SrpNemKor and Linked Open Data

ABSTRACT: This paper presents the possibilities of applying the Semantic Web principles and guidelines to the parallel corpus of novels *SrpNemKor*. First, we analyzed linking some bibliographic metadata about novels (the novel's author name and the title) with relevant records in the reference semantic databases. Next, based on the *SrpNemKor* content the bilingual dictionary *SrpNemLex* dataset was created as linked open data. At the end, the representation of novels from the corpus with Wikidata and the search with SPARQL queries with different possibilities for visualization of the obtained results are illustrated.
KEYWORDS: SrpNemKor, metadata, Semantic Web, Linked Open Data, RDF, SrpNemLex, Wikidata, SPARQL.

Jelena Andonovski
andonovski@unilib.rs
*University Library
"Svetozar Marković"
Belgrade, Serbia*

1 Introduction

Semantic Web or Web 3.0 is defined as a mechanism for structuring and linking data with precisely defined meaning, which forms interconnected data sets that can be searched through a unified search system. In this way the "Web of Data" is transformed into a "Web of Doculments",[1] creating from the Web a global knowledge database. In this concept the meaning of data expressed in natural language is explicitly annotated, so that machines can read it and correctly understand it (Rurarz et al. 2021). The key paradigm of the Semantic Web is "Linked Data" (LD), which allows its principles and technologies (Janik, Scherp, and Staab 2011) to become available and applicable for various applications, and data sets to become interconnected and available in open access, Linked Open Data (LOD). (Radulovic et al. 2015).

---

1. Linked Data. W3C, accessed 10.3.2023.

The concept of the Semantic Web and open linked data technologies expand the traditional web by using a standard markup language and similar processing tools, where RDF (Resource Description Framework)[2] plays a significant role and makes more efficient information retrieval solutions possible (Shah et al. 2002; Stanković and Davidović 2021).

The Linked Open Data initiative has been developing for more than a decade bringing together diverse resources and data types in a cloud technology. The greatest part of the cloud are knowledge bases, one of which is Wikidata[3]. Wikidata is a multilingual, general-purpose knowledge base developed by the Wikimedia Foundation to enable the retrieval and storage of multi-language structured data from Wikipedia in an interoperable machine-readable format. (Vrandečić and Krötzsch 2014). It is used not only to structure data from various Wikimedia projects, but also in numerous applications based on the Semantic Web, and we used it in our research as well. The LOD cloud is divided into nine subclouds[4] that are further divided into subcategories. One of the sub-clouds that is of particular importance for our research is *linguistics*[5]. It brings together or describes linguistic resources that can be used for research purposes in various fields of linguistics or within the field of natural language processing (McCrae 2016; Cimiano et al., 2020a). Sub-cloud *linguistics* (Chiarcos and Pareja-Lora 2019; Chiarcos et al. 2020; Cimiano et al., 2020b) is divided into seven categories: corpora, lexicons and dictionaries, terminologies, thesauri and knowledge bases, linguistic resource metadata, linguistic data categories, typological databases and other.

In the paper, we analyzed the possibilities of applying the principles and guidelines of the Semantic Web to the content of the parallel corpus *SrpNemKor*. First, we linked bibliographic metadata about novels (author's name and title) with corresponding records in reference knowledge bases, next, based on the content of the corpus, a bilingual dictionary of general type, *SrpNemlex*, was prepared as a set of open linked data and, finally, we analyzed an example from Wikidata for one of the novels from the corpus and the possibility to search with SPARQL queries.

---

2. Resource Description Framework, accessed 10.3.2023.

3. Wikidata, accessed 10.3.2023.

4. The Linked Open Data Cloud, accessed 20.4.2023.

5. The Linked Open Data Cloud. Subclouds by domain. Linguistics, accessed 20.4.2023.

2   Parallel corpus *SrpNemKor*

Parallel Serbian-German literary corpus was created during the work on dissertation "Linked Open Data and language resources in creating Serbian-German literary corpus" (Андоновски 2019). The corpus contains a total of 14 parallelized novels, of which seven are by Serbian writers that have been translated into German and seven by German-speaking writers (in this case, we mean writers from the territory of Germany and Austria) that have been translated into Serbian. After parallelization, the corpus was placed in the *Bibliša* digital library[6] as a separate *SrpNemKor* collection[7] divided in two sub-collections, *Novels written originally in German* and *Novels written originally in Serbian*.

Bibliša is a digital library developed within Belgrade Natural language Processing (NLP) group and today maintained by the Association for Language Resources and Technologies (JeRTeX).[8] It was created with the aim of improving the search of multilingual digital libraries of electronic journals and other parallelized texts (Stanković et al. 2012). *Bibliša* enables multilingual full-text search of parallel collections with a possibility to enhance and refine users' queries both morphologically and semantically by using various lexical resources (Stanković et al. 2015; Stanković et al. 2016; Stanković et al. 2017). Initially, *Bibliša* contained Serbian-English parallel collections and it was searchable only in these two languages. In order to enable the search of the collection also in German and make use of the existing query expansion tools, one of the terminological resources, the multilingual database *Termi*, has been enhanced[9] with Serbian-German equivalent list of lexical units with synonyms in both languages (more about in the next section). We used this same list for generating bilingual Serbian-German dictionary *SrpNemLex* as a set of linked open data (more about it in Section 4.2).[10]

---

6. Библиша, accessed 20.4.2023.

7. SrpNemKor, accessed 20.4.2023.

8. Association for Language Resources and Technologies – JeRTeh, accessed 20.4.2023.

9. Termi is a multilingual terminological database lunched as a support for the development of terminological dictionaries in various domains (mathematics, computer science, mining, library science, computational linguistics, etc.). It enables query enhancement in Bibliša; accessed 20.4.2023.

10. Billingual dictionary SrpNemLex, accessed 20.4.2023.

3 Serbian-German translation equivalents list

Serbian-German translation equivalents list of lexical units was extracted from *SrpNemKor* aligned textual collection (Andonovski, Šandrih, and Kitanović 2019). For this purpose we used the BilTE (Bilingual Terminology Extraction)[11] system described in (Krstev et al. 2018) and developed by JeRTeX. It was originally designed for the extraction of bilingual terminology based on parallel bilingual text collections, terminological lists of the source language and a system for the extraction of multi-word terms of the target language. In our case, the source language is German, and the target language is Serbian. For these purposes, we used the following language resources:

1. The aligned Serbian-German literary textual collection *SrpNemKor* with 48.004 aligned sentences (Андоновски 2021);
2. A list of lexical items for German as a source language that was prepared using two sources. The first source was the open German WordNet *Open-de-WordNet*[12] (*Open DE WordNet Initiative* 2023) with approximately 120,000 units. The second one was a list of the most frequent German words available in Wiktionary,[13] with approximately 10,000 units. These two lists were united and we got preliminary list of German lexical units. For better results, we performed a word-by-word lemmatization on a source list of lexical units using German model available in spaCy.[14] The model performs lemmatization using an underlying lookup table of 355,354 single-word units. After lemmatization and duplicate elimination, the final list contained 27,638 distinct German lexical units;

---

11. BilTE, accessed 20.4.2023.

12. The *Open-de-WordNet* initiative was launched with the idea of creating a German-language Wordnet that would become part of the wider multilingual Wordnet environment, be fully open access and usable within the Natural Language Toolkit processing package. The first version of *Open-de-WordNet* was made on the basis of a lexicon of synonyms in the German language that is in open access OpenThesaurus (OpenThesaurus German synonym lexicon) and Open Multilingual WordNet English resource, and published in 2017. More about this initiative at Open-de-WordNet, while the thesaurus is available at Open-de-WordNet and can be used under the licence Creative Commons Attribution Share Alike 4.0 International.

13. Wiktionary:Frequency lists, accessed 20.4.2023.

14. spaCy is a free open-source library written in Python and Cython for Named Entity Recognition in English, German, Spanish, Portuguese, French, Italian, Dutch and in multilingual texts.

3. An extractor of multi-word lexical units for a target language (in our research Serbian), *LeXimir*. It is a system for multi-word units extraction for Serbian, based on e-dictionaries and local grammars described in (Stanković et al. 2016). Since this system extracts multi-word units, we separately produced the "Bag of Words"[15] from the merged Serbian texts using Unitex.[16] A resulted list contained 94,802 single-word and 48,159 multi-word units.After that, we performed lemmatization using electronic morphological dictionaries for Serbian (Krstev 2008; Крстев 2019; Stanković et al. 2018) after which duplicates were eliminated and a list of 77,297 different lexical units in the Serbian language was obtained.

Using machine translation tool GIZA++[17] (Och and Ney 2003) (Koehn, Och, and Marc 2003) a list of German-Serbian translation equivalents of multi-word units was produced, which after filtration contained 14,142 potential candidates. These equivalents were manually evaluated and a final list of 3,984 correct equivalents was obtained. The evaluated translation equivalents were further analyzed, synonym candidates for both languages identified and imported into *Termi* database (864 synonyms in Serbian and 791 synonyms in German). Some examples of translation equivalents are given in Table 1, while Table 2 are illustrates some examples of synonyms for both languages.

Source language (German)	Target language (Serbian)	English
Haltestelle	аутобуска станици	bus station
Hoffnungslos	безнадежан случај	hopeless case
Allee	алеја	alley
alt Frau	старица	old woman

Table 1. German-Serbian translation equivalents of lexical units.

---

15. Bag of Words represents a set of words in one text with the calculated frequency of their occurrence in that text independent of their grammatical forms and word order in the sentence.

16. Unitex, accessed 20.4.2023.

17. GIZA++, accessed 20.4.2023.

Lexical units in Serbian	Synonyms in German
потпун (engl. *complete*)	absolut, gänzlich, total, völig, vollkommen, vollständig
базен (engl. *swimming pool*)	Bassin, Becken, Schwimmbad, Schwimmbecken
мајка (engl. *mother*)	Mama, Mutter, Mutti
Lexical units in German	**Synonyms in Serbian**
Augenblick (engl. *moment*)	трен, тренутак момент, час
fremd (engl. *foreign*)	туђ, туђински, непознат
grenzenlos (engl. *limitless*)	безграничан, бескрајан

Table 2. Synonyms of German-Serbian translation equivalents of lexical units.

4 *SrpNemKor* and Linked Open Data

In this section we analyze possibilities of applying principles and guidelines of Semantic Web to the corpus *SrpNemKor*. In Section 4.1 we present the linking of bibliographic metadata with relevant records in reference databases, and in Section 4.2 a bilingual dictionary of general type as a set of linked open data.

4.1 Relating bibliographic metadata

Collection *SrpNemKor* is described with appropriate metadata in digital library *Bibliša*, which is described in more details in (Андоновски 2019, 2021). We will analyze here only the linking of metadata "author's name" and "title of work" with records in three authority files and one general knowledge base that are part of the LOD cloud, namely: Virtual International Authority File – VIAF,[18] Gemeinsame Normdatei – GND,[19] Library of Congress

---

18. Virtual International Authority File – VIAF
19. Gemeinsame Normdatei – GND

Name Authority File – LCNAF[20] and knowledge base Wikidata. In each of the mentioned databases, the records have their own unique identifiers (ID) through which the linking was done. The appropriate identifiers were inserted into *Bibliša* metadata schema and grouped together depending on whether they refer to the author's name or the title of the novel. Figure **??** (left) shows the metadata structure in *Bibliša* for the novel "My Awards" (Meine Preise) by Thomas Bernhard, where we see that the element "Refs" groups the ID numbers of the records in the mentioned databases in the manner and in the order in which they subsequently appear in the user environment. In Figure 1 (right) we see the *Bibliša* panel where next to the name of the author or the title of the work there are links to the paths of the records in the mentioned databases.
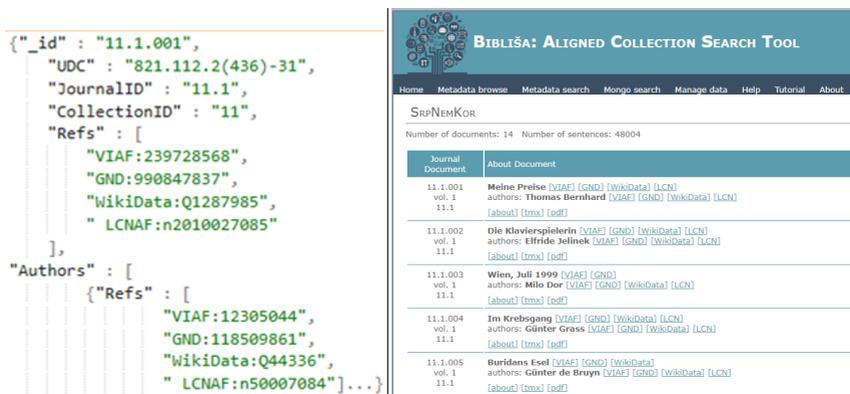
Figure 1. Linking metadata in *Bibliša*'s metadata schema with the records in semantic databases (left); Display of links in the *Bibliša*'s interface with the records in the semantic databases.

4.2 Bilingual dictionary as linked open data dataset

The German-Serbian electronic dictionary of general type, *SrpNemlex*, which we present here as a set of open linked data, was generated by exporting and transforming the list of German-Serbian translation equivalents with

---

20. Library of Congress Name Authority File – LCNAF

synonyms (described in section 3) from the database *Termi*. For preparing dictionary as linked open data set we used Lexicon Model for Ontologies (OntoLex) (Cimiano, McCrae, and Buitelaar 2016; McCrae 2016; Bosque-Gil and Gracia 2019) developed by Ontology-Lexicon community group (OntoLex).[21] The model is based on the RDF format and intended for the description of dictionaries and other linguistic resources containing lexicographic data, in the form of an information ontology on the web (Buitelaar et al. 2011). It follows the principle of *semantics by reference* (McCrae et al. 2012) "in the sense that the semantics of a lexical entry is expressed by reference to an individual, class or property defined in an ontology" (Bosque-Gil and Gracia 2019). The OntoLex model is an upgrade of the model *lemon* (**Gracia2015**; Gracia et al. 2018), which was used as a previous version, and many classes and properties for describing destinations and translation equivalents have been retained from it.

The whole German-Serbian translation equivalents list of lexical units was exported from *Termi* in *OntoLex* model using dotNetRDF,[22] library and the obtained results were presented in Turtle/RDF[23] syntax. In the previous version of OntoLex, *lemon*, three separate files were obtained after export: a file of lexical units (entries) in the source language (in our case German), a file of lexical units (entries) in the target language (in our case Serbian) and a file of translation equivalents (Андоновски 2019). Now, only one file was generated containing a metadata header and three data sets: a data set of entries in German, a data set of entries in Serbian and a data set of translation equivalents.

Entries in both languages are grouped by the *lime:entry* property into separate data sets. All entries in the data sets are arranged in alphabetical order and have an alphabetic identifier consisting of the written form of the entry, information about its Part-Of-Speech and the language. For example, :AUGENBLICK-N-DE (entry from German dataset) or :TREN-N-SR (entry from Serbian dataset) (eng. *moment*). After each set of data, in our example in Serbian and German, each entry is additionally described as follows: a part of speech is described with property *lexinfo:partOfSpeech*, an entry identifier with property *ontolex:Form* and a written form of entry with property *ontolex:writtenRep* with information about a language using its code after @. The property *ontolex:LexicalEntry* indicates that it is an entry that belongs to a certain data set. In both languages, German and

---

21. Ontology-Lexica Community group
22. dotNetRDF
23. Turtle

Serbian, there are some letters with diacritics (in Serbian – č, ć, š, ž, đ, in German – ö, ä, ü) and they are in *SrpNemLex* dictionary represented with codes (web encoding). For example, %C4%8Das – čas (ENG. *moment*), obu%C4%87ar – obućar (ENG. *shoemaker*), %C3%96ffentlichkeit – Öffentlichkeit (eng. *publicity*). Table 3 presents some entry examples in *SrpNemLex* dictionary in RDF/Turtle syntax.

Entry example in German dataset	Explantion
:Augenblick-n-de *lexinfo:partOfSpeech* lexinfo:noun *a ontolex:LexicalEntry;* *ontolex:Form* :Augenblick-x-de-form. :Augenblick-n-de-form *ontolex:writtenRep* "Augenblick"@de.	part of speech – value from Lexinfo ontology entry entry identifier written form of entry
Entry example in Serbian dataset	Explantion
:%C4%8Das-n-sr *lexinfo:partOfSpeech* lexinfo:noun; *a ontolex:LexicalEntry;* *ontolex:Form* :%C4%8Das-n-sr-form. :%C4%8Das-n-sr-form *ontolex:writtenRep* "čas"@sr.	part of speech – value from Lexinfo ontology entry entry identifier written form of entry

Table 3. Entry examples in dictionary SrpNemLex in RDF/Turtle.

During the generation of translation equivalents, each entry in the data sets is assigned an alphanumeric identifier from which the identifier of the translation equivalent is formed, for example, :112383-SR-SENSE-112383-DE-SENSE-TRANS. The translation equivalents are then grouped into a separate data set via the *vartrans:Translation* class, after which each translation equivalent is individually described so that all the entries that are part of it are defined, in our example the entries in German and Serbian. Each entry in the translation equivalent is individually described using the *ontolex:LexicalSense* class as follows: the *skos:definition* property describes the meaning of the entry in the form of a textual definition with information about the language, if exists, the *ontolex:isLexicalizedSenseOf* property specifies the entry in the format as specified in the corresponding language dataset, and the properties *vartrans:source* and *vartrans:target* indicate the datasets of the source and target languages, i.e. the datasets to which the entries from the translation equivalent belong. Table 4 presents one example

of a translation equivalent from the dictionary *SrpNemLex*, while Figure 2 shows a graphic representation of the same translation equivalent.

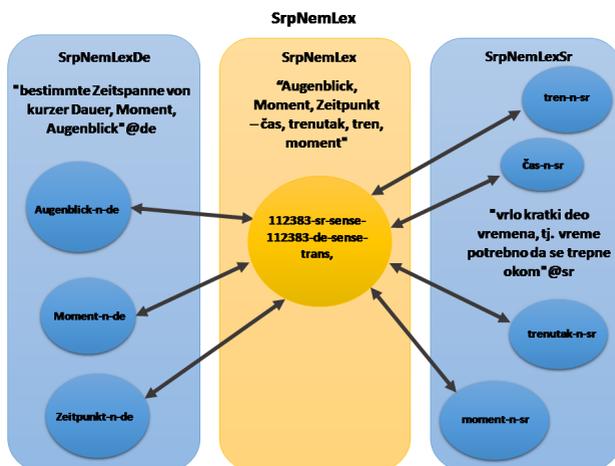| Translation equivalent example	Explanation
:112383-de-sense *a ontolex:LexicalSense;*	entry from German dataset
skos:definition "bestimmte Zeitspanne von kurzer	entry definition
Dauer, Moment, Augenblick"@de;	
ontolex:isLexicalizedSenseOf	entry format in German dataset
:Augenblick-n-de, :Moment-n-de, :Zeitpunkt-n-de	
:112383-sr-sense *a ontolex:LexicalSense;*	entry from Serbian dataset
skos:definition "vrlo kratki deo vremena,	entry definition
tj. vreme potrebno da se trepne okom"@sr;	
ontolex:isLexicalizedSenseOf	entry format in Serbian dataset
:%20tren-n-sr, :%20trenutak-n-sr,	
:%C4%8Das-n-sr, :moment-n-sr.	
:112383-sr-sense-112383-de-sense-trans	translation equivalent –
a vartrans:Translation;	class *vartrans:Translation*
vartrans:source :112383-sr-sense;	source language dataset (German)
vartrans:target :112383-de-sense.	target language dataset (Serbian)

Table 4. An example of a translation equivalent from the dictionary *SrpNemLex* in RDF/Turtle.

In addition to the ontologies provided by the OntoLex model for generating and structuring dictionaries (@prefix ontolex:, @prefix vartrans:, @prefix lime:) it is also possible to use additional ontologies to add more lexicographic explanations such as part of speech, definitions, and the like to dictionary entries and the translation equivalents they form. In our example we used ontologies LexinfoLexinfo[24] and SKOS[25] (@prefix lexinfo: and @prefix skos:). Lexiinfo (Buitelaar et al. 2009) is an ontology that enables the annotation of linguistic properties of entries. It is partially derived from DatCatInfo registry,[26] the Data Category Repository (DCR) developed according to

---

24. Lexinfo
25. Simple Knowledge Organization System
26. DatCatInfo

Figure 2. A graphic representation of a translation equivalent from the dictionary SrpNemLex.

ISO 12620:2019 standard[27] as a replacement for ISOcat metadata registry[28] (Kemps-Snijders et al. 2008). In our example, we used Lexinfo to define entry's part of speech. In the previously mentioned examples, the part of speech of the given entry was defined using the property *lexinfo:partOfSpeech* where its value LEXINFO:NOUN indicated that a word described is a noun.

The second ontology we used is SKOS (Isaac and Summers 2009) for expressing the meaning of the entry in the form of a definition. In the aforementioned examples of translation equivalents, the meaning of the entries was specified through the property *skos:definition*, whose value is a string of characters with the code for the language in which the definition is written. In our examples in table 4 there are definitions in both Serbian and German:

- *skos:definition* "bestimmte Zeitspanne von kurzer Dauer, Moment, Augenblick"@de;
- *skos:definition* "vrlo kratki deo vremena, tj. vreme potrebno da se trepne okom"@sr;.

---

27. ISO 12620:2019

28. The official registry of information about linguistic categories of data on the Web and an important resource for research and development in various linguistic disciplines.

In order for the structure of the dictionary to be complete, it is necessary to create appropriate metadata that provide all the necessary information about the resource. For the purposes of this work, we created a header with basic metadata (short description of the dictionary – *dc:description*, metadata about where the dictionary is currently stored – *dc:source*, title of the dictionary – *dc :title* and authors or collaborators who worked on it – *dct:contributor*), while complete metadata will be created in later research using the LexMeta model[29] for the description of lexical resources (Lindemann, Labropoulou, and Klaes 2022). An example of the current metadata header for the *SrpNemLex* dictionary looks like this:

> *:SrpNemLexTrans dc:description* "This is the lemon version of the bilingual dictionary for German and Serbian languages. Bilingual lemon dictionary contains lexical units with synonyms and was generated from German-Serbian parallel corpus. Corpus is the subject of the doctoral dissertation Linked Open Data and Language Resoruces in preparing Serbian-German Literary Corpus. For German-Serbian lemon bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element."@en;
> *dc:source* <http://llod.jerteh.rs/id/>;
> *dc:title* "German-Serbian lemon Bilingual dictionary";
> *dct:contributor* "Branislava Šandrih", "Cvetana Krstev", "Jelena Andonovski", "Olivera Kitanović", "Ranka Stanković";

5   SrpNemKor in Wikidata

In this section, we will illustrate the representation of one of the novels from the *SrpNemKor* corpus with Wikidata and the search using SPARQL queries with different possibilities for visualizing the obtained results. At the time of the preparation of the *SrpNemKor* corpus, all the authors represented in the corpus were already represented in Wikidata, as well as six novel titles as a literary work. With this research, we started the preparation of Wikidata for other novels from the corpus according to the principles used in the WikiELTeC project[30] (Ikonić Nešić, Stanković, and Rujević 2021) realized on the corpus of Serbian novels SrpELTeC (Trtovac, Milnović, and Krstev

---

29. LexMeta, accessed 9.5.2023.

30. WikiELTeC; Википедија:Википројекат WikiELTeC, accessed 21.4.2023.

2022).[31] SrpELTeC was developed in the scope of the COST action „Distant Reading for European Literary History". [32]

5.1 Wikidata

Wikidata is document-oriented, with items representing topics, concepts, or objects. With a personal account, new items can be created or existing items can be edited in the language selected at the beginning, which defines the default language. Each item in Wikidata has a unique identifier, the so-called permanent identifier (QID), which is assigned automatically when the record is created and cannot be changed later. The identifier is in the form of a positive integer preceded by the capital letter Q, which allows humans and machines to understand the differences between items with the same or similar name but different meaning. The item identifier is associated with the title and description, it can have several aliases and a certain number of statements representing the properties and their values for the item being described. (Станковић and Давидовић 2021). A statement is an ordered triple, RDF triple: subject (item) – predicate (property) – object (value) (Schreiber and Raimond 2014).[33] An item (Q) is a subject of a statement and can be any topic (person, object, place, concept), for example, Q571 (book). Property (P) is a predicate of a statement and expresses some characteristic of the item, for example, P123 (publisher), and value is an object, additional descriptive information of a statement. Statement value can be: a string literal, for example, "1 November 1926", some other item, for example, Q3711 (Belgrade), or URL address (Hernández, Hogan, and Krötzsch 2015; Erxleben et al. 2014).

5.2 *Buridan's ass* in Wikidata

To illustrate an example from our corpus in Wikidata, we have chosen the novel "Buridan's ass" by Gunter de Bruyn. At the beginning of the research, we determined that in Wikidata there are the following items related to this novel: the literary work "Buridanov Magarac: a novel"[34] with which a connection has already been established in *Bibliša* when linking metadata

---

31. srpELTeC: Serbian Literary Text Collection (ELTeC)

32. CA16204 – Distant Reading for European Literary History (DISTANT-READING)

33. RDF 1.1 primer 2014

34. Buridan's ass : novel

(Subsection 4.1), philosophical paradox regarding free will (Q340003)[35], a film "Backhouse Bliss" from 1980 by Herrmann Zschoche based on the novel with same title (Q1532735)[36] and Günter de Bruyn (Q62753)[37] the author of the novel with which a link has already been established in *Bibliša* when linking metadata (Subsection 4.1). According to the experience from the WikiELTeC project, it is necessary that for each novel there is an item described by the statement "is (P31) a literary work (Q7725634)" and an item (or items) described by the statement "is (P31) an edition (Q3331189)", which must be mutually linked via properties "issue" (P747) or "issue from" (P629). The literary work, novel, is connected with editions using property P747 (has edition or translation), while the editions of the novel are connected with a literary work using property P629 (edition or translation of).

As already mentioned, „Buridan's ass: novel" as an item „literary work" already existed in Wikidata, and was described by the following properties:

- "instance of" (P31),
- "title" (P1476),
- "derivative work" (P4969),
- "form of a creative work" (P7937),
- "author" (P50),
- "language of a work or a name" (P407),
- "publication date" (P577),
- "GND ID" (P227).

For the purposes of our research, we have added values for the properties:

- "has edition or translation" (P747),
- "narrative location" (P840),
- "set in period" (P2408) and
- "characters" (P674).

In table 5 we have given a comparative view of statements in Wikidata for items "Buridan's ass: novel" (Q110637016), the literary work, and "Günter de Bruyn" (Q62753), the author, in natural language, their annotation with ID identifiers and in the form of Wikidata statements. In both items, the subject is not repeated, statements with the same subject are separated by a semicolon (";"), while the last statement is followed by a full-stop ".". Also,

---

35. Buridan's ass
36. Backhouse Bliss
37. Günter de Bruynн

we see that the object values are most often other items that have their own QID, except for the properties "date of issue" (P577), "date of birth" (P569) and "date of death" (P570) whose values are literals (respectively: "1968", "November 1, 1926" and "October 4, 2020") and properties P227 whose value is the ID number of the entity in the normative file GND and represents a unique identifier, "4575022-1", that is, "118516388".

Buridan's ass is a literary work written by Günter de Bruyn published 1968 in German. Its GND ID is 45750221.	

Buridan's ass (Q110637016) is (P31) a literary work (Q7725634). written by (P50) Günter de Bruyn (Q62753). published (P577) „1968". in (P407) German (Q108). GND ID (P227) is „4575022-1".	Q110637016
	    P31 Q7725634;
	    P50 Q62753;
	    P577 „1968";
	    P407 Q108;
	    P227 „4575022-1".

Günter de Bruyn was born on 1st November 1926 in Berlin, and he died on 4th October 2020 in Bad Saarow. He was a writer and librarian, member of Academy of Arts, Berlin and German Academy for Language and Literature. He's GND ID is „118516388".	

Günter de Bruyn (Q62753) was born on (P569) "1st November 1926". in (P19) Berlin (Q64). and he died on (P570) "4th October 2020". in (P20) Bad Saarow (Q49421). He (Q62753) was (P106) a writer (Q36180) and a librarian (Q182436), member of (P463) Academy of Arts, Berlin (Q414110). and German Academy for Language and Literature (Q459620). He's (Q62753) GND ID (P227) is „118516388".	Q62753
	    P569 „1st November 1926";
	    P19 Q64;
	    P570 „4th October 2020";
	    P20 Q49421;
	    P106 Q36180;
	Q182436;
	    P463 Q414110;
	Q459620;
	    P227 „118516388".

Table 5. Comparative representation of statements in natural language, using identifiers and in the form of Wikidata statements.

In addition to the "literary work" item, we also created the "edition" item. In our research, we are talking about the German-Serbian parallel corpus, which was created by parallelizing texts in German and Serbian. In our example, the corpus edition of the novel "Buridan's ass" was created on the basis of the edition in German (in our example from 1992) and the edition in Serbian (in our example from 2007). For this reason, a representation of Wikidata was made for three editions of the item "Buridan's ass: novel" (Q110637016): Buridan's ass: Novel, SrpNemKor Edition (Q117220537),[38]

---

38. Buridan's ass : Novel : SrpNemKor Edition

Figure 3. Property "has edition or translation" (P747) in item "Buridan's ass: novel" (Q110637016).

and Buridan's ass: 1992 Edition (Q117221046)[39]and Buridan's ass: 2007 Edition (Q117220339).[40]

For each "edition" item, we used the following properties to describe it:

- "instance of" (P31),
- "title" (P1476),
- "form of creative work" (P7937),
- "edition or translation of" (P629),
- "author" (P50),
- "publisher"(P123) (for items Q117221046 and Q117221046)
- "place of publication" (P291),
- "language of work or name" (P407),
- "publication date" (P577),

---

39. Buridan's ass : 1992 Edition)
40. Buridan's ass : 2007 Edition

- "set in period" (P2408),
- "narrative location" (P840),
- "number of pages" (P1104),
- "published in" (P1433).

The items Q117221046 and Q117220339 are additionally described with property "publisher" (P123), while item Q117220537 is additionally described with properties "full work available at URL" (P953) and "number of words" (P6570).

Item Q110637016 is linked to items Q117220537, Q117221046 и Q117220339 using property P747 ("has edition or translation", Figure 3), while they are linked with the same item Q110637016 using property P629 ("edition or translation of"). The item Q117220537 is additionally connected with items Q117221046 and Q117220339 using property P629 (Figure 4).
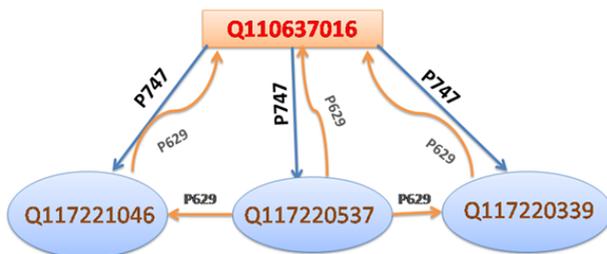
Figure 4. "Has edition or translation" (P747) and "edition or translation of" (P629) properties in Wikidata for the novel *Buridan's ass.*

Qualifiers and references can be assigned to object values that provide additional information about the object value itself. They are also in the form of an ordered RDF triple. For example, in the Figure **??** we see the qualifiers next to the values of the "has edition or translation" property (P747) and these are: "language of the work or name", "publication date" and "place of publication". In this example, the mentioned qualifiers provide users with more detailed information about the editions of the selected literary work.

References can take the form of a classic bibliographic citation, or they can be web addresses and referenced datasets, each of which can be a Wikipedia entry. In our example, we will see the references already in the

first statement of the item "Ginter de Bruyn" (Q62753) where the following references are listed: one indicates as the download source Wikipedia in Russian, and the other indicates BnF (Bibliothèque nationale de France) authorities[41] with information: "retrieved" (10 October 2015) and "reference URL" (`http://data.bnf.fr/ark:/12148/cb118858431`) (Figure 5).

Figure 5. Reference examples in the item "Günter de Bruyn" (Q62753)
.

5.3 SPARQL queries

It is possible to search metadata in Wikidata user environement by SPARQL queries[42] and to help users, there are numerous query examples that can be used so that users who are not familiar with the SPARQL language can satisfy their information needs.[43] Query results can be visualized in a various ways: table, image grid, graph builder, map, line chart, bar chart, scatter chart, area chart, bubble chart, tree map, tree, timeline, dimensions or graph. We will illustrate the use of SPARQL queries with two examples. First example lists all editions of the novel *Buridan's ass* with the following metadata:

---

41. BnF authorities
42. Simple Protocol and RDF Query Language
43. Wikidata Query Service

author, place of publication, publication date and form of a creative work.[44]
We presented the results in two ways: as a table (Figure 6)[45] and as a graph
(Figure 7).[46]

First example:

```
#defaultView:Table #defaultView:Graph
SELECT ?edition ?editionLabel ?author ?authorLabel ?pubPlace ?pubPlaceLabel
?publicationDate ?workForm ?workFormLabel
WHERE { #edition (P629) of Buridan's Ass (Q110637016)
    ?edition wdt:P629 wd:Q110637016;
        #author (P50) Günter de Bruyn(Q62753)
        wdt:P50 ?author;
        #place of publication
        wdt:P291 ?pubPlace;
        #publication date
        wdt:P577 ?publicationDate;
        #work form
        wdt:P7937 ?workForm.
SERVICE wikibase:label
{bd:serviceParam wikibase:language "en,[AUTO_LANGUAGE],en".} }
LIMIT 100
```

edition	editionLabel	author	authorLabel	pub_Place	pub_PlaceLabel	publicationDate	workForm	workFormLabel
wd:Q117220339	Buridan's ass : 2007 Edition	wd:Q62753	Günter de Bruyn	wd:Q3711	Belgrade	January 1, 2007	wd:Q8261	novel
wd:Q117220537	Buridan's ass : Novel : SrpNemKor Edition	wd:Q62753	Günter de Bruyn	wd:Q3711	Belgrade	January 1, 2019	wd:Q461183	text corpus
wd:Q117221046	Buridan's ass : 1992 Edition	wd:Q62753	Günter de Bruyn	wd:Q1794	Frankfurt am Main	January 1, 1992	wd:Q8261	novel

Figure 6. Visualization of the first SPARQL query in the form of a table.

The second example lists all characters from the novel *Buridan's ass* with
the following metadata: title and author.[47] As a default view we choose tree
(Figure 8).[48]

Visualization of SPARQL queries as a graph enables not only to see
statements (item-property-value) that are results of a particular query, but
also it is possible to see all set of statements and their links that exist in the
obtained result. Figure 9 shows the graph of the novel *Buridan's ass.*

44. SPARQL query, first example
45. First SPARQL query visualization as a table
46. First SPARQL query visualization as a graph
47. SPARQL query, second example
48. Second SPARQL query visualization as a tree
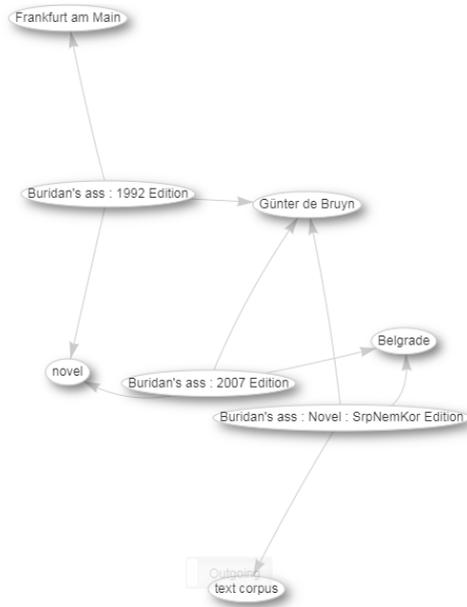
Figure 7. Visualization of the first SPARQL query in the form of an interactive graph.

Second example:

```
#defaultView:Tree
SELECT ?characters ?charactersLabel ?gender ?genderLabel ?novel ?novelLabel
?creator ?creatorLabel
WHERE { #characters (P1441) in Buridan's ass (Q110637016)
    ?characters wdt:P1441 wd:Q110637016;
        #gender (P21)
        wdt:P21 ?gender;
        #present in work Buridan's ass
        wdt:P1441 ?novel;
        #creator (P170) Günter de Bruyn
        wdt:P170 ?creator.
SERVICE wikibase:label
{bd:serviceParam wikibase:language "en,[AUTO_LANGUAGE],en".} }
LIMIT 100
```

Figure 8. Visualization of the first SPARQL query in the form of a tree.
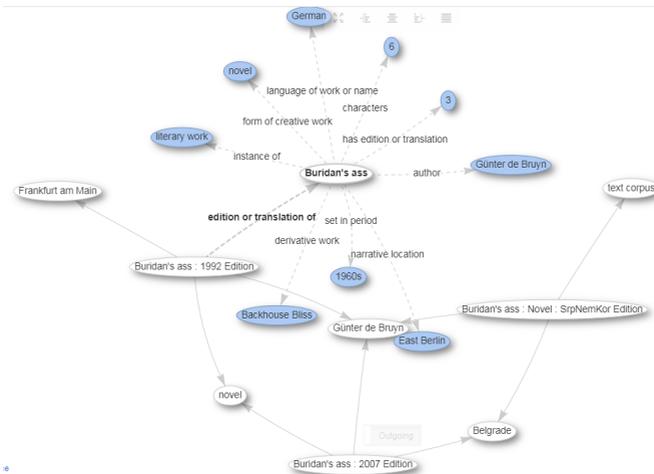
Figure 9. The graph of an item "Buridan's ass: novel" (Q110637016)
.

6 Conclusion and further work

In the paper, we demonstrated how certain principles and guidelines of the Semantic Web can be applied to a parallel corpus of literary texts, some of which are written in Serbian. Linking metadata from a corpus to records in relevant semantic databases from the LOD cloud, generating a bilingual dictionary as a set of linked open data based on the corpus content or illustrating metadata from the corpus with Wikidata are just some of the ways to display the content of a parallel corpus in a wide network of linked open data.

In addition to the research results presented in this paper, there are more possibilities of applying the Semantic Web to a linguistic resource presented in this paper. We will continue to work on the developed *SrpNemlex* dictionary, primarily in order to create of a more detailed header with metadata about the dictionary itself and to publish it in the LOD cloud. Moreover, we will research and analyse the frequency of occurrences of terms from the dictionary in the *SrpNemKor* corpus and the Corpus of Contemporary Serbian SrpKor13.[49].

---

49. Корпус савременог српског језика

We also plan to continue to import to Wikidata metadata for remaining novels from *SrpNemKor* relying on the experience obtained from the novel "Buridanov's Donkey" and applying the automatic metadata loading method that was used for the srpELTeC corpus. In addition, the annotation of the named entities in the corpus and the possibilities of their linking with the appropriate knowledge bases will be analyzed. In this way the content of the corpus will become part of a wider environment that enables easier information retrieval and different ways of displaying search results. The possibility to access content in the Serbian language from the wider LOD cloud will contribute to a better representation of the Serbian language on the Web.

References

Andonovski, Jelena, Šandrih, and Olivera Kitanović. 2019. "Bilingual lexical extraction based on word alignment for improving corpus search." *The Electronic Library* 37 (4): 722–739.

Bosque-Gil, Julia, and Jorge et al. Gracia, eds. 2019. *The OntoLex Lemon Lexicography Module: Final Community Group Report 13 September 2019.* Accessed 13.3.2023, https://www.w3.org/2019/09/lexicog/. W3C.

Buitelaar, Paul, Philipp Cimiano, Peter Haase, and Michael Sintek. 2009. "Towards linguistically grounded ontologies." In *ESWC2009, LNCS 5554, Towards linguistically grounded ontologies,* edited by L. Aroyo et al., 111–125. Preuzeto 14.03.2019, https://link.springer.com/content/pdf/10.1007/978-3-642-02121-3_12.pdf. Berlin, Heidelberg: Springer.

Buitelaar, Paul, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. 2011. "Ontology lexicalisation: The *lemon* perspective." In *WS 2 Workshop Extended Abstracts, 9th Int. Conf. on Terminology and Artificial Intelligence, TIA 2011, Paris, 10 November 2011,* 33–36. http://oa.upm.es/9772/1/Ontology_Lexicalisation.pdf.

Chiarcos, Christian, Bettina Klimek, Christian Fäth, Thierry Declerck, and John Philip McCrae. 2020. "On the Linguistic Linked Open Data Infrastructure." In *Proc. of the 1st International Workshop on Language Technology Platforms,* 8–15. Marseille, France: ELRA. ISBN: 979-10-95546-64-1. https://aclanthology.org/2020.iwltp-1.2.

Chiarcos, Christian, and Antonio Pareja-Lora. 2019. "Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction." In *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, 1–17. The MIT Press.

Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020b. *Linguistic Linked Data*. Springer International Publishing.

Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020a. "Linguistic Linked Open Data Cloud." In *Linguistic Linked Data: Representation, Generation and Applications*, 29–41. Springer, Cham. https://doi.org/10.1007/978-3-030-30225-2_3.

Cimiano, Philipp, John P. McCrae, and Paul Buitelaar. 2016. *Lexicon Model for Ontologies: Community Report, 10 May 2016: Final Community Group Report 10 May 2016*. Preuzeto 13.3.2023, https://www.w3.org/2016/05/ontolex/. W3C.

Erxleben, Fredo, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. "Introducing Wikidata to the Linked Data Web." In *The Int. Semantic Web Conf.– ISWC 2014*, edited by Peter et al. Mika, 50–65. Cham: Springer International Publishing. ISBN: 978-3-319-11964-9.

Gracia, Jorge, Marta Villegas, Asunción Gómez-Pérez, and Núria Bel. 2018. "The apertium bilingual dictionaries on the web of data." *Semantic Web* 9 (2): 231–240.

Hernández, Daniel, Aidan Hogan, and Markus Krötzsch. 2015. "Reifying RDF: What Works Well With Wikidata?" In *11th Int. Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2015) At the 14th Int. Semantic Web Conf. (ISWC2015)*, 32–47.

Ikonić Nešić, Milica, Ranka Stanković, and Biljana Rujević. 2021. "Serbian ELTeC Sub-Collection in Wikidata." *Infotheca* 21 (2): 60–87. https://doi.org/10.18485/infotheca.2021.21.2.4.

Isaac, Antoine, and Ed Summers. 2009. *Simple Knowledge Organization System Primer*. Working Group Note, W3C. Preuzeto 14.3.2023, https://travesia.mcu.es/server/api/core/bitstreams/3f55d0de-b342-4d5d-8016-013b52e81707/content.

Janik, Maciej, Ansgar Scherp, and Steffen Staab. 2011. "The Semantic Web: Collective Intelligence on the Web." Preuzeto 10.3.2023, http://www.ansgarscherp.net/publications/pdf/J07-JanikScherpStaab-TheSemanticWeb-Preprint.pdf, *Informatik-Spektrum* 34 (5): 469–483.

Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. "ISOcat: Corralling data categories in the wild." In *Proc. of the 6th Int. Conf. LREC'08,* 887–891. Preuzeto 15.03.2019, http://www.lrec-conf.org/proceedings/lrec2008/pdf/222_paper.pdf. ELRA.

Koehn, Philipp, Franz Josef Och, and Daniel Marc. 2003. "Statistical phrase-based translation." In *Proc. of the 2003 Conf. of the North American Chapter of the ACL on Human Language Technology-Volume 1,* 1:48–54. Association for Computational Linguistics.

Krstev, Cvetana. 2008. *Processing of Serbian – Automata, Texts and Electronic dictionaries.* Belgrade: Faculty of Philology, University of Belgrade,

Krstev, Cvetana, Branislava Sandrih, Ranka Stanković, and Miljana Mladenović. 2018. "Using English Baits to Catch Serbian Multi-Word Terminology." In *Proc. of the 11th Int. Conf. LREC 2018,* edited by Nicoletta Calzolari et al., 2487–2494. Preuzeto 15.03.2019, http://www.lrec-conf.org/proceedings/lrec2008/pdf/222_paper.pdf. ELRA.

Lindemann, David, Penny Labropoulou, and Christiane Klaes. 2022. "Introducing LexMeta: a metadata model for lexical resources." In *Dictionaries and Society. Proc. of the XX EURALEX Int. Congress, 12–16 July 2022, Mannheim, Germany,* 310–320. IDS-Verlag.

McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, and Jorge et al. Gracia. 2012. "Interchanging lexical resources on the Semantic Web." Preuzeto 14.03.2023, https://link.springer.com/content/pdf/10.1007%2Fs10579-012-9182-3.pdf, *Language Resources and Evaluation* 46 (4): 701–7019.

McCrae, John Philip et al. 2016. "The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud." In *Proc. of the 10th Int. Conf. LREC'16,* 2435–2441. ELRA. https://doi.org/10.5167/uzh-132353.

Och, Franz Josef, and Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models." *Computational linguistics* 29 (1): 19–51.

Open DE WordNet Initiative. 2023. Pristupljeno 13.3.2023, https://ikum.mediencampus.h-da.de/projekt/open-de-wordnet-initiative/.

Radulovic, Filip, María Poveda-Villalón, Daniel Vila-Suero, Víctor Rodríguez-Doncel, Raúl García-Castro, and Asunción Gómez-Pérez. 2015. "Guidelines for Linked Data generation and publication: An example in building energy consumption." *Automation in Construction,* no. 57, 178–187.

Rurarz, Daniel, Aneta Poniszewska-Maranda, Iryna Ivanochko, and Oksana Urikova. 2021. "Use of Semantic Networks to Search for Information." In *Developments in Information & Knowledge Management for Business Applications,* 1:93–199. Springer, Cham. https://doi.org/10.1007/978-3-030-62151-3_3.

Schreiber, Guus, and Yves Raimond. 2014. *RDF 1.1 Primer.* W3C Note. Https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/. W3C, June.

Shah, Urvi, Tim Finin, Anupam Joshi, R Scott Cost, and James Matfield. 2002. "Information Retrieval on the Semantic Web." In *Proceedings of the eleventh international conference on Information and knowledge management,* 1:461–468. https://doi.org/10.1145/584792.584868.

Stanković, Ranka, and Lazar Davidović. 2021. "Infotheca (Q25460443) in Wikidata." *Infotheca - Journal for Digital Humanities* 21 (1): 87–98. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.1.5. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.1.5_en.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries – from File System to *lemon* Based Lexical Database." In *Proc. of the 11th Int. Conf. LREC 2018 - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018), Miyazaki, Japan, May 7-12, 2018,* edited by John P. et al McCrae. ELRA. http://lrec-conf.org/workshops/lrec2018/W23/pdf/3_W23.pdf.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Dalibor Vorkapić. 2015. "A bilingual digital library for academic and entrepreneurial knowledge management." In *Proc. of 10th International Forum on Knowledge Asset Dynamics-IFKAD,* 1764–1777. Preuzeto 13.03.2023, http://poincare.matf.bg.ac.rs/\unhbox\voidb@x\protect\penalty\@M\cvetana/biblio/IFKAD_RS_CK_BL_DV-fin.pdf.

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. "Rule-based Automatic Multi-word Term Extraction and Lemmatization." In *Proc. of the 10th Int. Conf. LREC'16),* 507–514. ELRA. https://aclanthology.org/L16-1081.

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, and Miloš Utvić. 2012. "A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals." In *Proc. of the 8th Int. Conf. LREC 2012,* edited by Nicoletta et al. Calzolari, 1710–1717. Preuzeto 16.04.2016, http://www.lrec-conf.org/proceedings/lrec2012/pdf/375_Paper.pdf. Istanbul: ELRA.

Stanković, Ranka, 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,* edited by Andrea Calì, Dorian Gorgan, and Martín Ugarte, 112–123. Springer International Publishing. ISBN: 978-3-319-53640-8.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2022. "The Serbian Part of the ELTeC – from the Empty List to the 100 Novels Collection." *Infotheca* 21 (2): 7–25. https://doi.org/10.18485/infotheca.2021.21.2.1.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: a free collaborative knowledgebase." *Communications of the ACM* 57 (10): 78–85. https://doi.org/10.1145/2629489.

Андоновски, Јелена С. 2019. "Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса." Докторска дисертација, Универзитет у Београду, Филолошки факултет.

Андоновски, Јелена С. 2021. *Паралелни српско-немачки корпус књижевних текстова: израда, проналажење информација и семантички веб.* Београд: Универзитетска библиотека „Светозар Марковић“.

Крстев, Цветана. 2019. "О одабиру одредница за електронски речник српског језика и њиховом повезивању." In *Научни састанак слависта у Вукове дане,* 48:133–147. Београд : Филолошки факултет, Међународни славистички центар. https://doi.org/10. 18485/msc.2019.48.3.ch7.

Станковић, Ранка, and Лазар Давидовић. 2021. "Инфотека (Q25460443) у Википодацима." Accessed 17.3.2023, https://infoteka.bg.ac.rs/ojs/ index.php/Infoteka/article/view/2021.21.1.5_sr, *Инфотека* 21 (1): 93–104.