

Impressum

FOR THE EDITOR:

Prof. Aleksandar Jerkov, PhD

University Library "Svetozar Marković"

Faculty of Philology, University of Belgrade

office@unilib.rs

EDITOR:

Faculty of Philology, University of Belgrade

University Library "Svetozar Marković"

Serbian Academic Library Association

EDITOR IN-CHIEF:

Prof. Cvetana Krstev

cvetana@matf.bg.ac.rs

MANAGING EDITOR:

Aleksandra Trtovac, PhD

University Library "Svetozar Marković"

aleksandra@unilib.rs

EDITOR OF ONLINE EDITION:

Jelena Gavrilović, PhD

University Library "Svetozar Marković"

andonovski@unilib.rs

EDITORIAL BOARD:

Prof. Aleksandra Vraneš, PhD, Prof. Aleksandar Jerkov, PhD, Prof. Cvetana Krstev, PhD, Prof. Biljana Dojčinović, PhD, *Faculty of Philology, University of Belgrade*; Prof. Elisabeth Burr, PhD, *Institut für Romanistik, Universität Leipzig*; Prof. Vladan Devedžić, PhD, *Faculty of Organization Sciences, University of Belgrade*; prof. Milena Dobрева, PhD, *Faculty of Media and Knowledge Sciences, University of Malta*; Tomaž Erjavec, PhD, *Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana*; Prof. Svetla Koeva, PhD, *Institute for Bulgarian Language, Bulgarian Academy of Sciences*; Prof. Denis Maurel, PhD, prof. Agata Savary, PhD, *Université Francois Rabelais de Tours*; Prof. Ivan Obradović, PhD, *Faculty of Mining and Geology, University of Belgrade*; Prof. Gordana Pavlović Lažetić, PhD, prof. Duško Vitas, PhD, *Faculty of Mathematics, University of Belgrade*; Prof. Katerina Zdravkova, PhD, *Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje*

ISSN 1450-9687 (print edition)
ISSN 2217-9461 (online edition)

Belgrade, Vol. 23, No. 1, July 2023

WEB PORTAL:

Jelena Gavrilović, PhD

University Library "Svetozar Marković"

DESIGN AND PREPRESS:

Infotheca Team

REDACTOR OF REFERENCES AND UDC:

Nataša Dakić, PhD

University Library "Svetozar Marković"

DOI REDACTOR:

Miloš Utvić, PhD

Faculty of Philology, University of Belgrade

JOURNAL REDACTION:

Journal Infotheca

11000 Belgrade, Bulevar kralja Aleksandra 71

+381 11 3370-211

infothecajdh@gmail.com

PRINTED BY:

Mamigo plus

Belgrade

Printing of this issue of Infotheca was supported by the Association for Language Resources and Technologies JeRTeh and Wikimedia Serbia through the project "Wikidata in Serbian novels (input, visualization, and analysis)"

Journal is published twice a year

Contents

Scientific papers

The Language of Food and Its Dictionary Duško Vitas	7
SrpNemKor and Linked Open Data Jelena Andonovski	33
It-Sr-NER: Web Services for Recognizing and Linking Named Entities in Text and Displaying Them on a Web Map Olja Perišić, Ranka Stanković, Milica Ikonić Nešić, Mihailo Škorić	61
Automatic Assessment of Short Answers Using Latent Semantic Analysis Teodora Mihajlov	78
The Chicago Akkadian Dictionary: Contexts and Perspectives Marko Vitas	107

Professional papers

Fourth Summer Datathon on Linguistic Linked Open Data Tijana Radović, Ranka Stanković	117
---	-----

Reviews

Theoretical Basis and Creation of a Parallel Corpus: Delineating New Perspectives in the Processing of Library Material Dragana Stolić	125
The Fight for Literacy Begins in “Studenjak”: the Promotion of the Publication Introduction to Information Literacy Anđelija Lazić	129

The Language of Food and Its Dictionary

UDC 030:004.738.5]:023-051

DOI 10.18485/infotheca.2023.23.1.1

ABSTRACT: A dictionary of the language of food, as a dictionary of a special sublanguage in the Serbian language, does not exist to this day. Too ordinary or too prone to changes, the description of the lexicon of this sublanguage escapes the attention of language researchers, and confuses lexicographers. The paper analyzes examples of lexicographic processing of individual lexemes of this language and shows that the description of the culinary lexicon cannot be solved by standard lexicographic procedures. A digital corpus consisting of different types of texts describing the language of food is presented, including cookbooks, ethnographic and historical studies, literary works, etc. It is shown that the language of food evolves rapidly and continuously, and its evolution is closely related to changes in society. Finally, the author advocates the creation of an encyclopaedic dictionary of the language of food and indicates its basic properties.

KEYWORDS: language of food, domain corpus, lexicography.

PAPER SUBMITTED: 18 April 2023

PAPER ACCEPTED: 22 May 2023

Duško Vitas

vitas@matf.bg.ac.rs

University of Belgrade

Faculty of Mathematics

Belgrade, Serbia

Translated by: Jelena D. Bajić

1 Introduction

The language used when speaking about food is so quotidian, so ordinary that almost everything that is said becomes immediately understandable to everyone. Such a view relies on the fact that a plate contains, except in extraordinary circumstances, the usual dish, whose ingredients are well-known, forming part of the “traditional” cuisine. It is likely that these are the reasons why the investigation of the language of food and especially its

temporal, spatial, cultural and social evolution has been largely overlooked in our midst.

This language, or more precisely sublanguage in Harris' sense (Kittredge and Lehrberger 1982), at first glance features a limited lexicon and a narrow choice of syntactic constructions, which intensifies the impression that its nature is simple. However, unlike numerous other sublanguages, the language of food is constantly changing, above all, under new cultural influences and changes in food fashions. On the other hand, despite all of its simplicity, the transfer of culinary content from one language to another is exceptionally complex, since the conceptualization of food differs from language to language, not only in the structure of the dishes that get prepared, but also in what is considered to be acceptable (and customary) choice of ingredients and dishes. This transfer also involves a constant process of naming the foodstuffs and dishes adopted from other places (Радојичић 2015).

In order for something to become food, it must have (acquire) its linguistic expression. Each element constituting this language, starting from the ingredients, moving to the ways of their preparation and ending in the prepared dish has its own name, thus naming and differentiating items in the language of food in a way precedes the food itself.¹ Let us note immediately that the lexical elements of this language are, as a rule, abstract notions in the sense that they are devoid of their features. As we will see, lexicography is faced with this nature of the lexis of food when a culinary term needs to be defined. For example, the lexeme *tomato*, when it appears in a recipe, denotes just a general concept and in most cases does not provide precise information as to the kind of tomato. However, this fruit can have additional attributes which will yield different results when preparing food and that aspect evades lexicographers' attention. A lexicographic description will not, except in extraordinary circumstances specify the kind of tomato – JABUČAR (*apple-like tomato*), VOLOVSKO SRCE (*beef tomato*), ČERI (*cherry tomato*), etc. – or the meaning of the expression *tubed tomato paste*, *peeled tomatoes*, *tomato soup*, *tomato broth*, etc. A description of the relation between the meaning of these words and their usage (at any given moment) relies on the collective notions related to the meaning of the concept *tomato*.

In this paper, we will be interested in the issue of evolution of the language of food in Serbian ever since the appearance of the first cookbooks to the

1. An example of the influx of unnamed foodstuffs and dishes is illustrated in (Радојичић 2015, 104) where a case of importing “30 spices totally unknown as far as our market is concerned” that might get a name and be accepted in the local cuisine is quoted.

present day. By defining the issue in question in such a way, the description of the language of food gets narrowed down, primarily, to townspeople's cuisine in the regions populated by Serbs, as illustrated by cooking handbooks. Beyond that framework, countryside eating habits that are believed to follow tradition and customs, in terms of food preparation, experience changes in an indirect and belated manner. These changes, coming from the urban cuisine have to do both with the technology of preparing food and the range of foodstuffs and dishes. This claim can be supported by the depictions of the contemporary eating habits in rural areas (Радуловачки 1996; Лазих 2000; Радованових 2011) that attest to the integration of foodstuffs, dishes and cooking techniques present in antique cooking handbooks that older ethnographic sources (Тројанових 1983; Ердељанових 1908; Ђорђевић 1928) were not familiar with, though.

Cookbooks in Serbian, often featuring words such as “Serbian” or “national” in their titles resulted, above all, from the attempts to express in that language culinary notions coming from the areas with developed cooking skills and traditions. Thus, Midžina's handbook Поповић-Мицина (1891) named VELIKI SRPSKI KUVAR (*Big Serbian Cookbook*),² provides primarily a description of the Viennese cuisine, while VELIKI NARODNI KUVAR (*Big National Cookbook*) by Aleksandra Рустановић (1938) endeavours to evoke dishes belonging to the French and Russian cuisine in the numerous recipes in Serbian. It is clear that texts of this kind are faced with the problem of naming new and unfamiliar concepts referring either to ingredients or prepared meals. In that sense, the description of the language of food bears witness to the process of modernization and internationalization of the domestic cuisine. By comparing the content of cooking handbooks with the range of fast or convenience food on offer in Belgrade in early 21st century (Витас 2018) we can see that the contemporary range of “fast food” dishes, which are nowadays viewed as being ordinary, sometimes even traditional were unfamiliar in the first half of the 20th century and later, as well.

2. The 3rd edition of Midžina's *Cookbook* consists of two parts. The first part (up to the page 367) is the same as the 2nd edition, while the second part titled *Supplement* includes contributions no longer authored by Midžina herself. The references in this paper mentioning the nd edition relate to the first part of the rd edition, in fact.

Such nature of the language of food was recognized as early as in *Big Serbian Cookbook* by Katarina Popović Midžina, who was quick to make the following point in the introduction to the first edition:³

Let's say a few words now about "artificial expressions" too, namely, foreign words that we must use out of necessity in this book of ours. Since, we Serbs inherited very few dishes and desserts, no one will hold it against us or find fault with us if we use German, French and foreign names in general. However, we will also keep those names that were passed on by our ancestors and remained unscathed in the process.

Today's generation is accustomed to affluence and luxury and keep it up in terms of dishes too. There are thousands of new dishes for which we have no names, despite the fact that we have come pretty long way as far as books are concerned. A good and comprehensive Serbian cooking dictionary that we could consult as the need arises is still missing (Поповић-Мицина 1891, 11).

The "artificial expressions" mentioned here refer to the neologisms coined by Midžina in order to render into Serbian primarily recipes belonging to the Viennese cuisine. The text of this book itself provides an indication as to what could be featured in a "good and comprehensive Serbian cooking dictionary". Striving to describe her recipes, Midžina resorts to giving synonyms, or foreign equivalents, so as to explain to the users what the recipes are about, like in the examples below:⁴

- PODAJ NA TRPEZU SA ZEJTINOM — ULJEM I SIRČETOM — ОСТОМ⁵ (*Put it on the table with oil — oil and vinager - vinager*) (44);
- BELA BUNDEVA ILI KORABA — КЕЈЕРАБА (*White squash or turnip cabbage — kohlrabi*) (360);
- АРАПСКА SMOLA - GUMIARABIKUM (*Arabian resin — gum arabic*) (357);
- O UKRASIMA ZA GOVEDINU. Rindfleisch – Garnirung (*On Beef garnish. Rindfleisch – Garnirung*) (73).

3. This foreword has been quoted verbatim in all later editions of this cooking handbook.

4. The examples in this paper have been extracted by using the Unitex system (Paumier et al. 2021) ([Unitex/GramLab – Open Source Corpus Processing Suite](#)) over the digital version of the source in question.

5. *Zejtin* and *ulje*, on the one hand and *sirće* and *ocat* on the other are synonyms, meaning oil and vinegar respectively.

The phenomenon of the evolution of the language of cooking is discussed in the 1911 fourth edition of this *Cookbook* (Поповић-Миџина 1911, 367), as well, where the publisher points out:

The needs of a contemporary household require that it too should be recast in a more modern guise, and in order to do it, we have purchased permanent publishing rights to this book from the legal heirs of the late Katarina Popović-Midžina, so as to be able to edit this useful local book independently, providing it with a direction followed by the forward-looking housewives today.

Less than 20 years after the appearance of the second edition, at the beginning of the 20th century, the new publisher thoroughly changed and modernized the content of this book. Therefore, even at the time when “culinary fashions” were not placed under pressure of globalization and the internet, cooking and the related culinary products had been subjected to rapid changes. An insight into the scope of the changes can be obtained by comparing the 3rd and 4th editions (Витас 2022), namely, there are new recipes for the dishes having the same name, some dishes are “outdated” and are no longer featured in the new edition, while there are also some new hereto unknown (or undescribed) ones.

A century and a half after the first edition of Midžina’s Cookbook, the language of food still awaits its own “good and comprehensive” dictionary. However, a dictionary of this kind must not be a mere listing of the vocabulary used in the domain of cooking. Instead, it should also provide a description of meaning that in addition to the contemporary and “usual” meaning includes its evolution from the moment of appearance of the first cooking handbooks. The descriptions of the lexis of cooking in the local lexicographic practice and its use in the wider context of talking about food i (Section 2) will be the first to be examined in this light. In order to make these different descriptions of the language of food the topic of an analysis relying on the contemporary information technology methods, we will provide a brief outline of the corpus consisting of Serbian cooking handbooks published from the mid-19th century to the present, as well as other texts depicting our eating habits and highlight the issues related to corpus organization and exploitation (Section 3). We will wrap up (Section 4) by providing a framework for building an (encyclopaedic) dictionary of cooking and outline the directions of further research (Section 5).

2 The language of food in dictionaries

The lexis of the language of food has been described primarily in the large dictionaries of Serbian (RSANU 1959–2019; RMS 1976; RSJ 2007). These dictionaries provide, for the most part, depending on the time of publishing, a description of an inventory of Serbian entries playing part in expressing cooking-related content. But the impact of that quiet evolution of the language of food results in the lexemes described in them not featuring new concepts and new meanings.

The qualifier KUV(ANJE, KULINARSTVO) (*cooking, culinary art*) has been proposed for the purpose of describing the lexis of the language of cooking, but as it has been shown in (СТИЈОВИЋ, СТАНКОВИЋ, and САОБО 2017; Радовановић 2017), assigning this qualifier in general dictionaries creates a number of problems and dilemmas. A search for the entries marked with this qualifier yields only a small number of them, e.g. АСПИК (*aspic*), МАРИНАДА (*marinade*), МЕДУЈЕЛО (*relevé*), ГРЕНАДИРМАРШ (*Grenadier March*), КРОМПИРАЧА (*potato pie*)..., mostly originating from foreign languages or regional naming of dishes. The description of the entries that are indispensable, a widely known fixture of the language of cooking does not include this qualifier. For example, the entry ГОВЕДИНА (*beef* in (RMS 1976) is described as *beef meat* without the qualifier KUV,⁶ while the entry РИНФЛАЈШ (*beef garnish*) in the same dictionary is JELO OD GOVEDEG MESA, КУВАНА ГОВЕДИНА (*beef dish, cooked beef*) having the qualifier KUV. Similarly, the entry МУСАКА (*moussaka*) does not feature the qualifier KUV in (RSANU 1959–2019), despite being described as *a dish made of chopped meat, potatoes or other vegetables and eggs browned in the oven*. Separating part of the wealth of lexical information in Serbian defining the concepts relevant to the language of cooking is possible by using other qualifiers, such as АГР(и)culture), БОТ(any), etc. (СТИЈОВИЋ, СТАНКОВИЋ, and САОБО 2017). This heuristic procedure, however, yields an approximate result only, as shown by the following example. Fruit is described as a botanical notion, while different varieties of a fruit are classified as an agronomic one in (RSANU 1959–2019). So, the entry ЈАБУКА (*apple*), namely, its basic meaning gets described in the same source as a botanical term:

6. The same entry, ГОВЕДИНА (*beef*) is defined in (RSANU 1959–2019) as ГОВЕДЕ МЕСО (*beef meat*), which is considered to be its basic meaning, while all examples refer to *cooked beef*. RSJ (2007) provides the synonym ГОВЕДИНА, as well.

JABUKA ž 1. BOT. vočka Malus iz f. Rosaceae sa mnogo vrsta, koja rađa mesnate, okruglaste za jelo ukusne plodove; plod te biljke. (*apple f. 1. bot. fruit from the Malus family with many varieties having succulent, roundish, edible and tasty fruits; fruit of that plant.*)

However, an apple variety – KOŽARA (lit. *leathery (skinned)*) – gets described as an agronomic term:

KOŽARA ž 2. AGR. a. vrsta jabuke sa debelom korom. (*kožara f. 2. agr. a. an apple variety with thick skin.*)

Note that this definition of *kožara* is ambiguous: if in the definition of KOŽARA, apple gets replaced by the above definition of apple, thick skin can refer both to the apple tree and its fruit. On the other hand, the qualifiers AGR, BOT and the like are also used when describing entries not belonging to the language of food, since their referents are not consumed as food, representing agronomic terms for certain plant growing procedures or plant illnesses, for example.

The only excerpted culinary source in the (RSANU 1959–2019) material is the third edition of Midžina's *Cookbook*. The examples from this handbook illustrate more than three hundred entries, but only a couple of them like ASPIK (*aspic*), BAJCOVATI (*to soak*), MADŽUN (*grape syrup*), MUTVARA (*schmarrn*), NABOCKATI (*to prick*), NAZADEVATI (*to stuff*), ZAVARITI (*to marinade*), ZAVEZAČA (*ravioli-like dish*) feature the qualifier KUV. Entries like BATAK (*drumstick*), BELANJAK (*egg white*), BUBREŽNJAK (*loin*), DEVDIR (*colander*), GARNIRATI (*garnish*), ISPRŽITI (*fry*), JETRENICA (*beef steak*), etc. do not include the qualifier KUV, although their usage is illustrated by an example from Midžina and it is clear that they belong to the language of food (too). A particularly interesting example is provided by the entry JETRENICA, a neologism introduced by Midžina so as to replace German *lun-genbraten* or *beef steak* (Витас 2022). This entry is defined and illustrated in RSANU (1959–2019) in the following way:

JETRENICA: 2. v. bubrežnjak (1 a) — Najslađe je jetrenica — lungebraten. (*jetrenica: 2. see bubrežnjak (1a) — jetrenica — lungebraten is the tastiest.*)

However, BUBREŽNJAK (1a) that in Midžina refers to a different piece of meat (with different German equivalents) is defined in RSANU (1959–2019) as *kidneys with tallow and meat around them* and illustrated by an example

from Midžina: UZMI OD MLADOG GOVEČETA BUBREŽNJAK (*Take the loin of a veal*).⁷

A different example is provided by the entry MESNJAČA whose principal meaning is defined in RSANU (1959–2019) as:

MESNJAČA ž pokr. 1. vrsta kobasice od mesa; isp. mesnatica, mešnjača, (*mesnjača* m. regional. 1. a kind of meat sausage; cf. mesnatica, mešnjača, / meaty sausage,)

which is not found in Midžina or the majority of other books on cooking. For instance, in (RK 1915; Мирковић 1922; Марковић 1959),⁸ the expression MES(AT)A KOBASICA (*meat(y) sausage*) is used for this kind of sausage. This expression appears in RSANU (1959–2019) not as a separate term, but as part of examples for the entries NADENUTI (*to lard*) and KRTINAST (*meaty – without bones*) and it seems most likely to be an unrecorded synonym of MESNJAČA. But, in the cookbook (HrL 1878), whose author is Midžina's contemporary, and in (Пустановић 1938) as well, MESNJAČA is the only term found.⁹ Automatic recognition of the entry MESNJAČA as belonging to the domain of food, based on the above-mentioned definition and having the same meaning as MESNA KOBASICA seems like a time-consuming, if not an impossible task, to say the least.

Some entries appearing in the third edition of Midžina's *Cookbook*, such as BUTERTAJG (*butter puff pastry*), DEMERMAVIŠ (*dough cutter*), KATARCELTL (*caramel candy*)..., are among the ones omitted from the culinary terms in the above dictionaries. In the fourth thoroughly revised edition of that cookbook (Витас 2022), some of these words are featured, including the illustrative examples that define them:

- ... onaj URECKANI MALI TOČAK, kojim se seče testo („*demermaviš*“) (...*that jagged little wheel used for cutting dough “demermaviš”*) (Поповић-Мицина 1911, 498);¹⁰

7. The full text of this example is as follows: UZMI OD MLADOG GOVEČETA BUBREŽNJAK — NIERENSTUCK, NIERENBRATEN. (*Take the loin of a young veal — Nierenstuck, Nierenbraten*).

8. Hereinafter, (Марковић 1959) will be referred to as PATIN KUVAR (*Pata's Cookbook*).

9. Пустановић (1938) further specifies MESNJAČE, SVEŽE, POZNATE KAO BRATVURST... (*Mesnjača (sausages), fresh ones, known as bratwurst*).

10. MAFIŠNJAK, another term for DEMERMAFIŠ, is found in RSANU (1959–2019), without the qualifier KUV, featuring the following explanation: TOČKIĆ ZA REZANJE I UKRAŠAVANJE MAFIŠA (*a small wheel for cutting and decorating mafiš pastries*).

- ... UMEŠENA VEĆ GOTOVA TESTA OD SLATKOG MASLA („*butertajga*“, Butterteig) (...*premade sweet butter dough –Butterteig*) (169).

On the other hand, KATARCELTL was illustrated as early as in the third edition, in a separate chapter, titled BONBONI-KATARCELTL (*Candy-Caramel*).¹¹

Excerpting the fourth edition (Поповић-Миџина 1911) means finding words that enriched the lexicon of cooking in the meantime, namely, the words Midžina herself, had not been using in previous editions. The entries BARDIRATI, BLANŠIRATI, FRIKASIRATI, HAŠE, DOBATNJAK, KARIČICE, RAZNOLIJE, absent from (Поповић-Миџина 1891) are among the examples of this kind. However, they are used in the fourth edition. The definitions or equivalents in foreign languages are stated in many cases:

- svaki njen komad pre zgotovljavanja BELITI („*blanširati*“ blanchiren) (*each piece is to be blanched before cooking* (“to blanch”, *blanchiren*));
- BARDIRATI (Bardiren) znači obložiti meso slaninom (*to bard (bardiren) means to cover meat with bacon*);
- HAŠE je jelo od iseckana mesa ili ribe (*Haché is a dish made of meat or fish cut into pieces*);
- FRIKASIRATI znači omanje komade mesa prelitati začinenim umokcem koji se gotovi od kuvana iseckana kroz sito proceđena mesa (*fricasser means to soak small pieces of meat in a dressing with spices made of cooked and chopped meat drained in a colander*);
- BUHAVICE („*krapfne*“) sa cimetom (*doughnuts with cinnamon*);¹²
- jedan batak sa KARABATKOM ili *dobatnjakom* (*a drumstick with the upper leg or thigh of poultry*);¹³
- RAZNOLIJE (Mixed Pickles);
- ISPRŽENE KARIČICE LUKA ili *karičice* („*ringlice*“) (/ *fried onion rings*).

11. KATARCELTL have been confirmed in Jakov Ignatović's novel TRPEN SPASEN (*Patience is a Virtue*) with a somewhat different orthography: KARTACETL.

12. The entry BUHAVICA is featured in RSANU (1959–2019) but this meaning is not present.

13. In the second edition, Midžina does not differentiate between BATAK (*drumstick*) and KARABATAK (*upper leg (of poultry)*) as separate pieces of meat, while in (Мирковић 1922, 23) a ШИКЕН is simply cut into “nice, tasty pieces “. The word KARABATAK is used in (Ердељановић 1908).

This includes the adjectives derived, for example, from the names of foreign cheeses: GORGONZOLSKI, GROERSKI, EMENTALSKI, PARMEZANSKI...¹⁴ (of or related to *Gorgonzola*, *Gruyère*, *Emmental*, *Parmesan*...) that are not present in RSANU (1959–2019).

Ethnographic sources (Тројановић 1983) and (Ердељановић 1908) were carefully excerpted for the purpose of compiling the dictionary RSANU (1959–2019), but although around a hundred excerpted examples refer to the language of food, just a small number of them features the qualifier KUV.¹⁵

In the dictionary of foreign words (Клајн and Шипка 2006) more attention has been dedicated to marking the lexemes belonging to the domain of cooking by using the qualifier KULIN(arstvo) (*culinary*). The following definitions apply to some of the examples found in this dictionary:

- MUSAKA KULIN. jelo od tanko izrezanog krompira, patlidžana ili tikvice sa iseckanim ili mlevenim mesom koje se peče u pećnici. tur. musaka od ar. musaqqa natopljen.¹⁶ (*mousaka culin. A dish made of thinly sliced potatoes, eggplants or courgettes with chopped or minced meat fried in the oven. tur. musaka from ar. musaqqa soaked*);
- BLANŠIRATI (...) KULIN. kratko (pro)kuvati (povrće ili voće), (po)pariti. fr. blanchir izbeljivati. (*to blanch (...) culin. to cook (vegetables or fruit), steam for a short period of time fr. blanchir.*).

But, there too, certain food ingredients or cookware do not have the marker *kulin* as in the examples:

14. Third edition lists the form ПАРМАЗАНСКИ. Including such adjectives is not unusual because, for instance, the forms КАЌКАВАЛЕВ and КАЌКАВАЛЈЕВ (belonging to *kashkaval*) are listed in RSANU (1959–2019).

15. The entries DOLMA and KRUMPIR-PIRJAN (*braised potato*) are the only ones in these sources to be marked with the qualifier KUV. The entry MEDŽGANIK is described as *a dish made of cooked, strained and mashed beans, bean paste* while another alternative name of this dish – MEŽGANIK, is marked as KUV and defined as (*dish*) *made of cooked, mashed beans*.

16. In (Поповић-Мицина 1891, 45) the moussaka recipe is given under the name MODRI PATLIDŽAN ZA 6 OSOBA (*Eggplant for 6 People*), in (Поповић-Мицина 1911) it appears under the name MUSAKA (“MODRI PATLIDŽAN”) (*moussaka “eggplant”*), while Марковић (1959) features no less than 30 different recipes containing the word *musaka*, e.g. *asparagus, strained kohlrabi, sour kraut, new lettuce moussaka*, etc. This example further illustrates the difficulty of defining culinary products: dictionary definitions most often refer to the most frequent, usual type of a dish, without describing the concept itself.

- PARADAJZ (...) BOT. jestiva povrtarska biljka, *Solanum lycopersicum*, crveni patlidžan, rajčica. nem. *Paradiesapfel* rajska jabuka (*tomato* (...) *bot. edible vegetable plant, Solanum lycopersicum, crveni patlidžan / red tomato, rajčica / tomato. ger. Paradiesapfel apple from paradise*);
- ŠERPA (...) niska okrugla posuda u kojoj se čuvaju namirnice. Nem. *Scherbe* (*pot* (...) *shallow, round dish for storing food. Ger. Scherbe*).

The missing qualifier KUV/KULIN, and the way of defining the elements of the language of food do not prevent an intelligent dictionary reader from understanding the meaning of a certain word, based on its definition. However, such instances of the lack of precision are, most likely a serious obstacle, both in terms of dictionary formalization and automatic extraction of a sub-dictionary describing the language of food. A query that does not refer to the presence of a certain entry or its forms in the dictionary, but to other markers, such as qualifiers, for example cannot yield a reliable result.

A comprehensive catalogue of the vocabulary of cooking could also be found in multilingual dictionaries aimed at the hospitality industry, e.g. (Banićević and Popović 2010; Jovanović 1970). These dictionaries are primarily intended for compiling multilingual menus and they, as a rule, consist of lists featuring names of dishes in two or more languages. These lists, obtained by excerpting different sources provide an insight into the equivalent dish names in different languages, especially dishes belonging to the international cuisine. But, such dictionaries offer no information about the sources of the entries selected to be featured in them, which makes their actual use in (Serbian) language questionable in certain cases. For instance, in (Banićević and Popović 2010), we come across exotic entries like ČORBA OD KENGUROVOG REPA (*kangaroo tail broth*) (67) or KONSOME OD PERAJA AJKULE (*shark fin consommé*) (72). In special cases, the entry is accompanied by main ingredients, as in the example ČORBA SA DRONJCIMA (MASA OD BRAŠNA I JAJA ZAKUVANA U VRUĆI BUJON) dough (*flake broth – a mixture of flour and eggs poured into a hot bouillon*) (69). Thus, in the case of numerous entries, it turns out to be hard or impossible to get a clue as to their meaning based on dictionary information. Still, this lexicographic effort offers an insight into the issue of transfer of culinary content from one language to another. For instance, the French dish *cailles sous le cendre* is translated into Serbian as PREPELICA POD SAČOM (175), and into English as *quails baked in ash*; therefore, the Serbian translation introduces a regional concept of preparing food under a large metal lid (SAČ) that is not present either in French or in English. Conversely, even when the Serbian equivalent is linguistically precise, it cannot convey the meaning of the original entry. So,

the corresponding Serbian translation equivalent of the French term *potage cultivateur*¹⁷ (Banićević and Popović 2010, 79) is ZEMLJORADNIČKA ČORBA (*farmer's broth*), but this French traditional cuisine recipe is not known in our region.

The way in which the material for the above-mentioned Serbian language dictionaries is collected via partial excerption of selected sources, followed by a choice of a certain number (not all) hits obtained in this manner, raises the question of entry relevance. Namely, since the information regarding entry frequency is missing, all entries are of equal importance, at the same level. Moreover, as shown by the above examples, partial excerption results in a number of (potential) entries becoming excluded from the dictionary material.

Finally, the very structure of the language of food evades such a description of its lexis. The name of a dish makes sense only if it is accompanied by a recipe detailing how it is to be prepared. The entries belonging to the language of food would have to include, in addition to a description of meaning, a wide array of information establishing their social and cultural status, temporal and territorial distribution, dietary comments, etc.

These remarks make it clear that conceiving and compiling a cooking dictionary describing different aspects of the language of food, requires setting up a wider corpus that incorporates diverse sources, as well as defining a more complex microstructure of a dictionary article.

3 The corpus of the language of food

Although the lexis of the language of food has been largely described in Serbian dictionaries, this description is indirect, as seen in the previous section and the information giving indications as to its evolution is missing. Particularly, the complex processes determining the nature of the collective taste, as part of the national identity cannot be described on the basis of the existing lexicographic descriptions. As an illustration of the complexity of this process, it is possible to make a comparison between contemporary notions about the content of the Serbian cuisine, for instance, from a corresponding Wikipedia article,¹⁸ with the range of dishes from older cookbooks and similar sources. Thus, the salad section in Wikipedia features eight “typical” salads that are part of the contemporary view of “traditional salad”, which

17. *Potage cultivateur*

18. *Serbian cuisine on Wikipedia*

are, however, still missing from Pata's Cookbook in that shape or form. Besides, dictionaries do not provide data on the frequency of the individual concepts belonging to the language of food, which would be considered to be a reliable criterion of their being widely known (accepted). So, the above article on salads mentions *TRLJANICA*, but the web or Serbian language corpora¹⁹ show that the widespread knowledge about this salad is insignificant compared to others, such as *SRPSKA* (*Serbian*), *ŠOPSKA* (*Shopska – Bulgarian*) or *RUSKA SALATA* (*Russian Salad – Salad Olivier*). *TRLJANICA* is, therefore, a concept i.e. local cuisine dish that the wider public is not familiar with.²⁰

The things that define the concepts related to dish names are recipes for their preparation (as precise as possible). The definitions beginning with *dish made with flour...* or *with milk...* that are a regular occurrence in lexicographic descriptions give only the most generalized impression of the dish type. On the other hand, a recipe depends, among other things, on the time when it was devised, as well as on the wider social context in which it was applied. The gradual integration of the concept of German *SOUP* into the contemporary Serbian menu lasting almost a whole century can serve as an illustration of that. Starting from the first cookbook in Serbian by hieromonk Jerotej Draganović Драгановић (1855), where the *clear beef soup* recipe (Продановић-Младенов 1984, 359) is named *O GOVEDINI* (*On Beef*) and Midžina's cookbooks, where it is named *GOVEDA ČORBA*, *BULJON* (*beef broth, bouillon*) and later *BELA GOVEDA ČORBA* (*white beef soup*) to Pata's Cookbook, where it appears as *GOVEDA SUPA* (*beef soup*), separate from *GOVEDI BUJON* (*beef bouillon*), one and the same concept, accompanied by similar recipes, changes its name, but not its significant position, as compared to other "liquid dishes" (as described in dictionaries). The advance of this dish from the Serbian cuisine in Austria-Hungary to other parts of Serbia was gradual and the term *SUPA* itself took long to become accepted in Serbia at the time.²¹

19. In the **Corpus of the Contemporary Serbian Language** *TRLJANICA* appears just two times, while the **web-corpus of the Serbian Language** features 20 instances, mainly originating from the web page of **Vranjske novine**.

20. A dish that is locally well-known at one point in time can quickly assume dominant position on the menu of a certain period, under the influence of food fashions.

21. In Stevan Sremac's 1904 short story *JUNAK DANA* (*Hero of the Day*), the argument about the term *SUPA* was heated: I tu bi se penzioner Stipsa prosto zgranuo, toliko bi ga ta jedna reč, „supa“ nervirala! (*And then Thrifty, the pensioner*

The question now is whether it would be possible to set up a corpus of the language of food which would allow monitoring different parameters determining the evolution of eating habits over the last 150 years. For each query, such a corpus would have to provide information about the time when a certain food-related concept appears or disappears, the manner in which it is propagated in different social strata, the changes in diet under the influence of technologies and external cultural influences, etc. Searching the corpus would provide a more precise insight into the changes of the language of food and the changes of the collective taste alike. When did a certain word with a certain meaning appear, when did it change its meaning, when did it disappear? What are the changes in its frequency in the period covered by the corpus?

The sources for compiling this kind of corpus collected up to the present moment are heterogenous. They include different written sources available, among which the primary ones are as follows:

- Cooking handbooks, published or in manuscript;
- Ethnographic studies on food and drink;
- Historical studies, especially private life histories;
- Cooking and gastronomy textbooks;
- Dictionaries of various kinds;
- Cooking monographs;
- Excerpts from literary works (including translations into Serbian) describing eating habits;
- Statistical representations;
- Other information-rich contributions like menus, etc.

It is clear that this kind of material must be properly prepared in advance, so as to make it possible to search the corpus using complex queries. This primarily applies to detailed metadata about the sources themselves and labelling relevant classes of named entities in them. For example, Тројановић (1983) describes customs present in a wide area, defined either by the common name of a territory (Serbia, Montenegro, Herzegovina, Macedonia, Slavonia, the Balkan peninsula, etc.) or a region indicator (Srem, Pirot District, Vasojević District, etc.) or very precisely, by stating the place where a certain custom was encountered (Pirot, Tetovo, Kladovo, Aleksinac, Rudnik, Mostar, Zlatibor, etc.). It is clear that, in that case, a concept belonging to the language of food gets specified in two ways: by the source

would simply become flabbergasted, because he was so irritated by that particular word “soup”!)

itself (Тројановић 1983) and the location where it was recorded (e.g. Vasojević District). The entities referring to measurements in recipes pose a particular problem themselves as a result of different unit systems, as well as the lack of precision in stating measures in cooking handbooks (Krstev et al. 2017).

The roles of individual sources listed above are not one and the same.

Recipe books list names of dishes and indirectly the ingredients that the featured dishes consist of. As a rule, they reflect eating habits of the middle class at a certain point in history where the recipes are often strongly influenced by foreign culinary skills. Still, these handbooks too include certain recipes typical of the traditional cuisine (for instance, in connection with the customs related to religious holidays). The numerous recipes found in such books provide no details as to their acceptance (in the population).

An important piece of information given by the cooking handbooks are instructions on the way in which a dish is prepared at a certain point in time and that is, as pointed out above, the true definition of the dish name. Moreover, the language of these handbooks reveals how certain cooking concepts got their names in Serbian. For example, the term ŠNICLA (*steak*) instead of KOMADIĆ (*small piece*) (Поповић-Миџина 1891) or KRIŠKA (*slice*) (HrL 1878) appears as late as in the cookbook by Мирковић (1922), while the cooking method of ПОХОВАНЈЕ (*frying in breadcrumbs or batter*), which used to be subsumed under the more general term PRŽENJE (*frying*), is named as late as in (Рустановић 1938).

Cookbooks also provide information about the range of foodstuffs available at a certain time (Витас 2017), as well as an insight into the problem of naming new ingredients.

The organization of the structure of cooking handbooks in order to make performing complex searches of their contents possible is another question to consider. Namely, a cooking handbook can be viewed as a separate item illustrating the cuisine of a period, and also as a collection of separate, independent recipes. In the latter case, linking the same or similar recipes from different handbooks enables us to follow their evolution. The presence of a dish in different handbooks separated by long stretches of time is an indication of the level of its integration into the national cuisine. On the other hand, among the wide variety of recipes offered by such handbooks, some recipes appear only once: these dishes that may have resulted from the inspiration of their authors are singular occurrences and do not form part of the usual range of widely accepted meals. For example, the recipe PEČENO PRASE U ASPIKU (*roasted young pig in aspic*) found only in (Мирковић 1922,

59) is more of an epitome of a display of gargantuan excess than part of the standard cuisine.

The recipes themselves feature a description of the necessary ingredients, preparation procedures, utensil requirements, etc., which paints a picture of the state of cooking abilities and skills at a certain period. On the other hand, these elements of the language of food too undergo changes from one handbook to another under the influence of new kinds of foodstuffs and food preparation techniques. For example, Jerotej Draganović's cookbook still does not mention ŠTEDNJAK (*cooker*), while as early as the first edition of Midžina, some twenty years later, cooking with the help of a cooker goes without saying.

The missing parameter in these sources is the confirmation of a concept independently from culinary or ethnographic descriptions. For example, among the myriad of broths, mentioned in the different editions of Midžina, some are amply confirmed in other sources, while in the case of others, there are no examples, except those included in these handbooks, showing that they were prepared. In (Драгановић 1855), there is a recipe for ČORBA OD ČOKOLADE (*chocolate broth*) that is not present in later cookbooks. In Midžina (second edition) OSTRIGE (*oysters*) are mentioned as many times as DUNAVSKI SOM (*Danube catfish*),²² therefore, we have reason to wonder whether, in the mid-19th century Novi Sad, oysters were as commonly eaten as catfish.

The information about the frequency of a concept cannot be found in cookbooks either. This kind of data can be obtained indirectly in the extracts from the corpus of literary works or newspaper articles describing eating customs, as shown, for example, in (Vitas 2022) or in the rare surviving restaurant or individual feast menus, e.g. (Чолак-Антић 2004; Костић 2019).

Monographic publications such as (Mijo 2012; Onfre 2002; Пелапрат 1971; Kostjuković 2007), describing culinary habits in other places are of special importance to the extent to which they show how the content from other culinary cultures, whose numerous elements will be subsequently integrated into the Serbian eating repertory can be expressed in Serbian. Such

22. In (Поповић-Мицина 1891), the word OSTRIGA (*oyster*) in its different forms appears 12 times, SOM *catfish* 13 times, while its synonyms, SOMOVINA or SOMSKA RIBA (*catfish flesh*) are not present. In (Поповић-Мицина 1911) SOM (*catfish*) appears 16 times, SOMSKA RIBA 47 times, SOMOVINA is absent, while OSTRIGA (*oyster*) appears 32 times, including LAŽNE OSTRIGE (*false oysters*).

foodstuffs or dishes get described in the translation of the foreign source first, and only later included in the Serbian menu.

The gradual integration of new ingredients and dishes into what the local cuisine has to offer can be illustrated via the example of inclusion of sea fish and sea food into the Serbian menu repertory. In the third edition of her cookbook, Midžina mentioned MORSKA RIBA (*sea fish*) remarking that she had never cooked them (Поповић-Мицина 1891, 149), while there was no mention of LIGNJE (*squids*). The fourth significantly altered edition mentions sea fish that come salted or dried “therefore half-prepared” (Поповић-Мицина 1911, 409), but without stating the individual fish varieties. In Rustanović’s cookbook (Рустановић 1938, 349) there are instructions for preparing sea fish that differentiate between blue, white and wild sea fish. The latter group also includes LIGNI (*squid*) or KALAMARI (*calamari*) in addition to LIGNICA,²³ HOBOTNICA (*octopus*) and SIPA (*cuttlefish*). Moreover, the cookbook features recipes for their preparation. This cookbook features ORADA (*sea bream*) too, “one of the tastiest kinds of sea fish”, which is also referred to as ZLATNA (*gold – Gilt-head bream*) in the picture on the page 350. In Pata’s cookbook the selection of sea fish is limited to SARDELE (*sardines*) and SKUŠA (*mackerel*), but the author says that “sea fish is regularly available in our open-air markets” (Марковић 1959, 93). Sea bream is missing from her book, but the list of sea fish includes BRANCIN (*sea bass*) (47). It mentions KALAMARI (*calamari*) (or OLIGNJI and ULINJE) as well. A vast selection of sea fish would appear in (Ненадовић 1978) (ZUBATAC (*common dentex*), BRANCIN (*sea bass*), ORADA (*sea bream*), CIPOL (*flathead grey mullet*), MURINA (*moray...*), together with KALAMARI (*calamari*). MORSKI RAKOVI (*crayfish*) that used to be present in earlier cookbooks, no longer appear in newer books on cooking, while the concept of MORSKI PLODOVI (*sea food*) is not recorded in any of the analyzed cooking handbooks. This example shows how the food that does not originate from the local sources gets gradually integrated into the national menu through the development of the appropriate lexis.

4 Towards a dictionary of cooking

The described corpus of texts on cooking provides a foundation for constructing a lexical base depicting the lexicon of the sublanguage of food in Serbian. In view of the content of the corpus, different types of dictionaries

23. A possible variation of LIGNJA (*squid*).

describing certain aspects of this sublanguage can be exported from such a base (dictionary of foodstuffs, historical dictionary describing the evolution of culinary terms, multilingual dictionary, no-fat cookbook, etc. (Stanković, Krstev, et al. 2018; Stanković, Stijović, et al. 2018).

In an earlier stage of processing of the texts belonging to the corpus, it is necessary to solve, above all, the issue of a high degree of variability of the names used, especially in older texts, describing the same concept, referring either to the ingredients, pre-made products or preparation procedures and the necessary utensils. This problem can be solved either by establishing an appropriate semantic network (Vujičić-Stanković, Krstev, and Vitas 2014), or via a collection of graphs (in a manner in which they are used in the Unitex system, for example) whose content depicts the content of individual nodes of the network assigning to it a canonical node representative (as a transducer output). Thus, plant-based ingredients have different names, depending on author and time of appearance. ŠARGAREPA (*carrot*), whose name was taken from Hungarian is mentioned in cooking handbooks under the following names as well:

- ŽUTA REPA, ŠARGAREPA (Драгановић 1855; RK 1915; Мирковић 1922);
- ŽUTA REPA – ŠARGAREPA, MRKVA (*žuta repa*), ŽUTA MRKVA (*šargarepa*) (Поповић-Мицина 1891, 1911);
- ŽUTA REPA (HrL 1878);
- MRKVA and ŠARGAREPA (Марковић 1959; AFŽ 1952; Ненадовић 1978; Продановић-Младенов 1984);
- MRKVA (Jokić et al. 1983);
- MRKVA, ŠARGAREPA, KAROTA (Рустановић 1938);
- ŠARGAREPA (Стојанић 2004; Рашић 1999).

In addition to these, in (Јазић 2000) there are also the names ŽUTA ZELEN, ŠANGAREPA, CRVENA MRKVA. The beginnings of use of this plant in the Serbian cuisine in the early 20th century are discussed in (Косрић 2006):

“Naši seljaci [...] nisu gajili tada spanać, grašak, paradajz, salatu, ŠARGAREPU i ostalu zelen. Sve su to zemunske piljarice donosile za kuće u kojima je već vladala zapadnjačka kuhinja.” (*Our peasants [...] weren't growing spinach, peas, tomatoes, lettuce, carrots or other vegetables at that time. All that was brought by the green grocers in Zemun for the households that had already been dominated by the western cuisine.*) (397)

In the Eltec-corpus²⁴ there is only one mention of ŽUTA REPA (*yellow turnip*),²⁵ while ŠARGAREPA and MRKVA are absent. According to the corpus of the contemporary Serbian language,²⁶ ŠARGAREPA (with the frequency 667), is the dominant term nowadays, MRKVA (56) appears more rarely, and KAROTA (3) and ŽUTA REPA (1) are quite rare. By bringing together these different names for ŠARGAREPA (*carrot*) and other regional ones recorded in RSANU (1959–2019) (besides Latin *Daucus carota hortensis*, there are also *mrkvela*, *mrkuca*, *merlen*, etc.) to form a node in the semantic network (or a node in a Unitex graph) it is possible to neutralize the variability of names in the course of indexing or performing searches of the corpus. The described semantic network construction would make neutralization of lexical variability possible, thus simplifying solving the next complex task – identification of identical (or similar) recipes.

The content of morphological electronic dictionaries allows recognizing a certain number of culinary concepts, as described in (Крстев and Јазућ 2015; Vujičić-Stanković, Krstev, and Vitas 2014), via semantic markers associated with the dictionary entries. For the purposes of indexing the corpus it would be necessary to broaden and refine the set of markers in order to specify particular features of entities. For example, the marker *Zool* describes zoological terminology, while the marker *Anim* describes raw foodstuffs, but these markers cannot be used to automatically distinguish foodstuffs of animal origin the way it is done in cooking handbooks (kinds of meat, fish, poultry, etc.)

The recipes featured in cooking handbooks are either variations of an already well accepted dish or occasional dishes that are not repeated in other places and whose existence has not been confirmed elsewhere. Thus, for instance, PILEĆI PAPRIKAŠ (*chicken stew*) or GOVEDI GULAŠ (*beef goulash*) are present in almost all above-mentioned books on cooking. Therefore, they can be said to belong to the range of national dishes, despite their Hungarian origins, while the recipe for PATKA U MALAGINOM UMOKCU (*duck in Malaga dip*) is featured only in early editions of Midžina, disappearing as early as in the fourth edition. Some dishes disappear in time, or undergo changes and others get integrated into the recognizable range of dishes that make up the national cuisine. MRKA or CRNA ČORBA (*dark or black broth*), for example, present in early books on cooking, used to be an important dish at feasts that

24. **SrpELTeC** – The corpus consisting of a hundred Serbian novels first published in the period from 1840 and 1920.

25. In Draga Gavrilović's novella BABADEVOJKA (*Old Maid*) from 1887.

26. **SrpKor13** – The Corpus of Contemporary Serbian

was gradually reduced to a modest bone soup, eventually disappearing from the repertory of the described dishes altogether. A recognizable element of the national cuisine today, RAŽNJICI (*shish-kebab*) appears for the first time, as late as in (Поповић-Мицина 1911).

The decision regarding the concepts forming part of the traditional repertory of dishes can be made by using the frequencies found in the different corpora of the contemporary Serbian language SrpKor13 and the ELTeC-corpus as a starting point. Thus, PASULJ (*beans*), PEČENJE (*roast*), PAPRIKAŠ (*stew*), GULAŠ (*goulash*), ČEVAPČIĆ or ČEVAP (*grilled minced meat fingers*), DUVEČ (*sautéed rice with meat and vegetables*), SARMA (*stuffed cabbage leaf rolls*), SARMICA (*small stuffed leaf rolls*), PODVARAK (*stewed sauerkraut*), KAPAMA (*lamb stew*), AJVAR (*red pepper chutney*), etc.²⁷ are among the high frequency dishes in the first of the two corpora that are confirmed in the other. Some of these terms have undergone a change of meaning over time.²⁸ Certain old dishes have disappeared, e.g. JANIJA (*lamb stew*) and PAPAŽJANIJA (*beef and pork stew*) while new ones have appeared: PLJESKAVICA (*grilled meat patties*), ŠNICLE (*steak*). The same applies to the ingredients used, e.g. SUSAM (*sesame*).

The structure of the dictionary articles in the database should be conceived in such a way so as to make accumulating different kinds of information related to the entries possible. An example to be emulated in that respect is the lexicon (Zirojević 2019) where the featured foodstuffs are illustrated not only by lexicographic, but also various other information, including etymological, historical and nutritional data. The selection of entries for a dictionary of this kind would have to incorporate, in addition to the foodstuffs and dishes present in the Serbian cuisine, culinary techniques and equipment.

A description of dishes in such a database can be accompanied not only by linguistically relevant data, but also by outstanding examples of recipes (including their development over time), as well as chosen, primarily, literary examples that often define the social, cultural and geographic context in which a certain dish appeared.

27. This list largely corresponds to the dishes prepared by cooks in Kruševac around 1910. (Ђорђевић 1928, 203)

28. In older sources the term AJVAR denotes KAVIJAR (*caviar*), and ČEVAP denotes pieces of meat on a spit.

5 Conclusion

The neglected sublanguage of food in Serbian, as described in this paper, presents a multifaceted research challenge. Possibly the most comprehensive rationale for such an undertaking was given by Escoffier in his introduction to the famous *Larousse gastronomique* (Montagné 1949; Courtine 1986): “The history of the table of a nation is a reflection of the civilization of that nation. To show the changes in the order and serving of meals from century to century, to describe and comment on the progress of the French cuisine, is to paint a picture of the many stages through which a nation has evolved since the distant times.”²⁹

The available dictionaries, private life histories, ethnographic writings, books on cooking and other available sources provide only well-known, accessible parts of a complex mosaic in which the Serbian gastronomy developed. On the other hand, the dishes that are nowadays considered to be an integral part of the Serbian traditional table conceal the continuum of events determining the current dominant local taste. The road traveled by today’s “traditional” dish, from a regional one and even more often under pressure of cultural influences from elsewhere has largely remained hidden.

The basis that would finally ensure the provision of the material for a “good and comprehensive dictionary of cooking” is the described concept of a corpus of the language of food. In view of the contemporary dictionary form, it is clear that it will emerge and get constructed as a particular, possibly multimedia, database before any paper version. Its excerpts can yield specific or general dictionaries of cooking, but also, besides that, such information technology structures make possible continuous updating and development of the inner structure by adding new content (Stanković, Krstev, et al. 2018). Apart from the dictionary, as an occasional overview of the state of the database, the opportunity for long-term monitoring of the eating habits and customs would allow to get a clearer insight into what is indeed part of the usual – traditional? – eating habits of the local population.

The applications of this kind of database are wider than lexicography. They can be perceived in other research areas, especially in culture studies, as

29. ENTREPRENDRE d’écrire l’histoire de la table d’un peuple, exposer les modifications qui, de siècle en siècle, furent apportées dans son ordonnance et ses services, décrire et commenter les progrès de sa cuisine, c’est brosser un tableau suggestif de la civilisation de ce peuple en le suivant dans les étapes qu’il parcourut, depuis l’époque lointaine. (English translation from (Montagné 1978).)

well as historiographical and ethnographic studies, along with the everyday practice of the hospitality industry and tourism.

Acknowledgment

The preparation of the English version of this paper was supported by the Society for Language Resources and Technologies.

References

- AFŽ. 1952. *Мали кувар : са конзервирањем воћа и поврћа и разним практичним саветима*. Београд : Главни одбор АФЖ за Србију.
- Banićević, Marta, and Magdalen Popović. 2010. *Rečnik ugostiteljstva : (srpski-engleski-nemački-francuski-italijanski-ruski)*. Београд : Udruženje naučnih i stručnih prevodilaca Srbije.
- Courtine, Robert. 1986. *Dictionnaire de cuisine et de gastronomie, I-II*. Paris: Larousse.
- HrL. 1878. *Велики српски кувар : с обзиром на нове мере за српске домаћице, и оне, које желе то бити као и за све гостионичаре / саставила из разних књига а према искуству своје и својих пријатељица*. Панчево : Књижара Браће Јовановића.
- Jokić, Nedeljko, Milorad Tekić, Gavro Maletić, Andrija Paradžik, Olga Ašanin, and Nijaz Halilović. 1983. *Recepture za pripremanje jela u Jugoslovenskoj narodnoj armiji*. Београд : Vojnoizdavački zavod.
- Jovanović, Ksenija. 1970. *Francuski u restoranu : priručnik za radnike u ugostiteljstvu*. Београд : Kolarčev narodni univerzitet.
- Kittredge, Richard, and John Lehrberger. 1982. *Sublanguage: Studies of language in restricted semantic domains*. de Gruyter.
- Kostjuković, Elena. 2007. *Zašto Italijani vole da pričaju o jelu*. Београд : Paideia.
- Krstev, Cvetana, Duško Vitas, Miloš Utvic, and Branislava Šandrih. 2017. "The New Clothes for an Old Cookbook." In *Proceedings of 8th Language & Technology Conference, November 17-19, 2017, Poznań, Poland*, 174–178. <http://poincare.matf.bg.ac.rs/~cvetana/biblio/ltc-042-krstev.pdf>.

- Mijo, Kristijan. 2012. *Rečnik zaljubljenika u gastronomiju*. Beograd : Službeni glasnik.
- Montagné, Prosper. 1949. *Larousse gastronomique*. Paris: Larousse.
- Montagné, Prosper. 1978. *The New Larousse gastronomique – The Encyclopedia of Food, Wine & Cookery*. New York: Crown Publishers, Inc. ISBN: 0517531372.
- Onfre, Mišel. 2002. *Gurmanski um : filozofija ukusa*. Beograd : B. Kukić ; Čačak : Gradac, 2002.
- Paumier, Sébastien, Wolfgang Flury, Franz Guenther, Eric Laporte, Friederike Malchok, Clemens Marschner, Claude Martineau, Cristian Martinez, Denis Maurel, Sebastian Nagel, et al. 2021. “UNITEX 3.3 User Manual.”
- RK. 1915. *Ратни кувар у којем су упуства за припремање јела већим делом готовљена без меса а за ове ратне прилике : ручна књига за наше домаћице*. Ујвидек : С. Ф. Огњановић.
- RMS. 1976. *Речник српскохрватскога књижевног језика*. Нови Сад : Матица српска.
- RSANU. 1959–2019. *Речник српскохрватског књижевног и народног језика, књ. I - XXI*. Београд : Институт за српски језик САНУ. ISBN: 86-7025-309-7.
- RSJ. 2007. *Речник српског језика*. Нови Сад : Матица српска.
- Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. “Electronic Dictionaries – from File System to lemon Based Lexical Database.” In *Proc. of the Eleventh International Conference LREC 2018 – W23 6TH Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018)*, edited by J.P. McCrae, C. Chiarcos, T. Declerck, J. Gracia, and B. Klimek. Miyazaki, Japan.
- Stanković, Ranka, Rada Stijović, Duško Vitas, Cvetana Krstev, and Sabo Sabo. 2018. “The Dictionary of the Serbian Academy: from the Text to the Lexical Database.” In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, edited by J. Čibej, V. Gorjanc, I. Kosem, and S. Krek, 941–949. Ljubljana, Slovenia.

- Vitas, Duško. 2022. “From Onions to Champagne – Food and Drink in the SrpELTeC Corpus.” *Infotheca – Journal for Digital Humanities* 21 (2): 88–118. <https://doi.org/10.18485/infotheca.2021.21.2.5>.
- Vujičić-Stanković, Staša, Cvetana Krstev, and Duško Vitas. 2014. “Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain.” In *Proceedings of the Seventh Global Wordnet Conference*, 127–132.
- Zirojević, Olga. 2019. *Istočno-zapadna sofrа: mali kulturno-istorijski i kulinarski leksikon*. 2. izd. Beograd : Geopoetika izdavaštvo. ISBN: 978-86-6145-290-1.
- Витас, Душко. 2017. “Белешка о језику кулинарства.” In *Српски језик и његови ресурси: теорија, опис и примене*, edited by Р. Драгићевић and А. Милановић, 46:45–60. Научни састанак слависта у Вукове дане 3. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду. <https://doi.org/10.18485/msc.2017.46.3.ch3>.
- Витас, Душко. 2018. “Храна из нежељене поште : (анатомија језика брзе хране).” In *Српски језик и његови ресурси: теорија, опис и примене*, edited by Б. Ђорић and А. Милановић, 47:21–35. Научни састанак слависта у Вукове дане 3. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду. <https://doi.org/10.18485/msc.2018.47.3.ch2>.
- Витас, Душко. 2022. “Легенда о Мициној: прилог историји српских куварских приручника” [in Serbian]. 8, *Књижевство : часопис за студије књижевности, рода и културе* 12 (12). <https://doi.org/10.18485/knjiz.2022.12.12.8>.
- Драгановић, Јеротеј. 1855. *Србскій куварЪ: (по немачкомју кош-бухЪ)*. У Новом Саду : ТрошкомЪ Ігнаца Фухса, књижара.
- Ђорђевић, Тихомир. 1928. *Живот и обичаји народни. Књ. 17 Српски етнографски зборник*. Београд : Српска краљевска академија.
- Ердељановић, Јован, ed. 1908. *Српска народна јела и пића. Књ. 1*. Vol. 10. Београд : Српска краљевска академија.
- Клајн, Иван, and Милан Шипка. 2006. *Велики речник страних речи и израза*. Нови Сад : Прометеј.

- Костић, Ђорђе С. 2006. “Окрепљење тела.” In *Приватни живот код Срба у деветнаестом веку: од краја осамнаестог века до почетка Првог светског рата*, edited by Ана Столић and Ненад Макуљевић, 385–402. Београд: Clio.
- Костић, Ђорђе С. 2019. *Трпеза за уморне путнике : европски путописци о исхрани у Србији у 19. веку*. Београд : Геопоетика издаваштво.
- Крстев, Цветана, and Биљана Лазић. 2015. “Глаголи у кухињи и за столом.” *Научни састанак слависта у Вукове дане–Српски језик и његови ресурси: теорија, описи, примене* 44:3:117–135.
- Лазић, Анђелка. 2000. *Јела и пића у Мачви*. Шабац: А. Лазић.
- Марковић, Спасенија Пата. 1959. *Велики народни кувар [Патин кувар]*. Београд : Народна књига.
- Мирковић, Софија. 1922. *Велики српски кувар*. 3. поправљено и проширено изд. Нови Сад : Издавачка књижарница С. Ф. Огњановића.
- Ненадовић, Љубица. 1978. *Мали Патин кувар*. Београд : Народна књига.
- Пелапрат, Анри Пол. 1971. *Први кувар света : модерна француска и међународна уметност кувања*. Београд : Просвета.
- Поповић-Мицина, Катарина. 1891. *Велики српски кувар : са много лепих и врло вештачки израђених слика : за употребу српских домаћица*. 3. прегледано и умножено изд. Нови Сад : Српска књижара и штампарија браће М. Поповића.
- Поповић-Мицина, Катарина. 1911. *Велики српски кувар : са много лепих и врло вештачки израђених слика : српским домаћицама : са допунама најновијег практичног кувања*. 4. прегледано и умножено изд. Нови Сад : Учитељско деоничарско друштво „Натошевић“.
- Продановић-Младенов, Златија. 1984. *Српски кувар*. Beograd : Kolaričev narodni univerzitet Београд : Београдски издавачко-графички завод.
- Радовановић, Драгана. 2011. “Из кулинарске лексике Санада и Мокрина.” *Лексикологија, ономастика, синтакса: Зборник у част Гордани Вуковић*, 227–238.

- Радовановић, Драгана. 2017. “Од Вукове трпезе из његовог Рјечника до Матичиног Једнотомника.” In *Словенска терминологија данас*, edited by П. Пипер and В. Јовановић, 577–585. Београд : Српска академија наука и уметности : Институт за српски језик САНУ. ISBN: 978-86-7025-741-2. <https://dais.sanu.ac.rs/123456789/4548>.
- Радојичић, Драгана. 2015. *Дијалози за трпезом: антрополошки огледи о култури исхране*. 2. изд. Београд: Службени гласник : Етнографски институт САНУ.
- Радуловачки, Љиљана. 1996. *Традиционална исхрана Срба у Срему*. Нови Сад: Матица српска.
- Рашић, Атанасија. 1999. *Посни кувар на уљу*. Костолац : Манастир Рукумија.
- Рустановић, Александра. 1938. *Велики народни кувар*. Београд : Г. Кон.
- СТИЈОВИЋ, Рада, Ранка Станковић, and Олга Сабо. 2017. “Речник САНУ као база термилошких речника : на примеру речника кулинарства.” In *Словенска терминологија данас*, edited by П. Пипер and В. Јовановић, 109–123. Београд : Српска академија наука и уметности : Институт за српски језик САНУ. ISBN: 978-86-7025-741-2. <https://dais.sanu.ac.rs/123456789/1937>.
- Стојанић, Милијан. 2004. *Српски народни кувар*. Београд : Политика Newspapers / Magazines.
- Тројановић, Сима. 1983. *Старинска српска јела и пића*. Vol. 3. Београд : Просвета.
- Чолак-Антић, Тијана. 2004. “О posluženju i ponašanju за stolom u Beogradu.” *Kultura* 2 (109/112): 340–350.

SrpNemKor and Linked Open Data

UDC 811.163.41'322.2:811.112.2'322.2]:004.822

DOI 10.18485/infodhca.2023.23.1.2

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.

PAPER SUBMITTED: 30 March 2023

PAPER ACCEPTED: 9 May 2023

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 Documents",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 (Ruraz 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 Wikidata³. 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 subclouds⁴ 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 library⁶ as a separate *SrpNemKor* collection⁷ 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 enhanced⁹ 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. Bilingual 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 BiTE (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. BiTE, 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. <i>complete</i>)	absolut, gänzlich, total, völig, vollkommen, vollständig
базен (engl. <i>swimming pool</i>)	Bassin, Becken, Schwimmbad, Schwimmbecken
мајка (engl. <i>mother</i>)	Mama, Mutter, Mutti
Lexical units in German	Synonyms in Serbian
Augenblick (engl. <i>moment</i>)	трен, тренутак момент, час
fremd (engl. <i>foreign</i>)	туђ, туђински, непознат
grenzenlos (engl. <i>limitless</i>)	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 – LCN²⁰ 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.

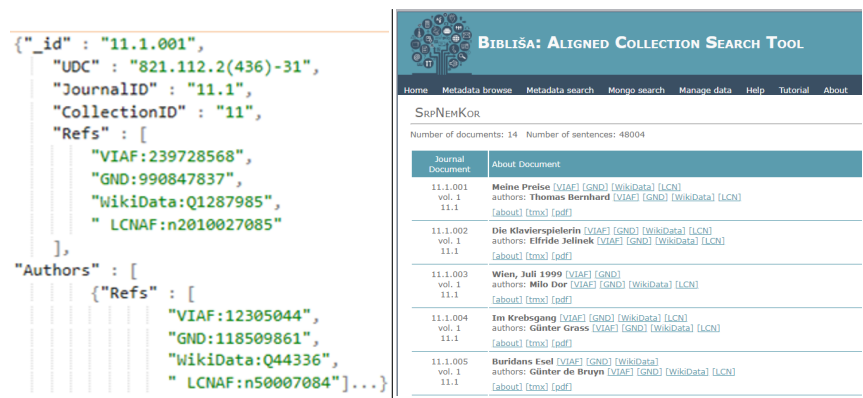


Figure 1. Linking metadata in *Bibliša*'s metadata schema with the records in semantic databases (left); Display of links in the *Bibliša*'s interface with the records in the semantic databases.

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 – LCN^{AF}

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/RDF²³ 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, ü, 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 <i>lexinfo:partOfSpeech</i> lexinfo:noun <i>a ontolex:LexicalEntry</i> ; <i>ontolex:Form</i> :Augenblick-x-de-form. :Augenblick-n-de-form <i>ontolex:writtenRep</i> "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 <i>lexinfo:partOfSpeech</i> lexinfo:noun; <i>a ontolex:LexicalEntry</i> ; <i>ontolex:Form</i> :%C4%8Das-n-sr-form. :%C4%8Das-n-sr-form <i>ontolex:writtenRep</i> "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 <i>a ontolex:LexicalSense</i> ; <i>skos:definition</i> "bestimmte Zeitspanne von kurzer Dauer, Moment, Augenblick"@de; <i>ontolex:isLexicalizedSenseOf</i> :Augenblick-n-de, :Moment-n-de, :Zeitpunkt-n- de	entry from German dataset entry definition entry format in German dataset
:112383-sr-sense <i>a ontolex:LexicalSense</i> ; <i>skos:definition</i> "vrlo kratki deo vremena, tj. vreme potrebno da se trepne okom"@sr; <i>ontolex:isLexicalizedSenseOf</i> :%20tren-n-sr, :%20trenutak-n-sr, :%C4%8Das-n-sr, :moment-n-sr. :112383-sr-sense-112383-de-sense-trans <i>a vartrans:Translation</i> ; <i>vartrans:source</i> :112383-sr-sense; <i>vartrans:target</i> :112383-de-sense.	entry from Serbian dataset entry definition entry format in Serbian dataset translation equivalent – class <i>vartrans:Translation</i> source language dataset (German) 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 LexinfoLexinfo²⁴ and SKOS²⁵ ([@prefix lexinfo:](#) and [@prefix skos:](#)). Lexinfo (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

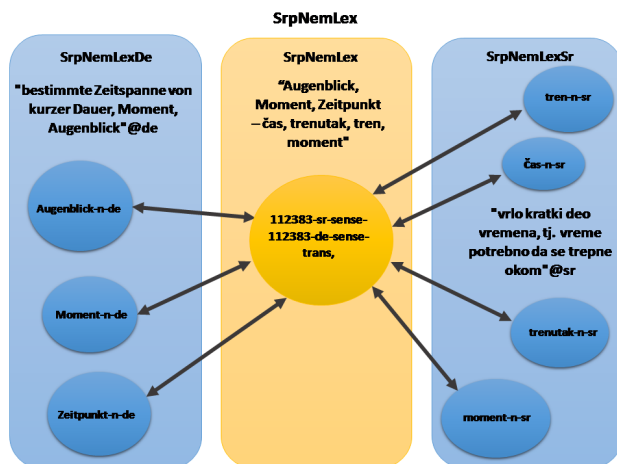


Figure 2. A graphic representation of a translation equivalent from the dictionary SrpNemLex.

ISO 12620:2019 standard²⁷ as a replacement for ISOcat metadata registry²⁸ (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 model²⁹ 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 gen-  
erated 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 An-  
donovski", "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 project³⁰ (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 Bruyn

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

Buridan's ass (Q110637016)

1968 novel by Günter de Bruyn 

[In more languages](#)

Configure

Language	Label	Description	Also known as
English	Buridan's ass	1968 novel by Günter de Bruyn	
Serbian	Буриданов марапац : роман	роман немачког писца Гјентера де Брујна	
German	Buridans Esel	Roman von Günter de Bruyn	
Croatian	No label defined	No description defined	

has edition or translation 

Buridan's ass : Novel : SrpNemKor Edition 

language of work or name Serbian

German

publication date 2019

place of publication Belgrade

0 references 

Buridan's ass : 1992 Edition 

language of work or name German

publication date 1992

place of publication Frankfurt am Main

0 references 

Buridan's ass : 2007 Edition 

language of work or name Serbian

publication date 2007

place of publication Belgrade

0 references 

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).

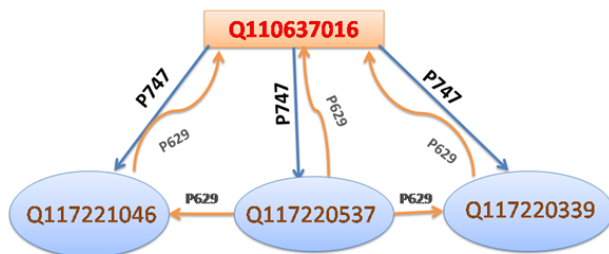


Figure 4. “Has edition or translation” (P747) and “edition or translation of” (P629) properties in Wikidata for the novel *Buridan’s ass*.

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) authorities⁴¹ with information: “retrieved” (10 October 2015) and “reference URL” (<http://data.bnf.fr/ark:/12148/cb118858431>) (Figure 5).

Günter de Bruyn (Q62753)

German writer (1926–2020) [edit](#)

[In more languages](#)

[Configure](#)

Language	Label	Description	Also known as
English	Günter de Bruyn	German writer (1926–2020)	
Serbian	Гинтер де Брујн	немањки писцац	
German	Günter de Bruyn	deutscher Schriftsteller	Günther de Bruyn
Croatian	No label defined	No description defined	

[All entered languages](#)

Statements

instance of	human edit
	2 references
	imported from Wikimedia project Russian Wikipedia
	stated in BnF authorities
	retrieved 10 October 2015
	reference URL http://data.bnf.fr/ark:/12148/cb118858431

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

5.3 SPARQL queries

It is possible to search metadata in Wikidata user environment by SPARQL queries⁴² 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
Q1117220339	Buridan's ass : 2007 Edition	wd:Q62753	Günter de Bruyn	Q:wd:Q3711	Belgrade	January 1, 2007	Q:wd:Q8261	novel
Q1117220597	Buridan's ass : Novel: SrpNemKor Edition	wd:Q62753	Günter de Bruyn	Q:wd:Q3711	Belgrade	January 1, 2019	Q:wd:Q461183	text corpus
Q1117221046	Buridan's ass : 1992 Edition	wd:Q62753	Günter de Bruyn	Q:wd:Q1794	Frankfurt am Main	January 1, 1992	Q: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

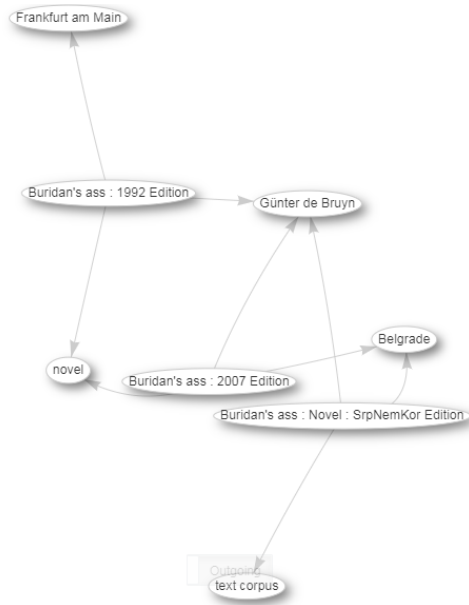


Figure 7. Visualization of the first SPARQL query in the form of an interactive graph.

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.

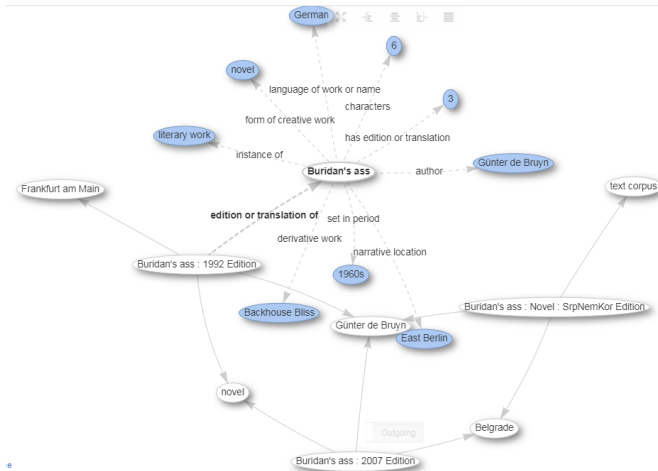


Figure 9. The graph of an item “Buridan’s ass: novel” (Q110637016)

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-TheSemantiWeb-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 Cimini, 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.medien-campus.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.

It-Sr-NER: Web Services for Recognizing and Linking Named Entities in Text and Displaying Them on a Web Map

UDC 81'322.3

DOI 10.18485/infotheca.2023.23.1.3

ABSTRACT: The paper will present the results of the project “It-Sr-NER: Web services for named entities recognition, linking and mapping,” in which teams from the University of Turin and the Society for Language Resources and Technologies JeRTeh participated, and whose goal was the development of the *It-Sr-NER* web service for named entity annotations in the text and displaying them on the map. Named entities in these services are names of persons, places, organizations, demonyms (ethnicities), events and works of art.

KEYWORDS: parallel corpora, named entities, NER, NEL, geoparsing, Serbian, Italian, web services.

PAPER SUBMITTED: 29 November 2022

PAPER ACCEPTED: 31 January 2023

Olja Perišić

olja.perisic@unito.it

Università degli Studi di Torino

Dipartimento di Lingue e

Letterature Straniere e

Culture Moderne

Turin, Italy

Ranka Stanković

ranka.stankovic@rgf.bg.ac.rs

University of Belgrade

Faculty of Mining and Geology

Belgrade, Serbia

Milica Ikonić Nešić

milica.ikonic.nesic@fil.bg.ac.rs

University of Belgrade

Faculty of Philology

Belgrade, Serbia

Mihailo Škorić

mihailo.skoric@rgf.bg.ac.rs

University of Belgrade

Faculty of Mining and Geology

Belgrade, Serbia

1 Introduction

The lack of tools and resources that enable the annotation, research and analysis of bilingually aligned (parallel) Italian-Serbian texts was the main motivation and inspiration for starting this project. In the didactics of foreign languages in Serbia, there is an almost complete absence of corpus tools

in teaching (Vitaz and Poletanović 2020), while on an individual level and through personal initiatives in the teaching of the Serbian language as a foreign language in Italy and the Italian language in Serbia, it has been shown that corpora in teaching are significant in many ways and that students are glad to accept and apply them in shared work, but also in independent research (Moderc 2015a; Perišić 2021). The rich and sophisticated morphology of the Serbian language provides for the declension of toponyms and other named entities that foreign students are not always able to recognize and reduce to their basic form. Some of the reasons are the same endings for the masculine and neuter gender in most cases, the presence of certain toponyms only in the plural form, the so-called *pluralia tantum* (Pljevlja, Divčibare, etc.), but also phonetic transcription of foreign names as well as some orthographic inconsistencies (Витас and Павловић-Лажетић 2008).

As part of the "Bridging Gaps" call of the European infrastructure for language resources and technologies CLARIN¹ (Common Language Resources & Technology Infrastructure), a team of experts from the University of Turin and the Society for Language Resources and Technologies JeRTeh joined forces and developed web services for annotating named entities in text, ensuring their linking with Wikidata,² as well as geoparsing, i.e. geolocation of recognized locations and their display on the map. Named entities in these services are names of persons, places, organizations, demononyms (ethnicities), events and works of art.

The main goal of the project is the realization and publication of web applications and services for monolingual and bilingual, parallel texts within the CLARIN infrastructure, as well as on the platform of the Society for Language Resources and Technologies JeRTeh. The project envisages the creation and publication of the Italian-Serbian corpus of 10,000 segments of extracted and aligned sentences, selected from classics of Italian and Serbian literature. The results of the project are not limited to the Serbian-Italian language combination, but the developed services can be applied to the processing of texts in twenty-four different languages.

The project initiator and team leader is Olja Perišić, a professor at the University of Turin (*Università degli Studi di Torino, Dipartimento di Lingue e Letterature Straniere e Culture Moderne*) where she teaches the Serbian language. On behalf of JeRTeh, the development of the services was led by prof. Ranka Stanković in cooperation with prof. Duško Vitas. More information about the project is also available on the website of the Society for

1. CLARIN

2. Wikidata

Language Resources and Technologies JeRTeh It-Sr-NER CLARIN compatible NER and geoparsing web services for parallel texts.³

2 Parallel corpus

In foreign language teaching, parallel corpora have proved to be an indispensable instrument in order to acquire morphosyntax and lexis. The parallel observation of two or more languages facilitates contrastive analysis, i.e. the observation of similarities and differences of language structures thanks to the large number of examples of sentences in context, as was noticed by Sinclair at an early stage of corpus linguistics: “The language looks rather different when you look at a lot of it at once.” (Sinclair 1991, 100). In teaching translation, parallel corpora enable word sense disambiguation and the definition of polysemic lexicon, which are not given enough space in bilingual dictionaries (Moderc 2015b; Perišić Arsić 2018). At the same time, it has been noticed that the number of representative, parallel corpora even for major world languages is insufficient (Granger 2018).

For all these reasons, in the first phase of the project, it was necessary to create an Italian-Serbian corpus of 10,000 aligned segments (sentences) excerpted from ten different novels. The novels are presented with random samples of segments in order to avoid copyright issues. Novels by Italian writers represented in the corpus are: Umberto Eco, *The Name of the Rose*; Carlo Collodi, *The Adventures of Pinocchio*; Elena Ferrante, *Those Who Leave and Those Who Stay*; Luigi Pirandello, *One, None and a Hundred Thousand*. Serbian writers are represented by five novels: Ivo Andrić, *Legends of Anika* and *The Bridge on the Drina*; Borisav Stanković, *Impure Blood*; Branislav Nušić, *Municipal child: the novel of an infant*; Danilo Kiš, *Garden, Ashes*. Considering that the main task of the project is to annotate the named entities, the corpus also includes translations into Italian and Serbian of Jules Verne’s novel *Around the World in Eighty Days*.

The novels were aligned and prepared in TMX (Translation Memory eXchange) format using the ACIDE application, an integrated environment for the development of parallel corpora (Obradović, Stanković, and Utvić 2008; Krstev and Vitas 2011). Figure 1 presents the first two translation units, marked with the label <tu> (translation unit), within which there are translation equivalents marked with the label <tuv> (translation unit variant). Paired, i.e. aligned segments in Italian and Serbian are numbered (n1,

3. <https://jerteh.rs/index.php/it-sr-ner-3/>

```
<tu>
  <tuv xml:lang="it" creationid="n1" creationdate="20220825T211907Z">
    <seg>Sposa giovanissima Stefano Carracci e gestisce con
    successo prima la salumeria nel nuovo rione, poi il negozio di
    scarpe a piazza dei Martiri.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n1" creationdate="20220825T211907Z">
    <seg>Veoma mlada se udaje za Stefana Karačija i uspešno
    upravlja isprva delikatesnom radnjom u novom rejonu, a potom
    obučarskom radnjom na Trgu mučenika.</seg>
  </tuv>
</tu>
<tu>
  <tuv xml:lang="it" creationid="n2" creationdate="20220825T211907Z">
    <seg>Elena comincia a scriverla nel momento in cui apprende
    che la sua amica d'infanzia, Lina Cerullo, solo da lei
    chiamata Lila, è sparita.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n2" creationdate="20220825T211907Z">
    <seg>Elena počinje da je piše kada sazna da je nestala njena
    prijateljica iz detinjstva, Lina Čerulo, koju samo ona zove
    Lila.</seg>
  </tuv>
</tu>
```

Figure 1. Example of a TMX output document

n2,...) and each has an attribute indicating the language: `xml:lang="it"` or `xml:lang="sr"`. The parallelization process with the ACIDE application, in addition to generating a TMX output document, also generates an HTML representation that can be seen in Figure 2.

Italian (it)	Serbian (sr)
n1 Sposa giovanissima Stefano Carracci e gestisce con successo prima la salumeria nel nuovo rione, poi il negozio di scarpe a piazza dei Martiri.	n1 Veoma mlada se udaje za Stefana Karačija i uspešno upravlja isprva delikatesnom radnjom u novom rejonu, a potom obučarskom radnjom na Trgu mučenika.
n2 Elena comincia a scriverla nel momento in cui apprende che la sua amica d'infanzia, Lina Cerullo, solo da lei chiamata Lila, è sparita.	n2 Elena počinje da je piše kada sazna da je nestala njena prijateljica iz detinjstva, Lina Čerulo, koju samo ona zove Lila.

Figure 2. Example of aligned segments of translation equivalents in Italian and Serbian

The corpus is published in the ILC4CLARIN B Center, with the unique identifier <http://hdl.handle.net/20.500.11752/OPEN-980>, so it is visible

through the VLO (Virtual Language Observatory) <https://vlo.clarin.eu/>. The zipped form of the corpus contains several parts: aligned bilingual, but also individual, monolingual versions. Automatically tagged named entities (as explained in Section 3) are also part of the published corpus. The corpus can also be accessed from the working *github* site <https://github.com/rankastankovic/It-Sr-NER/tree/main/corpus>.

In addition to being downloadable, the corpus containing complete novels from which the published version of 10,000 segments were extracted, is also available for searching on the Bibliša⁴ digital library. A feature of the Bibliša library is an advanced search with the possibility of morphological and semantic expansion of queries for the Serbian language. Figure 3 shows

metadata	n3330 E si riunivano per lo più dietro la casa, presso le finestre della stanza grande, perchè li erano del tutto isolati dai contadini e dagli ultimi venuti e potevano gemere liberamente e inosservati, uccisi e irrigiditi dal dolore esclamare: - Ahimè, bato!	n3330 A najviše ih je bilo iza kuće, do samih prozora velike sobe, jer tu su sasvim odvojeni od ovih varošana i od dolazećih i mogli slobodni, neopaženi, da hukću, ubijeni i zgrčeni: - Oh, bre, bato!
metadata	n31 Specialmente le vedove, i cui figli appena cresciuti si davano a spendere e a sprecare, invece di pensare alla casa e a sostituire il padre, il capo della famiglia, facevano paura a questi lor figli, ricordando il "loro bato" [Bato = fratello], come tutti lo chiamavano in famiglia, e minacciavano.	n31 Osobito udovice, čiji sinovi tek što nastali, pa mesto da preduzmu i počnu voditi brigu o kući, da zamene oca, domaćina, a oni počeli trošiti i rasipati - osobito su one te svoje sinove jednako njime, „ batom svojim", kako su ga svi u rodbini zvali, zastrašivale i pretile im:
metadata	n4780 Dede ha bisogno di compagnia, crescere da soli è brutto, è meglio darle un fratellino o una sorellina.	n4780 Za Dede će biti dobro da dobije drugaricu, nije lako odrastati sam, bolje je da ima batu ili sestricu.
metadata	n2565 Anche quelli di Sofia facevano a gara nel servire, stringendosi intorno a Marco, per dimostrare agli altri quanto bene volevano al loro nuovo amico, al loro bato.	n2565 A Sofkini opet, radi njih, da bi pred njima pokazali koliko oni vole i cene svoga novog prijatelja, tog njihovog „ batu ", svi su se utrkiivali u služenju, obletanju oko Marka.
metadata	n2523 «Nadia e suo fratello».	n2523 „Nadja i njen brat .”

Figure 3. An example of expanding the search in Bibliša

a panel with the results of a semantic expansion example where the query “brat” (“brother”) is automatically expanded with the synonym “bata” (little brother), and then their morphological forms are added to the query. Several segments can be seen in Figure 3.

4. Библиша

The option to view collections of parallel documents (Figure 4) is available with a limited number of segments for all users and with a larger number of segments for authorized users.



BIBLIŠA: ALIGNED COLLECTION SEARCH TOOL

Home

Metadata browse

Metadata search

Mongo search

Manage data

Help

Tutorial

About

ITSRKOR

Number of documents: 11 Number of sentences: 33394

Journal Document	About Document
14.1.001 vol. 1 14.1	Storia di chi fugge e di chi resta, L'a authors: Elena Ferante [about] [txtx] [pdf]
14.1.002 vol. 1 14.1	Il nome della rosa authors: Umberto Eko [about] [txtx] [pdf]
14.1.003 vol. 1 14.1	Uno, nessuno e centomila authors: Luidi Pirandelo [about] [txtx]
14.1.004 vol. 1 14.1	Le avventure authors: Car [about] [txtx]
14.2.001 vol. 14.2	L'ora di Ani authors: Ivo [about] [txtx]
14.2.002 vol. 14.2	Il ponte sul authors: Ivo [about] [txtx]
14.2.003 vol. 14.2	Il ponte sul authors: Ivo

ITSRKOR

Number of documents: 11 Number of sentences: 33394

Journal Document	About Document
14.1.001 vol. 1 14.1	Priča o onima koji odlaze i onima koji ostaju authors: Elena Ferante [detallojse] [txtx] [pdf]
14.1.002 vol. 1 14.1	Ime ruže authors: Umberto Eko [detallojse] [txtx] [pdf]
14.1.003 vol. 1 14.1	Jedan, nijedan i sto hiljada authors: Luidi Pirandelo [detallojse] [txtx] [pdf]

En/De/Sr/JB: (first 9 out of 3277 sentences) [pdf]

n1

Luidi Pirandello, Uno, nessuno e centomila

n2

I. Mia moglie e il mio naso.

n3

- Che fai? - mia moglie mi domandò, vedendomi insolitamente indugiare davanti allo specchio.

n4

- Niente, - le risposi, - mi guardo qua, dentro il naso, in questa narice.

n5

Premendo, avverto un certo dolorino.

n6

Mia moglie sorrise e disse:

n7

- Credevo ti guardassi da che parte ti pende.

n8

Mi voltai come un cane a cui qualcuno avesse pestato la coda:

n9

- Mi pende?

Srpski - (prvih 9 od 3277 rečenica) [pdf]

n1

Luidi Pirandelo, Jedan, nijedan i sto hiljada

n2

I MOJA ŽENA I MOJ NOS.

n3

- Šta to radiš? - upitala me je moja žena kada je videla da se zadržavam pred ogledalom duže nego obično.

n4

- Ništa - odvratio sam - gledam nešto ovde, u nosu, u ovoj nozdrvi.

n5

Kad pritisnem, tu me malo bolucka.

n6

Moja žena se nasmeđila i kazala je:

n7

- Misliš sam da gledaš na koju stranu ti se krivi.

n8

Otkrenuo sam se naglo, kao pas kad mu neko stane na rep:

n9

- Krivi se?

Figure 4. Browsing a collection of parallel documents ItSrKor

Parallel corpora are valuable for translation studies, while contrastive linguistics and the simple use of concordances facilitate the study of cross-linguistic phenomena (Figure 5). Students of the Italian language at the University of Belgrade, where prof. Saša Moderc teaches, as well as students of the Serbian language in Turin, where prof. Olja Perišić teaches, will be able to use the developed resources, considering that they are open and therefore available to other students and researchers.

66

Infotheca Vol. 23, No. 1, July 2023

2.1 Service implementation

It-Sr-NER services,⁵ stored in the CLARIN repository with the unique identifier <http://hdl.handle.net/20.500.11752/OPEN-981>, not only process monolingual texts (in 24 languages), but also successfully annotate bilingual texts in the form of translation memories in TMX format.

Apart from the development of the web services, the ultimate goal was the integration into the European infrastructure for language resources and CLARIN technologies, specifically the Language Resource Switchboard platform.⁶ The primary goal was to annotate named entities for Italian and Serbian languages, but it was extended to other languages for which the compatible spaCy⁷ models exist. Models trained using the spaCy library, downloaded for each language from the corresponding repository <https://spacy.io/models>, were used to label the named entities. For the Italian language, the `it_core_news_sm-3.4.0`,⁸ model was downloaded, trained on an automatically created corpus for the recognition of named entities, WikiNER,⁹ based on the text and structure of Wikipedia (Nothman et al. 2013), while for the Serbian language, a model trained on the corpus was implemented of old Serbian novels SrpCANNER (Šandrih Todorović et al. 2021), downloaded from the European Language Grid (ELG) platform.¹⁰

In addition to recognizing named entities, the goal of the application was also to link them with items in the open knowledge base Wikidata. The University Library of Mannheim (Universitätsbibliothek Mannheim, abbreviated UB Mannheim) developed spaCyOpenTapioca¹¹ for the task of linking named entities to concepts (items) in Wikidata in spaCy using OpenTapioca¹² (Delpuch 2019). The source code of the service and the application is available on the github repository. By using the spaCyOpenTapioca package, It-Sr-NER services can, in addition to recognizing and annotating named entities, link entities with items in wikidata.

As a final result, a web service was created that enables geoparsing, i.e. displaying recognized named entities that exist in wikidata on a map.

5. It-Sr-NER services

6. <https://switchboard.clarin.eu/tools>

7. <https://spacy.io/>

8. `it_core_news_sm-3.4.0`

9. WikiNER

10. ELG

11. spaCyOpenTapioca

12. OpenTapioca

Eight web services have been developed, four each for monolingual and bilingual resources, which enable:

1. NER - recognition of named entities according to the classes from Table 1 trained language models of the spaCy library;
2. NER+NEL - in addition to the recognition of named entities from the previous point, additional linking with wikidata of annotated entities, where possible by using the functions of the spacyOpenTapioca service, applied only for the recognized named entities, i.e. for the text inside the XML tag;
3. NEL - recognition and linking of named entities with wikidata relying on the recognition of named entities by the spacyOpenTapioca system, whereby the recognized named entity is annotated with the tag <WDT> and the class of the named entity with the *label* attribute.
4. Geoparsing - for those recognized named entities of LOC class that exist in wikidata, finding the latitude and longitude (geolocating) using the *geopy*,¹³ library, followed by displaying them on a map using the *folium*¹⁴ library.

NER class tag	Description of the entity class
PERS	Names, surnames, nicknames and their combinations (of real people and fictional characters, including gods and saints).
LOC	Continents, countries, regions, settlements, oronyms, bodies of water, names of celestial bodies, city locations.
ORG	Names of companies, political parties, educational institutions, sports teams, hospitals, museums, libraries, hotels, cafes, churches and shrines.
DEMO	Residents of countries, cities, regions or ethnic groups; derived adjectives from the name of the location.
EVENT	Names of events that recur regularly or happened once but they have their own name: natural disasters, revolutions, battles, wars.
WORK	Titles of books, plays, poems, paintings, sculptures, newspapers.

Table 1. Named entity classes

13. <https://pypi.org/project/geopy/>

14. <https://python-visualization.github.io/folium/>

Language-specific models often use different labels to denote classes of named entities. So, for example, LOC is usually used for the spatial location class, but one can also find GPE for geopolitical entities (English model), then LC (Korean model), as well as placeName and geogName (Polish model). In order to harmonize the labels of named entity classes between models for different languages, the labels of the classes denoting locations and geopolitical entities have been renamed to the LOC label, even if other labels (GPE, LC, placeName, geogName) were used. The mapping is done systematically for all classes and can be seen in the configuration file available at `lng_config.csv`.¹⁵

In addition, entity labels of the PERS class, which mark persons, were set as the basic tag into which corresponding labels from other models were mapped: PER (Italian), PRS (Swedish), PERSON (Macedonian), persNAME (Polish). The label NORP (nationalities or religious or political groups) of nationalities, political and religious groups from the Japanese and Finnish models, then NAT_REL_POL from the Romanian one, are mapped into the label of the class DEMO, which are marked with demonyms, ethnic relations (Stanković et al. 2021). Since some language models have a much richer set of named entity classes, for example English has 18 classes, Romanian 16, a column with a list of ignored labels is defined in the configuration file.

In addition to the mentioned classes with which the named entities are associated, the NER+NEL and NEL web services provide information about the entity type (attribute *label*), description (attribute *desc*) and a link in the Wikidata knowledge base (attribute *ref*).

It has already been mentioned that the input can be monolingual or bilingual text. In the case of processing bilingual resources, the input file must be a valid TMX document. Figure 5 shows the output of the NER+NEL service for bilingual resources, while Figure 6 shows the output of the NEL service, also for bilingual resources, but this link recognized named entities to items in wikidata using the spacyOpenTapioca library for both languages.

All project results: program code, web services, web application, as well as parallel corpora, have been published with open licenses, and as such can be freely used for research and commercial purposes.

15. `lng_config.csv`

```

<tu>
  <prop type="Domain"/>
  <tuv xml:lang="it" creationid="n54" creationdate="20220825T211907Z">
    <seg>Progettava di raggiungere <LOC ref="https://www.wikidata.org/wiki/Q90" desc="capital and largest city of France">Parigi</LOC> insieme ad altri suoi compagni, mi invitò ad andare con lei in automobile.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n54" creationdate="20220825T211907Z">
    <seg>Plan joj je bio da stigne u <LOC ref="https://www.wikidata.org/wiki/Q90" desc="capital and largest city of France">Pariz</LOC> zajedno sa drugim svojim kolegama, pozvala me je da joj se pridružim, išle bismo automobilom.</seg>
  </tuv>
</tu>

```

Figure 5. NER+NEL output for bilingual resources in TMX format

```

<tu>
  <prop type="Domain">
  <tuv xml:lang="it" creationid="n934" creationdate="20220825T211907Z">
    <seg>Ranko Mihailović è un giovane silenzioso e bravo, anch'egli frequenta la facoltà di legge a <WDT ref="https://www.wikidata.org/wiki/Q1435" label="LOC" desc="capital city of Croatia">Zagabria</WDT>, si vede già dinanzi una carriera nell'amministrazione e prende raramente e tiepidamente parte alle discussioni e ai dibattiti dei suoi amici sull'amore, la politica, le diverse concezioni della vita e l'ordine sociale.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n934" creationdate="20220825T211907Z">
    <seg>Ranko Mihailović, ćutljiv i dobroćudan mladić, koji studira pravo u <WDT ref="https://www.wikidata.org/wiki/Q1435" label="LOC" desc="capital city of Croatia">Zagrebu</WDT>, pomišlja već sada na činovničku karijeru i slabo i mlako učestvuje u drugarskim prepirkama i razgovorima o ljubavi, politici, i pogledima na život i društveno uređenje.</seg>
  </tuv>
</prop></tu>

```

Figure 6. NEL output for bilingual resources in TMX format

3 Methods of use

The described web services are available via the web application at the address <https://ners.jerteh.rs/>, as well as by using the requests module within the Python application¹⁶ as follows:

```

1 import requests
2 # choose language – lang, and feat
3 lang = "it" #@param ['ca', 'zh', 'hr', 'da', 'nl', 'en', 'fi', 'fr', 'de', 'el', 'it', 'ja', 'ko', 'lt', 'mk', 'nb', 'pl', 'pt', 'ro', 'ru', 'es', 'sv', 'uk', 'sr']
4 feat = "nel" #@param ['ner', 'nel', 'nernel', 'geo']
5 # use api
6 API_KEY = ["file", "data", "lng", "feat"]

```

16. <https://colab.research.google.com>

```
7 url = 'https://ners.jerteh.rs/api'
8 params = dict(key=API_KEY, data=data, lng=lang, feat=feat)
9 res = requests.get(url, params=params)
```

Figure 7 shows the integrated web services on the **Language Resource Switchboard** platform. The input data can be submitted as an XML file in case of bilingual resources, and for monolingual resources a text file can be passed or the text can be entered directly into the provided field of the web application form.

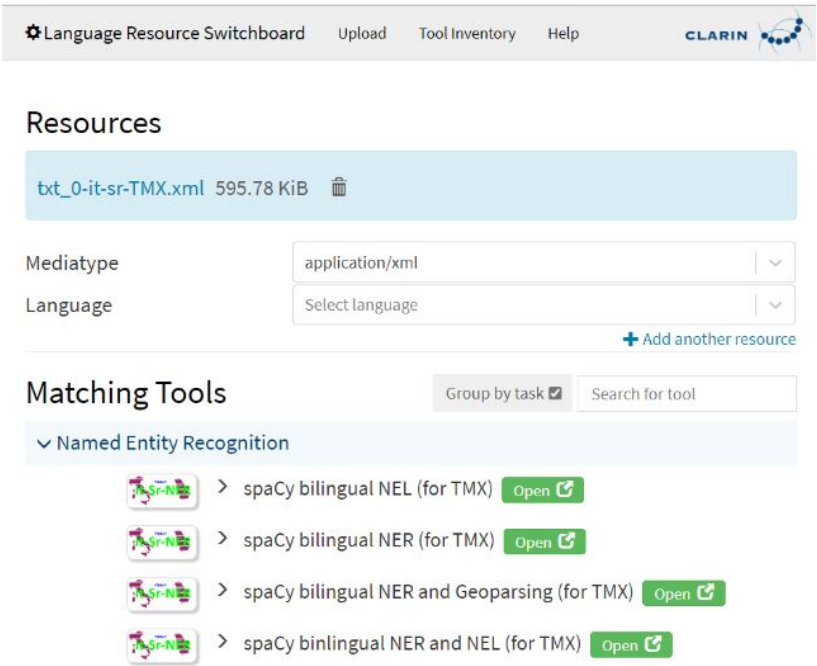


Figure 7. Presentation of the integrated web service on the CLARIN infrastructure

The developed services allow two different formats of display of processing results, HTML and XML documents as illustrated in the following example. Figure 8 shows the processing of text directly written in the textbox for

Serbian (left) and Italian (right). For both languages, a selection of NER services (number 1 for Serbian and number 5 for Italian) was presented, and as a result of the work, HTML and XML documents were presented (number 2 for Serbian and number 6 for Italian). Using the NEL service (number 3) for the Serbian language, the output is presented in the form of HTML and XML documents (number 4), where it can be seen that this service also provides a description of the entity by mouseover event on the entity.

Process plain text file, TMX, or a string 1

Upload a file here: No file chosen

Or type inpaste the text below:

Srbija je zemlja Balkana, glavni grad je Beograd, kroz koji protiču Dunav i Sava. Predsednik je Aleksandar Vučić.

Select processing Language: (if file is not a TMX)

Select parsing method:

Process plain text file, TMX, or a string 5

Upload a file here: No file chosen

Or type inpaste the text below:

La Serbia è un paese balcanico, la capitale è Belgrado, attraverso la quale scorrono il Danubio e la Sava. Il presidente del paese è Aleksandar Vučić.

Select processing Language: (if file is not a TMX)

Select parsing method:

Process plain text file, TMX, or a string 2

PERS LOC ORG DEMO WORK EVENT

Srbija je zemlja Balkana, glavni grad je Beograd, kroz koji protiču Dunav i Sava. Predsednik je Aleksandar Vučić.

Process plain text file, TMX, or a string 6

PERS LOC ORG DEMO WORK EVENT

La Serbia è un paese balcanico, la capitale è Belgrado, attraverso la quale scorrono il Danubio e la Sava. Il presidente del paese è Aleksandar Vučić.

Process plain text file, TMX, or a string 3

PERS LOC ORG DEMO WORK EVENT

Srbija je zemlja Balkana, glavni grad je Beograd, kroz koji protiču Dunav i Sava. Predsednik je Aleksandar Vučić.

Select processing Language: (if file is not a TMX)

Select parsing method:

Process plain text file, TMX, or a string 7

PERS LOC ORG DEMO WORK EVENT

La Serbia è un paese balcanico, la capitale è Belgrado, attraverso la quale scorrono il Danubio e la Sava. Il presidente del paese è Aleksandar Vučić.

Select processing Language: (if file is not a TMX)

Select parsing method:

Process plain text file, TMX, or a string 4

PERS LOC ORG DEMO WORK EVENT

Srbija je zemlja Balkana, glavni grad je Beograd, kroz koji protiču Dunav i Sava. Predsednik je Aleksandar Vučić.

Process plain text file, TMX, or a string 8

PERS LOC ORG DEMO WORK EVENT

La Serbia è un paese balcanico, la capitale è Belgrado, attraverso la quale scorrono il Danubio e la Sava. Il presidente del paese è Aleksandar Vučić.

Figure 8. Display of directly entered text processing

The NER+NEL service (number 7) for the Italian language, shows results under number 8, where it can be seen that this service also provides additional information about the entity using the description of the corresponding item in Wikidata. It can be seen that the XML documents of these two services (number 4 and number 8) differ by the labels of recognized named entities, i.e. NEL service is labeled with the label `<WDT>`, while the NER+NEL annotation uses the tags described in Table 1. Also, the at-

tributes of these two services differ, the NEL service has one more attribute than the NER+NEL service, which is the *label* attribute, describing tagged named entities in English, regardless of the text language. Descriptions in English are the most common, that is because wikidata is the most developed for English, so it is implemented in this version. In the next versions of the service, the approach will be modified so that the description language corresponds to the text language. The HTML preview of both services is the same.

PERS LOC ORG DEMO WORK EVENT		Download XML
n1	Abilissimi agenti di polizia sono stati inviati in <u>America</u> e in <u>Europa</u> , in tutti i principali porti d'imbarco e di sbarco: e sarà ben difficile quel galantuomo sfugga alla loro caccia.	n1 Policijski inspektori, vrlo vešti ljudi, poslani su u <u>Ameriku</u> i <u>Evropu</u> , u svim glavnim lukama i pristaništima: i taj će gospodin teško umaći.
n2	Perciò respinsi a forza il pensiero di <u>Nino</u> , presi <u>Pietro</u> sottobraccio, mormorai: sì, sposiamoci al più presto, voglio andar via di casa, voglio prendere la patente, voglio viaggiare, voglio avere il telefono, la televisione, non ho mai avuto niente.	n2 Zato skupih snagu i izbacih <u>Nina</u> iz misli, uzeh <u>Pjetra</u> pod ruku, promrmljah: da, hajde da se venčamo što pre, želim da odem od kuće, želim da položim vozački ispit, da putujem, da imam telefon, televizor, nikada ništa nisam imala.
n3	Il professor <u>Guido Airota</u> provò a persuadere il figlio, ci furono lunghe telefonate nel corso delle quali con un certo stupore sentii che il vecchio perdeva la calma, alzava la voce.	n3 Profesor <u>Gvido Ajrota</u> pokušao je da ubedi sina da povuče optužbu, vodili su dugo razgovore tokom kojih sam zapanjeno slušala kako njegovog oca izdaje strpljenje, vikao je s druge strane slušalice.
n4	La chiesa non era maestosa come altre che vidi in seguito a <u>Strasburgo</u> , a <u>Chartres</u> , a <u>Bamberga</u> e a <u>Parigi</u> .	n4 Crkva nije bila veličanstvena poput onih koje sam kasnije video u <u>Strazburu</u> , <u>Sartru</u> , <u>Bambergu</u> i <u>Parizu</u> .
n5	Si raccontava che l'arciprete, vedendo quello che accadeva nella kasaba, avesse tenuto nella chiesa di <u>Dobrua</u> una veglia di maledizione, indossando abiti rivoltati e con le candele accese alla rovescia.	n5 Pričalo se da je proto, videći šta se dešava u kasabi, držao u <u>dobruškoj</u> crkvi kletveno bdjenje, u izvrnutoj odeždi i sa svećama zapaljenim naopako.
n6	Crebbe allora l'importanza del ponte sulla <u>Drina</u> quale unico collegamento sicuro tra il pascialato bosniaco e la <u>Serbia</u> .	n6 U tim vremenima važnost mosta kao jedine sigurne veze između <u>bosanskog</u> pašaluka i <u>Srbije</u> neobično je porasla.

Figure 9. Display of bilingual text processing using the NER+NEL service

Figure 9 shows an example of the results of processing a bilingual text (submitted as a TMX document) on the CLARIN platform Language Resource Switchboard using the NER+NEL service. The possibility of displaying the description of the item (determiner) America (Q30) is illustrated, where it can be seen that the recognized named entity America is associated by underlined style.

As in the case of previous web services, geoparsing is available for both bilingual and monolingual resources, where only recognized named entities of LOC class by NER+NEL are displayed on the map. Figure 10 shows

geoparsing for monolingual resources for both languages, Serbian (left) and Italian (right).

For different languages (in this case: Italian and Serbian), there may be differences in the recognition of named entities, as well as differences in geoparsing, for several reasons:

- 1) For the given language, there is no item (headword) for the annotated entity in the knowledge base;
- 2) The named entity in Serbian is not recognized because the system does not recognize inflected forms for the Serbian language (such as cases different from the nominative singular: Србије, Београду, etc.);
- 3) The translation equivalents (Serbian and Italian) are not literal, so the named entity does not appear in one of the equivalents (see segment number 6 in Figure 9).

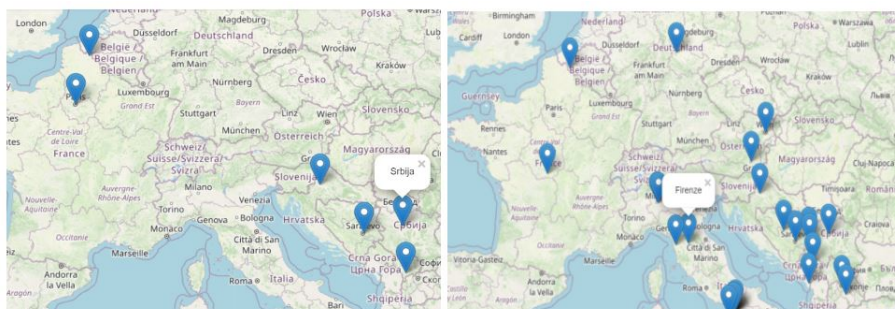


Figure 10. Geoparsing for Serbian language (left) and Italian language (right)

4 Conclusion

In the paper, we presented the results of the project It-Sr-NER: CLARIN compatible NER and geoparsing web services for Italian and Serbian parallel text, supported by the European infrastructure for language resources and technologies CLARIN. The initial motivation for launching a joint project of experts from the University of Turin and JeRTeh, the Society for Language

Resources and Technologies, was to improve the teaching of the Serbian and Italian languages through the creation and publication of a web service for annotating the named entities and displaying them on the map. The lack of specific language technologies for the Serbian language has for years been an obstacle in the application of the results achieved for languages with a larger number of speakers. Individual initiatives that see partial implementation of corpora in teaching are insufficient and do not represent a sufficient stimulus for researchers and teachers in the field of Serbian as a foreign language.

The main primary result of the project is the publication of a set of web services for monolingual and bilingual parallel texts on the CLARIN platform Language Resource Switchboard. At the same time, the project enabled the realization of secondary goals, which are equal in importance to the primary goal, such as the creation and publication of a parallel Italian-Serbian corpus and the construction of a web application and service on the platform of the Society for Language Resources and Technologies JeRTeh. A total of eight services were developed, four each for monolingual and bilingual resources. The text can also be processed by direct input into the provided field at the sentence level or a file entered by the user can be processed. At the same time, linking of named entities with Wikidata and geoparsing are provided. Although Serbian and Italian resources were the focus of the project, the developed services can process texts in 24 languages.

Further research will take place in the direction of organizing additional training in order to popularize web services, as well as their inclusion in teaching. One of the most important goals is to expand the corpus, as well as to improve the model for annotation of named entities and link them to knowledge bases.

Acknowledgment

The authors are thankful for providing the parallelization to: prof. Cvetani Krstev, prof. Duško Vitas, prof. Saša Moderc and Nikola Janković; to reviewers, as well as to the European infrastructure for language resources and technologies CLARIN, on the supported project entitled “It-Sr-NER: Web services for named entities recognition, linking and mapping” within the “Bridging Gaps” call.

References

- Delpeuch, Antonin. 2019. *OpenTapioca: Lightweight Entity Linking for Wiki-data*. <https://doi.org/10.48550/ARXIV.1904.09131>.
- Granger, Sylviane. 2018. “Has Lexicography Reaped the Full Benefit of the (Learner) Corpus Revolution?” In *Proc. of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, edited by Jaka Čibej et al., 17–24. Ljubljana University Press, Faculty of Arts, July.
- Krstev, Cvetana, and Duško Vitas. 2011. “An Aligned English-Serbian Corpus.” In *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality)*, edited by N. Tomović and J. Vujić, I:495–508. Belgrade: Faculty of Philology, University of Belgrade, December. ISBN: 978-86-6153-005-0.
- Moderc, Saša. 2015a. “Elektronski korpus srpskih književnih dela i njihovih prevoda na italijanski jezik” [in Serbian]. 15, *Analiz Filološkog fakulteta* 27 (2): 301–316. ISSN: 0522-8468. <https://doi.org/10.18485/analiff.2015.27.2.15>.
- Moderc, Saša. 2015b. “Su un modo di tradurre l’avverbio serbo “inacche” in italiano: il caso dell’equivalente “altrimenti”.” *Università di Belgrado. In Italica Belgradensia* 1:61–79.
- Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. “Learning multilingual named entity recognition from Wikipedia.” *Artificial Intelligence* 194:151–175. ISSN: 0004-3702. <https://doi.org/https://doi.org/10.1016/j.artint.2012.03.006>.
- Obradović, Ivan, Ranka Stanković, and Miloš Utvić. 2008. “Integrisano okruženje za pripremu paralelizovanog korpusa.” *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, 563–578.
- Perišić, Olja. 2021. “Corpora in the Classroom-the Case of the Serbian Language for Italian Speakers.”
- Perišić Arsić, Olja. 2018. “L’uso dei corpora nella didattica della traduzione: l’esempio del verbo serbo “prijati” e i suoi traduttori italiani,” 49–65.

- Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian NER&Beyond: The Archaic and the Modern Intertwined." In *Proc. of the Int. Conf. RANLP 2021*, 1252–1260. September. <https://aclanthology.org/2021.ranlp-main.141>.
- Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, and Mihailo Škorić. 2021. "Annotation of the Serbian ELTeC Collection." *Infotheca-Journal for Digital Humanitie* 21 (2): 43–59.
- Vitaz, Milica, and Milica Poletanović. 2020. "Data-Driven Learning: The Serbian Case." *EL.LE* (April): 409–422. <https://doi.org/10.30687/ELLE/2280-6792/2019/02/009>.
- Витас, Душко, and Гордана Павловић-Лажетић. 2008. "Ресурси и методе за препознавање именованих ентитета у српском." *Infoteka* 9 (1-2): 33–40.