Impressum

FOR THE EDITOR:
Prof. Aleksandar Jerkov, PhD office@unilib.bg.ac.rs
University Library "Svetozar Marković"
Faculty of Philology, University of Belgrade

EDITOR:
Faculty of Philology, University of Belgrade
University Library "Svetozar Marković"
Serbian Academic Library Association

EDITOR-IN-CHIEF:
Prof. Cvetana Krstev, PhD cvetana@matf.bg.ac.rs
Faculty of Philology, Department for Library and Information Science

MANAGING EDITOR:
Aleksandra Trtovac, PhD aleksandra@unilib.bg.ac.rs
University Library "Svetozar Marković"

EDITOR OF ONLINE EDITION:
Jelena Andonovski, PhD andonovski@unilib.bg.ac.rs
University Library "Svetozar Marković"

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

Belgrade, Vol. 19, No. 2, December 2019

JOURNAL REDACTION:
Journal Infotheca
11000 Belgrade, Bulevar kralja Aleksandra 71

+381 11 3370–211
infotheca@unilib.rs

PRINTED BY:
Mamigo plus
Belgrade

Journal is published twice a year

Contents

Scientific papers

Extraction and annotation of 'location names'

UDC 81'322.2
DOI 10.18485/infotheca.2019.19.2.1

ABSTRACT: Introduced as part of the Message Understanding Conferences dedicated to information extraction, Named Entity extraction is a well-studied task in Natural Language Processing. The recognition and the categorisation of person names, location names, organisation names, etc., is regarded as a fundamental process for a wide variety of natural language processing applications dealing with content analysis and many research works are devoted to it, achieving very good results. One of our objectives is the identification and automatic (or semi-automatic) annotation of location names in order to apply the most appropriate information extraction methods. The main objective concerns the combination and interoperability between symbolic and statistical NLP (Natural Language Processing) methods (symbolic rules, machine learning, and data mining). Our work consisted of recognising named entities and in particular locations with Unitex, annotating them with Brat, and correcting them manually. The recall and accuracy rates are very encouraging but the question remains: What is a location name?

KEYWORDS: location names, locative complement, annotation, information extraction, Unitex.
PAPER SUBMITTED: 07 October 2019
PAPER ACCEPTED: 28 November 2019

Tita Kyriacopoulou
tita@u-pem.fr
Claude Martineau
claude.martineau@u-pem.fr

University of Paris-Est
Laboratoire d'Informatique
Gaspard-Monge, France

Markarit Vartampetian
markaritvar@gmail.com
Paris Nanterre University
Paris, France

1 Introduction

In this paper we will present a brief review of our research carried out on French location names. This work is done in the context of named entity extraction and annotation and, in particular, extracting and annotating

locations in unstructured corpora. Over the last twenty years, considerable amount of work has emerged on this topic in the field of NLP (Natural Language Processing), MUC evaluations[1] (1987-1998), ACE program (Doddington et al., 2004), NIST's[2], and ATALA's[3] work. The main difficulty lies, no longer, in the named entity recognition, but in the named entity disambiguation and their correct interpretation; for instance, distinguishing the organisation *aéroport Charles de Gaulle* (Charles de Gaulle Airport) in *L'aéroport Charles de Gaulle va ouvrir ses portes à un nouveau terminal* (Charles de Gaulle Airport will open its doors to a new terminal) from the location *On arrive à l'aéroport Charles de Gaulle dans 15 minutes* (We arrive at Charles de Gaulle airport in 15 minutes). That is, the analysis must assign a category according to the context. Despite the work done so far, research on named entities still plays a central role in NLP and especially in the process of disambiguation of corpora.

It is worth nothing that the choice of Location Names was made because of the University Gustave Eiffel – the University of Marne-la-Vallée is part of it – which will be founded by January 1st 2020 and whose topic will be 'the City'.

First, we will present our research data, based on corpora works, and then the different categories of location names. Next, we will discuss the problem of ambiguity and we will try to make a clear distinction between location names and locative complements. Besides, we will describe our annotation method of location names and we will compare the results with those of the semi-automatic annotation[4] by use of the open source tool Gemini.

2 Collecting data

Within the framework of our research, the language resources consist of a combination of lexicons available in Unitex/GramLab,[5] and of freely avail-

[1] Message Understanding Conference
[2] National Institute of Standards and Technology
[3] Association pour le Traitement Automatique des Langues (French Association for Natural Language Processing)
[4] 'Nous entendons une annotation ... '(Brando et al., 2016)
[5] Unitex/GramLab

able corpora. These corpora were preprocessed and analysed in Unitex[6] and annotated in XML format. In particular, we have worked on four corpora[7]:

- Corpus de la presse
  Press corpus
- Le Tour du monde en 80 jours
  Around the world in eighty days
- 20.000 lieues sous les mers
  Twenty thousand leagues under the sea
- Extraits de 2001 à 2012 du Monde Diplomatique
  Monde Diplomatique extracts from 2000 to 2012

Table 1 lists the main features of these corpora. The parentheses in tokens and simple forms indicate the number of different forms.

Corpus	Press ...	Around ...	Twenty ...	Monde ...
sentence delimiters	1,756	3,652	7,349	8,387
tokens	100,642 (8,467)	165,239 (9,452)	339,425 (15,301)	479,808 (22,952)
simple forms	43,664 (8,404)	71,859 (9,422)	149,007 (15,268)	205,177 (22,860)
digits	3,742 (10)	438 (10)	1,047 (10)	11,892 (10)
simple-word entries	10,082	13,229	20,654	199,417
compound entries	2,039	2,099	3,123	5,586
unknown simple words	1,405	3,156	2,161	4,877

Table 1. Overview of corpora

It should be noted that the corpora do not consist of only narrative content. Some sentences constitute simple utterances of terms, recalling the inventory process:

[6] For the sake of conciseness we will be using Unitex instead of Unitex/GramLab in the rest of the article.

[7] The second and third corpora are novels written by Jules Verne.

(1) îles Salcette, Colaba, Éléphanta, Butcher
Salcette, Colaba, Éléphanta, Butcher islands

(2) au large des Pomotou, des Marquises, des Sandwich
off the coastline of Pomotou, Marquises, Sandwich

In example (1) note that, because of the enumerative process, the trigger word *îles* (*islands*) is present only in front of the first name, whereas in example (2) the same trigger word is absent. The enumerative sequence is introduced by *au large*. Besides, we notice that the words *Marquises* and *Sandwich* are ambiguous.

Similarly, forms with connectors (and, or, etc.) can be complex to analyze.

(3) cratères de l'Érébus et du Terror
Erebus and Terror craters

(4) affluents ou sous-affluents de la Little-Blue-river
tributaries or sub-tributaries of the Little-Blue-river

(5) côtes arabiques du Mahrah et de l'Hadramant
arabian coasts of Mahrah and Hadramant

In example (3), in addition to a coordination of names of volcanoes, there is a trigger word which is not the word *volcan* (volcano), but the word *cratère* (crater), which complicates even further the recognition. The example (4) contains a disjunction *affluents ou sous-affluents* (tributaries or sub-tributaries) which consists of trigger words and indicates that the following is a hydronym (river, stream) as in the phrase *les affluents de la Seine* (tributaries of the Seine). In *Little-Blue-river* the status of the hydronym is indicated by the word *river*. But in this case, *affluents* and *sous affluents* are not named entities, but entities of type hydronym in relation to the named entity *Little-Blue-river*.

Finally, example (5) constitutes yet another coordination for location names. However, the precise location is defined both by the noun *côtes* (coasts) and the toponymic adjective *arabiques* (arabian).

As the extraction systems appear to have been trained more in continuous texts rather than in non-continuous texts, it is difficult for them to deal with this non-narrative content.

The absence of context may occasionally lead to false recognition. For instance, in example (6) *aéroport de Montréal* clearly indicates a location. However, in example (7) we notice that the context *porte-parole* (spokeswoman) has a double impact on the interpretation. On the one hand,

it confers on the *aéroport de Montréal* an organisation 'role' and, on the other hand, it refers to *porte-parole* (the spokeswoman), but not to the organisation itself.

(6) Quand on arrive à l'aéroport de Montréal, ...
When we arrive at Montreal airport ...

(7) La porte-parole de l'aéroport de Montréal assure, pour sa part, ...
The spokeswoman for Montreal airport ensures, ...

3 Location Names

Regarding the named entities and, in particular, locations names, notable works have emerged on typologies, which allow us to define what we aim to recognise and extract. Several typologies have been proposed and they differ in the categories as well as in the structure of the elements and even in the typology itself. Sometimes, they also differ fundamentally in the definition of the notion of named entities

For instance in TEI[8] the sequence *à 20 km au nord de Paris* is annotated as follows:

```
<placeName>
    <measure>20 km</measure>
    <offset>au nord de</offset>
    <settlement type="city">Paris</settlement>
</placeName>,
```

whereas in ISTEX[9] it is annotated as:

```
<placeName>
    Paris
</placeName>.
```

3.1 Categories

According to the typologies in use, the number of named entity categories varies. However, certain categories are present in most typologies:

[8] TEI
[9] ISTEX

- Names of persons or saints: 'Alexander the Great', 'Saint Ambroise', 'Constantine the Great', 'Saint Demetrius.'
- Location names[10] (names of countries, cities, regions, etc.): 'Eastern Europe,'. Note that institutions, organisations and even events act as location names in some contexts, for instance *Aller au Salon du Livre de la Havane* (cf. Section 5).
- Names of mountains (oronyms) or hydronyms: 'Saint Basil Lake','the Black Sea.'
- Organizations: 'Agricultural Bank', 'Orsay Museum', 'Ministère de l'Agriculture'
- Authorial texts: 'Green Paper,' 'the National Anthem'
- Events: 'The French Revolution' 'The Olympic Games','Le 11 Septembre'

According to (Chinchor, 1998), named entities include proper nouns but also time expressions, such as 'Holy Week', 'Nautical Week', 'Black Friday'.

The task of named entity extraction consists of automatically recognising the NE in corpus, extracting and classifying them into categories such as *Person*, *Location*, *Organization*. As indicated by (Denis and Sagot, 2012), we can distinguish two ways of identifying an entity, either intrinsic where *France* denotes a place, or contextual, as in *France signed the treaty*, where *France* can be recognised as an organisation.

As it is very aptly described in (Hengchen et al., 2015), during the analysis of a corpus, various questions arise regarding the classification of named entities of type location. Even though the categories are well defined (at first, a location is not an organisation and vice versa), their rigidity requires the researcher to make subjective choices. *Paris* is used in a geographical sense (town), but Paris may also indicate a name, thus it can be logically categorised as location *LOC* and *PERSON*. But what about the term *Paris* in contexts referring to the town of Paris as an organisation *ORG*, e.g., in *Paris va organiser les jeux olympiques* (Paris will organise the Olympic Games)? Therefore, a term which has a fixed geographical location by nature, but represents a commercial enterprise, is it to be considered as a location or an organisation? In fact, this is the key question in this article

[10] In the Prolex dictionary provided with the Unitex distribution, all the location names are marked as toponym with a more specific feature (City, Country, Hydronym, etc.)

and to which we intend to answer.

Furthermore, a great number of named entities, especially organisations and/or locations, may appear as initialisms or acronyms: (Université Paris-Est Marne-la-Vallée UPEM, Made in USA), which complicates the interpretation. *CE*, for example, can indicate *conseil d'Etat* (Council of State), *conseil de l'Europe* (Council of Europe), *Comité d'Entreprise* (works council), etc. In Wikipedia there are more than ten interpretations for *CE*.

Named entity variations were also present in our corpora, e.g., *Muséum de Paris* which can be designated by different variations :

(8) Muséum National d'Histoire Naturelle de Paris
National Natural History Museum of Paris

(9) Muséum National d'Histoire Naturelle.
National Natural History Museum

(10) Muséum National de Paris
National Museum of Paris

(11) Muséum d'Histoire Naturelle
Natural History Museum

(12) Muséum de Paris
Museum of Paris

(13) MNHN
NNHM

Deleting certain terms influences considerably the accuracy of recognition. Moreover, the English translation generates additional ambiguity. In French, the word *Muséum* designates a scientific museum devoted to natural sciences. For other domains the word *Musée* is used more (e.g. Musée du Louvre (Louvre Museum), Musée Guimet (Guimet Museum), Musée des Arts Décoratifs (Museum of Decorative Arts), etc.). Hence, in French, *d'Histoire Naturelle* can be deleted without resulting in ambiguities, since the domain is indicated by the word *Muséum*. The location Paris is indicated either by the presence of the word *Paris* or by the adjective *National* because the other "Muséums" that exist (in Le Havre, Grenoble, La Rochelle, etc.) are not qualified as nationals. This example illustrates the fact that the accuracy of the recognition depends on the terms that are erased and that this erasability is language dependent. In English, *Natural History* cannot be erased without losing information.

4 What is a location name

Location names (according to typology *LOC*, *placeName*, *Loc.admi*, etc.) as other named entities may be more or less ambiguous and may depend on corpus but in a given context they can act as metonymy. Sometimes even we have translation problems, especially when a location name refers to different locations. For instance, *London* denotes a town in Great Britain as well as in Canada. However, the French translation differs: *Londres* designates the capital of Great Britain, but *London* refers to the town across the Atlantic.

Considering the following sentences (14) et (15):

(14) Marie va à Paris
Mary is going to Paris

(15) Paris va organiser les jeux olympiques en 2024
Paris will organise the Olympic Games in 2024

the town of *Paris* denotes a location in (14), but an administrative authority, namely an organisation, in (15). Hence, defining a location remains an acute problem.

(16) Marie habite Rue de Paris
Marie lives in Rue de Paris

Likewise, in (16) the sequence *Rue de Paris* does not indicate a precise location since *Rues de Paris* exist in Lille, Nice, and other cities including Paris. However, in (17) *Paris* denotes a sports team, and probably a football team.

(17) Paris a battu Lille 2-0
Paris beat Lille 2-0

Finally, it is known that location names have always inspired first name choices. Thus, *France* does not only constitute a country, but also a first name.

These issues are partly solved by Ester 2 initiative,[11] which proposes named entity sub-categories in addition to the traditional categories, as it is shown in the following examples:

[11] Ester 2 initiative

(18) Je suis stationné à côté de la <ent type ="loc.fac">mairie de Paris</ent>.
I am parked next to the city hall of Paris

(19) La course à la <ent type ="org.pol">mairie de Paris</ent> a commencé.
Paris mayor's race has begun

(20) Je me suis fait opérer à l'<ent type ="org.non-profit">hôpital Necker</ent>.
I had a surgery at Necker Hospital

Thus, *la ville de Paris* would be *LOC.admi*, namely a *LOC* entity type related to an administrative authority. Nevertheless, this type of classification is rarely exhaustive and even less integrated in information extraction systems which extract named entities without relating them to a given context.

Afterwards, we shall present a number of questions that need to be raised.

4.1 Named Entity or Extended Named Entity

Do we intend to recognise a Named Entity or an Extended Named Entity? Gaio and Moncla (2017) have used the concept of Extended Named Entity (ENE). Based on Jonasson's definition, an ENE refers to an entity built with a proper name, e.g., *Rue de Paris*, and possibly composed of one or more concepts, e.g., *La maire de Paris*.

We believe that when we have to extract and annotate locations, it is evident that we should recognise the ENE rather than the NE since in (21) *La maire de Paris* denotes a person.

(21) La maire de Paris
The mayoress of Paris

However, regarding the ENE, the referential nature of named entities can generate difficulties and pose problems that are not solved yet. For instance, Sagot (Sagot et al., 2012) argues that this situation is found in annotation cases of university names. Therefore, according to Sagot, *université*[12] *de Marne-la-Vallée* denotes a university located in Marne-la-Vallée and we should only annotate the town Marne-la-Vallée, whereas *Université*

[12] The word is written without a capital letter.

de Marne-la-Vallée refers directly to the organisation that the university constitutes and, as a result, we should annotate the whole entity as an organisation. Similarly, regarding the example of *Université de Montpellier*, since a unique organisation corresponding to this term does not exist, only *Montpellier* should be annotated as a town.

Hence, the above examples demonstrate that in some cases we need to recognise an ENE, e.g., *maire de Paris*, *rue de Paris*, *Université de Nantes*, whereas in other cases it is sufficient to recognise only an NE, e.g., *ville de Paris*. From our perspective, it is almost impossible to distinguish the trigger words that are part of the entity from those who do not without leaving out the trigger words that are part of the entity, e.g., in *Mer Morte* (Dead Sea). The solution proposed by Unitex with the use of graphs and advanced functions (use of contexts, weights, variables, etc.) seems – for us – to be the best compromise, which allows us to choose the annotation according to the needs of the applications and of the corpora.

From our perspective, this type of precision is necessary as we are interested in the context and in the exact and precise named entity annotation, with a view to using it in machine learning systems.

In this research, we decided to recognise the ENE even if *université de Paris* is not a unique reference and/or does not refer to a different era.

5 Ambiguities

As we have shown various examples of ambiguities concerning location names, we could say in brief, that the most problematic cases are:

1. distinguishing a location from a locative complement,
2. distinguishing a location or a locative complement from an organisation, and
3. distinguishing *Loc.admi* from *Loc.Person*.

The first two cases are related, among others, to the syntactic description, thus considering the named entity not as an isolated sequence (simple or complex), but as an element of a basic sentence. In bibliography, we often refer to the context, but the notion of context is more vague and almost never explicit.

For the first case, it is important to specify that we consider as a location every toponym generally used in a geographical sense – see example (22) –

whereas a locative complement answers to the question *where* – see examples (23) and (24).

(22) Paris est une belle ville
Paris is a beautiful city

(23) Claire va à Paris
Claire goes to Paris

(24) Claire dort à Paris
Claire sleeps in Paris

To push our analysis further consider the following examples:

(25) Claire se repose à Paris / Claire va à Paris
Claire rests in Paris / Claire goes to Paris

(26) Claire se repose. / *Claire va.
Claire rests. / *Claire goes.

(27) À Paris, Claire se repose. / *À Paris, Claire va.
In Paris, Claire rests / To Paris, Claire goes

In these examples, we observe that there are specificities between the verbal predicate *se reposer* (to rest) and *aller* (to go) and yet in both cases the locative complement *Paris* answers the question *where*. These differences emerge from the fact that in the case of the verb *se reposer à Paris* is a modifier whereas in the case of the verb *aller* is an argument. It is noteworthy, that in English this distinction results in two different translations (*to Paris*, *in Paris*). In order to make this distinction a syntactico-semantic analyser is necessary.

Regarding the second case, we consider as organisation names all references to an organisation; political, educational, financial, religious, associative, etc., which are annotated as *Loc.admi* in Ester 2. In this case, syntactic analysis is able to give solutions:

(28) La ville de Paris va organiser les jeux olympiques en 2024/Paris va organiser les jeux olympiques en 2024
The city of Paris will organise the Olympic Games in 2024/Paris will organise the Olympic Games in 2024

(29) Les jeux olympiques auront lieux à la ville de Paris / Les jeux olympiques auront lieux à Paris
Olympic Games will take place in the city of Paris/Olympic Games will take place in Paris

Thus, by applying the same tests as in examples (28) and (29) we observe that in example (29) the segments *ville de Paris* / *Paris* answer to the question *where* and, as a result, they constitute locative complements, which are considered as arguments by linguists (Gross, 1996) because a sentence like (30) is not acceptable.

(30) *Les jeux olympiques auront lieu
*Olympic Games will take place

However, not all ambiguities that we have encountered in our corpus can be solved by syntactic analysis.

Regarding the third case, disambiguating a *Loc.admi* (a *Loc* entity type related to an administrative authority) from a *Loc.person* (a *Loc* entity type related to a person) is more complicated. Thus, between :

(31) Paris organise les jeux olympiques de 2024
Paris will organise Olympic Games in 2024

(32) Paris organise une grande fête pour le 14 juillet
Paris will organise a great feast for the 14th of July

the syntactico-semantic analysis is not able to distinguish the town of *Paris* from a person. In example (31), extralinguistic information is necessary for a machine to know that a person cannot organise an event of this type. Hence, this sentence must be modified as *La ville de Paris* organise les jeux olympiques de 2024. Still, example (32) remains ambiguous even for a human being.

Note that all the ambiguities cannot be solved with syntactico-semantic analysis.

6 Methodology

In order to deal with the above problems, we used the following method: First, we created graphs using Unitex and we proceeded to an automatic

annotation in XML format. Then, we imported the annotated corpora using a script to Brat (Brat rapid annotation tool),[13] we validated the annotated corpora, and, finally, we corrected the annotations manually. Brat is a tool that allows us to visualise in a browser (Mozilla, Opera, etc.) the annotated corpora by highlighting the recognised patterns with a color associated with its named entity type.

The grammar of Figure 1 allows us to annotate the text by adding tags *Lieu* (<Lieu> </Lieu>) to the recognised patterns. At the bottom of the graph, we notice two paths that allow us to recognise ambiguous patterns containing a location name without tagging it. Figure 2 shows an extract of the results in concordance form. Figure 3 shows certain concordances in which the tag *Lieu* contains a precise type (country, region, hydronym, etc.). Figure 4 shows some sequences recognized by the grammar for purposes of disambiguation and therefore not annotated.

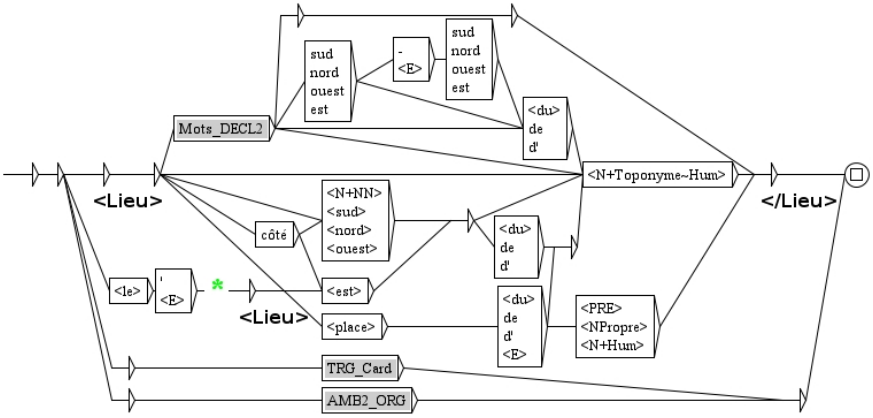

Figure 1. Grammar of location names annotation

.

We proceeded with two types of manual annotation. The first one aims at annotating arguments, whereas the second takes into account all location names. The grammars that have been constructed for the automatic annotation by now perform correctly for the second type of annotation and will

[13] Brat

be improved so as to be able to recognise only the arguments. As a result, during our evaluation of the results, we compared the manual annotation with all locations.

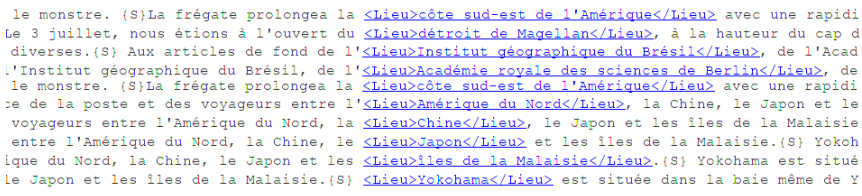

Figure 2. Concordances of location names

.

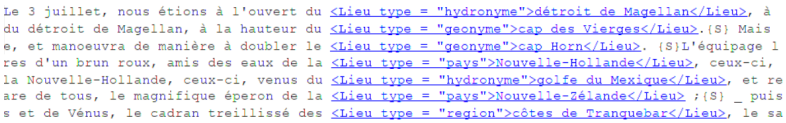

Figure 3. Concordances of location names with types

.

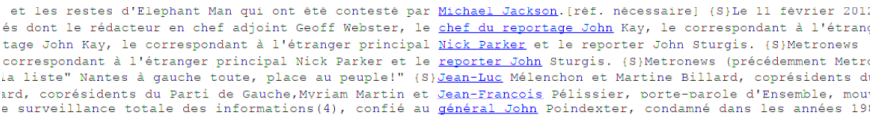

Figure 4. Concordances of recognized sequences not tagged as location names

.

7 Comparing automatic annotations with semi-automatic annotations

In this section, we compare the automatic annotation produced by Unitex with the semi-automatic annotation (manual correction of the automatic annotation by an experienced linguist) This task was done automatically by use of the Gemini[14] tool, developed in LIGM,[15] which allows us to calculate the *Precision*, the *Recall*, and the *F-measure* values:

[14] Gemini source on Github

[15] Laboratoire d'Informatique Gaspard-Monge

$$Precision = \frac{number\ of\ correctly\ matched\ location\ names}{number\ of\ terms\ matched\ as\ location\ names}$$

$$Recall = \frac{number\ of\ correctly\ matched\ location\ names}{number\ of\ actual\ location\ names}$$

$$F = 2 * \frac{precision * recall}{precision + recall}$$

The results obtained from our corpora are listed in Table 2. We notice that in the newspaper corpus the recall has increased. This can occur because of the various ambiguities in that corpus. As the corpus Monde Diplomatique is more homogeneous, it is affected less by this phenomenon.

Corpus	Press corpus	Around the ...	Twenty Thou ...	Monde Diplo ...
Precision	0.24403422	0.74358974	0.78455056	0.51353696
Recall	0.89144737	0.72432932	0.68691588	0.47702724
F-measure	0.38317427	0.73383317	0.73249409	0.49460927

Table 2. Results generated by the Gemini comparison tool

The following constitute true (positive, negative) and false (positive, negative) examples issued from our corpora and the application of grammars.

– True positive :

- ... on apprit qu'un steamer de la ligne de San Francisco de Californie à Shangaï avait revu l'animal, trois semaines auparavant, dans les ***mers septentrionales du Pacifique*** ...
  ... word came that a steamer on the San Francisco line sailing from California to Shanghai, had sighted the animal again, three weeks before in the ***northerly seas of the Pacific*** ...
- Mais, en attendant, il me fallait chercher ce narwal dans le ***nord de l'océan Pacifique***
  But in the meantime, I would have to search this narwhale in the ***northern Pacific Ocean***

- Tout deux se rendront au ***lycée Paul Cézanne d'Aix-en-Provence*** qui propose une section expérimentale de langues et cultures méditerranéennes.
  Both will go to ***Paul Cézanne d'Aix-en-Provence high school*** which proposes an experimental section of Mediterranean languages and cultures

- ... plusieurs instituts dédiés à l'enseignement du droit international virent le jour en Amérique et en Europe, tel l'***Institut universitaire de hautes études internationales de Genève*** ...
  ... several institutes dedicated to international legal studies have emerged in America and in Europe, such as the ***Graduate Institute of International Studies in Geneva*** ...

- Une centaine de manifestants ukrainiens ont pénétré samedi dans le ***principal bâtiment du ministère de l'Énergie*** ...
  About one hundred Ukrainian protesters entered the ***main building of the Ministry of Energy*** on Saturday ...

- C'est à 20h45mn qu'il a foulé le ***tarmac de l'aéroport international Léopold Sédar Senghor de Dakar***, à bord de la compagnie d'Air France.
  It was 20:45 when he crossed the ***tarmac of Léopold Sédar Senghor International Airport in Dakar***, onboard of the Air France company.

– True negative :

- ... ancien chef de l'ANC, devenu président de l'***Afrique du Sud***
  ... former leader of the ANC, became president of ***South Africa*** ...

- Cette ambiguïté, ou plutôt cette confusion, autour du mot « science » permet au sociologue Alain ***Touraine*** de disculper ...
  This ambiguity, or rather this confusion, around the word "science" allows the sociologist Alain ***Touraine*** to exonerate ...

- C'étaient des portraits, des portraits de ces grands hommes historiques dont l'existence n'a été qu'un perpétuel dévouement à une grande idée humaine, Kosciusko, le héros tombé au cri de Finis Polonioe, Botzaris, le Léonidas de la Grèce moderne, O'Connell, le défenseur de l'Irlande, ***Washington***, le fondateur de l'Union américaine ...
  They were portraits of these great men of history who had spent

their lives in perpetual devotion to a great human ideal, Kosciusko, the hero whose dying words had been Finis Poloniae, Botzaris, the Leonidas of modern Greece, O'Connell, Ireland's defender, ***Washington***, founder of the American Union ...

- Elles furent ensuite nommées Malouines, au commencement du dix-huitième siècle par des pêcheurs de ***Saint-Malo*** ...
  at the beginning of the 18th century, they were named the Malouines by fishermen from ***Saint-Malo*** ...
- Si aujourd'hui il est difficile d'imaginer la firme de Cupertino passer à ***4 Go de RAM*** un de ses mobiles ...
  If today it is difficult to imagine the Cupertino firm move one of his mobiles to ***4GB RAM*** ...

– False positive :

- ***Allons*** donc !
  Let's go!
- « ***Viens*** là »
  Come on!
- Et si l'on en croit la leçon du beau film Little ***Sénégal***
  And if we believe the lesson of the beautiful film Little ***Senegal***
- ***Paris 2002***
- Vêtus comme des hérauts du ***Moyen*** Age
  Dressed as heralds of the ***Middle*** Ages
- ***Né*** dans l'ombre du pouvoir
  Born in the shadow of power
- ***Mars***

– False negative :

- New Zealand's Role in the International Spy Network, Craig Potton Publishing, ***Nelson***, Nouvelle-Zélande, 1996.
- ... de ses voyages et des trois petites Nahîla nées à ***Deir-el-Assad*** ...
  ... about his travels and the three little Nahilas born at ***Deir-el-Assad*** ...

- ... et exploitant à ***Marly-le-Roi*** (78), Trappes (78), Les Clayes-sous-Bois (78), Asnières (92), Nanterre (92), Boussy-Saint-Antoine (91) et Epernay (51)
  ... and operator in ***Marly-le-Roi*** (78), Trappes (78), Les Clayes-sous-Bois (78), Asnières (92), Nanterre (92), Boussy-Saint-Antoine (91) et Epernay (51)
- Il ne s'est pas opposé à l'installation des prisonniers d'Al-Qaida sur la ***base militaire de Guantanamo Bay***
  He did not oppose to the installation of Al-Qaida prisoners at the ***Guantanamo Bay military base***
- L'auteur est née en 1968 à ***Niodor***...
  The author was born in 1968 in ***Niodor*** ...

8 Conclusion

Our first goal was the annotation of location names based on four corpora (469,707 words in total) by use of Unitex – and of Brat regarding the manual correction – in order to, ultimately, test the GEMINI tool, which is able to compare an automatic annotation with a semi-automatic one (automatic annotation manually corrected by an experienced linguist). But we have readily been confronted with the complexity of this named entity type, to the point of asking the question *what is a location name*. Our main contribution in this article was to improve the recognition of these named entity types without resolving all the ambiguities presented in our corpora. While most tools do not recognise a location name as a type *LOC* named entity if it refers to another type of entity, such as organisation, person or event, we have tried to disambiguate this type of information and annotate it according to the context and not to the ambiguities. We strongly believe that by integrating a syntactic analyser in our system, the results will ameliorate. However, we need to annotate more corpora and process more languages in order to achieve satisfactory results.

References

Brando, Carmen, Nathalie Abadie and Francesca Frontini. "Évaluation de la qualité des sources du Web de Données pour la résolution d'entités nommées". *Ingénierie des Systèmes d'Information* (2016), numéro spécial 'Web de données : publication, liage et capitalisation'

Chinchor, Nancy A. "Proceedings of the Seventh Message Understanding Conference (MUC-7) Named Entity Task Definition". In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, 1998. http://acl.ldc.upenn.edu/muc7/ne_task.html, version 3.5, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

Denis, Pascal and Benoît Sagot. "Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging". *Language Resources and Evaluation* Vol. 46, no. 4 (2012): 721–736. https://hal.inria.fr/inria-00614819

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel et al. "The Automatic Content Extraction (ACE) program-tasks, data, and evaluation". *Proceedings of LREC* Vol. 2 (2004)

Gaio, Mauro and Ludovic Moncla. "Extended Named Entity Recognition Using Finite-State Transducers: An Application To Place Names". In *The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2017)*. Nice, France, 2017. https://hal.archives-ouvertes.fr/hal-01492994

Gross, Maurice. *GRAMMAIRE transformationnelle DU FRANÇAIS: 1 - Syntaxe du verbe*, Cantilène, 1996

Hengchen, Simon, Seth van Hooland, Ruben Verborgh and Max De Wilde. "L'extraction d'entités nommées : une opportunité pour le secteur culturel ?". *Information, Données et Documents* Vol. 52, no. 2 (2015): 70–79

Sagot, Benoît, Marion Richard and Rosa Stern. "Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées". In *Traitement Automatique des Langues Naturelles (TALN)*, Antoniadis, Georges, Hervé Blanchon and Gilles Sérasset, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Vol. 2 - TALN, Grenoble, France, 2012. https://hal.inria.fr/hal-00703108

Recognition and normalization of temporal expressions in Serbian medical narratives[1]

UDC 811.163.41'322.2:61
DOI 10.18485/infotheca.2019.19.2.2

Jelena Jaćimović
jelena.jacimovic@stomf.bg.ac.rs

University of Belgrade
School of Dental Medicine
Belgrade, Serbia

ABSTRACT: The temporal dimension emerges as one of the essential concepts in the field of medicine, providing a basis for the proper interpretation and understanding of medically relevant information, often recorded only in unstructured texts. Automatic processing of temporal expressions involves their identification and formalization in a language understandable to computers. This paper aims to apply the existing system for automatic processing of temporal expressions in Serbian natural language texts to medical narrative texts, to evaluate the system's efficiency in recognition and normalization of temporal expressions and to determine the degree of necessary adaptation according to the characteristics and requirements of the medical domain.

KEYWORDS: Natural language processing, temporal reasoning, TIMEX, TimeML, clinical records.

PAPER SUBMITTED: 10 October 2019
PAPER ACCEPTED: 06 December 2019

1 Introduction

As a reflection of physical reality changes in human consciousness, or the way the human mind perceives and interprets events, time determines a person's understanding of the world. The temporal dimension is a central

[1] This paper is part of the research conducted within the author's doctoral dissertation, defended at the Faculty of Philology, University of Belgrade.

aspect not only of our daily lives but also of understanding the changes and problems that arise within other specific domains.

In the field of medicine, time also emerges as one of the essential concepts (Shahar, 1999; Augusto, 2005; Zhou and Hripcsak, 2007; Reeves et al., 2013), providing a basis for the proper interpretation and understanding of medically relevant information. The particular order of symptoms, the right time of therapy administration, duration, and frequency are significant only in a specific temporal context. Moreover, during the diagnosis, it is particularly important to know the chronological order of individual symptoms or their duration.

Without temporal dimension, it is almost unlikely to adequately present clinically relevant data, nor to accurately infer and make decisions based on them. Different medical interventions occur at one or more points in time (for example, *hirurška intervencija zakazana za* ***29.02.2000. g. u 9 h*** 'surgical intervention scheduled for 2/29/2000 at 9 am'). Furthermore, specific data, like laboratory test results or data on diagnoses and prescribed therapies, are valid only within an explicitly or implicitly defined period (such as *po dobijanju gore navedenog rezultata ordinirana je infuzija Tygacil 50 mg* ***10 dana*** 'after receiving the above result, Tygacil 50 mg infusion administered for 10 days' or *tokom* ***hospitalizacije*** *ordinirana konzervativna terapija* 'conservative therapy administered during hospitalization'). Besides, relevant clinical data and recommended or performed interventions are important medical concepts, which are often temporally related (e.g. *potrebno nekoliko HD dijaliza* ***posle*** *implantacije grafta* 'few hemodialyses required after graft implantation'). These temporal relations need to be discovered and resolved to correctly determine the chronology of events or conditions, which enables, for instance, the successful monitoring of disease development and the effectiveness of the therapy applied.

1.1 The importance of automatic processing of temporal information in medical narratives

Time is crucial for the representation of information contained in medical information systems. Such information is relevant to patients, containing accurate and comprehensive data on medical conditions, diagnoses, treatment courses and outcomes. Modern advances in medical information technology have made it possible to automate most of the processes involved in providing healthcare services, facilitating the collection and storage of patient information. The real need to store these correctly temporally oriented data exists,

first of all, because of the possibility of their later use in another context to improve the quality of medical care services, diagnostic and theoretical processes in medicine.

Electronic health records contain a certain amount of information provided in a structured form. The existence of predefined data types and their relations (patient's name or health insurance number, doctor's visit date, etc.) enables their machine processing. However, for a more intelligent decision support system, these data types present a deficient resource, as they lack key information related to clinical treatment courses and outcomes. This crucial information, such as disease condition and its course of development, is often recorded only in unstructured texts (like physicians' clinical notes), containing a wealth of clinically relevant information presented by complex, non-standard forms of natural language. In order to locate and extract them from medical narratives, natural language processing technologies are applied. For extracted medical concepts to be chronologically organized and used in this context, it is necessary to provide automatic extraction and interpretation of temporal information found in medical texts. The way of presenting and concluding based on time-oriented clinical data is equally important for both healthcare providers or medical professionals as well as automated systems designed for general electronic patient record search or decision support during making a diagnosis, determining therapy, or evaluating its effectiveness. For example, after the administration of „sublingual allergen-specific immunotherapy" to evaluate the clinical efficacy of treating allergic rhinitis and asthma in children, it would be useful to know whether nasal congestion and rhinorrhea symptoms were statistically significantly reduced after the first measurement, in the first six months of the therapy.

2 The nature of medical narrative texts and the challenges in automatic processing

Automatic processing of temporal expressions (TEs) of medical texts involves their identification and formalization in a language understandable to computers, which is not an easy task. Besides the complex nature of time and its possible structure model (linear, branching, or circular), this problem is also affected by some difficulties arising from the domain specificity and the language used in these texts.

As one of the most significant forms of medical narrative texts, clinical notes represent the physician's written reports of a visit and examination of

a patient, intending to collect all the information necessary to detect the disease and make an accurate diagnosis. These texts contain an abundance of temporal information and facts, listed in chronological order, from previous therapies and symptoms, through the current state, to future interventions. To enable reliable and accurate conclusions to be drawn from the collected data and the analysis, these texts must contain as much detail as possible. Their recording is time-consuming, and the doctors rarely have enough time. Since these texts are written by physicians for their colleagues and other medical professionals who have the same expert knowledge, a variety of non-standard terms and abbreviations are commonly used, providing a concise but sufficiently detailed description at the same time. Thus, in addition to the commonly used TEs, certain abbreviations of temporal meaning are often used in medical narratives to express time (for instance, `qid` from the Latin *quarter in die* – four times a day, `bid` from the Latin *bis in die* – twice a day, `post op` – postoperative, after a surgical operation). In medical narratives, like in any other natural language text, the same temporal information can be expressed in various forms (e.g. *January 3, 2007*; *03.01.07*; *01/03/2007*). An extensive review and classification of TEs found in discharge summaries, as one type of medical narrative text, are given in (Zhou et al., 2006). However, since clinical notes are written in a short time in a specific, medical language, characterized by fragmented and incomplete utterances, the lack of punctuation marks and formatting, as well as many spelling and grammatical errors, the number of forms of TEs that need to be recognized, due to their importance in interpreting clinically relevant data, becomes significantly higher (e.g. *03 01 07*; *03.01,2007*; *3.01..2007*; *0'3.01.2007*).

Another challenge in recognizing and formalizing the TEs of medical narrative texts is the fact that temporal information does not necessarily have to be explicit, but is often implicit and requires an interpretation or inferences based on general knowledge. In addition to absolute, relative TEs whose value is not explicitly stated and have to be taken from the context are also frequently used. These expressions require the value of another TE serving as an anchor for determining which particular time is meant. For example, the value of the expression *12. mart o.g.* 'March 12 this year' is relative to the date clinical note was written, while in the case of the *trećeg p.op dana* 'the third postoperative day', the calendar date of the mentioned surgery will serve as an anchor. In the statement „ranije lečena od dijabetične flegmone" 'previously treated for diabetic phlegmon' the temporal information processing system should determine if „ranije" 'previously' refers to a

period of, for instance, a week or a year earlier, requiring higher levels of analysis and resolution of co-references, which is an unresolved issue that has not yet been adequately discussed in the context of medical narrative data processing (Sun et al., 2013b).

Given the nature of medical narratives language, which is characterized by inconsistent application of grammatical and spelling rules, frequent use of abbreviations and text parts copied from other texts, it is necessary to collect and analyze as many different types of medical narrative texts as possible, to allow the development of a system that would perform high-response and precision automatic analysis. However, this is not a simple task, given the need to protect the right to the confidentiality of patients' health information. In addition to a large amount of useful, medically relevant information, the narrative texts of this domain also contain many personal details, including a patient's medical condition. For both ethical and legal reasons, when confidential clinical data are shared and used for research purposes, it is necessary to protect patient privacy and remove patient-specific identifiers through a process of de-identification. De-identification is focused on detecting and removing/modifying all explicit personal Protected Health Information (PHI) present in medical or other records, while still preserving all the medically relevant information about the patient. These PHI categories include names, geographic locations, elements of dates (except year), telephone and fax numbers, medical record numbers or any other unique identifying numbers, among others.

2.1 Previous work in TE medical narratives processing

Automatic processing of temporal information is one of the dominant natural language processing areas, although it could not be said to have had the same treatment in all domains of language use. Despite the progress made in the general domain characterized by the narrative style of a standard language, which is a rich source of TEs and relations examples, more active work on revealing temporal information contained in medical narratives has only been occurring for the last ten years (Lin et al., 2015). Nevertheless, the problem of representation and inference using some of the temporal aspects has attracted the attention of biomedical researchers for decades (Savova et al., 2009; Meystre et al., 2008; Augusto, 2005). Several studies aimed at processing the temporal information of medical texts were conducted even in the late 1980s (Johnson, 1987; Hirschman, 1981; Obermeier, 1985). The first system to identify words and linguistic forms that

carry temporal meaning in medical texts was developed at the University of New York under the LSP (Linguistic String Project) program (Sager, 1967). This system, designed to process standard English, is adapted to the needs of applying to medical texts - LSP-MLP (Linguistic String Project - Medical Language Processor), with the aim of identifying, analyzing and formalizing various forms of temporal information into a representation that will allow physicians to extract and summarize details on the symptoms, administration, and dosage of drugs, as well as reactions to their administration and possible side effects (Hirschman, 1981; Lyman et al., 1985). Also, one of the first systems for temporal information of medical texts analysis, known as GROK (Grammatical Representation of Objective Knowledge) (Obermeier, 1985), was developed using texts related to liver disease. The result of this system was a text representation based on knowledge of medical field key concepts, within which relevant medical events were extracted and chronologically ordered.

During the last decade of the 20th century, an increasing number of researchers have been concerned with the role of time in medicine and the temporality of medical narrative texts, publishing the achieved results both in medical and computer science journals as well as in conference proceedings (Keravnou, 1991; Goodwin and Hamilton, 1996; Combi and Shahar, 1997). The most significant system developed during that period at Columbia University in New York (Friedman et al., 1993), known as MedLEE (Medical Language Extraction and Encoding System), has been used since 1995 for the needs of the Presbyterian Hospital of Columbia in New York City. However, during this period, a large number of researchers interested in temporal resolution in the clinical domain dealt primarily with structured data in the form of explicitly encoded events (laboratory tests, doctor visits, etc.), which were dated and stored in databases. A comprehensive methodological overview of the approaches developed and the proposed standards for formalizing clinical-domain temporal information, which will enable the exchange of data among healthcare information systems, is given in (Augusto, 2005).

More recently, there has been an increasing interest of researchers in the effective use of temporal information of medical narratives and their incorporation into information extraction systems. Solving the problem of extraction of temporal information in the field of medicine, unlike the general domain, requires consideration of the system's design to be used for decision support in diagnosis, treatment determination, clinical data summarization, epidemiological studies conduction, etc. (Adlassnig et al., 2006). For example,

in (Denny et al., 2010) the authors developed a system that extracts the time and status of screening tests from electronic health records for the early detection of colorectal cancer. Similar ideas were applied to determine the status of the drugs used. To study patients' exposure to Warfarin[2] at the time of hospital admission, Liu and collaborators (2011) developed a system that extracts data on the use of certain medicines at a given time from patients' electronic health records. Concerning the automatic processing of temporal expressions of unstructured data, such as medical narratives, several papers provide a summary of the published literature (Meystre et al., 2008; Zhou and Hripcsak, 2007).

The problem of automatic processing of temporal information of general-domain narrative texts has been the focus of several international challenges, which have been addressed by a large number of researchers, primarily due to the existence of a sufficient number of publicly available annotated corpora necessary for the development of these systems. However, given the medical texts nature whose public use is not possible, more active involvement in the automatic processing of temporal information of medical texts was enabled in 2012 when for the i2b2 (Informatics for Integrating Biology and the Bedside) challenge was established the first English language clinical corpora annotated with temporal information (Sun et al., 2013a). The previous tasks of this challenge, which has been organized by the US National Center for Biomedical Informatics since 2007, were focused on automatic processing of clinical texts and problem solving (such as the deidentification of protected health information, extraction of medically relevant concepts and the relationships that exist between them, as well as identifying basic concept co-references).

The theme of the sixth i2b2 challenge, involving 18 teams from around the world, was the extraction of temporal information from clinical narrative texts such as the extraction of events (e.g. patient's medical problems, tests, therapies), TEs and temporal relations. Considering the widespread use of the TimeML specification language for temporal annotation, which also served as the basis for the development of ISO-TimeML standard (ISO, 2009), a modified version of this guideline, currently known as THYME-TimeML (Temporal Histories of Your Medical Events), developed specifically for the clinical domain was used (Styler IV et al., 2014a,b). More recently, several studies that have been addressing the possibility of customizing the

[2] Warfarin is an anticoagulant used to prevent heart attacks, strokes, and blood clots.

TimeML guidelines for temporal annotation of medical narratives (Savova et al., 2009; Galescu and Blaylock, 2012) and good clinical practice guidelines (Wenzina and Kaiser, 2014a,b) have achieved encouraging results. In addition to the TimeML adaptation for medical texts, some authors have also explored other options (Zhou et al., 2006; Tao et al., 2011).

The most successful i2b2 challenge systems in 2012 used already existing systems designed for recognition and normalization of TEs in newspaper texts, such as HeidelTime, SUTIME, GUTIME, NorMA systems, applying rule-based or hybrid approaches, combining hand-crafted rules and machine learning methods (Sun et al., 2013a). The analysis of the achieved results has shown that for most systems normalization of relative TEs is a more complex problem requiring further research.

Following the organized i2b2 challenge, the scientific community's interest in extracting clinical domain temporal information continues as part of the SemEval competition, held in 2015. As one of the tasks of the SemEval-2015 challenge, Clinical TempEval (Bethard et al., 2015) was a task aimed at identifying the extent and main features of TEs in clinical notes and pathology reports of the Mayo Clinic, Rochester, Minnesota. Clinical TempEval 2017 follows in the footsteps of the i2b2 2012 shared task, Clinical TempEval 2015, and Clinical TempEval 2016 in bringing timeline extraction to the clinical domain. As in past Clinical TempEvals, data were drawn from clinical notes and pathology reports for cancer patients at the Mayo Clinic. The key challenge for the Clinical TempEval 2017 task was the adaptation of the systems trained on data from colon cancer patients to make predictions on brain cancer patients.

Besides the mentioned international challenges, other systems have been developed to extract temporal information from medical narrative texts, primarily based on already existing general-purpose systems. For example, by modifying existing and creating new TARSQI rules, the Med-TTK system has been developed for use in the medical domain (Reeves et al., 2013). The normalization of recognized TEs is not covered by this system. Regarding normalization, the current problem is the normalization of the relative TEs of clinical texts (Sun et al., 2015).

For the most part, the research conducted in the field of automatic annotation of TEs of medical narrative texts relates to documents written in English. However, some efforts to adapt and develop systems for medical narratives in other languages, such as French (Hamon and Grabar, 2014), Chinese (Xiaojia et al., 2011), Swedish (Velupillai, 2014), Bulgarian (An-

gelova and Boytcheva, 2011; Boytcheva and Angelova, 2012; Boytcheva et al., 2012) have also been noted.

2.2 The aim of this paper

The automatic processing of temporal information of Serbian medical narrative texts has not been the subject of research so far. Considering the importance of time in interpreting medically relevant information, it would be useful to develop a system for the automatic processing of TEs of medical narrative texts, which would influence the development of other applications of automatic processing of medical texts, and consequently improve the quality of healthcare services, diagnostic and theoretical processes in medicine. Because medical domain narratives are stylistically very different from general natural language texts, natural language processing technologies developed for other domains cannot simply be applied to clinical domain texts, but require some modifications and adjustments.

Therefore, the aim of this paper is to apply the existing system for automatic processing of TEs in Serbian natural language texts to medical narrative texts, to evaluate the system's efficiency in recognition and normalization of TEs and to determine the degree of necessary adaptation according to the characteristics and requirements of the medical domain. Some *ad-hoc* solutions and preliminary evaluation of the system's performance conducted in this paper will serve as a basis for further research on the automatic processing of Serbian language in the medical field.

3 Recognition and normalization of TEs in the corpus of medical narratives

The automatic processing of TEs involves two stages. The first phase - recognition of TEs - concerns the identification of those fragments of text that carry temporal meaning and which represent the full range of TEs present in the texts, as well as the determination of their type. The second phase of TEs annotation is the normalization of their values, expressed explicitly or implicitly in the text.

Similar to TEs of standard Serbian, recognizing the TEs of medical narrative texts involves identifying and determining the range of linguistic expressions of absolute (e.g. *27. juna 2003. godine* 'June 27th 2003', *15:35 časova* '15:35 o'clock', *proleća 2015.* 'spring 2015', *tri nedelje* 'three weeks') or relative (e.g. *16. maja* 'May 16th', *sutra uveče* 'tomorrow evening', *sledećeg*

meseca 'next month', *nekoliko godina* 'a few years') temporal meaning, represented in medical texts by various formal units. Based on the TimeML guideline, the basic semantic types of TEs that need to be recognized are those that indicate a temporal location in the form of calendar date (DATE) or time of a day (TIME), duration (DURATION), and frequency (SET). The TEs' semantic classes QUANTIFIER and PREPOSTEXP, as suggested by the THYME annotation guideline, will not be taken into account at this time in identifying TEs because it is about applying an existing system rather than upgrading it.

The process of TEs' normalization refers to the interpretation of the recognized TE values, in a standardized form that complies with the ISO 8601 standard. All identified TEs are marked by inserting a `<TIMEX3>` tag with the following defined attributes:

- the attribute `type`, specifying the semantic class of the recognized expression (e.g. DATE, TIME, DURATION, SET, DATE.PERIOD);
- the attribute `temporalFunction`, indicating the absolute or relative value of the identified TE (e.g. false, true);
- the attribute `value`, representing the finite values of TEs in a standardized form (e.g. 2007-01-03T19:30, XXXX-12-23, P4M);
- the attribute `mod`, used to represent the meaning of quantified or modified TEs (e.g. START, MID, END, LESS_THAN, APPROX);
- the attribute `valueFromFunction`, used in the case of relative TEs, indicating the function that must be performed to calculate the absolute value of the identified expression (e.g. -1W, +1D);
- the attributes `quant` and `freq`, used to complete information about the meaning of TEs that indicate the frequency at which TE regularly reoccurs (e.g. `quant`="EVERY", `freq`="6X").

Several examples below illustrate the required semantic classes of TEs, as well as how to tag the results (Example 1).

Example 1.
(a)
23.12.2011. 'December 23, 2011'

```
<TIMEX3 type=‘‘DATE‘‘ temporalFunction=‘‘false‘‘ val=‘‘2011-12-23‘‘>
23.12.2011.</TIMEX3>
```

martu prošle godine 'March last year'

```
<TIMEX3 type=‘‘DATE‘‘ temporalFunction=‘‘true‘‘ val=‘‘XXXX-03‘‘
```

```
valueFromFunction=‘‘-1Y‘‘>martu prošle godine</TIMEX3>
```

(b)
07.10.2012. godine oko 16 h 'October 7, 2012 around 4 p.m.'

```
<TIMEX3 type=‘‘TIME‘‘ temporalFunction=‘‘false‘‘
val=‘‘2012-10-07T16‘‘
mod=‘‘APPROX‘‘>07.10.2012. godine oko 16 h</TIMEX3>
```

ujutru 'the morning'

```
<TIMEX3 type=‘‘TIME‘‘ temporalFunction=‘‘true‘‘
val=‘‘TMO‘‘>ujutru</TIMEX3>
```

(c)
više od godinu dana 'more than a year'

```
<TIMEX3 type=‘‘DURATION‘‘ temporalFunction=‘‘false‘‘ val=‘‘P1Y‘‘
mod=‘‘MORE_THAN‘‘>više od godinu dana</TIMEX3>
```

više godina 'more years'

```
<TIMEX3 type=‘‘DURATION‘‘ temporalFunction=‘‘true‘‘ val=‘‘PXY‘‘>
više godina</TIMEX3>
```

(d)
6 puta nedeljno 'six times a week'

```
<TIMEX3 type=‘‘SET‘‘ temporalFunction=‘‘false‘‘ val=‘‘P1W‘‘
freq=‘‘6X‘‘>6 puta nedeljno</TIMEX3>
```

svaki dan 'every day'

```
<TIMEX3 type=‘‘SET‘‘ temporalFunction=‘‘false‘‘ val=‘‘P1D‘‘
quant=‘‘EVERY‘‘>svaki dan</TIMEX3>
```

(e)
od 29.09.2009. do 03.10.2009. 'from 29.09.2009. to 03.10.2009.'

```
OD <TIMEX3 type=‘‘DATE.PERIOD‘‘>
<TIMEX3 type=‘‘DATE‘‘ temporalFunction=‘‘false‘‘
val=‘‘2009-09-29‘‘>29.09.2009.</TIMEX3>
DO <TIMEX3 type=‘‘DATE‘‘ temporalFunction=‘‘false‘‘
val=‘‘2009-10-03‘‘>03.10.2009.</TIMEX3>
</TIMEX3>
```

10-14 dana '10-14 days'

```
<TIMEX3 type=‘‘DURATION.PERIOD‘‘>
<TIMEX3 type=‘‘DURATION‘‘ temporalFunction=‘‘false‘‘
val=‘‘P10D‘‘>10</TIMEX3>
-
<TIMEX3 type=‘‘DURATION‘‘ temporalFunction=‘‘false‘‘ val=‘‘P14D‘‘>14
dana</TIMEX3>
</TIMEX3>
```

3.1 Finite-State methodology

The existing system for the automatic processing of TEs of Serbian newspaper texts is applied to medical narrative texts to identify and normalize the TEs contained in this document type. This system is a part of the Serbian system for named entity recognition (Krstev et al., 2014). It is a rule-based system that relies on lexical resources and handles both absolute and relative time. Its role is recognizing temporal expressions in unstructured texts and re-interpreting their temporal semantics in a standard format, according to the TimeML annotation guidelines, specified in (ISO, 2009).

Resources for natural language processing of Serbian are being developed using the finite-state methodology as introduced by Maurice Gross and LADL (Laboratoire d'Automatique Documentaire et Linguistique) laboratory (Gross, 1993). For the development and application of these resources the Unitex corpus processing system[3] was used (Paumier, 2016). The processing of TEs is carried out on a text having undergone sentence segmentation and part-of-speech tagging and morphological analysis. On the input text, general-purpose lexical resources (electronic dictionaries and dictionary finite-state transducers (FSTs)) are applied to tag text with lemmas, grammatical categories, and semantic features. After a text is being tagged this way, the system for TE recognition is applied.

The Serbian language temporal recognition system is based on the transducer cascade - CasSys (Friburger, 2002; Friburger and Maurel, 2004), which is integrated into the Unitex system. Cascade is a simple and effective way of organizing FSTs that may greatly increase the precision and speed of the system, as well as containment of ambiguity. The cascade for TE recognition currently consists of 16 transducers, whose role is to identify the expression,

[3] https://unitexgramlab.org

as well as to determine the range and type of each expression detected, following the TimeML scheme (DATE, TIME, DURATION and SET). The created graphs of this extensive grammar in the form of a series of transducers are designed to identify expressions that indicate calendar dates, times of day, duration and frequency of reoccurring times. The entire process (from the pre-processing of the text, through the creation of recognition rules, to the extraction of temporal information) is performed through the Unitex graphical user interface and the working environment.

Each transduction is defined by a set of patterns. For the most frequent alternative forms of dates and times represented in Serbian, corresponding FSTs were built and applied to text to recognize patterns described in the input alphabet. When the pattern was matched, the output alphabet specified the action to be taken. For instance, FST *Datum* in Figure 1 recognizes some possible date patterns that consist of a day (written using digits or letters) followed by a month (written in letters or Roman numerals) followed by a year (written using digits), as well as incomplete date expressions in which year is omitted. The output contains the TE described in the input and with the addition of a lexical tag that can be used in a subsequent FSTs. Semantic markers associated with recognized expressions provide useful information primarily regarding the type of the named entity (+time), as well as temporal expression (+date, +hour,[4] +duration, +set), as given in Example 2. Additional information concerning the type of TE is provided by semantic markers +abs and +rel used to indicate absolute and relative TEs, respectively. Lexical tags produced by FSTs, even though the most convenient for the use of subsequent FSTs in the cascade, are not useful for other applications and at the end are converted to the XML tags (Example 2).

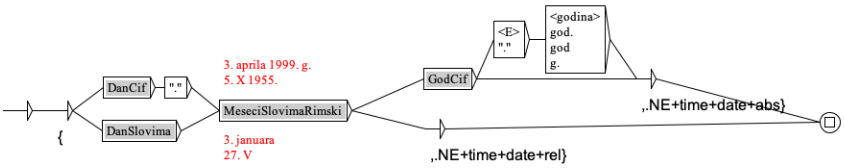

Figure 1. One path from FST *Datum* that recognizes complete and incomplete date expressions

[4] Corresponds to TimeML type TIME.

Example 2.
23.12.2011.
`{23.12.2011.,.NE+time+date+abs}`
`<time.date.abs>23.12.2011.</time.date.abs>`

3. januara 'January 3rd'
`{3. januara,.NE+time+date+rel}`
`<time.date.rel>3. januara</time.date.rel>`

Applied to a text in a predefined order, FSTs first track down patterns of dates, times, and durations that can be retrieved with a high degree of certainty, while the retrieval of more complex TEs is postponed (e.g. temporal ranges, conjoined expressions, combinations of calendar dates and times-of-day). Furthermore, there are a lot of appearances of numerals that do not necessarily have to refer to the time of a day, and the cascade provides the right context for disambiguation. For instance, numerals that occur together with some already tagged dates could be reliable indicators of some time patterns after which words *čas* or *sat* 'hour' do not appear (Figure 2).

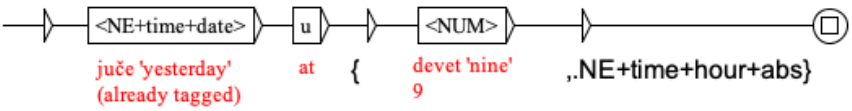

Figure 2. A simplified path in an FST that a numeral occurring after already marked date tags as time

The system responsible for the normalization of recognized TEs was also developed using the finite state methodology. As well as for TEs recognition, the Unitex software tool was used to create a collection of FSTs describing normalization rules and their application. Unlike the TEs recognition in which rules given by finite transducers are performed in a cascade order, whereby one transducer uses the results of previously applied ones, in the phase of TEs normalization rules are applied in the form of the FSTs local grammar to interpret the value of recognized expressions. Since the interpretation of TE value depends directly on its type, the local grammar aimed for normalization consists of four major transducers, each corresponding to

existing TE semantic classes. For instance, the FST given in Figure 3 illustrates the normalization of commonly used forms of relative calendar dates. The complete methodology used for recognition and normalization of TEs

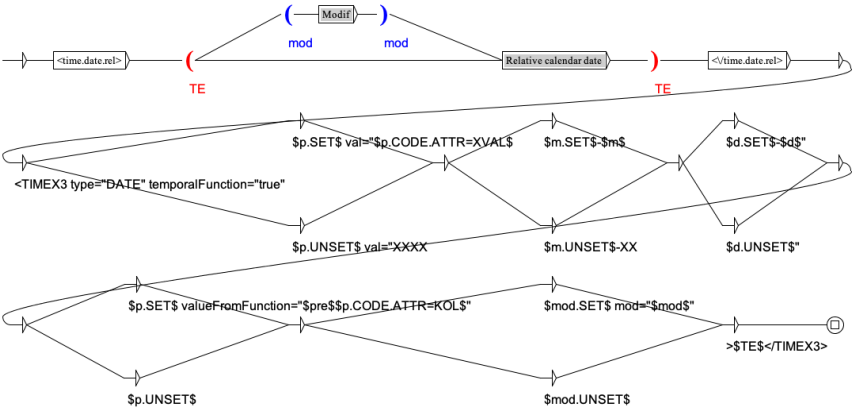

Figure 3. Transducer *Date relative* output

in Serbian natural language texts is described in detail in (Jaćimović, 2013, 2016).

3.2 Serbian medical narrative corpus

To evaluate the performance of the system in the automatic processing of TEs of medical narrative texts, a corpus of 150 randomly selected documents was used, consisting of discharge lists (n=100) and physician reports (n=50) of two teaching units of the School of Dental Medicine, University of Belgrade. The texts used for evaluation were previously automatically de-identified (Jaćimović et al., 2015), replacing personal and location names, as well as other health protected information with fictitious data, while TEs were not changed.[5] The selected texts were not used in the development

[5] Since these texts represent material that has not been fully de-identified, the medical records have been used with the consent of the School of Dental Medicine, University of Belgrade.

of the system and present entirely new material, suitable for preliminary evaluation.

Discharge lists and reports are unstructured, natural-language texts typed by physicians at the conclusion of hospital treatment or after multiple patient visits, and contain patient medical histories, current physical status, prescribed medication, laboratory test results, diagnostic findings, recommendations on the discharge of patients and other information relating to the patient's health. The dimensions of the corpus used, measured by the number of words and sentences, are presented in Table 1. Sentence segmentation was performed automatically (as described in 3.1), so the given numbers are considered only as approximate values, due to possible mistakes made during the text segmentation. The corpus used to evaluate the system, like all other medical narrative texts, is characterized by unfinished sentences, lack of punctuation marks, and an unusually large number of spelling and typographical errors, much larger than, for example, in the newspaper texts. Given the circumstances in which these texts originate, it is quite expected that many errors will occur, as one of the essential characteristics of medical narratives. Unlike newspaper texts written in standard Serbian, in which only 3-4% of words are unknown after applying electronic dictionaries, medical narrative texts contain slightly more than 20% of unrecognized words. This is understandable since these are specific domain texts characterized by terminology not yet included in the standard Serbian electronic dictionaries.

Corpus	**Number of words**	**Number of sentences**
Discharge lists	23.175	2.167
Physician reports	13.373	1.219
Total	36.529	3.386

Table 1. The dimensions of the corpus of medical narrative texts used

4 Implementation results and system performance evaluation

The Serbian system for automatic processing of TEs has been evaluated on the corpus of medical narratives previously described in Subsection 3.2.

Since the application of any automated information retrieval and extraction system results in some errors, it is necessary to determine their extent and type before drawing conclusions and conducting further research. Possible error types of an automated temporal processing system can occur concerning both recognition and normalization of the recognized TEs' values. First of all, in the recognition phase, errors occur in the form of missing slots, which are not and should have been recognized as a TE; then, incorrect slots, which represent expressions incorrectly identified as a TE; lastly, errors in the form of a wrongly defined range or type of TE. During the normalization of TE values, it is possible to misdefine or omit the attribute values used for the interpretation of recognized TE values. Source of errors may be the result of errors in the text itself or omitted or insufficiently defined rules of the grammar used. The extensive list of error types and their examples will be given in Subsection 4.1.

Automatically recognized and normalized values are controlled by adding the `proveraP` 'checkP' and `proveraN` 'checkN' attributes (explained in detail below) within the `<TIMEX3>` tag to define the type and source of the error (Table 2). All unrecognized expressions are assigned `<TIMEX3>` tags and a `proveraP` attribute whose value is `MISS`. The examination was performed by

Value	**Meaning**
`OK`	Correctly recognized/normalized
`UOK`	The TE range or some of the attributes are omitted or not correctly defined
`UOK/E`	The TE range or some of the attributes are not correctly defined or missed due to a text error
`NOK`	All attributes incorrectly defined (`NOKp` – recognition, `NOKn` – normalization)
`MISS`	Missed TE
`MISS/E`	Missed TE due to a text error

Table 2. The general values of the attributes used for evaluation (`proveraP` and `proveraN`)

one annotator (with experience in evaluating the named entity extraction system) based on a comprehensive instruction manual, with the control of the applied system's author. For all correctly recognized or normalized TEs,

the value `OK` is assigned to the attributes `proveraP` and `proveraN` (Example 3).

Example 3.
(a)
1978 GODINE
```
<TIMEX3 proveraP=‘‘OK‘‘ proveraN=‘‘OK‘‘ type=‘‘DATE‘‘
temporalFunction=‘‘false‘‘ val=‘‘1978‘‘>1978 GODINE</TIMEX3>
```

prošlog meseca ’last month’
```
<TIMEX3 proveraP=‘‘OK‘‘ proveraN=‘‘OK‘‘ type=‘‘DATE‘‘
temporalFunction=‘‘true‘‘ val=‘‘XXXX-XX‘‘ valueFromFunction=‘‘-1M‘‘>
prošlog meseca</TIMEX3>
```

(b)
oko 18 h ’around 6 p.m.’
```
<TIMEX3 proveraP=‘‘OK‘‘ proveraN=‘‘OK‘‘ type=‘‘TIME‘‘
temporalFunction=‘‘false‘‘ val=‘‘T18‘‘ mod=‘‘APPROX‘‘>oko 18
h</TIMEX3>
```

(c)
više od tri dana ’more than three days’
```
<TIMEX3 proveraP=‘‘OK‘‘ proveraN=‘‘OK‘‘ type=‘‘DURATION‘‘
temporalFunction=‘‘false‘‘ val=‘‘P3D‘‘ mod=‘‘MORE_THAN‘‘>više od
tri dana</TIMEX3>
```

(d)
svake godine ’every year’
```
<TIMEX3 proveraP=‘‘OK‘‘ proveraN=‘‘OK‘‘ type=‘‘SET‘‘
temporalFunction=‘‘false‘‘ val=‘‘P1Y‘‘ quant=‘‘EVERY‘‘>svake
godine</TIMEX3>
```

In situations where none of the attributes assigned at the recognition or normalization stage are correctly defined, the value `NOK` is assigned to the attributes used for evaluation. Given the structure of the system, which does not allow normalization of values of temporal expressions that were not previously recognized, missed expressions due to a text error can only occur with the recognition of TEs when `MISS/E` value is assigned to the attribute `proveraP`.

For incorrectly defined or omitted values of the <TIMEX3> tag, value UOK is assigned to the attribute used for evaluation (proveraP or proveraN), with an indication mark which attribute is incorrectly specified. A detailed view of the <TIMEX3> attribute marks used for evaluation purposes is illustrated in the Table 3. For instance, if the TE type is incorrect, the recognition

Attribute	**Mark**
type	t
range	o
value	v
mod	m
valueFromFunction/function	f
quant	q
freq/frequency	u

Table 3. <TIMEX3> attribute marks

attribute proveraP will be assigned the value UOKt, while in the case of the misdefined mod and valueFromFunction attributes, the proveraN attribute will be assigned a UOKmf value. If an incorrectly defined range or a type of recognized expression is the result of a text error, the proveraP attribute will be assigned UOKo/E, or UOKt/E value, respectively.

The possible attribute values for proveraP are as follows:

- OK – correctly determined type and precisely defined full range of TE;
- UOK – some of the attributes are incorrect;
  - UOKt (the expression type is not correctly determined, but the full range is accurately defined);
  - UOKt/E (the expression type is not correctly determined due to a text error, but the full range is accurately defined);
  - UOKo (the type of expression is accurately specified, but the full range is not correctly defined);
  - UOKo/E (the expression type is accurately specified but the full range is not correctly defined due to a text error);
- NOKp – incorrectly defined both the type and range of the TE;
- NOK – an expression that is incorrectly defined and tagged as a TE (the recognized expression is not TE);

- `MISS` – missed expression (not recognized, although it should have been);
- `MISS/E` – the expression was not recognized due to a text error.

The possible values for the `proveraN` attribute are the following (used `<TIMEX3>` attribute marks are explained in Table 3):

- `OK` – normalization is done correctly;
- `UOK` – some of the attributes is incorrect;
  - `UOKv` (attribute `value` is incorrect);
  - `UOKm` (attribute `mod` is incorrect);
  - `UOKf` (attribute `valueFromFunction` is incorrect);
  - `UOKmf` (attributes `mod` and `valueFromFunction` are incorrect);
  - `UOKvf` (attributes `value` and `valueFromFunction` are incorrect);
  - `UOKvm` (attributes `value` and `mod` are incorrect);
  - `UOKq` (attribute `quant` is incorrect);
  - `UOKu` (attribute `freq` is incorrect);
  - `UOKvq` (attributes `value` and `quant` are incorrect);
  - `UOKvu` (attributes `value` and `freq` are incorrect);
- `NOKn` – all normalization attributes are incorrectly defined.

The results of evaluation show that out of the total number of existing TEs (884), 772 (87.3%) TEs indicating the date, time, duration and frequency occur in the corpus of medical narrative texts, and 112 (12.7%) expressions that indicate periods (Table 4). In this corpus, 613 (69.3%) abso-

TIMEX3	**abs**	**rel**	**Total**
DATE	402	40	442
TIME	43	37	80
DURATION	168	22	190
SET			60
Total	613	99	772
DATE.PERIOD			109
DURATION.PERIOD			3
Total	613	99	884

Table 4. The proportion of the TE types in the corpus of medical narrative texts

lute TEs were found that convey the meaning of a point in time and duration

(DATE, TIME, and DURATION), while 99 (11.2%) belong to the group of relative expressions. Regarding the expressions indicating the frequency of occurrences in time, a total of 60 were identified (6.8%), which is expected for the narrative texts of the medical domain.

During the verification of TEs defined by the TimeML guideline, identification and tagging of expressions characteristic of medical narrative texts (indicating a specific period associated with an event, based on the THYME instruction) were manually performed for later research. Thus, in the corpus used, 26 terms were found belonging to the newly defined PREPOSTEXP class (Example 4).

Example 4.
(a)
Postoperativno sprovedena RT. 'Postoperatively conducted RT.'
```
<TIMEX3 type=‘‘PREPOSTEXP‘‘>Postoperativno</TIMEX3> sprovedena
RT.
```

(b)
Rezultat trombektomije dobar,ali trećeg p.op dana dolazi do ponovne tromboze.
'Thrombectomy result is good,but on the third p.op day thrombosis recurs.'
```
Rezultat trombektomije dobar,ali <TIMEX3 type=‘‘PREPOSTEXP‘‘
val=‘‘POD8‘‘>trećeg p.op dana</TIMEX3> dolazi do ponovne
tromboze.
```

(c)
... otpušten je iz bolnice 10. dana kao oporavljen ...
'... discharged from the hospital on 10^{th} day as recovered ...'
```
...otpušten je iz bolnice <TIMEX3 type=‘‘PREPOSTEXP‘‘
val=‘‘HD10‘‘>10. dana</TIMEX3> kao oporavljen ...
```

(d)
Međutim intraoperativno su nađeni nepovoljni uslovi ...
'However adverse conditions were found intraoperatively ...'
```
Međutim <TIMEX3 type=‘‘PREPOSTEXP‘‘>intraoperativno</TIMEX3> su
nađeni nepovoljni uslovi ...
```

(e)
Preoperativno korigovana kardiološka terapija.
'Preoperatively corrected cardiac therapy.'

```
<TIMEX3 type=‘‘PREPOSTEXP‘‘>Preoperativno</TIMEX3> korigovana
kardiološka terapija.
```

The evaluation of the system's success in recognizing and normalizing the TEs of Serbian medical narrative texts was performed based on standard measures for assessing the performance of the information retrieval and extraction system. The basic evaluation unit was a complete TE. The precision (p) of the system was observed, representing the ratio of the total number of correctly recognized TEs (`OK`) to the total number of recognized expressions (M) (Formula 1);

$$p = \frac{OK}{M} = \frac{OK}{OK + UOK + NOK} \tag{1}$$

next, response (r), as the ratio of the total number of correctly recognized TEs (`OK`) to the total number of existing expressions in the text (N) (Formula 2), and F_1 measure, as a common measure that reflects the harmonic mean of response and precision and always has a value that weighs less than the value of the response or precision achieved (Formula 3).

$$r = \frac{OK}{N} = \frac{OK}{OK + MISS} \tag{2}$$

$$F_1 = 2\frac{pr}{p + r} \tag{3}$$

The evaluation used a balanced F_1 measure, which favors neither response nor precision.

Regarding performance in recognizing TEs (determining range and type), the system achieved slightly higher accuracy (94%) than response (90%), with an overall F_1 measure of 92% (Table 5). The best results are obtained when identifying and determining the range and type of expressions that indicate a period or represent a sequence of multiple TEs that are in a particular relation. Errors made in this process are solely the result of typing errors in the text, or inconsistent application of spelling rules.

The system achieved the highest precision in determining the range and type of expressions indicating frequency, while at the same time the response in identifying this class of expressions was the lowest, with an achieved F_1 measure of 67%. A much lower response rate (86%) than the achieved accuracy (96%) also occurs in expressions that indicate duration, which, given the source of errors, may be explained by omissions in defining rules for identification of TEs. Regarding expressions with the meaning of a point in time

	N	M	OK	UOKt	UOKo	UOKo/E	MISS	MISS/E	NOK	*p*	*r*	F_1
DATE	442	419	393	0	17	9	14	7	0	0.94	0.95	0.94
TIME	80	80	67	1	5	5	2	0	2	0.84	0.97	0.90
DURATION	190	165	158	3	2	2	21	4	0	0.96	0.86	0.91
SET	60	30	30	0	0	0	28	2	0	1	0.5	0.67
Total	772	694	648	4	24	16	65	13	2	**0.93**	**0.89**	**0.91**
DATE.PERIOD	109	105	104	0	0	1	0	4	0	0.99	0.96	0.98
DURATION.PERIOD	3	3	3	0	0	0	0	0	0	1	1	1
Total	884	802	755	4	24	17	65	17	2	**0.94**	**0.90**	**0.92**

Table 5. Data on the results of the evaluation and the achieved system performance in recognizing the TEs of medical narrative texts based on existing semantic expression classes

i.e. calendar dates and times of the day, a better response (95% and 97%, respectively) to the achieved accuracy (94 % and 84 %) clearly indicates the need to modify the existing rules, with the aim of correct tagging the identified TEs range. Detailed information on the performance of the system in recognizing TEs is given in the Table 5, while a detailed error analysis will be conducted below (4.1). Column N indicates the total number of TEs that exist in the corpus of medical texts, while column M contains the total number of TEs identified by the system.

The results achieved with respect to the success of the TEs normalization process are very good (Table 6). The values of all correctly recognized

	N	M	OK	UOKv	MISS	MISS/E	NOK	*p*	*r*	F_1
DATE	419	419	419	0	0	0	0	1	1	1
TIME	80	80	78	2	0	0	0	0.98	1	0.99
DURATION	165	165	165	0	0	0	0	1	1	1
SET	30	30	30	0	0	0	0	1	1	1
Total	694	694	692	2	0	0	0	**0.997**	**1**	**1**

Table 6. Data on evaluation results and system performance achieved in normalizing TEs of medical narrative texts based on existing semantic expression classes

TEs are accurately normalized (the achieved response is 100%), assigning all the appropriate attributes necessary for their interpretation. Since for the expressions indicating the time of day, due to a wrongly defined range, incor-

rect values of the attribute `value` were entered, the precision in performing normalization is 99.7%. Column N shows the total number of TEs identified by the system in the corpus of medical texts (for which the normalization process is expected to be performed), while column M contains the total number of TEs normalized by the system.

4.1 Error analysis

Errors that occurred while applying the system, as well as its success in recognizing and normalizing TEs concerning existing semantic classes of TEs, are presented within Tables 5 and 6. The errors observed and the precision, response, and F_1 measure achieved in determining the range and assigning the necessary attributes, without considering the class to which the TE belongs, are illustrated by the data in the Table 7. Column N shows

	N	M	OK	UOK	UOKo/E	MISS	MISS/E	NOK	*p*	*r*	F_1
range	884	802	761	24	17	65	17	0	0.95	0.90	0.93
type	802	802	796	4	0	0	0	2	0.99	1	1
temporalFunction	694	694	694	0	0	0	0	0	1	1	1
value	694	694	692	2	0	0	0	0	0.997	1	1
mod	46	46	46	0	0	0	0	0	1	1	1
valueFromFunction	11	11	11	0	0	0	0	0	1	1	1
quant	12	12	12	0	0	0	0	0	1	1	1
freq	4	4	4	0	0	0	0	0	1	1	1

Table 7. Data on evaluation results and system performance achieved in determining the range and attributes of TEs

the total number of TEs that exist in the corpus of medical texts for which range determination and attribute assignment were expected, while column M contains the total number of TEs whose range and attributes were determined by the system. Considering the achieved F_1 measure, the biggest problem in the automatic processing of TEs, if not the only one, was identified already in the first step when recognizing and determining the range of expressions.

Expressions not identified and tagged by the system during processing make 9% of the total number of existing TEs in the corpus used (N). Almost a third of the missed expressions (n=28) are those expressing the frequency of the prescribed therapies (npr. *na **drugi dan*** 'on the second day', *infuzijom 25000 j / **24 h*** 'infusion of 25000 U/24 hr', *0,8 ml na **12 h*** '0,8 ml in 12 hrs', etc.). Omissions related to expressions that indicate calendar dates

and durations are primarily the results of the application of rules, which, for the sake of greater system precision, did not allow the identification of numerical terms that may be extremely ambiguous or unusual in the context of newspaper texts (Example 5).

Example 5.

(a)
Navodi da je 1995 a potom 96 operisala karcinom
'She states that in 1995 and then in 96 she had cancer surgery'
```
Navodi da je
<TIMEX3 proveraP=‘‘MISS‘‘ type=‘‘DATE‘‘>1995</TIMEX3>
a potom
<TIMEX3 proveraP=‘‘MISS‘‘ type=‘‘DATE‘‘>96</TIMEX3> operisala
karcinom
```

(b)
U 7 god. operisao ... 'In 7 yr. had surgery ...'
```
U <TIMEX3 proveraP=‘‘MISS‘‘ type=‘‘DURATION‘‘>7 god.</TIMEX3>
operisao ...
```

(c)
Pušač unazad oko 20 g. 'Smoker for about 20 y.'
```
Pušač unazad <TIMEX3 proveraP=‘‘MISS‘‘ type=‘‘DURATION‘‘>oko 20
g.</TIMEX3>
```

Omissions in the identification of TEs resulting from typographical errors make almost 17% of the total number of missed expressions. Some cases are illustrated with the Example 6.

Example 6.
emesec dana 'one omonth'
```
<TIMEX3 proveraP=‘‘MISS/E‘‘ type=‘‘DURATION‘‘>emesec dana</TIMEX3>

<TIMEX3 proveraP=‘‘MISS/E‘‘ type=‘‘DATE‘‘>18.o6.2009.</TIMEX3>

<TIMEX2 proveraP=‘‘MISS/E‘‘ type=‘‘DATE‘‘>26 07 2013</TIMEX3>
```

sptembru o.g. 'sptember this yr.'
```
<TIMEX3 proveraP=‘‘MISS/E‘‘ type=‘‘DATE‘‘>sptembru o.g.</TIMEX3>
```

svaakodnevno 'daaily'
<TIMEX3 proveraP=‘‘MISS/E‘‘ type=‘‘SET‘‘>svaakodnevno</TIMEX3>

(sati '(hours'
<TIMEX3 proveraP=‘‘MISS/E‘‘ type=‘‘SET‘‘>(sati</TIMEX3>

An analysis of errors that indicate an incorrectly defined range of the identified TE shows that 58.5% of them are indeed system flaws (Example 7), while the remaining 41.5% are errors due to existing spelling and typographical errors in the text (Example 8).

Example 7.
(a)
početkom septembra o.g. 'early September th.yr.'
<TIMEX3 proveraP=‘‘UOKo‘‘ proveraN=‘‘OK‘‘ type=‘‘DATE‘‘
temporalFunction=‘‘true‘‘ val=‘‘XXXX-09‘‘ mod=‘‘START‘‘>početkom
septembra</TIMEX3> o.g.

(b)
. . . povređen 22.12.2011. godine oko 01h 30 min.
'. . . injured 22.12.2011. around 01h 30 min.'
...povređen <TIMEX3 proveraP=‘‘UOKo‘‘ proveraN=‘‘OK‘‘
type=‘‘TIME‘‘ temporalFunction=‘‘false‘‘ val=‘‘2011-12-22T01‘‘
mod=‘‘APPROX‘‘>22.12.2011. godine oko 01h </TIMEX3> 30 min.

(c)
godinu i po dana 'year and a half'
<TIMEX3 proveraP=‘‘UOKo‘‘ proveraN=‘‘OK‘‘ type=‘‘DURATION‘‘
temporalFunction=‘‘false‘‘ val=‘‘P1.5Y‘‘>godinu i po</TIMEX3> dana

Example 8.
(a)
. . . po podnevnim časovima istog dana 'in the after noon same day'
...po podnevnim časovima <TIMEX3 proveraP=‘‘UOKo‘‘ proveraN=‘‘OK‘‘
type=‘‘DATE‘‘ temporalFunction=‘‘true‘‘ val=‘‘XXXX-XX-XX‘‘
valueFromFunction=‘‘=1D‘‘>istog dana</TIMEX3>

(b)
0'6.12.2012. 'December 6th, 2012'
0'<TIMEX3 proveraP=‘‘UOKo‘‘ proveraN=‘‘OK‘‘ type=‘‘DATE‘‘
temporalFunction=‘‘false‘‘ val=‘‘2012-12-06‘‘>6.12.2012.</TIMEX3>

Since TEs in the context of medical narrative texts are a special PHI category that is subject to secrecy, it is very significant to have the best possible response when identifying them, as well as to achieve high precision in range determination in order not to reveal some information. With this in mind, the results of the error analysis in determining the range of identified TEs show that in most cases the omissions made do not adversely affect the determination of the type of expression and the normalization of its value, which is extremely important for the de-identification process i.e. hiding actual dates and moving them for a randomly selected interval. For example, in determining the range of calendar dates, the most frequently omitted part is related to the element of the year expressed by the abbreviated form of the relative expression *o.g.* (*this year*) (Example 7.a). The identified part of the expression that has to be hidden *early September* is correctly defined as the relative calendar date, with correctly assigned values of the `value` and `mod` attributes.

Errors in determining the type of recognized TE account for only 0.7% of the total errors. In most cases, these are expressions that indicate the frequency of prescribed therapy, and they are recognized as expressions with a meaning of duration (Example 9).

Example 9.
... cefalosporinski preparat na 8 sati '... cephalosporin on 8 hours'
```
...cefalosporinski preparat na <TIMEX3 proveraP=‘‘UOKt‘‘
proveraN=‘‘OK‘‘ type=‘‘DURATION‘‘ temporalFunction=‘‘false‘‘
val=‘‘P8H‘‘>8 sati</TIMEX3>
```

Regardless of the wrong expression type, the form and value of the `value` attribute correspond to the defined expression form that indicates the frequency. Since the expression implies a repetition rate every 8 hours, the missing item is the `freq` attribute, whose value would be EVERY. Because the TimeML guideline suggests the `freq` attribute use only if the repetition rate is explicitly stated, the omitted attribute was not treated as an error this time.

Detecting and correcting errors in the form of *intruders*, or expressions that are misidentified as TEs, is especially important in the processing of medical narrative texts, particularly if the misspelled expression is an expression indicating some medically relevant information. The error analysis after the application of the system revealed only two such incorrectly identified expressions (Example 10).

Example 10.
. . . koštani stepenik u predelu rama leve orbite na oko 7 sati
'. . . bone step in the region of the left orbit frame at **around 7 o'clock**'
```
...koštani stepenik u predelu rama leve orbite na
<TIMEX3 proveraP=‘‘NOK‘‘ proveraN=‘‘OK‘‘ type=‘‘TIME‘‘
temporalFunction=‘‘false‘‘ val=‘‘T07‘‘ mod=‘‘APPROX‘‘>oko 7
sati</TIMEX3>
```

For example, in the field of maxillofacial surgery to diagnose a zygomatic fracture, an expression is used, which by analogy with the clockwise describes and localizes the occurrence of a bone step on the orbital frame (*at 5 o'clock position*).

Regarding other attributes defined by the TimeML guideline, the system made only two errors in determining the value of the most important attribute in the normalization process – `value` attribute. Both errors are the result of a wrongly defined range of TE (example 11), based on which the normalization of the value of the recognized expression was performed.

Example 11.
Detralex tbl 2 ujutro 'Detralex tbl 2 in the morning'
```
Detralex tbl <TIMEX3 proveraP=‘‘UOKo‘‘ proveraN=‘‘UOKv‘‘
type=‘‘TIME‘‘ temporalFunction=‘‘false‘‘ val=‘‘T02‘‘>2
ujutro</TIMEX3>
```

The number that was mistakenly identified as the absolute time of day and included in the range of expression is an indication of the amount of therapy prescribed, which should be tagged as a QUANTIFIER type according to the THYME manual.

Based on the calculated system performance as a traditional F_1 measure, the system error rate is calculated as E=1-F. However, since the importance of intruders and missed expressions has been minimized in the calculation of the harmonic response and precision (in that way the total error rate can never be greater than the error rate of any type of error separately), evaluation is also performed using the *Slot Error Rate* (SER). As a simple error measure, the SER equally weights different types of error directly, enabling the comparison of all systems against the fixed base. The SER is equal to the sum of all three types of errors (missed – `MISS`, intruders – `NOK` and non-matched – `UOK`) – divided by the total number of expressions existing in the reference corpus (N) (Formula 4).

$$SER = \frac{MISS + NOK + UOK}{N} = \frac{MISS + NOK + UOK}{OK + UOK + MISS} \qquad (4)$$

Based on the data in Tables 8, 9 and 10, it is clear that the error rate calculated based on the *SER* is about 50% higher than the error rate presented by the F_1 measure. The obtained *SER* values more emphasize the

	OK	**UOK**	**MISS**	**NOK**	F_1	***E***	***SER***
DATE	393	26	21	0	0.94	0.06	0.11
TIME	67	11	2	2	0.90	0.10	0.19
DURATION	158	7	25	0	0.91	0.09	0.17
SET	30	0	30	0	0.67	0.33	0.5
Total	648	44	78	2	0.91	0.09	**0.16**
DATE.PERIOD	104	1	4	0	0.98	0.02	0.05
DURATION.PERIOD	3	0	0	0	1	0	0
Total	755	45	82	2	0.92	0.08	**0.15**

Table 8. Values of the achieved F_1 measure and E and *SER* error measures in recognizing the TEs of medical narrative texts based on existing semantic expression classes

	OK	**UOK**	**MISS**	**NOK**	F_1	***E***	***SER***
DATE	419	0	0	0	1	0	0
TIME	78	2	0	0	0.99	0.01	0.03
DURATION	165	0	0	0	1	0	0
SET	30	0	0	0	1	0	0
Total	692	2	0	0	1	0.001	**0.003**

Table 9. Values of the achieved F_1 measure and E and *SER* error measures in normalizing the TEs of medical narrative texts based on existing semantic expression classes

system's errors in the process of recognizing and normalizing the TEs of medical narrative texts.

Since medical narrative texts differ stylistically from general domain texts (Sun et al., 2013b) and that one-third of these texts consists of incomplete

	OK	**UOK**	**UOK/E**	**MISS**	**MISS/E**	**NOK**	F_1	***E***	***SER***
range	761	24	17	65	17	0	0.93	0.07	**0.14**
`type`	796	4	0	0	0	2	1	0.004	**0.008**
`temporal-Function`	694	0	0	0	0	0	1	0	0
`value`	692	2	0	0	0	0	1	0.001	**0.003**
`mod`	46	0	0	0	0	0	1	0	0
`value-From-Function`	11	0	0	0	0	0	1	0	0
`quant`	12	0	0	0	0	0	1	0	0
`freq`	4	0	0	0	0	0	1	0	0

Table 10. Values of the achieved F_1 measure and E and *SER* error measures in determining the range and attributes of TEs

sentences that are difficult to identify with conventional language parsers (Savova et al., 2009), automatic processing of medical narrative texts requires a slightly different approach. Given that in the corpus of medical texts used, 26.4% of the total number of errors occur due to inconsistent application of grammatical and spelling rules, as well as the existence of more typographical errors, the rules used to recognize TEs need to be modified in order to improve both the precision and response of the automatic processing process.

Semantic classes of TEs that occur in newspaper texts also exist in medical narrative texts, the most common being calendar dates and expressions indicating duration. However, unlike the general domain, the use of expressions that indicate frequency is more common within medical texts, notably the frequency of the use of medication therapies. In the corpus used, even 7.7% of expressions of this type were identified, which is an unusually high percentage compared to their representation in newspaper texts. Although a large number of expressions (for which recognition rules were created) were not found in the corpus of medical texts, a particular class of TEs (PREPOSTEXP), specific only to medical domain texts, was identified.

5 Conclusion

The system for recognition and normalization of Serbian language TEs, developed on the examples from newspaper articles, has been applied to

medical domain texts to determine the success of its application in this field. The results of the evaluation indicate that the created system, without any prior preparation and adaptation of rules to this specific domain, performs automatic recognition and normalization of TEs of medical narrative texts of the Serbian language, with high precision (94%). Although the achieved precision in automatic processing of TEs is lower than the precision achieved in the general domain (99%) (Jaćimović, 2016), it indicates that the system can be successfully applied in this domain as well, by adjusting the recognition rules based on the analysis of errors identified by the evaluation. Regarding the achieved response, the system applied to medical narrative texts performs better results with respect to the general domain (90% and 80%, respectively), which is understandable given the nature of medical texts that, compared to newspaper texts, contain a limited number of TE forms.

The error analysis revealed some omissions, which were essentially the result of the absence of appropriate rules, or the application of the existing rules, not allowing the identification of some expressions due to the higher precision of the system. The modularity of the system allows us to easily refine the existing and incorporate new rules for identifying those forms that are characteristic of medical texts, which will affect the improvement of both precision and response. Regarding the normalization of recognized expressions, as in the case of newspaper texts, the identified errors indicate that this process depends entirely on the success of the recognition (correct determination of the TEs range and type) since the values of all correctly identified and tagged expressions were normalized accurately. Given that medical narrative texts belong to a group of texts characterized by the existence of a large number of spelling and typographical errors, it would be useful to implement the existing text correction system (Stanković et al., 2011) in the further work.

References

Adlassnig, Klaus-Peter, Carlo Combi, Amar K Das, Elpida T Keravnou and Giuseppe Pozzi. "Temporal representation and reasoning in medicine: research directions and challenges". *Artificial intelligence in medicine* Vol. 38, no. 2 (2006): 101–113

Angelova, Galia and Svetla Boytcheva. "Towards temporal segmentation of patient history in discharge letters". In *Proceedings of the Second Workshop on Biomedical Natural Language Processing*, 2011, 49–54

Augusto, Juan Carlos. "Temporal Reasoning for Decision Support in Medicine". *Artificial intelligence in medicine* Vol. 33, no. 1 (2005): 1–24

Bethard, Steven, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky et al. "Semeval-2015 task 6: Clinical tempeval". *Proc. SemEval* (2015)

Boytcheva, Svetla and Galia Angelova. "A workbench for temporal event information extraction from patient records". In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, 48–58. Springer, 2012

Boytcheva, Svetla, Galia Angelova and Ivelina Nikolova. "Automatic analysis of patient history episodes in Bulgarian hospital discharge letters". In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 77–81. Association for Computational Linguistics, 2012

Combi, Carlo and Yuval Shahar. "Temporal reasoning and temporal data maintenance in medicine: issues and challenges". *Computers in biology and medicine* Vol. 27, no. 5 (1997): 353–368

Denny, Joshua C, Josh F Peterson, Neesha N Choma, Hua Xu, Randolph A Miller et al. "Extracting timing and status descriptors for colonoscopy testing from electronic medical records". *Journal of the American Medical Informatics Association* Vol. 17, no. 4 (2010): 383–388

Friburger, N. and D. Maurel. "Finite-state transducer cascades to extract named entities in texts". *Theoretical Computer Science* Vol. 313, no. 1 (2004): 93–104

Friburger, Nathalie. "Reconnaissance automatique des noms propres: application à la classification automatique de textes journalistiques". PhD. thesis, Tours, 2002

Friedman, Carol, James J Cimino and Stephen B Johnson. "A conceptual model for clinical radiology reports.". In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 829. American Medical Informatics Association, 1993

Galescu, Lucian and Nate Blaylock. "A corpus of clinical narratives annotated with temporal information". In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 715–720. ACM, 2012

Goodwin, Scott D and Howard J Hamilton. "It's about time: an introduction to the special issue on temporal representation and reasoning". *Computational Intelligence* Vol. 12, no. 3 (1996): 357–358

Gross, Maurice. "Local grammars and their representation by finite automata". *Data, Description, Discourse. Papers on the English Language in honour of John McH Sinclair* (1993): 26–38

Hamon, Thierry and Natalia Grabar. "Tuning HeidelTime for identifying time expressions in clinical texts in English and French". *EACL 2014* (2014): 101–105

Hirschman, L. "Retrieving time information from natural language texts". In *Information retrieval research*, Oddy, RN, SE Robertson, CJ Van Rijsbergen and Williams P, 154–171. Butterworths, London, 1981

ISO. "ISO/DIS 24617-1 Language Resources Management - Semantic Annotation Framework (SemAF) - Part 1: Time and Events (SemAF-Time, ISO-TimeML)", Standard, International Organization for Standardization. Geneva, Switzerland, 2009

Jaćimović, Jelena. "Automatic Processing of Temporal Expressions in Serbian Natural Language Texts". In *35th Anniversary of Computational Linguistics in Serbia*, 57–69, 2013

Jaćimović, Jelena. "Automatsko prepoznavanje i normalizacija vremenskih izraza u nestrukturiranim novinskim i medicinskim tekstovima na srpskom jeziku". PhD. thesis, Univerzitet u Beogradu, Filološki fakultet, Beograd, 2016

Jaćimović, Jelena, Cvetana Krstev and Drago Jelovac. "A rule-based system for automatic de-identification of medical narrative texts". *Informatica* Vol. 39, no. 1 (2015): 43–51

Johnson, S. "Temporal information in medical narrative". In *Medical Language Processing: Computer Management of Narrative Data*, Sager, N, C Friedman and MS Lyman, Reading, 175–94. MA: Addison-Wesley, 1987

Keravnou, Elpida. "Medical temporal reasoning". *Artificial Intelligence in Medicine* Vol. 3, no. 6 (1991): 289–290

Krstev, Cvetana, Ivan Obradović, Miloš Utvić and Duško Vitas. "A system for Named Entity Recognition Based on Local Grammars". *Journal of Logic and Computation* Vol. 24, no. 2 (2014): 473–489

Lin, Ching-Heng, Nai-Yuan Wu, Wei-Shao Lai and Der-Ming Liou. "Comparison of a semi-automatic annotation tool and a natural language processing application for the generation of clinical statement entries". *Journal of the American Medical Informatics Association* Vol. 22, no. 1 (2015): 132–142

Liu, Mei, Min Jiang, Vivian K Kawai, Charles M Stein, Dan M Roden et al. "Modeling drug exposure data in electronic medical records: an application to warfarin". In *AMIA annual symposium proceedings*, Vol. 2011, 815–823. American Medical Informatics Association, 2011

Lyman, M, N Sager, C Friedman and E Chi. "Computer-structured narrative in ambulatory care: its use in longitudinal review of clinical data". In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 82–86. American Medical Informatics Association, 1985

Meystre, Stéphane M, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle et al. "Extracting Information from Textual Documents in the Electronic Health Record: a Review of Recent Research". *Yearb Med Inform* Vol. 35 (2008): 128–44

Obermeier, Klaus K. "Temporal inferences in medical texts". In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, 9–17. Association for Computational Linguistics, 1985

Paumier, Sébastien. *Unitex 3.1 User manual*, 2016, http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf?

Reeves, Ruth M, Ferdo R Ong, Michael E Matheny, Joshua C Denny, Dominik Aronsky et al. "Detecting temporal expressions in medical narratives". *International journal of medical informatics* Vol. 82, no. 2 (2013): 118–127

Sager, Naomi. "Syntactic analysis of natural language", *Advances in computers* Vol. 8 (1967), no. 153–188: 35

Savova, G, S Bethard, W Styler, J Martin, M Palmer et al.. "Towards temporal relation discovery from the clinical narrative", In *AMIA... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium*, 568–572, 2009

Shahar, Yuval. "Timing Is Everything: Temporal Reasoning and Temporal Data Maintenance in Medicine". In *Artificial Intelligence in Medicine*, Horn, Werner, Yuval Shahar, Greger Lindberg, Steen Andreassen and Jeremy Wyatt, *Lecture Notes in Computer Science*, Vol. 1620, 30–46. Springer Berlin Heidelberg, 1999

Stanković, Ranka, Ivan Obradović, Cvetana Krstev and Duško Vitas. "Production of Morphological Dictionaries of Multi-Word Units Using a Multipurpose Tool". In *Proceedings of the Computational Linguistics-Applications Conference*, Jassem, K, PW Fuglewicz, M Piasecki and A Przepiórkowski, Jachranka, 77–84. Poland: Polish Information Processing Society, 2011

Styler IV, William F, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan et al. "Temporal annotation in the clinical domain". *Transactions of the Association for Computational Linguistics* Vol. 2 (2014a): 143–154

Styler IV, William F, Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman et al. *THYME annotation guidelines*, 2014b

Sun, Weiyi, Anna Rumshisky and Ozlem Uzuner. "Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge.", *Journal of the American Medical Informatics Association* Vol. 20, no. 5 (2013a): 806–813

Sun, Weiyi, Anna Rumshisky and Ozlem Uzuner. "Temporal Reasoning over Clinical Text: the State of the Art", *Journal of the American Medical Informatics Association* Vol. 20, no. 5 (2013b): 814–819

Sun, Weiyi, Anna Rumshisky and Ozlem Uzuner. "Normalization of relative and incomplete temporal expressions in clinical narratives". *Journal of the American Medical Informatics Association* Vol. 22, no. 5 (2015): 1001–1008

Tao, Cui, Harold R Solbrig and Christopher G Chute. "CNTRO 2.0: a harmonized semantic web ontology for temporal relation inferencing in clinical narratives". *AMIA Summits Transl Sci Proc* Vol. 2011 (2011): 64–8

Velupillai, Sumithra. "Temporal expressions in swedish medical text–a pilot study". In *Proceedings of BioNLP*, 2014, 88–92

Wenzina, Reinhardt and Katharina Kaiser. "Towards the Application of TimeML in Clinical Guidelines". In *Modellierung im Gesundheitswesen. Tagungsband des Workshops im Rahmen der Modellierung 2014*, ICB-Research Report, 2014a, 37–48

Wenzina, Reinhardt and Katharina Kaiser. "Using TimeML to support the modeling of computerized clinical guidelines.". In *MIE*, 2014b, 8–12

Xiaojia, Zhou, Li Haomin, Lu Xudong and Duan Huilong. "Temporal expression recognition and temporal relationship extraction from Chinese narrative medical records". In *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on*, IEEE, 2011, 1–4

Zhou, Li and George Hripcsak. "Temporal reasoning with medical data—a review with emphasis on medical natural language processing". *Journal of biomedical informatics* Vol. 40, no. 2 (2007): 183–202

Zhou, Li, Genevieve B Melton, Simon Parsons and George Hripcsak. "A Temporal Constraint Structure for Extracting Temporal Information from Clinical Narrative". *Journal of biomedical informatics* Vol. 39, no. 4 (2006): 424–439

Old or new, we repair, adjust and alter (texts)

UDC 811.163.41'322.2: 004.9
DOI 10.18485/infotheca.2019.19.2.3

ABSTRACT: In this paper we present how e-dictionaries and cascades of finite-state transducers, as implemented in Unitex, can be used to solve three text transformation problems: correction of texts after OCR, restoration of diacritics and switching between different language variants.

KEYWORDS: text correction, OCR errors, diacritic restoration, language variants, electronic dictionary, finite-state transducers.

PAPER SUBMITTED: 13 October 2019
PAPER ACCEPTED: 08 December 2019

Cvetana Krstev
University of Belgrade, Faculty of Mathematics
cvetana@matf.bg.ac.rs

Ranka Stanković
University of Belgrade, Faculty of Mining and Geology
ranka.stankovic@rgf.bg.ac.rs

Belgrade, Serbia

1 Text mending – introduction to problems

Text mending is one of the simplest text transformation problems, when compared to speech recognition and generation, text summarization and machine translation. It is also one of the first problems posed to computers that did not involve calculation. Miller and Friedman (1957) wrote about the reconstruction of mutilated texts from the point of view of the information theory in order to calculate the redundancy in English texts. The problem discussed at that time was how many characters can be omitted while still allowing text reconstruction by humans.

Following this line of research the first practical solutions to correcting spelling errors emerged. The idea of a program that corrects spelling errors is reported by Blaire (1960). It consists of weighting the letters to create a four- or five-letter abbreviation for each word – if abbreviations match, the words are assumed to be the same. Blaire claims that "given only a vocabulary of a properly spelled words, the computer can correct most (including unanticipated) misspellings without human assistance." He also claims that programming common orthographic rules for English can not be successful.

Soon, many new approaches were tried in spelling checking programs and most of them included a correction module. Most of them were produced for English (one of the reasons was its notorious "illogical" spelling), and dealt

with errors produced by humans (typographic errors, author's ignorance) or machines during transmission and storage (Peterson, 1980).

Peterson, as many other authors, classifies typographic errors into four groups: insertions, deletions, replacements and reversals. However, texts produced by humans can also be mutilated in other ways: for many reasons, when typing, humans are sometimes compelled to use a degraded instead of a standard alphabet. This results in different types of errors. As usually diacritics are omitted, the process of transforming a text in a standard alphabet is usually called "diacritic restoration", and procedures that perform this task using various approaches were developed for many languages. (Krstev et al., 2018).

Errors produced during machine text input, for instance by Optical Character Recognition (OCR), are of a different type and different solutions were developed for detecting and correcting such errors. As early as in the late 1950s, Bledsoe and Browning (1959) wrote about solving the OCR problem by using a small dictionary with a probability assigned to each word. Despite the considerable improvement of OCR software in subsequent years the problem of OCR error detection and correction is still not considered solved (see, for example, (Kolak and Resnik, 2002)), especially for more "demanding" scripts and languages (Cyrillic, Arabic, etc.).

Transformation from one language variant to another is usually not perceived as an error/correction problem. In his US Patent (2004) Henton describes a voice system that transforms American English utterances for British listeners. The system includes spelling and lexicon normalization; the first is being solved with a set of rules, the second with a list of equivalences. Similar problems are sometimes tackled as translation problems, as for instance in the case of Arabic dialects (Salloum and Habash, 2012) – authors present a rule-based machine translation system that transforms dialectal Arabic to Modern Standard Arabic.

For Serbian, text mending problems were not often reported through scientific channels. Solutions were developed for spelling correction for various platforms (e.g. as an extension for LibreOffice) but very few scientific papers were published with the underlining procedure explained (an exception is (Ostrogonac et al., 2015)). The similar is true for problems of diacritic restoration, OCR errors correction and language variants transformation.

In this paper we present an approach to solving three text mending problems for Serbian: OCR errors, diacritics omission and language variant switching. The common characteristic of these problems is that they occur in a text systematically rather than occasionally. The approach works

at the word level – "incorrect words" or "suspicious words" are recognized by dictionaries (as not belonging to them), a problem specific solution is applied to transform them, and they are corrected if the offered solution is in a dictionary. The problems are usually solved locally, which means that only "incorrect words" and "suspicious words" are considered and rarely their context or more complex structures.

2 Correction of OCR errors

In the process of digitization printed books are scanned and then optical character recognition (OCR) is applied. A text that fully corresponds to the original is rarely obtained since OCR is prone to errors. The quality of the resulting text depends on various factors: the software used, quality of the paper and print of the original text, and its language and alphabet. OCR software today is of good quality compared to its first versions, even when produced for personal rather than professional use,[1] and it is applicable to a large number of languages and scripts, including Serbian Cyrillic. However, OCR of old printed books can still be a challenge due to various reasons: the use of old and non-standard fonts, the deterioration of paper, and if the book that is digitized comes from a library, which is often the case, handwritten additions from numerous users (underlying, redactor's marks, comments, etc.).

OCR errors that occur in scanned texts differ from typing errors, which can be divided into two groups, typographic errors that are the result of mistypes and cognitive errors caused by a misunderstanding of the correct spelling of a word (Kukich, 1992). Errors of the first type can be tackled in terms of keyboard key proximity, while errors of the second type are language dependent and are the result of user's (mis)understanding of relevant orthographic rules. Both types of errors can result in production of either valid (but not intended) or invalid words.

Akin to typing errors, OCR errors are local: one or more characters are erroneously recognized as different characters. However, contrary to typing errors, OCR errors are rarely occasional, as the same type of errors tend to be repeated in one text, while in some other text different type of errors may frequently occur. The erroneously recognized characters can be letters,

[1] For the project presented in this section we used ABBYY FineReader 12 Professional.

punctuation marks and digits. OCR errors can also produce valid words or non-valid words.

We will present the solution for detecting and correcting OCR errors developed for the compilation of the corpus of Serbian novels written and published in the period 1840–1920.[2] The novels selected for this corpus were mainly printed in Cyrillic script (only a few of them were in Latin script). When scanning, the recognition was reduced to only one script, Cyrillic or Latin, in order to avoid confusion between theese two scripts, e.g. Latin *B* with Cyrillic *B* (corresponding to Latin *V*). The majority of books were obtained from the University Library "Svetozar Marković", and the rest from other libraries and private collections.

First of all, the OCR software was set to recognize in novels printed in Cyrillic only the Cyrillic script. As a consequence, occasional phrases written in Latin script (and languages other than Serbian) could not be recognized correctly and had to be retyped manually. However, in this way confusing certain Cyrillic and Latin letters with similar graphical representation was avoided, e.g. a Cyrillic 'а' can be confused for a Latin 'a' (denoting the same letter) or a Cyrillic 'р' can be confused for a Latin 'p' (denoting different letters).

The analysis of frequently occurring OCR errors in Cyrillic texts showed that the following error types predominate:

- Some individual letters are mutually confused, for instance letters 'п', 'и', 'н', letters 'с' and 'е', letters 'с' and 'о' (but not 'е' and 'о');
- A group of letters can be confused for one letter or another group of letters, e.g. two letters 'га' and one letter 'ш'; two letters 'шп' and two letters 'ип';
- Letters can be confused for digits or punctuation or special marks. For instance, a digit '0' and a letter 'О'; a letter 'И' can be recognized as 11 (but not vice versa) or a letter 'љ' can be recognized as 'л>' (but not vice versa).

Our solution addresses only falsely recognized words that result in non-valid words. It follows four steps:

1. The text obtained by OCR is processed using Serbian morphological electronic dictionaries (SMD) (Krstev, 2008);

[2] This corpus is a part of the *European Literary Text Collection corpus* (ElTEC) developed in the scope of the COST action 16204 *Distant Reading for European Literary History* (d-reading).

2. All words not detected by dictionaries are marked as potential results of OCR errors;
3. In these words one or two letters (in general, characters) that were identified as prone to false recognition are replaced with other letter(s) thus generating candidate words. This process is repeated for all letters that were identified as frequent sources of confusion. Moreover, the process is applied to already generated candidates but avoiding the circular replacements (e.g. 'п' → 'и' → 'н' → 'п'). Example: if ппво occurs in a text as an unknown word, then 'п' → 'и' is applied twice which results in a string *ппво*_*ипво*_*пиво*_иво*, after that 'и' → 'н' is applied resulting in *ппво*_*ипво*_*пиво*_*иво* _*нпво*_*пнво*_*ниво*_*инво*_*иво*. After that rule 'н' → 'п' is not applied because it would generate same candidates.
4. Candidates are accepted if they represent words in SMD; others are rejected. For the above example, candidates пиво 'beer' and ниво 'level' would be accepted.

A few examples of the application of this procedure are given in Table 1.

Special attention is paid to hyphenated words. A hyphen in a Serbian OCR text can signify a word hyphenated at the end of the line or a hyphen in a multi-word. Our procedure first eliminates the hyphen and generates candidates by replacements, as illustrated by Table 1. For instance, in the case of Љу-бомнр, the hyphen is eliminated, and several candidates are produced: *_*Љубомпр*_***Љубомир***_*Љубомнр, one of which is accepted (the masculine first name *Ljubomir*). However, if this does not produce any valid word then the hyphen is retained and various replacements are applied to the component of the multi-word that is not a valid simple word (or to both of them). For instance, there are two incorrect letters in Нбрахпм-Хасан. With the elimination of a hyphen no valid candidates can be produced: *_*НбрахнмХаеапа*_*НбрахнмХасапа*_ *НбрахимХаеапа*_*НбрахимХасапа*_*НбрахнмХаеана*... If the hyphen remains, then corrections are attempted only for Нбрахпм since Хасан is a valid word (*Hasan*, a masculine first name): *_*Нбрахнм*_*Нбрахим*_*Нбрахпм*_... *Ибрахнм*_*Ибрахпм*_***Ибрахим***... One of the candidates is accepted – *Ibrahim*, a masculine first name.

The actual replacements are performed by finite-state transducers (FST) implemented in Unitex.[3] A separate FST is written for each replacement

[3] UnitexGramLab, a lexicon-based corpus processing suite.

a non-valid word	**candidates**	**the replacement in text**
One error – one candidate		
кнша	*_*киша*_*кшиа*_*кпша*_*кнша	кнша ⇒ +киша+
Two errors – one candidate		
прекннутп	*_*прскпиуги*_*прскпиути*_*прекпиуги*_...*прекинуги*_*прекинути*_*прскниуги*...*_*ирекииути*_*ирскннугн*_*ирскннутн*...	прекннутп ⇒ +прекинути+
One error – two candidates		
погрешпо	*_*нотрснио*_*нотрсшно*_*ногрснио*_*погренио*_*погрешно*_*потрсшио*_*иотрешио*_*потрешио*_*погрешио*_*потрснио*_*потрсшно*...	погрешпо ⇒ +погрешно+погрешио+
One error – no candidates		
срдски	*_*срдскп*_*срдскн*_*ердски*_*ердеки*_*ердскп*_*срлски	***срдски...

Table 1. All candidates are separated with an underscore and delimited with asterisks. The valid replacements in a corrected text are surrounded with plus signs, words that could not be corrected are marked with asterisks. The incorrect letters are given in red, the accepted candidates are in blue.

pair. These FSTs are, however, very simple since they only give characters to be incorrect and character(s) that can replace them (see the top FST in Table 2). This simple FST invokes another generic FST which corresponds to the type of replacement (see the bottom FST in Table 2), and there are just a few such FSTs. The FST presented at the bottom in Table 2) works for the case when two letters should be replaced with one or more letters. The upper path in this FST works like this: the beginning of the word that needs to be corrected (words preceded with `*_*`) is assigned to variable `$1$`, while the last two letters are assigned to variable `$3$`. If the value of variable `$3$` is equal to letters that should be replaced (`$3.EQUAL=ul$`) then the output is produced that indicates the path that produced it (`*%1*`), followed by

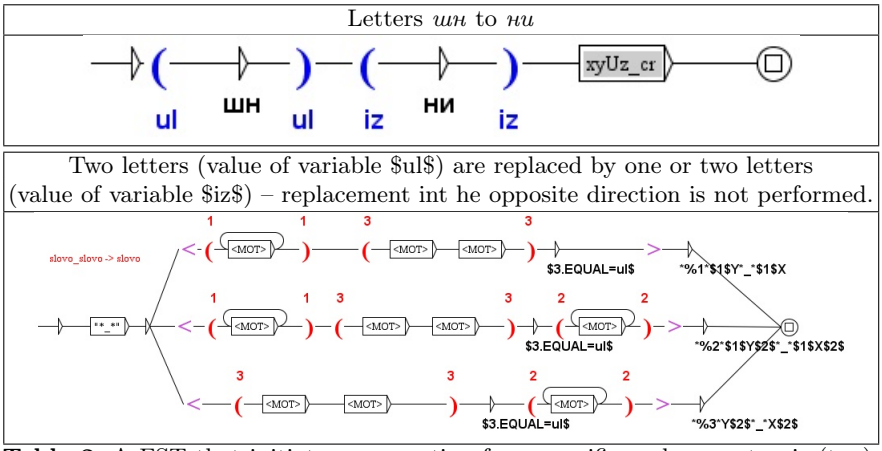

Table 2. A FST that initiates a correction for a specific replacement pair (top). A generic FST that performs a certain type of replacement (bottom).

two words: the original, input word (`$1$Y`) and the corrected, output word (`$1$X`).[4]

The application of these FSTs always leaves in a processed texts the original word to which a candidate is added. Also, FSTs for each replacement pair are iterated which enables the same corrections in one word: for instance, in сиежии three letters 'и' occur, first and second are incorrect and should both be replaced by 'н' yielding a valid word снежни – *snežni* 'snowy'.

Just a few general FSTs (bottom of Figure 2) were developed for a few types of corrections – one or two character replacement, in one or both directions. Specific FSTs (top of same figure) were developed many – for various confusion pairs that can occur in an OCR text. Moreover, as each texts can bring its own problems, new specific FSTs can be easily produced and added to processing.

[4] Both of these words are modified in order to suppress the infinite modifications of same characters – letters not used in Serbian Latin alphabet Y and X replace original letters and their replacement. After all modifications of the same pair are done, the original characters are restored.

Table 3 illustrates the whole process, from a rough OCR output to a clean text.[5]

A text after OCR
- Е. *нпје него број! Тебе ће неко *еад *шптатн шта ти хоћеш, а *пгга нећеш! Него. кажи ти мени. је ли теби *бнла позната моја наредба, којом се забрањује тумарање по турским кућама? — *Нпје. — Ја где си ти *бно за ово месец дана — У *болнници.
A text after automatic correction
— Е. +++пије+++није+++ него број! Тебе ће неко +++сад+++ +++ништи+++пишти+++питати+++ шта ти хоћеш, а ***пгга нећеш! Него. кажи ти мени. је ли теби +++била+++ позната моја наредба, којом се забрањује тумарање по турским кућама? — +++Није+++. — Ја где си ти +++био+++ за ово месец дана — У +++болници+++.
A text after reading and correcting
— Е, није него број! Тебе ће неко сад питати шта ти хоћеш, а шта нећеш! Него, кажи ти мени, је ли теби била позната моја наредба, којом се забрањује тумарање по турским кућама? — Није. — Ја где си ти био за ово месец дана — У болници.

Table 3. Three phases of a text correction: 1) Marking of incorrect/unrecognized words by e-dictionaries; (2) Automatic correction by offering possible candidates; (3) Elimination of incorrect candidates and correction of remaining errors by a reader.

3 Diacritic restoration

The problem of diacritic restoration in Serbian occurs only for texts that are using Latin script. Diacritic omission happens when the Serbian Latin alphabet is reduced to the 26 letter Latin alphabet that can be accommodated by ASCII character set, and affects five letters: *š*, *đ*, *č*, *ć* and *ž*. As a result, some letters or letter groups become ambiguous:

- *c* – can stand for *č* and *ć* (and *c*);
- *z* – can stand for *ž* (and *z*);

[5] The example text is from the novel by Lazar Komarčić "One devastated mind" (Један разорен ум).

– *s* – can stand for *š* (and *s*);
– *dj* – can stand for *đ* (and *dj*);
– *dz* – can stand for *dž* (and *dz*).

When compared with the problem of OCR errors, it can be observed that the repertoire of these "errors"[6] as well as their possible corrections is limited and is known in advance. Namely, candidates for correction are only those that are represented in dictionaries. Similar to the OCR errors – a word that contains letters *c*, *s*, *z* or a digraph *dj* can be both correct and incorrect because a diacritic is missing, e.g. *kuca* can be correct (small dog) or a diacritic can be missing *kuća* (house). However, contrary to the OCR solution presented in the previous section, such words are treated as potentially incorrect and candidates for correction are offered, if they exist. Words that contain neither letters *c*, *z*, *s* nor digraphs *dj* and *dz* will be treated as correct and will not be subject to any correction, e.g. *voda* (water). The preposition of our procedure is that text contains no diacritics.

For each potentially incorrect word in the text the procedure will offer a list of all possible candidates:

– This list may contain the original word because it exists in dictionaries: *kupaca* ⇒ *kupača* (*kupač* 'bather'), *kupaća* (kupaći 'bathing'), *kupaca* (kupac 'buyer');
– It need not contain the original word (because it is not in dictionaries): *jezice* ⇒ *ježiće* (*ježiti se* 'to bristle' and *ježić* 'diminutive of hedgehog'), *jeziče* (*jezik* 'tongue'), *ježice* (*ježica* 'female hedgehog');
– If the list contains only the original word, and no other candidates, it will be accepted as correct.

For all potentially incorrect words procedure ranks its candidates according to the frequency of their occurrence in the Corpus of Contemporary Serbian (SrpKor).[7] For instance, the candidates for *kupaca* occur with following frequencies: *kupača* → 207, *kupaća* → 6, *kupaca* → 2202.

In order to be able to offer ranked candidates in diacritic restoration a special dictionary was produced from the standard Serbian Morphological Dictionaries, in which the entries for the forms *kupaca*, *kupača* and *kupaća* are as follows:

```
kupaca,kupac.N+Hum:mp2v
```

[6] We will call them "errors" although they are done intentionally.
[7] Corpus of Contemporary Serbian

```
kupača,kupač.N+Hum:mp2v:ms2v:ms4v
kupaća,kupaći.A:aefs1g:aefs5g:aenp1g:aenp4g:akms2g...
```

The construction of the dictionary to be used in diacritic restoration (SMD_DR) from SMD is performed by the following steps:

- All word forms containing at least one diacritic and all word forms containing at least one letter *c*, *s*, or *z* or a digraph *dj* or *dz* are extracted from SMD;
- Diacritics are removed from each word form that contains them, while the original word form is saved as the value of a new marker `+CR=`;
- All information that is not needed for this procedure is deleted (lemma, POS, syntactic and semantic markers, etc.);
- Information about the frequency of the original word form in SrpKor is added;
- Information for same word forms is merged.

This process for the previously examples is illustrated in Table 4.

`kupaca,kupac.N+Hum:mp2v`	`kupaca,.X+CR=kupaca(21)`	⇒ SMD_DR
`kupača,kupač.N+Hum:mp2v`	`kupaca,.X+CR=kupača(2)`	
`kupaća,kupaći.A:aefs1g`	`kupaca,.X+CR=kupaća(1)`	
`kupaca,.X+CR=kupaca(21)_kupača(2)_kupaća(1)`		

Table 4. Production of an entry in SMD_DR

Relative frequencies in SrpKor were calculated and assigned to each candidate for correction in order to facilitate calculation and usage. For instance, relative frequency 1 was assigned to tokens that had in the corpus absolute frequency $[1, 10]$.More about this can be read in (Krstev et al., 2018).

The produced SMD_DR has almost one million entries (word forms that are candidates for correction), of which 95.1% have only one candidate for correction, 4.4% have two candidates, and remaining the 0.5% have 3 or more candidates. However, in a number of cases when more than one candidate exists, all candidates occur with the comparable frequency which makes it difficult to choose the right one despite the ranking. Some of difficult cases are:

– `čašu(10)\_času(28)\_ćasu(1)` (candidates are forms for *čaša* 'glass', *čas* 'hour/moment', *ćasa* 'bowl');
– `reci(24)\_reči(261)\_reći(145)` (candidates are forms for *reka* 'river', *reč* 'word', *redak* 'line' and *reći* 'to say');
– `veće(155)\_veče(34)\_vece(1)` (candidates are forms of *veće* 'council', *veći* 'bigger', *veče* 'evening' and *vece* 'WC').

In order to reduce the number of multiple candidates, procedure uses some additional resources.

– A list of 30 most frequent trigrams obtained from SrpKor in which at least one word would contain letters *c*, *s*, *z* or a digraphs *dj* if diacritics were removed; for instance, *zbog toga sto* 'because of that' ⇒ *zbog toga što*, thus avoiding multiple candidates for *sto* (*sto* 'table/hundred' and *što*, a functional word that can be an adverb, a pronoun or a conjunction);
– A list of 50 most frequent bigrams obtained from the SrpKor in which at least one word would contain letters *c*, *s*, *z* or a digraph *dj* if diacritics were removed; for instance, *znaci da* 'meaning that' ⇒ *znači da*, thus avoiding multiple candidates for *znaci* (a form of *znak* 'sign' and *značiti* 'to mean');
– A dictionary of multi-word units (MWU) (nouns, adjectives, adverbs, pronouns, conjunctions and interjections) obtained from a dictionary of more than 18,000 MWU lemmas; for instance, *Dobro vece* ⇒ *Dobro veče* 'Good evening' (avoiding multiple candidates for *vece*) or *ukrstene reci* ⇒ *ukrštene reči* 'cross-words' (avoiding multiple candidates for *reci*, but also resolving the preceding adjective *ukršten*). These dictionary contains all inflective forms,so, for instance, *ukrstenih reci* (the ginitive form) would be corrected as well – *ukrštenih reći-*.

After the application of the diacritic restoration procedure, a new version of the text is obtained which contains, for each word form in the original text that contains letters *c*, *s*, *z* or digraphs *dj*, *dz*, a list of zero[8] or more candidates obtained from the dictionary SMD_DR, or one candidate obtained from lists of trigrams or bigrams, or a dictionary of MWUs for a sequence of words. The result of the application of the procedure to a sample text is given in Table 5.[9]

[8] This list is empty if a word from neither with diacritics/digraphs nor without them exists in SMD.

[9] The excerpt is taken from the novel "Komo" by Srđan Veljarević.

the source text	the text with suggested corrections
KGB mu je ponudio da saradjuje s njima, i da ce mu onda knjige biti objavljivane. Brodski je odbio. I nije mogao da objavljuje. Posle nekog vremena predlozili su mu da napusti zemlju, i da ce tako biti najbolje, za njega i za drzavu. Brodski je seo u avion za Bec. Poneo je pisacu masinu, nesto odece, zbirku poezije Dzona Dona, i flasu votke, poklon za pesnika Vinstana Odna, koji ga je docekao na beckom aerodromu.	KGB mu je ponudio da *5a(sarađuje(25)) *2(i da će) mu onda knjige biti objavljivane. V_*5b(Brodski(2)_brodski(3)) je odbio. I nije mogao da objavljuje. V_*5b(pošle(1)_posle(1422)) nekog vremena *5a(predložili(12)) su mu da napusti *5b(zemlju(137)_žemlju(1)), *2(i da će) tako biti najbolje, za njega i za *5a(državu(81)). V_*5b(Brodski(2)_brodski(3)) je seo u avion za V_*5a(Beč(22)). Poneo je *4(pisaću mašinu(0)), *5a(nešto(515)) *5a(odeće(13)), zbirku poezije *5a(Džona(17)) Dona, i *5a(flašu(4)) votke, poklon za pesnika Vinstana Odna, koji ga je *5a(dočekao(12)) na *5a(bečkom(5)) aerodromu.

Table 5. The output of the procedure for diacritic restoration

The output presented in Table 5 is only an intermediate result which has to be further transformed into a corrected text. Information in this intermediate output – a list of candidates, their relative frequency in the reference corpus, an indication of the type of correction (e.g. *4 indicates the dictionary of MWUs, *2 the list of trigrams, *5 the dictionary SMD_DR, etc.) – can be used in the cleaning phase. The cleaning can be very simple by accepting all suggestions – when there is more than one candidate choose one with the higher frequency – or it can be more sophisticated. In the later case, when there is more than one candidate the procedure could choose one if it has a significantly higher relative frequency, e.g. at least 10 or 100 times. In the example given, for the two cases of multiple candidates: *5b(pošle(1)_posle(1422)) and *5b(zemlju(137)_žemlju(1)),[10] both criteria would chose *posle* and *zemlju*, respectively, which would be the right choice in both cases. The procedure can also look in the broader context and apply some decision rules. For instance, for a word form *celu* there are three candidates: čelu(95), ćelu(1), celu(44) (forms of *čelo* 'forehead/cello', *ćela* 'bold

[10] Actually, there is one more case of multiple candidates, V_*5b(Brodski(2)_brodski(3)), stemming from two different entries in SMD: *Brodski*, a surname, and *brodski* 'like a boat'; however, the correction result is the same – no correction, a word form *Brodski* remains.

head', *ceo* 'whole', respectively). If this word form appears after the preposition *za* that demands the genitive, the accusative, or the instrumental case, then *čelu* would be discarded (being the dative or the locative form of *čelo*) despite having the highest relative frequency. If some multiple candidates cannot be resolved by rules nor by frequencies, they remain in the text for the user to chose the right one.

4 Switching between two Serbian pronunciation variants

In Serbian, two standard variants of pronunciation are in use, Ekavian and Ijekavian. They differ in the reflection of the old Proto-Slavic phoneme (*jat*): in the Ekavian variant it is replaced predominantly by *e*, while in the Ijekavian variant it is replaced by syllables *ije/je*, and sometimes *i*. A Serbian text is usually written in one of these variants.

Sometimes it is desirable to transform text written in one pronunciation into another. Although this is not an issue related to errors, the problem is similar. Namely:

- When transforming an Ekavian text to Ijekavian, for each word containing the letter *e* it must be determined whether it is a reflection of the phoneme *jat* and that it should therefore be replaced by an Ijekavian variant containing *ije/je/i*;
- When transforming an Ijekavian text to Ekavian, for each word containing *ije/je/i* it must be determined whether it is a reflection of the phoneme *jat* and that it should therefore be replaced by an Ekavian variant containing *e*.[11]

The problem, though different in nature, has similarities with problems of OCR error correction and diacritic restoration:

- Like in the case of diacritic omission "errors" are limited to a small number of letters and/or syllables, which implies that a dictionary solution might be appropriate;

[11] The problem is further complicated by morphological alternations, e.g. *nežan* vs. *nježan* 'tender' and *leto* vs. *ljeto* 'summer'. In the case of Ijekavian *nj* and *lj* are not consonant groups but dighraphs whose corresponding Cyrillic letters are *њ* and *љ*. Although our procedure also deals with such cases they are not explained here for reasons of simplicity.

– Similar to problems of OCR error correction and diacritic restoration, an *e* in an Ekavian text word can be a reflection of *jat* (*reka* $\Longleftrightarrow$ *rijeka* 'river') or not (*zeka* 'bunny'); a syllables *ije/je* in an Ijekavian text word can be a reflection of *jat* (*snijeg* $\Longleftrightarrow$ *sneg 'snow'* and *mjesec* $\Longleftrightarrow$ *mesec* 'month/moon') or not (*sujeta* 'vanity' and *prijem* 'reception').

For the problem of switching between two pronunciations two systems were developed: the first one transforms an Ekavian text to its Ijekavian version, the other transforms an Ijekavian text to its Ekavian version. These systems work in the similar way as the system for diacritic restoration, and each of them uses its own dictionary for transformation.

Ekavian	Ijekavian	translation
`reka,N612+Ek`	`rijeka,N612+Ijk`	'river'
`zeka,N741+Zool`		'rabbit'
`sneg,N291+Ek`	`snijeg,N291+Ijk`	'snow'
`prijem,N1`		'reception'
`mesec,N9+Ek`	`mjesec,N9+Ijk`	'moon/month'
`sujeta,N600`		'vanity'

Table 6. Pronunciation markers in SMD

Dictionaries for variant transformation, as in the case of the dictionary for diacritic restoration SRP_DR were obtained from the standard SMD. Pronunciation variants in the standard SMD are marked but not connected (this is explained in more details in (Lazić and Škorić, 2019) in the same issue). The marker `+Ek` denotes Ekavian specific lemmas, and the marker `+Ijk` Ijekavian specific lemmas, while entries for lemmas that do not contain a reflection of *jat* do not have a variant marker, as illustrated in Table 6.

As explained in (Lazić and Škorić, 2019), specific rules were developed for linking corresponding Ekavian and Ijekavian entries. This enabled the production of two specific dictionaries: Ijk2Ek for transforming an Ekavian text to Ijekavian, and Ek2Ijk for the transformation in the other direction. Lemmas that are same in both pronunciations are not represented in these dictionaries because for them no transformation is needed - as in the case of the dictionary for diacritic restoration in which entries that do not contain

c, *s*, *z* and *đ* are omitted. Examples of entries in these two dictionaries are represented in Table 7.

Ijk2Ek	Ek2Ijk
`rijeka,.X+EK=reka(865)`	`reka,.X+IJK=rijeka(235)`
`rijekama,.X+EK=rekama(121)`	`rekama,.X+IJK=rijekama(34)`
`snijeg,.X+EK=sneg(473)`	`sneg,.X+IJK=snijeg(129)`
`snijega,.X+EK=snega(298)`	`snega,.X+IJK=snijega(97)`
`mjesec,.X+EK=mesec(2282)`	`mesec,.X+IJK=mjesec(644)`
`mjeseca,.X+EK=meseca(3922)`	`meseca,.X+IJK=mjeseca(3955)`

Table 7. Dictionary entries in dictionaries for transformation form Ijekavian to Ekavian and vice versa.

Frequencies incorporated in them for the selection of candidates are obtained from two different corpora – one containing only texts written in Ekavian pronunciation and the other containing only texts written in Ijekavian pronunciation. In the case of multiple corrections, they are merged in one entry, as in the SRP_DR dictionary. Specific problems may arise with multiple corrections when transforming texts in either direction:

- An Ijekavian form is a homograph of a form that does not contain Ijekavian *ije/je*:
  - `njega,.X+Ijk+EK=nega(56)_njega(5459)` (*njega/nega* 'nursing' vs. *njega* 'him');
  - `bolje,.X+Ijk+EK=bole(42)_bolje(4279)` (*bolje/bole* a form of *boljeti/boleti* '(they) hurt (aorist)' vs. *bolje* 'better').
- An Ekavian form is a homograph of a form that does not contain Ekavian *ije/je*:
  - `beg,.X+IJK=bijeg(78)_beg(15)` (*beg/bijeg* 'runaway' vs. *beg* 'bey');
- An Ekavian form has two or more different Ijekavian forms:
  - `cedila,.X+IJK=cjedila(0)_cijedila(5)` (the genitive singular form of *cjedilo* 'strainer' vs. the active participle of *cijediti* 'to strain');
  - `posede,.X+IJK=posjede(10)_posijede(0)_posijedje(0)_posjedje(0)` (forms of the noun *posjed* 'property' and tree different verbs: *posijedjeti* 'grow white', *posjedjeti* 'sit down', *posjesti* 'assign a seat').

– For an Ekavian form list of candidates more than one Ijekavian form exisr as well as a form that is not affected by different pronunciation:
  • `bega,.X+IJK=bijega(42)_bega(9)_bjega(0)` (*bega/bijega* 'runaway (genitive)' vs. *bega* 'bey (genitive)') vs. *bjega* 'to runaway (aorist)'.

Both systems use the same procedure for detecting words that should be "corrected" and producing lists of candidates for replacement, although, obviously, the resources they use are different. The results produced by two systems on two sample texts are presented in Table 8. In the case of Ijekavian source sample, three word forms had to be replaced – *vjerovatno* 'probably', *izmjene* 'change', *cijena* 'price' – and they were all correctly replaced by one offered solution. In the case of the Ekavian source sample, three different word forms had to be replaced – *posledica* 'consequence', *delu*, a form of *delo* 'work' and *deo* 'part', and *rešenje* 'solution'. Two forms were correctly replaced by one offered solution – *posledica* and *rešenje* – while the third form *delu*, which is homographous in Ekavian, has two different corresponding forms in Ijekavian – *dijelu* and *djelu* – and the system could not decide using rules what would be the right choice. Namely, a phrase *u nekom delu* is correct for both meanings of the form *delu* ('in some part' and 'in some work'). The frequency of use of forms *dijelu* and *djelu* in the Ijekavian corpus was not enough in favor of any of them (the right choice in both cases in the sample is *dijelu*).

5 Implementation and results

All systems for text mending presented in sections 2–4 use similar resources and are built using similar solutions:

– Electronic dictionaries are used to detect words that are candidates for change and to offer possible corrections. They are also used to build special dictionaries for solving some concrete problems (diacritic restoration, language variant switching).
– Detecting words that are candidates for change as well as the production of lists of candidates for replacement is done by finite-state transducers implemented in Unitex software (Paumier et al., 2016).
– All presented systems consist of two independent parts, both of which are implemented as cascades of finite-states transducers – in these cascades each FST works on a texts produced by a FST that directly precedes it

Ijekavian ⇒ Ekavian	
the source text	the output text
"Na Medicinskom fakultetu u Foči ove godine u planu je multidisciplinarni program pa će moći da konkurišu i studenti sa srodnih fakulteta, a što se tiče ostalih, vjerovatno bi u narednom periodu moglo doći do izmjene Pravilnika po tom pitanju. Zasad ostaje ovako", kaže on. Cijena školarine za godinu kreće se od 2.500 KM pa naviše.	"Na Medicinskom fakultetu u Foči ove godine u planu je multidisciplinarni program pa će moći da konkurišu i studenti sa srodnih fakulteta, a što se tiče ostalih, verovatno bi u narednom periodu moglo doći do izmene Pravilnika po tom pitanju. Zasad ostaje ovako", kaže on. Cena školarine za godinu kreće se od 2.500 KM pa naviše.
Ekavian ⇒ Ijekavian	
the source text	the output text
„Ne razmatramo mogućnost da Tanjug ponovo počne radi kao vladin medij, nego ispitujemo sve okolnosti koje će dovesti do nekih od mogućih posledica, a to je da Tanjug bude vladin medij u nekom delu, da Tanjug uopšte ne bude vladin medij u bilo kom delu i neko treće rešenje koje je između ta dva", kaže ministar.	„Ne razmatramo mogućnost da Tanjug ponovo počne radi kao vladin medij, nego ispitujemo sve okolnosti koje će dovesti do nekih od mogućih posljedica, a to je da Tanjug bude vladin medij u nekom *(dijelu(751)_djelu(59)), da Tanjug uopšte ne bude vladin medij u bilo kom *(dijelu(751)_djelu(59)) i neko treće rješenje koje je između ta dva", kaže ministar.

Table 8. The output of two procedures for switching from Ijekavian to Ekavian and vice versa.

and produces a new text for a FST that follows. These cascades are also implemented in Unitex (Friburger and Maurel, 2004). The first cascade produces an intermediate result with lists of all possible candidates, as illustrated in Table 5 in Section 3. The second cascade eliminates some or all multiple candidates. These two cascades are independent in all systems, which makes it easy to produce the second cascade as strict, relaxed or somewhere in between.

A similar approach that also relies on electronic dictionaries and FSTs implemented in Unitex was used for vowel restoration in Arabic (Neme and

Paumier, 2019). The approach presented in this paper is specific since it offers a solution for solving three different tasks.

Not all of the systems presented were fully evaluated. The system for OCR correction is difficult to evaluate since each text (especially in the case of relatively old books) poses different problems and different lists of solutions may be offered. However, results after correcting some thirty novels show that after applying the system for OCR error correction the number of errors (precisely, unknown words) may be reduced by 5% to up to almost 90%, which depends, naturally, on the initial number of errors

A thorough evaluation results for the system for diacritic restoration was presented in details in (Krstev et al., 2018). The authors showed that on the average the precision of the system is $P = 98.93\%$, the recall $R = 94.94\%$ and $F_1 = 96.90\%$ when calculated on all occurrences (tokens) of words that were candidates for correction ("suspicious" words).

The system for transforming texts from Ekavian to Ijekavian pronunciation and vice versa was not evaluated since it is not fully developed yet. Namely, the Ijekavian corpus that was used to calculate frequencies needs to be enlarged and enriched with more versatile texts.

6 Future Work

In this paper we presented three operational systems for three text mending tasks for Serbian: correction of OCR errors, diacritic restoration and switching between two pronunciations. Although these systems were already successfully used for various tasks (for corrections of OCR errors see (Jaćimović, 2019) and for diacritic restoration see (Petković, 2019)) there is still much to be done. The treatment of unknown words in processed texts as well as the elimination of multiple candidates needs to be further developed. Options of using a hybrid approach that would merge a dictionary with machine-learning will be explored. Finally, a user-friendly interface that will enable the use of these systems on the Web is already under development.

Acknowledgment

This research was partly supported by the Ministry of Education, Science and Technological Development through projects 138006 and III47003.

References

Blair, Charles R. "A program for correcting spelling errors". *Information and Control* Vol. 3, no. 1 (1960): 60–67

Bledsoe, W. W. and I. Browning. "Pattern Recognition and Reading by Machine". In *Papers Presented at the December 1-3, 1959, Eastern Joint IRE-AIEE-ACM Computer Conference*, IRE-AIEE-ACM '59 (Eastern). New York, NY, USA: ACM, 1959, 225–232. http://doi.acm.org/10.1145/1460299.1460326

Friburger, Nathalie and Denis Maurel. "Finite-state transducer cascades to extract named entities in texts". *Theoretical Computer Science* Vol. 313, no. 1 (2004): 93–104

Henton, Caroline G. "Automated transformation from American English to British English", 2004, uS Patent 6,738,738

Jaćimović, Jelena. "Textometric methods and the TXM platform for corpus analysis and visual presentation". *Infotheca – Journal for Digital Humanities* Vol. 19, no. 1 (2019): 30–54. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2019.19.1.2_en

Kolak, Okan and Philip Resnik. "OCR error correction using a noisy channel model". In *Proceedings of the second international conference on Human Language Technology Research*, 257–262. Morgan Kaufmann Publishers Inc., 2002

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

Krstev, Cvetana, Ranka Stanković and Duško Vitas. "Knowledge and Rule-Based Diacritic Restoration in Serbian". In *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, Koeva, Svetla. Sofia, Bulgaria: Institute for Bulgarian Language "Prof. Lyubomir Andreychin", 41–51. Bulgarian Academy of Sciences, 2018

Kukich, Karen. "Techniques for automatically correcting words in text", *Acm Computing Surveys (CSUR)* Vol. 24, no. 4 (1992): 377–439

Lazić, Biljana and Mihailo Škorić. "From DELA based Dictionary to Leximirka Lexical DataBase". *Infotheca – Journal for Digital Humanities* Vol. 19, no. 2 (2019): 00–00, https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2019.19.2.4_en

Miller, George A and Elizabeth A Friedman. "The reconstruction of mutilated English texts". *Information and Control* Vol. 1, no. 1 (1957): 38–55

Neme, Alexis Amid and Sébastien Paumier. "Restoring Arabic vowels through omission-tolerant dictionary lookup". *Language Resources and Evaluation* (2019): 1–65

Ostrogonac, Stevan, Branislav Popović and Robert Mak. "The Use of Statistical Language Models for Grammar and Semantic Error Handling in Spell Checking Applications for Serbian". In *12 th International Conference on Electronics, Telecommunications, Automation and Informatics, ETAI*, 2015

Paumier, Sébastien, Sebastian Nagel Marschner and Johannes Stiehler. "UNITEX 3.1 User Manual". *Université Paris-Est Marne-la-Vallée.* (2016)

Peterson, James L. "Computer Programs for Detecting and Correcting Spelling Errors". *Commun. ACM* Vol. 23, no. 12 (1980): 676–687. http://doi.acm.org/10.1145/359038.359041

Petković, Ljudmila. "Creation and Analysis of the Yugoslav Rock Song Lyrics Corpus from 1967 to 2003". *Infotheca – Journal for Digital Humanities* Vol. 19, no. 1 (2019): 5–29. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2019.19.1.1_en

Salloum, Wael and Nizar Habash. "Elissa: A dialectal to standard Arabic machine translation system". In *Proceedings of COLING 2012: Demonstration Papers*, 2012, 385–392

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

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) (**?**). 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[U+2010]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

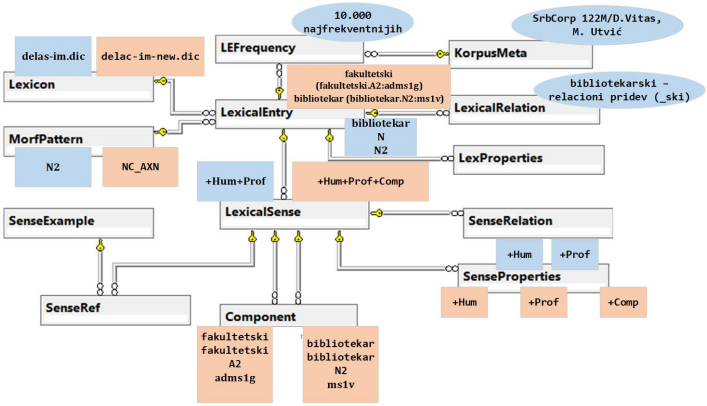

Figure 1. Leximirka's lexicographic database model

.

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.

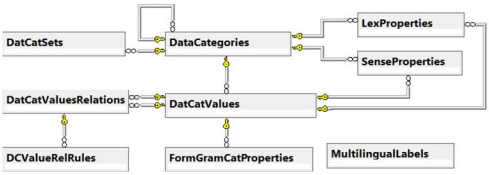

Figure 2. A lexicographic database model for data category information

.

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: brijeg+Ijk→ breg, bezbjednost+Ijk→ bezbednost, sljedeći+Ijk→ sledeći .

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: $sterilizovati, V + DER = ZovatiSati \rightarrow sterilisati$. 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: $sterilizovan, A + DER = ZovatiSati \rightarrow sterilisan; sterilizovanje, N + DER = ZovatiSati \rightarrow sterilisanje$.

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: filozofije+DER=ZS and filosofije+DER=SZ or string omission: halva+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-

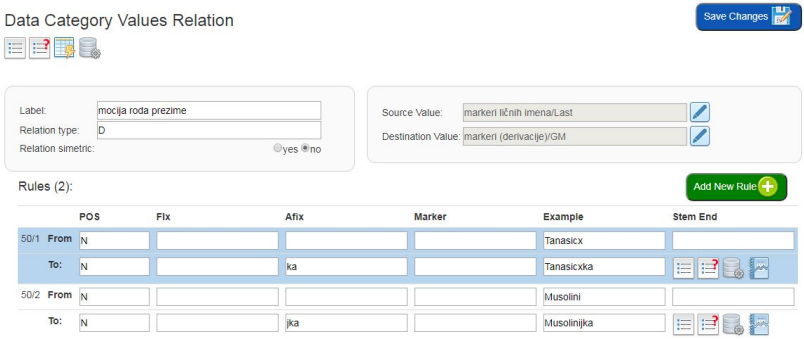

Figure 3. Data Category Values Relation panel

.

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.

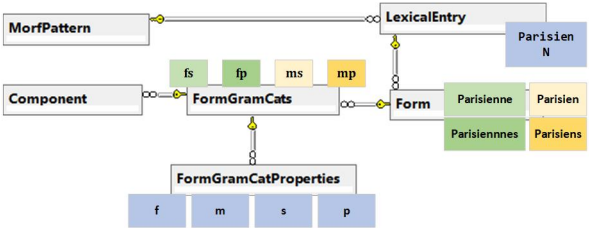

Figure 4. Flective forms in Leximirka database model

.

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

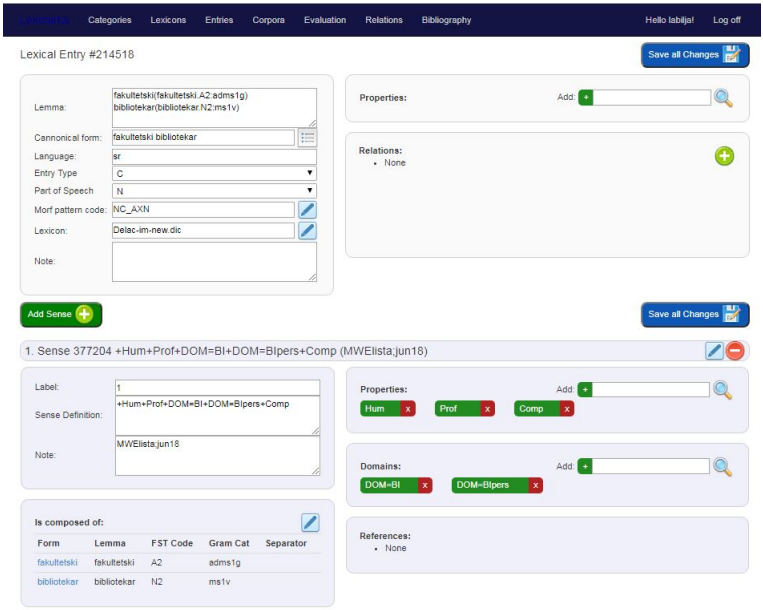

Figure 5. The view of a lexical entry "*bibliotekar*" for logged-in users

.

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).

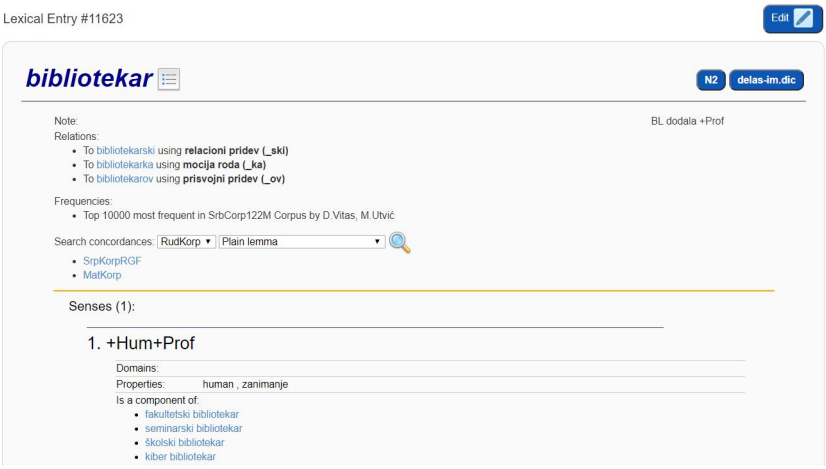

Figure 6. Editing panel of a lexical entry "*fakultetski bibliotekar*"

.

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.

Vebran Web Services for Corpus Query Expansion

UDC 004.738.52:811.163.4'373
DOI 10.18485/infotheca.2019.19.2.5

ABSTRACT: This paper discusses the development of the Vebran web services and their application to corpus search improvements. The Vebran web services are used to consult external lexical resources for Serbian (mainly electronic morphological dictionaries and Serbian Wordnet) and expand user queries to retrieve more relevant results from Serbian corpora.

KEYWORDS: corpus search, web service, Serbian lexical resources, query expansion.

PAPER SUBMITTED: 18 October 2019
PAPER ACCEPTED: 13 December 2019

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology

Miloš Utvić
milos.utvic@fil.bg.ac.rs
University of Belgrade
Faculty of Philology

Belgrade, Serbia

1 Introduction

It is generally known that users typically formulate very short instead of long and complex queries. The lemmatization of corpora enables easy retrieval of all inflected forms for a lemma, but one can not be sure in the accuracy of automatic lemmatization.

In order to achieve retrieval effectiveness as regards these short queries, we need special techniques, because a straightforward keyword match may not always be adequate. The main objectives are 1) to upgrade the existing web interfaces for searching through language resources and 2) to enable querying language resources supported with available lexical resources.

The language resources to be searched are various digital libraries and corpora, but in this paper, we will focus on corpus case study. The query expansion will rely on different lexical resources to support the search: 1) morphological electronic dictionaries; 2) WordNets and 3) terminological databases. The new web services named Vebran are developed as an upgrade of the previous service wsQueryExpand (Stanković, 2009; Stanković et al., 2012). The Vebran is implemented as Restful API with several new functions, but the biggest improvement is a set of functions specially prepared to support query expansion based on lexical resources.

Sections 2 and 3 describe language resources for Serbian, corpora that we can search and lexical resources that Natural Language Processing (NLP) applications can consult. Vebran web services and their usage of lexical resources are discussed in Section 4. Morphological and semantic corpus query expansion, based on Vebran web services, is described in detail in subsections 4.2 and 4.3 respectively. Finally, in Section 5 we conclude with some remarks concerning the future work.

2 Corpora

Numerous corpora are developed within research activities of Human Language Technology (HLT) Group at the University of Belgrade and the Language Resources and Technologies Society (JeRTeh):

- monolingual general corpora: Corpus of Contemporary Serbian (versions SrpKor2003 and SrpKor2013)[1] and its subset SrpLemKor[2];
- SrpEngKor[3], aligned English-Serbian corpus including subcorpus SELFEH (Serbian-English Law Finance Education and Health) with documents on finance, health, law and education;
- SrpFranKor[4], aligned French-Serbian corpus;
- SrpNemKor[5], aligned German-Serbian corpus;
- RudKor[6], a specialized monolingual corpus of texts from the mining domain, etc.

The query expansion will be demonstrated using examples in two Serbian corpora: SrpKor2013 (cf. 2.1) and RudKor (cf. 2.2). Both corpora are morphologically tagged in the sense that each token is associated with the information about the corresponding part of speech and lemma as described in (Утвић, 2011, 2014). The TreeTagger software tool (Schmid, 1997, 1999) was used for automatic morphological annotation of both corpora. The TreeTagger language parameter file for Serbian was created as a derivative of a system of Serbian Morphological electronic Dictionaries (SMD, cf. 3.1), authored by Cvetana Krstev and Duško Vitas (Krstev, 2008).

[1] http://www.korpus.matf.bg.ac.rs
[2] http://www.korpus.matf.bg.ac.rs/SrpLemKor/
[3] http://www.korpus.matf.bg.ac.rs/SrpEngKor
[4] http://www.korpus.matf.bg.ac.rs/SrpFranKor
[5] http://jerteh.rs/biblisha/
[6] http://147.91.181.179/ cqp/cqpweb

2.1 SrpKor2013

The Corpus of Contemporary Serbian (SrpKor) was established in 2002, at the initiative of Professor Ljubomir Popović, with the aim of enabling researchers to consult the chosen collection of Serbian texts via the Internet (Vitas and Krstev, 2012). The first version of SrpKor, SrpKor2003, has not been morphologically annotated.

SrpKor2013 is the current version of SrpKor (Utvić, 2014), used as a reference and general purpose corpus containing over 122 million corpus words. It includes literary texts of Serbian writers in the XX and XXI centuries, as well as scientific and popular science texts from different domains (natural and social sciences), administrative and general texts. The general texts represent articles from the daily newspapers "Politika","Večernje Novosti", "Danas", texts from magazines "Danica", "Ebit", "Ekonomist", "Glasnik", "NIN", "Ilustrovana politika", "Kalibar", "Moje srce", "Mostovi", "Pravoslavlje", "Svet", "Teološki pogledi", "Trn", "Viva", "Republika", texts from the internet portal "Peščanik". Some of the texts are translations, most of which are literary texts, while a smaller part are translations of general texts. Apart from being morphologically annotated, the corpus texts are provided with corresponding bibliographic description, information concerning the functional style to which the text belongs, as well as an indicator whether corpus text is written in Serbian or represents a translation from another language.

The SrpKor2013 is not structurally annotated, although some or all levels of the text structure (section, title, paragraph, sentence) are annotated in some particular corpus texts, especially those which are part of aligned corpora.

The SrpKor2013 corpus is used by more than 700 users, mostly Slavists.

2.2 RudKor

Systematic collection and preparation of texts from the mining domain started with English-Serbian alignment of articles in a bilingual journal "Podzemni radovi", followed by mining projects, law regulations, PhD theses and textbooks from the mining domain. Texts are gathered and organized in the ROmeka@RGF digital library (Tomašević et al., 2018). The RudKor corpus originated from ROmeka@RGF digital library to enable various linguistic and terminological research, including extraction of terms and other tasks in the field of knowledge engineering (Утвић et al., 2018). The RudKor contains 344 different texts with a total size of 5.4 million words.

Mining terminology is introduced in the system of Serbian Morphological Dictionaries (cf 3.1). In order to allow the extraction of specific concepts and relations between concepts by creating lexical masks, new semantic markers relevant to the field of mining have been integrated (Обрадовић et al., 2017).

2.3 Corpora tools

Three different systems for diverse types of usage scenarios are used in this research:

- Unitex (Paumier, 2016; Krstev, 2008);
- Open Corpus Workbench (OCWB) (Evert and The OCWB Development Team, 2019) and web-based graphical user interface CQPweb (Hardie, 2012);
- NoSketch Engine (Rychlý, 2007).

Unitex[7] is open source software for an analysis of textual data, corpus processor with user-friendly interface, language resources distributed out-of-the-box and set of functions that can be used from other software systems. The Unitex NLP engine is based on automata-oriented technology, allowing users to 1) Compile rules and dictionaries as finite-state machines; 2) Use variables instanced with a part of the text or with any characters; 3) Use regular expressions and graphs of automata and transducers for searching and extraction; and 4) Build cascades of rules.

Corpora SrpKor2013 (cf. 2.1) and RudKor (cf. 2.2) can be searched by OCWB, while a search of RudKor is also available through NoSketch Engine.

This paper presents the process of upgrading the existing corpus search web interfaces of OCWB and NoSketch Engine in order to enable corpus query expansion.

3 Lexical resources

In order to improve the current corpus search capabilities based on linguistic annotation, it is necessary to consult external lexical resources. The following lexical resources have been developed for Serbian by the HLT Group at the University of Belgrade and JeRTeh Society:

- System of Serbian morphological electronic dictionaries (Unitex DELA format);

[7] https://unitexgramlab.org/

– Semantic network WordNet for Serbian;
– Terminological databases Termi, RudOnto, GeolISS.

3.1 Serbian morphological resources

The system of morphological electronic dictionaries of the Serbian language (SMD) (Krstev, 2008) is the core for the morphological expansion. SMD follows the methodology and format (known as DELAS/DELAF) that was developed in LADL (Laboratoire d'Automatique Documentaire et Linguistique) under the guidance of Maurice Gross. The format of a DELAS-type dictionary basically consists of simple word lemmas, each accompanied by inflectional class code. Every inflectional class code is associated with a corresponding finite-state transducer responsible for the generation of all inflectional forms of DELAS lemma. Thus, the DELAS-type dictionary with finite-state transducers for inflection enables the production of a DELAF-type dictionary which consists of all inflectional forms of DELAS-lemmas with their corresponding grammatical information. The Serbian morphological dictionary of simple words contains 190,000 lemmas which yield the production of approximately 2.4 million different inflected forms for lemmas and about 7.6 million forms with associated grammatical categories. At present, the dictionary of compounds has about 18,000 lemmas covering different parts of speech.

Lexical data have been migrated from textual e-dictionaries to a lexical database. After years of development, SMD, developed as a system of textual files, have become a large and complex lexical resource. An on-line application for dictionary development and management, based on a central lexical data repository (lexical database) is developed offering various possibilities for improvement of SMD, e.g. control of data consistency and introduction of explicit relations between lexical entries, automatic generation of dictionary candidates. The new version of service Vebran (cf. 4) is using this database for morphological expansion (Stanković et al., 2018).

The automatic procedure was used to transfer data from the existing dictionaries into the lexical database and to store all information about lemma and form entries as structured data. DELAS-lemma entries are generally mapped to entries in tables LexicalEntry and LexicalSense (Figure 1). A lemma, its corresponding PoS and inflectional class code (defining all inflected forms) are stored in the LexicalEntry table, while associated syntactic, semantic, domain and other types of markers are separated. Identical lexical entries from DELAS sharing the same inflectional class (e.g. `vez,N297`)

are merged into one LexicalEntry, while associated markers that differentiate senses are recorded in the LexicalSense and SenseProperties tables.

All inflected forms of lemma `vez` (`vez`, `veza`, `vezu`, `veze`, `vezom`, `vezovi`, `vezova`, `vezovima`, `vezove`) are stored in the table Forms, 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 table. For instance, the form `vezom` has one set of grammatical categories assigned to it `:ms6q` (the instrumental case, singular), while three sets of grammatical codes (`:ms2q`, `:mw2q`, `:mw4q`) are assigned to the form `veza` (the genitive case, singular and paukal, as well as accusative paukal). In addition, sets of grammatical categories are represented as individual categories in the table FormGramCatProperties, as presented on the left side of Figure 1. More details about multi-word units mapping, markers and relations between lexical entries can be found in (Stanković et al., 2018).

3.2 Semantic networks

The Serbian wordnet (SWN) has been developed in the scope of the Balkanet project following the model adopted for the EuroWordnet project (Krstev et al., 2004). More than 22,000 synsets built by app. 28,000 literals are Princeton WordNet (PWN), except for 532 Balkan specific concepts that are connected with other Balkan languages, and 155 Serbian specific concepts that remain unconnected with other languages. Since the core of wordnets developed for Balkan languages was produced by translation of the basic synsets in the PWN 3.0, the hypernym/hyponym relations in SWN mirror its hierarchical structure. Other relations are implemented more freely, depending on specific lexicalizations in Serbian. These relations include antonymy, meronymy, as well as some cross-part of speech relations (XPoS), such as causes and be_in_state.

A synset ENG30-14473222-n (Figure 2), visualized by Hydra[8] (Rizov and Dimitrova, 2016), defined as *Nečije opšte okolnosti ili uslovi u životu (uključujući sve što vam se dešava)* ("your overall circumstances or condition in life (including everything that happens to you)") has three literals: *okolnosti*, *sudbina* and *sreća* ("circumstances", "destiny", "luck"). An example of hypernym to ENG30-14473222-n is *uslov* ("condition") and the corresponding hyponym is *srećne okolnosti* ("lucky circumstances"). All mentioned relations can be used for query expansion.

[8] https://dcl.bas.bg/bulnet/

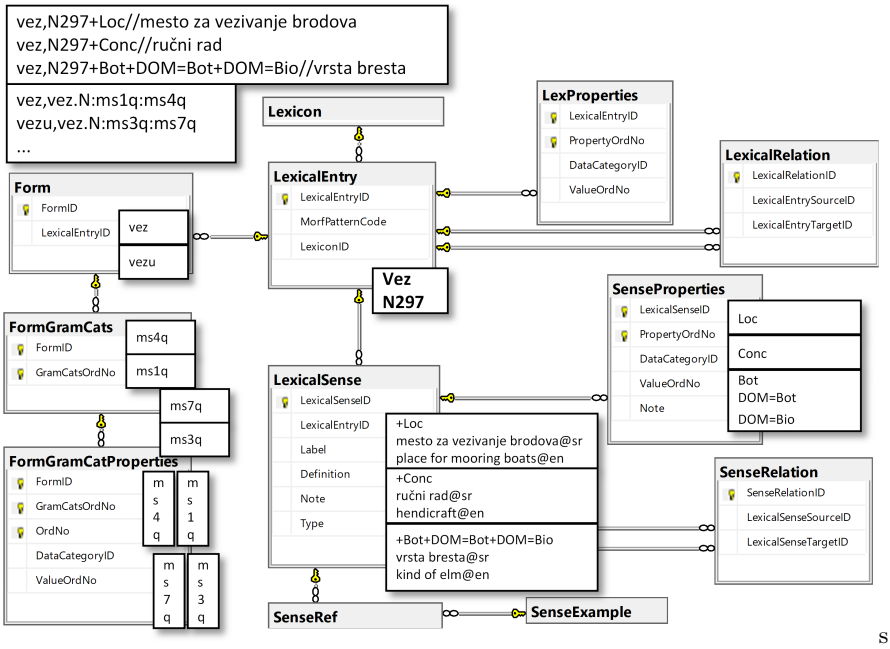

Figure 1. DELA to lexical database mappings

.

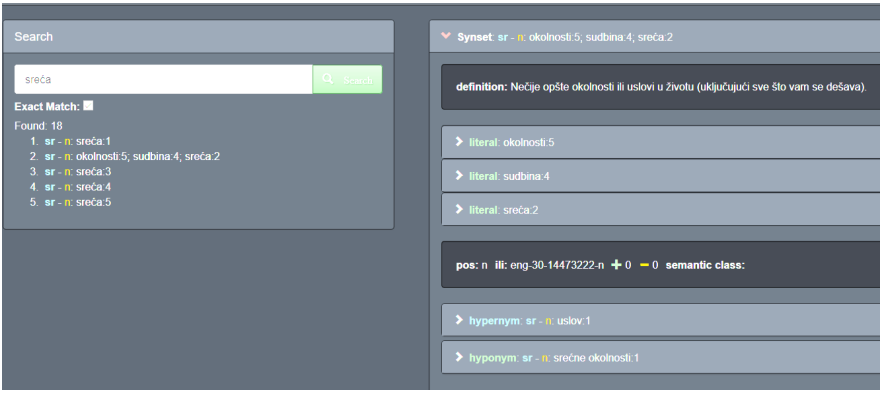

Figure 2. Wordnet synsets with literal "sreća"

.

3.3 Terminological resources

Termi[9] application supports the development of terminological dictionaries in various fields (mathematics, agronomy, mining), as well as the processing and presentation of terms in Serbian and English. Termi has been recently supplemented by the general Serbian-German bilingual dictionary extracted from 14 contemporary novels (Andonovski et al., 2019). All verified terms are available in public view, while additional terms are internally available to authorized users, according to their role and domain. A hierarchical display of the vocabulary terms is available for each domain. Besides its name, each term has its synonyms, abbreviations, description and bibliography. In case that the description of a term contains a LaTeX fragment, the fragment will be interpreted, which helps in the presentation of mathematical formulae.

Rudonto is a terminological resource developed to support knowledge management in mining engineering, focusing on the application in mining equipment and mine safety domains. Through export to several specific formats, RudOnto ontologies offer the possibility of generating stand-alone terminological resources or ontologies from specific sub-fields (sub-domains) (Kolonja et al., 2016).

GeolISSTerm represents the core of GeolISS (GEOLogical Information System of Serbia), and it is implemented as an aggregation of geological vocabularies, collections of terms and text definitions of entities thought to exist in a domain or collections of possible values for properties. The terms in the vocabularies are used to classify observations/interpretations, or to specify attribute values. GeolISSTerm is organized as a taxonomy with definitions for each entry, accompanied by synonyms, bibliographical references, equivalent terms and definition in another language (presently only English equivalents of definitions are in the database).

Externally developed Dictionary of library and information sciences[10] encompasses the terminology of theory and practice of librarianship and information sciences and a wide range of close or related fields, in Serbian, English and German languages. The languages in this dictionary have equal status. Online version currently includes 40,000 entries, but for the implementation of our web service, the older version is used with 23,400 entries (11,300 in English and 12,100 in Serbian, 910 definition or annotation terms which belong to library standards, and 2,200 acronyms of international and national entities). The intention of this dictionary is to be the useful elec-

[9] http://termi.rgf.bg.ac.rs/
[10] http://rbi.nb.rs

tronic resource of information for Library and Information Science professionals, for scientists and students, as well as for library users with different interests.

4 Vebran Web Services

Vebran Web Services enable users to search corpora using query syntax which is not supported by back-end query processors of OCWB (CQP) and NoSketch Engine (Manatee) in the following way:

- user can request that particular term X in a given query should be replaced with lemmas or word forms of terms which are semantically related to X in some manner (synonyms of X, antonyms of X, meronyms of X, hyponyms of X, hypernyms of X);
- user can request that particular lemma X in a given query should be replaced with its inflectional paradigm, i.e. with all word forms of lemma X.

4.1 Services architecture

The architecture of query expansion via Vebran Web Services (Figure 3) resembles a typical client-server architecture. Corpus web search interfaces, OCWB/CQPweb and NoSketch Engine / Bonito, perform the role of clients for Vebran Web Services, requesting a set of lemmas or word forms related to a given term X.

Firstly, clients need to get authorization in order to send their requests to Vebran Web Services. In this case, clients use an access token to identify themselves (Figure 4). The only parties that should ever see the access token are the client itself, the authorization server, and the resource server. The client should ensure that the storage of the access token is not accessible to other clients on the same device. The access token can only be used over a https connection, since passing it over a non-encrypted channel would make it trivial for third parties to intercept. The token endpoint is where apps make a request to get an access token for a user.

After successful authorization, clients are allowed to send a request specifying the term X and the relation (semantic or morphological) which should exist between X and the requested lemmas or word forms. Based on the client's request, Vebran services consult external lexical resources (see Section 3) and generate a response to the client. Communication with Vebran Web Services is based on RESTful technology, implemented in the Microsoft

MVC.Net web application framework which uses the model–view–controller pattern.

Clients are open source software (OCWB/CQPweb and NoSketch Engine / Bonito have been implemented in PHP and Python respectively) and their source code has been adapted to:

- send requests to Vebran Web Services,
- receive a response from Vebran Web Services,
- expand a given user query with a response from Vebran Web Services producing a new query with syntax acceptable by back-end query processors (OCWB/CQP and NoSketch Engine / Manatee).

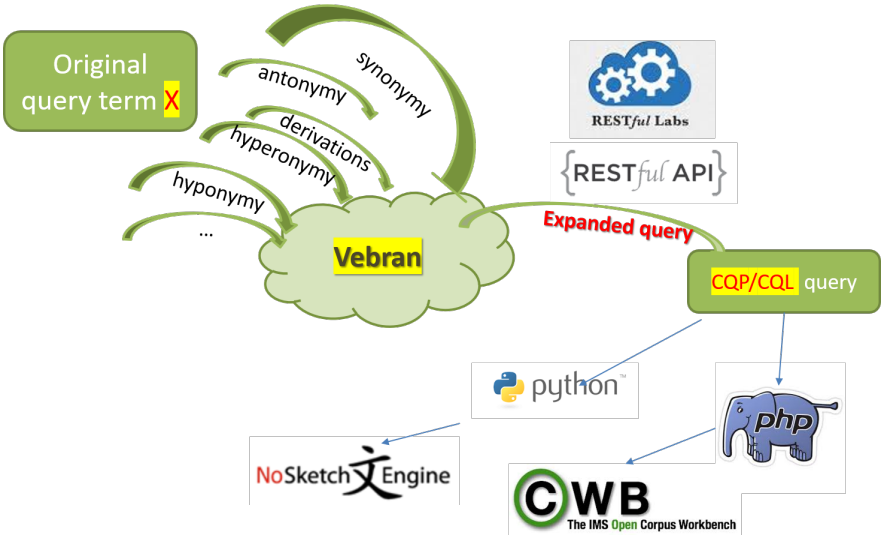

Figure 3. Query expansion architecture

.

4.2 Morphological expansion

Morphological query expansion is an alternative to queries based on a part-of-speech annotation of the corpus. That alternative is necessary to

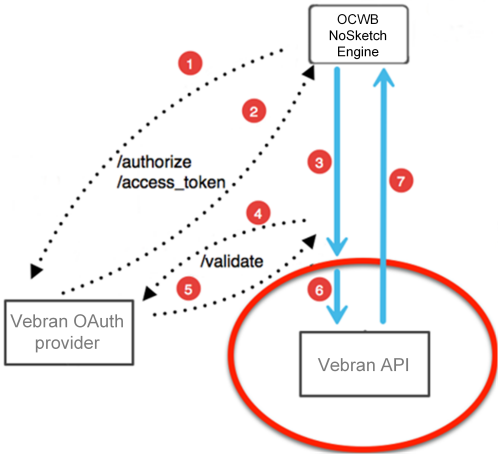

Figure 4. Oauth 2.0 Access Token Enforcement

.

overcome recall problems caused by tagging errors and limitations imposed by the format of a TreeTagger full-form lexicon. Each entry of the TreeTagger full-form lexicon contains one-word form and a sequence of tag-lemma pairs that could correspond to that word form (Schmid, 1997). TreeTagger full-form lexicon does not allow the possibility of a lexicon entry with two or more tag-lemma pairs corresponding to the same word form and having the same PoS tag, while having different lemmas. However, it is a common case that different short-length lemmas in Serbian, e.g. nouns `tat` ("thief") and `tata` ("dad") have homograph word forms (`tati`, `tatom`, `tate`, `tatu`, `tata`) causing that lexicon entries with these forms cannot contain both tag-lemma pairs `(N, tat)` and `(N, tata)` where `N` is PoS tag denoting noun. Thus, creator of full-form lexicon has to choose which tag-lemma pair will keep and the choice is commonly based on the automatic process which randomly favors one lemma over another. As a result, the query `[lemma="tata"]` in SrpKor2013, automatically PoS-tagged by TreeTagger, gives only 30 results and only word types `tatama` (poor recall). Actually, forms of lemma `tata` are far more frequent in Serbian (SrpKor2013, too) than forms of lemma `tat`.

OCWB (Evert and The OCWB Development Team, 2019) and NoSketch Engine (Rychlý, 2007) treat corpus as a table where :

– each row represents either a particular corpus position (token) or an XML structure tag;
– only the first column (called `word`) is mandatory and represents token values or XML tags;
– if the table includes more columns then each column represents a specific type of information (linguistic and non-linguistic) associated with a corresponding token.

The names of corpus columns, also called *positional attributes* by OCWB and NoSketch Engine, are used in queries in the form

[positionalAttribute="regularExpression"].

Besides the column `word`, SrpKor2013 also uses positional attributes `pos` (part of speech) and `lemma`, while RudKor uses positional attributes `tag` (part of speech) and `lemma`.

The general idea behind the morphological expansion is to replace lemma X in a given user query with the corresponding inflected forms of X in the specified alphabet(s) and, optionally, with restrictions regarding grammatical categories. Inflected forms are stored in LeXimirka database and originate from Unitex DELAF and DELACF dictionaries, described in Section 3.1. The inflection of multiword units is additionally supported by the rule based system. The system supports different alphabets and character encodings (the aurora alphabet and ISO-8859-1 character encoding for SrpKor2013 corpus, Serbian Latin alphabet and UTF-8 character encoding for RudKor corpus), in order to enable query expansion for different corpora.

There are several functions of Vebran Web Services which handle the morphological expansion:

– `delaf`[11],
– `obliciZaCQP`[12],
– `delafs`[13].

Function `delaf` expects input parameters (Table 1) via a POST request as a JSON (JavaScript Object Notation) structure like Figure 5 representing all plural word forms of lemma/noun `sreća` ("happiness") in Serbian using Serbian Latin alphabet. The appropriate output result is in the form of a regular expression: `sreć(a|ama|e)`, that can effectively retrieve all inflected plural forms as requested.

[11] http://hlt.rgf.bg.ac.rs/vebran/api/delaf/
[12] http://hlt.rgf.bg.ac.rs/vebran/api/obliciZaCQP/
[13] http://hlt.rgf.bg.ac.rs/vebran/api/delafs/

```
{
  lema:'sreća',
  alphOut:'L',
  lngIn:'sr',
  lngOut:'sr',
  POS:'N',
  GramCats:'p',
  fleksije:false,
  dlfByLemma:false
}
```

Figure 5. JSON structure of a request for all plural word forms of lemma/noun `sreća` ("happiness") in Serbian using Serbian Latin alphabet

Parameter	Value examples	Description
lema	sreća	Requested lemma X
alphOut	C	Alphabet of the output result C-Cyrillic, L-Latin, A-Aurora, combinations like CL, CA, LA, CLA are allowed
lngIn	sr	language of a given lemma X, by default sr (Serbian)
lngOut	sr	language of the function result, by default sr (Serbian)
POS	N	part of speech for a given lemma X
GramCat	p	morphological constraints as SMD data category values: s-singular, p-plural, 1..7 — case restrictions, etc.
fleksije	false	indicator whether the result contains only lemmas X or their inflectional forms as well (true, false)
dlfByLemma	false	indicator of output format (forms grouped by lemma or not)

Table 1. Parameters for morphological expansion requests (function `delaf`)

Function `obliciZaCQP` requires lemma as an input parameter, while part of speech is optional. The function is adapted to SrpKor which uses aurora alphabet. The function result is a regular expression (CQP and Manatee syntax) using aurora alphabet. The example (Figure 5) with alphOut:'A' would return `srecx(a|ama|e|i|o|om|u)`.

Function `delafs` uses the same input parameters, but generates output in the form of a list: `sreća; srećama; sreće; sreći; srećo; srećom; sreću; срећа; срећама; среће; срећи; срећо; срећом; срећу`. This format is used for query expansion in digital libraries Romeka and Bibliša, since their query processors require such forms.

Web search interfaces for OCWB and NoSketch Engine, CQPweb and Bonito respectively, have been adapted to accept and preprocess user query with an expanded syntax which is not supported by OCWB and NoSketch Engine corpus search engines. For morphological expansion fake positional attribute `flemma` is introduced allowing a user to request the inflectional paradigm of a lemma, e.g. `tata`, with a query `[flemma="tata"]`. Preprocessing includes:

- extraction of `flemma` value (e.g. `tata`);
- sending a request to Vebran Web Services (similar to Figure 5);
- using Vebran Web Services response to generate the final query (e.g `[word="tat(a|ama|e|i|om|u)"]`) adjusted with allowed syntax of the query processor and
- sending the final query to the query interpreter.

In case of `[flemma="tata"]` vs. before used `[lemma="tata"]`, we get 2,171 results in SrpKor2013 (100% recall) instead of earlier 30 results. However, due to homographs, it is possible that some retrieved forms do not correspond to the given lemma. Actually, `[flemma="tat"]` would produce similar results, but most of them would not be relevant.

A similar example can be found in RudKor through NoSketch Engine for lemma: `kap` ("a drop") vs. lemma `kapa` ("a cap"). The example page of search results for `kapa` (Figure 6), shows that the first and the last concordance line correspond to lemma `kap`.

Although query expansion produces 100% recall, precision is reduced due to homographs.

The problems with query expansion in case of multi-word units (MWUs) described in (Утвић et al., 2019) have recently been resolved. Different solutions need to be applied for a case where all components of an MWU are the same as their lemmas, e.g. `leksički resurs` ("lexical resource") and for

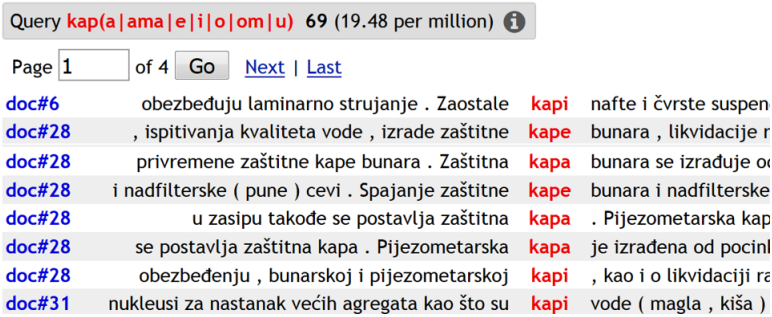

Figure 6. Expanded query in NoSketch Engine

.

a case where some components are not lemmas but instead some other inflected forms, like `leksička baza` ("lexical database"). For example, if the web service response contains the MWU `leksički resurs`, the generated query `([lemma="leksički"][lemma="resurs"])` would appear to succeed because both the leksički and the resurs are lemmas of the corresponding lexemes. However, if the MWU `leksička baza` can be found in the web service response, the generated query `[lemma="leksička"][lemma="baza"]` would find nothing since `leksička`, the inflected form of the lexeme `leksički`, is not a lemma of that lexeme.

Another problem is caused by the fact that MWUs may contain some components that inflect as a part of MWU and some that do not, e.g. `jato ptica` ("flock of birds"), where the first component inflects and the second does not (as a part of MWU), so generated query should be `([lemma="jato"][word="ptica"])`.

The solution for both problems includes web service function `MWUzaCQP`[14] which uses SMD to get information about MWUs, as well as an implementation of additional heuristics to process the out-of dictionary MWUs (Table 2):

1) If an MWU given by a user is found in SMD, its inflectional code is analysed and appropriate transformation applied. For instance, the inflectional code AXN associated with the MWU `leksički resurs` means that MWU contains three components, an adjective (A) that inflects, a separator (X) that does not inflect and a noun (N) that inflects. Another example is inflectional code N2X associated with MWU `jato ptica` where only the

[14] http://hlt.rgf.bg.ac.rs/vebran/api/MWUzaCQP/

first component (a noun) inflects. Web service answers client's request based on an inflectional code of MWU using markers `C:` (compound) `_L` (lemma) and `_W` (word), where the last two indicate whether marked string should be associated with the positional attribute `lemma` (it inflects as a part of MWU) or with the positional attribute `word` (it does not inflect as a part of MWU).

2) If SMD does not contain an MWU as a compound, each component of MWU is analysed separately. A component of MWU which has been found to be a lemma is associated with marker `_L` (e.g. `ptica pevačica` in Table 2). If a component has been found to be an inflected form different from its lemma, component is then replaced with the corresponding lemma and marked with `_L` (e.g. `leksička relacija` in Table 2). If a component cannot be found in SMD, it is treated as a word form that does not inflect (as a part of MWU) and it is marked with `_W`.

Requested MWU	Inflectional code	Service response	Generated query
leksički resurs	AXN	C:leksički_L resurs_L	[lemma="leksički"] [lemma="resurs"]
leksička baza	AXN	C:leksički_L baza_L	[lemma="leksički"] [lemma="baza"]
jato ptica	N2X	C:jato_L ptica_W	[lemma="jato"] [word="ptica"]
ptica pevačica	?	C:ptica_L pevačica_L	[lemma="ptica"] [lemma="pevačica"]
leksička relacija	?	C:leksički_L relacija_L	[lemma="leksički"] [lemma="relacija"]
slobodan kao ptica	A3XN2	C:slobodan_L kao_W ptica_L	[lemma="slobodan"] [word="kao"] [lemma="ptica"]
album za slike	N4X	C:album_L za_W slike_W	[lemma="album"] [word="za"] [word="slike"]

Table 2. Examples of queries with MWUs

The presented heuristics to process queries with MWUs do not use recursive query expansion, that is, the inflectional paradigm of components is not produced in the form of a regular expression containing a

union of all word forms. Instead, the generated query uses positional attribute `lemma` and therefore a linguistic annotation provided by TreeTagger (last column of Table 2). An alternative would be to create a temporary query with expanded syntax, that is, with fake positional attribute `flemma` and then use a corresponding web service again as many times as there are components that inflect as a part of MWU. In that case, an example of the final query which searches for an inflectional paradigm of MWU `ptica pevačica` would be: `[word="ptic(a|ama|e|i|o|om|u)"] [word="pevačic(a|ama|e|i|om|u)"]`.

In the example `slobodan kao ptica` ("free as a bird") the first and the last component inflect as parts of MWU, so the final query would be `[word="slobod(an|na|ne|ni|nih|nim|no|nog|noj|nom|nome|nu)"]`[15] `[word="kao"] [word="ptic(a|ama|e|i|o|om|u)"]`. The morphological category degree has been restricted to the value "positive" eliminating comparative and superlative forms from inflected forms of the adjective.

In example `album za slike` ("photo album") only the first component has inflected forms: `[word="album(|a|e|i|ima|om|u)"] [word="za"] [word="slike"]`.

An extra fake positional attribute `mwulemma` allows a user to request an inflectional paradigm of an MWU lemma, e.g. `leksički resurs`, with a query `[mwulemma="leksički resurs"]`.

4.3 Semantic expansion

The general idea to expand an original query with word forms related to term X is based on the use of semantic and terminological resources to find other terms such that there exists a given semantic relation (synonymy, antonymy, hyperonymy, meronymy) between those terms and the term X.

Web service function `sinonimi/post`[16] receives a JSON structure as an input and returns the synonyms of a given lemma. For example, function `sinonimi/post` for the lemma `sreća` returns `S:околности;S:срећа|срећама|среће|срећи|срећо|срећом|срећу;` `S:судбина|судбинама|судбине|судбини|судбино|судбином|судбину` as a set of corresponding synonyms and their inflected forms using Serbian Cyrillic alphabet.

More details for Semantic expansion can be found in (Утвић et al., 2019)

[15] Actually, this regular expression also includes `noga` i `nima`, but they are omitted to avoid text longer than length of line.

[16] http://hlt.rgf.bg.ac.rs/vebran/api/sinonimi/post

5 Conclusion

The paper describes Vebran web services that enable corpus query expansion and support corpus search that combines linguistic annotation of the corpus with external lexical resources. The presented approach allows search results to include the inflectional paradigm of the lexemes in a given user query, as well as the word forms semantically related to them (synonyms, antonyms, hyperonyms, etc.) where semantic relations are available through the semantic network (wordnet). The emphasis in this paper is on morphological query expansion supported by Vebran web services. Services consult external lexical resources and produce regular expressions that improve recall of the retrieved inflected forms for a given lemma. The described hybrid approach was successfully tested by modifying the web interface of corpus search tools OCWB and NoSketch Engine. Vebran web services are currently available only to the authorized users and applications. Further improvement of web services will include better support for multi-word units and a more flexible combination of different query parameters.

Acknowledgment

This research was partly supported by the Ministry of Education, Science and Technological Development through projects ON-178006 and III 47003.

References

Andonovski, Jelena, Branislava Šandrih and Olivera Kitanović. "Bilingual Lexical Extraction based on Word Alignment for Improving Corpus Search". *The Electronic Library* Vol. 37, no. 2 (2019): 722-739

Evert, Stefan and The OCWB Development Team. *CQP Query Language Tutorial*, 2019, the IMS Open Corpus Workbench (CWB 3.4.16), May 2019. Accessed August 1, 2019. http://cwb.sourceforge.net/files/CQP_Tutorial.pdf

Hardie, Andrew. "CQPweb — Combining Power, Flexibility and Usability in a Corpus Analysis Tool". *International Journal of Corpus Linguistics* Vol. 17, no. 3 (2012): 380–409

Kolonja, Ljiljana, Ranka Stanković, Ivan Obradović, Olivera Kitanović and Aleksandar Cvjetić. "Development of terminological resources for expert knowledge: a case study in mining". *Knowledge Management Research & Practice* Vol. 14, no. 4 (2016): 445–456. https://doi.org/10.1057/kmrp.2015.10

Krstev, Cvetana. *Processing of Serbian — Automata, Text and Electronic Dictionaries*. Belgrade: Faculty of Philology, 2008

Krstev, Cvetana, Gordana Pavlović-Lažetić and Ivan Obradović. "Using Textual and Lexical Resources in Developing Serbian Wordnet". *Romanian Journal of Information Science and Technology* Vol. 7, no. 1–2 (2004): 147–161

Paumier, Sébastien. *Unitex 3.1 User Manual*, 2016, accessed August 1, 2019. http://releases.unitexgramlab.org/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf

Rizov, Borislav and Tsvetana Dimitrova. "Hydra for Web: A Browser for Easy Access to Wordnets". In *Proceedings of the Eighth Global Wordnet Conference, Research Institute for Artificial Intelligence, Romanian Academy*, 339–343, 2016

Rychlý, Pavel. "Manatee/Bonito — A Modular Corpus Manager". In *First Workshop on Recent Advances in Slavonic Natural Language Processing*, Sojka, P. and A. Horák, 65–70. Brno: Masaryk University, 2007

Schmid, Helmut. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In *New Methods In Language Processing*, Jones, D. B. and H. Somers, Chapter 12, 154–164. Routledge, 1997

Schmid, Helmut. "Improvements in Part-of-Speech Tagging with an Application to German". In *Natural Language Processing Using Very Large Corpora*, Armstrong, S. et al. *Text, Speech and Language Technology*, Vol. 11, Chapter 12, 154–164. Dordrecht: Springer, 1999,

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac and Miloš Utvić. "A Tool for Enhanced Search of Multilingual Digital Libraries of E-Journals". In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC),(Istanbul, Turkey*, 2012, 1710–1717

Stanković, Ranka. "Modeli ekspanzije upita nad tekstuelnim resursima". Phdthesis, Univerzitet u Beogradu, Matematički fakultet, Beograd, 2009

Stanković, Ranka, Cvetana Krstev, Biljana Lazić and Mihailo Škorić. "Electronic Dictionaries - from File System to lemon Based Lexical Database". In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, McCrae, J. P., C. Chiarcos, T. Declerck, J. Gracia and B. Klimek, 48–56. 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. https://doi.org/10.1108/EL-11-2017-0239

Утвић, Милош. "Анотација Корпуса савременог српског језика". *Инфотека* Vol. XII, no. 2 (2011): 39–51

Utvić, Miloš. "Izgradnja referentnog korpusa savremenog srpskog jezika". Phdthesis, Univerzitet u Beogradu, Filološki fakultet, Beograd, 2014, accessed August 1, 2019. https://fedorabg.bg.ac.rs/fedora/get/o:10061/bdef:Content/download

Утвић, Милош. "Листе учестаности Корпуса савременог српског језика". In *Научни састанак слависта у Вукове дане. Српски језик и његови ресурси: теорија, опис и примене. 3/43. научни састанак слависта у Вукове дане, Београд, 12-15. IX 2013.*, Милановић, А., Ж. Станојчић and Љ. Поповић, Vol. 43/3, 241–262. Београд: МСЦ, Универзитет у Београду, Филолошки факултет, 2014

Утвић, Милош В., Иван М. Обрадовић, Ранка М. Станковић, Александра Ђ. Томашевић and Биљана Ђ. Лазић. "Изградња специјалних корпуса савременог српског језика на примеру корпуса из области рударства". In *Српски језик и његови ресурси: теорија, опис и примене. 3/47. научни састанак слависта у Вукове дане, Београд, 2017.*, Ћорић, Б. and А. Милановић, Vol. 47/3, 103–118. Београд: МСЦ, Универзитет у Београду, Филолошки факултет, 2018. https://doi.org/10.18485/msc.2018.47.3.ch7

Утвић, Милош В., Ранка М. Станковић, Александра Ђ. Томашевић, Михаило Ђ. Шкорић and Биљана Ђ. Лазић. "Претрага корпуса заснована на употреби екстерних лексичких ресурса путем веб-сервиса". In *Српски језик и његови ресурси: теорија, опис и примене. 3/48. научни састанак слависта у Вукове дане, Београд, 2018.*, Ћорић, Б. and А. Милановић, Vol. 48/3, 279–298. Београд: МСЦ, Универзитет у Београду, Филолошки факултет, 2019. https://doi.org/10.18485/msc.2019.48.3.ch12

Vitas, Duško and Cvetana Krstev. "Processing of Corpora of Serbian Using Electronic Dictionaries". *Prace Filologiczne* Vol. 63 (2012): 279–292

Обрадовић, Иван, Александра Томашевић, Ранка Станковић and Биљана Лазић. "Увођење доменских и семантичких маркера за област рударства у српске електронске речнике", In *Српски језик и његови ресурси: теорија, опис и примене. 3/46. научни састанак слависта у Вукове дане, Београд, 2016.*, Драгићевић, Р. and А. Милановић, Vol. 46/3, 147–158. Београд: МСЦ, Универзитет у Београду, Филолошки факултет, 2017. http://doi.fil.bg.ac.rs/pdf/eb_ser/msc/2017-3/msc-2017-46-3-ch10.pdf

Extraction of Bilingual Terminology using Graphs, Dictionaries and GIZA++

UDC 81'322.2
DOI 10.18485/infotheca.2019.19.2.6

ABSTRACT: In science, industry and many research fields, terminology is rapidly developing. Most often, a language that is "lingua franca" for most of these areas is English. As a consequence, for many fields, domain terms are conceived in English, and are later translated to other languages. In this paper, we present an approach for automatic bilingual terminology extraction for English-Serbian language pair that relies on an aligned bilingual domain corpus, a terminology extractor for a target language and a tool for chunk alignment. We examine the performance of the method on a Library and Information Science domain. The obtained results, as well as the application that implements the method, are available on-line.
KEYWORDS: terminology extraction, terminology validation, GIZA++, graphs, Unitex, text classification.

PAPER SUBMITTED: 30 September 2019
PAPER ACCEPTED: 20 December 2019

Branislava Šandrih
branislava.sandrih@fil.bg.ac.rs
University of Belgrade
Faculty of Philology

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology

Belgrade, Serbia

1 Introduction

In science, industry and many research fields, terminology is rapidly developing. Most often, a language that is "lingua franca" for most of these areas is English. As a consequence, for many fields, domain terms are conceived in English, and are later translated to other languages. It does not happen rarely that a certain term is translated either as a short explanation of its meaning, or the translation is specifically adapted as an utterance in the language in which it is translated to (i.e. as a word in a target language). An example that demonstrates both cases is an English word "a screenshot", from the computer science. In Serbian, this term is either translated as *snimak*

ekrana (namely, *a photo of a current state of the screen*) or as a "skrinšot" (i.e, the word is transcribed). It is not uncommon that even experts from a certain field have difficulties while translating texts that contain domain terminology. As in the example with a "debugger", the transcribed version is adopted for everyday use in Information Technologies domain.

It is a challenge to produce and maintain up-to-date terminology resources, especially for an under-resourced language, such as Serbian. Today, Serbian terminology is transferred mainly from English, since it is better developed for many scientific and technological domains. Purely manual production of terminological resources is not the solution due to rapid changes both in research fields and corresponding terminology.

Multi-Word Expressions (MWEs) are lexical units composed of more than one word, which are syntactically, semantically, pragmatically, and/or statistically idiosyncratic (Baldwin and Kim, 2010). MWEs represent a class of linguistic forms spanning conventional word boundaries that are both idiosyncratic and pervasive across different languages (Constant et al., 2017).

As Baldwin and Kim (2010), among others, have pointed out, the question of what constitutes a word is surprisingly complex, and one reason for this is the predominance of elements known as MWEs in everyday language. They consist of several words (in the conventionally understood sense) but behave as single words to some extent.

An illustration is given in (Constant et al., 2017), with a MWE *by and large*, that has roughly equivalent meaning and syntactic function to adverb *mostly*. Among the problematic characteristics of this expression are (1) syntactic anomaly of the part-of-speech (POS) sequence preposition + conjunction + adjective, (2) non-compositionality: semantics of the whole that is unrelated to the individual pieces, (3) non-substitutability of synonym words (e.g., by and big), and (4) ambiguity between MWE and non-MWE readings of a substring by and large (e.g., *by and large we agree* versus *he walked by and large tractors passed him*).

Due to all these difficulties, tackling MWEs represents a special challenge. This paper aims at MWEs since terminology consists mainly of Multi-Word Terms (MWTs). MWTs are domain-specific MWEs. Terms consisting of a single word are mainly referred to as Single-Word Terms (SMTs).

In this paper, we describe an approach for obtaining bilingual terminology pairs automatically, initially proposed in (Krstev et al., 2018) and demonstrated on English-Serbian language pair. In this first approach, we performed and discussed only one setting of the experiment. After evalua-

tion, we recognised a need to examine several settings of the experiment, which are conducted and discussed in the later text.

The proposed approach is based on the following hypothesis:

> On the basis of bilingual, aligned, domain-specific textual resources, a terminological list and/or a term extraction tool in a source language, and a system for the extraction of *terminology-specific Multi-Words Terms* in a target language, it is possible to compile a bilingual aligned terminological list.

This paper is organised as follows. An overview of previous work on this topic is given in Section 2. Lexical resources and tools that were used in the experiments in Subsection 3. The proposed approach is thoroughly explained in Section 4. Results and a discussion are given in Section 5. A Web application that implements the proposed technique is presented in Section 6. Finally, conclusions and directions for future work are given in Section 7.

2 Related Work

Over the past years, in order to compile bilingual lexica, researchers used various techniques for MWT extraction and alignment that differ in methodology, resources used, languages involved and purpose for which they were built.

Bilingual lexica were compiled for different language pairs: English/French (Bouamor et al., 2012; Hamon and Grabar, 2016; Hazem and Morin, 2016; Hakami and Bollegala, 2017; Semmar, 2018), English/Spanish (Oliver, 2017), English/Arabic (Lahbib et al., 2014; Naguib Sabtan, 2016; Hewavitharana and Vogel, 2016), English/Urdu (Hewavitharana and Vogel, 2016), English/Italian and English/German (Arcan et al., 2017), English/Slovene (Vintar and Fišer, 2008), English/Croatian, Latvian and Lithuanian (Pinnis et al., 2012), English/Chinese (Xu et al., 2015), English/Hebrew (Tsvetkov and Wintner, 2010), English/Ukrainian (Hamon and Grabar, 2016), English/Greek (Kontonatsios et al., 2014), English/Romanian (Pinnis et al., 2012; Kontonatsios et al., 2014), Bengali/Hindi/Tamil/Telugu (Irvine and Callison-Burch, 2016), Slovak/Bulgarian (Garabík and Dimitrova, 2015) and Italian-Arabic (Fawi and Delmonte, 2015).

In several cases, the bilingual lists of MWTs were compiled in order to improve statistical machine translation (SMT) of an existing machine translation system (Bouamor et al., 2012; Tsvetkov and Wintner, 2010; Naguib Sabtan, 2016; Irvine and Callison-Burch, 2016; Semmar, 2018; Hewavitharana and Vogel, 2016; Arcan et al., 2017; Oliver, 2017), for the development of an existing language resource in a target language on the basis of a corresponding resource in a source language (e.g. used for development of the Slovenian WordNet (Vintar and Fišer, 2008) based on English WordNet), or for the presentation of bilingual correspondences between two languages (e.g. correspondences between Slovak-Bulgarian parallel corpus (Garabík and Dimitrova, 2015)).

Some approaches request parallel sentence-aligned data (Arcan et al., 2017; Garabík and Dimitrova, 2015; Bouamor et al., 2012; Semmar, 2018), while others perform the extraction on comparable corpora (Xu et al., 2015; Hazem and Morin, 2016; Hewavitharana and Vogel, 2016; Pinnis et al., 2012). For the technique used in (Naguib Sabtan, 2016), groups of aligned sentences (verses) were used. In (Irvine and Callison-Burch, 2016) authors performed two experiments, the first one relying on the existence of a bilingual dictionary with no parallel texts and the second one requiring only the existence of a small amount of parallel data.

In order to compile a bilingual lexicon for a specific domain, we combined and compared several settings. Besides using only a parallel sentence-aligned corpus, we conducted an experiment where sentences from the corpus were extended with a bilingual list of inflected word forms from a general-purpose dictionary, similarly as in (Tsvetkov and Wintner, 2010).

We compared different configurations for the extraction of domain terminology on both, source and target, sides. For the source side, we compare two cases. In the first case, we use an existing bilingual domain dictionary, similarly as in (Vintar and Fišer, 2008; Hakami and Bollegala, 2017; Kontonatsios et al., 2014). In the second case, we obtain source terminology using an existing term extractor, similarly to some other authors (Pinnis et al., 2012; Hamon and Grabar, 2016; Arcan et al., 2017).

For the extraction of terminology on the target side, we apply morphological and statistical analysis. A similar approach was taken by other authors (Bouamor et al., 2012; Lahbib et al., 2014; Fawi and Delmonte, 2015; Hamon and Grabar, 2016; Naguib Sabtan, 2016; Semmar, 2018).

3 Lexical Resources and Tools

As previously mentioned in Section 1, the approach proposed in (Krstev et al., 2018) relies on several lexical resources and tools:

i A sentence-aligned domain-specific corpus involving a source and a target language, denoted as $S(text.align) \leftrightarrow T(text.align)$. In this paper we refer to this tool as LIS-CORPUS.
As a textual resource, twelve issues with a total of 84 papers were aligned at the sentence level resulting in 14,710 aligned segments (Stanković et al., 2017; Stanković et al., 2014).[1] The Serbian part has 301,818 simple word forms (41,153 different), while the English part has 335,965 simple word forms (21,272 different).

ii A list of terms in the source language, denoted as $S(term)$.
This list can be either an external resource from the same domain or extracted from the text.
As an external resource, we used the Dictionary of Librarianship: English-Serbian and Serbian-English. It was developed by a group of authors from the National Library of Serbia.[2] In this paper we refer to this tool as LIS-DICT.
We also tried to extract terms on the source side. For this purpose, we decided to use an open-source software tool, FlexiTerm (Spasić et al., 2013). It automatically recognises MWTs from a domain-specific corpus, based on their structure, frequency and collocations. In this paper we refer to this tool as ENG-TE.
Three other MWT extractors were considered for obtaining English MWTs: TextPro[3] (Pianta et al., 2008), TermSuite[4] (Cram and Daille, 2016) and TermEx2.8.[5] Evaluation performed on the list of terms extracted by all four extractors and evaluated as potential MWU terms showed that FlexiTerm outperformed the other three.

iii A list of terms in the target language, denoted as $T(term)$.

[1] Biblisha

[2] A more enhanced version of this dictionary, available on-line, contains 40.000 entries (approximately 14.000 in Serbian, 12.400 in English and 14.000 in German). We used the version obtained from the authors for research purposes.

[3] TextPro (former KX toolkit)

[4] TermSuite is the Open Source and UIMA-based application drawn out from the European project TTC Terminology Extraction

[5] TermEx

This list can be either an external resource from the same domain or obtained from the text.

The only system developed specifically for the extraction of MWTs from Serbian texts is a part of LeXimir (Stanković et al., 2016), a tool for management of lexical resources. LeXimir consists of two modules for the terminology extraction. The first module is a rule-based system relying on e-dictionaries and local grammars developed in Unitex,[6] that are implemented as finite-state transducers (FST). The second module implements various statistical measures used for ranking of term candidates. In this research the system was tuned to recognise 26 most frequent syntactic structures, which were previously identified by an analysis of several Serbian terminological dictionaries and the Serbian e-dictionary of MWUs (Krstev, 2008). In this paper we refer to this tool as Serb-TE. Some of these structures are `A_N_Prep_N` in *republički zavod za statistiku* 'republic office for statistics' or `A_N_(A_N)`$_{gen}$ in *statistički godišnjak republičkog zavoda* 'statistical yearbook of the republican institute' where `A` stands for an adjective, `N` for a noun and `PREP` for a preposition. Each of these components can be a single word or a MWU. Our system was used in a mode in which all possible MWTs in a word sequence are recognised, and not only the longest one. For instance, for the sequence *studija slučaja u primeni mašinskog učenja* 'case study in application of machine learning' the recognised terms would be: *studija slučaja* 'case study', *studija slučaja u primeni* 'case study in application', *mašinskog učenja* 'machine learning' and the longest match would be *studija slučaja u primeni mašinskog učenja*. The list of the most frequent classes is presented in (Krstev et al., 2018).

We have also prepared an additional resource, namely a set of aligned and inflected English-Serbian single and multi-unit word forms (denoted as bi-list). We used two bilingual lexical resources that we processed with LeXimir: (a) Serbian Wordnet (SWN),[7] which is aligned to the Princeton WordNet (Princeton WordNet, 2010), and (b) an English-Serbian list containing general lexica.

The production of bi-list was done in several steps:

1. First, a parallel list from SWN and PWN containing aligned English/Serbian Single and Multi-Word literals was compiled. This list was then merged with the bilingual list yielding a new list.

[6] Unitex/GramLab, a lexical-based corpus processing suite

[7] Serbian WordNet

2. To each Serbian noun, verb or adjective from the merged list we assigned its inflected forms obtained from the Serbian morphological e-dictionaries (Krstev, 2008). These inflected forms have various grammatical codes assigned to them, which were used in the final step.
   As mentioned earlier, 26 most frequent syntactic structures, grouped in 14 graphs labelled: AXN, 2XN, N2X, N4X, AXN2X, NXN, AXAXN, N6X, AXN4X, 2XAXN, AXN6X, 12N8X, 2XAXN2X and 2XAXN4X were used for terminology extraction. In this notation, A stand for and adjective, N for noun, X for a component that do not inflect. The separators like space and hyphen are also labeled as X and nX is short notation for repetition of X n times.
   As an example, we present a FST for extraction of type N4X, meaning that the first component is noun that inflects, followed by two words that do not inflect. The N4X graph is shown in Figure 1, which shows two paths that recognize two syntactic structures:
   – N_Prep_Np, noun (1^{st} component that inflects) followed by prepositional phrase (3^{rd} component agrees in case with a preposition), as in examples: *lista sa podacima* (list with data), *mašina za pranje* (washing machine), *ugovor o radu* (work contract);
   – N_Ngi_Ngi NxAg(i)xNg(i), 1^{st} component inflects; the second and the third component (noun or adjective) are in genitive case (such as *izrada geološke karte* (creation of a geological map)) or instrumental case (such as *etiketiranje vrstom reči* (Part-of-Speech tagging))
   The graph output consists of 4 values for each recognised MWU, separated by ";": graph label (grf04a or grf04b), followed by a label that indicates grammatical number (sin or plu); followed by recognised form ($n.INFLECTED\ p\ np$ or $n.INFLECTED\ ng1\ ng2$) and lemmatised inflective component followed by constant components ($n.LEMMA\ p\ np$ or ($n.LEMMA\ ng1\ ng2$). An example would be: "grf04b;plu;ciljeva pronalaženja informacija;cilj pronalaženja informacija;" (goal of finding information).
   A Software solution for multi-word units extraction displayed in Figure 2 offers possibilities for general NLP processing on selected corpus (applying lexical resources, generating bag of words and extraction of unknown words), extraction of selected syntactic patterns applying specified options and further processing (lemmatisation, calculation of statistical measures, support for manual evaluation and final evaluation report). For automatic extraction and lemmatisation, the system calls Unitex command-line functions in the background to apply appropriate graphs.

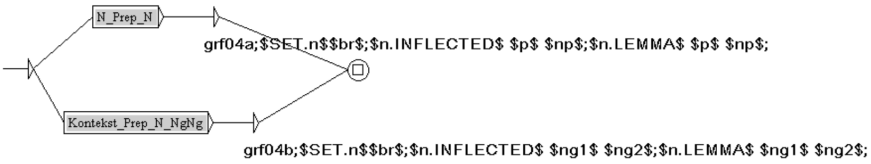

Figure 1. A FST for extraction of type N4X

3. A similar procedure was performed for English nouns, verbs and adjectives from the bilingual list. In order to obtain inflected forms with grammatical categories we used the English morphological dictionary from the Unitex distribution and the MULTEX-East English lexicon.[8]
4. In the final step Serbian and English inflected word forms were aligned taking into account the corresponding grammatical codes, which were previously harmonised to the best possible extent.
   For example, the grammatical category codes in the Serbian dictionary are a/b/c, for the positive/comparative/superlative forms. The Unitex English dictionary does not have a code for the positive, while the codes for the comparative and superlative are C and S, respectively. The second English dictionary followed the MULTEXT-EAST specification, using p/c/s as codes. Thus the Serbian codes a/b/c were mapped to English codes ϵ/C/S and p/c/s, respectively.

4 Terminology Extraction

In our experiments the source language is English, and the target language is Serbian. For input and processing we used resources and tools described in Section 3. As the aligned corpus (Input i) we used LIS-CORPUS alone, or augmented with bilingual pairs from the BI-LIST. For the extraction of English terms (Input ii) we used the English side of the dictionary LIS-DICT in one series of experiments, and term extractor ENG-TE in the other, while the extraction of Serbian terms (Input iii) was done by SERB-TE.

With the notation introduced in Section 3, the extraction procedure consists of the following steps:

[8] MULTEX-East English lexicon

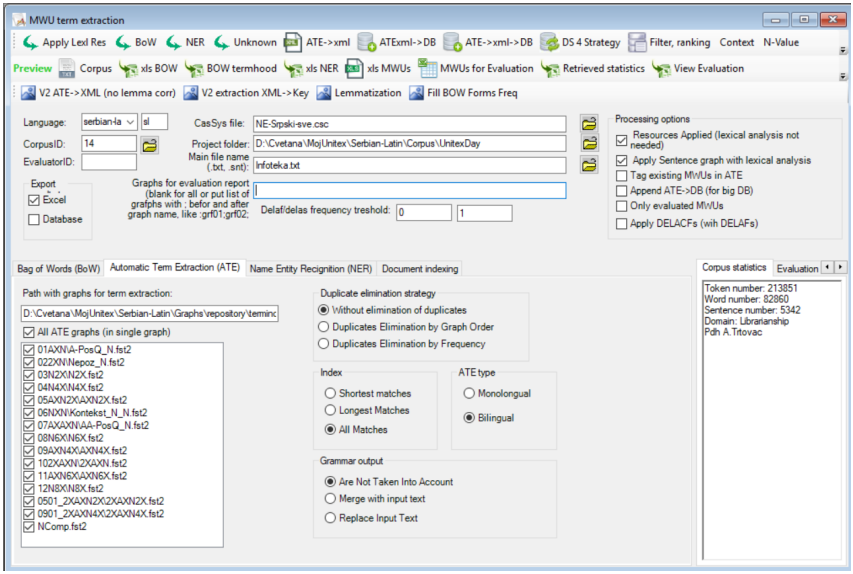

Figure 2. Software solution for MWT extraction

i Aligning bilingual chunks (possible translation equivalents) from the aligned corpus. We will denote aligned chunks by $S(align.chunk) \leftrightarrow T(align.chunk)$.

The alignment of chunks began with pre-processing using MOSES (Koehn et al., 2007) to perform tokenisation, truecasing and cleaning. In the next step a 3-gram translation model was built using KenLM (Heafield, 2011), followed by the training of this translation model. For the purpose of word-alignment, phrase extraction, phrase scoring and creation of lexicalised reordering tables, GIZA++[9] (Och and Ney, 2000) was used, together with the *grow-diag-final* symmetrisation heuristic (Koehn et al., 2003).

Each pair of aligned chunks from this list also contained information about inverse and direct phrase translation probability.[10] We have initially discarded all aligned chunks that did not have at least one of these probabilities greater than 0.85, simultaneously eliminating punctuation

[9] Statistical Machine Translation toolkit

[10] The way phrase translation probabilities are determined

marks. Chunks that consisted of punctuation marks and digits only were also discarded.

Afterwards, we provided a Bag-of-Words (BoW) representation for English terms from the LIS-DICT, i.e. from ENG-TE, and removed stop words from it, producing a list mainly populated with content words. Then we lemmatised each token from the BoW. Aligned chunks in which the English part did not have at least one lemmatised content word from the BoW list were eliminated.

ii Keeping only chunks (from the previous step) in which the source part of the chunk matches a term in the list of domain terms in the source language remain: $S(align.chunk) \sim S(term) = \{(s_1, s_2) : s_1 \sim s_2\}$, where the symbol $\sim$ denotes the relation "match" (explained later).

iii Keeping only chunks (from the previous step) in which the target part of the chunk matches a term in the list of extracted MWTs in the target language remain: $T(align.chunk) \sim T(term) = \{(t_1, t_2) : t_1 \sim t_2\}$.

The relation "match" ($\sim$) is defined as follows: if a chunk is represented by an unordered set of distinct words obtained from the chunk after removal of stop words, the two chunks match if they are represented by the same set. For example, if there is one "dictionary words" chunk and another "words from dictionary" chunk, their corresponding set representations are {dictionary, words} and {words, dictionary}, respectively ('from' should be discarded as functional word). Since these two sets are equal, these two chunks match.

Let two candidate pair chunks be "reči iz rečnika" (translated as 'words from dictionary') and "reč o rečniku" (translated as 'dictionary words'). Considering the specific application, these two chunks should match. If observed as unordered set of distinct content words, these chunks can be written as {reči, rečnika} and {reč, rečniku} ("iz" and "o" are prepositions, meaning *from* and *about*, and should be discarded as a functional word). Conceived like this, these two sets are different. For the best possible matching, chunks have to be normalised. This especially applies for highly inflectional languages, such as Serbian. In this specific case, Simple-Word lemmatisation within MWTs is needed. This means that each word from a MWT has to be replaced by a corresponding lemma from the available morphological e-dictionaries for Unitex (Krstev, 2008). For example, a word "reči" is a noun, has feminine gender, is in plural and is in nominative case. A lemma for any noun is singular, and is nominative case, namely "reč" for this case. The words "rečnika" and "rečniku" are also both nouns, but in genitive and dative case, respectively. After single-word lemmatisation, both of these words

are replaced with their lemma "rečnik". After this lemmatisation, for both chunks, set representations are {reč, rečnik} and {reč, rečnik}, and they, therefore, match.

A list of the resulting matched source and target terms $S(term) \leftrightarrow T(term)$, obtained from the aligned chunks, was retrieved as:

$$S(term) \leftrightarrow T(term) = \{(s,t) :$$
$$s \in S(term) \sim S(align.chunk) \wedge$$
$$t \in T(term) \sim T(align.chunk) \wedge$$
$$(s,t) \in (S(align.chunk) \leftrightarrow T(align.chunk)\}$$

5 Results and Discussion

The input preparation steps as well as processing consist of several components developed in C# and Python that are interconnected to work in a pipeline. The pipeline relies on existing tools for the extraction of English MWTs (ENG-TE) and Serbian MWEs (SERB-TE) implemented in LEXIMIR (Stanković et al., 2016) and on GIZA++ for word alignment, while all other components are newly developed.

In our experiments we combined each of the three following parameters, all related to the preparation of the input, where each parameter comes in two options, thus obtaining 8 different experimental settings:

1. The input domain aligned corpus (Input i) consists of:
   (a) the aligned corpus LIS-CORPUS;
   (b) the aligned corpus LIS-CORPUS extended with the bilingual aligned pairs BI-LIST (LIS-CORPUS+);
2. The list of domain terms for the source language (Input ii) is
   (a) the source language part of LIS-DICT including SWTs;
   (b) the output of the extractor ENG-TE applied to the source language part of the aligned input corpus;
3. The extraction of the set of MWTs in the target language by SERB-TE (Input iii) was done:
   (a) on the target language part of the aligned chunks (CHUNK);
   (b) on the target language part of the aligned input sentences (TEXT).

The summary of results obtained by our system for 8 experiment settings is given in Table 1. We refer to the experiments using the labels introduced above.

The numbers in the columns represent the following results:

Input and GIZA++ output results

A Number of entry pairs in LIS-DICT, i.e. English terms extracted by ENG-TE;

B Number of lines obtained from GIZA++ phrase table, after preprocessing steps;

C Number of distinct, lemmatised Serbian MWTs extracted from the target language part of the aligned chunks (for CHUNK) or from the target language part of the aligned input corpus (for TEXT).

Table 1. Numerical data that describes the results of the term extraction system

Experiment			A	B	C	I	II	III	IV
LIS-DICT	LIS-CORP	CHUNK	17,889	240,253	26,719	6,646	1,141	647	173
		TEXT			49,632		1,531	770	240
	LIS-CORP+	CHUNK		496,787	45,813	11,740	2,508	1,105	301
		TEXT			50,644		2,500	1,075	362
ENG-TE	LIS-CORP	CHUNK	3,169	215,317	35,226	5,063	2,233	x	x
		TEXT			49,632		2,233	x	x
	LIS-CORP+	CHUNK		446,979	44,885	8,164	3,333	x	x
		TEXT			50,644		3,310	x	x

Additional filtering of results obtained by GIZA++[11]

I Number of the aligned chunks after initial filtering using English terms (Processing ii): $(S(align.chunk) \sim S(term))$, where the list of English terms depends on the choice of parameter 1 (the English part of LIS-DICT or obtained from the corpus by using ENG-TE for extraction).

II Number of aligned chunks after subsequent filtering using Serbian terms (Processing iii) : $(S(term) \sim S(align.chunk)) \wedge (T(term) \sim T(align.chunk)) \wedge (S(align.chunk) \leftrightarrow T(align.chunk))$.

[11] To keep it simple, in the following notation, we refer to sets as to single representative terms, e.g. when we write $S(align.chunk)$, we refer to one term from that list.

III Number of new term pairs after filtering, namely those that do not already exist in LIS-DICT — these term pairs were obtained by selecting filtered chunks in which the Serbian part of the chunk does not match a term in the Serbian part of LIS-DICT ($(T(align.chunk) \not\sim T(term.list))$) (applicable only when LIS-DICT is used in the experiment);

IV Number of term pairs after filtering that already exist in LIS-DICT – these term pairs were obtained by selecting filtered chunks in which the Serbian part of the chunk matches a term in the Serbian part of ($T(align.chunk) \sim T(term.list)$) (also applicable only for (LIS-DICT) experiments).

In order to assess the efficiency of our approach, we have first evaluated all extracted pairs manually. Evaluation results showed that a number of new term pairs were retrieved. When LIS-DICT was used as a source of English terminology, 364 English terms from the dictionary were linked to new Serbian translations yielding 428 new term pairs. Among all term pairs retrieved using ENG-TE for extraction, 538 were supported by LIS-DICT, while among all term pairs retrieved using LIS-DICT for extraction, 168 were also retrieved with ENG-TE. A detailed evaluation procedure and results were described in (Šandrih et al., 2019).

6 BiLTe Web Application

In this Section, a Web application[12] that implements the proposed technique for terminology extraction is presented. The tool is freely available for online use.

The Web application consists of three modules: 1) input, 2) alignment and post-processing and 3) results module. Each module is briefly described and shown in the following Subsections.

Input Module

First, a user has to upload two sentence-aligned text files. Files must have the same names. File extensions should differ and indicate language (e.g. medicine.en and medicine.sr). These files are later fed into GIZA++.

Afterwards, a user has to upload a list of English terms. The first line should contain a header, and each line should contain one term.

[12] BILingual Terminology Extraction

Finally, a user has to upload a list of terms in Serbian (not necessarily MWUs). The first line is a header, each line contains a term and its frequency (for filtering later), separated with | ("pipe" character).

The interface of this module is displayed in Figure 3.

Figure 3. Input module of the BiLTe Web application

Alignment and Post-Processing Module

Aligning with GIZA++ yields a so called "phrase-table".

The alignment works in the following way. GIZA++ reads the two input texts in parallel. Whenever two bilingual chunks appear together, their co-occurrence is written into text file (dubbed *f_phrases*). Afterwards, *f_phrases* is sorted in two ways (by the target term and by the source term), and that's how two tables are obtained.

As earlier mentioned in Section 4, in step (ii), we discard candidates with direct or inverse probabilities lower than the threshold. After this, two post-processing steps follow. The first step is filtering by discarding terms that are out of the domain. This step is followed by a lemmatisation of English chunks with WordNet (Princeton WordNet, 2010) and Serbian chunks with e-dictionaries for Serbian (Krstev, 2008) (Procedure, i).

The interface of this module is displayed in Figure 4.

Figure 4. Pre-processing and Alignment module of the BiLTe Web application

Results Module

The basic steps of this module are: 1) keeping only candidates present in the English list (Procedure, ii), 2) performing intersection with Serbian extracted MWUs (Procedure, iii) and 3) additional filtering (optional) of bad candidates from the previous step.

The interface of this module is displayed in Figure 5.

Figure 5. The module for obtaining results of the BiLTe Web application

7 Conclusion

We conclude that the best results, in terms of quantity and quality of the obtained pairs, were achieved when input sentences were enhanced with additional bilingual pairs, and when extraction of Serbian terms was performed on the Serbian part of the aligned corpus, instead of aligned chunks. We will continue to experiment with these settings. Moreover, we intend to enrich BI-LIST with newly produced pairs. Our experiments also show that both methods of extraction produce some different pairs of equivalent terms. In our future work we will use not only both methods, when a dictionary for a source language becomes available, but also terms obtained from several different extractors. Another indented work is the integration of lemmatisation procedure into the bilingual extraction, already developed and implemented in monolingual MWU extraction, as described in (Stanković et al., 2016).

We intend to apply the same approach to other domains — mining, electric power system and management — for which aligned domain corpora have already been prepared. Of course, the enrichment of sentence-aligned domain-specific corpora, bilingual word lists and monolingual dictionaries of MWTs are long-term activities.

Acknowledgment

This research was partly supported by the Ministry of Education, Science and Technological Development through projects ON-178006 and III47003.

References

Arcan, Mihael, Marco Turchi, Sara Tonelli and Paul Buitelaar. "Leveraging Bilingual Terminology to Improve Machine Translation in a Computer Aided Translation Environment". *Natural Language Engineering* Vol. 23, no. 5 (2017): 763–788

Baldwin, Timothy and Su Nam Kim. "Multiword Expressions". *Handbook of Natural Language Processing* Vol. 2 (2010): 267–292

Bouamor, Dhouha, Nasredine Semmar and Pierre Zweigenbaum. "Identifying Bilingual Multi-Word Expressions for Statistical Machine Translation". In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard et

al.. Istanbul, Turkey: European Language Resources Association (ELRA), 2012

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch et al. "Multiword Expression Processing: A Survey". *Computational Linguistics* Vol. 43, no. 4 (2017): 837–892

Cram, D. and B. Daille. "Terminology Extraction with Term Variant Detection". In *Proceedings of ACL-2016 System Demonstrations*, 13–18, 2016.

Fawi, F. and R. Delmonte. "Italian-Arabic Domain Terminology Extraction From Parallel Corpora". In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, Vol. 130, Accademia University Press, 2015

Garabík, Radovan and Ludmila Dimitrova. "Extraction and Presentation of Bilingual Correspondences from Slovak-Bulgarian Parallel Corpus". *Cognitive Studies | Études cognitives* no. 15 (2015): 327–334

Hakami, H. and D. Bollegala. "A Classification Approach for Detecting Cross-lingual Biomedical Term Translations". *Natural Language Engineering* Vol. 23, no. 1 (2017): 31–51

Hamon, T. and N. Grabar. "Adaptation of Cross-lingual Transfer Methods for the Building of Medical Terminology in Ukrainian". In *Proceedings of the* 17^{th} *International Conference on Intelligent Text Processing and Computational Linguistics (CICLING2016)*, LNCS. Springer, 2016

Hazem, Amir and Emmanuel Morin. "Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora". In *Proceedings of COLING 2016, the* 26^{th} *International Conference on Computational Linguistics: Technical Papers*, 3401–3411, 2016.

Heafield, Kenneth. "KenLM: Faster and Smaller Language Model Queries". In *Proceedings of the* 6^{th} *Workshop on Statistical Machine Translation*, 187–197. Association for Computational Linguistics, 2011.

Hewavitharana, Sanjika and Stephan Vogel. "Extracting Parallel Phrases from Comparable Data for Machine Translation". *Natural Language Engineering* Vol. 22, no. 4 (2016): 549–573

Irvine, Ann and Chris Callison-Burch. "End-to-end Statistical Machine Translation with Zero or Small Parallel Texts". *Natural Language Engineering* Vol. 22, no. 4 (2016): 517–548

Koehn, Philipp, Franz Josef Och and Daniel Marcu. "Statistical Phrase-based Translation". In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Vol. 1, 48–54. Association for Computational Linguistics, 2003.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico et al.. "Moses: Open Source Toolkit for Statistical Machine Translation". In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics, 2007

Kontonatsios, G., M. Claudiu, Korkontzelos I., Thompson P and S. Ananiadou. "A Hybrid Approach to Compiling Bilingual Dictionaries of Medical Terms from Parallel Corpora". *Statistical Language and Speech Processing* Vol. 8791 (2014): 57–69

Krstev, Cvetana. *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade, 2008. https://hal.archives-ouvertes.fr/hal-01011806

Krstev, Cvetana, Branislava Šandrih, Ranka Stanković and Miljana Mladenović. "Using English Baits to Catch Serbian Multi-Word Terminology". In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, chair), Nicoletta Calzolari (Conference, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al., 7–12. Paris, France: European Language Resources Association (ELRA), 2018. http://www.lrec-conf.org/proceedings/lrec2018/pdf/384.pdf

Lahbib, W., I. Bounhas and B. Elayeb. "Arabic-English Domain Terminology Extraction from Aligned Corpora". In *On the Move to Meaningful Internet Systems (OTM 2014) Conferences, Confederated International Conferences : CoopIS, and ODBASE 2014, Amantea, Italy, October 27-31, 2014, Proceedings*, Robert Meersman, Tharam Dillon Michele Missikoff Lin Liu Oscar Pastor Alfredo Cuzzocrea & Sellis Timos, Hervé Panetto, 745–759. Springer Berlin Heidelberg, 2014,

Naguib Sabtan, Yasser Muhammad, "Bilingual Lexicon Extraction from Arabic-English Parallel Corpora with a View to Machine Translation". *Arab World English Journal* Vol. 7, no. 5 (2016): 317–336. http://search.ebscohost.com.proxy.kobson.nb.rs:2048/login.aspx?direct=true&db=edb&AN=115896070&site=eds-live

Och, Franz Josef and Hermann Ney. "Improved Statistical Alignment Models". In 38th *Annual Meeting on Association for Computational Linguistics*, 440–447. Stroudsburg, PA: Association for Computational Linguistics, 2000.

Oliver, Antoni. "A System for Terminology Extraction and Translation Equivalent Detection in Real Time: Efficient use of Statistical Machine

Translation Phrase Tables". *Machine Translation* Vol. 31, no. 3 (2017): 147–161

Pianta, E., C. Girardi and R. Zanoli. "The TextPro Tool Suite". In *Proceedings of 6th edition of the Language Resources and Evaluation Conference*, 2008

Pinnis, Marcis, Nikola Ljubešic, Dan Stefanescu, Inguna Skadina, Marko Tadic et al.. "Term Extraction, Tagging, and Mapping Tools for Under-resourced Languages". In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, 20–21. 2012

Princeton WordNet, 2010

Semmar, Nasredine. "A Hybrid Approach for Automatic Extraction of Bilingual Multiword Expressions from Parallel Corpora". In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, chair), Nicoletta Calzolari (Conference, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al.. Paris, France: European Language Resources Association (ELRA), 2018

Spasić, Irena, Mark Greenwood, Alun Preece, Nick Francis and Glyn Elwyn. "FlexiTerm: a Flexible Term Recognition Method". *Journal of Biomedical Semantics* Vol. 4, no. 1 (2013): 27. https://doi.org/10.1186/2041-1480-4-27

Stanković, Ranka, Cvetana Krstev, Nikola Vulović and Biljana Lazić. "Biblisha", 2014

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić and Aleksandra Trtovac. "Rule-based Automatic Multi-word Term Extraction and Lemmatization". In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard et al. Paris, France: European Language Resources Association (ELRA), 2016

Stanković, Ranka, Cvetana Krstev, Duško Vitas, Nikola Vulović and Olivera Kitanović. *Keyword-Based Search on Bilingual Digital Libraries*, 112–123. Cham: Springer International Publishing, 2017. http://dx.doi.org/10.1007/978-3-319-53640-8_10

Tsvetkov, Yulia and Shuly Wintner. "Extraction of Multi-word Expressions from Small Parallel Corpora". In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 1256–1264. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. http://dl.acm.org/citation.cfm?id=1944566.1944710

Vintar, Špela and Darja Fišer. "Harvesting Multi-Word Expressions from Parallel Corpora". In *Proceedings of the 6th International Conference on*

Language Resources and Evaluation (LREC 08), Marrakech, Morocco: European Language Resources Association (ELRA), 2008. http://www.lrec-conf.org/proceedings/lrec2008/

Šandrih, Branislava, Cvetana Krstev and Ranka Stanković. "Two Approaches to Compilation of Bilingual Multi-Word Terminology Lists from Lexical Resources". *Natural Language Engineering*, 2019

Xu, Yan, Luoxin Chen, Junsheng Wei, Sophia Ananiadou, Yubo Fan et al.. "Bilingual Term Alignment from Comparable Corpora in English Discharge Summary and Chinese Discharge Summary". *BMC bioinformatics* Vol. 16, no. 1 (2015): 149

Food as Text

UDC 811.163.41'322.2 811.163.4'374
DOI 10.18485/infotheca.2019.19.2.7

ABSTRACT: This paper aims to describe the initial steps in building a diachronic and interdisciplinary corpus that would present different aspects of the language of food in the Serbian language. This corpus consists of culinary recipes, ethnographic and anthropological studies and other different testimonies about food in Serbia dating from the second half of the 19th century onward. Such corpus offers insights into eating habits in Serbia and how they changed under different cultural influences. The problems of automatic processing of this corpus are discusses, some experiments that rely on the use of existing lexical resources are presented, and directions for future work are given.

KEYWORDS: domain corpus, culinary, food, computing, electronic dictionary.

PAPER SUBMITTED: 19 December 2019
PAPER ACCEPTED: 25 December 2019

Duško Vitas
vitas@matf.bg.ac.rs

University of Belgrade
Faculty of Mathematics
Belgrade, Serbia

Translated from Serbian:
Ana Popović

1 Introduction

This paper aims to describe the initial steps in building a corpus that would present different aspects of the language of food in the Serbian language. Food is talked about from various perspectives, from the daily preparation of meals to the consideration of the nature of food from the point of view of a number of scientific disciplines or gastronomic accounts. This language, due to its everyday use, seems at first sight quite ordinary. Yet when considered over an extended period of time, it is possible to identify complex layers of different cultural and socio-economic influences to which it is a testament.

The fact that the language of food has been neglected is illustrated, among other examples, by the fact that in the Dictionary of the Serbian Academy of Sciences and Arts (RSJ, 2011), in the twenty volumes published

to date, only about fifty entries have the usage label *kuv.* meaning "kuvarski termin, kulinarstvo" ("culinary term, cooking"). In the dictionary's last published volume, for instance, the entry for the word *pasulj* (*beans*) is labelled as a botanical and agronomic term, but not a culinary one. The definition of the primary meaning says "...leguminous plant... used in nutrition as a vegetable, the seed of this plant, a dish made with this plant, *grah*". It should be noted that there are two distinct meanings here: *leguminous plant* and its *seed* are botanical and agronomic concepts, while a *dish* made with this plant is a culinary concept. Moreover, most examples given for the primary meaning refer to the dish,[1] and only one each to its botanical and agronomic meanings. As further testimony to the complexity of the term, I will cite a personal experience – namely, I recently attended a three-hour scholarly discussion between biologists, chemists, physicochemists, doctors and agronomists on the properties of leguminous plants, whose most important representative is the *bean*. Only once did the discussion on the complex biochemical properties of this plant family touch on *beans* as a dish (in the form of *prebranac*).

Traditional lexicographical studies of the language of food in Serbian are based either on ethnolinguistic studies or on the excerption of culinary terms primarily from existing dictionaries. Of the extensive body of literature on the subject, we will cite here only the most recent. One of the ethnolinguistic works describing the culinary vocabulary of a particular geographical region in great detail, but without a replicable and precise specification of the corpus used in the study, is (Mirilov, 2016). A different approach has been used in (Radonjić, 2016), where culinary vocabulary is described based on descriptions in the Matica Srpska Dictioinary (RSJ, 2011), also drawing on other dictionaries as well as cookbooks. As the relevant material does not include appropriate labels, the excerption had to be done "by hand" (cf. footnote 13) and, inevitably, partially. Both approaches deal with the "ordinary" names of dishes, although the concept of "ordinary" or common is very hard to determine without an adequate referential corpus.

In addition to lexicographical and ethnolinguistic sources about the language of food, numerous other sources describe or testify to dietary habits from different perspectives – ethnological, anthropological, historical, sociological, philosophical... Although viewed from different perspectives, the

[1] A search of the Corpus of Contemporary Serbian Language confirms that the entry refers primarily to the dish rather than the plant (http://www.korpus.matf.bg.ac.rs/korpus/).

described object remains the same – foodstuffs, dishes and ways of preparing them. The paper aims to explore the possibilities of forming an electronic corpus that would present these various views of food and enable not only a lexical search, but also a search that would include historical and geographical information on the origins of the vocabulary used, and also about its evolution. The fluctuations in the occurrence of the word *bulgur* can serve as an example. It appears in (Тројановић, 1983),[2] where it says that "in our country...*bungur* [3] is also highly prized", but it does not occur even once in Midžina's cookbook (Поповић Мицина, 1878). About half a century later, in (Марковић, 1959) there is no mention of *bulgur*, but (Zirojević, 2019) provides a detailed description of *bulgur* in a separate section entitled "Three (not) forgotten foods", as the food had become fashionable again, due to the influence of medical research.[4]

Such a corpus could help synthesise information scattered across different papers, dictionaries, studies and cookbooks, information that is not only linguistic and lexicographical, but also belongs to the domain of the humanities, e.g. ethnology, anthropology or history. Thus, the corpus could provide the basis for an encyclopedic study of the language of food in the Serbian language.

In addition to the collection of material, there is the question of instruments for its processing. On the one hand, these are lexical resources enabling the adequate indexing of the corpus, and on the other, a complex corpus administration system which, apart from standard functions, must also establish connections not only within the corpus material but also with external sources such as geographical maps or encyclopedic sources. Moreover, preliminary processing should include text normalization, in terms both of vocabulary and of other markers used (e.g. the system of measurement).

A part of these resources for Serbian have already been constructed over recent years through the development of specific dictionaries and grammars for the Unitex system[5] (Крстев and Лазић, 2015) and additions to culinary terms of the WordNet type semantics network for the Serbian language (Вујићић Станковић and Пајић, 2014). Part of the necessary material – a corpus of cooking recipes collected from the web – is described in (Vujičić-Stanković et al., 2014). This corpus gives partial insight into the

[2] The study was first published in 1898.

[3] A dialect variant of *bulgur*

[4] In the Corpus of the Contemporary Serbian Language, *bulgur* occurs ten times (all ten occurrences are after 2008).

[5] Grammar-based corpus processing suite

contemporary cooking vocabulary, but it lacks certain information, notably on the socio-cultural (e.g. rural – urban), geographical and historical origin of recipes. A more complex organization of culinary material can be found in (Stošić et al., 2017), where different culinary sources are organised as a multimedia document.[6]

2 The structure of the corpus

The current version of the language of food corpus comprises different types of written sources on culinary topics. In other words, the corpus consists of the following types of texts:

- terminological monolingual and multilingual dictionaries and glossaries (for instance, (Banićević and Popović, 2010), (Vukov, 1954);
- cookbooks in Serbian (such as (Поповић Мицина, 1878) or (Марковић, 1959)) or translated into Serbian, e.g. (Пелапрат, 1973);
- doctoral dissertations addressing culinary topics, e.g. (Mirilov, 2016);
- Serbian-language recipe collections collected from the web, including user comments if available;
- culturological, ethnographic and anthropological studies (such as (Montanari, 2011), (Тројановић, 1983), (Zirojević, 2019), (Радуловачки, 1996) and (Милорадовић, 2014));
- newspaper articles from general sources or specialised food magazines;
- a collection of menus (restaurant menus or brochures);
- monographs on food (e.g. (Mijo, 2012) or (Onfre, 2002));
- historical accounts (e.g. (Фотић, 2005));
- excerpts from literary works in which food is mentioned (e.g. excerpts from (Петрониje, 1976), (Игњатовић, 1949) or (Балзак, 1933)), and
- textbooks (e.g. (Портић, 2011))

This list provides insight into the heterogeneous character of the corpus, because the nature and inner organization of the sources described varies widely. Apart from being different in character, the texts are not equally accessible: certain types of sources described are not archived and are impossible to find. A case in point are restaurant menus testifying to changes in urban dietary habits: [7] a valid and temporally well spaced sample of such

[6] A students' multimedia document "Back Then Eating Was Good"

[7] (Голубовић, 2007) provides an exhaustive list of Belgrade hotels, restaurants and inns from the mid-19th century, but with hardly any mention of the dishes they served. One of the few descriptions cites an excerpt from Miloš Crnjanski's *Belgrade* (pp. 46-47).

documents is not available to date. The importance of these marginalised documents is explained in (Витас, 2018): the habitual diet of a certain urban population cannot be inferred from cookbooks. On the other hand, the language of food varies depending on the type of source. Descriptions range from the "algorithmicised" text of recipes or the simple language of cookbooks to sophisticated consideration of flavours or complex cultural-historical influences the result of which is a certain dish. This stylistic heterogeneity also raises the question of weighting, which would enable the valorization of certain types of views of the corpus.

The structure of the documents in the corpus is also heterogeneous. Some texts have a normal, ordinary structure, while others require specific tagging. For one thing, the texts of recipes are non-uniform. Contemporary recipes, collected from recipe websites, tend to have a precise structure comprising the name of the dish, important ingredients, required quantities, time necessary for the preparation of the dish, etc., followed by the preparation procedure (sometimes broken down into steps). Older recipes, on the other hand, tend to state the name of the dish and the preparation procedure which includes those parts that are stated separately in contemporary recipes. In Section 3, we discuss the possibility of automatically translating the old into the new structure. Menus are also specifically structured in the form of lists or lists of lists (when ingredients are listed).

In addition to the normalization of specifically structured documents, there is the issue of the lexical normalization of the corpus. Namely, lexical variations stem both from the distribution of regional or archaic words, and from the nonstandard way in which new culinary terms are adopted from foreign sources. This problem has been partly solved in (Vujičić-Stanković et al., 2014). Examples include the word *ožica* in (Тројановић, 1983) for *kašika* (spoon), or *prevrtača* in (Ердељановић, 1908) for *palačinka* (crepe/pancake). In (Поповић Мицина, 1878) we find that different ways of *pohovanje* (frying in breadcrumbs/batter) are reduced to *prženje* (frying) (e.g. *pržene teleće nožice*), or that *musaka* referred to in (Ердељановић, 1908) is *modri patlidžan za 6 osoba* (aubergine for six persons). The tracing of terminological evolution in a corpus thus conceived means that cases such as these will also have to be included in the search.

Finally, in such heterogeneous sources, sections relevant for the corpus have been presented in different ways. Unlike the class of recipes, which give a straightforward description of a dish, in different types of studies recipes or language about the use of a particular food is inserted into a wider context that is not necessarily relevant. The separation of parts that constitute a

recipe from the parts that describe the processing of foods or flavours would have to be the subject of separate tagging which would need to be described.

3 Computational issues of processing cooking texts

A corpus conceived in such a way is not only important for ethnographic and lexicographical research, but also offers interesting material for different computational experiments which we will briefly describe here. These experiments would potentially lead to the defining of a formalised language of cooking, whose immediate application would be in programmable cooking robots,[8] and in query systems intended for assisting in food preparation.[9] Some of the tasks are applications that, apart from being standalone applications, could be used in the normalization and tagging of corpus texts.

Let us consider some of these tasks that would have to be resolved for the Serbian language in order for these applications to be built. One group of tasks involves supplementing cooking resources already developed for the Serbian language, and another regards systems for analyzing and transforming corpus texts.

The first group of required resources includes supplementing constructed cooking dictionaries which are partly described in (Vujičić-Stanković et al., 2014), (Стијовић et al., 2016). Specific regional, archaic and similar vocabulary has to be processed and added to the system of electronic dictionaries so that the corpus can be properly tagged. The vocabulary necessary for recognizing specific named cooking entities occurs here as a separate subsystem. Approximate measurements are analysed and discussed in (Krstev et al., 2014), but, in addition to these, there are hidden entities in the texts whose value has to be inferred (or checked) indirectly. Examples include information on preparation time, temperature etc., but also a group of names of ingredients, preparation procedures and dish names that include proper nouns in their basic or derived form (Krstev et al., 2019), (Вујићић Станковић and Пајић, 2015).

Semantic relationships, which are partly built into the system of electronic dictionaries for the Serbian language and also into the Serbian WordNet, would have to be significantly expanded with qualifiers such as "re-

[8] Cooking robots available today are capable of preparing dishes for which the procedure has been previously defined, but the user cannot instruct them to prepare his/her own recipe (Витас and Крстев, 2016).

[9] Such a system was proposed through IBM's *Chef Watson* system

gional", "archaic", etc., as well as data on energy value, allergens etc. including references to corresponding encyclopedic content. It should be noted that words of similar meaning can be used synonymously, but can also refer to a substitute for a certain ingredient, which represents a separate relationship within the semantic network. For example, *puter*, *buter*, *maslo* and *maslac* are part of the same WordNet synset,[10] but *margarin* belongs to a different synset. Due to their interchangeability, a separate relationship is needed to indicate possible ingredient substitution.

Multi-word units in the language of cooking, in addition to their usual meaning, can also be the name of a new culinary concept, particularly in terms of names of dishes. Their structure both at the level of ingredients and cooking utensils, and that of dishes, has not been the subject of separate lexicographical processing in Serbia. In the case of culinary innovations, which usually represent "unusual" dishes, the entire description of ingredients can appear as a lexical unit, for instance, *pohovani bri sa kulijem od šumskog voća* (breaded fried brie with forest fruit coolie) or *pate od čvaraka sa musom od kozijeg sira* (pork rillette with goat cheese mousse). This phenomenon of a lack of a "standard" lexeme for a certain dish occurs in other languages as well, as shown by analyses provided in (Gerhardt et al., 2013).

An extremely complex lexicographical issue is the question of multilinguality. Namely, WordNet, through a system of interlingual indices, enables the pairing of equivalent concepts. But the question of cooking recipes in a certain culture is beyond the possibilities of equivalence description at the synset level. It is a specific transfer from one language to another that depends to a great extent on the local (or national) system of food concepts. For example, *musaka*[11] is a Balkan concept that exists in all European languages, but the local manner of preparation varies to such an extent that it is not always certain that it refers to the same type of dish. On the other hand, there are conceptually similar dishes prepared with somewhat different ingredients outside the Balkan region (e.g. Italian *lasagna* or French *hachis parmentier*). This raises the question of identification of dishes where the preparation procedure is very similar although both the ingredients and the names of dishes vary between languages. One solution is connecting the

[10] A synset groups together words with similar meanings into a semantic network graph node. This node is connected to other synsets by semantic relations.

[11] In (Courtine, 1986) *musaka* is defined as a Balkan dish of alternating layers of aubergine slices and minced lamb, usually with a coating white sauce. Unlike this simplified presentation, (Марковић, 1959) gives more than 30 different recipes for this dish.

name of the dish to the family of recipes it is represented by, and then establishing equivalence between similar procedures in different languages. Such a procedure could lead to the establishment of an abstract schema of dishes that would enable lexical transfer from one language into another.[12] Nevertheless, a specific restriction on the internationalization of cuisine, despite culinary globalization, are local ingredient availability as well as local culinary techniques and gastronomic habits. As an example, *baget(a)* can be bought in most Belgrade bakeries but, apart from its name, it has little in common with the French *baguette*. Another issue is establishing a connection, whenever possible, between a certain dish and its historical and geographical origins. In (Поповић Мицина, 1878), for instance, we can find examples such as *teleće pržoljice* (*Kalberne Schnitzel*, *teleće šnicle*), this being the only occurrence of the word *šnicla* in the cookbook. The word *pržoljica*, however, has a broader meaning, because there are also *ovnujske pržoljice – Cotelette – rebarca*. In this case we can see the German and French influences on the forming of contemporary culinary names (*šnicla*, *kotlet*). Interestingly, none of the three names (*pržoljica*, *šnicla*, *kotlet*) occur in either (Тројановић, 1983) or (Ердељановић, 1908), while (Марковић, 1959) lacks only *pržoljica*: *šnicla* and *kotlet* had probably displaced *pržoljica* as a result of culinary refinement.

In addition to these requirements in the description of lexical structures and relations in culinary language, it is possible (partly starting from existing resources) to develop applications necessary for corpus normalization. Among such applications are programs that analyse possible combinations of ingredients, taking into account the frequency of their co-occurrence. For example, *breskva* (peach) and *ananas* (pineapple) are only very rarely connected to *so* (salt). A similar application would be one that analysed similar (or identical) recipes where the same name can involve different ingredients and procedures and vice versa, where different names are used for the same dish made from identical (or similar) ingredients and prepared using the same procedures.

Useful applications that would improve existing searchable recipe collections are numerous, and we will mention them only as an idea for some interesting computational experiments:

[12] Conceptual equivalence of two dish names in different languages might be established by identifying a similar procedure using different ingredients.

- an application that determines the level of difficulty of preparing a certain recipe based on an analysis of the operator (cooking verbs) and ingredients;
- an application that determines the level of difficulty of preparing a certain recipe based on an analysis of the operator (cooking verbs) and ingredients;
- an application that determines the appropriate season for preparing a recipe based on the seasonal availability of ingredients;
- an application that automatically determines the course (starter, main course...);
- an application that identifies recipes appropriate for a certain diet or the risks of consuming a certain dish or ingredient (e.g. tolerance of pungency on the Scoville scale is subject to individual and regional differences).

4 On the processing of the corpus and initial results

A corpus thus conceived transcends the possibilities of the usual ways of processing in view of the complex structure of heterogeneous documents and the nature of possible queries. In other words, a search that, in addition to key words, involves other resources – dictionaries and semantic networks, metadata and different inferred or external data – necessitates a system that, in addition to a corpus processing function, enables other types of analysis. In the experimental phase of constructing such a corpus, a software solution could involve the application of the Unitex (Paumier, 2016) system in the stage of corpus preprocessing and tagging, and a system such as TXM[13] for storing and searching texts (Jaćimović, 2019).

We will limit ourselves here to the initial stage of preparation: the normalization and tagging of texts for the future corpus. This stage involves the intensive use of the above-mentioned resources developed for Unitex. Most collected texts have been scanned and read (by means of OCR), using the system described in (Krstev and Stanković, 2019) (Krstev, švalje), and they have been semi-automatically corrected including minimal TEI-tagging (in accordance with the requirements of the EITec project).[14]

As part of a preliminary analysis of possible applications of the corpus, we will examine relations between two groups of texts, one that looks at the

[13] The TXM package, intended for textometry, in addition to a corpus processing system (IMS CQP) also integrates other means of analyzing textual data, such as the R system.

[14] Distant Reading for European Literary History (COST Action CA16204)

phenomenon of food from the perspective of ethnology and anthropology, and another consisting of cookbooks and testifying to culinary practices in a certain period. The texts selected for the experiment are listed in Appendix 6, and hereafter we will refer to them using the letters and numbers assigned to them in the Appendix. These texts cover the period from the late 19th century to the early 21st century, and they provide accounts of dietary habits of the period when they originated and of the environment they describe. Texts A1-A4 describe dietary habits in rural areas, while others testify to urban culinary practices. The original year of publication, and basic data about the size of individual texts in terms of number of words, the number of different words and unrecognised words (as analysed by an electronic dictionary system) are given in Table 1).

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**B1**	**B2**	**B3**
year	1896	1908	1908	1996	2014	2018	1878	1915	1959
total	33,913	19,266	13,899	24,914	74,633	43,189	121,531	83,08	399,358
different	9,220	4,477	3,302	7,964	18,648	12,032	9,367	1,709	16,542
repetition	3.68	4.3	4.21	3.12	4.00	3.56	12.98	4.86	24.14
err	956	338	480	442	479	1,189	1,107	79	175

Table 1. Quantitative data on sources

The relationship between the total number of occurrences of simple words and of different simple words shows that the level of repetition in cookbooks is high, which indicates that the basis of the language of cooking – ingredients and procedures – is very limited. The number of unrecognised words in the err category is extremely low for B2 and B3 because these are texts that have been thoroughly processed, and the vocabulary entered into the electronic dictionary. The character of other unrecognised words varies: they come partly from foreign languages (in transcribed or non-transcribed form), but mostly from specific local vocabulary, primarily in the names of dishes, and to a lesser extent from non-standard variations in the names of ingredients, procedures or dishes. Table 2 shows the most common unrecognised words or frequent forms that have not been included in the appropriate inflectional

class of the entry, and their frequency. In A1, for instance, in this stratum we come across the word *supraška* (meaning *vreo pepeo* [hot ashes]) in different forms 18 times, but also the sibilarised stem *surutci* eight times (while other forms are included in the pattern <*surutka*>).

Looking at the unrecognised words in Table 2, we notice oppositions in terms of rural-urban, archaic-modern and also regional differences. On the other hand, we note that the context of the unrecognised words, if they occur in multiple sources, indicates different meanings. Thus *vodnjika* in A1 refers to an old drink, like *jabukovača* [apple brandy], so common, according to the author, that it does not need a description; in A2 it is made from pears, and in A4 it is the same as *jabukovača*; it does not occur in other sources. This confirms that along with the names of dishes (and drinks) it is also necessary to state the procedure, because different products are labelled with the same name. Also, linking these sources to the literary part of the corpus makes it possible to interpret the literary content. For example, in Rastko Petrović's novel "Burleska gospodina Peruna, boga groma" we come across both *vodnjika* and *supraška* from A1.

Let us now look at the distribution of some culinary tags in electronic dictionaries described in (Крстев and Лазић, 2015). Let us take as an example the occurrence of plant names (as foodstuffs and as dishes) in certain sources tagged in the e-dictionary as *Bot* and *Food*.

In A1 grains predominate (*žito* (36) [cereal], *pšenica* (24) [wheat], *ječam* (21) [barley], *ovas* (15) [oat], *raž* (7) [rye], *proso* (6) [panicgrass], *krupnik* (5) [spelt], ...), *kukuruz* (17) [corn] and *bundeva* (20) [pumkin], adding up to over half of the total number of words thus tagged. Fruit and vegetables, such as *kupus* (9) [cabbage], *luk* (8) [onion], *pasulj* (5) [bean], *pirinač* (3) [rice], *krompir* [potato] (3),[15] *jabuke* (3) [apple] i *šljive* [plum] (2), are rarely mentioned. This disregard for vegetables in A1 probably stems from the author's aim to describe the simplest, most traditional diet that could still be attested in the late 19th century.[16] Unlike in A1, in A2 vegetables take precedence over grains: in the text we find *luk* (*beli* [garlic] and *crni* [onion]) (71), *paprike* (45) [pepper], *kupus* (38) [cabbage], *krompir* (23) [potato], *pasulj* (22) [bean], *zelje* (18) [garden patience], *tikve* (but not *bundeve*) (18), *boranija* (11) [green bean], *patlidžan* (*crveni* and *zeleni* = *paradajz* [tomato]) (11), *pečurke* and *gljive* (10) [mushroom], *krastavac* (10) [cucumber], *pirinač*

[15] A1 states that in the mid-nineteenth century potato is cultivate "rarely, and more from curiosity then necessity".

[16] "For these reasons, we shall describe more thoroughly the simple foods of our people, and thus also help westerners to contemplate their past through us."

A1	A2	A3	A4	A5	A6	B1	B2	B3
supraška (18)	kalenica (28)	(is|na|pre|u) križati (49)	komlov (6)	pomander (28)	dirhem (10)	(za|u|0) prigati (239)	(za|u) prigati (40)	"A la..." (6)
kačkaval (17)	prazi luk (16)	(po|is) prigati (46)	trgančići (5)	fish and chips (25)	kabak (8)	rem (118)	zapraška (15)	drinks (5)
jagurt (11)	izmeljati (14)	bungur (12)	uslatko (4)	fast-food (9)	fasulj (7)	zemičkin (104)	rem (10)	(lon|šort) drinks (4)
varica (9)	gruvanica (8)	nešesta (12)	saćurica (4)	garum (7)	patlidžajn (6)	buavan (90)	gris (9)	prevru (2)
kanavac (8)	(na|is) križati (8)	ovlaš(e) (11)	švargl (3)	trufa (7)	obiber-ćufter (6)	parme-isati (81)	zanski (9)	kvascom (2)
surutci (8)	kupusnik (5)	mohuna (10)	ukiselo (3)	princes-krofna (4)	pasuljica (6)	morunin (46)	obiberi-sati (6)	vips (2)
bungur (7)	vodnjika (5)	ćeškek (9)	vanil-krancle (3)	taan (3)	piper (6)	brkljača (41)	šerpenjica (4)	češnjom (2)

Table 2. The most frequently unrecognised words in the samples

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**B1**	**B2**	**B3**
total	33,913	19,266	13,899	24,914	74,633	43,189	121,531	8,308	399,358
<Bot+Food>	335	438	307	535	1466	1299	3252	263	13736
%	1%	2.3%	2.2%	2.1%	2%	3%	2.68%	3.2%	3.5%

Table 3. The share of plant foodstuffs in the samples

(10) [rice]... and the following fruit: *šljive* (25) [plum], *grožđe* (10) [grape], *jabuke* (8) [apple], in addition to a number of other fruit and vegetables with lower frequencies. In A3, although it was published in the same period as A2, there are several notable differences. Here also *luk* [onion] occurs with the highest frequency (54), followed by *patlidžan* (*crveni* and *modri* [tomato and aubergine]) (23), *kupus* [cabbage] (20), *pšenica* (18) [wheat], *zelje* (16) [garden patience], *pirinač* (13) and *oriz* (11) [rice], *bungur* (12) [bulgur], *kopriva* (10) [nettle], *krompir* (10) [potato], *tikve* (10) [pumpkin], and then *grah* [bean], *mohuna* (=*zeleni grah*) [bean] and *buranija* (together 24) [green bean], *bamje* (6) [okra], *paprika* (4) [pepper] and other less frequently mentioned vegetables. More frequent fruit are *orah* (7) [walnut], *grožđe* (6) [grape] and *jabuka* (5) [apple]. In A4, published a century after the previous sources, and in another region, a change in the frequency of plant foodstuffs is noticeable. Here, *krompir* [potato] has the highest frequency (56), followed by *luk* (39) [onion], *kupus* (31) [cabbage], *paprika* (31) [pepper], *pasulj* (30) [bean], *žito* (27) [cereal], *bundeve* and *tikve* (which are different, taken together 21 [pumpkin and squash]), *paradajz* (17) [tomato], *kukuruz* (16) [corn], *tikvice* (9) [zucchini], *zelje* (7) [garden patience], *boranija* (6) [green bean]... while the most frequent fruit are *jabuke* (30) [apple], *grožđe* (29) [grape], *višnje* (23) [sour cherry], *šljive* (20) [plum], etc.

In contrast to these ethnographic descriptions of plant foods in rural areas, in the anthropological study A5, as a result of the internationalization of urban cuisine, other plants also appear, displacing to a greater or lesser extent those used in rural areas. Thus, apart from *krompir* (77) [potato], we can find *paradajz* (76) [tomato], *bosiljak* (65) [basil], *špargle* (59) [asparagus] and *masline* (57) [olive] as a direct association with Mediterranean cuisine, or *bob* (62) [broad bean], *avokado* (52) [avocado], *jagode* (53) [strawberry], and *kesten* (49) [chestnut] as a testament to new sophisticated or discriminating

tastes. Plants mentioned in the previous sources (A1-A4) have been pushed into the background: *paprika* (53) [pepper], *kukuruz* (39) [corn], *patlidžan* (37) [aubergine], *kupus* (30) [cabbage], *luk* (30 generally as *beli* or *crveni*) [onion and garlic]... while *pasulj* [bean] occurs only 9, and *bundeva* [pumpkin], *tikvice* [zucchini] and *zelje* [garden patience] only 2-3 times. In A6, which gives an account of the history of foods, most frequent are *krompir* (203) [potato], *kukuruz* (155) [corn] and *paprika* (126) [pepper] although they are all imports from Central and South America. Among the most frequent foods are also *pasulj* [bean], in various forms (*grah*, *boranija*, *buranija* – 205 times in total), *patlidžan* (94) [aubergine], *tikva* and *bundeva* (82) [pumpkin and squash], *paradajz* (71) [tomato], *luk* (62) [onion], *heljda* (62) [buckwheat], *spanać* (59) [spinach], *pirinač* (26) [rice].

In cookbooks, B1-B3, the plant base comprises *luk* (*beli* [garlic] and *crni* [onion]), *krompir* [potato], *prininač* [rice], *pasulj* [bean] and *kupus* [cabbage], while the percentage of *paprika* [pepper] and *paradajz* [tomato][17] has changed significantly between B1 and B3, as shown in Table 4.

	luk onion	**krompir** potato	**pirinač** rice	**kupus** cabbage	**paprika** pepper	**pasulj** bean	**patlidžan** aubergine	**paradajz** tomato
B1	328	160	111	93	65	61	46	0
B2	47	31	24	29	28	31	31	3
B3	1901	728	437	442	61	203	198	604

Table 4. The most frequent plant ingredients in cookbooks

The listed frequencies actually indicate the ratio of foods within a certain source. Thus *krompir* [potato], almost entirely neglected in A1, has greater frequency than any other plant foodstuff in A5 and A6, and this is also corroborated by the most recent sources (Витас, 2018). In cookbooks, which testify to urban cuisine, potatoes are already – besides onions[18] – the most

[17] *Paradajz* [tomato] is mentioned in B1 only as *crveni patlidžan*, while in B2 both names can be found.

[18] According to (Вивијен, 2008), "the onion,... which is usually served raw, is the plague of Serbian cuisine (p. 179).

important plant food in B1 in the late 19th century. It is interesting to note the evolution of names as a result of observed differences between foods. In A2 *patlidžan* always refers to *paradajz* [tomato], while in B5 it is the generic name for *modri patlidžan* [aubergine] and *paradajz* [tomato]. In A5 it is used in its present meaning (aubergine), while in A6 it is distinguished by colour: green, blue, black and red. It is also uncertain what *bundeva* and *tikva* refer to in certain authors, that is, whether they denote the same plant [pumpkin and squash]. These examples suggest that the names of foodstuffs, just like the names of dishes, are both subject to change over time and geographically determined, and sometimes insufficiently defined.

In the preparatory phase, relying on e-dictionary tags, it is possible to tag texts using morphological grammars in Unitex. One such grammar is shown in Figure 1.

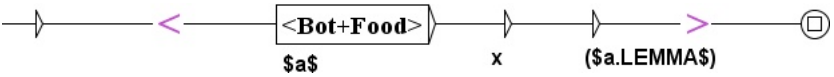

Figure 1. Morphological grammar

.

This finite-state automaton finds in a text the forms of words tagged as Bot and Food in the dictionary, places them in the `$a$` variable, and then adds to each form the name of the grammar by which it has been recognised (`BOT`), the lemma (`$a.CODE.LEMMA$`) and the semantic codes in the entry's field of syntactic-semantic properties (`$a.CODE.SEM$`). An extract of the concordances of different plant names in A6, generated by the grammar, is given in Figure 2.

Similar grammars have been formed for other features in the dictionary such as tags for courses (`Course`), drinks (`Drink`), ingredients (`Ing`), meals (`Meal`), approximate measures (`MesApp`) and utensils (`Uten`). These grammars have been collected into the TAG automaton (Figure 3) which then tags in the text those words that have any of these features.

An example of the application of these grammars to the A6 sample (p. 154) gives the following tagged sequence:

> Tu se daje i recept, pa tako saznajemo još da se boranija (`boranija, Bot+...+Food`) tada pripremala od spanaća (`spanać, Bot+...+Food+DOM=Culinary`), pirinča (`pirinač, Bot+...+Food`),

re svega, jelima s kukuruzom, pasuljem, amarantom i žalfijom.{S} I danas je kukuruz u Meksiku g
mu se dodaju razni začini (nana, čubar, anis, mirođija, lovor) koji olakšavaju varenje.{S} Uz t
lenim mahunama (lubija sabz), s kiselim artičokama (kangar) i s vrganjima (karč), uz obavezni d
amošnjih oblika i značenja: artićočina, artićok, antričok, antrićok, ratićokovina, ratićok, art
ričok, antrićok, ratićokovina, ratićok, artišoka, artičouka, artičoha i dr. {S}Nalazio se na me
), bundevama, paprikama, paradajzom, sa avokadom i mesom kojim su raspolagali (uključujući glod
iška ljuta turšijara, paradajz paprika, babura, kurtovka, somborka, slonova surla... {S}Za ljut
k od krasta i od uhobolje, dok je gorki badem sa šećerom lek za one koji pljuju krv.{S} Živi jo
d crvenih i crnih patlidžana, paprike i bamije.{S} Crni i crveni patlidžani ulazili su, uz luk
raši i tako dalje{S} Dodaje se, obično, bamiji, pasulju, boraniji, patlidžanu, krompiru, kupusu
crveni patlidžani ulazili su, uz luk i bamiju, i u posnu papazjaniju.{S} A poput paprike, i fr
u, pa čak i ispod palmi i kraj plantaža banana.{S} I dok u razvijenim zemljama njegova proizvod
dobro i dosta crna luka, i sa glavicom bela luka i dve tri ljute paprike, mete u lonac da se s
korijandera, kumin, aleva paprika, so, biber (uz druge začine, brašno i prašak za pecivo).{S}
ompleks „koji čine kukuruz, pasulj (ili bob) i tikvice (ili bundeve); ove poslednje leže u podn
ture koja počinje da zamenjuje zatečeni bob, od čijih su se zrna gotovile „lepe posne pihtije".
lim mlekom. {S}Za Vuka Karadžića (1818) boranija je „die noch grtinen Fisolen (lat. phaseoli vi
} U Sremu, uz pekmez od šljiva, jabuka, bresaka, ringlova, drenjina, <!-- p n = 144 --> grožđa,
ečenka, kuhanica – pominju se i strana: bundeva, dulek, jurget, kabak, misirača.{S} Ime ove pos
jne, patlidžarnik, pa pečeni sus luk, i buranija.{S} A ova poslednja spravljala se tako što se
zovu na čeruiš.{S} Pripremaju se kao i buranija od tikve." <!-- p n = 155 --> I danas u Bosni
kuhinji bilo je sedam jela u kategoriji buranije: sa spanaćem (esfenađ), s tikvicama (kadu), s
buranija. {S}A priča počinje upravo sa buranijom, tačnije, sa Buran bint Hasan ibn Sahl, ženom
i sam navodi tri: buraniju od tikvica, buraniju od mahuna i buraniju od poriluka. {S}Pišući u
ističe Alija Lakišić, i sam navodi tri: buraniju od tikvica, buraniju od mahuna i buraniju od p
an, na salatu, a ponajviše u tepsiju na buraniju, ili kako mnogi zovu na čeruiš.{S} Pripremaju
raniju od tikvica, buraniju od mahuna i buraniju od poriluka. {S}Pišući u Tradicionalnoj kuhinj
u Bosni Luka Grđić Bjelokosić pominje i buraniju. „Pravi se ponajviše od tikve, a može se način
ak luka, paradajza, gljiva, šargarepe i celera, naziva našim imenom zakuska. {S}Budući da u pot
o jelo pasuljica (sa šećerom, orasima i cimetom). {S}Ova mahunarka koristi se u mnogim lokalnim
kuhinji. {S}Može se koristiti zajedno s cimetom, kardamonom, korijanderom, karanfilićem, bibero
edoniji ovako: „Usitni se dobro i dosta crna luka, i sa glavicom bela luka i dve tri ljute papr
zma.{S} Budući da utiče na metabolizam, čaj od osušenih mahuna treba da ubrzava i mršavljenje.{
oga svrstava u grupu vitamina R Otuda i čaj od heljde, zbog svog visokog sadržaja rutina, utiče
bliku čaja za ublažavanje kašlja, kao i čaj od cvetova i lišća kod arterioskleroze, a svežim li
kamena, a zaustavlja i krvarenje.{S} Uz čaj od kukuruzne svile, koristi se i njen tečni ekstrak
koristili su se cvetovi heljde u obliku čaja za ublažavanje kašlja, kao i čaj od cvetova i lišć
sno da mu se dodaju razni začini (nana, čubar, anis, mirođija, lovor) koji olakšavaju varenje.{
a, drenjina, <!-- p n = 144 --> grožđa, dinja, kajsija, kuva se još i od šipaka.{S} Bez šećera
z od šljiva, jabuka, bresaka, ringlova, drenjina, <!-- p n = 144 --> grožđa, dinja, kajsija, ku
z „slatkiš od ukuvane zgusnute šire, od dudinja, grožđa ili drugog voća".{S} I za Đorđa Popović
, sitno izrezani komadi tikve i, retko, dunja.{S} Ovakvom „recelju" se tokom „pečenja" dodaje š
ene domaće varijante: elda, elsa, elja, eljda, jeda, jejda, jelda.{S} I Hrvati imaju heljdu, a
iku kod nas se već dugo koristi i naziv feferoni, a u novije vreme sve više se probija i čili,

Figure 2. Extract of the concordances obtained by the grammar in Fig. 1 from source A6

.

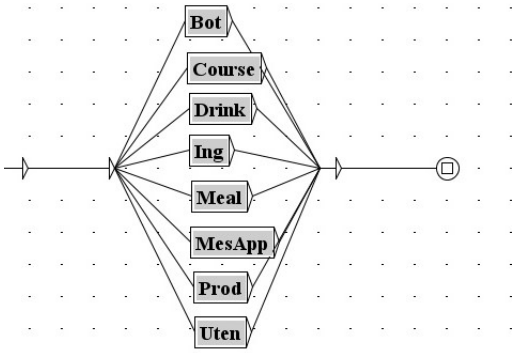

Figure 3. TAG automaton

.

> crnog luka (`crni luk, Comp+Bot+...+Food+Alim`) i maslinovog ulja (`PROD: maslinovo ulje,N,Comp+Conc+Prod+Food`). Servirala se ohlađena, s kiselim mlekom (`PROD: kiselo mleko,N,Comp+Ek+Conc+Food+Prod+Course`).[19]

The names of plant ingredients and foodstuffs have been recognised and they have been assigned word class tags (N) and semantic codes. Multi-word lexical units (e.g. *crni luk* [onion]) have been recognised. However, it can be seen from the text that *boranija* [green bean], apart from being the name of a plant, can also be the name of a dish: in Bosnia, *boranija* is the common name for different dishes (here from *spanać* [spinach] and other ingredients). Such examples suggest the need for tagged texts to be additionally analysed so that a certain meaning can be temporally and territorially specified.

We must note that, due to differences in the language of food, the corpus should be broken down into thematic micro-wholes instead of whole texts. Namely, in group A texts, besides documenting the language of cooking, there are significant parts that are "empty" in this respect. On the other

[19] Recipes are also given, and thus we find out that at the time, grean bean (`boranija, Bot+...+Food`) was prepared from spinach (`spanać, Bot+...+Food+DOM=Culinary`), rice (`pirinač, Bot+...+Food`), onion (`crni luk, Comp+Bot+...+Food+Alim`) and oliv oil (`PROD: maslinovo ulje,N,Comp+Conc+Prod+Food`). It was served cold, with soured milk (`PROD: kiselo mleko,N,Comp+Ek+Conc+Food+Prod+Course`).

hand, the results for a certain search key are scattered throughout a text. Figure 4 shows a histogram of occurrences of the noun *luk* in B3. Almost all the occurrences of this noun are in the first half of the text, and then about 3% towards the end of the text, in the section on preparing vegetable preserves. This indicates that the second half of the sample can be excluded from the search with this key. The breaking up of the text into micro-wholes would lead to a partitioning of the entire text into small sections which would have key words relevant to the section as their meta-data.

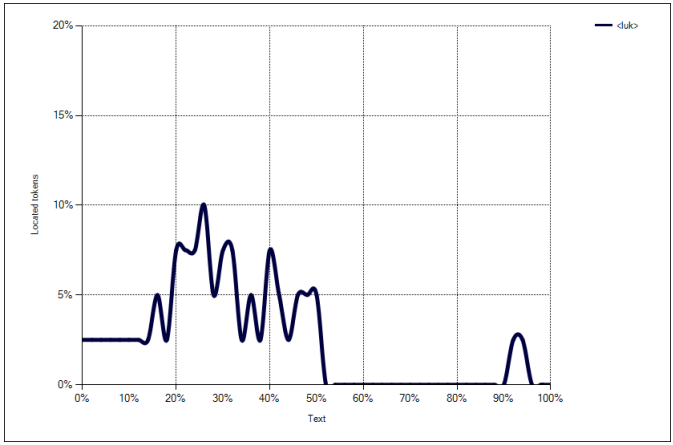

Figure 4. The distribution of occurrences of the noun luk in B3

.

5 Conclusion

The paper presents elements for the design of a corpus of the sub-language of food in Serbian and analyses some of the directions for its development, and also looks at the problems of text selection and annotation. It suggests possible applications which can be developed starting with the language documented in the corpus. In addition to providing searchable material for various disciplines, the construction of such a corpus would provide a basis for compiling an encyclopedia of dietary culture among the Serbian people,

and at the same time a means of transcending various mystifications about culinary traditions.

References

Banićević, Marta and Magdalena Popović. *Rečnik ugostiteljstva : (srpski-engleski-nemački-francuski-italijanski-ruski) [Hospitality industry dictionary: (Serbian-English-German-French-Italian-Russian)]*. Udruženje naučnih i stručnih prevodilaca Srbije, 2010

Courtine, Robert. *Dictionnaire de cuisine et de gastronomie*. Larousse, 1986

Gerhardt, Cornelia, Maximiliane Frobenius and Susanne (eds.) Ley. *Culinary Linguistics*. John Benjamins Publishing, 2013

Jaćimović, Jelena. "Textometric methods and the TXM platform for corpus analysis and visual presentation". *Infotheca - Journal for Digital Humanities* Vol. 19, no. 1 (2019): 30–54. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2019.19.1.2_en

Krstev, Cvetana and Ranka Stanković. "Old or new, we repair, adjust and alter (texts)". *Infotheca - Journal for Digital Humanities* Vol. 19, no. 2 (2019), [the same issue]

Krstev, Cvetana, Staša Vujičić Stanković and Duško Vitas. "Approximate Measures in the Culinary Domain: Ontology and Lexical Resources". In *Proceedings of the 9th Language Technologies Conference IS-LT*, 2014, 38–43

Krstev, Cvetana, Denis Maurel and Duško Vitas. "Serbian Language Integration in Prolexbase Multilingual Dictionary". *Infotheca - Journal for Digital Humanities* Vol. 18, no. 2 (2019): 29–52, https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2018.18.2.2_en

Mijo, Kristijan. *Rečnik zaljubljenika u gastronomiju [Dictionnaire amoureux de la gastronomie / Christian Millau]*. Službeni glasnik, 2012

Mirilov, Ružica. "*Kulinarska terminologija Vojvodine* [Culinary Terminology in Vojvodina]". PhD. thesis, Filozofski fakultet, Univerzitet u Novom Sadu, 2016

Montanari, Masimo. *Hrana kao kultura [Food as Culture]*. Sandorf, 2011

Onfre, Mišel. *Gurmanski um : filozofija ukusa [La raison gourmande / Michel Onfray]*. Gradac, 2002

Paumier, Sebastian. *Unitex 3.1 User Manual*, Université Paris-Est, 2016. https://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf

Radonjić, Danijela. "*Kulinarska leksika u savremenom srpskom jeziku* [Culinary Lexica in Contemporary Serbian]". PhD. thesis. Filološki fakultet, Univerzitet u Beograd, 2016

RSJ. *Речник српског језика [Dictionary of Serbian Language]*. Матица српска, 2011

Slapšak, Svetlana. *Leteći pilav : antropološki eseji o hrani [Flying Pilaf: Anthropological Essays on Food]*. Beograd: Biblioteka XX vek, 2014

Stošić, Anđela, Marina Milošević, Sandra Spasić, Dragica Dragosav, Ana Đorđević et al. "Working on the Multimedia Document "Al' se nekad dobro jelo"". *Infotheca* Vol. 17, no. 2 (2017): 97–118. http://infoteka.bg.ac.rs/pdf/Eng/2017-2/infoteka-2017-17-2-5.pdf, 5

Vujičić-Stanković, Staša, Cvetana Krstev and Duško Vitas. "Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain". In *Proceedings of the Seventh Global Wordnet Conference*, 2014, 127–132

Vukov, Lazar. *Imenik jela i pića : na srpskom, francuskom, nemačkom i engleskom jeziku) [Directory of dishes and drinks: in Serbian, French, German and English]*. Zadruga, 1954

Zirojević, Olga. *Istočno-zapadna sofra: mali kulturno-istorijski i kulinarski leksikon [East-West sofra: a small cultural-historical and culinary lexicon]*. Beograd: Geopoetika izdavaštvo, 2019

Балзак, Оноре де. *Голицаве приче [Les contes drolatiques / Balzac, Honoré de]*. Космос, 1933

Вивијен, Херберт. "Србија – рај сиромашног човека [Serbia - poor man's paradise]". In *Београд у деветнаестом веку – из дела страних писаца*, 163–199. Библиотека града Београда, 2008

Витас, Душко and Цветана Крстев, "Оглед из гастрономатике [Experiments in gastronomatics]". In *Теме језикословне у србистици кроз дијахронију и синхронију [Linguistic topics in Serbian through diachrony and synchrony]*, Дражић, Јасмина, Исидора Бјелаковић and Дејан Средојевић, 1–10. Нови Сад: Филозофски факултет, 2016

Витас, Душко М. "Храна из нежељене поште : (анатомија језика брзе хране [Spam Food: (Anatomy of Fast Food Language)])". In *Српски језик и његови ресурси: теорија, опис и примене*, Ђорић, Божо and Александар Милановић. *Научни састанак слависта у Вукове дане*. Vol. 47, Chapter 2, 21–35. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду, 2018. http://doi.fil.bg.ac.rs/pdf/eb_ser/msc/2018-3/msc-2018-47-3-ch2.pdf

Вујићић Станковић, Сташа and Весна Пајић. "Формирање доменског корпуса – кулинарска лексика [Preparation of a Domain Corpus – Culini-

ray Lexica]". In *Научни састанак слависта у Вукове дане – Српски језик и његови ресурси: теорија, опис и примене*, Vol. 43, Chapter 3, Међународни славистички центар, 2014, 51–59

Вујићић Станковић, Сташа and Весна Пајић. "Употреба властитих имена у кулинарском домену [The use of proper names in the culinary domain]". In *Научни састанак слависта у Вукове дане – Српски језик и његови ресурси: теорија, опис и примене*, Vol. 44, Chapter 3, Међународни славистички центар, 2015, 137–142

Голубовић, Видоје. *Механе и кафане старог Београда [Pubs and taverns of old Belgrade]*. Београд: Службени лист, 2007

Грђић-Бјелокосић, Лука. "Српска народна јела у Херцеговини и у Босни". In *Српска народна јела и пића, Књ. 1 [Serbian folk dishes and drinks, book 1]*, Ердељановић, Јован. Београд: Српска краљевска академија, 1908

Ердељановић, Јован (прир.). *Српска народна јела и пића, Књ. 1 [Serbian folk dishes and drinks, book 1]*. Београд: Српска краљевска академија, 1908

Игњатовић, Јаков. *Милан Наранџић*. Матица српска, 1949

Крстев, Цветана and Биљана Лазић. "Глаголи у кухињи и за столом [Verbs in the Kitchen and at the Table]". In *Научни састанак слависта у Вукове дане – Српски језик и његови ресурси: теорија, опис и примене*, Vol. 44, Chapter 3. Међународни славистички центар, 2015, 117–136

Марковић, Спасенија-Пата. *Велики народни кувар [Big Folk Cookbook]*. Београд: Народна књига, 1959

Милорадовић, Софија (прир.). *Обредна пракса - речима о храни : на материјалу из српских говора Војводине [Ritual practice - words about food: on material from the Serbian speeches of Vojvodina]*. Матица српска, Одељење за књижевност и језик, 2014

Мијатовић, Станоје М. "Српска народна јела (са прилогом о пићима) у Левчу и Темнићу", In *Српска народна јела и пића, Књ. 1 [Serbian folk dishes and drinks, book 1]*, Ердељановић, Јован. Београд: Српска краљевска академија, 1908

Огњановић, С. Ф. (прир.). *Ратни кувар у којем су упуства за приређивање јела већим делом готовљена без меса а за ове ратне прилике : ручна књига за наше домаћице [A war cookbook in which instructions are mostly for meat-free dishes, and for these war times: a Handbook for our housewives]*. Уjвидек, 1915

Пелапрат, Анри Пол. *Први кувар света : модерна француска и међународна уметност кувања) [L'Art culinaire moderne / Henri-Paul Pellaprat]*. Prosveta, 1973

Петроније, Гај Арбитер. *Сатирикон [Satyricon]*. Српска књижевна задруга, 1976

Поповић Мицина, Катарина. *Илустровани српски "Велики кувар" са 194 слике [Illustrated Serbian "Big Cookbook" with 194 illustrations]*. Нови Сад: Српска народна задружна штампарија, 1878

Портић, Милијанко. *Гастрономски производи [Gastronomic products]*. Природно-математички факултет, Департман за географију, туризам и хотелијерство, 2011

Радуловачки, Љиљана. *Традиционална исхрана Срба у Срему [Traditional nutrition of Serbs in Srem]*. Матица српска, 1996

Стијовић, Рада, Олга Сабо and Ранка Станковић. "Речник САНУ као база терминолошких речника (на примеру речника кулинарства) [SASA dictionary as a base of terminology dictionaries (illustrated by culinary lexica)]". In *Међународни научни симпозијум Словенска терминологија данас (Београд, САНУ, 11-13. мај 2016)*, 2016, 109–123

Тројановић, Сима. *Старинска српска јела и пића [Old-Fashioned Serbian Dishes and Drinks]*. Београд: Просвета, 1983

Фотић, Александар (прир.). *Приватни живот у српским земљама у освит модерног доба [Private life in Serbian countries in the dawn of modern times]*. Clio, 2005

6 Appendix

A1	(Тројановић, 1983)
A2	(Мијатовић, 1908)
A3	(Грђић-Бјелокосић, 1908)
A4	(Радуловачки, 1996)
A5	(Slapšak, 2014)
A6	(Zirojević, 2019)

Table 5. The group of ethnographic and anthropological texts

B1	(Поповић Мицина, 1878)
B2	(Огњановић, 1915)
B3	(Марковић, 1959)

Table 6. The group of textbooks

Author Guidelines

All *Infotheca* articles are published both in English and Serbian in the same issue. Authors should submit their articles in one of the languages; only after the notification of acceptance the translated article is expected (for Serbian authors; for all other authors translation from English to Serbian is provided by the journal). Except the printed edition, all articles are also published in the online edition in open access.

PAPER CATEGORIZATION

For documents accepted for publishing which are subject to review, the following categorization in the Journal applies:

1. Scientific papers:
   - Original scientific paper (containing previously unpublished results of authors' own research acquired using a scientific method);
   - Review paper (containing original, detailed and critical review of a research problem or a field in which authors' contribution can be demonstrated by self citation);
   - Preliminary communication (original scientific work in progress, shorter than a regular scientific paper);
   - Disquisition and reviews on a certain topic based on scientific argumentation.
2. Scientific articles presenting experiences useful for advancement of professional practice.
3. Informative articles can be:
   - Introductory notes and commentaries;
   - Book reviews, reviews of computer programs, data bases, standards etc.
   - Scientific event, jubilees.

Papers classified as scientific must receive at least two positive reviews. The opinions of the Editorial Committee do not have to correspond to those expressed in the published papers. Papers cannot be reprinted nor published under a similar title or in a changed form.

ELEMENTS OF MANUSCRIPTS

For scientific or professional papers the following data should be provided:

1. Papers should not normally exceed 15 A4 pages, Times New Roman 12pt. For longer articles the authors should contact the journal editors.

2. Names and surnames of all authors should be written in the sequence in which they will appear in a published paper.
3. After each author's full name, without titles and degrees, an e-mail address should be specified as well as the full and official name of his or her affiliation. (For large organizations full hierarchy of names should be specified, top down).
4. The submission date should be provided.
5. The authors should suggest the category of their paper but the Editor-in-Chief is responsible for the final categorization.
6. An informative abstract not normally EXCEEDING 200 WORDS that concisely outlines the substance of the paper, presents the goal of the work and applied methods and states its principal conclusion, should accompany the paper. The abstract should be supplied in both languages used for publication. In the abstract, authors should use the terms that, being standard, are often used for indexing and information retrieval.
7. Authors should supply at least 3 but not more than 10 keywords separated by commas that designate main concepts presented in the paper. The list of keywords should be supplied in both languages used for publication.
8. If paper derives from a Master's thesis or Doctoral dissertation authors should give the title of the thesis or dissertation, as well as a date of its submission and names of responsible institutions.
9. If the paper presents the results of authors' participation in some project or program, authors should acknowledge the institution that financed the project in a special section "Acknowledgment" at the end of the article, before the "Reference" section. The same section should contain acknowledgment to individuals who helped in the production of the paper.
10. If the paper was presented at a Conference but not published in its Proceedings, this should also be stated in a separate note.
11. Authors can use footnotes, while endnotes are prohibited; however, too long footnotes should be avoided. Authors can add appendices to their paper.
12. The referenced material should be listed in the section "References" at the end of the paper. In the reference list authors should include all information necessary for locating the referenced work. All items referenced in the text should be listed here; nothing that was not referenced in the text should appear in this section.

EDITING CONVENTIONS FOR ACCEPTED PAPERS

1. Papers should be prepared and submitted using LaTeX(the journal style and all packages can be downloaded from the journal web site). Authors that are not familiar with LaTeXcan prepare their papers using Word, as .doc, .docx, .rtf or .txt documents. These authors should not use any special formatting – the final formatting and transformation to LaTeXwill be done by the Infotheca team.

2. The papers written in Serbian should use CYRILLIC alphabet because they will be printed in that script. The only exceptions are those parts of the text for which the use of the other script, such as Latin, is more appropriate. All scripts should be represented using Unicode encoding, UTF-8 representation.
3. Title of the paper should not be written in capital letters. The authors should keep the length of titles reasonable – preferably less than 90 characters. For all titles authors should provide a shorter title that will be used for page headers.
4. Italic type may be used to emphasize words in running text, while bold type or italic bold type can be used if necessary. Underlined text should be avoided. Please do not highlight whole sentences or paragraphs.
5. Paper can be divided in sections and subsections, but more than two levels of the section headings should be avoided. All sections and subsections will appropriately numbered. Appendices, if any, should come at the end of the paper and they will also be appropriately labeled. If using lists, do not use more than two levels of nesting.
6. All paragraphs should be separated by one empty line (one Enter).
7. Authors should avoid too wide tables keeping in mind that the journal is published on A5 paper and. All tables, illustrations, diagrams and photographs should not be wider than 72.5 mm (the width of one column) or (exceptionally) 150 mm (the width of the page). All illustrations should be prepared in some lossless format, for instance .png, .tif or .jpg and their resolution should be at least 300 dpi.
8. The authors are kindly requested to add (if possible) the link to the screen from which a screenshot was taken. When taking a screen shot of a part of some screen, authors are advised to use the Zoom possibility of the browser or other program. For diagrams that are produced with Excel, please provide the original .xls document.
9. All tables, illustrations, diagrams and photographs should be prepared as separate files, both in black-and-white for printing and in color for the on-line version. Captions that should be below tables, illustrations, diagrams or photographs should remain in the text. Each file should have the same name as the file containing the main text, followed by the type of material to which the ordinal number in the text is added. For instance, the file containing the fourth figure of the paper "Example" should be named Example_figure_4.
10. Please add additional document(s) that explain some specific aspects of formatting required for your paper, for instance, formulas prepared in LaTeXin a .pdf format.
11. URL addresses that appear in the paper should be placed in footnotes; the date when the site was visited should be given.

REFERENCES AND CITATION

1. Referenced material should be listed at the end of the text, within the unnumbered section References. The reference section should be complete; references should not be omitted. This section should not contain any bibliographic information not referenced in the main text. Referenced items should not be mentioned in footnotes.
2. Entries in the reference list should be ordered alphabetically by authors or editors names, or publishing organizations (when no authors are identified). If this list contains several entries by the same authors, these entries should be ordered chronologically.
3. For preparation of a reference list use Chicago Manual of Style reference list entry (www.chicagomanualofstyle.org).
4. Full names of journals, and not their short titles or acronyms, should be specified. Use the 10-point type for entries in the reference list.
5. All authors, whether they prepare their articles using LaTeXor Word, will prepare all the items from their References section using BibTeX templates that are given for all the examples at the Infotheca web site (http://infoteka.bg.ac.rs/index.php/sr/upu-s-v-z-u-r).