E-Connecting Balkan Languages

Cvetana Krstev, Ranka Stanković, Duško Vitas, Svetla Koeva

E-Connecting Balkan Languages | Cvetana Krstev, Ranka Stanković, Duško Vitas, Svetla Koeva | Proceedings of the Workshop Workshop on Multilingual resources, technologies and evaluation for Central and Eastern European Languages, 17 September 2009, eds. C. Vertan, S. Piperidis, E. Paskaleva and Milena Slavcheva | 2009 | |

http://dr.rgf.bg.ac.rs/s/repo/item/0001597

E-Connecting Balkan Languages

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

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

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

Svetla Koeva
Dep. of Computational Linguistics
Institute for Bulgarian
svetla@dcl.bas.bg

Abstract

In this paper we present a versatile language processing tool that can be successfully used for many Balkan languages. This tool relies for its work on several sophisticated textual and lexical resources that were developed for most of Balkan languages. These resources are based on several *de facto* standards in natural language processing.

Keywords

Query expansion, e-dictionaries, wordnets, proper names, aligned texts

1. Introduction

The software tool WS4LR (shortened for WorkStation for Language Resources) is being developed by the Language Technology Group organized at the Faculty of Mathematics for several years now. Its first version was introduced in 2004 [8] and it dealt mainly with harmonizing various heterogeneous lexical resources. Subsequently, many new features were added, particularly those that helped in the production and exploration of aligned texts on the basis of the incorporated lexical resources [9]. The new tool WS4QE (shortened for Work Station for Query Expansion) was developed on the basis of WS4LR that enables expansion of queries submitted to the Google search machine [10]. The integrated lexical resources enable modifications of users queries for both monolingual and multi lingual search.

When presenting WS4LR and WS4QE we have always stressed that although they have been mainly used for Serbian they are by no means language dependent as long as compatible lexical resources exist for any two languages. Nevertheless, a full potential of these tools was until now used only for Serbian, and in bilingual context, for Serbian and English.

In this paper we will show that tools WS4LR and WS4QE are truly independent both from Serbian, for which they were initially developed, and from English which seems to be in the background of many natural language processing tools. The main presupposition for the usage of these tools for other languages is the existence of textual and lexical resources developed in the same methodological framework. Since this prerequisite is satisfied for Bulgarian, and to some extent for some other Balkan languages (Greek, Romanian, etc), we will show that WS4LR and WS4QE can be successfully used for them.

2. Integrated Language Resources

In order to prove the usability of WS4LR and WS4QE for languages other then Serbian and English we used various resources, both textual and lexical. In the following sections we will briefly present these resources, what methodological framework was used for their development, and how they were integrated for their successful usage.

2.1 Textual Resources – Aligned Texts

The aligned texts as a special form of multilingual corpora were in focus of many projects in past couple of decades. A systematic approach to the development of multilingual corpora was initiated within the Multext project, which subsequently included East-European languages through the Multext-East project [5]. In meantime many multilingual corpora were compiled, from large corpora usually fully automatically prepared comprising from texts in some limited technical domain [18], to more versatile literary corpora [5] that are often more modest in size but minutely prepared.

The main textual resource used to explore WS4LR is Jules Verne's novel *Around the world in eighty days*. This text was chosen for various reasons. First of all, the text is available in digital form for the majority of European languages, including Balkan languages. Regarding its content, it represents a suitable text for different types of analysis, especially in the domain of named entity recognition (geographical concepts and different measures). Besides that, it was already used for some interesting research, e.g. multi-word tagging [13] and building models for machine translation [21]. Finally, from the practical point of view its suitability stems from the fact that it presents the sample text for the French distribution of the Unitex system [15].

Versions of the novel in fifteen languages have been acquired, but not all of these texts have yet been aligned; Among already aligned texts are French original and translations in English and four Balkan languages – Serbian, Bulgarian, Greek, Romanian.

In the preparatory phase each translation was marked in accordance with the TEI-standard in XML, and the title (<head>), paragraph (<p>) and segments (<seg>) were included as units of text logical layout. At the beginning of the alignment process all segments coincided with sentences automatically tagged by Unitex. The XAlign system [1] was used for the alignment process. Starting from the French version, the goal of the alignment was to establish 1:1 relations on the segment level with all other languages. In order to achieve this goal and after manually checking all aligned segments, some of them had to be divided in smaller units, and some were grouped in larger units. Thus we arrived at the total of 4409 segments in all texts. This way, the missing segments or the inconsistencies between the source text and its translations were in most of the cases identified. In the following example the English segment is given only for the sake of translation.

```
<tu id=" n2941">
  <seg lang="en">
      <s id="Verne80days.n2941">
          Between Omaha and the Pacific the railway crosses a territory which is still infested by Indians and wild beasts, and a large tract which the Mormons, after they were driven from Illinois in 1845, began to colonise.</s></seg>
  <seg lang="fr">
      <s id="Verne80days.n2941">
          Entre Omaha et le Pacifique, le chemin de fer franchit une contrée encore fréquentée par les Indiens et les fauves, -- vaste étendue de territoire que les Mormons commencèrent à coloniser vers 1845, après qu'ils eurent été chassés de l'Illinois.</s></seg>
  <seg lang="sr">
      <s id="Verne80days. n2941">
          Između Omahe i  Tihog okeana pruga prolazi kroz predeo u kome  još ima Indijanaca i divljih zveri - prostranu  zemlju koju su počeli naseljavati mormoni oko  1845. godine, kada su ih prognali iz države Ilinois.</s> </seg>
  <seg lang="bg">
      <s id="Verne80days. n2941">
          Между Омаха и Тихия океан железопътната линия прекосява район, все още населяван от индианци и диви зверове. Това е обширна територия, която мормоните са започнали да колонизират около 1845 г., след като са били прогонени от щата Илинойс.</s></seg>
  <seg lang="gr">
      <s id="Verne80days. n2941">
          Ανάμεσα στην Ομάχα και στον Ειρηνικό, το τρένο διασχίζει περιοχές όπου συχνάζουν ακόμα Ινδιάνοι και αγρίμια - τεράστια εδαφική έκταση την οποία αρχισαν να αποικίζουν οι μορμόνοι μετά το 1845, οπότε κυνηγήθηκαν από το Ιλινόις.</s></seg>
  <seg lang="ro">
      <s id=" Verne80days.n569">
          între Omaha şi Pacific drumul de fier trece printr-o regiune populatã încã de indieni şi fiare, - vastã întindere pe care mormonii au început s-o colonizeze pe la 1845 dupã ce au fost izgoniţi din Illinois.</s>
</tu>
```

2.2 Morphological Dictionaries in LADL Format

Morphological dictionaries are a necessary resource in various phases of the automatic analysis of text. The tool WS4LR expects morphological dictionaries to be in the format known as DELAS/DELAF presented in [2] that was developed in LADL (*Laboratoire d'Automatique Documentaire et Linguistique*) under the guidance of Maurice Gross. The format of a DELAS-type dictionary basically consist of simple word lemmas accompanied with inflectional class codes which enable production of a DELAF-type dictionary which consists of all inflectional forms with their grammatical information. In Unitex environment one finite-state transducer responsible for generation of all inflectional forms of each DELAS lemma corresponds to each inflectional class code. The Serbian morphological dictionary of simple words contains 121,000 lemmas which yield the production of approximately 1,450,000 different lexical words. Close to 87,000 simple lemmas belong to general lexica, while the remaining 34,000 lemmas represent various kinds of simple proper names [11]. The Bulgarian Grammar dictionary (DELAS dictionary) consists of 127,000 lemmas distributed as follows: app. 85,000 simple lemmas belong to general lexis, app. 6,000 lemmas represent domain specific lexis and app. 36,000 lemmas are simple proper names. The corresponding DELAF dictionary consists of app. 1,260,000 entries [7].

2.3 Semantic Networks - Wordnet

Semantic networks, seen as one important node in the hierarchy of ontologies, are used more and more in various phases of the automatic analysis of text. The tool WS4LR expects them to be in the form of wordnets, that is, nodes representing sets of synonymous word (synsets) which are linked by various semantic relations. The first built wordnet was English wordnet, so-called Princeton Wordnet (PWN), having today approximately 140,000 synsets. Due to its remarkable size and successful inclusion in various computer-based applications it is considered as a de facto standard upon which wordnets for many other languages were built. One successful application of this concept was achieved by Balkanet project which was funded by European Commission from (2001-2004). In the scope of this project development of wordnets for the Balkan languages was initiated [20]: Bulgarian, Greek, Romanian, Serbian, and Turkish. The important feature of these wordnets is that they are all aligned with PWN via the Interlingual index (ILI) [22]. Namely, ILI consists of concepts, while wordnets represent lexicalization of concepts in various languages and the way they are connected.

Serbian wordnet today consists of more then 15,000 synsets built by app. 25,000 literals. All of them are linked to PWN, except for 532 Balkan specific concepts that are connected with other Balkan languages, and 155 Serbian specific concepts that remain unconnected with other languages. Bulgarian wordnet consists of more then 31,000 synsets built by more than 66,000 literals. The synsets are linked with the PWN as well, again there are 436 Balkan specific concepts shared with other Balkan languages and

182 Bulgarian language specific concepts. Both Serbian and Bulgarian wordnets, as well as wordnets for other Balkan languages, are in WS4LR represented using the common XML schema.

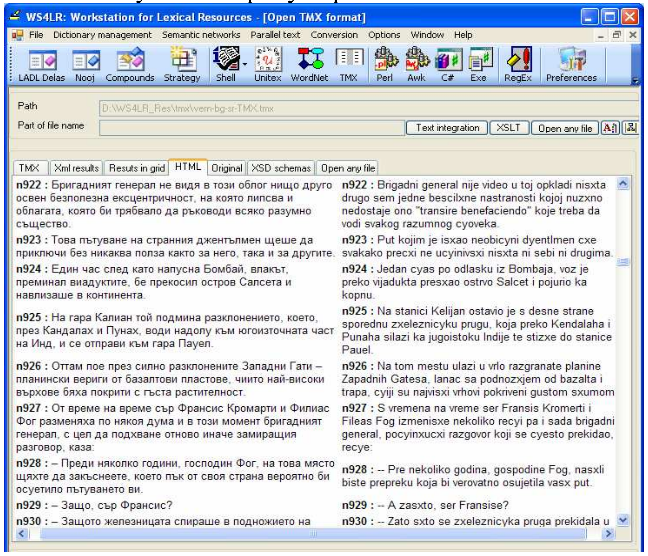
2.4 Prolex Database

The *Prolex project* was initiated in 1990s with the study of toponyms in French with aim of appropriately processing proper names in natural language applications [16]. This work has been pursued by development of a Serbian version, which finally led to the design and construction of a relational multilingual dictionary of Proper Names, Prolexbase, in a form of relational database [19]. This model is based on two main concepts: the *pivot* (that represents the *conceptual proper name*) at a language independent level and the *prolexeme* (the projection of the pivot onto particular language) that is a set of lemmas that includes the name, but also its aliases (variations in orthography, abbreviated forms, acronyms, etc.) and its derivatives. For instance, if meronymy relation is established between concepts ‘New York’ and ‘United States of America’, then their Serbian Latin equivalents *Njujork* and *Sjedinjene Američke Države*, Serbian Cyrillic equivalents *Њујорк* and *Сједињене Америчке Државе*, and Bulgarian equivalents *Ню Йорк* and *Съединени американски щати* are connected automatically.

3. Using WS4LR with Aligned Texts

The WS4LR module that works with aligned texts expects them to be in Translation Memory eXchange (TMX) format[1]. It can also transform texts previously aligned by XAlign into that format but also in several other formats: textual, XML and tabular. This is particularly important since XAlign has been integrated into Unitex software starting from its version 2.1. Besides, the user can also produce various visualization of aligned texts by applying appropriate XSLT transformations. Thus visualized texts user can freely browse. One such visualization is represented in Figure 1.

Browsing, however, is not a particularly successful form of text exploration. WS4LR module for aligned texts offers users to pose different forms of queries that can be automatically expanded by using various bilingual lexical resources presented in previous section. WS4LR offers to a user the possibility to expand the query morphologically, semantically, but also to another language. If the first language is Serbian, the second language can be English, Bulgarian, or any other. A user can choose two working languages by adjusting parameters in the “Preferences” manu of WS4LR. Besides, WS4LR provides further possibilities for a user to control the query formulation, since in addition to expansion it also offers a narrowing of the query. Namely, a user can reject some of the automatically offered query expansions.



Path D:\WS4LR_Res\tmx\vern-bg-sr-TMX.tmx

Part of file name

n922 : Бригадният генерал не видя в този облог нищо друго освен безполезна ексцентричност, на която липсва и облагата, която би трябвало да ръководи всяко разумно същество.	**n922 :** Brigadni general nije video u toj opkladi nisxta drugo sem jedne bescilxne nastranosti kojoj nuzxno nedostaje ono "transire benefaciendo" koje treba da vodi svakog razumnog cyoveka.
n923 : Това пътуване на странния джентълмен щеше да приключи без никаква полза както за него, така и за другите.	**n923 :** Put kojim je isxao neobicyni dyentlmen cxe svakako precxi ne ucyinivsxi nisxta ni sebi ni drugima.
n924 : Един час след като напусна Бомбай, влакът, преминал виадуктите, бе прекосил остров Салсета и навлизаше в континента.	**n924 :** Jedan cyas po odlasku iz Bombaja, voz je preko vijadukta presxao ostrvo Salcet i pojurio ka kopnu.
n925 : На гара Калиан той подмина разклонението, което, през Кандалах и Пунах, води надолу към югоизточната част на Инд, и се отправи към гара Пауел.	**n925 :** Na stanici Kelijan ostavio je s desne strane sporednu zxeleznicyku prugu, koja preko Kendalaha i Punaha silazi ka jugoistoku Indije te stizxe do stanice Pauel.
n926 : Оттам пое през силно разклонените Западни Гати – планински вериги от базалтови пластове, чиито най-високи върхове бяха покрити с гъста растителност.	**n926 :** Na tom mestu ulazi u vrlo razgranate planine Zapadnih Gatesa, lanac sa podnozxjem od bazalta i trapa, cyiji su najvisxi vrhovi pokriveni gustom sxumom
n927 : От време на време сър Франсис Кромарти и Филиас Фог разменяха по някоя дума и в този момент бригадният генерал, с цел да подхване отново иначе замиращия разговор, каза:	**n927 :** S vremena na vreme ser Fransis Kromerti i Fileas Fog izmenisxe nekoliko recyi pa i sada brigadni general, pocyinxucxi razgovor koji se cyesto prekidao, recye:
n928 : – Преди няколко години, господин Фог, на това място щяхте да закъснеете, което пък от своя страна вероятно би осуетило пътуването ви.	**n928 :** -- Pre nekoliko godina, gospodine Fog, nasxli biste prepreku koja bi verovatno osujetila vasx put.
n929 : – Защо, сър Франсис?	**n929 :** -- A zasxto, ser Fransise?
n930 : – Защото железницата спираше в подножието на	**n930 :** -- Zato sxto se zxeleznicyka pruga prekidala u

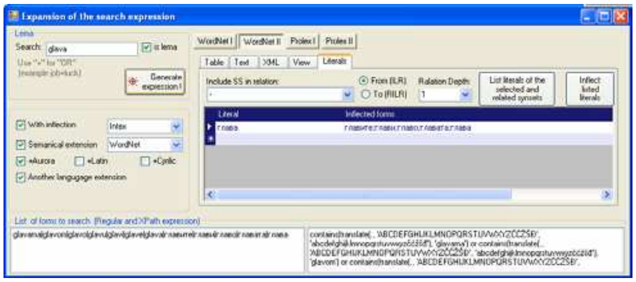
Figure 1. The HTML view of the aligned Bulgarian-Serbian text

Users queries can be semantically expanded by wordnets and by Prolex database. WS4LR obtains semantic expansion of a query by means of wordnet of the first language (Serbian wordnet – SWN in the case of our examples), selecting all synsets containing a given word and offering them to the user. This provides a user with an insight to all concepts the keyword pertains to, through sets of synonyms used for these concepts. A user then gets the possibility to delete some of these synsets if she/he decides that they pertain to concepts which are not of interest at that particular moment. Also, a user can formulate a bilingual query by adding the second language to it. Namely, WS4LR can for a given set of concepts identify all corresponding concepts in the second language wordnet by using the ILI. Thus, for an expanded Serbian query, one could obtain the corresponding expanded query in Bulgarian. The form used to bilingually expend a simple query *glava* ‘head’ with Bulgarian *глава* is presented in Figure 2. The semantic expansion is obtained by checking the box “Semantic extension” in this form and by choosing the appropriate resource (Wordnet in this case), while the bilingual expansion is obtained by checking the box “Another language extension”.

In the same form user can choose to morphologically inflect all chosen keywords in both languages. If she/he wishes to do so the box “With inflection” should be checked. Morphological expansion is performed by Unitex modules that use morphological dictionaries of simple words as well as inflectional transducers. This options works only if a particular query keyword is listed in the morphological dictionary of the corresponding language. If it is not so, the aligned text will be searched only with the original keyword. As shown in Figure 2, the automatically

[1] For details on TMX format see http://www.lisa.org/tmx/tmx.htm

added inflected forms of chosen keywords are presented in an editable form in which some of these inflected forms can be deleted or modified. For instance, Serbian word *put* 'path' has two forms of plural: *putevi* and *puti*. The second one is restricted to poetical usage and a user can choose to delete it from the expended query if the working text is not of that kind.



Search: glava ☑ is lema

☑ With inflection: Intex
☑ Semanical extension: WordNet
☑ +Aurora ☐ +Latin ☐ +Cyrilic
☑ Another language extension

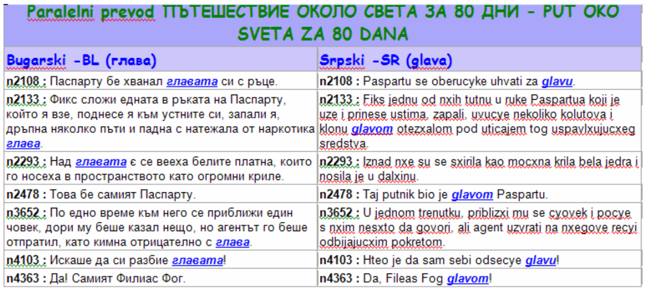
Include SS in relation: - ● From (ILR) ○ To (RILR) Relation Depth: 1

Literal	Inflected forms
глава	главите;глави;главо;главата;глава

List of forms to search (Regular and XPath expression)

Figure 2. The original query keyword *glava* is shown in the upper left corner. The chosen query expansions are shown on the left side. The query expended by Bulgarian wordnet is shown on the right side, together with the automatically obtained list of inflected forms that can be edited. Two fields at bottom show the final query set.

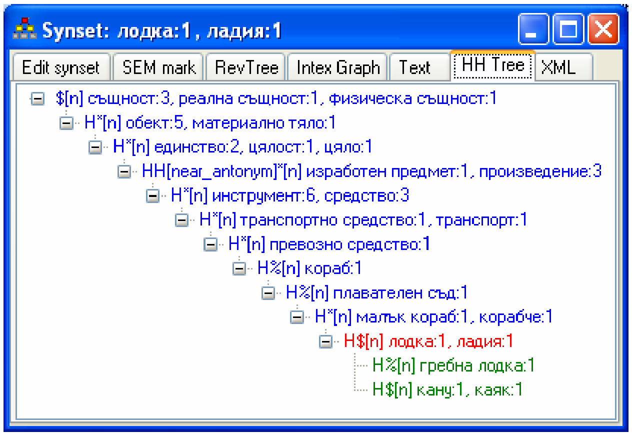
Finally, when a query is launched, the result is obtained with all retrieved occurrences highlighted (see Figure 3)



Paralelni prevod ПЪТЕШЕСТВИЕ ОКОЛО СВЕТА ЗА 80 ДНИ - PUT OKO SVETA ZA 80 DANA	
Bugarski -BL (глава)	**Srpski -SR (glava)**
n2108 : Паспарту бе хванал *главата* си с ръце.	**n2108 :** Paspartu se oberucyke uhvati za *glavu*.
n2133 : Фикс сложи едната в ръката на Паспарту, който я взе, поднесе я към устните си, запали я, дръпна няколко пъти и падна с натежала от наркотика *глава*.	**n2133 :** Fiks jednu od nxih tutnu u ruke Paspartua koji je uze i prinese ustima, zapali, uvucye nekoliko kolutova i klonu *glavom* otezxalom pod uticajem tog uspavlxujucxeg sredstva.
n2293 : Над *главата* є се вееха белите платна, които го носеха в пространството като огромни криле.	**n2293 :** Iznad nxe su se sxirila kao mocxna krila bela jedra i nosila je u dalxinu.
n2478 : Това бе самият Паспарту.	**n2478 :** Taj putnik bio je *glavom* Paspartu.
n3652 : По едно време към него се приближи един човек, дори му беше казал нещо, но агентът го беше отпратил, като кимна отрицателно с *глава*.	**n3652 :** U jednom trenutku, priblizxi mu se cyovek i pocye s nxim nesxto da govori, ali agent uzvrati na nxegove recyi odbijajucxim pokretom.
n4103 : Искаше да си разбие *главата*!	**n4103 :** Hteo je da sam sebi odsecye *glavu*!
n4363 : Да! Самият Филиас Фог.	**n4363 :** Da, Fileas Fog *glavom*!

Figure 3. Some representative examples of aligned segments with keywords *glava* and *глава* and their inflectional forms in HTML format.

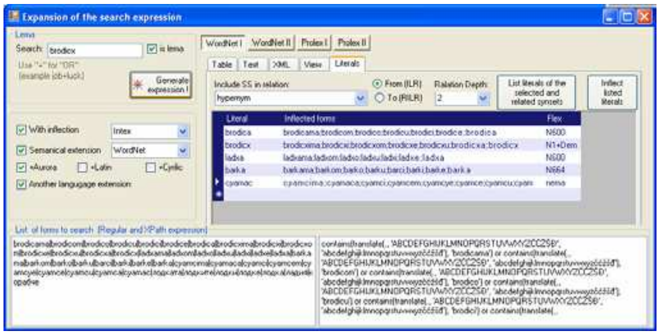
The query can be further semantically expanded by the choice of a particular semantic relation (e.g. hypernymy/hyponymy), in which case synsets pertaining to hypernyms/hyponyms of concepts from the initial group will also appear among the query set. This feature will be illustrated by the query which starts with the Serbian keyword *brodić* 'small boat'. We would like to perform the bilingual search with semantic expansion. The chosen Serbian keyword belongs to only one synset {brodica:1, brodić:1} whose corresponding Bulgarian synset is {лодка:1, ладия:1}. Figure 4 shows that these synsets are deep in the hypernymy/hyponymy hierarchy. In such situation expending query with hypernym synsets can be useful.



```
⊟ $[n] същност:3, реална същност:1, физическа същност:1
  ⊟ H*[n] обект:5, материално тяло:1
    ⊟ H*[n] единство:2, цялост:1, цяло:1
      ⊟ HH[near_antonym]*[n] изработен предмет:1, произведение:3
        ⊟ H*[n] инструмент:6, средство:3
          ⊟ H*[n] транспортно средство:1, транспорт:1
            ⊟ H*[n] превозно средство:1
              ⊟ H%[n] кораб:1
                ⊟ H%[n] плавателен съд:1
                  ⊟ H*[n] малък кораб:1, корабче:1
                    ⊟ H$[n] лодка:1, ладия:1
                        H%[n] гребна лодка:1
                        H$[n] кану:1, каяк:1
```

Figure 4. Hypernym/hyponym wordnet hierarchy of the Bulgarian synset {лодка:1, ладия:1}. The corresponding Serbian synset belongs to the similar tree.

Figure 5 shows the query expansion form in which the original query *brodić* is expanded not only with a literal from its corresponding synset, that is *brodica*, but also with the literals from synsets belonging to the hypernym branch of length two, that are {barka:1, čamac:1, čun:1} 'boat' and {lađa:1} 'vessel'.



Search: brodicx ☑ is lema

☑ With inflection: Intex
☑ Semanical extension: WordNet
☑ +Aurora ☐ +Latin ☐ +Cyrilic
☑ Another language extension

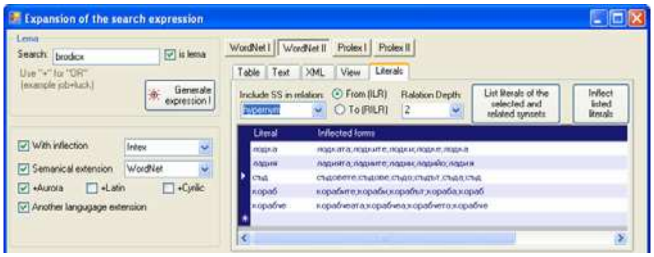
Include SS in relation: hypernym ● From (ILR) ○ To (RILR) Relation Depth: 2

Literal	Inflected forms	Flex
brodica	brodicama;brodicom;brodico;brodicu;brodici;brodice;brodica	N600
brodicx	brodicxima;brodicxi;brodicxom;brodicxe;brodicxu;brodicxa;brodicx	N1+Dem
ladxa	ladxama;ladxom;ladxo;ladxu;ladxi;ladxe;ladxa	N600
barka	barkama;barkom;barko;barku;barci;barki;barke;barka	N664
cyamac	cyamcima;cyamaca;cyamci;cyamcem;cyamcye;cyamce;cyamcu;cyam	nema

List of forms to search (Regular and XPath expression)

Figure 5. In the query expansion form a user can choose the type of semantic relation for the expansion and the length of the path with this relation she/he wishes to pursue.

Since in this case bilingual search is initiated a user can perform the same semantic expansion for the second language, presented in Figure 6. Two Bulgarian literals thus obtained are *плавателен съд* and *малък кораб* which are multi-word units. Since inflection of multi-word units for Bulgarian is not yet integrated in WS4LR, as will be explained in the final section, a user can choose to delete it from the final query set or to keep only the nouns *съд* and *кораб*, as we have done in our example search.



Search: brodicx ☑ is lema

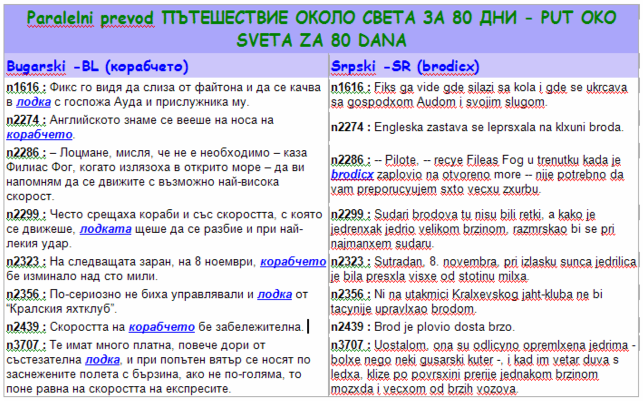
☑ With inflection: Intex
☑ Semanical extension: WordNet
☑ +Aurora ☐ +Latin ☐ +Cyrilic
☑ Another language extension

Include SS in relation: hypernym ● From (ILR) ○ To (RILR) Relation Depth: 2

Literal	Inflected forms
лодка	лодката,лодките,лодки,лодко,лодка
ладия	ладията,ладиите,ладии,ладийо,ладия
съд	съдовете,съдове,съдо,съдът,съда,съд
кораб	корабите,кораби,корабът,кораба,кораб
корабче	корабчетата,корабчета,корабчето,корабче

Figure 6. The semantic expansion in the second language – Bulgarian – using hypernym relation

The results obtained by this query are very interesting and show by themselves the potential this tool offers for various linguistic and literary researches. This query retrieved 129 aligned segments, each of which contained at least one of the keywords from the produced query set in at least one of the languages. It comes as a surprise that only 8 of these segments contained query keywords in both languages. This is mainly due to the fact that adjectives *плавателен* and *малък* were omitted from Bulgarian keywords thus broadening the query on Bulgarian side too much. There were 5 segments with a keyword *съд*, with two occurrences of *плавателен съд* 'vessel'; to none of them corresponded a Serbian wordnet equivalent *lađa*. There were also 90 occurrences of *кораб* among which there was not one *малък кораб*; in this case, however, Serbian equivalent for *кораб* was almost unmistakably *brod*, as suggested by both wordnets.

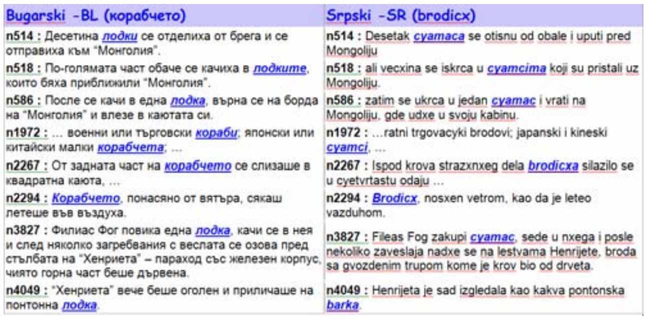


Paralelni prevod ПЪТЕШЕСТВИЕ ОКОЛО СВЕТА ЗА 80 ДНИ - PUT OKO SVETA ZA 80 DANA

Bugarski -BL (корабчето)	Srpski -SR (brodicx)
n1616 : Фикс го видя да слиза от файтона и да се качва в *лодка* с госпожа Ауда и прислужника му.	n1616 : Fiks ga vide gde silazi sa kola i gde se ukrcava sa gospodxom Audom i svojim slugom.
n2274 : Английското знаме се вееше на носа на *корабчето*.	n2274 : Engleska zastava se leprsxala na klxuni broda.
n2286 : – Лоцмане, мисля, че не е необходимо – каза Филиас Фог, когато излязоха в открито море – да ви напомням да се движите с възможно най-висока скорост.	n2286 : -- Pilote, -- recye Fileas Fog u trenutku kada je *brodicx* zaplovio na otvoreno more -- nije potrebno da vam preporucyujem sxto vecxu zxurbu.
n2299 : Често срещаха кораби и със скоростта, с която се движеше, *лодката* щеше да се разбие и при най-лекия удар.	n2299 : Sudari brodova tu nisu bili retki, a kako je jedrenxak jedrio velikom brzinom, razmrskao bi se pri najmanxem sudaru.
n2323 : На следващата заран, на 8 ноември, *корабчето* бе изминало над сто мили.	n2323 : Sutradan, 8. novembra, pri izlasku sunca jedrilica je bila presxla visxe od stotinu milxa.
n2356 : По-сериозно не биха управлявали и *лодка* от "Кралския яхтклуб".	n2356 : Ni na utakmici Kralxevskog jaht-kluba ne bi tacynije upravlxao brodom.
n2439 : Скоростта на *корабчето* бе забележителна.	n2439 : Brod je plovio dosta brzo.
n3707 : Те имат много платна, повече дори от състезателна *лодка*, и при попътен вятър се носят по заснежените полета с бързина, ако не по-голяма, то поне равна на скоростта на експресите.	n3707 : Uostalom, ona su odlicyno opremlxena jedrima - bolxe nego neki gusarski kuter -, i kad im vetar duva s ledxa, klize po povrsxini prerije jednakom brzinom mozxda i vecxom od brzih vozova.

Figure 7. A few examples of a partial retrieval

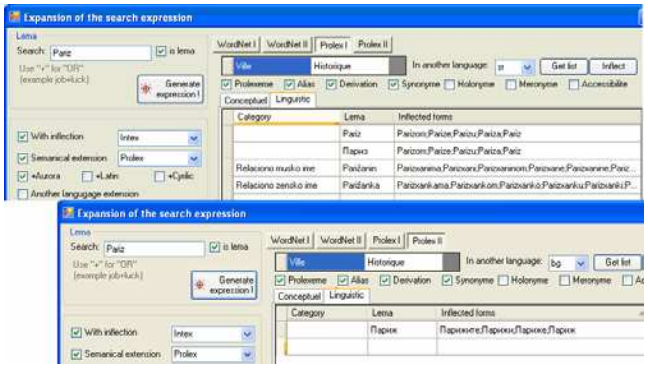
Figure 7 shows some examples of a partial retrieval. First (n1616) and third (n2286) segments in this sample occur due to the fact that the reference to a 'boat' is missing in one of the languages. The other segments show that Serbian *brod*, besides corresponding to English *ship* and Bulgarian *кораб*, is also a generic notion and should probably be added to the hypernym synset (segments n2274, n2356 and n2439). On the other hand Serbian *jedrilica* and *jedrenjak* 'sailing vessel' are in Bulgarian translated with a "sister" synsets *кораб* or *корабче* instead of using a more specific Bulgarian word *платноход* (segments n2299 and n2323). In the last example (n3707), in Bulgarian a rather arbitrary choice *лодка* is made for a more specific type of a vessel referred to in Serbian as *kuter* 'cutter'.



Bugarski -BL (корабчето)	Srpski -SR (brodicx)
n514 : Десетина *лодки* се отделиха от брега и се отправиха към "Монголия".	n514 : Desetak *cyamaca* se otisnu od obale i uputi pred Mongoliju
n518 : По-голямата част обаче се качиха в *лодките*, които бяха приближили "Монголия".	n518 : ali vecxina se iskrca u *cyamcima* koji su pristali uz Mongoliju.
n586 : После се качи в една *лодка*, върна се на борда на "Монголия" и влезе в каютата си.	n586 : zatim se ukrca u jedan *cyamac* i vrati na Mongoliju, gde udxe u svoju kabinu.
n1972 : ... военни или търговски *кораби*; японски или китайски малки *корабчета*; ...	n1972 : ...ratni trgovacyki brodovi; japanski i kineski *cyamci*, ...
n2267 : От задната част на *корабчето* се слизаше в квадратна каюта, ...	n2267 : Ispod krova strazxnxeg dela *brodicxa* silazilo se u cyetvrtastu odaju ...
n2294 : *Корабчето*, понасяно от вятъра, сякаш летеше във въздуха.	n2294 : *Brodicx*, nosxen vetrom, kao da je leteo vazduhom.
n3827 : Филиас Фог повика една *лодка*, качи се в нея и след няколко загребвания с веслата се озова пред стълбата на "Хенриета" – параход със железен корпус, чиято горна част беше дървена.	n3827 : Fileas Fog zakupi *cyamac*, sede u nxega i posle nekoliko zaveslaja nadxe se na lestvama Henrijete, broda sa gvozdenim trupom kome je krov bio od drveta.
n4049 : "Хенриета" вече беше оголен и приличаше на понтонна *лодка*.	n4049 : Henrijeta je sad izgledala kao kakva pontonska *barka*.

Figure 8. All occurreneces of a full retrieval

Figure 8 shows eight examples of the full retrieval. In one of these examples (n1972) for the Serbian *čamac* the near synonym in Bulgarian *корабчето* is used (as determined by wordnets). In two cases (n2267 and n2294) for the Serbian *brodić* the near hypernym *корабчето* is used, while in five cases (n514, n518, n586, n3827, n4049) for the Serbian *čamac* and *barka* the near hyponym *лодка* is used. This is not an unexpected result; it only proves that searching with the help of semantic networks, on web for instance, can be useful, which is the ultimate goal of our experiments.



Expansion of the search expression

Lema

Search: Pariz ☑ is lema

Use "+" for "OR" (example job+luck)

☑ With inflection: Intex

☑ Semanical extension: Prolex

☑ +Aurora ☐ +Latin ☐ +Cyrilic

☐ Another language extension

Ville | Historique — In another language: sr

☑ Prolexeme ☑ Alias ☑ Derivation ☑ Synonyme ☐ Holonyme ☐ Meronyme ☐ Accessibilite

Category	Lema	Inflected forms
	Pariz	Parizom;Parize;Parizu;Pariza;Pariz
	Париз	Parizom;Parize;Parizu;Pariza;Pariz
Relaciono musko ime	Parižanin	Parizxanima;Parizxani;Parizxaninom;Parizxane;Parizxanine;Pariz...
Relaciono zensko ime	Parižanka	Parizxankama;Parizxankom;Parizxanko;Parizxanku;Parizxanki;P...

Expansion of the search expression

Lema

Search: Pariz ☑ is lema

Use "+" for "OR" (example job+luck)

☑ With inflection: Intex

☑ Semanical extension: Prolex

Ville | Historique — In another language: bg

☑ Prolexeme ☑ Alias ☑ Derivation ☑ Synonyme ☐ Holonyme ☐ Meronyme ☐ Ac

Category	Lema	Inflected forms
	Париж	Парижките;Парижки;Парижке;Париж

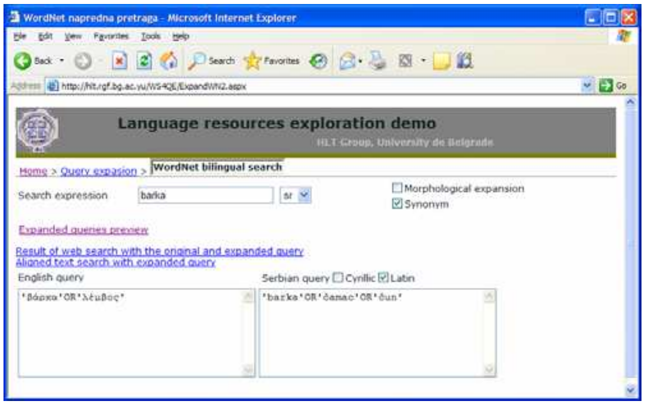
Figure 9. Prolex based semantic expansions

When search is performed not by common keywords but by proper nouns then query expansion with Prolex database offers more possibilities. Semantic relations incorporated in this database are adapted to proper names. Here, user can choose to expand his query both on the conceptual and the linguistic level. It can be seen in Figure 9 how a query launched with a pivot *Paris* is linguistically expanded in two languages. The morphological expansion can be chosen here as well and it is performed in the same way and using the same methods as for common words. In the given example, query expansion for Serbian gives more results since Prolex database for Bulgarian has only some sample entries.

4. Additional Possibilities

We have illustrated in the previous section by the Serbian and Bulgarian pair the functions of WS4LR for working with aligned texts. It can be successfully used for other Balkan languages as well. Wordnets were being developed through Balkanet project for Greek, Romanian and Turkish, which enabled the experiments with semantic query expansions for those languages as well. For Greek [12] and Romanian [3], morphological dictionaries in LADL format were also developed – however, these resources were not at our disposal so we could not experiment with morphological expansion for these languages.

The possibility and the need for some of the functions developed within WS4LR to become also available on the web led to the development of the WS4QE web application for lexical resources. This application is still under development, but some of its functions can already be used. Numerous user functions are envisaged for this tool, but the largest set is related to the expansion of queries submitted to the search engine Google, and they have already been implemented. In fact, they are very similar to those presented in the previous section. The only difference is that expanded queries are not applied to an aligned text but are rather forwarded to the search engine. Figure 10 shows such an retrieval that starts with the Serbian keyword *barka* 'boat' and is further expended by the Serbian synset {barka:1, čamac:1, čun:1} and Greek corresponding synset {βάρκα:0, λέμβος:0}. Figure 11 represents the first results retrieved by such an expanded query by Google.


WordNet napredna pretraga - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go
Language resources exploration demo
HLT Group, University de Belgrade

http://hlt.rgf.bg.ac.yu/WS4QE/ExpandWN2.aspx

Home > Query expasion > WordNet bilingual search

Search expression: barka sr

□ Morphological expansion
☑ Synonym

Expanded queries preview

Result of web search with the original and expanded query
Aligned text search with expanded query

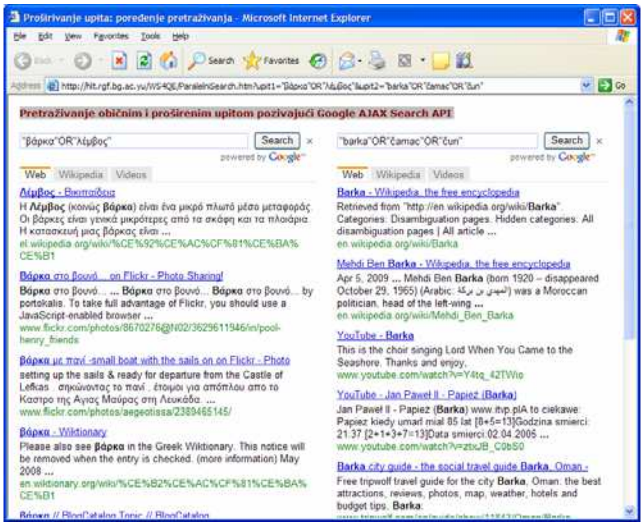
English query: 'βάρκα'OR'λέμβος'

Serbian query □ Cyrillic ☑ Latin: 'barka'OR'čamac'OR'čun'

Figure 10. Bilingual query expansion with WS4QE – example of Serbian and Greek

5. Further Work

Our main concern for the future work is adequate processing of multi-word units. That is, we would like our tool to treat multi-word units in the same way as simple words and to inflect them correctly upon request. The first version of this approach was presented in [10]. Although this version gave promising results for Serbian, it was hardwired into the tool itself so that it was not easy neither to modify Serbian module nor to apply it to other languages. With a new approach that relies on feature structure description of particular language morphology [6] and widely uses XML technology the portability to other languages will be much easier [17]. On a more practical level, our aim is enrich our lexical resources, first of all the Prolex database since we plan to use it in a translation environment [14]. It is our wish to work in a future with a true aligned Balkan text – that is, a text originally written in some Balkan language and translated to other Balkan languages.


Proširivanje upita: poređenje pretraživanja - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go
Search
powered by Google
Web
Wikipedia
Videos

http://hlt.rgf.bg.ac.yu/WS4QE/ParallelnSearch.htm?upit1="βάρκα"OR"λέμβος"&upit2="barka"OR"čamac"OR"čun"

Pretraživanje običnim i proširenim upitom pozivajući Google AJAX Search API

"βάρκα"OR"λέμβος"

Λέμβος - Βικιπαίδεια
Η **Λέμβος** (κοινώς **βάρκα**) είναι ένα μικρό πλωτό μέσο μεταφοράς. Οι βάρκες είναι γενικά μικρότερες από τα σκάφη και τα πλοιάρια. Η κατασκευή μιας βάρκας είναι ...
el.wikipedia.org/wiki/%CE%92%CE%AC%CF%81%CE%BA%CE%B1

Βάρκα στο βουνό ... on Flickr - Photo Sharing!
Βάρκα στο βουνό... ... **Βάρκα** στο βουνό... **Βάρκα** στο βουνό... by portokalis. To take full advantage of Flickr, you should use a JavaScript-enabled browser ...
www.flickr.com/photos/8670276@N02/3629611946/in/pool-henry_friends

βάρκα με πανί -small boat with the sails on on Flickr - Photo
setting up the sails & ready for departure from the Castle of Lefkas . σηκώνοντας το πανί , έτοιμοι για απόπλου απο το Καστρο της Αγιας Μαύρας στη Λευκάδα. ...
www.flickr.com/photos/aegeotissa/2389465145/

βάρκα - Wiktionary
Please also see **βάρκα** in the Greek Wiktionary. This notice will be removed when the entry is checked. (more information) May 2008 ...
en.wiktionary.org/wiki/%CE%B2%CE%AC%CF%81%CE%BA%CE%B1

Βάρκα // BlogCatalog Topic // BlogCatalog

"barka"OR"čamac"OR"čun"

Barka - Wikipedia, the free encyclopedia
Retrieved from "http://en.wikipedia.org/wiki/**Barka**". Categories: Disambiguation pages. Hidden categories: All disambiguation pages | All article ...
en.wikipedia.org/wiki/Barka

Mehdi Ben Barka - Wikipedia, the free encyclopedia
Apr 5, 2009 ... Mehdi Ben **Barka** (born 1920 – disappeared October 29, 1965) (Arabic: المهدي بن بركة) was a Moroccan politician, head of the left-wing ...
en.wikipedia.org/wiki/Mehdi_Ben_Barka

YouTube - Barka
This is the choir singing Lord When You Came to the Seashore. Thanks and enjoy.
www.youtube.com/watch?v=Y4tq_42TWio

YouTube - Jan Pawel II - Papiez (Barka)
Jan Pawel II - Papiez (**Barka**) www.itvp.plA to ciekawe: Papiez kiedy umarl mial 85 lat [8+5=13]Godzina smierci: 21:37 [2+1+3+7=13]Data smierci:02.04.2005 ...
www.youtube.com/watch?v=ztxJB_C0b50

Barka city guide - the social travel guide Barka, Oman -
Free tripwolf travel guide for the city **Barka**, Oman: the best attractions, reviews, photos, map, weather, hotels and budget tips. **Barka**:

Figure 11. Results of a query bilingually expanded by Wordnet

6. References

[1] P. Bonhomme, T. M. H. Nguyen, S. O'Rourke. XAlign: l'aligneur de Langue & Dialogue, http://www.loria.fr/equipes/led/outils/ALIGN/align.html, 2001.

[2] B. Courtois, M. Silberztein (eds.). *Dictionnaires électroniques du français.* Langue française. 87, Larousse, Paris, 1990.

[3] D.-M. Dimitriu. *Grammaires de flexion du roumain en format DELA,* Rapport interne 2005-02 de l'Institut Gaspard-Monge, CNRS, 2005.

[4] T. Erjavec and N. Ide. The MULTEXT-East Corpus. In *LREC'98*, Granada, pp. 971-974, 1998.

[5] A. Gelbukh, G. Sidorov, J.-A. Vera-Félix. A Bilingual Corpus of Novels Aligned at Paragraph Level. In proc. *FinTAL*-2006. *Lecture Notes in Artificial Intelligence*, no. 4139, Springer-Verlag, pp. 16–23, 2006.

[6] ISO 24610. *Language resource management – Feature Structures*, ISO/TC 37/SC 4, 2005.

[7] S. Koeva. M*odern language technologies – applications and perspectives,* in: Lows of/for language, Hejzal, Sofia, 2004, 111- 157, 2004.

[8] C. Krstev, et al. Combining Heterogeneous Lexical Resources, in Proc. of the Fourth International Conference LREC, Lisbon, Portugal, May 2004, vol. 4, pp. 1103-1106, 2004.

[9] C. Krstev, R. Stanković, D. Vitas, I. Obradović. *WS4LR: A Workstation for Lexical Resources*, Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 2006, pp. 1692-1697, 2006.

[10] C. Krstev, R. Stanković, D. Vitas, I. Obradović, The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines, in *Proceedings of the Sixth Interantional Conference on Language Resources and Evaluation* (LREC'08), Marrakech, Morocco, 28-30 May 2008, European Language Resources Association (ELRA), 2008.

[11] C. Krstev. *Processing of Serbian*, Faculty of Phylology, University of Belgrade, Belgrade, 2008.

[12] T. Kyriacopoulou. Les dictionnaires électroniques : Morphologie et syntaxe. Le cas du grec moderne, *Proceedings AILA 1990*, Chalcidique, 1990.

[13] E. Laporte, T. Nakamura, S. Voyatzi. A French Corpus Annotated for Multiword Nouns, in: *Towards a Shared Task for Multiword Expressions* (MWE 2008), in scope of the *Sixth Interantional Conference on Language Resources and Evaluation* (LREC'08), http://multiword.sourceforge.net/download/MWE2008-papers/8_Laporte.pdf, 2008.

[14] D. Maurel, D. Vitas, C. Krstev, S. Koeva. Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian, in *Bulag - Bulletin de Linguistique Appliquée et Générale*, Les langues slaves et le français : approches formelles dans les études contrastives, eds. A. Dziadkiewicz & I. Thomas, No. 32, pp. 55-72, Presses Universitaires de Franche Comté, Besancon, 2007.

[15] S. Paumier. *Unitex 2.1 User Manual,* http://www-igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf, 2008.

[16] O. Piton, D. Maurel. Beijing frowns and Washington takes notice: Computer Processing of Relations between Geographical Proper Names in Foreign Affairs, *Fourth International Workshop on Applications of Natural Language to Data Bases (NLDB'00),* Versailles, 28-30 juin (Actes p. 66-78), 2000.

[17] R. Stanković. Improvement of Queries using a Rule Based Procedure for Inflection of Compounds and Phrases. *Polibits (37) 2008, Special section: Natural Langugage Processing, Journal of Research and Developement in Computer Science and Engeneering, ed. Grigori Sidorov,* Centro Innovacion y Desarrollo Tecnologico en Computo, Instututo Politecnico Nacional, Mexico, pp. 14-20, 2008.

[18] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th LREC Conference*, Genoa, Italy, 22-28 May, 2006, pp.2142-2147, 2006.

[19] M. Tran, D. Maurel. Prolexbase : Un dictionnaire relationnel multilingue de noms propres, Traitement automatique des langues, Vol. 47-3, 2006.

[20] D. Tufiş (ed.). *Special Issue on BalkaNet Project*, Romanian Journal on Information Science and Technology. Bucureşti: Publishing house of the Romanian academy, Vol. 7, No.1-2, 2004.

[21] D. Tufiş, S. Koeva, T. Erjavec, M. Gavrilidou, and C. Krstev. Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In M. Tadić, M. Dimitrova-Vulchanova and S. Koeva (eds.) *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, pp. 145-152, Dubrovnik, Croatia, September 25-28, 2008.

[22] P. Vossen (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers, 1998.