Electronic Dictionaries – from File System to *lemon* Based Lexical Database

Ranka Stanković, Cvetana Krstev, Biljana Lazić, Mihailo Škorić
{Faculty of Mining and Geology, Faculty of Philology } University of Belgrade
{Djušina 7, Studentski trg 3} Belgrade, Serbia
{ranka.stankovic, biljana.lazic, mihailo.skoric}@rgf.bg.ac.rs, cvetana@matf.bg.ac.rs

Abstract

In this paper we present our approach for lexical data migration from textual e-dictionaries to a lexical database. After years of development, Serbian Morphological Dictionaries (SMD), developed as a system of textual files, have become a large and complex lexical resource. As a consequence, LeXimir, the application that has been used for SMD development and management, was no more suitable. We thus started developing an on-line application for dictionary development and management, based on a central lexical data repository (lexical database). In this paper we present the model for the SMD lexical database developed following the *lemon* model, and the thesaurus of data categories, to be used for enabling links to other (lexical) data. The new database offers various possibilities for improvement of SMD, e.g. control of data consistency and introduction of explicit relations between lexical entries. Besides the procedure used for mapping the existing data model to the new one, we present sets of rules developed to establish relations between lexical entries. We also present some additional improvements – automatic generation of dictionary candidates, with their lexical and derivation variants. This automatic procedure enabled migration of all 26 simple word and 15 multi-word unit Serbian dictionary files with more than 150,000 lexical entries.

Keywords: lexical database, lemon, electronic dictionaries, lexical model, lexical relations

1. Introduction

An application dubbed WS4LR (Krstev et al., 2006), subsequently upgraded and renamed LeXimir (Stanković, Ranka and Krstev, Cvetana, 2016), was designed and implemented for the purpose of further development and management of morphological electronic dictionaries of Serbian (SMD), presented in more details in Section 3.. However, with the growing number of dictionary developers, and given the variety of dictionaries and information stored in them (proper names, domain-specific terms, etc.), the need arose for a more robust application. The main shortcoming of LeXimir, being a desktop application, was that dictionary updates by one user could not be synchronized with other users in real time. Thus, we decided to develop a web application for dictionary management, and enhance the development environment from singe-user to multi-user. In addition to that, LeXimir did not offer support for complex constraints that the development of large dictionaries with rich information needs. The format used in LeXimir did not support the establishment of relations between lexical entries, nor cross-linking with other lexical models, such as Serbian WordNet, another important lexical resource for Serbian (Koeva et al., 2008). This was the main motivation for transforming SMD dictionaries from the existing file system to a *lemon* based lexical database. The model for this lexical database was developed in compliance with the state of the art standards for lexical resources. In this paper we describe how the lexical database was designed following the *lemon* model. We also present how dictionaries were automatically improved and enriched by introducing new lexical entries and/or lexical relations, and by checking the existing ones.

An NLP lexicon has little in common with human-oriented e-dictionary. Data structures in these two types of e-dictionaries are quite different. However, it proved to be very useful to use NLP applications and components in human-oriented e-dictionaries. There are also some NLP-lexicons that can be used by humans. One of such examples is WordNet. A growing number of e-dictionaries pinpointed the need for data standardization, interchange and reusability. In addition to that, the development of the Semantic Web emphasized the importance of enriching ontologies with lexical information. These developments motivated the NLP community to join efforts in standardization. The resulted are widely-used guidelines and standards for dictionary description and lexical databases such as TEI (Tutin and Véronis, 1998), LexInfo (Cimiano et al., 2011), LMF (Francopoulo, 2013), *lemon* (McCrae et al., 2011) etc. The *lemon* model was implemented in several well-known and widely used resources (BabelNet, DBpedia, WordNet), proving that it can be useful in bringing complementary lexical resources together within a single framework.

2. Related work

In order to develop a concrete and general model of dictionaries, it is essential to distinguish between the formal model itself and the encoding or database schema that may ultimately instantiate it (Ide et al., 2000). Having in mind interoperability and standardization issues, three options for the lexical model were considered. The first one were TEI (Text Encoding Initiative) Guidelines for dictionary description. TEI is a widely accepted standard for text encoding that proposes solutions for many text types, one of them being dictionaries. However, it seems that TEI is more often used for traditional human-oriented digitized dictionaries (Khemakhem et al., 2017, Bański et al., 2017).

The second option considered was the LMF (Lexical Markup Framework) model, as it pays special attention to language resources interoperability and re-usability. It provides description of lexical objects, including morphological, syntactic and semantic aspects (McCrae et al., 2012).

This model offers special solutions for the description of lexical information that is used in NLP. Many papers present examples of converting different lexical resources, such as monolingual (Attia et al., 2010) and bilingual (Maks et al., 2008) lexicons, to LMF based multifunctional and reusable electronic lexical databases. LeXimir provided for export of e-dictionaries to XML files compliant to LMF model, but further exploitation of these files was not implemented, neither for lexical database development nor for further processing (Stanković et al., 2013).

Finally we considered the *lemon* model (Lexicon Model for Ontologies), which was derived from LMF, and has been designed for ontology lexicons on the Semantic Web. It is aimed at enriching the conceptualization represented by a given ontology by means of a lexico-terminological layer (McCrae et al., 2012). In order to enable sharing on the semantic web, and for interface with tools *lemon* is based on RDF. Its semantic modeling is more lightweight than that of LMF. One of the advantages is that grammatical annotations are obtained by the use of separate linguistic description ontologies (ISOcat (Kemps-Snijders et al., 2008), GOLD (Farrar and Langendoen, 2003), LexInfo (McCrae et al., 2011)).

The *lemon* approach has been successfully used for comprehensive NLP resources (Bosque-Gil et al., 2016, Villegas and Bel, 2015). The *lemon* model was also implemented in well-known resources such as BabelNet and DBpedia. A paper dealing with WordNet conversion to lemon model (McCrae et al., 2011) demonstrated that *lemon* is an interchange format that can be used to bring complementary lexical resources together under a single framework. The main advantage of the *lemon* model for the research outlined in this paper was its support for linking with other (lexical) data and the possibility to access data by using the standardized SPARQL query language. The model presented is based on the *lemon* model, but some modifications and extensions were necessary to enable full migration of complex grammatical structures and numerous inflected forms for Serbian. MULTEX-East lexicons (Krstev et al., 2004) represent another important NLP lexical resource for Serbian, besides Serbian WordNet. However, both of them are not comparable with SMD either in size or in content, which is why SMD was chosen as the first lexicon for Serbian to be converted into a lexical database.

3. Morphological electronic dictionaries

Morphological electronic dictionaries of Serbian for NLP are being developed for many years now (Vitas et al., 1993) (Krstev, Cvetana and Vitas, Duško, 2015). They cover general lexica, proper names (persons and toponyms), general knowledge (famous or fictitious persons, places and organizations), and domain terminology. For practical reasons they are kept in a number of files, according to different criteria.

These dictionaries are in the so-called DELA format: in the dictionary of lemmas each lemma is described in full detail, so that the dictionary of forms containing all necessary grammatical information can be generated from it, and subsequently used in various NLP tasks (Courtois and Silberztein, 1990). A dictionary of lemmas can contain simple-word lemmas (DELAS) or multi-word lemmas (DELACF), producing, respectively, a dictionary of simple-word forms (DELAF) or multi-word forms (DELACF).[1] Traditionally, dictionaries of lemmas are prepared and maintained as one or more textual files, while dictionaries of forms are generated automatically, also as textual files. The structure of a simple word lemma is:

`lemma,POS#fst[+Marker]*`

Mandatory parts of this structure are a lemma, its POS, and identification of a finite-state transducer that will produce all lemma's inflected forms with associated grammatical information (e.d. case, number, gender, etc.). Markers are not mandatory, but they are nevertheless assigned to the majority of lemmas. Formally, they can be of two types:

- *switches*: if a marker of this type is present, then it indicates that a lemma has a certain feature, but if it does not exist, that indicates the absence this feature for the lemma. For instance, the marker `+Hum` indicates that a lemma represents a human being (e.g. `profesorka,N661+Hum` – 'woman professor', as opposed to `krava,N601` – 'cow');
- *attribute/value pairs*: an attribute indicates the type of the feature, while a value makes it more specific. For instance, in the marker `+DOM=Math` the attribute `+DOM` indicates that a lemma is related to a certain domain, whereas the value specifies this domain to be mathematics (e.g. `diedar,N3+DOM=Math`). Values assigned to a certain attribute can belong to a closed set (e.g. `+CC2=RS` is a two character country code marker assigned, for instance, to geopolitical names), or to an open set (e.g. `+Val=Vaughn` is assigned to a surname *Von*, Serbian transcription of the English surname *Vaughn*).

Semantically, markers can be of various types:

- *semantic/ontology* – these markers denote lemmas as belonging to a certain ontological class, e.g. `+Hum` (humans), `+Body` (body parts), etc.;
- *syntactic* – these markers provide some syntactic information about a lemma, e.g. a marker `+Ref` assigned to a verb indicates it is a reflexive verb;
- *pronunciation* – these markers are assigned to lemmas specific to a certain pronunciation, e.g. `+Ek` for Ekavian, `+Ijk` for Ijekavian pronunciation;
- *derivation* – these markers are assigned to lemmas derived from other lemmas, e.g. the marker `+GM` assigned to *profesorka* 'woman professor' denotes that it is derived from *profesor* 'professor' by gender motion; [2]

[1] Serbian e-dictionaries, SMD, have reached a considerable size: they comprise more than 150,000 simple-word lemmas, generating more than 5 million forms and 18,000 multi-word lemmas.

[2] In the lexical database described in this paper these markers are converted from switch to attribute/value markers, e.g. `+DER=GM`.

- *variation* – these markers indicate that a lemma has a variant, and how this variation is produced. In Serbian, many words have lexical variants that do not bear any specific meaning – they may be preferred in certain regions or in certain period of time (Klajn, 2005, Stanojčić and Popović, 2008). For instance, *afirmisati* and *afirmirati* 'to establish' are two such variants, to which markers `+VAR=SatiRati` and `+VAR=RatiSati` are assigned, respectively;

- *domain* – these markers indicate the domain of use of lemmas to which they are assigned;

- *information* – these markers provide some additional information about a lemma, e.g. the lemma *deci*, shortened for 'deciliter', has a marker `+SI=dl` assigned to it, indicating that its abbreviation in the International System of Units is *dl*.

Relations can exist between certain markers. For example, the hyperonymy/hyponymy relation exists between semantic markers: river (`+River`), which is a hydronym (`+Hyd`), which is a geographic concept (`+Top`), and thus all three are assigned to the lemma *Dunav* 'Danube'. Some lemmas are related by some sort of "inverse" relation, which indicates that if one lemma has a certain feature, then at least one other lemma exists with an "inverse" feature. These relations are sometimes explicitly encoded by appropriate markers (e.g. variation and pronunciation markers presented before), while in most cases they are implicit. For instance, lemmas for *profesorka* and *profesorica*, both meaning 'woman profesor' are derived from *profesor*, and they both have a marker `+GM`, while lemma for *profesor* does not have a marker indicating that forms derived from it by gender motion exist.
All the entries in a DELAF dictionary of forms are in the following format:

```
form,lemma[:categories]*
```

where `form` is a simple word form of a `lemma`, represented by its DELAS entry form, and `:categories` are the possible grammatical categories of the word form, each category represented by a single character code (Krstev and Vitas, 2007).
LeXimir, a tool for development and maintenance of e-dictionaries enabled development of Serbian morphological dictionaries in the past decade. However, with the enhancement of dictionaries and enrichment of their content some serious drawbacks of this tool became evident. Besides being a desktop application, discouraging cooperative work, it also does not have appropriate support for the treatment of duplicates (e.g. should *atlas* be one lemma or two lemmas that have same inflectional behavior, one denoting a book with maps and having markers `+Conc+Text`, the other denoting a type of a fabric and having markers `+Conc+Mat`). The consistency check is missing as well (e.g. can a marker `+Hum` be assigned to a lemma whose grammatical category `q` indicates it is inanimate, like *lonac* 'pot'?), as well as a check establishing the correctness of "inverse" relations (e.g. does a variant lemma *duhan* indicated by the marker `+VAR=VH` assigned to a lemma *duvan* 'tobacco' exist?). Finally, the lack of all these features was an impediment to production of special purpose dictionaries: for instance, for purely morphological dictionaries, *atlas* should be one lemma, while for dictionaries aiming at semantic processing, two lemmas are necessary.

4. The Model and Implementation of the Lexical Database

The main goal of the research presented in this paper was to produce a central lexical repository that will enable multiuser distributed management of lexical data, overcoming the main problem of the existing solution – local, single-user editing of dictionaries in textual form. The new lexical database should also enable of its content export in various formats. The Unitex[3] format for DELA dictionaries (dictionaries of lemmas and dictionaries of inflected forms presented in the previous section), supported by LeXimir, will be only one of the formats supported by the lexical database. The database will also provide for automatic production of dictionary editions for different profiles of users: full dictionaries, public-domain oriented, filtered by different criteria (e.g. pronunciation: Ekavian and Ijekavian), etc.
In the new lexical database model for Serbian Morphological Dictionaries, based on the *lemon* model, main classes for lexical entries, morphological, syntactic and semantic features are controlled by the internal thesaurus of data categories, outlined in (Krstev et al., 2010). During the whole period of the development of Serbian morphological dictionaries, the corresponding metadata were documented by a simple textual file. This file was the base for the creation of a dictionary of markers, that is, data categories and their values (Figure 1). Transition to the database that supports the control of field domains revealed inconsistencies among markers: same markers used for different purposes, different markers used for the same purpose, missing markers, markers associated to wrong categories, etc. Presently, there are 23 semantic markers in the database (e.g. `+Hum` for human beings), 17 syntactic (e.g. `+Ref` for reflexive verbs), 24 grammatical (`V` for verbs), with a total of 836 different values. There are also special domain markers (at present 104), which relate the lexical entry (and a particular sense) to its domain of use. For instance, the lexical entry *jezik* 'language, tongue' has three different senses (presently recorded in SMD), and their textual representation in DELA format is:

```
jezik,N9+DOM=Ling//communication media
jezik,N9+Conc+Body+DOM=Anatomy//body part
jezik,N9+Conc+Food+Prod+DOM=Culinary//food
```

Each of these entries is connected to a different domain (linguistics, anatomy, and culinary, respectively).
Since the use of *lemon* is complemented with LexInfo, as an ontology of types, values and properties to be used with the *lemon* model (partially derived from ISOcat), one of the goals was to map categories used in existing SMD to LexInfo, as a catalog of data categories (e.g., to denote gender, number, part of speech, etc.).

[3]Unitex is a lexically-based corpus processing suite that offers strong support for finite-state processing using morphological dictionaries –`http://unitexgramlab.org/`

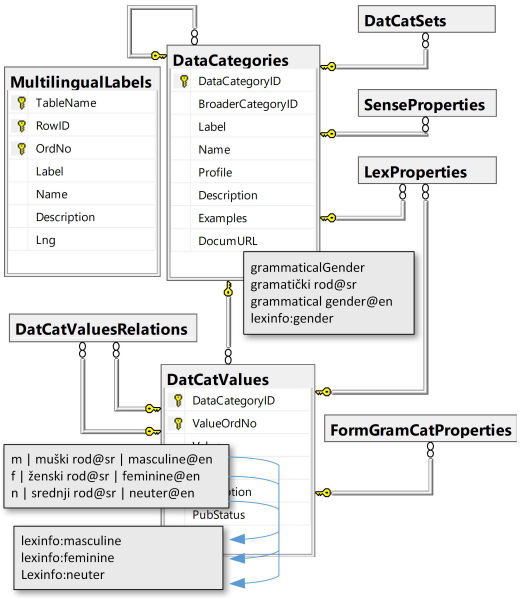

Figure 1: Data categories (markers) dictionary.

The main class of the core of the lexicon model is the class `LexicalEntry`, representing a unit of analysis of the lexicon, which encompasses a set of inflected forms that are grammatically related, and a set of base meanings that are associated with all of these forms (Figure 2). A lexical entry is a (single) word, multi-word expression, acronym or affix with a single part-of-speech, a morphological pattern, or a set of senses.

The `LexicalRelation` class relates lexical variants (for instance, *istorija* and *historija* 'history'), full forms and their abbreviation (e.g. *kilogram* and *kg*), derivationally related lexical entries (e.g. *istorija* and *istorijski* 'relating to the study of history'), and different pronunciations (Ekavian *dete* and Ijekavian *dijete* 'child'). `LexicalSense` is used to represent a particular sense of a lexical entry (e.g. for instance, three senses of *jezik*), and link a lexical entry with an ontology by connecting the set of markers denoting one sense with individual markers in the `SenseProperties` table.[4]

`SenseRelation` provides for connecting various senses with others that are narrower, broader, synonymous and so on, while `SenseRef` and `SenseExample` contain information about provenance and usage.

For languages with rich morphology, such as Serbian, the maintenance of dictionaries of inflected word forms is very important. For instance, inflected forms of *jezik* are: *jezik*, *jezika*, *jeziku*, *jeziče*, *jezikom*, *jezici*, *jezike*, *jezicima*. In the model presented, the table `Forms` is used to store all forms that are inflected from a lemma, together with sets of grammatical categories assigned. Since one lexical form can represent one or more grammatical realization of a lexical entry, it is described with one or more sets of grammatical categories stored in `FormGramCats`. For instance, the form *jezikom* has one set of grammatical categories assigned to it `:ms6q` (the instrumental case, singular), while two sets of grammatical codes are assigned to *jezika*: `:ms2q` and `:mp2q` (the genitive case, singular and plural). In addition, sets of grammatical categories are represented as individual categories in the table `FormGramCatProperties`, as presented in the left side of Figure 2.

The class `Forms` is used in the *lemon* model to indicate a non-semantic relationship between two lexical entries, for instance, cases when a term is derived from another term: "lexical" and "lexicalize". In the model presented, the class `LexicalEntry` is used for canonical forms of different variants, and the class `LexicalRelation` for relations between variants.

Dictionary production in different formats is also envisaged. For instance, compiled dictionaries to be used by Unitex, or textual inflected files to be further utilized by users. RDF serialization (e.g. Turtle, RDF/XML) is under development, and Linguistic Linked Open Data (LLOD) publishing will also be supported, while the same lexical database will be used for query expansion Web APIs used for information retrieval and indexing support. The master lexical data repository is stored in a relational database management system, but the use of triple-stores, e.g. graph databases Neo4J and DBGraph, is being investigated. The use of triple-stores will be read-only in this phase of development, and they will be used for querying and linking to external resources, while CRUD (create, read, update, delete) operations will remain in the relational system, given the required stability and the implementation experience so far.

5. Migration of Dictionary Data

The procedure for transferring data from existing dictionaries into the lemon-based model is integrated in the existing tool for dictionary management LeXimir, in order to support parallel development for a certain period of time, and to enable smooth transition of development environment. The database contains all currently used markers, but these markers have not a "flat" structure anymore, but rather a hierarchical structure that can serve as a controller for domains of some fields in a database.

As previously mentioned, DELAS dictionaries are distributed in more than 40 files for practical reasons, and information about the file a lemma comes from is stored in the `Lexicon` table for development purposes. Lemma entries from a DELAS dictionary are generally mapped to entries in `LexicalEntry` and `LexicalSense` (Figure 2), where a lemma, its POS, the inflectional class (governing production of all inflected forms) are stored in the `LexicalEntry` table, while associated markers – syntactic, semantic, domain and other – can be separated if needed. Identical lexical entries from DELAS sharing the same inflectional class are merged into one `LexicalEntry`, while their semantic markers indicating different senses are separated into more entries. For instance, in the new database `jezik,N9` is an entry in the table `LexicalEntry`, while associated markers that differentiate senses are recorded in the `LexicalSense` table. Entries that are part of a MWU, which is entered in the same tables `LexicalSense` and `LexicalEntry`, are

[4]The terms *class* and *table* are used respectively to indicate a model class and a physical table in a database.

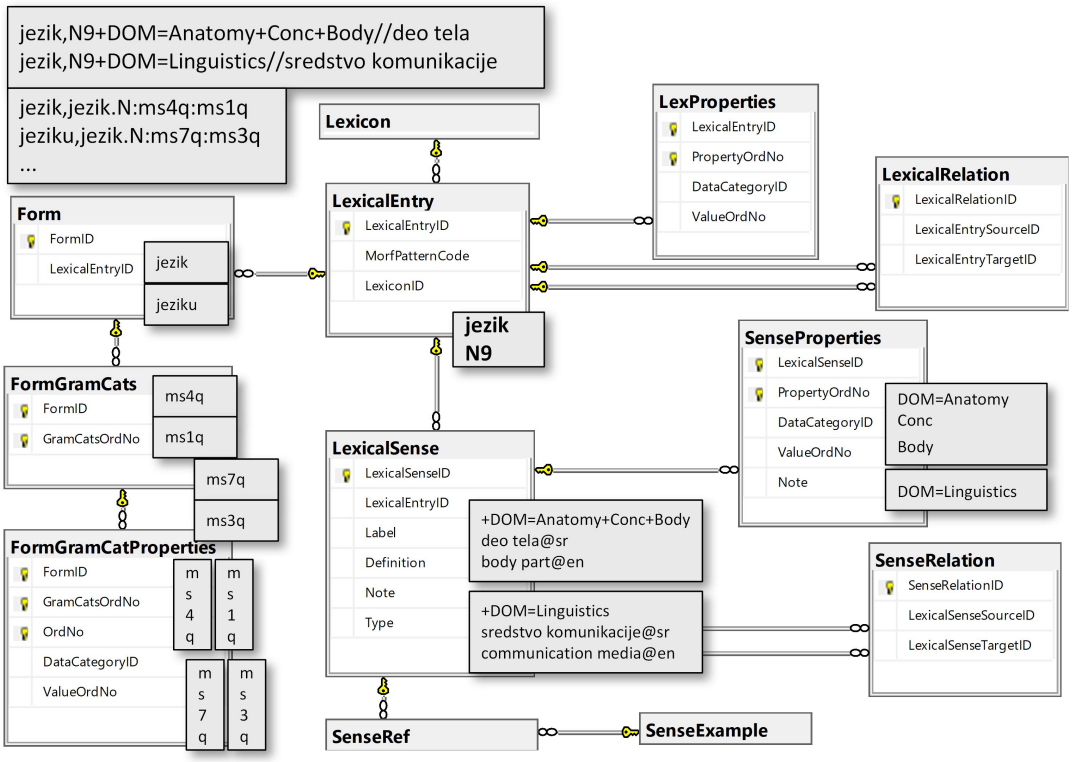

Figure 2: Lexical database core model.

related with the corresponding MWU. Examples of such entries are (simplified):

```
maternji jezik +DOM=Ling      'mother tongue'
jezik za zube  +DOM=Anatomy
     'tongue behind teeth (keep mouth shut)'
teleći jezik   +DOM=Culinary 'veal tongue'
```

The same example in the *lemon* form is:

```
lex_jezik a ontolex:LexicalEntry;
 lexinfo:partOfSpeech lexinfo:Noun;
 jezik ontolex:morphologicalPattern :N9;
 form_jezik ontolex:writtenRep "jezik"@sr;
 ontolex:canonicalForm :form_jezik;
 ontolex:sense :jezik_sense1;
 ontolex:sense :jezik_sense2;
 ontolex:sense :jezik_sense3.

:jezik_sense1 a ontolex:LexicalSense;
 dct:subject
  <http://dbpedia.org/page/Linguistics>;
 ontolex:reference
 <http://dbpedia..../Category:Language>.
:jezik_sense2 a ontolex:LexicalSense;
 dct:subject
  <http://dbpedia.org/page/Cooking>;
 ontolex:reference
  <http://dbpedia.../Tongue_(foodstuff)>.
:jezik_sense3 a ontolex:LexicalSense;
 dct:subject
  <http://dbpedia.org/page/Anatomy>;
 ontolex:reference
  <http://dbpedia.org/page/Tongue>.
```

Sense linking to WordNet synsets is planned, but is not yet implemented.

6. Dictionary improvement based on lexical variations and derivations

6.1. Corrections and Additions

The newly implemented lexical database (presented in Section 4.) introduced new possibilities for the improvement of valuable existing resources. Besides relatively trivial task of finding and correcting all incorrect markers (mostly typos), duplicate markers (denoting same concepts), it enabled the conversion of all markers that indicate links between lemmas (see Section 3.) into true relations between lexical entries. For instance, dictionary entry for *kućica* 'small house' had a marker for the diminutive `+Dem` assigned to it, but no indication of its basic form; at the same time, for the dictionary entry *kuća* 'house' it was not possible to determine whether it had a diminutive and if so, what it was.

Two approaches were used to establish relations between lexical entries. The first approach was used for explicit inverse relations, mostly for lexical variants or two different pronunciations, Ekavian and Ijekavian. In this approach one or more target lemmas are constructed based on the type of the relation, using some simple string matching and replacement, and the newly constructed lemmas had to (a) exist in dictionaries; and (b) have an inverse marker.

For instance, verbs *afirmisati* and *afirmirati* are two variants (the first one being preferred today in Serbian) of the same verb 'to establish'. Similarly, *hleb* and *leb* are two variants of the same noun 'bread' (the second one being

non-literary). The Ijekavian lemma for the Ekavian lemma *devojka* 'girl' is *djevojka*. These lemma pairs were recorded in DELAS entries of e-dictionaries, in the following manner:

```
afirmirati,V1+Imperf+Perf+Tr+Iref+Ref
           +VAR=RatiSati
afirmisati,V21+Imperf+Perf+Tr+Iref+Ref
           +VAR=SatiRati
hleb,N81+VAR=H0+Ek+Conc+Course+Food
     +DOM=Culinary
leb,N81+VAR=0H+Ek+Conc+Food+Prod
     +DOM=Culinary
devojka,N617+Hum+Ek
djevojka,N617+Hum+Ijk
```

The marker `+VAR=RatiSati` indicates that the suffix *-rati* can be substituted in the lemma *afirmirati* by the suffix *-sati* to produce the lemma *afirmisati*, which is recorded in e-dictionaries and has an inverse marker `+VAR=SatiRati` assigned to it. The marker `+VAR=H0` indicates that an *h* can be deleted in the lemma *hleb* to produce the lemma *leb*, which has an inverse marker `+VAR=0H` assigned to it in e-dictionaries. The marker `+Ijk` indicates that the reflection of an Old Slavic *yat* can be substituted in the Ijekavian lemma *djevojka* by *e* to produce the Ekavian lemma *devojka*, which has an inverse marker `+Ek` in e-dictionaries. It should be noted that for each lemma pair all other markers assigned to them are identical. Also, it is sometimes irrelevant which is the initial lemma used for producing the other lemma by substitution/deletion (the first example), while in some other cases one of the lemmas in a pair is a better initial choice (in examples above, lemma containing *h* for markers `+VAR=H0/+VAR=0H`, and Ijekavian lemma for markers `+Ek/+Ijk`). In the first two cases a *variation relation* is established between the pair of lexical entries, while in the third case it is a *pronunciation relation*. Namely, entries for which a lexical variant exists have a special marker in the form `+type=value` where, in this case, `+VAR` indicates a variation marker and `value` indicates a type of variation, which also gives a hint how one variant can be derived from another. As a rule, these relations should be inverse, as is the case with examples given above. With dictionaries maintained as textual files, one could rely only on a developer to enforce this rule.

The second approach is used for implicit inverse relations: a lemma that has a derivation marker is used to generate the source lemma, origin of the derivation, which is then sought in dictionaries. The generation is sometimes quite simple, as is the case with verbal nouns (gerunds) that are derived from most of imperfective verbs (marker `+Imperf` in DELAS dictionaries), and marked with `+VN`. The simple rule here is, to remove verbal noun suffix *-nje* and add an infinitive suffix *-ti*. Also, adjectives (past participles) are derived from most transitive verbs (marker `+Tr` in DELAS dictionaries), and marked with `+PP`. This procedure would establish a *derivation relation* between two above-mentioned verb variants, as well as respective verbal nouns and adjectives, starting from the following four e-dictionary entries:

```
afirmiranje,N300+VN+VAR=RatiSati
afirmiran,A6+PP+VAR=SatiRati
afirmisanje,N300+VN+VAR=SatiRati
afirmisan,A6+VAR=SatiRati
```

However, these verbal nouns and adjectives also come in variation pairs, so a *variation relation* is established between them also by using a procedure similar to the one described above.

6.2. Procedures for establishment of relations

We have developed a set of Unitex graphs, SQL procedures and C# tools to automate the task of explicit linking of existing entries. Even though our main goal was to connect existing entries, these automation tools introduced new possibilities for further expansion and annotation of dictionaries, including detection of missing markers and production of new entries. Here we will present, in more detail, two approaches that have been applied to actually connect lexical entries; first, the approach applied to produce and connect derived entries, and then the approach to connect lexical variants.

Establishing derivation relation is, in general, far from simple. So far, we have dealt with possessive adjectives derived from surnames, leaving other cases – diminutives, augmentatives, relational adjectives, gender motion, and so on – for future work. E-dictionaries contain a large number of surnames, both typical Serbian surnames (close to 18,000) and surnames of foreign origin transcribed according to Serbian orthography (close to 7,500). Possessive adjectives are often derived from surnames, e.g. *Lazić* ← *Lazićev* and *Ešton* ← *Eštonov* (Serbian transcription for 'Ashton'), as well as, in some cases, feminine nouns, *Lazić* ← *Lazićka* and *Ešton* ← *Eštonka* (women with surnames *Lazić* and *Ešton*, respectively). However, only a small number of these related lemmas were actually recorded in e-dictionaries (850 possessive adjectives and 25 feminine surnames). To systematically produce these derived lemmas we developed finite-state transducers (16 different FSTs), similar to those used for inflection, to derive possessive adjectives and feminine counterparts from all surnames, if they exist. One such FST is presented in Figure 3 and it derives a possessive adjective *Černijev* and a feminine surname *Černijka* from a surname *Černi* 'Czerny' (and 332 more surnames, mostly those ending with *i*). Derivation markers `+Pos` and `+GM` are added to the produced lemmas together with codes of inflectional transducers that should be applied to them (`A1` and `N661` in this case).

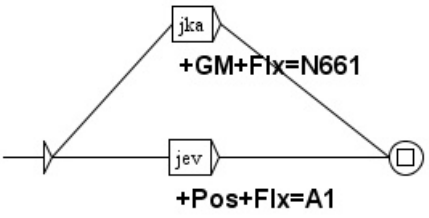

Figure 3: A FST for the derivation of a possessive adjective and a feminine counterpart from a surname belonging to a class of surnames.

As a result, the dictionary entry for the surname *Černi* produced two more derived and connected lemmas – the value

of the marker +BASE= explicitly establishes a *derivation relation* between the original and derived lemmas.

```
Černi,N1064+NProp+Hum+Last+SR
Černijev,A1+Pos+NProp+Hum+Last+SR
         +BASE=Černi_N1064
Černijka,N661+GM+NProp+Hum+Last+SR
         +BASE=Černi_N1064
```

The comparison of generated lemmas with those already in the dictionaries proved the correctness of this approach. In this way we generated more than 24,000 possessive adjectives and nearly 20,000 feminine counterparts of surnames that are all connected to basic lemmas.

For treatment of lexical variants, as well as simple derivation processes, as described in Subsection 6.1., we applied a different approach. In SMD, 5,592 lexical entries are annotated with one of 86 VAR markers.

This procedure, based on the set of rules, will be illustrated with two rule sets for variations: suffix_variations (124 rules) and affix_variations (44 rules) of a single lexical entry, but a similar approach is used for other types of variations, as well as for some simple derivation relations and pronunciation relations. Each rule is represented with the following set of attributes:

1. *RelationName* is a unique rule name and its identification, built upon a unique combination of other attribute values (e.g. *VAR=IratiOvati_V_V*);

2. *RelationType* is a type a variation: suffix_variations, affix_variations, etc.;

3. *SuffixFrom / SuffixTo* indicates suffixes (suffix_variations) or substrings (affix_variations) that a source/target lexical entry must contain;

4. *MarkerFrom / MarkerTo* is a required dictionary marker that a source/target lexical entry must have;

5. *Group* relates rules that are used in pairs.

The group attribute is used to relate a rule with its pair that generates a lexical entry in the opposite direction, e.g. from *oksidirati* → *oksidovati* can be generated, and conversely, from *oksidirati* → *oksidovati* (both meaning the same – *to oxidate*). In this way rule groups were introduced containing rules that come in pairs.

An example of a rule from this rule set is *VAR=ArisatiIrati_V_V*, which is applied to a verb that ends in *-arisati* and contains the marker *VAR=ArisatiIrati* (e.g. *komentarisati* 'to comment'). The rule can be used to generate its variation with suffix *-irati* and an inverse marker – *VAR=IratiArisati* (e.g. *komentirati*). This rule is in a group with five other rules: *VAR=ArisatiIrati_N_N* that generates a noun variation, *VAR=ArisatiIrati_A_A* that generates an adjective variation and three others with inverse markers *VAR=IratiArisati_V_V, VAR=IratiArisati_N_N,* and *VAR=IratiArisati_A_A*. A POS is an important part of these rules since it dictates the *SuffixTo* and *SuffixFrom* values which differ from rule to rule. For example, *-arisati* and *-irati* are related verb suffixes, *-arisanje* and *-iranje* are corresponding noun suffixes (*komentarisanje* vs. *komentiranje* 'commenting') and *-arisan* and *-iran* (*komentarisan* vs. *komentiran* 'commented') are adjective suffixes.

The second rule set (affix_variations) locates candidates that have a certain substring (one or more letters, but also an empty string indicating that a substring may be omitted) anywhere in the lexical entry, and an appropriate marker. For these rules a POS is irrelevant, but must be the same in both the origin and the target lexical entry. There are 22 two-rule inverse groups, which gives a total of 44 rules in this rule set. One example is the rule *VAR=0H* that describes lexical entries in which the letter *h* is missing and can be inserted to obtain a variant, for example *ladan* vs. *hladan* 'cold'. The corresponding inverse rule from the same group is *VAR=H0* indicating that a letter *h* may be omitted. The rule *VAR=CS* operates in a similar way, but in this case the operation is not omission/insertion but substitution – the letter *s* may be replaced with the letter *c*, thus generating, for example, *sufinanciranje* from *sufinansiranje* 'co-financing'.

These rules are not too successful for finding candidates for dictionary expansion because a large number of possible candidates may be generated due to unspecified position of the substring on which the rule operates. For example, *sufinansiranje* with *AffixFrom* being letter *s* and *AffixTo* being letter *c* can result in any of the following: *sufinanciranje, cufinansiranje* and *cufinanciranje*, with only the first one, in this case, being correct.

Developed rules were used to solve three subtasks:

1. Finding lexical entries that are missing in the dictionary (provided that their existence is indicated by markers of existing entries);

2. Finding lexical entries that exist in the dictionary, but lack the expected lexical marker (which is indicated by a marker assigned to a related existing entry);

3. Finding two lexical entries that exist in the dictionary and are expectedly marked (indicating a relation between them).

For the first option, a generated target entry becomes a candidate for a new lexical entry in the dictionary; for the second, a candidate for a marker annotation of an entry is generated; while for the third, a relation is established between two related lexical entries. This procedures also found a few errors in already assigned +VAR markers.

6.3. Statistics and Evaluation Results

The first subtask returned a total of 103 new candidates for dictionary entries through the suffix_variations rule set, of which 50 were accepted and 53 rejected. This may not seem as a very good result, but analysis revealed that the majority of the rejected candidates were actually marked with an incorrect +VAR marker, e.g. IratiOvati/OvatiIrati instead of a CiratiKovati/KovatiCirati. After these markers were corrected, 50 new candidate entries were accepted and only 3 were rejected. For the set of affix_variations rules, 119 candidates were returned for the first subtask, only 38 (32%) of them suitable. Most of affixes are very short (one

letter) and it is not easy to detect which letter should be affected by a rule if several of them occur in a single entry. Most of the rejected candidates were found due to unspecified number of replacements and their position (in cases when there is more than one replacement in the marked lexical entry).

The second subtask found only 35 lexical entries with missing markers. Since in each case both related entries existed in the dictionaries, and one is a possible variation of the other, there is just a small margin for errors. It was confirmed that all but one of the candidates were correct, and that this one occurred because one lexical entry variant was a homograph of another entry.

The third, and most important subtask, established relations between lexical entries using the produced rule sets to find properly marked pairs of entries (both having `+VAR` markers and a POS needed to activate a specific rule that generates their pair). A total of 5,129 symmetric relations was established, 4,411 through the suffix_variations rule set, and 718 through the affix_variations rule set. Frequencies of the most common variations used to connect entries are presented for suffix variations in Figure 4 and for affix variations in Figure 5.

Similar procedures are produced to connect some derivationally related entries (e.g. verbs and verbal nouns and adjectives) and to produce explicit inverse relations from originally implicit ones (in DELAS format).

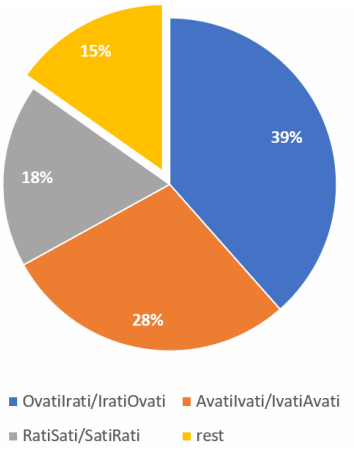

Figure 4: The frequency of established connections by relations from the suffix variations rule set.

7. Conclusion

In this paper we presented a new database model, developed upon the *lemon* model, as well as its application for migrating electronic morphological dictionaries from a single-user file system to a multi-user environment based on a relational database management system. The new lexical data model implemented as a lexical database has various advantages over the previously used file-based system. The introduced logical constraints will prevent omission of markers and enable their controlled use in the future. This will facilitate the enrichment of existing lexical entries with new markers and lexical relations, as we plan to establish as

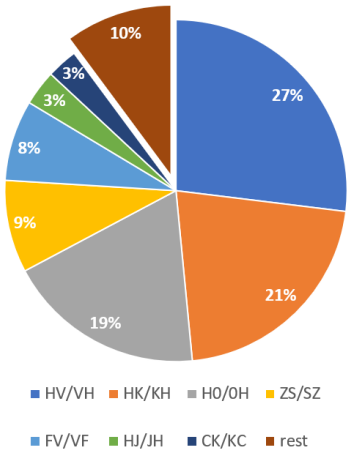

Figure 5: The frequency of established connections by relations from the affix variations rule set.

many explicit relations between lexical entries as possible, on the basis of information already given in SMD.

We adapted the *lemon* model in order to transfer all information stored in existing electronic dictionaries. Therefore, in our model the class `Form` is used for inflected forms instead of variant forms, which is important for Serbian as a highly inflective language. Also, we adapted the *lemon* model to store all existing markers as a thesaurus of data categories and their values, which enabled linking them to LexInfo and other ontologies, like SUMO. Mapping of grammatical categories as well as their values from existing dictionaries to LexInfo, using the catalog of grammatical categories that is complemented with the *lemon* model, is almost complete: for instance, grammaticalGender → lexinfo#gender, while m → lexinfo#masculine, f → lexinfo#feminine, n → lexinfo#neuter. However, for some categories the appropriate mapping still needs to be defined. The mapping of semantic markers to SUMO has also started, for instance `+DOM=Bot` → FloweringPlant and `+DOM=Culinary` → Cooking, but an exact match is not always possible. Future activities also include the use of linked data principles to enable open publishing and linking of language resources on the Web, integrating them with Linguistic Linked Open Data. After that novel application for dictionary management are planned, which will enable not only dictionary development and maintenance, but also their export to different dictionary schemata and formats, to support various NLP application needs.

The first part of the evaluation of the presented model was successfully completed, since all existing data were stored in the new database. Cross-linking was initiated, and some data-inconsistencies were detected and resolved. However, the final evaluation report will be given once the application is fully developed and database exploitation stars. Given that language resources for more than 22 languages, currently distributed with Unitex/GramLab, were developed in the same DELA format and that the presented migration approach is language independent, it is safe to say that it will prove useful for other languages as well.

8. Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #III 47003 and 178003.

9. Bibliographical References

Attia, M., Tounsi, L., and Van Genabith, J. (2010). Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic. Technical report, The NCLT Seminar Series, DCU, Dublin, Ireland.

Bański, P., Bowers, J., and Erjavec, T. (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden, the Netherlands, 19—21 September 2017*, pages 485 – 494.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado-de Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case. In Ilan Kernerman, et al., editors, *Proc. of GLOBALEX'16 workshop at LREC'16, Portoroz, Slovenia*, pages 65–72. European Language Resources Association, May.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

Courtois, B. and Silberztein, M. (1990). *Dictionnaires électroniques du français*, volume 87 of *Langue français*. Larousse, Paris.

Farrar, S. and Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT international*, 7(3):97–100.

Francopoulo, G. (2013). *LMF Lexical Markup Framework*. John Wiley & Sons.

Ide, N., Kilgarriff, A., and Romary, L. (2000). ITRI-00-30 A Formal Model of Dictionary Structure and Content. In *Proceedings of EURALEX 2000*, pages 113–126. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2000.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. E. (2008). ISOcat: Corralling Data Categories in the Wild. In *LREC*.

Khemakhem, M., Foppiano, L., and Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In Iztok Kosem, et al., editors, *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden, the Netherlands, 19 – 21 September 2017*, pages 598–613, Leiden, Netherlands, September.

Klajn, I. (2005). *Gramatika srpskog jezika*. Zavod za udžbenike.

Koeva, S., Krstev, C., and Vitas, D. (2008). Morpho-semantic relations in wordnet–a case study for two slavic languages. In *Proceedings of Global WordNet Conference 2008*, pages 239–253. University of Szeged, Department of Informatics.

Krstev, C. and Vitas, D. (2007). Extending the Serbian E-dictionary by using lexical transducers. In *Formaliser les langues avec l'ordinateur : De INTEX à Nooj*, pages 147–168.

Krstev, C., Vitas, D., and Erjavec, T. (2004). MULTEXT-East resources for Serbian. In *Zbornik 7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije 9-15 Oktober 2004, Ljubljana, Slovenija, 2004*. Erjavec, Tomaž and Zganec Gros, Jerneja.

Krstev, C., Stanković, R., Vitas, D., and Obradović, I. (2006). WS4LR: A Workstation for Lexical Resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 1692–1697.

Krstev, C., Stanković, R., and Vitas, D. (2010). A Description of Morphological Features of Serbian: a Revision using Feature System Declaration. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Maks, I., Tiberius, C., and van Veenendaal, R. (2008). Standardising Bilingual Lexical Resources According to the Lexicon Markup Framework. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, pages 1723–1727.

McCrae, J., Spohr, D., and Cimiano, P., (2011). *Linking Lexical Resources and Ontologies on the Semantic Web with Lemon*, pages 245–259. Springer Berlin Heidelberg, Berlin, Heidelberg.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.

Stanković, R., Obradović, I., and Utvić, M. (2013). Developing termbases for expert terminology under the TBX standard. In *35th Anniversary of Computational Linguistics in Serbia*, pages 12–26. University of Belgrade, Faculty of Mathematics.

Stanojčić, Ž. and Popović, L. (2008). *Gramatika srpskog jezika*. Zavod za udžbenike.

Tutin, A. and Véronis, J. (1998). Electronic dictionary encoding: Customizing the TEI guidelines. In *Proc. Euralex*.

Villegas, M. and Bel, N. (2015). PAROLE/SIMPLE 'lemon' ontology and lexicons. *Semantic Web*, 6:363–369.

Vitas, D., Pavlović-Lažetić, G., and Krstev, C. (1993). Electronic dictionary and text processing in Serbo-Croatian. *Sprache–Kommunikation–Informatik*, 1:225.

10. Language Resource References

Krstev, Cvetana and Vitas, Duško. (2015). *Serbian Morphological Dictionary - SMD*. University of Belgrade, HLT Group and Jerteh, Lexical resource, 2.0.

Stanković, Ranka and Krstev, Cvetana. (2016). *LeXimir*. University of Belgrade, HLT Group, Software Toolkit, 2.0.