From DELA based dictionary to Leximirka lexical database

Biljana Lazić, Mihailo Škorić

From DELA based dictionary to Leximirka lexical database

UDC 811.163.41'322.2 811.163.4'374
DOI 10.18485/infotheca.2019.19.2.4

ABSTRACT: In this paper, we will present an approach for transforming morphological dictionaries from a DELA text format to a lexical database dubbed Leximirka. Considering the benefits of storing data within a database when compared to storing them in textual files, we will outline some of the functionalities that the database has made possible. We will also show how hand-made rules that use category labels lexical entries are marked with can be used to link lexical entries. The initial morphological dictionaries were Serbian Morphological Dictionaries. However, we will show multilingual application of Leximirka using French Morphological Dictionaries.
KEYWORDS: morphological dictionaries, language resources, Leximirka.

PAPER SUBMITTED: 30 August 2019
PAPER ACCEPTED: 28 December 2019

Biljana Lazić
biljana.lazic@rgf.bg.ac.rs

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs

University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

1 Introduction

Prof. Dr. Dusko Vitas and Prof. Dr. Cvetana Krstev started working on the development of Serbian morphological dictionaries more than 25 years ago (Vitas, 1993; Krstev, 1997; Vitas et al., 1993). Morphological dictionaries represent a significant linguistic resource for languages with rich flexion. Therefore, Serbian morphological dictionaries represent a significant resource for Serbian language processing. The importance of this resource is in its multiple applications. Although Serbian morphological dictionaries (SMD) were initially developed for Unitex[1], which enables various complex queries with regular expressions or FSA, their main importance is their reusability. They were used for the basic tasks of word processing, automatic recognition

[1] Unitex is cross-platform Corpus Processing Suite to retrieve data.

of terms, the extraction of time expressions and advanced search of text repositories and libraries.

The morphological dictionaries were developed in the DELA text format (fr. Dictionnaires électroniques du LADL[2]) which will be discussed in Section 2.1. As the dictionaries have grown over the years, in terms of both the number of lexical entries and participants who have assisted in their development, a need for a more efficient system for managing and developing dictionary emerged. For many years, the dictionaries were maintained with the help of the desktop application Leximir and stored in several textual files. The need for an online application based on lexicographic database emerged. In response to these needs, a new application symbolically named Leximirka was developed. Leximirka is not a language dependent application and there is no obstacle for it to be applied in purpose of maintaining DELA dictionaries of other languages.

The Section 3.1 will present more detailed reasons for complex transition to a lexicographic database and to an online application. In Sections 3.2 and 3.3 we will give a description of Leximirka's lexicographic database model and data categories model. An overview of the application segments will be provided in Section 4.1. The possibilities for establishing relations among lexical entries in the database will be introduced in Section 4.2. Multilingual application of Leximirka based on French lexical entries will be presented in Section 4.3. Ideas for further work on application development will be presented in Section 5.

2 Electronic dictionaries

2.1 The DELA text format

Serbian morphological dictionaries are electronic dictionaries primarily intended for machine use. This type of dictionary was first developed for the French language under the influence of linguist Maurice Gross and it is one of the first electronic dictionaries, used before the database notion. These dictionaries also exist for many other languages: German, Bulgarian, Polish, Greek, Russian etc. The system of morphological dictionaries is based on the theory of finite-state automata, namely on morphological and local grammars in the form of finite-state transducers that generate all morphological forms of words in the dictionary (Krstev, 2008).

[2] Laboratoire d'Automatique Documentaire et Linguistique.

Morphological dictionaries consist of both simple and multiword units. The basic components of the simple word morphological vocabulary system are DELAS (fr. DELA de formes simple) and DELAF (fr. DELA de formes Fléchies) (Courtois and Silberztein, 1990). A single DELAS entry consists of a word lemma and its inflective, semantic, and syntactic properties. Here is an example of a simple vocabulary entry for the word lemma "`bibliotekar`" (librarian):

```
bibliotekar,N2+Hum+Prof
```

Record starts with the lemma "`bibliotekar`" , followed by a code defining its Part-of-Speech and type of its inflection paradigm "`N2`", and an optional list of semantic markers for the human being "`+Hum`" and the profession "`+Prof`". DELAF dictionary used for its production consists of automatically generated entries representing all inflectional forms of a DELAS dictionary. It consists of a word form, its lemma, word type designation, semantic markers and a set of grammar categories. Here is an example of one line from a DELAF dictionary:

```
bibliotekarom,bibliotekar.N+Hum+Prof:ms6v
```

This entry starts with the inflective form "`bibliotekarom`" followed by the lemma "`bibliotekar`" and "`N`" (noun). Semantic markers are inserted from the corresponding entry in a DELAS dictionary. Behind the colon, there is a list of grammatical categories defined: the masculine gender "`m`", the singular number "`s`", instrumental case - "`6`" and animateness tag for living beings - "`v`".

The basic components of the morphological dictionaries of multiword units are the dictionaries DELAC (fr. DELA de formes composés) and DELACF (fr. DELA de formes Composées Fléchies). The following is an example of a DELAC dictionary entry for "`fakultetski bibliotekar`" (faculty librarian) (Savary, 2009):

```
fakultetski(fakultetski.A2:adms1g) bibliotekar(bibliotekar.N2:ms1v),
NC\_AXN+Hum+Prof+DOM=BI
```

The part before the comma "`fakultetski(fakultetski.A2:adms1g) bibliotekar(bibliotekar.N2:ms1v)`" represents the lemma. The precise morphological information about a particular component of a MWU is given in parenthesis. This is followed by the PoS and inflective class label "`NC_AXN3`" of MWU, which models the relations between MWU constituents

using FSTs. A list of semantic markers for a human being "`+Hum`", for profession "`+Prof`", for a compound word "`+Comp`", and a tags "`+DOM=BI`" (BI stands for a library and information science domain).

The following is an entry from the DELACF dictionary describing one inflectional form of multiword unit: "`fakultetske bibliotekare, fakultetski (fakultetski.A2:adms1g) bibliotekar (librarian.N2:ms1v).NC:mp4`". The first part - "`fakultetske bibliotekare`" is a word form which is followed by the lemma "`fakultetski (fakultetski.A2:adms1g) bibliotekar (librarian.N2:ms1v)`". This is followed by the code for compound nouns "`NC`" and grammatical categories for the masculine "`m`", the plural "`p`", accusative case "`4`" and the animateness tag for living beings - "`v`". For the sake of simplicity, semantic and domain markers were omitted in the example.

All types of dictionaries were stored and used in the form of textual files whose number has grown significantly (with over a 100 of them at the moment).

2.2 The TEI, the LMF standards, Lemon, Data categories

While choosing a lexicographic database model, care was taken to standardize the data from the morphological dictionaries and to make them interoperable and reusable. Three standards for lexical information have been considered: Guidelines for Electronic Text Encoding and Interchange, Text Encoding Initiative (TEI)[3], Lexical Markup Framework (LMF)[4] and the *Lemon* model[5]. Although Chapter 9 of the TEI Guidelines addresses the issue of dictionary encoding, they only recently address the specificities of ontologies and web resources. The lexicographers' view is that TEI guidelines are more appropriate for encoding traditional dictionaries intended for human use. This does not mean that the situation is not about to change, because there is an interest in linking to the Simple Knowledge Organization System (SKOS) ontologies (Declerck et al., 2010) and the *Lemon* model within the community that uses TEI. At the same time, a new version of the vocabulary chapter, called TEI Lex-0 (Bański et al., 2017), is currently being developed. On the other hand, the LMF and *Lemon* models are more adapted for dictionaries used for natural language processing - NLP.

[3] TEI
[4] LMF
[5] Lemon

The LMF prescribes a standardized framework for recording linguistic information in computer lexicons and is based on the Standard ISO 24613: 2008 (Language Resource Management - Lexical Markup Framework - LMF). LMF is designed for lexicons specially designed for Natural Language Processing and Machine-Readable Dictionaries. LMF specification is represented as a subset of UML (Unified Modeling Language) language that provides linguistic description. The LMF consists of mandatory Core package and additional packages: Morphology Extension, NLP Multiword Expression Patterns, Machine Readable Dictionary, NLP syntax, NLP Semantic Extension and NLP Multilingual Notations. LMF is suitable for encoding morphological, semantic and grammatical aspect of lexical entry. The *Lemon* was modeled after the LMF, but with the idea of compensating the LMF shortcomings in dealing with externally standardized vocabularies and ontologies (e.g. by defining morphological categories and synsets) (McCrae et al., 2012). The *Lemon* model is concise, descriptive, modular and RDF based. At the time of making Leximirka database, *Lemon* model consisted of five modules: Ontology-lexicon interface – ontolex, Syntax and Semantics – synsem, Decomposition – decomp, Variation and Translation – vartrans and Linguistic Metadata – lime. The most commonly used module is ontolex that describes lexical entry (morphological, semantic and ontological description).

3 Transition to lexical database

3.1 Motivation

Automatisation of the management of Serbian Morphological Dictionaries started with the implementation of the Workstation for Lexical Resources WS4LR (Krstev et al., 2006). This single user desktop application later renamed Leximir has various useful functions. It is possible to distribute vocabularies in multiple files, extract subsets of lemmas according to various information assigned to DELAS entries. The application used several Unitex modules that enable the production of DELAF forms for each selected DELAS form (for paradigm checking) or the production of a complete DELAF dictionary from a chosen DELAS file. Opportunities for working with dictionaries of MWUs are also available. The most important one is the automatic generation of the complex DELAC lemmas from a simple list of their basic forms.

After years of working on dictionary expansion, the number of lexical records and categories (semantic and syntactic tags) has increased, as has

the number of files, but also contributors, participating in the workflow. The Leximir application which is first and foremost a desktop application, cannot support multi-user work. There was also a need for controlled entry and verification of data to avoid duplicates and inconsistencies of tags within the lexical record. In response to these needs, first a lexicographic database and then the application Leximirka was created.

The migration of DELA dictionary data into the Leximirka database was done using specific procedures developed in the Leximir application, for practical reasons. The idea was to use the Leximir application until the Leximirka was fully prepared. The Section 3.2 provides more information on how the automatic mapping of DELA dictionary data into a database was conducted.

During the transfer of data to the database some errors were discovered, that inevitably occur when working without automatic control of data entry. It happened that category labels were misspelled due to typographical errors; markers were used for the same concept and one marker for two different concepts, while some lexical records were not marked with adequate markers.

3.2 Leximirka's lexicographic database model

The lexicographic database model of Leximirka, shown in Figure 1, is guided by the *Lemon*, LMF and Data Category Registry catalog (Stanković et al., 2018). The implemented model provides the ability to store lexicographic information in the provided tables and interconnect them with the help of relations.

Figure 1 illustrates the representation of lexical data from DELAS and DELAC dictionary entries (Section 2) within the database model. The boxes marked blue contain data from the DELAS lexical entry "`bibliotekar`" (librarian), while the boxes marked orange contain information that corresponds to the "`fakultetski bibliotekar`" (faculty librarian) DELAC entry. Oval boxes contain additional information that was not recorded in the textual version of DELA dictionaries, for which a place was defined in the database model. The **LEFrequence** table shows that the word "`bibliotekar`" is among the 10,000 most frequent words in the Serbian Corpus of the Serbian Language SrbCorp (version of 122 million words by Duško Vitas and Miloš Utvić)[6]. Information about the Corpus is stored in the **KorpusMeta** table. The **LexicalRelation** table stores information

[6] Corpus of the Serbian Language – SrbCorp

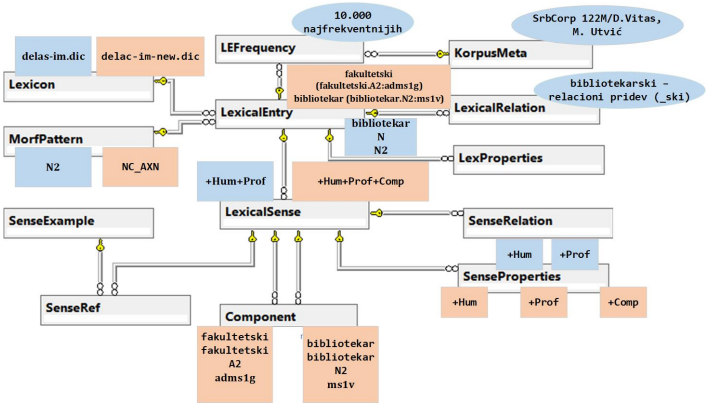

Figure 1. Leximirka's lexicographic database model

.

about relations with other entries in the same database. In this case, this table contains information about an established relation of the lexical entry "`bibliotekar`", called "`relacioni pridev`" (relational adjective), as well as a description of the rules for establishing this relation "`_ski`", with the other lexical entry "`bibliotekarski`" (like librarian). The **LexProperties** table is used to store the values of markers assigned at the lexical entry level. Information about the complete lexical entry (lexical entry identifier, lemma, canonical form, record type, part of speech, morphological class, etc.) is in the **LexicalEntry** table. Inflective class information is in the **MorfPattern** table, while the information about the dictionary to which the lexical entry belongs is in the **Lexicon** table. For one entry in the **Lexicon** table, that is one dictionary, one or more records of the **LexicalEntry** table are connected. This means that one or more lexical entries are part of one dictionary. The meanings of lexical entries are placed in the **LexicalSense** table, while the individual categories that define the meaning are placed in the **SenseProperties** table. A single representation of the **LexicalEntry** table can have multiple meanings or be related to multiple **LexicalSense** table entries. For instance identical lexical entries that share the same morphological class are stored in the same **LexicalEntry** table, but markers that specify their different meanings are stored separately in **LexicalSense** table. Such example would be the noun "`jezik`" (language) that represents the part of the body ("`+DOM=Anatomy`"), but also tool for communication ("`+DOM=Ling`").

The **SenseRef** table stores information about the bibliographic source from which the example of usage originated, while the example itself is in the **SenseExample** table. The **Component** table is used for multiword units, ie. DELAC dictionary, for a precise description of the form of lexical entry components.

3.3 Data categories

The main classes for lexical notations, morphological, syntactic, and semantic categories are controlled by the internal thesaurus of the data categories. Figure 2 shows a model of a lexicographic database that stores various categories of data, that is, grammatical, general, derivational, pronunciation, variational, syntactic, domain, and semantic markers.

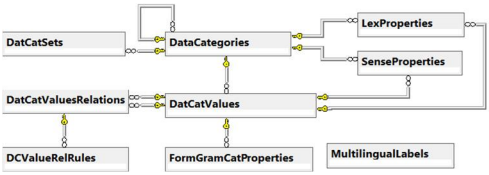

Figure 2. A lexicographic database model for data category information

.

The **DataCategories** table stores information about marker categories, that is, marker type information. The table is linked to itself, allowing for hierarchical representation of categories that are suitable for controlling entries. For example, the category "`derivative markers`" consists of the categories "`derivative noun markers`" and "`derivative markers`". The category "`derivative noun markers`" consists of the markers for the masculine gender "`MG`", the feminine gender "`FG`", the neutral gender "`NG`", etc. It is clear that marker for the feminine gender is derivative noun marker which is the type of derivative marker. Entries in the **DatCatSets** table define which part of speech the category applies to. If the part of speech is not significant for a particular data category, the record in the **DatCatSets** table has the value "`MOT`".

The marker value is written to the **DatCatValues** table. Multiple marker values from the same category form one category that is a record

in the **DataCategories** table. Values "`+Ijk`" and "`+Ijk`" form category "`izgovor`" (pronunciation). The role of **DatCatValuesRelations** table is to enable relationships between markers themselves. An example of such a relationship would be the connection between the Ekavian and the Iekavian entries[7], ie. the link between the "`+Ek`" and "`+Ijk`" markers. The **DCValueRelRules** table describes the set of connection rules that make up one type of relation. Marker values that mark meaning-level records are found as records in the **SenseProperties** table. These are semantic and domain markers. Marker values used to mark LexicalEntry-level records are in the **LexProperties** table. Examples of lexical entry markers are the marker "`+Tr`" for transitive or "`+Iref`" for ireflexive verbs. The **FormGramCatProperties** table contains the grammar categories that occur in the DELAF dictionary. Examples of such grammar codes are "`1`" for the nominative case, "`2`" for the genitive case, "`m`" for the masculine gender, "`f`" for the feminine gender, etc. The **MultilingualLabels** table was created with the idea of presenting meta-language that is used for description of labels, eg. labels and its description could be described in Serbian, English, French, etc. Currently only Serbian language is in use.

4 Leximirka application

4.1 Interface

Leximirka application (http://leximirka.jerteh.rs//) is intended for two types of users covering a wide range of users. Those without a registered account can use it for searching, while registered users can access the management and development interface of the Dictionary.

Unregistered users can search the Leximirka lexicographic database by querying using Latin script. The presentation of the retrieved data is limited to the basic set of data expressed mainly in natural language. Registered users, in line with their privileges, can access different segments of the Leximirka application:

- data categories (option Categories),
- dictionaries (option Lexicons),
- lexical entries (option Entries),
- corpora (option Corpora),

[7] Ekavian dialect the reflection of the Old-Church Slavonic "Jat" is an "e",while in Iekavian it can be "je", "ije" or "i".

– evaluation (option Evaluation) and
– relations (option Relations).

The Data Category segment provides an overview of all the data categories in two ways: tabular and tree-level hierarchical form. Users with the highest level of privileges can edit the existing categories or add new ones that will be used in the dictionaries.

The Lexicons segment offers the ability to view entries, edit metadata about individual dictionaries, add new or export individual dictionaries.

Through the part of the application dedicated to lexical entries it is possible to add new and edit existing entries, as well as search them by lemma or data category markers. Lexical entries that match the specified search criteria appear as rows in the table. The registered user has access to multiple corpus searches (in the MatKorp and SrpKorpRGF corpora). The Mining Corpus (RudKorp) (Tomašević et al., 2018) that can be searched by some predefined queries that retrieve a word searched for in a context. This predefined query could replace "Plain lemma" in the drop-down menu. For example, if a lexical entry describes a noun, the predefined query "`AN`" retrieves occurrences (concordances) in which the word described by an entry follows an adjective. This example is shown in the Appendix - Figure 5. Detailed description of the view panel of a lexical entry is also provided in in the Appendix after the Figure 5. Editing panel for multiword unit can be found in the Appendix of this paper - Figure 6.

The Corpus-related segment is used to access the search for available corpora.

The Evaluation segment is produced to enable the evaluation of the automatically obtained list of candidates for dictionary entries. It is left to the evaluator to decide whether a candidate word meets criteria to enter SMD and it is mainly intended for the creation of lexicons of multiword units or terminology lexicons. The Relations segment is used to define and execute a set of rules necessary to establish relations between pairs of lexical entries 4.2.

4.2 Application example: Establishing relations between lexical entries

The modeled and populated lexicographic database has enabled the automatic connecting of lexical entries. In order to accomplish this task, various procedures were developed using different means: relational query language

for managing SQL databases, FST in Unitex, and C# programming language.

There are several types of relations in the Serbian Morphological Dictionary. Generally, they can be classified as variation and derivation relations, with addition of one pronunciation relation "Ek-Ijk" which is a relation that connects lexical entries of the Ekavian and Iekavian pronunciation (e.g. "`devojka`" vs. "`djevojka`" (girl), "`leto`" vs. "`ljeto`" (summer)).

The pronunciation relation "Ek-Ijk" can be established only if the record containing the Iekavian entry is marked with "`+Ijk`" marker and the record containing the corresponding Ekavian entry has "`+Ek`" marker. This relation can be applied to various PoS. Several rules were defined in order to establish it. In Ekavian dialect the reflection of the Old-Church Slavonic "Jat" is an "e", while in Iekavian it can be "je", "ije" or "i" which can also modify preceding phoneme: *l+je* → *lje* ; *n+je* → *nje* etc. For that reason rules are applied to Iekavian entries since the reflection of "jat" is easier to detect in them.

Some rules are: bri̲jeg+Ijk→ bre̲g, bezbje̲dnost+Ijk→ bezbe̲dnost, sl̲jedeći+Ijk→ sle̲deći .

Variations include relations that connect two lexical entries that represent variant forms, i.e. there is no difference in their meaning, they are only stylistically marked, e.g. "sterilisan" and "sterilizovan" (sterilized), "kava" and "kafa" (coffee), "sufinansiranje" and "sufinanciranje" (cofinanced), etc. At present, there are 43 different variation relations in the Leximirka application and database.

Variant lemmas have appropriate markers assigned that define a rule for establishing a relation. The large part of these relation stems from verbs of foreign origin and way they were adapted to Serbian. One of such pairs and a rule it triggers is: $sterili\underline{zovati}, V + DER = ZovatiSati \rightarrow sterili\underline{sati}$. Since for many of these verbs gerunds (verbal nouns) exist as well as adjectives derived from past participles, the similar rules are applied for them: $sterili\underline{zov}an, A + DER = ZovatiSati \rightarrow sterili\underline{s}an; sterili\underline{zov}anje, N + DER = ZovatiSati \rightarrow sterili\underline{s}anje$.

A number of other variation relations were established that are PoS independent, and were appropriately marked in DELA dictionaries. They include string substitution like in: filo̲zofije+DER=ZS and filo̲sofije+DER=SZ or string omission: h̲alva+DER=Ho and alva+DER=oH (halvah). For them similar connecting rules were expressed.

Derivative relations include those that link derivationally linked lexical entries. These types of relationships include: surname gender motion, e.g. "Škorić" and "Škorićka", verbal nouns from verbs, e.g. "cvetanje" from "cve-

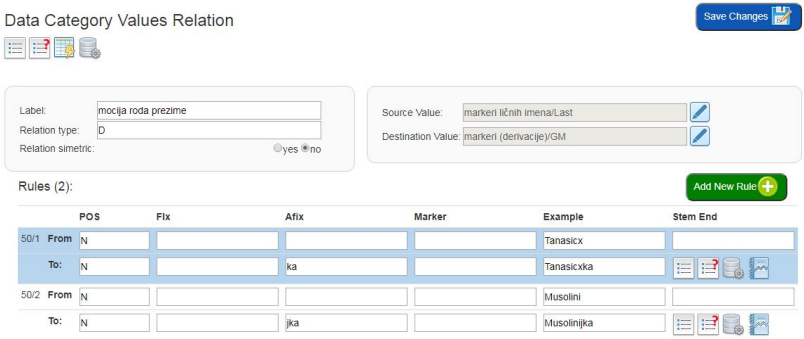

Figure 3. Data Category Values Relation panel

.

tati" (to flower), diminutives, e.g. "kućica" from "kuća" (house), and many others. The connection of these derivationally related entries was enabled by the existence of appropriate markers in DELAS dictionaries, for given examples "`+GM`", "`+VN`" and "`+Dem`", respectively. At present, 21 derivative relations have been established through the Leximirka application.

The functioning of rules that connect derivational entries will be illustrated with gender motion for surnames. Entries "`Škorić`" (`Škorić,N28+NProp+Hum+Last+SR`) and "`Škorićka`" (`Škorićka,N661+GM+NProp+Hum+Last+SR+BASE=Sxkoricx\_N28+DerivAut`) were connected using surname gender motion (_ka) derivational relation based on the rule that the starting and target lemmas should both be nouns with the target record having the suffix "`ka`". Figure 5 illustrates the panel in Leximirka used for connecting entries and the first rule (with blue background) is one that is used to connect "`Škorić`" and "`Škorićka`". The lemmas are connected only if both are marked with `"+Last"` (for surname) and the second lemma has a `"+GM"` marker (gender motion) in DELAS dictionaries. More about these procedures can be read in the paper (Stanković et al., 2018).

Relations are defined through the relationship management segment of the Leximirka application, by filling in general information and by setting a set of rules that more closely define those relations. The rules themselves represent the criteria that both lexical entries must satisfy. Criteria can be set regarding part of speech, inflectional class, affix, or used markers.

4.3 Multilingual application example

In order to prove language independence of Leximirka database, we will show examples of usage on five lexical entries from French Morphological Dictionary:

```
(1) Paris,Paris.N+PR+Toponyme+Ville:ms:fs
(2) Parisien,Parisien.N+PR+Hum+Toponyme+Ville:ms
(3) Parisienne,Parisien.N+PR+Hum+Toponyme+Ville:fs
(4) Parisiennes,Parisien.N+Hum+Toponyme+Ville:fp
(5) Parisiens,Parisien.N+Hum+Toponyme+Ville:mp
```

If we look at the first lexical entry (1), the lemma "`Paris`" and the part of speech "`N`" for a noun are placed in the **LexicalEntry** table of the Leximirka database. The table **LexicalSense** stores information on markers - personal noun "`+PR`", toponyme "`+Toponyme`" and city "`+Ville`". The name of the dictionary that contains this lexical entry "`Prolex-Unitex.dic`" is stored in the table **Lexicon**. These tables are shown in the Figure 1. The form "`Paris`" is the same for singular and both male and female gender and it is stored in the **Form** table. The combination of grammatical categories ("`ms`" – (m) male (s) singular and (fs) – (f) female (s) singular) is stored in the **FormGramCats** table while the separated categories are stored in the **FormGramCatProperties** table.

Lexical entries (2), (3), (4) and (5) represent demonyms for the city of Paris. They are different from the first entry by marker for human being "`+Hum`". All of them represent inflected forms of lemma "`Parisien`". The second entry "`Parisien`" (Parisian) has the same form as lemma and it represents male gender singular, its plural is represented by the lexical entry (5) and the form "`Parisiens`". Form "`Parisienne`" represents female gender singular and form "`Parisiennes`" is its plural. Figure 4 shows how lemma, its inflected forms and grammatical categories are stored in Leximirka database. The lemma "`Parisien`" is in the **LexicalEntry** table and all inflected forms are in the **Form** table. Every form is colored in the same manner as grammatical categories that describe it and all of combinations are stored in **FormGramCats** table. Each separated grammatical category is in **FormGramCatProperties** table. This example approves that the same database can be used for information from different morphological dictionaries in DELA format. The only difference comparing to Serbian example is that Serbian nouns use morphological class that is written in **MorfPattern** table.

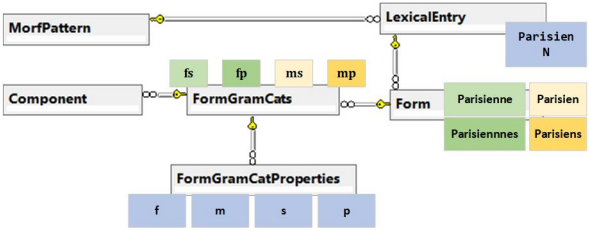

Figure 4. Flective forms in Leximirka database model

.

It is possible to establish derivative relation that links the city with its inhabitant (demonym) between entry (1) and entry (2). This relation is established by the rule that consists of markers "`+Toponyme+Ville`" in the first entry and marker "`+Hum`" in the second entry but also of suffix "`ien`" in the second entry. This rule also finds following pairs of lexical entries in French Morphological Dictionary "`Péone`" and "`Péonien`", "`Plélauff`" and "`Plélauffien`" etc. This relation that links the city with its inhabitant can be enriched with adding other rules. Similarly, other relations can be drawn.

5 Conclusion

The newly established system for managing Serbian Morphological Dictionaries based on the lexicographic database and the online application Leximirka has more advantages over the previously used system Leximir based on DELA dictionary textual files. As noted in the paper, the new system has brought about many advantages in terms of entry control, automatic vocabulary enrichment, multiuser work, and the establishment of relationships among lexical entries. The plan is to add new rules and establish new relations among lexical entries. In the future, work will be focused on defining the format for exporting vocabularies according to user needs, as well as developing a segment dedicated to corpora. The plan is to link lexical records to the WordNet for the Serbian language. It is also envisaged to prepare the data for display in the form of Linked Open Data on the web, which would enable connection with other lexical resources. Since the application is independent of the language for which it is used, it is expected that Leximirka will be used for other languages for which e-dictionaries in DELA format exist.

References

Bański, Piotr, Jack Bowers and Tomaž Erjavec. "TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms". In *Proceedings of eLex 2017 conference: Electronic lexicography in the 21st century*, 485–94. Brno: Lexical Computing CZ s.r.o., 2017, accessed September 1, 2018. https://elex.link/elex2017/wp-content/uploads/2017/09/paper29.pdf

Courtois, Blandine and Max Silberztein. "Dictionnaires électroniques du français". *Langue française* Vol. 87, no. 1 (1990): 11–22

Declerck, Thierry, Karlheinz Mörth and Eveline Wand-Vogt. "A SKOS-Based Schema for TEI-Encoded Dictionaries". In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 414–17, 2010, accessed September 1, 2018. https://www.researchgate.net/publication/265297624_A_SKOS-based_Schema_for_TEI-encoded_Dictionaries

Gross, Maurice. "The construction of local grammars". *In Finite State Language Processing eds. Emmanuel Roche and Yves Schabs* (1997): 329–354, accessed September 1, 2015. https://halshs.archives-ouvertes.fr/halshs-00278316/document

Krstev, Cvetana. "Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije". Phdthesis, Univerzitet u Beogradu, Matematički fakultet, 1997

Krstev, Cvetana. *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade, 2008

Krstev, Cvetana, Ranka Stanković, Duško Vitas and Ivan Obradović. "WS4LR - a Worksation for Lexical Resources". In *Proceedings of the Fifth Interantional Conference on Language Resources and Evaluation*, 1692–1697, 2006. http://poincare.matf.bg.ac.rs/~cvetana/biblio/Krstev_467_new.pdf

McCrae, John, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck et al.. *The Lemon Cookbook*, 2012, accessed September 1, 2018. http://lemon-model.net/lemon-cookbook.pdf

Paumier, Sébastien. *Unitex User Manual*, Université Paris-Est Marne-la-Vallée, 2016

Savary, Agata. "Multiflex: A Multilingual Finite-State Tool for Multi-Word Units". In *Implementation and Application of Automata, 14th International Conference, CIAA 2009, Sydney, Australia, July 14-17, 2009. Proceedings*, 237–240, 2009. URL https://doi.org/10.1007/978-3-642-02979-0_27

Stanković, Ranka, Cvetana Krstev, Biljana Lazić and Mihailo Škorić. "Electronic Dictionaries – from File System to lemon Based Lexical Database". In *Proceedings of the 11th International Conference on Language Resources and Evaluation - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018)*, McCrae, John P., Christian Chiarcos, Thierry Declerck, Jorge Gracia and Bettina Klimek. Paris, France: European Language Resources Association (ELRA), 2018

Tomašević, Aleksandra, Ranka Stanković, Miloš Utvić, Ivan Obradović and Božo Kolonja. "Managing mining project documentation using human language technology". *The Electronic Library* Vol. 36, no. 6 (2018): 993–1009, URL https://doi.org/10.1108/EL-11-2017-0239

Vitas, Duško. "Matematički model morfologije srpskohrvatskog jezika (imenska fleksija)". Phdthesis, Univerzitet u Beogradu, Matematički fakultet, 1993

Vitas, Duško, Gordana Pavlovic-Lažetić and Cvetana Krstev. "Electronic dictionary and text processing in Serbo-Croatian". In *Sprache - Kommunikation - Informatik: Akten des 26. Linguistischen Kolloquiums, Poznań 1991*, 225–232. Berlin: De Gruyter, 1993.

A Appendix

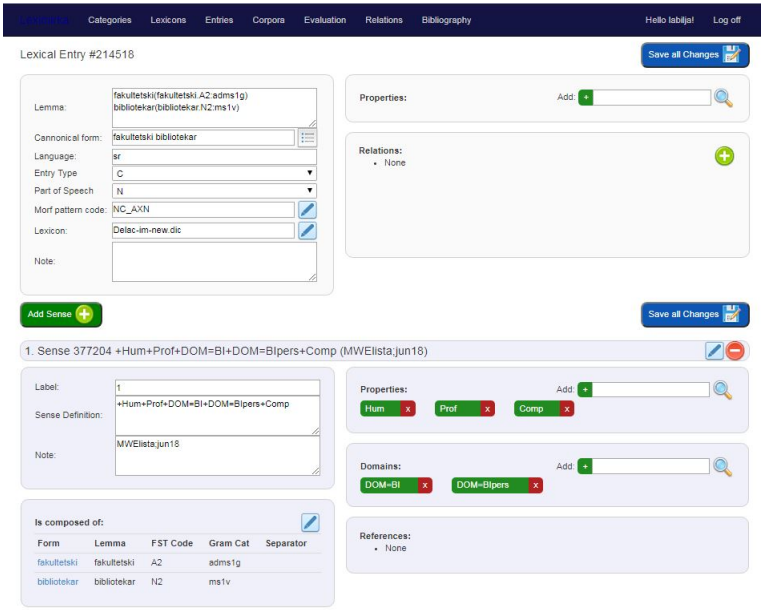

Figure 5. The view of a lexical entry "*bibliotekar*" for logged-in users

.

In Figure 5, the lexical entry "`bibliotekar`" is represented in a view a registered user is given when viewing entries. Unlike an unregistered user who sees only a lemma, its related entries, frequency in the SrpKor, sense expressed in natural language, and reference to multiword units in which the current lemma is a component, the registered user can see all the inflected lemma forms. In addition, a registered user sees editor's notes along with the record, and markers and/or domain tags. In Figure 5, there is a button for displaying all its inflectional forms to the right of the lemma. In the same row there are the shortcuts to the list of all lexical records that are in the same dictionary (delas-im.dic) and the shortcut to the list of lexical records that share the same inflective paradigm (`N2`). In the upper right corner there is the button to access the record edit panel (Edit button).

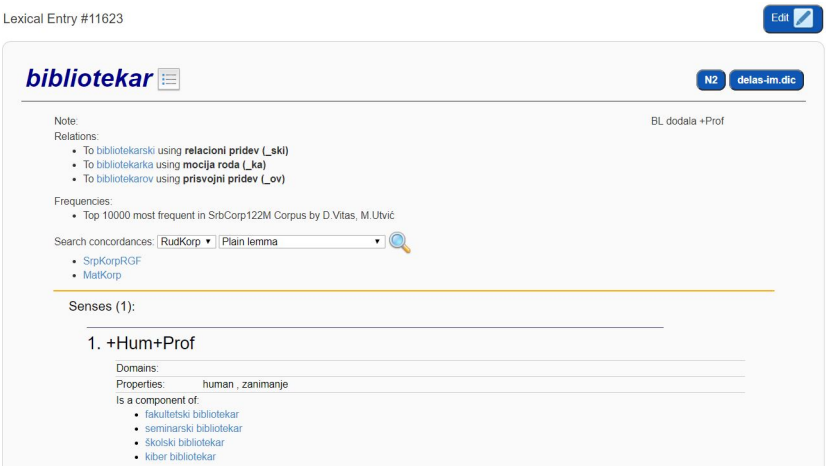

Figure 6. Editing panel of a lexical entry "*fakultetski bibliotekar*"

.

A panel for editing a multiword unit is illustrated with the example "`fakultetski bibliotekar`" (in Figure 6). On this panel privileged users can edit the entry information from the database, as well edit, add or remove other properties. The panel consists of two visually separate parts, the upper part refers to the lemma and the lexical record in general, while the lower part refers to senses.