Approximate Measures in the Culinary Domain: Ontology and Lexical Resources

Cvetana Krstev,* Staša Vujičić Stanković,† Duško Vitas†,

* Faculty of Philology, University of Belgrade
Studentski trg 3, 11000 Belgrade, Serbia
cvetana@matf.bg.ac.rs
† Faculty of Mathematics, University of Belgrade
Studentski trg 16, 11000 Belgrade, Serbia
(stasa,vitas)@matf.bg.ac.rs

Abstract

Language resource development is extremely important for Serbian, as a less-resourced language, to take it into the digital era. In our research we focused on the culinary domain, given the increasing popularity of linguistic processing of culinary content. We provide a detailed description of the language resources – electronic morphological dictionaries, the WordNet semantic network, and a corpus of Serbian written culinary recipes, developed during our earlier work, as well as our latest efforts in enriching morphological dictionaries and WordNet with approximate measure terminology and developing an approximate measure ontology. The paper presents the issues related to detecting and categorizing the approximate measures from the culinary domain to be marked with new domain-specific semantic markers and populate the ontology, and indicates the benefits that language resources gain after addressing them.

Približne mere v kulinariki: ontologija in leksikalni viri

Za srbščino kot jezik s pomanjkljivo jezikovno opremljenostjo je razvoj jezikovnih virov izrednega pomena, saj bo le tako uspešno prešla v digitalno dobo. V naši raziskavi smo se osredotočili na področje kulinarike, saj je mogoče zaslediti vedno večje zanimanje za jezikoslovno obdelavo besedil s tega področja. V prispevku podamo natančen opis jezikovnih virov – računalniških morfoloških slovarjev, semantičnega leksikona WordNet in korpusa srbskih kuharskih receptov, ki so bili razviti v predhodnem delu, kot tudi naše trenutne raziskave o razširitvi morfoloških slovarjev in leksikona WordNet s terminologijo približnih mer in izgradnjo ontologije približnih mer. Prispevek predstavi problematiko identifikacije in kategorizacije približnih mer iz domene kulinarike, ki jih je treba označiti z novimi, domensko specifičnimi semantičnimi markerji in vključiti v ontologijo, ter pokaže na prednosti za jezikovne vire, ki jih prinese razreševanje te problematike.

1. Introduction

It seems that the culinary domain is one of the rare domains in which the general public and the scientific community are equally interested today. The first claim can be easily supported by a number of web sites, which offer a huge number of recipes, in many languages, searchable by different criteria, and often populated by users. A number of such sites exist in Serbian as well. Moreover, many TV shows worldwide are devoted to the art of cooking. In addition to popular magazines, the publishing of culinary books and manuals is still flourishing: from at least 70 to more than 200 such works are published each year in Serbia and recorded by the National Library of Serbia.

On the other hand, Various aspects of the culinary domain continuously attract the research community. The existence of various scientific institutions[1] and many scientific publications[2] from the domain can serve as evidence. The new application from IBM "Chef Watson with Bon Appétit" uses Watson's capabilities to explore big data to create new recipes.[3]

It is obvious that such an attractive and vivid domain is interesting for information processing (Mori et al., 2012; Dufour-Lussier et al., 2012; Wiegand et al., 2012; Ahnert, 2013; Nedovic, 2013), as it offers a lot of resources in the form of written and spoken texts. Obviously, it also has to be supported by information technologies, like ontologies (Cantais et al., 2005; Batista et al., 2012; Kim, 2012), in order to build various applications (for example systems based on ontologies like FOODS (Snae and Bruckner, 2008), TAAABLE (Badra et al., 2008) or Global Track&Trace Information System (Pizzuti et al., 2014)).

In this paper, we will first present Serbian language resources that are not only used for the processing of texts from the culinary domain, but also benefit from it (Section 2). Next, we will present one specific aspect of this domain, namely the use of measures, in recipes (Section 3) with special emphasis on the approximate, more informal, measures that are not listed in formal standards or professional manuals (e.g. 'a pinch of', 'small bunch', 'clove of' etc.) (Section 4). Section 5 presents an approximate measures ontology and gives the details about how they are covered in the Serbian resources. Finally, we will show how an adequate treatment of measures helps in the processing of texts from the culinary domain and give some directions for the future work (Section 6).

2. Serbian Language Resources in the Culinary Domain

2.1. The Corpus of Serbian Written Culinary Recipes

For the purpose of research into the culinary domain, we created a corpus of approximately 14,000 culinary recipes (more than 1.5 million word forms) in Serbian (both pronunciations – Ekavian and Ijekavian), written in

[1] One of the most important is IEHCA – Institut européen d'histoire et des cultures de l'alimentation, in Tours, France.
[2] The IEHCA catologue can be consulted at http://www.portail.scd.univ-tours.fr and the selected scientific bibliography at http://www.foodbibliography.eu.
[3] See http://www.research.ibm.com/software/IBMResearch/multimedia/Cognitive-Cooking-Fact-Sheet.pdf.

the Latin script. The recipes were drawn from *Recepti*[4] and other similar Serbian culinary Internet portals mentioned above.

As any web user interested in food preparation can post her/his recipes on these sites, their content, regarding both their style and syntax, is not strictly controlled. Therefore different types of errors were identified. The most frequent one that cannot be automatically corrected, at least not in all cases, is the omission of Latin script diacritics or their replacement with digraphs,[5] which introduces a number of homographic forms. To resolve this problem, we did not include in our corpus any recipe that does not feature at least one Serbian Latin script letter with diacritics. In the remaining recipes, we managed to recover some of the missing diacritics with the help of Serbian e-dictionaries (described in the next subsection).

2.2. Serbian Electronic Dictionaries

The basic resources for natural language processing of Serbian consisting of electronic (e-)dictionaries and local grammars are being developed using the finite-state methodology as described in (Courtois et al., 1990). The main role of these resources is text tagging. Each word form in an e-dictionary is equipped with the following information: (a) lemma; (b) Part-of-Speech (PoS); (c) set of values of grammatical categories pertinent to a PoS; (d) set of markers – syntactic, semantic, dialectic, derivational, domain etc. – describing a lemma. As reported in (Krstev, 2008), the system of Serbian electronic dictionaries covers both general lexica and proper names, and its present version is derived from 131,000 simple form lemmas and 13,000 compound lemmas (a.k.a. multi word units). In addition to that, a collection of finite-state transducers (FSTs) has been developed to support tagging that recognizes multi-word units belonging to open sets, e.g. multi word numerals and other numerical expressions (Krstev & Vitas, 2006).

Semantic marker	**Description**
+Culinary	culinary domain
+Food	food
+Alim	alimentation (e.g. *žito* ‘wheat’)
+Prod	product (e.g. *brašno* ‘flour’)
+Course	course (e.g. *torta* ‘cake’)
+Ing	ingredient (e.g. *so* ‘salt’)
+Meal	meal (e.g. *čajanka* ‘tea party’)
+Uten	utensil (e.g. *ekspres-lonac* ‘express pot’)
+Taste	taste (e.g. *aromatizovan* ‘flavored’)
+WoP	way of preparation (e.g. *nadevati* ‘stuff’ and *nadeven* ‘stuffed’)
+Cond	condition (e.g. *taze* ‘fresh’)
+MesApp	approximate measure (e.g. *kriška* ‘slice’)

Table 1. The semantic markers in Serbian e-dictionaries related to the culinary domain.

In order to improve the processing of texts from the culinary domain, we enlarged our e-dictionaries with new lemmas from this domain and systematically added the appropriate semantic markers to all lemmas identified as related to the domain. For this task, we used both our corpus (subsection 2.1) and the Serbian WordNet (subsection 2.3), as described in (Vujičić Stanković et al., 2014). When adding new entries we took care about language variants or pronunciation, e.g. Ekavian *belo vino* and Ijekavian *bijelo vino* ‘white wine’, so they were added into dictionaries no matter which form was actually occurring in the corpus. The set of markers is presented in Table 1. As a result, our e-dictionary now has 2,923 lemmas from the culinary domain – 1,607 simple lemmas and 1,316 compound lemmas.

2.3. Serbian WordNet

The development of WordNet for Serbian started in 2001 as a part of the BalkaNet Project.[6] As part of this project, EuroWordNet, corresponding to Princeton WordNet 2.1 (Fellbaum, 2010), was expanded by adding Balkan languages: Bulgarian, Greek, Romanian, Serbian, and Turkish. In 2004, at the end of the BalkaNet project, the Serbian WordNet (SWN) contained 7,000 synsets (Tufis et al., 2004). In the years that followed, the development continued, primarily on a voluntary basis. At present, the SWN is related to the Princeton WordNet 3.0 (PWN) and contains more than 21,200 synsets. The culinary domain is one of the domains that has been systematically filled – some characteristic branches in the hypernym/hyponym hierarchy were transferred from PWN to SWN by volunteering students and then used to automatically fill the gaps in Serbian e-dictionaries; and vice versa, lemmas from a Serbian e-dictionary and their culinary markers were used to fill the gaps in the SWN with Serbian-specific concepts (for more details see (Vujičić Stanković et al., 2014)). As a results of this procedure, the SWN has around 1,800 culinary concepts today, almost 550 of which are Serbian-specific concepts.[7]

2.4. Serbian Named Entity Recognition System

The Serbian Named Entity Recognition (NER) system is a handcrafted rule-based system that relies on comprehensive lexical resources for Serbian implemented in Unitex[8] (Krstev et al., 2013). It recognizes most major types of NEs: names of persons, locations and organizations, temporal expressions, and numeric expressions, including measures, money, amount, and percentage. For recognition of some types of named NEs, e.g. personal names and locations, e-dictionaries and the information in them are crucial; for others, like temporal expressions, local grammars in the form of FSTs that try to capture a variety of syntactic forms in which a NE can occur had to be developed. However, for all of them, local grammars were developed that use the wider context to disambiguate ambiguous occurrences, as much as possible. The latest version of the Serbian NER system is organized as a cascade of transducers, which means that several FSTs are applied on a text, one after the other. Each of them recognizes some sub-type of NEs, adds an

[4] Recepti: http://www.recepti.com/.
[5] Letters *č* and *ć* are used as *c*, *ž* as *z*, *š* as *s*, while *đ* is replaced by *dj*.
[6] BalkaNet: http://www.dblab.upatras.gr/balkanet/index.htm.
[7] SWN: http://resursi.mmiljana.com/Default.aspx.
[8] Unitex: http://www.igm.univ-mlv.fr/~unitex/.

appropriate tag to a text, which the FSTs applied subsequently can use. The use of cascades enables, among other things, the distinction between amount expressions and other expressions that use numerals, like measurement expressions.

Measurement and amount expressions, and to some extent temporal expressions are the most interesting for application to a corpus of culinary recipes. Our NER system recognizes the measurement expressions in which metric and U.S. units are used (in the form of simple words, compounds, and abbreviations) and a count of units of measure is expressed by numerals consisting of digits, words, and their combination. The recognized expressions represent either exact values, ranges of values or approximate values. One example is: *parče tvrdog sira od oko 100-150g* 'a piece of hard cheese about 100-150g'. Our NER system recognizes as amount expressions the phrases in which a numeral is followed by a count noun (possibly preceded by one or more adjectives) that agrees with it in the values of grammatical categories.

The evaluation results of our NER system against a news corpus were very good: precision 0.98, recall 0.94 and F-measure 0.96 for all NEs measured in tokens, precision 0.96, recall 0.88 and F-measure 0.92 measured in types. For measurement expressions precision was 0.99, recall 0.97 and F-measure 0.98 measured in tokens, while for types it was: precision 0.97, recall 0.94 and F-measure 0.96 (Krstev et al., 2013). However, there were not many such expressions in the analyzed corpus, only 289 of them in a 155,000 words from corpus. Our NER system recognized 48,531 measurement expressions and 65,749 amount expressions in our recipe corpus. We have not yet performed an evaluation on this new text type, but we expect that the performance is not as good.

3. Units of Measure in the Culinary Domain

One characteristic of all the recipes is extensive use of measurement expressions. Full understanding of these expressions is crucial for culinary professionals as "food costing, recipe size conversion, recipe development, and cost control" depend on it (Blocker & Hill, 2007). Moreover, it helps to calculate the quantity of food that should be prepared in order to obtain portions of the right size, because most foods shrink during preparation (Jones, 2008). Kitchens in different environments (restaurants, schools, hospitals, etc.) have special considerations regarding quantities and nutrition values (Edelstein, 2008). In their everyday life, people want to calculate the calories in the food they are preparing.

In order to achieve this, it is necessary to know what the units of measure are and how they relate to each other. The list of units of measure used in cooking given in (Edelstein, 2008) includes: units of length, volume and mass (metric and U.S. units, and their rates), temperature (Celsius and Farenheit), as well as the relation between standard scoop and can sizes. In (Jones, 2008), count as a unit of measure is listed as well. Blocker & Hill (2007) divide measure units into customary (such as graduated measures and nested measuring cups) and proper measures.

Culinary recipes written by users for other users (and not professionals for other professionals) are specific in the use of units of measure. Count and standard units of measure are used together with many informal units. As a preliminary step in our research, we have analyzed our corpus in order to obtain a general understanding of the units of measure used in the Serbian recipes. For that purpose, we used the tools described in subsection 2.4.

First we turned to standard units of measure. As expected, U.S. units of measure – *inč* 'inch', *unca* 'ounce', *stepen Farenhajta* 'degree Farenheit', etc. – are not used at all. As far as units of length are concerned, only centimeters are used, usually in the part of the recipe that describes the procedure: *Testo razviti na 1 cm debljine* 'Roll out the dough to become 1 cm thick', *Pleh veličine 20cm x 28cm podmazati* 'Oil the pan size 20cm x 28cm'. Centimeters are used only occasionally in the part of recipes that lists ingredients: *7 kotleta debljine oko 2 cm* '7 chops around 2 cm thick', *Jedan komad rebara širok 10 do 20 cm* 'One piece of ribs 10 to 20 cm wide'.

Degrees Celsius are the only measure of temperature, used, although the closer description *Celzijus* is rarely mentioned – only six times in our corpus: *Ugrejati pećnicu na 200 stepeni Celzijusa* 'Warm the oven to 200 degrees Celsius'. This unit of measure is predominantly used to describe the preparation phase, and only a few times to describe preservation of food: *Idealna je temperatura čuvanja oko 10 stepeni C* 'Ideal storage temperature is 10 degrees C'.

Finally, for describing food preparation units of time are used as well: *minut* 'minute', *sat*, *čas* 'hour', *dan* 'day': *Koru sušiti 100 minuta* 'Dry thin dough for 100 minutes', *Čuvajte ga u frižideru 2-3 dana* 'Keep it in refrigerator 2-3 days'.

As can be expected, units of counting are frequently used as a unit of measure, either to designate an exact quantity – *2 velika krompira* '2 big potatoes', *tri cijela jajeta* 'three whole eggs' – or as an approximate quantity – *nekoliko crnih maslinki* 'a few black olives'.

At this moment, we are interested only in the units of measure specifying the ingredients used in recipes. We observe that this information is often expressed by units of measure that are used more or less informally and are not listed in professional manuals; however, they have to be taken into consideration in order to accomplish the tasks previously mentioned (e.g. to automatize conversion from approximate measures to standard measures).

4. Approximate Units of Measure in the Culinary Domain

Our first goal was to produce an extensive list of the approximate units of measure that are used in the culinary domain. In order to do that, we have used all the resources for Serbian described in the previous section.

Our first task was to retrieve the approximate units of measure from our culinary corpus. To achieve this, we had to distinguish between count and uncount units of measure. In the latter case we have taken the following approach:

> *If a noun is preceded neither by a numeral nor by a unit of measure and is followed by a noun in the genitive case that refers to food (and possibly preceded by an adjective in the corresponding case, gender, number and animacy) then it can be said it refers to an uncount unit of measure.*

Only a few were found in the produced concordances: *prstohvat* and *na vrh noža*, both meaning a very small

amount ‘a pinch of’. For retrieving count approximate units of measure, we have taken the following approach:

> *If an amount expression is followed by a noun in the genitive case referring to some kind of food (and possibly preceded by an adjective in the corresponding case, gender, number and animacy) then we can presume that the noun used in the amount expression refers to a count approximate unit of measure.*

Our goal was to retrieve as many expressions as possible that contain approximate units of measure; however, we were not aiming at comprehensiveness that can result in too many false retrievals. Thus, we deliberately omitted the cases where amount expressions were not used at all. Our experiment with proper units of measure showed that units of measure in the culinary domain are seldom used without numerals – actually, just 16 such cases were found, e.g. *decilitar pavlake* ‘a deciliter_of cream’. We can safely presume that we will retrieve the majority of the approximate units of measure by following our approach. FST modelling it is shown in Figure 1.

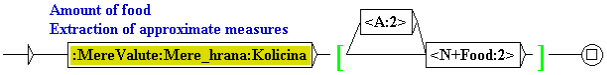

Figure 1. A FST that retrieves approximate units of measure. Agreement conditions are left out for simplicity purposes.

The FST in Figure 1 retrieved 15,521 lines of concordances; a few lines are shown in Figure 2. Only the candidates for units of measure can form part of concordance keywords, which facilitates the inspection of a large number of candidates and concordance lines. This is made possible by the use of contexts in graphs – the noun to which a unit of measure applies is used for retrieval but is not part of a keyword (green brackets in Figure 1). The same goes for numerals that are restricted to a context in the sub-graph **Kolicina** (the yellow box). The pattern <N+Food:2> retrieves all nouns in the genitive case related to food – both simple words (*vinobran* ‘potassium metabisulfite’) and compounds (*mladi luk* ‘fresh onion’).

3	**vezice**	crnog mladog luka
3	*small bunches*	*fresh onion*
jednu	**vezicu**	iseckanog peršunovog lista
one	*small bunch*	*chopped parsley leaves*
1	**vezu**	seckanog peršunovog lista
1	*bunch*	*chopped parsley leaves*
½	**vrećice**	praška za pecivo
½	*small pack*	*baking powder*
1	**vršna kašičica**	praška za pecivo
1	*peak small spoon*	*baking powder*
dva	**zrna**	suvog grožđa
two	*grains*	*raisins*
8-10	**zrnaca**	crnog bibera
8-10	*small corns*	*black papper*

Figure 2. A sample of the produced concordance lines.

The produced concordances were further analyzed by a volunteering student whose task was to select the candidates that represent approximate units of measure and mark those that are synonymous with other units of measure and/or are used only with some particular kind of food. As a result of this process, we obtained 106 approximate units of measure – 96 simple words and 10 compounds.

5. Approximate Measures Ontology and its Relation to Serbian Lexical Resources

A number of different ontologies of quantities and units of measure have been developed for different domains. For example, units of measurement ontology for biological and biomedical domains,[9] OASIS Quantities and Units of Measure Ontology Standard[10] for use across multiple industries, EngMath[11] for mathematical modeling in engineering, or Quantities, Units, Dimensions and Data Types Ontologies[12] and Ontology of Units of Measure (Rijgersberg et al., 2013) for a vast variety of quantitative research purposes, etc. The characteristic feature of these ontologies is that their scope is limited to formal measures, most frequently based on technical standards. Our goal is to develop an ontology for the informal measures specific to the culinary domain discussed in the above sections.

In order to enable semantic tagging of the approximate units of measure in the culinary domain, we modeled the OWL ontology. The ontology was modeled in the OWL 2 web ontology language[13] using the Protégé 4.3 tool,[14] because it makes it possible to establish a connection between classes and instances.

As to the discussed observations about the approximate measures in culinary recipes, we chose to use the introduced semantic categories as ontology classes, and the extracted units as instances. On the basis of the approximate units of measure extracted from our corpus, we introduced the following sub-classes of the top class *PribliznaMera* ‘ApproximateMeasure’: *Kontejner* ‘Container’, *Porcija* ‘Portion’, *DeoOd* ‘PartOf’, *Celina* ‘Whole’, and *Skupina* ‘Set’. Additionally, we proposed the object relationship property *jeManja* ‘isSmaller’ and the inverse property *jeVeca* ‘isBigger’ to signify that an approximate unit of measure is a smaller or bigger unit than another one from the same class. These classes with some instances from the class *Skupina* ‘Set’ are shown in Figure 3: *vezica* ‘small bunch’, *veza* ‘bunch’, *šaka* ‘handful’, *red* ‘row’; and the relationship property *jeManja*: *vezica jeManja veza* ‘small bunch *isSmaller* bunch’.

The analysis of concordances revealed that some approximate units of measure are used only for some particular kinds of food (or a restricted set), like *čen belog luka* ‘clove of garlic’ and *ploča lisnatog testa* ‘plate of puff pastry’ or *ploča lazanji* ‘plate of lasagna’, which is enforced in our ontology by introducing appropriate data properties *jeJedino* ‘isOnly’ and *jeIskljucivo*

[9] Units of measurement ontology: http://www.obofoundry.org/cgi-bin/detail.cgi?id=unit.
[10] OASIS QUOMOS: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=quomos.
[11] EngMath: http://www-ksl.stanford.edu/knowledge-sharing/papers/engmath.html.
[12] QUDT: http://www.qudt.org/.
[13] OWL 2: http://www.w3.org/TR/owl2-overview/.
[14] Protégé: http://protege.stanford.edu/.

'isExclusively'. On the other hand, there are instances that belong to more than one class. Such is the case with *zrnce* that belongs to the class *PartOf* meaning 'grain' when referring to *biber* 'pepper' or *mak* 'poppy' and to the class *Portion* meaning 'small spherical amount of' when referring to *puter* 'butter'.

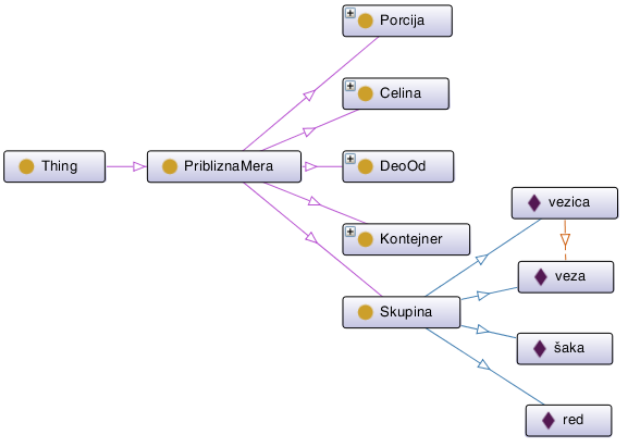

Figure 3. The hierarchy of approximate measures ontology classes, some instances, and their relationships.

Another very important aspect in ontology development is the possibility to designate that two or more instances refer to the same object. For example, *čen* and *češanj* '*clove of garlic*', and *štangla* and *rebro* 'bar of chocolate' should be treated as the same unit in culinary recipes. In the Serbian language, *šoljica za kafu*, *kafena šoljica*, *šoljica* and *kafena šolja* are different expressions for 'coffee cup'. It is sufficient to designate that information and the property of one of these instances, e.g. *šoljica jeManja šolja* 'cup *isSmaller* mug' in the ontology, for the reasoner to infer that the same is true for the other three instances (see Figure 4).

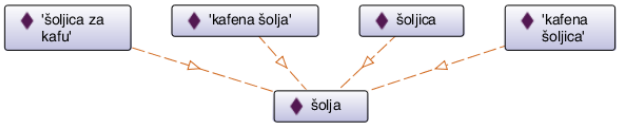

Figure 4. The same instances to which the property *isSmaller* is assigned.

Semantic marker	Description	Number of instances
+MesApp	approximate measure	106
+Cont	container (e.g. *supena kašika* 'soup spoon'	33
+Por	portion (e.g. *kriška* 'slice')	33
+Part	part of (e.g. *glavica* 'head')	30
+Wh	whole (e.g. *štapić* 'stick')	7
+Set	set (e.g. *veza* 'bunch')	4

Table 2. Semantic markers for approximate units of measure, typical representatives of classes and the number of instances in classes (some measures are in more than one class).

The ontology contains 7 classes, two object properties, two data properties, and 106 individuals. The ontology classes were used in the creation of a domain-specific e-dictionary of approximate units of measure. The new semantic markers and some representative instances from the classes are presented in Table 2.

Finally, we manually checked all the selected approximate units of measures against the SWN, and all those not already in it were added. This was not a straightforward task, because approximate measures in the culinary domain do not have a particular place in the PWN, and thus they do not have it in the SWN either. The only exception, to a certain extent, is 'containerful'. During this process, some units of measure were moved from one class to another that better corresponded to the PWN. For instance, *šaka* 'handful' was originally put in the class *Set*, but was afterwards moved to the class *Container* (because, 'handful' is a hyponym of 'containerful' in the PWN).

The work we have done is fully justified by the data presented in Table 3. In our culinary corpus, we have counted the expressions that use the units of measure applied to nouns representing some kinds of food by using the appropriate graphs. We cannot give an estimate of recall, but precision is very high (around 100% for the first column).[15] It can be seen that almost 45% of these expressions use approximate units of measure.

Units	With numerals	Without numerals	Total
Standard units	12,966	16	12,982
Approximate units	7,431	2,933	10,364

Table 3. Statistics of the use of units of measure in our culinary corpus.

Although this kind of knowledge could be, to some extent, represented in e-dictionaries and semantic networks, ontologies are much more suitable for useful reasoning. Moreover, the contribution of this ontology is the possibility of its integration in a comprehensive culinary domain recipe ontology on which we are currently working. To be more specific, in most cases, the culinary recipe structure is as follows: the name of the recipe (i.e. the meal that is in the focus of the recipe), the name of the author of the recipe, the part with listed ingredients that are required for recipe preparation together with the quantities, preparation description (usually listed in the steps that give a detailed account of the necessary utensils and way of preparation directions), and additional information like a summary of the recipe's nutritional values, preparation time or the level of preparation difficulty. Through detailed analysis of each of these parts, we came to a conclusion that it is necessary to develop a number of ontologies suitable for individual parts, which will later be integrated into a comprehensive culinary domain recipe ontology to represent the knowledge of the culinary domain.

[15] The produced concordances can be inspected at: http://poincare.matf.bg.ac.rs/~stasa//concordances/.

6. Conclusion

Our job is not yet finished, because there are still parts of the food ontology, e-dictionaries and WordNet that have to be filled. One major part still missing is related to the ways of preparation of food. However, the parts already developed can help in this. For instance, adjectives (and verbs) related to food preparation can be retrieved using the following procedure:

> *If an adjective derived from a verb past participle is preceded by numeric expressions with units of measure (standard and approximate) and followed by a noun in the genitive case that refers to food (and possibly preceded by an adjective in the corresponding case, gender, number and animacy) then it can be an adjective referring to a way of preparation of food.*

A graph developed following this approach retrieves with an almost 100% precision 1,805 concordance lines from our corpus related to adjectives (and the corresponding transitive verbs) we are looking for. From these, we have selected 85 adjectives and the corresponding verbs related to the culinary domain that vary from very general ones, like *pripremljen* 'prepared' and *pripremiti* 'prepare' to very specific ones like *pošećeren* 'sugared' and *pošećeriti* 'add sugar'. By these verbs as seeds for retrieval of more verbs and by modelling more procedures like this, we plan to prepare an exhaustive list of adjectives and verbs related to the culinary domain.

As was discussed in the previous section, we also plan to develop different ontologies like foodstuffs ontology, food product ontology, kitchen utensil ontology etc., in our future work in order to integrate all of them in one comprehensive culinary ontology and test in various applications.

7. 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 as part of the project no. 178006, financed by the Serbian Ministry of Science.

8. References

Ahnert, S.E., 2013. Network analysis and data mining in food science: the emergence of computational gastronomy. In *Flavour* **2**:4. URI: http://dx.doi.org/10.1186/2044-7248-2-4.

Badra, F., R., Bendaoud, R., Bentebibel, P.A., Champin, J., Cojan, A., Cordier, S., Després et al., 2008. Taaable: Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In *9th European Conference on Case-Based Reasoning-ECCBR 2008, Workshop Proceedings*, 219-228.

Batista, F., Pardal, J.P., Vaz Nuno Mamede, P., and R. Ribeiro, 2006. Ontology construction: cooking domain. *Artificial Intelligence. Methodology, Systems, and Applications* 4183 (2006): 213-221.

Blocker, L., and J. Hill, 2007. *Culinary Math*. John Wiley & Sons.

Cantais, J., Dominguez, D., Gigante, L., Laera, and V. Tamma, 2005. An example of food ontology for diabetes control. In *Proceedings of the International Semantic Web Conference 2005 workshop on Ontology Patterns for the Semantic Web*.

Chakkrit, S., and M. Bruckner, 2008. FOODS: a food-oriented ontology-driven system. In *Digital Ecosystems and Technologies. DEST 2008. 2nd IEEE International Conference on*. IEEE.

Courtois, B., L. M., Silberztein, et al., 1990. Dictionnaires électroniques du français. *Langue française*, 87(1):3–4. Armand Colin.

Dufour-Lussier, V., F., Le Ber, J., Lieber, T., Meilender, and E. Nauer, 2012. Semi-automatic annotation process for procedural texts: An application on cooking recipes. *arXiv preprint arXiv:1209.5663*.

Edelstein, S., 2008. *Managing Food and Nutrition Services: For the Culinary, Hospitality, and Nutrition Professions*. Jones & Bartlett Learning.

Fellbaum, C.. 2010. *WordNet*. Springer.

Jones, T., 2008. *Culinary Calculations: Simplified Math for Culinary Professionals*. John Wiley & Sons.

Kim, E., 2012. Korean Food Ontology. URL: http://kr-med.org/icbofois2012/fois/posters/kim_a4.pdf.

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

Krstev, C., and D. Vitas, 2006. Finate State Transducers for Recognition and Generation of Compound Words. In *Proceedings of the 5th Slovenian and 1st International Conference Language Technologies*, IS-LTC 2006, Ljubljana, Slovenia, October, 2006, eds. T. Erjavec and J. Žganec Gros, 192-197, Institut "Jožef Stefan".

Krstev, C., I., Obradović, M., Utvić, and D. Vitas, 2013. A system for named entity recognition based on local grammars. *Journal of Logic Computation* 24(2): 473-489, Oxford Journals, doi:10.1093/logcom/exs079.

Mori, S., T., Sasada, Y., Yamakata, and K. Yoshino, 2012. A machine learning approach to recipe text processing. In *Computers workshop (CwC)*:29.

Nedovic, V., 2013. Learning recipe ingredient space using generative probabilistic models. In *Proceedings of International Joint Conference of Artificial Intelligence Workshops*, 13-18.

Pizzuti, T., G., Mirabelli, M.A., Sanz-Bobi, and F. Goméz-Gonzaléz, 2014. Food Track & Trace ontology for helping the food traceability control. *Journal of Food Engineering* 120 (2014): 17-30.

Rijgersberg, H., van Assem, M., and J. Top, 2013. Ontology of units of measure and related concepts. *Semantic Web 4*, no. 1 (2013): 3-13.

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

Vujičić Stanković, S., C., Krstev, and D., Vitas, 2014. Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain. In *Proceedings of the seventh Global Wordnet Conference*.

Wiegand, M., B., Roth, and D. Klakow, 2012. Knowledge Acquisition with Natural Language Processing in the Food Domain: Potential and Challenges. In *Proceedings of the ECAI-Workshop on Cooking with Computers (CWC)*: 46-51.