SCIENTIFIC PAPER

Wordnet-Based Serbian Text Categorization

UDC 811.163.41'322.2

Jelena Graovac
jgraovac@matf.bg.ac.rs
University of Belgrade, Faculty of Mathematics, Department for Computer Science

ABSTRACT: A Serbian text categorization technique, based on the Serbian wordnet is presented. The author is guided by the hypothesis that the inclusion of morphological, syntactic and semantic information contained in lexical resources can improve the process of text documents categorization in Serbian, as one of morphologically rich languages. Ebart-3 corpus is used for driving experiments. It is a collection of newspaper articles in Serbian divided into three categories: Economics, Politics and Sport. The method is based on lists of representative synsets (for each category) from the Serbian wordnet and category assignment function, defined on the basis of these lists. Selection of representative synsets is based on the significance weight measure of a synset for the considered category. Inflection problem in Serbian is solved by means of the system of morphological dictionaries for Serbian. In order to evaluate the presented technique, micro- and macro-averaged Precision, Recall and F1 measures are used. For comparison purpose, another technique based on wordnet-encoded semantic domains is also developed. Instead of well-chosen synsets, representative lists for categories consist of all synsets that belong to semantic domains corresponding to the considered categories. The results show that the technique based on well-chosen synsets outperforms the technique based on semantic domains, although the main reason for enriching wordnet by semantic domains is its even more successful application in natural language processing tasks, especially in text categorization.

KEYWORDS: Natural Language Text Categorization, Serbian Wordnet, the System of Morphological Dictionaries for Serbian

DATE OF SUBMISSION: 13 November2013
DATE OF ACCEPTANCE: 18 March 2014

1. Introduction

We live in a world where the Internet and digital recording have made huge amounts of raw data available to the public. A frustrated management information systems executive a long time ago said: "Computers have promised us a fountain of wisdom but delivered a flood of data" (Frawley et al. 1992). Documents in their textual semi-structured data formats (or raw data), with different content and quality are rarely useful. It is necessary to prepare these raw data for analysis, to transform them into information and to transform information into invaluable knowledge. Data mining, also known as knowledge-discovery in databases, is an interdisciplinary subfield of computer science which aims at automatic or semi-automatic analysis of large quantities of data in order to extract previously unknown interesting patterns. It can

be defined as nontrivial extraction of implicit, previously unknown, and potentially useful information from data. One of the fundamental tasks in Data mining is Categorization.

Text categorization is the task of classifying unlabeled natural language documents into a predefined set of categories. This may be done manually, but this is time-consuming and expensive. Due to the widespread availability of fast computers, automatic classification of documents has become the key approach to efficient organizing and processing large amounts of information and to knowledge discovery. Some of the most commonly used machine learning techniques that have been applied to automatic text classification are: K Nearest Neighbors (kNN), Support Vector Machines (SVM), Decision Trees, Bayesian classifier, Neural Networks, Hidden Markov Models (HMM) etc. Most information (common estimates say over 80%) is currently stored as natural language (language that people use for everyday communication) text documents.

Serbian language belongs to the group of morphologically rich languages. It uses two alphabets – Cyrillic and Latin, it has phonologically based orthography, the rich morphological system, free word order, special placement of enclitics, and complex agreement system. From the computational point of view, all these characteristics have to be taken into consideration before attempting to process Serbian written texts (Vitas and Krstev 2005).

Lexical resources for Serbian have been developed within the Human Language Technologies Group at the Faculty of Mathematics, University of Belgrade (Vitas et al. 2003). The motivation for this work is the question of how the information contained in the rich lexical resources can be efficiently utilized in order to solve the problem of text documents categorization in Serbian. This paper presents an improved variant of the text categorization technique based on Serbian wordnet, presented in (Pavlović-Lažetić and Graovac 2010).

The rest of the paper is organized as follows. Section 2 shows some background information about the problems of text categorization and document representation. This section also presents lexical recourses for Serbian used in this work – Serbian wordnet and the system of morphological dictionaries for Serbian. Section 3 gives a brief discussion of related work. Data corpus in Serbian used for text categorization process is presented in Section 4 and Section 5 presents evaluation metrics used to assess the performance of the technique. Section 6 describes technique and presents categorization procedure. Significance weight measure of a synset for a given category and the category assignment function are defined. Sections 7 and 8 report on experimental results obtained by the presented new wordnet-based text categorization technique and its comparison with a similar technique based on wordnet-encoded semantic domains. Section 9 concludes the paper.

2 Background

2.1. Text Categorization Problem

Formally, categorization of text consists of associating a Boolean value to each pair $(d_j, c_i) \in D \times C$, where D is the set of text documents and C is the set of categories. The value T (True) is then associated to the pair (d_j, c_i) if the text document d_j belongs to the category c_i while the value F (False) will be associated to it otherwise. The goal of the categorization of text is to approximate the unknown target function $\bar{\phi}: D \times C \rightarrow \{T, F\}$ (that describes how documents ought to be classified) by means of a function $\phi: D \times C \rightarrow \{T, F\}$ called the classifier, such that ϕ and $\bar{\phi}$ "coincide as much as possible" (Sebastiani 2002).

Text categorization process includes three main components:

1. Document representation.
2. Building a text classifier using algorithms that learn classification patterns from a large number of examples (training dataset).
3. The classifier evaluation on the new text documents (testing dataset).

2.2. Document Representation

Document representation is one of the preprocessing steps essential for text categorization. We need an effective document representing model to build an efficient classification system. The role of the document representation component is to represent text document so as to facilitate machine manipulation but also to retain as much information as needed. A text document d_j is usually represented as a vector of term weights $d_j = \{w_{1j}, w_{2j}, ..., w_{|T|j}\}$ where T is the set of terms that occur at least once in at least one document of the training set, and $0 \leq w_{kj} \leq 1$ represents, loosely speaking, how much term t_k contributes to the semantics of the document d_j (Sebastiani 2002). Choosing a suitable level of text analysis on which to base the definition of terms is a trade-off between semantic expressivity and representational complexity. We distinguish between five levels of text analysis (Joachims 2002):

1. Sub-word level: decomposition of words and their morphology.
2. Word level: words and lexical information.
3. Multi-word level: phrases and syntactic information.
4. Semantic level: the meaning of text.
5. Pragmatic level: the meaning of text with respect to context and situation.

The n-grams approach (where text is represented on the sub-word level) has often been used for indexing (Graovac 2012; Graovac to appear 2014), but the most widely-used approach for indexing is a commonly called bag-of-words approach (Lewis and Ringuette 1994). In this model, a text document is represented as the bag (multiset) of its words, disregarding word order but keeping multiplicity. One of the main challenges in the bag-of-word document representation is high dimensionality of data vectors. Therefore there is a need for dimensionality reduction. Two main approaches are used for reducing dimensionality: *feature selection*, which is used to select the most relevant attributes (words) and *feature extraction*, which is used for combining attributes into a new reduced set of relevant features for building robust learning models. The most popular feature selection methods are: Stop Words Elimination, Word Frequency – Inverse Document Frequency, Mutual Information, $\chi 2$ test, Gini Index, Expected Cross Entropy etc. while the most popular feature extraction methods are: Stemming, Lemmatization, Thesaurus, Latent Semantic Indexing, Conceptual Indexing etc.

2.3. Lexical Resources for Serbian

One of the main tasks of the Natural Language Processing Group at the Faculty of Mathematics, University of Belgrade is the development of various lexical resources. Among them the two most important ones are: the Serbian wordnet (SWN) developed in the scope of the Balkanet project and the system of morphological dictionaries for Serbian (SMD) (Krstev et al. 2004) in Unitex format (http://igm.u-pem.fr/~unitex/) (Sébastien 2002)

Wordnet

Wordnet (also known as Princeton WordNet, PWN) is a manually constructed lexical system developed by George Miller and his colleagues at the Cognitive Science Laboratory at Princeton University. Its aim was to serve as a sort of a mental lexicon that can be used in the scope of psycholinguistic research projects (Fellbaum 2010). A traditional dictionary lists lexical items alphabetically, giving definitions for each sense. Wordnet, in contrast, is based on word meaning; all of the words that can express a given sense are grouped together in a *synonym set*, or *synset*. The outstanding multilingual initiative is EuroWordNet (EWN). It introduced multilingualism into the semantic network of concepts by building wordnets for seven European languages in a manner similar to PWN, and aligning them by interconnecting synsets representing the same concept in different languages by an Inter-Lingual-Index, or ILI (Vossen 1998). Along the same lines, the goal of BalkaNet project (Tufiş et al. 2004) is the development of aligned semantic networks for Bulgarian, Greek, Romanian, Serbian, and Turkish, while at the same time extending the existing network for Czech, initially developed within the EWN. The main aim of the BalkaNet project was the development of a

INFOtheca

modern language resource for Balkan languages that would enable new access to information that is expressed within Balkan languages (Krstev et al. 2004).

Serbian wordnet

SWN (http://korpus.matf.bg.ac.rs/SrpWN) represents a lexical semantic network for Serbian language (Krstev et al. 2004). Structure of SWN is basically the same as the structure of the PWN. It is organized through the nodes and the relationships established between those nodes, which are termed synsets in Wordnet. Each synset contains a group of synonymous words or literals (denoted by a "literal string"), followed by a "sense tag" which represents the specific sense of the literal string in that synset (as in any explanatory dictionary where an entry corresponding to a word is followed by a number of its possible meanings). Different senses of a word are in different synsets. The meaning of the synsets is further clarified by short defining glosses (definitions and/or example sentences).

SWN is divided, according to the part-of-speech, into nouns, verbs, adjectives, and adverbs. Table 1 shows synsets distribution of SWN by part-of-speech, as of January 31, 2013.

Part-of-speech	Serbian wordnet
Nouns	14765
Verbs	2104
Adjectives	1380
Adverbs	117
In total	18366

TABLE 1: Distribution of synsets in SWN, as of January 31, 2013.

The nominal part of the SWN is organized as hierarchies of nodes, which are established on the basis of the relation of subordination (*hyponym*) and superordination (*hypernym*) between the meanings represented by corresponding nodes. One synset is subordinated to another synset not only if it has all the features of the superordinated synset, but also if it has some specific additional features as well (Krstev et al. 2004). There is a level in the hierarchies of nominal synsets, somewhere in the middle, where most of the distinguishing features are attached. It is referred to as the base level of the noun lexicon, and synsets at this level are "basic synsets". These synsets are neither too specific nor too general. In (Graovac and Pavlović-Lažetić 2008), a measure of productivity of a synset is introduced, representing the extent of the hierarchy that the considered synset, belonging to it, effectively represents.

In addition to the *hyponym* and *hypernym* relations, there are also holo_part and holo_member relations. For example, synset {pas:C1x, pseto:1, domacxi pas:1} (in English {dog:1, domestic dog:1, Canis familiaris:1}) is connected by the relation *hyponym* with the synset {pas:C1} (in English {canine:2, canid:1}), synset {rep:2a} (in English {flag:7}) is connected by the relation *holo_part* with the synset {pas:C1x, pseto:1, domacxi pas:1} (in English {dog:1, domestic dog:1, Canis familiaris:1}) and this synset is connected by the relation *holo_member* with the synset {cyopor:1} (in English {pack:6}). Figures 1 and 2 show an idealized model and XML representation of the part of the SWN that illustrates this example. Another important relation between noun and adjective synsets is antonymy (*near_antonym*) that connects the synsets that have (almost) opposite meaning. Also important relation that is established between the synsets is the relation that connects the concepts that are lexicalized by different parts-of-speech. The important relation that connects a noun synset with an adjective synset is the relation *be_in_state*. An example is the synset {cyistocxa:1} (in English {cleanness:1}) that is connected with the adjective synset {cyist:1a} (in English {clean:1}). The synset {cyist:1a} (in English {clean:1}) is connected by the relation *near_antonym* with the synset {prlxav:1} (in English {dirty:1, soiled:1,unclean:1}). This synset is again in relation to the noun synset {prlxavsxtina:1, necyistocxa:1} (in English {dirtiness:1, uncleaness:1}) through the relation *be_in_state*, while this synset is, in turn, again connected by the relation *near_antonym* with the initial synset {cyistocxa:1} (in English {cleanness:1}). If relations that are established between verbal synsets are added, such as *causes* that connect, for

example, synsets {uspraviti:1, podignuti:3} (in English {stand:10; stand up:2; place upright:1}) and{stajati:1a} (in English {stand:1; stand up:4}) it is clear that the wordnet represents a large semantic network (Krstev et al. 2004). Table 2 shows distribution of all relations between synsets in SWN, as of January 31, 2013.

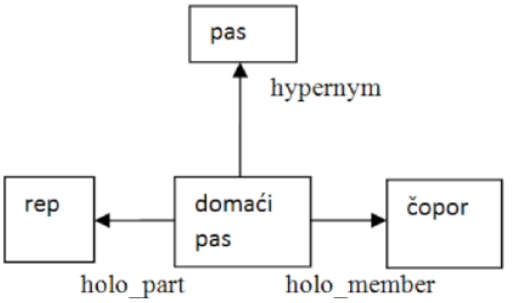

FIGURE 1: Part of the SWN–idealized model

The structure of PWN was enhanced several times with additional information in order to make it even more usable in various natural language applications (especially in text classification). One of these extensions is related to the introduction of semantic domains that provide a natural way to establish semantic relations between the meanings of concepts. Semantic domains are areas of human interests such as sports, economics or politics, which exhibit their own terminology and lexical coherence. Wordnet synsets have been annotated with at least one semantic domain label, selected from a set of about two hundred labels structured according to the WordNet Domain Hierarchy (http://wndomains.fbk.eu/hierarchy.html). Serbian wordnet has not been expanded by semantic domains but they can easily be obtained from the PWN. An example of a synset {sport:1, bavlxenxe sportom:X} (in English {sport:1, athletics:1}) is presented in Figure 3, using VisDic graphical application (Horák and Pavel 2004). The left side of the picture shows this synset in PWN and the right side shows the same synset in SWN. This synset is annotated in PWN by semantic domain "sport" so we will consider that the same domain is associated with the corresponding synset in SWN.

```
<SYNSET>
      <ID>ENG30-02083346-n</ID>
      <SYNONYM>
            <LITERAL>pas</LITERAL>
      </SYNONYM>
      <DEF>Bilo koji od raznovrsnih
           sisara koji obicyno imaju
           dugu nxusxku i kandye.
      </DEF>
      <POS>n</POS>
</SYNSET>

<SYNSET>
      <ID>ENG30-02084071-n</ID>
      <SYNONYM>
            <LITERAL>pas</LITERAL>
            <LITERAL>pseto</LITERAL>
            <LITERAL>domacxi pas</LITERAL>
      </SYNONYM>
      <DEF>Pripadnik Canis familiaris,
           srodan vuku, pripitomlxen od
           preistorijskog doba;
           postoje mnoge rase.
      </DEF>
      <POS>n</POS>
      <ILR>ENG30-02083346-n
            <TYPE>hypernym</TYPE>
      </ILR>
      <ILR>ENG30-07994941-n
            <TYPE>holo_member</TYPE>
      </ILR>
</SYNSET>

<SYNSET>
      <ID>ENG30-02158846-n</ID>
      <SYNONYM>
            <LITERAL>rep</LITERAL>
      </SYNONYM>
      <DEF>Upadlxivo oznacyen ili oblikovan
           zadnxi deo.</DEF>
      <POS>n</POS>
      <ILR>ENG30-02084071-n
            <TYPE>holo_part</TYPE>
      </ILR>
</SYNSET>

<SYNSET>
      <ID>ENG30-07994941-n</ID>
      <SYNONYM>
            <LITERAL>cyopor</LITERAL>
      </SYNONYM>
      <DEF>Grupa zxivotinxa koje love.</DEF>
      <POS>n</POS>
</SYNSET>
```

FIGURE 2: Part of the SWN – XML representation

INFOtheca

Relation	Number of occurrence
Hypernym	16886
holo_member	3879
eng_derivative	2926
holo_part	1560
near_antonym	762
category_domain	738
Derived	662
be_in_state	287
similar_to	244
also_see	226
holo_portion	209
verb_group	170
region_domain	82
Subevent	78
Causes	64
derived_gender	38

Relation	Number of occurrence
derived_pos	45
usage_domain	15
Particle	10
derived_vn	2

TABLE 2: Distribution of relationsbetween synsets in SWN, as of January 31, 2013

The System of Morphological Dictionaries for Serbian

The system of morphological dictionaries for Serbian (SMD) (http://korpus.matf.bg.ac.rs/SrpMD/) (Krstev 2008, Vitas 2003) in Unitex format, consists of a dictionary of simple lemmas (DELAS), a dictionary of compounds (DELAC), the corresponding dictionaries of word forms (DELAF), and morphological finite-state automata that model certain classes of lemmas. The information that has to be

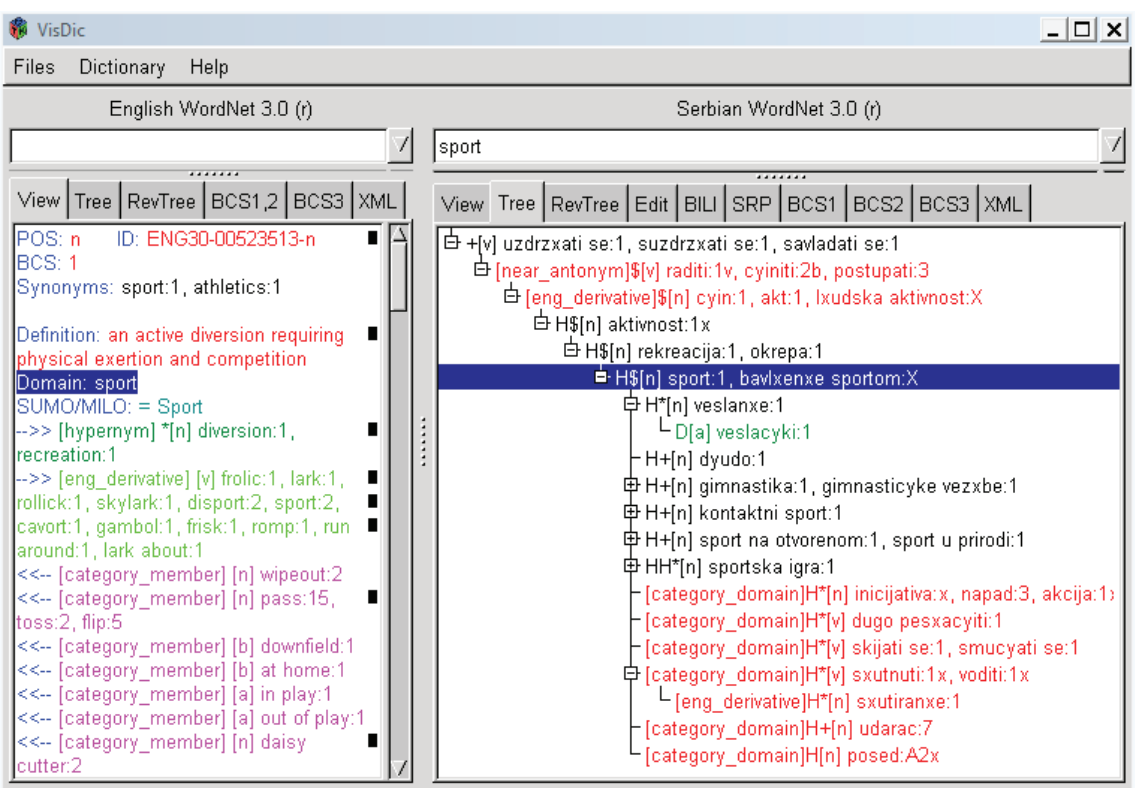

FIGURE 3: Example of synset {sport:1} in SWN and PWN, using VisDic software

assigned to every entry in the DELAS dictionary is the part-of-speech and the code of the inflectional class (for inflectional lemmas). Optional morphosyntactic, semantic and information on dialect can also be added. An example of an entry in the dictionary of simple lemmas (DELAS) is (Obradović and Stanković 2007)

devojcyin, A1+Pos+Ek (1)

which means that the lemma devojcyin (in English girl's) is an adjective belonging to inflectional class A1. The adjective is possessive (+Pos), in ekavian pronunciation (+Ek). The information assigned to a lemma in DELAS can be used by Unitex (Sébastien 2002) to formulate complex queries. For example, the query <A+Pos-Ek> would retrieve all the possessive adjectives from a text that do not belong to the Ekavian pronunciation (Krstev et al. 2004). The entries in DELAS can be enriched by derivational links that group together entries belonging to the same derivational nest. This kind of information is given after an underscore sign. For example,

devojcyin, A1+Pos+Ek_N=4ka (2)
devojka, N618+Hum+Ek_A=2cyin

The information in the first line states that the adjective devojcyin is linked to the noun entry and also indicates the way to identify this noun in the dictionary. Conversely, the information in the second line links the noun to the adjective. Moreover, the morphosyntactic information, preceded by a plus sign, can describe the type of derivational relation between two entries. In the example

(2), the adjective *devojcyin* is the possessive adjective of the noun *devojka* (in English *girl*). This information in the DELAS dictionary can be used by finite transducers to lemmatize the text using any lemma, arbitrarily chosen, from the derivational nest (Krstev et al. 2004). DELAS dictionary can also be used to obtain all the inflected forms of a word.

3. Related Work

One of the best known wordnet-based text categorization techniques applied to a corpus in English, which served as motivation for the method developed and presented in this paper, was introduced by Scott and Matwin (1998). They presented a categorization procedure that requires three passes through the corpus. During the first pass, tagger assigns a part-of-speech tag to each word in the corpus. During the second pass, all nouns and verbs are looked up in PWN and a global list of all synonym and hypernym synsets is assembled. Infrequently occurring synsets are discarded, and those that remain form the feature set. During the third pass, the density of each synset (defined as the number of occurrences of the synset in the PWN divided by the number of words in the document) is computed for each example, resulting in a set of numerical feature vectors. The calculations of frequency and density are influenced by the value of a parameter *h* that controls the height of generalization. This parameter can be used to limit the number of steps upward through the hypernym hierarchy for each word. Association of a text document with a specific category is defined using the obtained density of concepts. For instance, in the case of two categories, *history* and *taxes*, algorithm has learned a simple rule saying that if the synset *possession* has a low density, the document probably belongs in the history category. The hyponyms of possession include words such as ownership, asset, and liability – the sort of words often used during discussions about taxes, but rarely about history.

In (Rodrigues et al. 2000) Wordnet is used for improving text categorization methods based on Neural Networks. This technique was applied to Reuters-21578 newspaper collection. Using Wordnet for text categorization was also presented in (Rosso et al. 2004). In these two papers, Wordnet is used only for obtaining synonyms of the words.

The text categorization technique that will be presented in this paper is an improved variant of the technique presented in (Pavlović-Lažetić and Graovac 2010), where Ebart-5 corpus was used. This corpus consists of articles in the

INFOtheca

columns sport, economics, politics, culture and entertainment, chronicle and crime. The presented algorithm required following steps: 1. key words for each column and each category are identified as the most frequent words in a set of articles from the given column/category (training set); 2. SWN synsets containing the chosen key words, along with all of their hyponyms, are assigned to the corresponding categories; 3. Category assignment functions are defined for an article (from the test set) in different ways, the simplest being the maximum number of occurrences of literals from the hierarchy rooted in the synsets assigned to the category, maybe filtered by domains. Inflection problem was solved algorithmically (assuming that two words with long enough common prefix are the same word in different inflectional forms) and a dictionary was not used for that purpose.

4. Data corpus

The first step in the machine-learning text categorization process is collecting text documents into a corpus and dividing them into training and test datasets. In this paper Ebart-3 corpus, a subset of the Ebart corpus is used. Ebart (www.arhiv.rs) is the largest digital media corpus in Serbia with almost one million news articles from daily and weekly newspapers archived from early 2003 onwards. Within it, complete editions of fifteen daily and weekly newspapers published in Serbia are stored, as well as selected articles from the biggest local weekly. The current archive is classified into thematic sections following the model of regular newspaper columns (e.g. "Sport", "Politics", "Economics", "Chronicle", "Culture", "World", "Society", etc.). Ebart-3 corpus consists of articles from the Serbian daily newspaper "Politika" that belong to columns "Sport", "Economics" and "Politics", published from 2003 to 2006. There are 3366 such articles. This dataset is single-labeled and it is split into the training and test datasets in the ratio 2:1. Fig. 4 shows the distribution of this corpus.

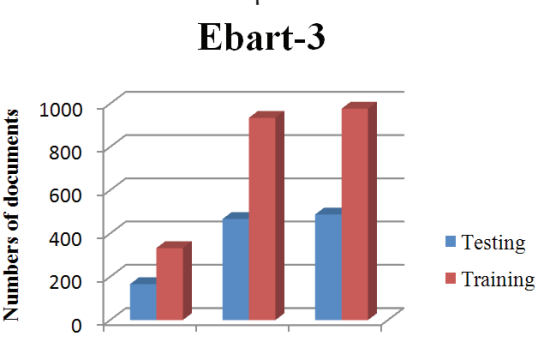

FIGURE 4: Distribution of the Ebart-3 corpus

5. Evaluation Metrics

Evaluation process consists of comparing the category known in advance with those proposed by the classifier. Most of evaluation metrics for two-category problems are built over a 2 x 2 confusion matrix as illustrated inTable 3. From this matrix, four simple measures can be directly obtained: TP and TN denote the number of positive and negative cases correctly classified, while FP and FN refer to the number of misclassified positive and negative examples, respectively.

Category C		Expert Judgment	
		Yes	No
Classifier	Yes	TP	FP
Judgment	No	FN	TN

TABLE 3: The confusion matrix for two-category problem

Typical evaluation metrics that come from information retrieval, *Precision* and *Recall*, are defined in terms of these sets, as follows (Baeza-Yates et al. 1999):

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}$$

Precision (or purity) is defined as the proportion of positive cases that are actually correct while *Recall* is the percentage of correctly classified positive examples. One of the most widely-used measures that combines Precision and Recall and gives both of them equal importance is the F1-measure introduced by van Rijsbergen (1979):

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The presented evaluation measures are applicable to two-category problems. When we have more than two categories, in order to obtain a single measure for the evaluation of a classification as a whole, the evaluation measures need to be averaged overall the categories. There are two ways to do this. Micro-average measure is the global calculation of measure considering all the documents as a single dataset, regardless of categories and macro-average measure is the average on measure scores of all the categories. Macro-average measure gives equal weight to each category, while micro-average measure is per document function, so it is heavily influenced by larger categories. In this paper we used micro- and macro-averaged Precision, Recall and F1 measures.

6. Text Categorization Procedure

Let us denote the categories Economics, Politics and Sport of Ebart-3 training dataset, and documents that belong to them as follows: $E = \{d_1, d_2, ..., d_{N_E}\}$, $P = \{d_1, d_2, ..., d_{N_P}\}$ and $S = \{d_1, d_2, ..., d_{N_S}\}$, where $N_E = 333$, $N_P = 935$ and $N_S = 977$ are the numbers of documents that belong to these categories, respectively.

Training phase of the text categorization procedure goes through the following steps:

• *For each category, list the basic form of words that occur in at least one document of that category in the training set, arranged by descending frequency*. The frequency of the word means total number of occurrences of all inflection forms of that word. For instance, the word "sport" has the following inflection forms defined in the Serbian morphological dictionary: "sport", "sporta", "sportu","sporte", "sportom", "sportovi", "sportova", "sportovima","sportove". We will assume that the word "sport" appears 90 times in category Sport if 90 is the total number of occurrences of all inflection forms of this word, in all the documents that belong to the category Sport in the training dataset.

• *Select the key words for each category*. In the list of basic words defined in the previous step, assign a part-of-speech tag to each word: nouns, verbs, or any of the other eight parts-of-speech in the Serbian language (pronouns, adjectives, numerals, adverbs, prepositions, interjections, particles or conjunctions). Discard from the list all the words that are not nouns or verbs. Select key words from the obtained list as the words that are distinctive for the considered category (most frequent for that category and not so frequent for other categories).

• *For each category, define candidates for category representative list of synsets from Serbian wordnet*. Candidates for category representative list of synsets from Serbian wordnet are chosen as synsets that encompass as many of the selected key words for that category, obtained in the previous step.

DEFINITION 1: A word x is *encompassed* by a synset s in SWN, if x (in its basic form) is contained in the synset s as a synonym string literal or along with all synsets that are connected with s by some of the chosen lexical or semantic relationships.

For this purpose, the most frequent relationships in SWN are taken into consideration: semantic – *hyponym/hypernym*, *holo_part*, *holo_member*, *near_antonym* and *category_domain* and lexical – *derived*, *eng_derivative*. Figure 3 illustrates example for the synset {sport:1, bavlxenxe sportom:X} (in English {sport:1, athletics:1}). From this figure we can see that this

INFOtheca

synset is superordinate to synsets {veslanxe:1}, {dyudo:1}, {gimnastika:1, gimnasticyke vezxbe:1},{kontaktni sport:1}, {sport na otvorenom:1, sport u prirodi:1} and {sportska igra:1} (they are connected by relationship *hypernym / hyponym*). With synsets {inicijativa:x, napad:3, akcija:1x}, {dugo pesxacyiti:1}, {skijati se:1, smucyati se:1},{sxutnuti:1x, voditi:1x}, {udarac:7} and {posed:A2x}, synset {sport:1, athletics:1} it is connected by relationship *category_domain*. Synset {veslanxe:1} is in relationship *derived* with synset {veslacyki:1}, and synset {sxutnuti:1x} is connected by relationship *eng_derivative* with synset {sxutiranxe:1}. So, some word x is encompassed by synset {sport:1, athletics:1} if x is equal to some of literal strings in synset {sport:1, athletics:1} or any other synset that is(directly or indirectly) connected with it (some of these synsets are mentioned above).

There are a lot of string literals in SWN that do not appear in any document of the corpus. Thereby, we will define the term – *active literals*.

DEFINITION 2: *Active literals* are literals that in some of its inflected forms appear in at least one document in the corpus.

When we talk about string literals that are encompassed by some synset, we consider only active string literals. So, in this step, for each category candidates will be selected for the category representative list of synsets, together with all active string literals encompassed by these candidates, maybe filtered by domains.

• *For each candidate for the category representative list of synsets, calculate how significant it is for the considered category*. In order to ensure proper selection of synsets for category representative list, we need to define a measure for determining the significance of synset candidate for the considered category. It will be calculated as follows: Let k_S be a synset candidate for the Sport representative list. The significance weight of this candidate is defined as a variant of *tf-idf* (term frequency – inverse document frequency) measure, and it is calculated as follows

$$SignificanceWeight(k_S) = tf(k_S, S) * idf(k_S)$$

where:

$$tf(k_S, S) = AveragedDensityByLiteral(k_S, S)$$

$$idf(k_S) = \log(N/\mathrm{df}\, k_S + 0.01)$$

N is the number of all documents in Ebart-3 corpus and $df\ k_S$ is the number of documents that contain at least one active literal string encompassed by synset k_S, maybe filtered by some domains.

Let M_S be a number of active literal strings encompassed by synset k_S, maybe filtered by some domains:

$$k_S = \{l_1, l_2, ..., l_{M_S}\}$$

Then the following holds:

$$tf(k_S, S) = AveragedDensityByLiteral(k_S, S) = \frac{\sum_{i=1}^{M_S} AveragedDensity(l_i, S)}{M_S}$$

where:

$$AveragedDensity(l_i, S) = \frac{\sum_{j=1}^{N_S} Density(l_i, d_j)}{N_S}$$

$Density(l_i, d_j)$ represents the number of occurrences of literal string l_i in the document d_j divided by the total number of words in that document. N_S is the number of documents that belong to the category Sport in the training set.

• *For each category, define category representative list of synsets from the Serbian wordnet*. For each category, based on the value of significance weight measure, synset candidate will be added to the category representative list (if the significance weight value is greater than some threshold) or not. Experimental results show that in the case of Ebart-3 corpus, 3 is a good choice for the threshold. For each category, we will select about ten synsets (with highest significance weight values, greater than threshold number) to be in the category representative list. By this, the training phase is complete.

The steps of the testing phase are as follows:
For each test document and each category

calculate the category assignment function. Category assignment function between test document and category is defined as document density of all active literal strings encompassed by all synsets from the category representative list, maybe filtered by some domains. Formally, category assignment function can be calculated as follows: Let $E = \{k_{E_1}, k_{E_2}, ..., k_{E_n}\}$ to be a representative list of synsets for category Economics, $P = \{k_{P_1}, k_{P_2}, ..., k_{P_n}\}$ to be a representative list for Politics and $S = \{k_{S_1}, k_{S_2}, ..., k_{S_n}\}$ to be a representative list of synsets for Sport category. Category assignment function for category Sport (the similar stands for Economics and Politics), is calculated as follows:

$$CategoryAsignmentFunction(d, Sport) = \sum_{k_S \in S} \sum_{l \in k_S} Density(l, d)$$

Where $Density(l, d)$ represents the number of occurrences of literal string l (in some of its inflected forms) in the document d, divided by the total number of words in that document. Test document is assigned to the category which has the greatest value of *CategoryAsignmentFunction*.

IMPLEMENTATION DETAILS: For obtaining semantic domains from PWN and literal strings from synsets we used *eXist* XML database. For browsing wordnet we used VisDic graphical application and for categorization procedure we used the software package *WordnetCategorization*, developed by the author of this paper.

POLITICS

A	B	C	D
izbor:4		63.78	1
partija:1a, stranka:1		30.97	3
parlament:1, skupsxtina:1		28.97	3
medxunarodan:1		28.94	1
politika:1b		22.69	1
ministar:1		18.52	41
poglavar drzxave:1, sxef drzxave:1		12.37	5
rat:1x,ratno stanxe:1		10.66	1
zajednica:1a, drusxtvo:1a	anthropology	8.90	8
politicyko telo:X	politics	10.66	1
svrstavanxe:2, alijansa:1, koalicija:1a	politics	8.90	8
glas:6,glasanxe:1		8.20	11
narod:1 nacija:1	politics	7.12	6
podrsxka:1y, potpora:2a	politics	6.47	6

TABLE 4: Representative list of synsets for the category Politics. **A** - Synsets Serbian - English, **B** - Domains for filtering, **C** - Synsets significance weight, **D** - Number of active literals

7. Experimental results

The category representative lists for the categories Economics, Politics and Sport obtained by the procedure presented in the previous section are presented in Tables 4-6. For each category and each synset from the list, tables show the values of significant weight measure of synset, number of active literals encompassed by that synset and optional domains for filtering.

During the process of forming category representative lists of synsets, it is necessary to pay attention to the relation of total numbers of active literals associated to each category (literals encompassed by all synsets from the category representative list). If a particular category has assigned significantly more literals than others, then that category may be wrongly favored. In the case of lists presented in Tables 4-6, the total numbers of active literals associated to categories Economics, Politics and Sport are 66, 59 and 69, respectively. Note that one literal can belong to more than one category.

ECONOMICS

A	B	C	D
banka:2		127.77	1
kredit:3		42.48	1
trzxisxte:2b, berza:x		41.74	2
prodaja:1, prodavanxe:1		19.65	5
industrija:1a, manufaktura:3	enterprise	19.44	5
ustanova:1, institucija:1	economy enterprise	16.80	4
dug:1		10.12	2
poduzecxe:2, preduzecxe:2	enterprise	4.99	26
novcyana jedinica:1	economy	3.53	25

TABLE 5: Representative list of synsets for the category Economics. **A** - Synsets Serbian - English, **B** - Domains for filtering, **C** - Synsets significance weight, **D** - Number of active literals

SPORT

A	B	C	D
klub:1b, udruzxenxe:1x		30.23	15
sezona:2, doba:2b		25.75	2
trijumf:2, pobeda:1b		22.44	2
tim:y, kolektiv:2b, ekipa:1a		19.25	2
skor:1, rezultat:2		16.86	3
lopta:2a		16.63	3
takmicenxe:1, nadmetanxe:1, natecanxe:1		8.68	2
igra:y	sport	7.2	1
oprema:1a, deo opreme:X, pribor:1a	sport	5.47	12
takmicyar:1		5.06	13
sport:1, bavlxenxe sportom:X		3.19	27

TABLE 6: Representative list of synsets for the category Sport. **A** - Synsets Serbian - English, **B** - Domains for filtering, **C** - Synsets significance weight, **D** - Number of active literals

For each testing document (from the Ebart-3 test dataset) and each category, value of CategoryAsignmentFunction is calculated. The category which achieved the maximum value of this function is assigned to a test document. Since the corpus is single-labeled, one of the possible problems is assigning more than one category to test document. So, we can distinguish two approaches: optimistic – if a document really belongs to one of the assigned categories we assume that the document is properly classified, and realistic – if a document really belongs to one of the assigned categories, we assume that the document is properly classified for corresponding category and not properly assigned for other categories determined by the classifier. The only difference between these two approaches is the number of false positives. Obtained results for optimistic approach are presented in Table 7 and for realistic approach in Table 8.

For comparison purpose, we developed a similar text categorization technique based on semantic domains associated with synsets in the wordnet. The only difference between this technique and the technique based on well-chosen synsets is the way of defining category representative lists. Instead of well-chosen synsets, in the case of domain-based technique category, representative list consists of all synsets that belong to semantic domains corresponding to the considered category. The domains of interest in the case of Ebart-3 corpus are:

- Economics: economy (economy, banking, enterprise, money, tax), commerce, industry.
- Politics: politics, anthropology.
- Sport: sport (sport, badminton, baseball, basketball, football, golf, soccer, tennis, volleyball, skiing, rowing, swimming, diving, athletics, boxing, fishing, hunting, and bowling).

Experimental results obtained by the technique based on semantic domains are presented in Table 9 (realistic approach). Number of active literals associated to categories Economics, Politics and Sport are 249, 111, and 66, respectively. Due to the large differences between numbers of literals associated to categories, category Economics is wrongly favoring (302 documents are false positives). Therefore we performed another experiment where the number of active literals per category is limited to 66 highest frequency string literals for corresponding category (as the number of literals in the category Sport). Results are presented in Table 10.

INFOtheca

A - Precision, **B** - Recall, **C** - F1-measure

	TP	FP	FN	A	B	C
S	450	37	38	92.40	92.21	92.31
E	130	34	36	79.27	78.31	78.79
P	425	45	42	90.43	91.01	90.72
Macro-average				87.37	87.18	87.27
Micor-average				89.65	89.65	89.65

TABLE 7: Results of wordnet-based text categorization technique based on well-chosen synsets – optimistic approach.

	TP	FP	FN	A	B	C
S	450	66	38	87.21	92.21	89.64
E	130	103	36	55.79	78.31	65.16
P	425	85	42	83.33	91.01	87.00
Macro-average				75.45	87.18	80.60
Micor-average				78.15	90.00	79.83

TABLE 8: Results of wordnet-based text categorization technique based on well-chosen synsets – realistic approach.

	TP	FP	FN	A	B	C
S	376	46	112	89.10	77.05	82.64
E	152	302	14	33.48	91.57	49.03
P	335	85	132	79.76	71.73	75.54
Macro-average				67.45	80.12	69.07
Micor-average				66.59	76.98	66.59

TABLE 9: Results obtained by wordnet-based text categorization technique based on semantic domains - realistic approach

	TP	FP	FN	A	B	C
S	402	51	86	88.74	82.38	85.44
E	147	274	19	34.92	88.55	50.09
P	365	108	102	77.17	78.16	77.66
Macro-average				66.94	83.03	71.06
Micor-average				67.85	81.53	67.85

TABLE 10: Results obtained by wordnet-based text categorization technique based on semantic domains with limited numbers of literals – realistic approach

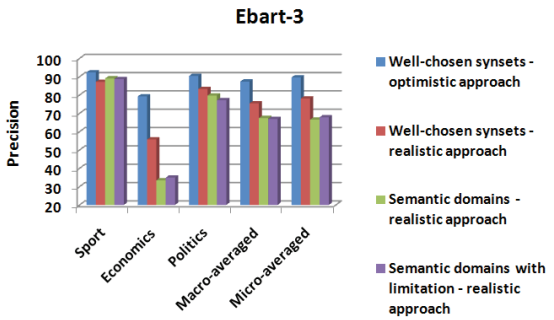

FIGURE 5: Comparison of wordnet-based text categorization techniques on Ebart-3 corpus, in term of Precision

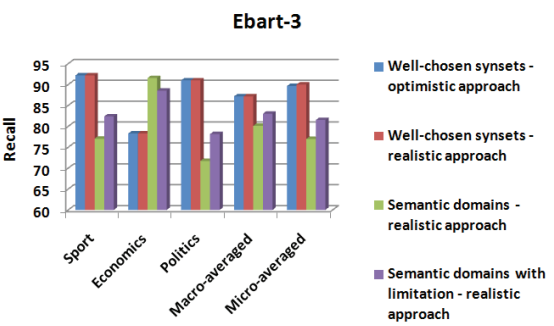

FIGURE 6: Comparison of wordnet-based text categorization techniques on Ebart-3 corpus, in term of Recall

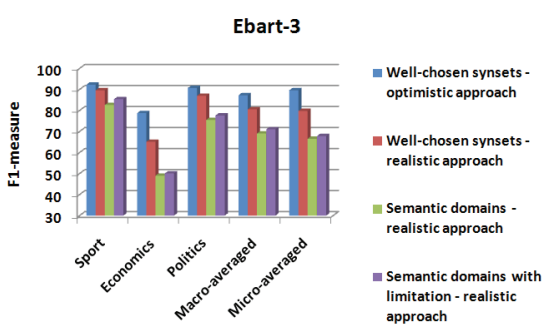

FIGURE 7: Comparison of wordnet-based text categorization techniques on Ebart-3 corpus, in term of F1-mesure

Comparison of results obtained by the two presented techniques – the one based on well-chosen synsets and another based on semantic domains, are presented in Figures 5, 6 and 7, in term of Precision, Recall and F1-measure, respectively. We conclude that the technique

based on well-chosen concepts outperforms the one based on semantic domains, where slightly better results are obtained when we limit the number of literals per category.

9. Conclusion

The main aim of this study is to examine how morphological, syntactic and semantic information contained in lexical resources for Serbian language can be effectively utilized in order to improve text classification. A wordnet-based text categorization technique for Serbian language is presented. The Ebart-3 corpus, a collection of newspaper articles in Serbian divided into three categories: Economics, Politic and Sport, is used. The technique is based on category representative lists of well-chosen synsets from the Serbian wordnet and category assignment function, defined on the basis of these lists. Inflection problem in Serbian is solved with the help of the system of morphological dictionaries of Serbian. The results show that the technique based on well-chosen synsets outperforms similar technique based on synsets that belong to the corresponding semantic domains, although the main reason for enriching wordnet by semantic domains is its even more successful application in natural language processing tasks, especially in text categorization.

We believe that results obtained by the technique presented in this paper would be much better if the corpus consisted of longer documents (containing a greater number of words). In this case realistic approach would be more similar to the optimistic approach (there would be a smaller number of cases in which more than one category is associated with one test document). Also, this technique would achieve better results if the corpus used a non-standard and expanded vocabulary.

Although this technique is developed for Serbian, it can be applied to any other language that has the same lexical recourses developed. Our aim is to test this technique on other corpora: in Serbian, with longer documents and richer vocabulary, or in some other languages.

References

Baeza-Yates, R., B. Ribeiro-Neto, et al. 1999. *Modern information retrieval, vol. 463*. ACM press New York.

Fellbaum, Christiane. 2010. Wordnet: An electronic lexical database. In *Theory and Applications of Ontology: Computer Applications*, eds. Roberto Poli, Michael Healy and Achilles Kameas, 231-243. Dordrecht : Springer.

Frawley, William J, Gregory Piatetsky-Shapiro and Christopher J. Matheus. 1992. Knowledge discovery in databases: An overview. *AI magazine*, 13(3): 57.

Graovac, Jelena and Gordana Pavlović-Lažetić. 2008. Productivity of concepts in Serbian Wordnet. In *Proceedings of the 11th International Multiconference Information Society - IS 2008*, vol. C, eds. Tomaž Erjavec and Jerneja Žganec Gros, 86–91. Ljubljana : Institut "Jožef Stefan".

Graovac, Jelena. 2012. Serbian text categorization using byte level n-grams. In *Proceedings of CloBL 2012: Workshop on Computational Linguistics and Natural Language, 5th Balkan Conference in Informatics, Novi Sad, Serbia, September 16-20, 2012*, eds. Zoran Budimac, Mirjana Ivanović and Miloš Radovanović, 93–97. Novi Sad : Faculty of Sciences, Department of Mathematics and Informatics.

Graovac, Jelena. To appear 2014. A variant of n-gram basedlanguage-independent text categorization. *Intelligent Data Analysis*, 18(4).

Horák, Aleš and Pavel Smrž. 2004. VisDic–wordnet browsing and editing tool. In *Proceedings of the Second International WordNet Conference - GWC 2004, Brno, Czech Republic, January 20 –23*, eds. Petr Sojka et al. 136–141. Brno : Masaryk University.

Joachims, Thorsten. 2002. *Learning to classify text using support vector machines: methods, theory and algorithms*. Berlin : Springer.

Krstev, Cvetana, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. Using textual and lexical resources in developing Serbian wordnet. *Romanian Journal of Information Science and Technology*, 7(1–2): 147–161.

Krstev, Cvetana, Bojana Đorđević, Sanja Antonić, Nevena Ivković-Berček et al. 2008. Kooperativan

INFOtheca

rad na dogradnji srpskog wordneta. *Infoteka*, 9(1): 57–75.

Krstev, Cvetana, Dusko Vitas, Ranka Stankovic, Ivan Obradovic and Gordana Pavlovic-Lazetic. 2004. Combining Heterogeneous Lexical Resources. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, eds. Maria Teresa Lino, 1103-1106. Paris : European Language Resources Association.

Krstev, Cvetana. 2008. *Processing of Serbian: automata, texts and electronic dictionaries*. Belgrade : Faculty of Philology, University of Belgrade.

Lewis, David D. and Marc Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, ed. Theo Pavlidis, 1-14. Las Vegas : Information Science Research Institute, University of Nevada

Obradović Ivan and Ranka Stanković. 2007. Integracija heterogenih tekstualnih resursa. U *Zbornik radova međunarodnog simpozijuma Razlike između bosanskog /bošnjačkog, hrvatskog i srpskog jezika*, ed. B. Tošović, 596–616. Graz.

Pavlović-Lažetić, Gordana and Graovac Jelena. 2010. Ontology-driven conceptual document classification. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval Valencia, Spain, October 25-28, 2010*, eds. Ana L. N. Fred and Joaquim Filipe, 383-386. Valencia : SciTePress.

Rodrıguez, Manuel de Buenaga, Hidalgo, Jose Maria Gomez and Agudo and Belen Diaz. 2000. Using WordNet to complement training information in text categorization. In *Recent Advances in Natural Language Processing II: Selected Papers from RANLP'97*, eds. Nicolas Nicolov and Ruslan Mitkov, 353-364. Amsterdam : John Benjamins Publishing Company.

Rosso, Paolo, Edgardo Ferretti, Daniel Jimenez and Vicente Vidal. 2004. Text categorization and information retrieval using wordnet senses. In *Proceedings of the Second International WordNet Conference - GWC 2004, Brno, Czech Republic, January 20 –23*, eds. Petr Sojka et al. 299-304. Brno : Masaryk University.

Van Rijsbergen. 1979. *Information Retrieval*. London : Butterworths.

Scott, Sam and Stan Matwin. 1998. Text classification using wordnet hypernyms. In *Usage of WordNet in Natural Language Processing Systems: Proceedings of the Workshop ; 16th August 1998*, ed. Sandra Harabagiu, 38–44. Stroudsburg : Association for Computational Linguistics.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1): 1–47.

Sébastien, Paumier. 2002. *Manuel d'utilisation du logiciel Unitex*. Champs-sur-Marne : Université de Marne-la-Vallée.

Tufis, Dan, 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.

Vitas, Duško, G. Pavlović-Lažetić, Cvetana Krstev, Lj. Popović and I. Obradović.2003. Processing Serbian written texts: an overview of resources and basic tools. In *Workshop on Balkan Language Resources and Tools, 21 Novembar 2003, Thessaloniki*, eds. S. Piperidis and V. Karkaletsis, 97–104. Thessaloniki : Greek Computer Society.

Vitas, Duško and Cvetana Krstev. 2005. Regular derivation and synonymy in an e-dictionary of Serbian. In *Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Pozna, Poland*, ed. Zygmunt Vetulani, 139-143. Poznań : Wydawnictwo Poznańskie.

Vossen, Piek. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Boston : Kluwer Academic.