[17] from the novel "The Adventures of Pinocchio"[18] and discuss some of its parts. 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. A substring of the `nif:Context` can be: a single word, a sentence, or a named entity that is linked to the relevant `nif:Context` resource via `nif:referenceContext`. Beginning and end indices refer to the string content (sentence) represented by the context. The previous and the next sentence are references as well as a list of words.

```
<http://url/it1.txt#char=105530,105643>
 a nif:RFC5147String, nif:Sentence,
 nif:String;
 nif:anchorOf "- Dunque , compar
  Geppetto ,  - disse il falegname
  in segno di pace fatta , - qual è il
  piacere che volete da me ?" ;
 nif:beginIndex "105530"  ;
 nif:endIndex "105643"  ;
 nif:nextSentence
  <http://url/it1.txt#char=105644,105851>;
 nif:previousSentence
  <http://url/it1.txt#char=105356,105529>;
 nif:referenceContext
  <http://url/it1.txt>;
 nif:word
  <http://url/it1.txt#char=105530,105531>,
  <http://url/it1.txt#char=105532,105538>,
  ...
  <http://url/it1.txt#char=105642,105643>;
  dct:identifier "585" .
```

The following classes: `nif:Word`, `nif:Phrase`, `nif:Sentence` represent a segment of a text, depending on the unit of annotation. The property `nif:referenceContext` points to the respective `nif:Context` instance of the text segment. The segment position inside the context is specified using the `nif:beginIndex` and `nif:endIndex` properties. The actual text segment can be specified using the `nif:anchorOf` property.

The following listing presents triplets for tokens (words). Apart from text segments (indices), additional grammatical information and relations can be included. The information about the part of speech can be linked using the `nif:posTag` property, while for the canonical form the `nif:lemma` property is used. Previous and next words are linked with the following properties: `nif:previousWord` and `nif:nextWord`. To link a word or a named entity with its sentence the `nif:sentence` property is used.

```
<http://url/it1.txt#char=105541,105547> a
 nif:RFC5147String, nif:String, nif:Word;
 nif:anchorOf "compar";
 nif:beginIndex "105541";
 nif:endIndex "105547";
 nif:lemma "compar";
 nif:nextWord
  <http://url/it1.txt#char=105548,105556>;
 nif:oliaCategory olia:CommonNoun ;
 nif:posTag "NOUN";
 nif:previousWord
  <http://url/it1.txt#char=105539,105540>;
 nif:referenceContext
   <http://url/it1.txt> ;
 nif:sentence
  <http://url/it1.txt#char=105530,105643>.
```

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), various ontologies are utilized, including `olia:Person` from Olia ontology, `wdt:Q5` from Wikidata, and `dbo:Person` from DBpedia.

```
<http://url/it1.txt#char=105007,105015> a
 nif:RFC5147String, nif:String, nif:Word;
 nif:anchorOf "Geppetto";
 nif:beginIndex "105007";
 nif:endIndex "105015";
 ...
 itsrdf:taClassRef olia:Person,
  wd:Q5, dbo:Person ;
 itsrdf:taIdentRef wd:Q1428120 .
```

Figure 3 presents the transformation of novels into aligned TMX-XML, annotation in *TSV* files (NER+NEL) into RDF (NIF).

[15] Uncompressed files are accessible at: URL, with a CCA 4.0 International license. Zipped files will be available also at CLARIN, the European Language Grid portal, and other language repositories.

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

[17] "So, Compare Geppetto, - said the carpenter as a sign of peace made, - what pleasure do you want from me?"

[18] Le-avventure-di-Pinocchio.xml

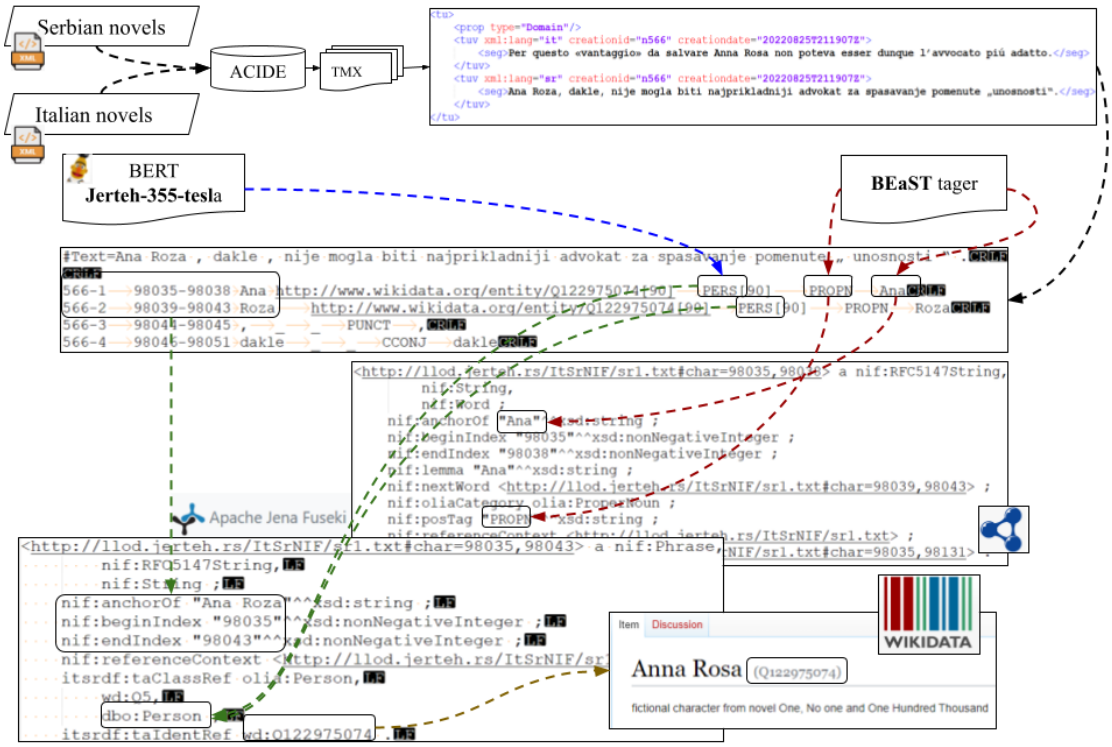

Figure 3: Workflow of the transition from a novel to LOD

5. Quering It-Sr-NER using SPARQL

Apache Jena Fuseki (Apache Software Foundation, 2023) is used for the management of the RDF graphs in the form of *ttl* files. Dataset *ItSrNIF* was created by uploading all files which generated 1,002,834 triples for 1000 sentences in each language. The Italian part of the corpus has 36,457 words, 1036 persons (`wd:Q5`), 310 toponyms (`wd:Q7884789`), 30 organizations (`wd:Q43229`), while Serbian part has 33,514 words, 901 persons, 257 toponyms, 31 organization.

The following query presents SPARQL query in Fuseki presenting retrieved result with aligned sentences.

```
SELECT ?sr ?srt ?it ?itt
WHERE { ?sr a nif:Sentence;
    nif:anchorOf ?srt .
   ?it a nif:Sentence;
   nif:anchorOf ?itt .
   ?it skos:closeMatch ?sr .
}
```

The query retrieves Serbian sentences represented by variables `?sr` (sentence ID) and `?srt` (sentence itself), Italian sentences represented by variables `?it` and `?itt`, while the query constraint demanding a link of a type `skos:closeMatch` between the sentence identifiers `?sr` and `?it` ensures that sentences are translation equivalents.

Figure 4 presents a Fuseki screenshot with SPARQL query for counting and presenting aligned named entities in Serbian, Italian, and their Wikidata URI.

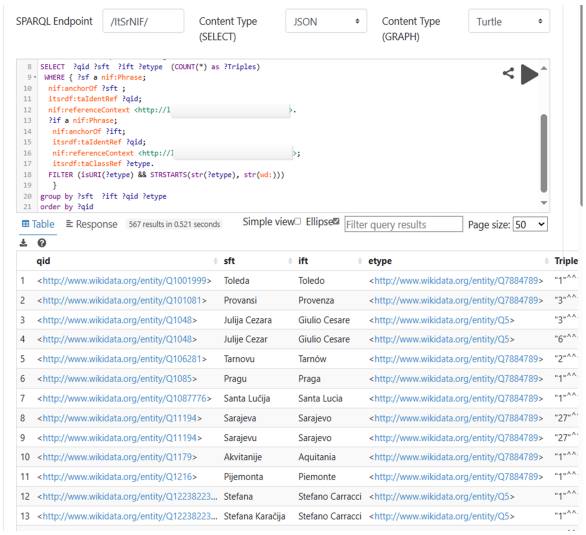

Figure 4: SPARQL query with aligned sentences

6. Discussion

The presented research connects the previous results from the fields of Digital Humanities (Ikonić Nešić et al., 2022) and Linked Data (Hell-

Lng	*txt*	*tsv*	*ttl*	Fuseki
it	0.17	1.7	24.4	/
sr	0.19	1.5	26.5	/
All	0.36	3.2	51.0	317

Table 3: Size of files in MB. ***txt*** - plain text, ***tsv*** - tab separated INCEpTION format (POS, lemmas, NER, NEL), ***ttl*** - NIF files, **Fuseki** -whole repository.

mann et al., 2012; Brümmer, 2015; Alexiev and Casamayor, 2016; Cimiano et al., 2020b) which are traditionally considered separate areas of research. Parallel corpora are widely used in translation studies, while Linked Data focuses on interlinking and integrating diverse datasets. The integration of parallel resources with the broader Linked Data ecosystem, described in this paper, contributes to the efforts to bridge the gap between these two areas.

We are aware that NIF has some potential downsides, one of which is a high degree of verbosity. Therefore, the scalability issues for such kinds of data should be carefully planned. Table 3 gives an overview of the differences in size that can be expected for different levels of annotation and formats, taking as an example the data set with 1000 sentences. It can be seen that the size of NIF files is 16 times larger than *TSV* version with similar information, while the Fuseki repository size for both languages and for the same dataset is more than 6 times larger than the repository with *ttl* files.

The presented pipeline transforming parallel corpus into NIF-linked data (Figure 3), offers several benefits: multilingual research and translation, cross-lingual information retrieval, multilingual information extraction, cultural and societal insights, and bridging language barriers. In summary, the benefits of parallel corpus NIF linked data, extend to various domains, including machine translation, linguistics, language learning, and cross-lingual information access, making it a valuable resource for researchers, businesses, and individuals seeking to bridge language gaps and expand their global reach. Analyzing parallel corpus data in a distributed environment using federated SPARQL queries can reveal cultural and societal differences in how topics are discussed and portrayed across languages.

The greatest benefits will be in the field of translation, encompassing teaching and lexicography, especially in resolving cases of lexical anisomorphism. This phenomenon results not only from linguistic asymmetry but also from cultural differences, so this insight can be valuable for cross-cultural studies and international business strategies. The varied lexical realization of a concept or its lack of lexicalization creates lexical gaps that can be identified, understood, and translated by applying targeted translation strategies. These strategies are made possible through data linking with other layered multilingual resources. Through this approach, the semantic essence of every word can be grasped, beginning from individual concepts and extending to their functional manifestation within the context.

7. Conclusion

One way to achieve semantic interoperability is by leveraging parallel corpora and incorporating NEL. By representing parallel corpora as linked data, we can establish links between equivalent concepts or entities in different languages, thereby enhancing cross-lingual information exchange. This paper demonstrated NEL for people, organizations, and locations by linking their references in texts to their corresponding entries in Wikidata. By linking these entities to standardized identifiers or ontologies, the interoperability of data is greatly improved. Incorporating NEL into parallel corpora as linked data not only enhances cross-lingual interoperability but also fosters better integration with the broader semantic web. When parallel corpora are exposed as linked data, they become part of the larger network of linked open data, allowing for a more comprehensive and coherent exchange of information. Further research will include NEL model training (Upadhyay et al., 2018), as well as the publication of all 10,000 aligned segments in NIF.

8. Acknowledgements

The authors extend their gratitude to Prof. Cvetana Krstev for her invaluable contributions. This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA and COST Action NexusLinguarum (CA18209)

9. Bibliographical References

Vladimir Alexiev and Gerard Casamayor. 2016. FN goes NIF: integrating FrameNet in the NLP interchange format. In *Proc. of the LDL 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, pages 1–10.

Cristina Bosco, F Dell'Orletta, S Montemagni, Manuela Sanguinetti, Maria Simi, et al. 2014. The Evalita 2014 Dependency Parsing task. In *Proc. of the 4th Int. Workshop EVALITA 2014*, pages 1–8. Pisa University Press.

Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. 2016. DBpedia abstracts: a large-scale, open, multilingual NLP training corpus. In *Proc. of the 10th Int. 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.

Christian Chiarcos. 2012. POWLA: Modeling linguistic corpora in OWL/DL. In *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, 2012. Proc. 9*, pages 225–239. Springer.

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

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

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020b. Linked Data-Based NLP Workflows. *Linguistic Linked Data: Representation, Generation and Applications*, pages 197–211.

Thierry Declerck, Jorge Gracia, and John P McCrae. 2020. COST Action "European" network for Web-centred linguistic data science"(NexusLinguarum). *Proc. del Lenguaje Natural*, 65:93–96.

Titika Dimitroulia. 2023. Corpora and literary translation. In *Advances in Corpus Applications in Literary and Translation Studies*, pages 103–118. Taylor & Francis.

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, 2013, Proc., 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 Int. Conference, EKAW 2012, 2012. Proc. 18*, pages 446–449. Springer.

Susan Hunston. 2002. *Corpora in applied linguistics*. Cambridge University Press.

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

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 *Proc. of the 8th Workshop on Linked Data in Linguistics the 13th LREC*, pages 7–16, France. ELRA.

Milica Ikonić Nešić, Ranka Stanković, and Biljana Rujević. 2021. ELTeC Corpus in Wikidata. *Infotheca - Journal for Digital Humanities*, 21(2).

Ronald Jenn and Amel Fraisse. 2022. Benefits of a Corpus-based Approach to Translations: The Example of Huckleberry Finn. In *Advances in Corpus Applications in Literary and Translation Studies*, pages 176–190. Routledge.

Fahad Khan, Christian Chiarcos, Thierry Declerck, et al. 2022. A Survey of Guidelines and Best Practices for the Generation, Interlinking, Publication, and Validation of Linguistic Linked Data. In *Proc. of the 8th Workshop on Linked Data in Linguistics the 13th LREC*, pages 69–77. ELRA.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2020. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6982–6993. Association for Computational Linguistics.

Cvetana Krstev. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries*. Faculty of Philology of the University of Belgrade.

Cvetana Krstev and Duško Vitas. 2011. An Aligned English-Serbian Corpus. In *ELLSIIR*, volume I, pages 495–508, Belgrade. Faculty of Philology, University of Belgrade.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Riccardo Moratto and Defeng Li. 2022. *Advances in Corpus Applications in Literary and Translation Studies, Introduction*, pages 1–9. Taylor & Francis.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175. AI, Wikipedia and Semi-Structured Resources.

Ivan Obradović, Ranka Stanković, and Miloš Utvić. 2008. Integrisano okruženje za pripremu paralelizovanog korpusa. *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, pages 563–578.

Olja Perisic, Stanković Ranka, Ikonić Nešić Milica, Škorić Mihailo, et al. 2023. It-Sr-NER: CLARIN Compatible NER and GeoparsingWeb Services for Italian and Serbian Parallel Text. In *Selected Papers from the CLARIN Annual Conference 2022, Czechia, 2022*, pages 99–110. Linköping University Electronic Press.

Robert Sanderson, Paolo Ciccarese, and Herbert Van de Sompel. 2013. Designing the W3C open annotation data model. In *Proc. of the 5th Annual ACM Web Science Conference*, pages 366–375.

Helmut Schmid. 2013. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154.

Ranka Stanković, Christian Chiarcos, Miloš Utvić, and Olivera Kitanović. 2023. Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data. In *Proc. of the 4th Conference on Language, Data and Knowledge*, pages 180–191.

Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2017. Keyword-based search on bilingual digital libraries. In *Semantic Keyword-Based Search on Structured Data Sources*, pages 112–123, Cham. Springer International Publishing.

Ranka Stanković, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. Parallel Bidirectionally Pretrained Taggers as Feature Generators. *Applied Sciences*, 12(10).

Ranka Stanković, Christian Chiarcos, and Milica Ikonić Nešić. 2024. Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora. In *Book of Abstracts of the UniDive 2nd general meeting, 8-10 February 2024, Naples*.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint Multilingual Supervision for Cross-lingual Entity Linking. In *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2495, Brussels, Belgium. Association for Computational Linguistics.

Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, XVIII:279–292.

Federico Zanettin. 2012. *Translation practices explained: translation-driven corpora*. St Jerome Publishing.

Mihailo Škorić. 2024. Roberta: A robustly optimized bert pretraining approach. *Infotheca - Journal for Digital Humanities*.

10. Language Resource References

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

David Huynh. 2022. *OpenRefine*. Metaweb Technologies, Inc, https://openrefine.org/.

Explosion. 2022. *it_core_news_sm spaCy pipeline model*. spaCy, https://github.com/explosion/spacy-models/releases/tag/it_core_news_sm-3.4.0.

Milica Ikonić Nešić and Saša Petalinkar and Mihailo Škorić and Ranka Stanković. 2024. *Jerteh-355 Tesla - model for Named Entity Recognition.* Hugging Face, https://huggingface.co/Tanor/sr_pln_tesla_j355.

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 Duško Vitas. 2006. *SrpMD - Serbian morphological dictionaries*. ELG, https://live.european-language-grid.eu/catalogue/lcr/17355, 1.0.

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.

Magnus Manske. 2019. *QuickStatements*. Free Software Foundation, https://quickstatements.toolforge.org/, 2.0.

Perišić, Olja and Stanković, Ranka and Ikonić Nešić, Milica and Škorić, Mihailo and Vitas, Duško and Krstev, Cvetana. 2022a. *It-Sr-NER*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Perišić, Olja and Stanković, Ranka and Vitas, Duško and Krstev, Cvetana and Moderc, Saša. 2022b. *It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian*. CLARIN-IT, http://hdl.handle.net/

20.500.11752/OPEN-980. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Heiden Serge. 2020. *The TXM Platform*. National Center for Scientific Research, https://txm.gitpages.huma-num.fr/textometrie/.

Ranka Stanković and Olivera Kitanović and Nikola Vulović and Cvetana Krstev. 2018. *Bibliša: Aligned Collection Search*. JeRTeh, http://biblisha.jerteh.rs/.

Ranka Stanković and Mihailo Škorić and Biljana Lazić and Cvetana Krstev. 2021. *Leximirka lexical database*. ELG, https://live.european-language-grid.eu/catalogue/tool-service/17356.

Duško Vitas and Cvetana Krstev and Ranka Stanković and Miloš Utvić and Mihailo Škorić. 2021. *SrpKor4Tagging*. ELG, https://live.european-language-grid.eu/catalogue/corpus/9295, 1.0.

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