Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain

Staša Vujičić Stanković
Faculty of Mathematics
University of Belgrade
Serbia
stasa@matf.bg.ac.rs

Cvetana Krstev
Faculty of Philology
University of Belgrade
Serbia
cvetana@matf.bg.ac.rs

Duško Vitas
Faculty of Mathematics
University of Belgrade
Serbia
vitas@matf.bg.ac.rs

Abstract

In this paper we present three lexical resources for Serbian that are crucial for the development of applications in the culinary domain based on natural language processing. The first two of them — Serbian WordNet and morphological e-dictionaries — have already been in development for some time, while the third one – a corpus of culinary recipes -– has been developed specifically for this purpose. In this paper, we present how we use each of these resources to correct and enlarge the other two. We use various automatic procedures, but manually check all the results.

1 Introduction and Motivation

In recent years, linguistic processing of culinary content has become increasingly popular. One of the main reasons for this is the emergence of a large amount of content related to the culinary domain on the Internet. Culinary linguistics (Gerhardt et al., 2013) emerged from the fact that both food and language are present in everyday life. From the perspective of natural language processing, in addition to knowledge representation, culinary linguistics comprises different types of reasoning. Providing these types of processing for the Serbian written texts was the motivation for our research.

WordNet (WN) has been recognized as one of the most important resources for the development of natural language processing applications (information extraction, information retrieval, question answering applications etc.). Accordingly, enriching and enhancing WN using different lexical resources, and vice versa, has become one of the central tasks (Agirre et al., 2000; Agirre et al., 2001; Nimb et al., 2013). Nowadays, with the increasing popularity of the Semantic Web to which WN is closely associated, a lot is being done on enhancing its expressiveness by introducing new relations between concepts (Ruiz-Casado et al., 2007) or new categories (Montoyo et al., 2001).

For the development of any kind of natural language processing application for Serbian written texts from the culinary domain, it was essential to enrich both the Serbian WordNet (SWN) and electronic dictionaries with the appropriate terms from the domain. There were similar efforts taken for other languages where authors addressed the problem of enriching WN related to some specific domains (Vintar and Fišer, 2011; Navigli and Velardi, 2002), but the suggested approaches were different from the one proposed in this paper. Additionally, to the best of our knowledge there is no research dealing with these problems related to Serbian WordNet, although some research related to culinary domain were proposed in (Milićević, 2013), but for different purposes.

Our motivation for WN and electronic dictionaries domain-specific enrichment was to provide a basis for the development of language resources and more complex natural language processing applications in the culinary domain. Language resources of particular interest for this specific domain are recipe, food, meal and other ontologies. Related applications should provide extraction of the relevant concepts, attributes and relations from the recipe corpus in order to overcome standard querying by keywords, and provide advanced search, based on criteria and queries.

The goal of our (informal) culinary project is to develop application where user could query recipes in Serbian; for example, by number of calories according to some diet, even though this information is not explicitly stated in the recipes themselves, but in specially developed ontology. Other search criteria could be related to some special condition of the user health and nutritional information related to the food contained in recipes, in which case it is necessary to include food nutritional information or substitutions in ontologies, etc.

To that end, our first task was to enhance and upgrade the existing lexical resources for Serbian – SWN and morphological electronic dictionaries, and to build a corpus that we can use for terminology extraction. The organization of this paper is as follows: The details of the corpus of culinary recipes in Serbian that we created for the purposes of this research are presented in Section 2. In Section 3 and Section 4 we provide an overview of the current versions of the SWN and electronic dictionaries for Serbian, respectively, with special emphases on the terms related to the culinary domain, as well as, on the newly introduced concepts and domain-specific semantic markers. WN and electronic dictionaries enrichment process and the results obtained are presented in Section 5. Finally, some conclusions and thoughts on future work are given in Section 6.

2 Details of the Culinary Text Corpus

For the purpose of harvesting domain-specific terminology, we created corpus of Serbian written culinary recipes in the Latin script. Due to the growing amount of culinary content, such as recipes, various tips and descriptions, the corpus was formed from web texts.

There are numerous free programs for downloading text from web pages, that give satisfactory results — like BootCaT.[1] But besides the text that is displayed to users, we were interested in maintaining the original structure of web pages, as well. Therefore, for the purposes of our research, we decided to develop programs adjusted to particular web pages, their content and also the meta-data that could be used in our ongoing work. These individually tailored programs were implemented in the Java programming language that provides support for text processing using regular expressions.

The texts have been collected from several leading national websites from the culinary domain like Recepti[2], Kuhinjica[3] etc. The created text corpus contains approximately 14,000 recipes, which consist of approximately 1,600.000 simple word forms. However, since much of the culinary content on the Web is user-generated we discovered that we could not use everything that was collected for our purpose. Namely, when using the Latin script users sometimes tend to ignore diacritics which renders the produced texts unusable for linguistic processing. Such omissions cannot be corrected automatically, because they increase the homography of forms – e.g. *vece* itself can represent a word of the language (colloquial for WC), but we may also presume that it is missing one of two possible diacritics: *veće* ‘bigger’ or *veče* ‘evening’. Therefore, we discarded all recipes that did not contain any Serbian-specific letters with diacritics. Since the resulting corpus still contained quite a number of errors, due to careless typing, we corrected some of the frequently occurring ones, like the use of the digraph *dj* instead of the letter *đ*, and the digraph *dz* instead of *dž*. As we did not want to introduce new errors by applying simple find/replace, we corrected only unknown words that became known Serbian words after correction (according to Serbian e-dictionaries, see Section 4).

3 Serbian WordNet

The production of the SWN was initiated together with the Bulgarian, Greek, Romanian, and Turkish versions by the BalkaNet project. The structure of all these WNs corresponded to the structure established by the EuroWordNet project and they were all linked to the Princeton WordNet (PWN), through the so-called Interlingual Index (version 2.1 at the end of the project). Besides, all BalkaNet WNs were developed following the *expand model* (Fellbaum, 2010), which means that synsets from the PWN were translated into target languages, and the relations between synsets were transferred as well (a hypernym/hyponym as a rule, other if applicable). At the end of the BalkaNet project, the SWN had 7,000 synsets, covering basic concept sets 1 and 2, and most of the concepts from the subset 3 (Tufis et al., 2004).

After the end of the BalkaNet project, the development of the SWN continued, but at a much slower pace, since there was no project to support it. The development mostly relied on volunteer work of its chief editor and numerous Masters and PhD students who followed the same expand model in their work. Due to such circumstances, the choice of the synsets to be transferred was not concept-dependent, but rather domain-dependent, because chief editor wanted to make the most of

[1] http://www.bootcat.sslmit.unibo.it
[2] http://www.recepti.com
[3] http://www.kuhinjica.rs

the specific knowledge and interests of her volunteers. As a result, the Serbian WordNet was enlarged to almost 20,000 synsets.

Before the beginning of the (informal) culinary project, concepts belonging to the culinary domain were not given special attention. However, 393 such concepts were already present in the SWN, 99 of which belong to basic concept sets and 91 to Balkan- or Serbian-specific concepts.

4 Electronic Dictionaries for Serbian

The development of Serbian e-dictionaries follows the methodology and format known as DELA presented for French in (Courtois et al., 1990). The role of electronic dictionaries, covering both simple words and multi-word units (MWUs), and dictionary finite-state transducers (FSTs), is text tagging as part of various natural language applications. Each such e-dictionary of forms consists of a list of entries supplied with their lemmas, morphosyntactic, semantic, and other information. The forms are, as a rule, automatically generated from the dictionaries of lemmas containing the information that enables the production of forms. The system of Serbian e-dictionaries covers both general lexica and proper names and all inflected forms are generated from 130,500 simple forms and 10,500 MWU lemmas (Krstev, 2008). Approximately 28.5% of these lemmas represent proper names: personal, geopolitical, organizational, etc.

Most of the word forms in the Serbian morphological e-dictionaries are supplied not only with the values of the grammatical categories, but also with the additional markers that are inherited from the lemmas from which they are generated. These markers can be grammatical (the marker +MG for the natural masculine gender, as opposed to the grammatical gender, e.g. in *muškarčina* 'macho'), derivational (+Pos for possessive adjectives, e.g. *bikov* 'belonging to a bull, taurine'), dialectic (+Ek for the Ekavian pronunciation, e.g. *devojka* 'girl'), domain specifying (+Math for mathematics, e.g. *mnogougao* 'polygon'), and semantic (+Hum for humans, e.g. *drug* 'friend'). Some of the semantic markers are redundant, e.g. the marker +Top (for geographic names) is superfluous if the marker +Gr (for settlements) is present. However, we keep them all for processing purposes – if a geographic name is needed, we do not have to list all their types.

Some of these markers were systematically added to the dictionary entries to which they apply, while others were conceived later and added systematically only to the entries included in the dictionaries at some later stage. The latter was the case for words from the culinary domain. Before starting the enrichment process, there were 218 simple word entries with the semantic marker +Food, and 217 multi-word entries. All entries with the +Food marker should also have been assigned the +Conc marker (for *concrete object*, as a more general category), but this was not the case either: 32 simple entries and 20 multi-word entries were missing it. Naturally, at this moment we still do not know how many entries in e-dictionaries are missing the +Food marker, because supplying as many entries as possible with it is one of the goals of our project.

4.1 Domain Specific Semantic Markers for Serbian Electronic Dictionaries

The concepts and the terminology specific to the culinary domain required introduction of a new domain marker and more refined semantic markers. Table 1 provides an overview of the newly proposed semantic markers, that could be used individually or in combination. Naturally, the domain marker +Culinary is assigned to all the lemmas from the culinary domain. All other markers are used in combination with the +Conc marker, except the +MesApp marker for approximate measures often used in cooking, like *prstohvat* 'an amount between fingers, a pinch'. Similarly, the +Food marker is assigned with all other markers except +MesApp and +Uten, that is asigned to utensils used in food preparation and serving. The +Erg marker is assigned to the names of man-created items that have the status of trademarks. It can be assigned to both food *tabasko* 'Tabasco' and utensils *teflon* 'Teflon'. It goes without saying that in the culinary domain these names are used loosely and because of that often with the lower-case initial in Serbian. Namely, if recipe states that *campari* 'Campari' should be used, it is understood that if not available, it can be replaced by some similar liqueur. The marker +Erg is used outside the culinary domain, as well, e.g. *rols-rojs* 'Rolls Royce'.

In addition to these semantic markers that are already added to the Serbian e-dictionary, in further research, we intend to address the terminology related to food condition, food taste, as well as the way of food preparation, for which we have dedicated new semantic markers – +Cond, +Taste, and +WoP, respectively, that are related mainly to adjectives and verbs. At this point, they are not included in the dictionary (except for some newly added entities), and their systematic adding would be an objective of our future work.

Semantic marker	**Description**
+Culinary	culinary domain
+Food	food (e.g. *senf* 'mustard')
+Alim	aliment (e.g. *mleko* 'milk
+Prod	product (e.g. *sirće* 'vinegar')
+Meal	meal (e.g. *doručak* 'breakfast')
+Course	course (e.g. *puding* 'pudding')
+Uten	utensil (e.g. *šolja* 'cup')
+Erg	ergonym (e.g. *rokfor* 'Roquefort')
+MesApp	approximate measures (e.g. *kašičica* 'spoonful')
+Taste	taste (e.g. *slatkokiseo* 'sweet-sour')
+WoP	way of preparation (e.g. *dinstati* 'to stew'; *dinstanje* 'stewing')
+Cond	condition (e.g. *bajat* 'stale')

Table 1: The overview of newly proposed semantic markers.

5 Enrichment Process

The process of enriching both the Serbian WordNet and Serbian e-dictionaries proceeded in several steps:

1. Manual translation of as many synsets from the culinary domain as possible belonging to the PWN.

2. Inspection of unknown words resulting from the application of Serbian e-dictionaries to the corpus of recipes in search of new entries.

3. (Semi-)automatic production of new simple word and multi-word entries for e-dictionaries with all applicable markers, derived from the synsets, in the SWN, belonging to the culinary domain.

4. (Semi-)automatic addition of all missing markers in e-dictionaries, based on the synsets in the SWN belonging to the culinary domain.

5. (Semi-)automatic addition of new culinary and/or Serbian-specific concepts to the SWN and manual correction.

Steps one and two were performed by three graduate Library and Information Science students well-educated in the field of information search. Their role in step one was to investigate specific branches in the PWN and transfer into the SWN all concepts recognized in Serbian. The branches of interest were '*food, nutrient*' related to aliments, products, drinks, meals and courses, and '*kitchen utensil*' and '*tableware*' related to utensils. The role was not very precise, but students took their job seriously and translated everything for which they could find evidence. As a result, the SWN now has all concepts related to fruits, as the PWN, although hardly anybody in Serbia has ever heard of some of them (e.g. *durian* 'durian' and *žabotikana* 'jaboticana'), let alone tasted them. The same principle could not always be applied -– for instance, quite a number of fish species represented in the PWN are completely unknown in Serbia (e.g. *scup*, *sailfish*, *sucker*, etc.). It should be stressed that the students supplied a definition for each introduced sysnet, which is in line with the strategy applied for the development of the SWN from the beginning – practically all its synsets have a definition. Everything produced by the students was double checked by chief SWN editor.

Step two was equally imprecise. The students' task was to recognize, in the long list of unknown words in the corpus of recipes comprising of 9,100 word forms, all those for which they knew the meaning without further consultation. All chosen entries were assigned the appropriate markers, as well as, codes for inflectional paradigms, which was done manually for simple words and automatically for MWUs.

Step three consisted of two tasks. First, we produced new candidates for e-dictionaries of simple and MWUs automatically by inspecting the synsets belonging to the already mentioned hierarchies, choosing those that were not in e-dictionaries already. These new candidates were all supplied with the appropriate markers which were derived from the position of a synset in a hierarchy. For instance, the new candidate *fondi* 'fondue' belongs to the hierarchy {dish:2}, {nutriment:1,...}, {food:1, nutrient:1}, {substance:1, matter:1}, and therefore the suggested markers for it were +Conc, +Food, +Course (and +Culinary, as a domain marker). The second task consisted of manual checking of all new candidates and their markers. A good number of candidates were rejected for several reasons. There were duplicates (a literal belonging to several synsets, e.g. *brizle* is connected to {neck sweetbread:1, throat sweetbread:1} and to {sweetbread:1, sweetbreads:1}) for which there should be only one entry in the e-dictionaries. There were literals irrelevant to e-dictionaries, because they were of a descriptive nature and not really lexicalized (e.g. *grožđe sa glatkom kožom* corresponding to {fox grape:1, slip-skin grape:1}. In a few cases, a literal from the chosen hierarchies did not actually belong to the culinary domain (e.g. *Poslednja večera* corresponding to {Last Supper:1, Lord's Supper:2} that belongs to the branch {food:1, nutrient:1}). The markers themselves have also to be checked and if necessary corrected. For instance, *pomfrit* 'french fries' has as a hypernym {vegetable:1}, and thus it obtained the marker +Alim; however, we believed that +Prod was more appropriate.

The fourth step was performed in a similar way as the previous one, except that we considered now only the entries already in e-dictionaries missing some or all appropriate markers. The produced list of enhanced entries had also to be considered carefully in order not to add markers to wrong entries. For instance, suggested new markers for the entry *baba* 'baba' were +Conc, +Food, +Course, while the entry already in the dictionary corresponded to *baba* 'grandmother'. Similarly, the entry *luk* 'bow' obtained markers +Food+Conc+Alim intended only for the entry *luk* 'onion'.

In step five, we used new entries for e-dictionaries, produced in step two, to create new synsets in the SWN. These entries include either the concepts specific to Serbia, like *afusali*, a type of grapes very popular in Serbia, or too specific concepts that were missing in the PWN, like *friteza* 'deep fryer'. Since they were already assigned semantic markers, we used them to find the right place for the appropriate synsets. In the case of MWUs, we could do even more, because many of them contained as a unit a literal from a hypernym synset: *vatrostalna činija* 'fireproof bowl' i *zdenka sir* 'zdenka cheese, a popular cheese' are a kind of a bowl and a kind of cheese, respectively, and they could be pushed further down the hierarchy. The position of every newly added synset was checked manually and corrected if necessary.

At the end of this phase we obtained the following results:

- The SWN was enlarged by translating 1,404 synsets from the culinary domain from the PWN to the SWN, to contain a total of 1,797 such synsets;
- Serbian e-dictionaries of simple words were enlarged by 636 entries, 246 of which were obtained from the SWN and 390 from the culinary corpus.
- Serbian e-dictionaries of MWU were enlarged by 612 simple entries, 514 of which were obtained from the SWN and 98 from the culinary corpus.
- The full set of the appropriate markers was assigned to 735 simple word and 125 multiword entries.
- 450 specific concepts from the culinary domain were added to the SWN.

6 Conclusion and Future Work

We have completed the first phase of enrichment of the SWN and Serbian e-dictionaries. The next phase will consist of the following steps:

1. (Semi-)automatic detection in the corpus of all words belonging to the culinary domain and e-dictionaries that are still not assigned all applicable markers and manual marker selection and assignment.
2. (Semi-)automatic detection in the corpus of other MWU terms belonging to the culinary domain.
3. Extension of our approach to other PoS synsets and dictionary enties.

In order to complete this phase, we will rely on various local grammars, some of which were already developed for Serbian for different purposes (Krstev et al., 2011).

Acknowledgements

We would like to thank the following PhD students at the Faculty of Philology, University of Belgrade for their help in enhancing the SWN with synsets from the culinary domain: Biljana Đorđević, Jelena Andonovska and Katarina Stanišić. This research was conducted through the project 178006, financed by the Serbian Ministry of Science.

References

Eneko Agirre, Olatz Ansa, Eduard Hovy, and David Martínez. 2000. Enriching very large ontologies using the WWW. *arXiv preprint cs/0010026*.

Eneko Agirre, Olatz Ansa, Eduard Hovy, and David Martinez. 2001. Enriching WordNet concepts with topic signatures. *arXiv preprint cs/0109031*.

Blandine Courtois, Max Silberztein Ladl, et al. 1990. Dictionnaires électroniques du français. *Langue française*, 87(1):3–4.

Christiane Fellbaum. 2010. *WordNet*. Springer.

Cornelia Gerhardt, Maximiliane Frobenius, and Susanne Ley. 2013. *Culinary Linguistics: The chef's special*, volume 10. John Benjamins Publishing.

Cvetana Krstev, Duško Vitas, and Aleksandra Trtovac. 2011. Orwells 1984 — the Case of Serbian Revisited. In *Proc. of 5th Language & Technology Conference*, pages 25–27.

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

Maja Milićević. 2013. Genre-based BootCaT corpora for morphologically rich languages. *BOTWU - BootCaTters of the world unite! A workkshop on the BootCaT toolkit, Forli, 24 June 2013*. `http://botwu.sslmit.unibo.it/download/milicevic.pdf`.

Andrés Montoyo, Manuel Palomar, and German Rigau. 2001. WordNet Enrichment with Classification Systems. In *Proc. of WN and Other LRs: Applications, Extensions and Customisations Workshop.(NAACL-01) The 2nd Meeting of the North American Chapter of the ACL*, pages 101–106.

Roberto Navigli and Paola Velardi. 2002. Automatic Adaptation of WordNet to Domains. In *3rd International Conference LREC, Las Palmas, Canary Islands, 29-31 May 2002*, pages 1023–1027.

Sanni Nimb, Bolette S Pedersen, Anna Braasch, Nicolai H Sørensen, and Thomas Troelsgård. 2013. Enriching a wordnet from a thesaurus. *Lexical Semantic Resources for NLP*, page 36.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2007. Automatising the Learning of Lexical Patterns: an Application to the Enrichment of WordNet by Extracting Semantic Relationships from Wikipedia. *Data & Knowledge Engineering*, 61(3):484–499.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Špela Vintar and Darja Fišer. 2011. Enriching Slovene WordNet with domain-specific terms. *Translation: Computation, Corpora, Cognition*, 1(1):29–44.