Creating Links between Cooking Recipes

Vujičić Stanković, Staša and Pajić, Vesna

Abstract: ***Increasing popularity of development of applications in the culinary domain brings linguistic processing of culinary content in focus. As cooking is a constantly evolving process, the more complex and user-friendly systems are required. With that in mind, we have created lexical resources for Serbian language related to culinary domain, and used them for developing the system that could meet different user requirements. One of them is to find out more information about preparing some parts of a cooking recipe that are mentioned in some recipe preparation description. The paper presents the method for creating this type of links between cooking recipes from domain specific corpus in Serbian.***

Index Terms: ***Natural Language Processing*, *Language Resources*, *Culinary Domain***

1. Introduction

In this paper we want to give quick overview of important Serbian language resources in the culinary domain that have been used for development of a system for automatic processing and retrieval of text recipes, ReceptiX. This system, based on a corpus of culinary texts, generates all possible answers to requirements from a query made by a user. In the frame of this system, a specific method, which serves for creation of links between different recipes, has been implemented. It is used when a user reviews a text of a recipe and notices that, in preparing description features, there is some part, which already had appeared in another recipe, but with additional or different explanation. In our system for automatic processing and retrieval, we were able to implement a sophisticated preprocessing method that allows a user to discover related recipes for a certain recipe that meets his requirements.

The paper is organized as follows: Section 2 presents Serbian language resources used for the processing of texts from the culinary domain. Section 3 introduces an advanced search system ReceptiX. Section 4 explains our approach for establishing links between culinary recipes, implemented as one of the system's functionalities. Section 5 concludes the paper.

2. Serbian Language Resources in the Culinary Domain

2.1 The Corpus of Serbian Written Cooking Recipes

Domain-specific corpus of Serbian written cooking recipes has been created to provide a basis for studying the culinary lexica, as well as for the development and evaluation of language resources and applications in this field [1].

Recipes were retrieved from several Serbian culinary web sites intended for collecting and searching recipes, like *Recepti*[1], *Kuhinjica*[2], *Veliki kuvar*[3] etc. These sites generally contain recipes that have been posted by the users without additional proofreaders' corrections, so they consist of language which is prone to irregularities. For example, one of the most common errors is the omission of diacritics when using the Latin script, where letter *ž* is used as *z*, letter *š* as *s*, and letters *č* and *ć* are used as *c*, which renders the produced texts unusable for linguistic processing. All recipes that have been written without the use of diacritics are eliminated from the corpus.

The created text corpus contains approximately 14,000 recipes with approximately 1,600.000 simple word forms.

2.2 Serbian Electronic Dictionaries and Culinary Domain Lexica

Electronic dictionaries are intended for automatic processing of texts. They are being developed for many years now for the Serbian language, and as reported in [2] its present version is derived from a total of almost 150,000 lemmas out of which 91% are simple form and 9% are multi-word units.

Electronic dictionaries of Serbian cover both general lexica and proper names. The process of

Manuscript received October 29, 2016. This work was supported as part of the projects no. 178006 and no. III47003, financed by the Serbian Ministry of Science.

Staša Vujičić Stanković is with the University in Belgrade, Faculty of Mathematics, Serbia (e-mail: stasa@matf.bg.ac.rs).

Vesna Pajić is with the University in Belgrade, Faculty of Agriculture, Serbia (e-mail: svesna@agrif.bg.ac.rs).

[1] Recepti: http://www.recepti.com

[2] Kuhinjica: http://www.kuhinjica.rs

[3] Veliki kuvar: http://velikikuvar.com

enlarging electronic dictionaries with lemmas from the culinary domain extracted from domain-specific corpora is described in [3]. All of these lemmas also have been marked with appropriate domain specific semantic markers. For example, *meso* ‘meat’ has been marked with semantic markers +Conc+Food+Prod+DOM=Culinary to denote that it is concrete (+Conc) food (+Food) product (+Prod) from the culinary domain (+Dom=Culinary). After this process electronic dictionaries had 2,923 lemmas from the culinary domain – 1,607 simple and 1,316 compound lemmas.

The list of verbs related to the culinary domain, together with some basic information about them, is given in [4].

Domain-specific corpus has been used for detecting and categorizing different units of measure that are used in recipe descriptions. It has been noticed that beside standard units of measure, users frequently use more informal measures that are not listed in formal standards or professional manuals, like *češanj* ‘clove of’ or *prstohvat* ‘a pinch of’. Electronic dictionaries have been enlarged with a total of 105 approximate measures marked with newly proposed domain-specific semantic markers related to measures – 95 simple and 10 compound lemmas [5].

2.3 Serbian WordNet and Culinary Domain

The development of semantic network WordNet for the Serbian language has been started in the framework of the project BalkaNet[4]. At the end of the BalkaNet project, the Serbian WordNet contained about 7,000 synsets [6]. At present, it is related to the Princeton WordNet 3.0 and contains more than 21,200 synsets.

The procedure of Serbian WordNet enrichment with culinary domain lexica based on Serbian electronic dictionaries and domain-specific culinary corpus is described in [3]. At the end of this procedure, Serbian WordNet has been enlarged with almost 1,800 culinary concepts.

2.4 Ontologies related to Culinary Domain

In order to enable semantic tagging of the culinary domain lexica, as well as calculating recipe similarity, three ontologies have been created [7].

The approximate measure ontology (with 7 classes, two object properties, two data properties, and 105 individuals), has been created during the process of analyzing approximate measures used in the description of cooking process [5], while the food ontology (with 161 classes and 1091 instances), and the ontology of ingredients that can be used as mutual replacements in the culinary domain (with one class and 266 instances) have been developed during the development of the ReceptiX system (Section 3).

3. ReceptiX System for Automatic Processing and Retrieval of Cooking Recipes

The above introduced culinary corpus, identified culinary lexica, enriched WordNet, enlarged electronic morphological dictionaries, and developed domain ontologies have been used for the development of an advanced search system, ReceptiX [7].

The system generates all possible answers to inquiries made by users, like choosing ingredients user wants or does not want to cook with, which kind of dish wants to make, or determining the similarity of cooking recipes as the similarity of text that describe recipe preparation, in terms of identical or similar steps, and ingredients that are the same or are interchangeable [8].

4. Method for Creating Links between Cooking Recipes

The requirement to create links between cooking recipes occurs when a user reviews a recipe text, and observes that, in its preparation appears part for which there is a potential recipe with additional or different explanation.

In the above described system user could make this type of connection using “ReceptiX” button placed below the recipe text. After the button is used, text sequences from the preparation description that correspond to some recipe title existing in the corpus, become links to these recipes, thus enableing their viewing.

Figures 2 and 3 show an example of a recipe before and after creating links. Text sequences *bešamel sos* ‘bechamel sauce’ and *musaka* ‘moussaka’ (Figure 3) have become links to recipes that explain how to prepare those meals.

As the first step in the process of creating links between cooking recipes, we have decided to produce lemmatized text for the recipe that is being processed, and to produce lemmatized recipes’ titles for all recipes from corpus. For that purpose, we created final state transducer in the Unitex system [9] (Figure 1).

Lexical masks are listed in its boxes that recognize different types of words such as nouns (indicated by <N>), verbs (marked with <V>), pronouns (marked with <PRO>), adjectives (marked with <A>) and numbers (marked with <NUM>). Recognized substring of a text sequence is stored in variable $promenljiva$, which is used in transducer output to produce lemma of recognized word ($promenljiva.LEMMA$). In this case, finite state transducer output is applied in a special “Replace mode” where the output replaces the sequences that have been recognized in the text, while unrecognized

[4] BalkaNet: http://www.dblab.upatras.gr/balkanet/index.htm.

sequences are transcribed in its original form (as is in the example presented below, where the incorrectly written word *rastopjenom* is just transcribed in the lemmatized text).

Due to the morphological electronic dictionaries, one word form could correspond to a larger number of lemmas, hence these replacements are not unique (for example, the word *dodati* 'to add' could be recognized as an adjective and replaced with *dodat* 'additional' or as a verb when it is replaced with *dodati* 'to add'). However, in such cases, recognized sequence is always replaced with the first lemma that is listed in the electronic dictionaries, so the replacement would be the same every time, although not necessarily accurate. For further steps in creating links between recipes, where we are using lemmatized forms of both recipe texts and each lemmatized recipes' title, this approximated solution is good enough.

The resulting lemmatized recipe text and each lemmatized recipes' title are further processed in Unitex post-processing step by removing the appearance of characters '-' and multiple whitespaces produced during Unitex text processing. These characters are also removed from the original text preparation and original titles.

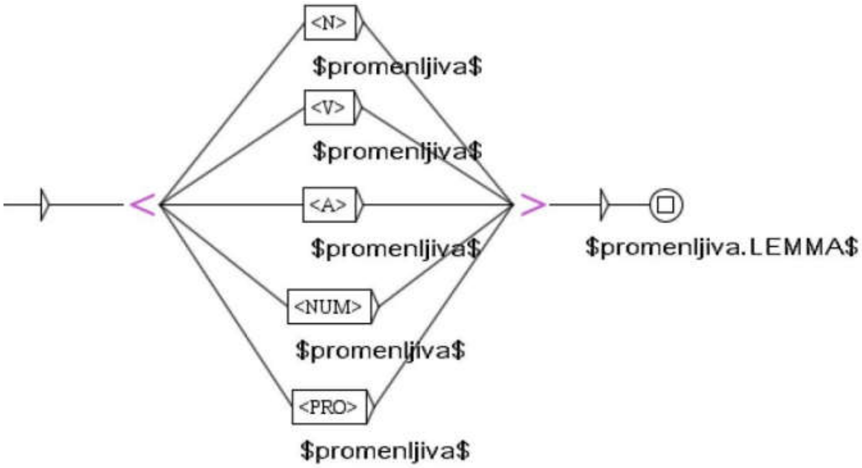


Figure 1. Final state transducer for text lemmatization

In the next step, one by one recipes' title is processed by checking if its lemmatized form occurs as a substring of the lemmatized preparation text for the recipe. If it occurs, the position of the substring is used to calculate the number of spaces before that position, i.e. the number of words in the lemmatized preparation text that precede the first occurrence of the substring. Given that the original and lemmatized texts generally do not have the same number of characters, this information is used to determine the position in original text where recipe title appears, and to mark this appearance as the link to the appropriate full recipe text.

For example, for the recipe showed in the Figure 2 lemmatized title text is *bešamel musaka* 'bechamel moussaka', while the lemmatized preparation text is:

> *makaroni obariti. luk izdinstati sa mleveno meso i začiniti dodatak za jelo. bešamel-sos napraviti na sledeći način: upržiti 2 kašika brašno na rastopjenom puter, dodat 2 dl mleko i kad provreti i dobiti određen gustina, dodat parče sir za topljenje, žumance i malo senf. kačkavalj izrendati. u pouljen plesti staviti bešamel, zatim polovina pripremljen makaroni, meso, rendan kačkavalj, preostali makaroni, preliti bešamel i preostali rendan kačkavalj. musaka staviti u rerna i peći na 200 stepen oko 20 minut. kad se prohladiti musaka iseći na kocka i poslužiti topao jesti se moći jesti i hladno.*

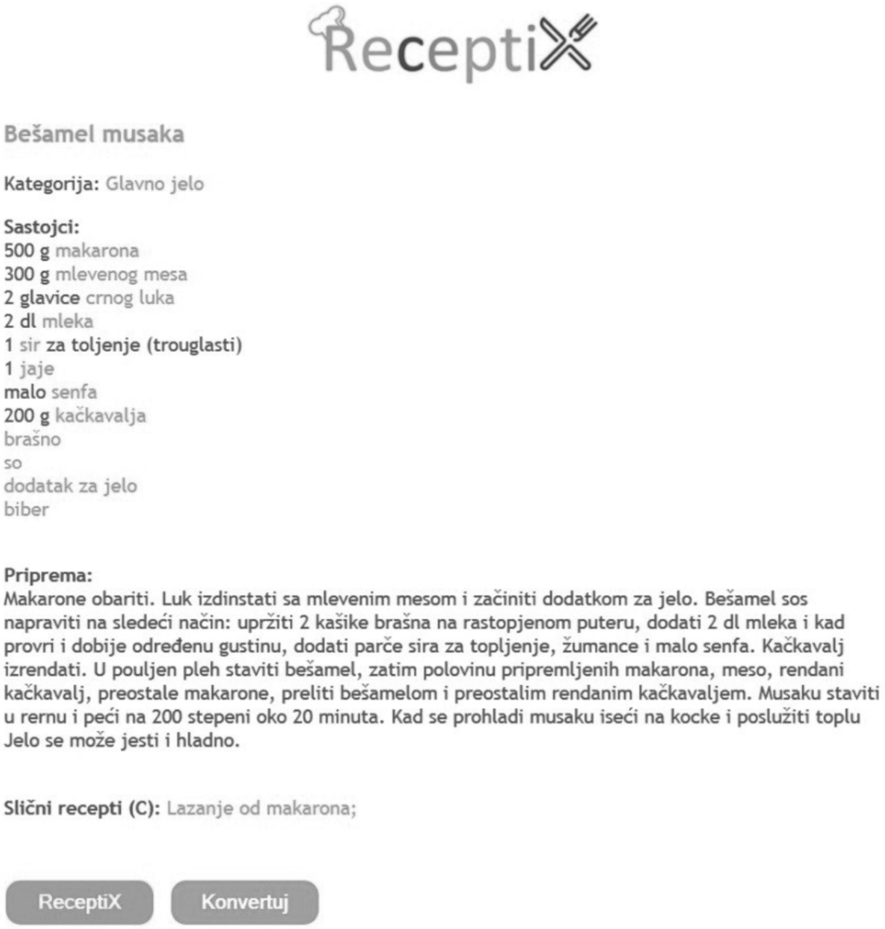
and after Unitex post-processing step:

> *makaroni obariti. luk izdinstati sa mleveno meso i začiniti dodatak za jelo. bešamel sos napraviti na sledeći način: upržiti 2 kašika brašno na rastopjenom puter, dodat 2 dl mleko i kad provreti i dobiti određen gustina, dodat parče sir za topljenje, žumance i malo senf. kačkavalj izrendati. u pouljen plesti staviti bešamel, zatim polovina pripremljen makaroni, meso, rendan kačkavalj, preostali makaroni, preliti bešamel i preostali rendan kačkavalj. musaka staviti u rerna i peći na 200 stepen oko 20 minut. kad se prohladiti musaka iseći na*

> *kocka i poslužiti topao jesti se moći jesti i hladno.*

where the occurrence of lemmatized recipe titles substrings *bešamel sos* ‘bechamel sauce’ and *musaka* ‘moussaka’ obtained after Unitex post-processing are shown underlined.

Ordinal numbers of lemmatized recipe titles initial words position in lemmatized preparation texts are 13, 68 and 83. In the same positions are marked titles in the original preparation text associated with links to appropriate recipes (Figure 3).



Bešamel musaka

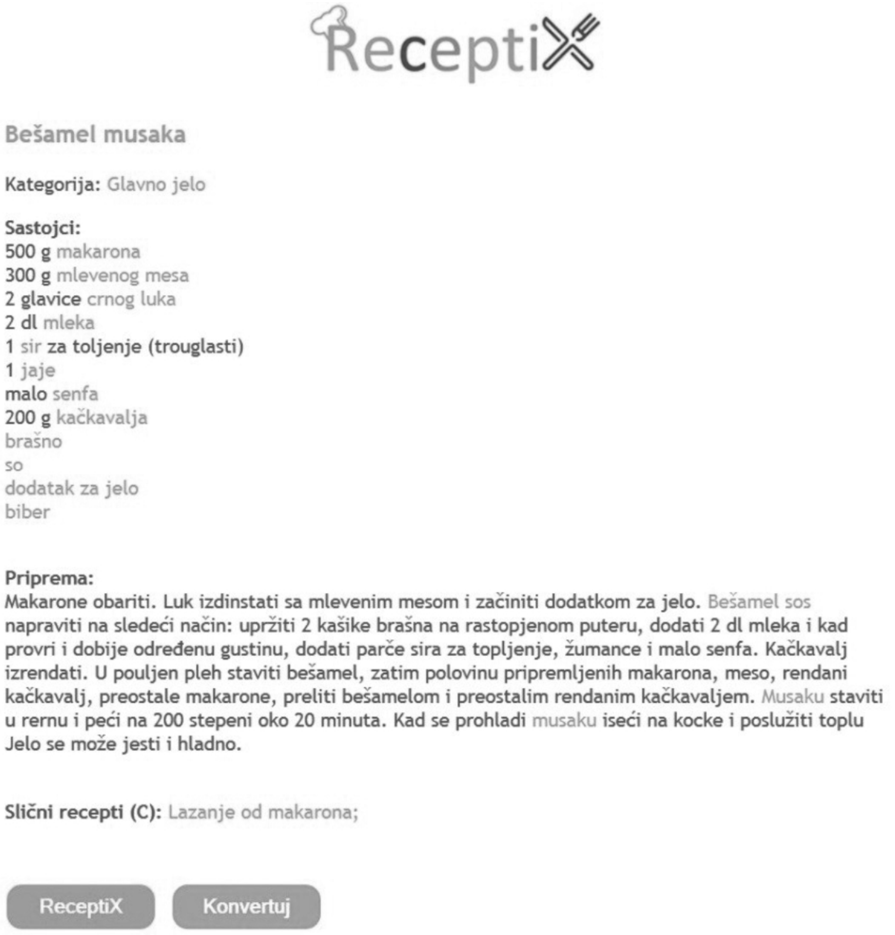
Kategorija: Glavno jelo

Sastojci:
500 g makarona
300 g mlevenog mesa
2 glavice crnog luka
2 dl mleka
1 sir **za toljenje (trouglasti)**
1 jaje
malo senfa
200 g kačkavalja
brašno
so
dodatak za jelo
biber

Priprema:
Makarone obariti. Luk izdinstati sa mlevenim mesom i začiniti dodatkom za jelo. Bešamel sos napraviti na sledeći način: upržiti 2 kašike brašna na rastopjenom puteru, dodati 2 dl mleka i kad provri i dobije određenu gustinu, dodati parče sira za topljenje, žumance i malo senfa. Kačkavalj izrendati. U pouljen pleh staviti bešamel, zatim polovinu pripremljenih makarona, meso, rendani kačkavalj, preostale makarone, preliti bešamelom i preostalim rendanim kačkavaljem. Musaku staviti u rernu i peći na 200 stepeni oko 20 minuta. Kad se prohladi musaku iseći na kocke i poslužiti toplu Jelo se može jesti i hladno.

Slični recepti (C): Lazanje od makarona;

Figure 2. The cooking recipe text before marking links to other recipes.



Bešamel musaka

Kategorija: Glavno jelo

Sastojci:
500 g makarona
300 g mlevenog mesa
2 glavice crnog luka
2 dl mleka
1 sir **za toljenje (trouglasti)**
1 jaje
malo senfa
200 g kačkavalja
brašno
so
dodatak za jelo
biber

Priprema:
Makarone obariti. Luk izdinstati sa mlevenim mesom i začiniti dodatkom za jelo. Bešamel sos napraviti na sledeći način: upržiti 2 kašike brašna na rastopjenom puteru, dodati 2 dl mleka i kad provri i dobije određenu gustinu, dodati parče sira za topljenje, žumance i malo senfa. Kačkavalj izrendati. U pouljen pleh staviti bešamel, zatim polovinu pripremljenih makarona, meso, rendani kačkavalj, preostale makarone, preliti bešamelom i preostalim rendanim kačkavaljem. Musaku staviti u rernu i peći na 200 stepeni oko 20 minuta. Kad se prohladi musaku iseći na kocke i poslužiti toplu Jelo se može jesti i hladno.

Slični recepti (C): Lazanje od makarona;

Figure 3. The cooking recipe text with marked links to other recipes.

5. CONCLUSION

This paper shows how different domain specific linguistic resources and tools can be integrated in order to provide users with a way to take benefit of improved cooking recipes representation. In order to improve the quality we have implemented the method that helps users discover recipes closely related to the one that they are preparing at the moment. It is based on extracting recipe titles automatically from recipe texts and their linking to appropriate full cooking recipe text, using an application based on natural language processing methods. Creating those links will make it possible, especially for beginner cooks, to use advantages of a more user-friendly interface that offers additional explanations and cooking instructions leading to easier overcoming of cooking challenges or broadening cooking knowledge and skills. Although there are systems that meet complex user requirements of similar types (for example, [10]), the system with requirements presented in this paper is the first one for the automatic natural language processing of the cooking recipes written in the Serbian language.

REFERENCES

[1] Vujičić Stanković, S., Pajić, V., “Formiranje domenskog korpusa – kulinarska leksika,” *Faculty of Philology*, *University of Belgrade*, Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opis i primene 43(3), 2014, pp. 51-59.

[2] Krstev, C.,“Processing of Serbian – Automata, Texts and Electronic Dictionaries,” *Faculty of Philology*, *University of Belgrade*, Belgrade, Serbia, 2008.

[3] Vujičić Stanković, S., Krstev, C., Vitas, D., “Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain,” Proceedings of the Seventh Global Wordnet Conference, 2014, pp. 127-132.

[4] Krstev, C., Lazić, B, “Glagoli u kuhinji i za stolom,” *Faculty of Philology*, *University of Belgrade*, Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opis i primene 45(3), 2015, pp. 117-135.

[5] Krstev, C., Vujičić Stanković, S., Vitas, D., “Approximate Measures in the Culinary Domain: Ontology and Lexical Resources,” *Institut "Jožef Stefan"*, Proceedings of the 9th Language Technologies Conference IS-LT 2014, 2014, pp. 38-43.

[6] Tufis, D., Cristea, D., Stamou, S., “BalkaNet: Aims, Methods, Results and Perspectives, A General Overview,” Romanian Journal of Information Science and Technology, 7(1-2), 2004, pp. 9-43.

[7] Vujičić Stanković, S., “Ontology based Information Extraction (Model for the Serbian Language),” doctoral dissertation, *Faculty of Mathematics*, *University of Belgrade*, Belgrade, Serbia, 2016.

[8] Vujičić Stanković, S., Pajić, V., “Automatsko utvrđivanje sličnosti kuvarskih recepata upotrebom metoda

ekstrakcije informacija," *Faculty of Philology*, *University of Belgrade*, Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opis i primene 45(3), 2015, pp. 47-59.

[9] Paumier, S., "Unitex 3.1 User Manual," http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf. 2014.

[10] Bilow, R., "How IBM's Chef Watson actually works," http://www.bonappetit.com/entertaining-style/trends-news/article/how-ibm-chef-watson-works. 2014.