OntoLex Publication Made Easy: A Dataset of Verbal Aspectual Pairs for Bosnian, Croatian and Serbian

Ranka Stanković, Maxim Ionov, Medina Bajtarević, Lorena Ninčević

OntoLex Publication Made Easy: A Dataset of Verbal Aspectual Pairs for Bosnian, Croatian and Serbian

Ranka Stanković[1], **Maxim Ionov**[2], **Medina Bajtarević**[3], **Lorena Ninčević**[4]
[1]University of Belgrade, Faculty of Mining and Geology, Serbia, ranka.stankovic@rgf.bg.ac.rs
[2]University of Cologne, Cologne Center for eHumanities, Germany, mionov@uni-koeln.de,
[3]Bosnia and Herzegovina, medina.bajtarevic@gmail.com,
[4]University of Zagreb, Faculty of Humanities and Social Sciences, Croatia, lnincevi@ffzg.unizg.hr

Abstract

This paper introduces a novel language resource for retrieving and researching verbal aspectual pairs in BCS (Bosnian, Croatian, and Serbian) created using Linguistic Linked Open Data (LLOD) principles. As there is no resource to help learners of Bosnian, Croatian, and Serbian as foreign languages to recognize the aspect of a verb or its pairs, we have created a new resource that will provide users with information about the aspect, as well as the link to a verb's aspectual counterparts. This resource also contains external links to monolingual dictionaries, Wordnet, and BabelNet. We believe it will be useful for research in the field of aspectology, as well as machine translation and other NLP tasks. Using this resource as an example, we also propose a sustainable approach to publishing small to moderate LLOD resources on the Web, both in a user-friendly way and according to the Linked Data principles.

Keywords: verbal aspect, Linguistic Linked Open Data, BCS

1. Introduction

One of the most difficult properties for the learners of Slavic languages is the verbal aspect. When speaking or writing, the L2 learners/speakers of BCS have to choose whether the appropriate aspect in the given context is perfective or imperfective. After that, they have to recall the form for the appropriate aspect and then conjugate it accordingly. When reading or listening, the learners have to recognize the aspect used. The lack of learning resources makes this task even more difficult. Therefore, we have decided to create a resource to help BCS learners learn the verbal aspect.

1.1. Motivation

Generally, there is a lack of language learning resources for Bosnian, Croatian, and Serbian. The available digital dictionaries do provide information about a verb's aspect, however, their aspectual counterparts are not included in the entry. The users simply need to know the counterpart. Moreover, if learners encounter a new verb, there are no ways to distinguish its aspect. Indeed, perfectivization happens mostly by prefixation, but there can be a secondary imperfectivization in which we would have a verb with a prefix but with a suffix added subsequently, which then makes it imperfective. Sometimes perfective verbs are longer, and other times imperfective verbs are longer. There can be cases where both verbs are of the same length, e.g. *'odmarati se'* (to be resting, imperfective) and *'odmoriti se'* (to rest, perfective). There are also cases where the perfective pair is formed with the prefix but it is also made reflexive. There are no ways for learners to know all this, and this is the reason why the aspect can be one of the most disliked features of Slavic languages for learners.

In addition to being a language learning resource, our dataset can also be used in research. For instance, retrieving all the verbs derived with a certain prefix can be used for much more systematic research of the nuanced meanings given by it.

With the aspectual information provided, our resource can also be used to aid in NLP applications, such as machine translation. Sometimes, the aspectual information is not translated well, and it is not possible to infer additional temporal information. Aspect also conveys information about the duration of an event, completion, or frequency, and if it is not translated well, much of the meaning is lost. Therefore, having aspectual information included in the translation process could greatly enrich it. It would also greatly improve temporal reasoning in NLP.

For example, Google Translate gives the following translations of the sentences below. In example (1) the aspect is perfective and the translation is correct. However, example (2) is imperfective and the translation should be *'Were you reading the book?'* or *'Have you been reading the book?'* Although there are indeed contexts in which the translation in the example (2) would be appropriate. In example (3), even with the aspectual marker *'how long'*, which makes the aspect imperfective,

LDL: Linked Data in Linguistics @ LREC-COLING 2024, pages 108–114

we get a perfective translation.

(1) *'Jesi li pročitala knjigu?' – 'Have you read the book?'*

(2) *'Jesi li čitala knjigu?' – 'Have you read the book?'*

(3) *'Koliko dugo si čitala knjigu?'– 'How long did you read the book?'*

1.2. About Aspect

Aspect is a grammatical category that provides information on whether an action is completed, repeated, or in progress. Slavic languages can have perfective and imperfective verbal aspects. To simplify, the perfective aspect is used for single, completed actions, while the imperfective aspect is used to express actions that are ongoing, habitual or repeated. Perfective verbs are formed from their imperfective counterparts mostly by adding a prefix. Imperfective verbs are formed from perfective ones mainly by adding suffixes. There are also biaspectual verbs that can be used as both perfective and imperfective.

Generally, when prefixes are added, the meaning of the root is changed. Semantic changes caused by prefixes can be neutral, sublexical, and genuine lexical modification (Sussex and Cubberley, 2006). Richardson (2007) calls them purely perfectivizing, superlexical, and lexical prefixes, respectively. Most verbs usually have only one neutral prefix (Sussex and Cubberley, 2006). For example, in the case of *'pisati'* (write, imperfective) prefix *'na-'* would be the neutral prefix which produces *'napisati'* (to complete writing, perfective). If we add the prefix *'pre-'* - we would get *'prepisati'* (to copy by writing, perfective) which changes the meaning of the verb slightly and that would be a sublexical change.

For our resource, we decided to collect only the perfectives formed by purely perfectivizing and sublexical prefixes. Since this resource is made for learners of these languages, we decided not to match the verbs that have undergone a considerable semantic modification. For example, we decided not to match *'ispraviti'* (to correct, to align, perfective) with *'praviti'* (to make, imperfective). In cases like this, we decided that there is no sense match. These decisions were made using native speaker intuition with a subsequent check for archaic verbs.

2. Related Work and Resources

A similar resource has been created for the Russian — Database of Russian Verbal Aspect (Borik and Janssen, 2012). It is a part of the Open Source Lexical Information Network (OSLIN) for Russian (Janssen, 2005).[1] Verbs with the same derivational base are linked to each other and classified as perfective, imperfective or biaspectual. However, there are no definitions of the verbs and no links to other resources. The database, however, provides clusters of verbs, that is, a list of all verbs that are morphologically or aspectually related to a base verb.

Samardžić and Miličević-Petrović (2013) have proposed a learner-friendly dictionary of verbal aspects in Serbian, but so far the resource has not been created. In later work, Samardžić and Miličević (2016) have also proposed a framework for the automatic classification of verbal aspect. A dataset of 2000 verbs based on this framework has been created for testing automatic classification. We are planning on using this resource in the future for our database, as it progresses.

3. Data Modelling with OntoLex-Morph

Since we originally decided to produce and publish our dataset as LLOD, we chose OntoLex-Lemon (McCrae et al., 2017) to model it, since it is the *de facto* standard for publishing lexical resources in RDF in accordance to the Semantic Web standards. The central element of the model is a `LexicalEntry`, which corresponds to lexemes or dictionary entries (Fig. 1).

Each verb in a pair should be modelled as such and there should be a relation between them, showing that they form an aspectual pair.

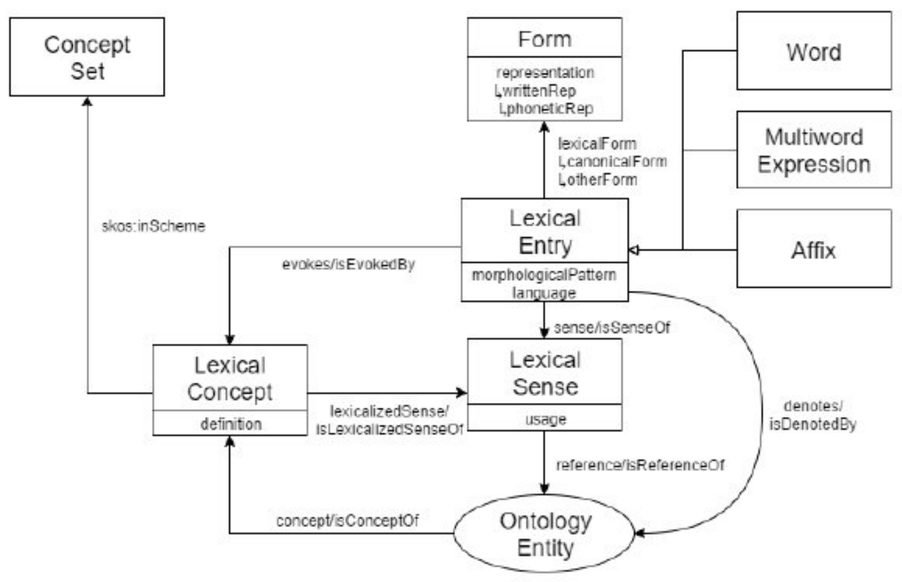

Figure 1: OntoLex-Lemon core model

Generally, to express lexico-semantic relations between two lexical entries, one would use a subclass of a class `LexicalRelation` from the *vartrans* module. But since this relation describes word formation, we decided to use the new OntoLex-Morph module (Chiarcos et al., 2022), which has a way of expressing the derivational relations between words.

[1] `http://ru.oslin.org/`.

OntoLex-Morph (Fig. 2) consists of three parts: derivation (left), inflection (right), and information on how to generate new forms, both for inflection and derivation (top).

In order to represent this relation, we need to create a `WordFormationRelation` between the two entries and additionally specify a `DerivationRule` that provides information on how to create a written representation of a canonical form of the target entry. Below is an example of a relation between an imperfective verb *'parati', to tear apart* and its perfective counterpart *'proparati', to tear apart successfully*:

```
:parati a ontolex:LexicalEntry ;
        ontolex:canonicalForm [
          ontolex:writtenRep "parati"@sr
        ] .
:rel_pro_parati a morph:WordFormationRelation ;
                vartrans:source :parati ;
                vartrans:target :proparati ;
                morph:WordFormationRule
                                  :pro_pref_rule.

:pro_prefix_rule a morph:DerivationRule ;
                 morph:replacement [
                   morph:source "^" ;
                   morph:target "pro"
                 ] .
```

Additionally, we create `Morph` objects for each prefix and add information about them to the corresponding rules:

```
:pro_prefix a ontolex:Morph, lexinfo:Prefix ;
            rdfs:label "pro-"@sr ;
            morph:grammaticalMeaning [
              lexinfo:aspect lexinfo:Perfective
            ] .

:pro_prefix_rule morph:involves :pro_prefix_morph .
```

Having this along with triples describing the original lexical entries, we can either generate lexical entries of the pairs (pre-generate or create them on the fly every time they are requested) using the provided rules, or create them any other way (e.g., if we extract both entries from a database).

Here is an example of a SPARQL CONSTRUCT query that adds the rest based on the data described above:[2]

```
# PREFIXes are removed for brevity
CONSTRUCT {
  ?new_entry a ontolex:LexicalEntry ;
             ontolex:canonicalForm ?new_form ;
             decomp:subTerm ?prefix, ?source_entry .
  ?new_form a ontolex:Form ;
            ontolex:writtenRep ?new_string ;
}
WHERE {
  ?source_entry ontolex:canonicalForm ?source_form ;
  ?source_form ontolex:writtenRep ?base_string .

  ?wfRel a morph:WordFormationRelation ;
         vartrans:source ?source_entry ;
         vartrans:target ?new_entry ;
         morph:WordFormationRule ?rule .

  ?rule a morph:DerivationRule ;
      morph:replacement/morph:source ?srcPattern;
      morph:replacement/morph:target ?dstPattern;
      morph:involves ?prefix .

  ?prefix morph:grammaticalMeaning [
      ?pred ?obj ;
  ] .

  BIND(URI(CONCAT(STR(?new_entry), "_form"))
           AS ?new_form)
  BIND(REPLACE(?base_string, ?srcPattern, ?dstPattern)
           AS ?new_string)
}
```

The output of the query for this example is the following:

```
:proparati a ontolex:LexicalEntry;
    ontolex:canonicalForm :proparati_form;
    decomp:subTerm :pro_prefix, :parati.
:proparati_form a ontolex:Form;
    ontolex:writtenRep "proparati"@sr.
```

4. Dataset Description and Conversion

For this paper, we decided to limit the database to aspectual pairs only formed by perfectivization. We extracted prefixes and the verbs they modify from Leximirka (Stanković et al., 2021), a lexicographic database with a web application for developing, managing and exploring lexicographic data in Serbian. It enables lexical entry control, automatic vocabulary enrichment, multiuser work, and establishment of relations among lexical entries.

Figure 3 shows the source data in the interface: on the left there is an imperfective verb *'raditi'* and it can be seen that it has several perfective pairs linked to it: *'proraditi'*, *'izraditi'*, *'zaraditi'*, *'naraditi'*, *'odraditi'*, *'doraditi'*. In the lower part of the panel, there are available markers: '+Imperf+Tr+It+Iref' meaning (i) imperfective, (ii) can be both transitive and intransitive, and (iii) non-reflexive. For the verb on the right, *'naraditi'*, markers are '+Perf+It+Ref' meaning (i) perfective, (ii) intransitive, and (iii) reflexive.

The rule-based system enables automatic linking between lexical entries in several different ways. One of them related to the linking perfective-imperfective pairs was used for this research.

Leximirka is interlinked with corpora, enabling developers and users to consult concordances and frequency lists for each lexical entry, being single- or multi- word unit, and its collocations. Currently, Leximirka supports Serbian Morphological Dictionaries (Krstev and Vitas, 2006) (Krstev, 2008), but it can support any other language as long as lexical data conform to the lexical database model (Stanković et al., 2018).

The Leximirka data category thesaurus controls the morphological, domain, syntactic, and semantic features that describe the lexical entries. The pronunciation markers are 'Ijk' for ijekavian words and 'Ek' for ekavian-specific words. For example, child *'dijete'* has a marker 'Ijk' for ijekavian, *'dete'* has 'Ek' for the ekavian pronounciation. In addition

[2]The fragment described in this section, SPARQL query for generating derived forms and the full dataset are available at https://github.com/max-ionov/aspect-db/tree/main/public/docs/ldl2024/.

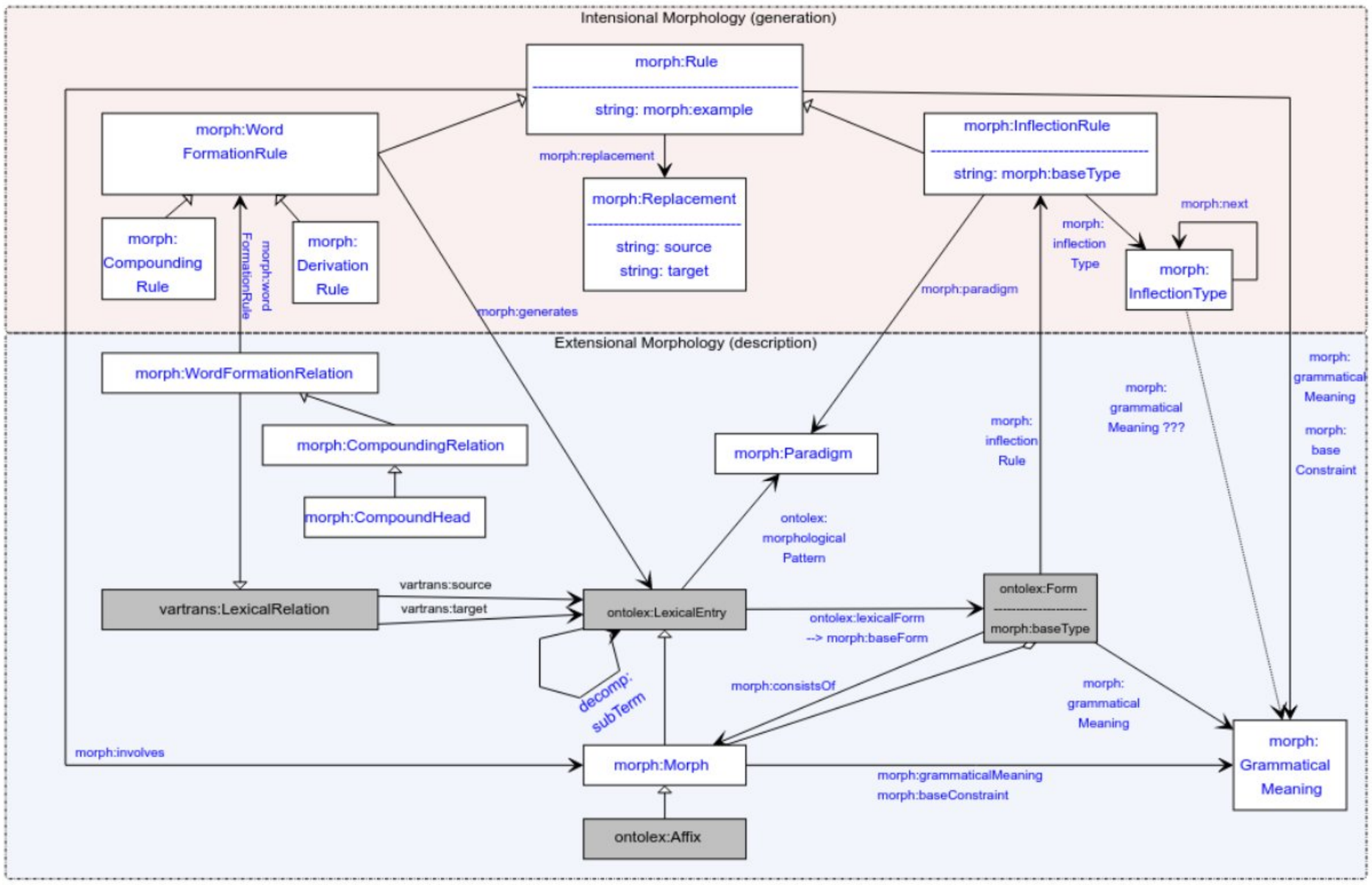

Figure 2: OntoLex-Morph draft model

to that, there is a marker ‘Cr’ for words specific for verbs. Croatian. For example, a masculine form for *film director*, *‘redatelj’*, has a ‘Cr’ marker to indicate that it is specific to Croatian, and *‘reditelj’* is a Serbian equivalent. For imperfective forms, the marker ‘Imperf’ is used, and for perfective forms, ‘Perf’.

So far, we have included verb prefixes *‘pro’*, *‘od’*, *‘ot’*, *‘iz’*, *‘is’*, and *‘na’*. We decided to treat *‘od’* and *‘ot’* as two separate lexical entries in the morph module, even though they might be considered different phonetic variants of the same prefix. The data was extracted from Leximirka using handcrafted rules that were used to establish links between lexical entries. Namely, in SrpMD (Krstev and Vitas, 2006) (Krstev, 2008) markers ‘+Perf’ and ‘+Imperf’ were consistently assigned to appropriate verbs. That information was used in a set of rules for prefixation to establish explicit links between lexical entries. It can be seen that a total of 480 entry pairs were established, and the relation using the rule for the prefix ‘pro’ has 102 pairs.

After the initial extraction, we manually validated the results to check if the verbs also exist in Croatian and Bosnian. For each pair, both verbs were checked separately. The validation was based on native speakers’ intuition with the help of lexicographic resources for archaic verbs. We then performed a manual sense check, in which we made sure that the perfective verbs formed by the prefixes were indeed the pairs of the imperfective

We did not rely on information about verbal reflexivity from Leximirka, but instead, it was manually annotated. More specifically, there were some pairs in which an imperfective verb is not reflexive, but its perfective counterpart is, e.g. *‘raditi’* (work, imperfective) and *‘naraditi se’* (work a lot and get tired, perfective).

The validated results were converted to OntoLex, according to the data model described in Section 3. The total number of entries for each language can be found below:

Language	lexical entries
Bosnian	1071
Croatian	1140
Serbian	1222

Table 1: Number of lexical entries

After converting the dataset, we linked the entries to monolingual Croatian and Bosnian dictionaries. For Croatian, we have linked the entries to the entries on Hrvatski jezični portal (https://hjp.znanje.hr/), and for Bosnian, we have linked the entries to the entries on Sandžakpedija (https://rjecnik.sandzak.com/). There are also links to Croatian Wordnet and BabelNet extracted through ‘Sintaksno-semantički okvir’ (http://www.ss-framework.com/) which is

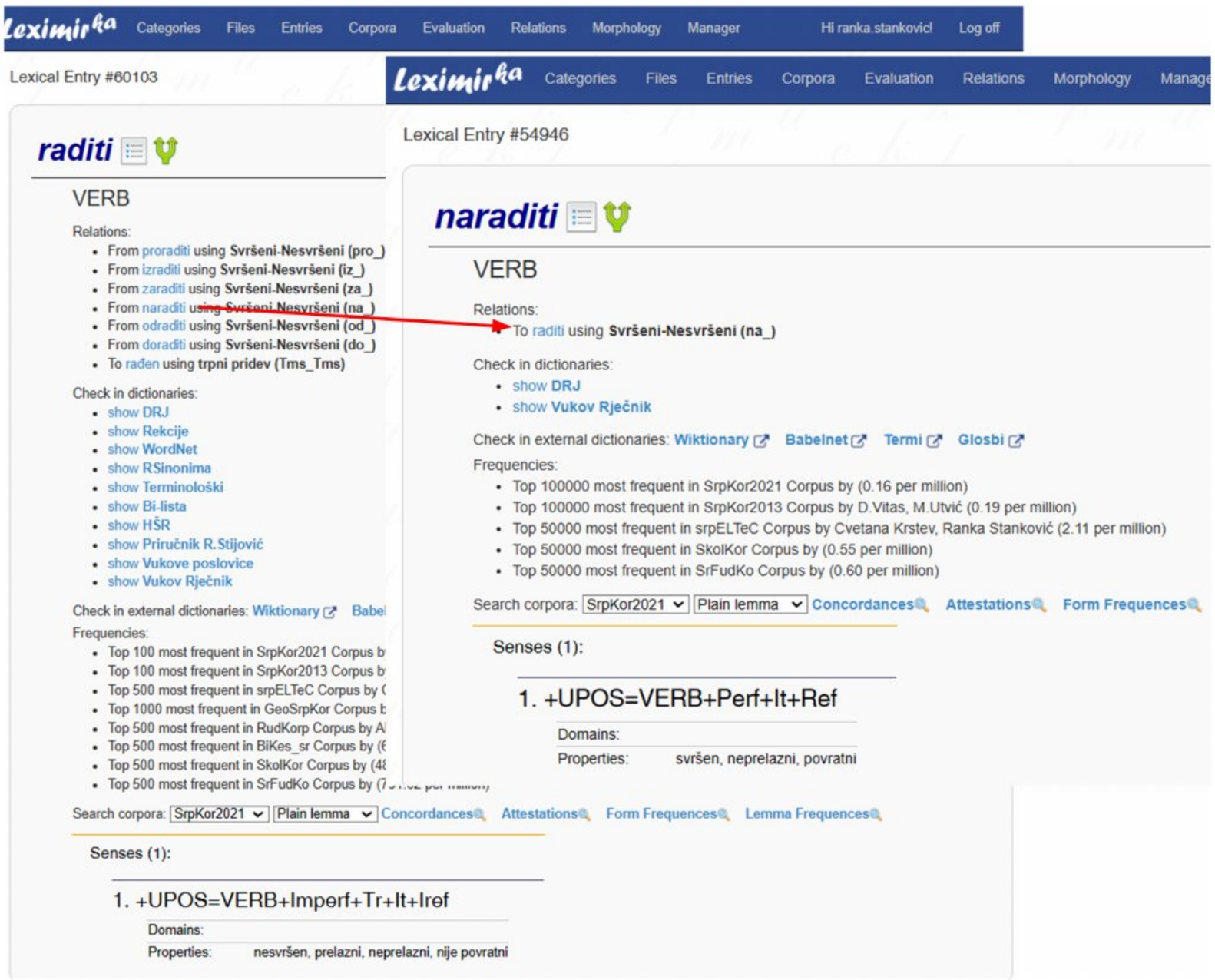

Figure 3: Leximirka: lexical entry for imperfective lexical entry left and one of the perfective pairs on the right side.

itself a part of the LLOD cloud (Orešković, 2019).

5. Deployment as a Static LLOD Resource

One of the problems that often arises in discussions about the adoption of LLOD technology is the lack of infrastructure and high technical requirements for publishing datasets in a way that they can be easily queried (Chiarcos, 2021; Gromann et al., in press, p. 27). The problem can be summarised as follows:[3]

- Data consumers want to be able to access the data in a convenient way, without downloading data dumps and setting up LD infrastructure on their side;
- Data providers do not want to have the burden of supporting SPARQL endpoints, or do not have the resources for sustainable long-term solution.

Most proposals to remedy this argue towards large infrastructures that could take the technological burden, and some already do (e.g., Databus,[4] TriplyDB,[5] and Semantic Media Wiki[6]). On the opposite side, Linked Data Fragments[7] is an effort to put computational load on the side of the end-user, without them needing to pay the price of setting up the infrastructure (Heling and Acosta, 2020).

We argue that this direction — moving the computation to the side of the end-user — has more promise than relying on big infrastructure projects, both from the theoretical and practical sides: On the theoretical side, this is much more aligned with the decentralisation spirit of the World Wide Web and the LOD cloud; on the practical side, creating small independent services that do not have much technical requirements would make publishing Web Data more accessible.

More specifically, our approach is to use statically generated web pages with serialised RDF datasets and client-side SPARQL engine that queries these local datasets (with federated queries to remote ones, if needed).

In this way, just by serving static web sites, data providers can simultaneously distribute their datasets in three different ways with (i) different levels of availability, (ii) usability, and (iii) oriented towards different groups:

[3] These points were confirmed by a poll conducted at a plenary meeting of the COST Action NexusLinguarum which hosted both data consumers and data providers.

[4] https://databus.dbpedia.org/.

[5] https://triply.cc/

[6] https://www.semantic-mediawiki.org/

[7] https://linkeddatafragments.org/

- RDF dumps that can be downloaded and used independently most availability, least usable for end-users, oriented towards people who need unrestricted access to data for parsing or converting;
- A remote endpoint that can be loaded or queried with a SPARQL engine less availability, more usable for end-users, oriented towards people who want to query the data;
- Web page with predefined functionality defined by the data provider least available, most usable for the end-users. For people who want to use the service provided on top of the data.

This approach significantly lowers the requirements to host a website showcasing a dataset in RDF. Due to the fact that many organisations provide free hosting solutions for static web sites, it is not necessary to have access to a server with specialised software installed. And unlike relying on specialised solutions like Triply, this does not create vendor lock since there are many options for hosting a static site.

To showcase our approach, we deployed the database of aspectual pairs as a static website hosted on GitHub Pages.[8]

The page presents the project and allows interaction with the data: searching for an aspectual pair for a verb and for getting all the verbs that use a certain prefix for perfectivisation. Both functions correspond to a SPARQL query that is being run on live data on the side of the client. The dataset does not need to be downloaded and the user does not need to set anything up, since JavaScript-based SPARQL engine Comunica[9] queries the remote file.

The website is being regenerated with every commit to the repository, and using the dynamic routing system, VitePress,[10] the static site generator that we use generates static pages that dereference local URIs of the dataset.

6. Conclusion and Future Work

In this paper, we have presented a new openly available language learning resource for learning verbal aspect in Bosnian, Croatian, and Serbian. The resource is available as an RDF dump, as an endpoint and as an interface, built on top of the endpoint.

Currently, the dataset consists of aspectual pairs in which perfective forms are formed by prefixation. In the future, we will expand this resource to include other types of word formation, for example, imperfectivization done by suffixation.

The second result of this paper is a proposed approach to democratise publishing LLOD datasets by using client-based RDF technology. We believe that this is the way to increase the number of accessible usable LLOD resources and help their adoption for end-user applications.

The possible limits of this approach — how much data can be handled in this way and how performant it can be — is still an open question.

Regardless of that, this approach is a working solution for small- and medium-scale datasets, so it could be adopted by student and small research projects as a way to present and preserve the results.

7. Acknowledgements

This research was supported by the COST Action NexusLinguarum (CA18209) – “European network for Webcentered linguistic data science”. The authors also want to thank the organizers and the lecturers of the 5th Summer Datathon on Linguistic Linked Open Data held in Zaprešić, Croatia in June 2023, where this project started.

8. Bibliographical References

Olga Borik and Maarten Janssen. 2012. A database of russian verbal aspect. In *Proceedings of the conference Russian Verb, St. Petersburg, Russia*.

Christian Chiarcos. 2021. Get! Mimetypes! Right! In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 5:1–5:4, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Christian Chiarcos, Katerina Gkirtzou, Anas Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022. Computational morphology with ontolex-morph. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, page 78.

Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Verginica Mititelu, Liudmila Mockiene, Michael Rosner, et al. in press. Multilinguality and LLOD: A Survey Across Linguistic Description Levels. *Semantic Web Journal*.

[8] https://ionov.me/aspect-db/.
[9] https://comunica.dev/.
[10] https://vitepress.dev/.

Lars Heling and Maribel Acosta. 2020. Cost- and robustness-based query optimization for linked data fragments. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I 19*, pages 238–257. Springer.

Maarten Janssen. 2005. Open source lexical information network. In *Third international workshop on generative approaches to the lexicon*, pages 400–401. Citeseer.

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

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Marko Orešković. 2019. *An Online Syntactic and Semantic Framework for Lexical Relations Extraction Using Natural Language Deterministic Model*. Ph.D. thesis, University of Zagreb. Faculty of Organization and Informatics.

Kylie R Richardson. 2007. *Case and aspect in Slavic*. Oxford University Press.

Tanja Samardžić and Maja Miličević. 2016. A framework for automatic acquisition of croatian and serbian verb aspect from corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Tanja Samardžić and Maja Miličević-Petrović. 2013. Constructing a learner-friendly corpus-based dictionary of serbian verbal aspect. *Primenjena lingvistika*, 14:77–89.

Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. Electronic dictionaries–from file system to lemon based lexical database. In *6th Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science*.

Roland Sussex and Paul Cubberley. 2006. *The slavic languages*. Cambridge University Press.

9. Language Resource References

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

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