The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines

Krstev Cvetana, Stanković Ranka, Vitas Duško, Obradović Ivan

The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines | Krstev Cvetana, Stanković Ranka, Vitas Duško, Obradović Ivan | LREC 2008: Conference on Language Resources and Evaluation, Marrakesh, Morocco, May 2008 | 2008 | |

http://dr.rgf.bg.ac.rs/s/repo/item/0004838

The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines

Cvetana Krstev[1], Ranka Stanković[2], Duško Vitas[3], Ivan Obradović[4]
[1] professor, Faculty of Philology, Belgrade, [2] assistant, Faculty of Mining and Geology, Belgrade
[3] professor, Faculty of Mathematics, Belgrade, [4] professor, Faculty of Mining and Geology, Belgrade
E-mail: cvetana@matf.bg.ac.yu, ranka@rgf.bg.ac.yu, vitas@matf.bg.ac.yu, ivano@rgf.bg.ac.yu

Abstract

In this paper we present how resources and tools developed within the Human Language Technology Group at the University of Belgrade can be used for tuning queries before submitting them to a web search engine. We argue that the selection of words chosen for a query, which are of paramount importance for the quality of results obtained by the query, can be substantially improved by using various lexical resources, such as morphological dictionaries and wordnets. These dictionaries enable semantic and morphological expansion of the query, the latter being very important in highly inflective languages, such as Serbian. Wordnets can also be used for adding another language to a query, if appropriate, thus making the query bilingual. Problems encountered in retrieving documents of interest are discussed and illustrated by examples. A brief description of resources is given, followed by an outline of the web tool which enables their integration. Finally, a set of examples is chosen in order to illustrate the use of the lexical resources and tool in question. Results obtained for these examples show that the number of documents obtained through a query by using our approach can double and even quadruple in some cases.

1. Introduction

When delivering a query to a web search engine the user is typically interested in information available on the web related to a particular topic. The result of this query is a selection of web pages the search engine determines as relevant to the query. The information the user is interested in can generally be expressed in terms of *concepts*, abstract ideas or mental symbols that denote objects in a given category or class of entities, interactions, phenomena, or relationships between them. On the other hand, concepts are lexicalized by one or more synonymous *words* (simple or compound). For example, the concept of a "housing that someone is living in" is lexicalized by the word "house", but also by "dwelling", "home", "domicile", "abode", "habitation" or "dwelling house". Hence, the concept a web query pertains to is in practice very often formalized by a Boolean OR combination of words, which the user believes best describe the concept in question, e.g. "house OR home OR domicile".

It goes without saying that the choice of words used in a query are of crucial importance for the relevance of the results delivered by the search engine. At the first glance, the main problem lies in the fact that the user, when composing a query, might omit some words related to the concept, thus reducing system *recall*. A simple query expansion by adding the omitted words would seemingly resolve this problem. However, the expansion of the set of words describing a concept in a query, although contributing to the recall in general, has and adverse effect. Namely, due to the fact that many words are homonymous or polysemous, adding new words to the query might reduce *precision*. Given this trade-off between recall and precision, words used in a query have to be very carefully selected in order to attain an optimal balance between the two.

The problem is further complicated when searches are performed for highly inflective languages such as Serbian, which, moreover, equally uses two alphabets, Cyrillic and Latin. Some of the search engines, such as Google, have tackled the problem of inflection, and Google queries for Serbian are now expanded with the usage of some sort of a stemmer. However, this approach solves the inflection problem only partially and the solution is far from systematic. As is often the case with stemmers, Google expands the query by including not only (some) inflective forms but also related words. For example, a Google query with the Serbian word *prevodilac* 'translator' also offers web pages containing the word *prevod* 'translation', while the query with *javno mnjenje* 'public opinion' also offers pages containing the word *javnost* 'populace'. As it could be expected, this kind of approach works poorly for verbs. For instance, a query with *slati poruku* 'to send a message' returns only pages that contain the verb *slati* in the infinitive form, or the verbal noun *slanje* 'sending' and omits numerous pages on the Web containing other verb forms like, for instance, *šaljem poruku* '(I) send a message'. In some cases, unrelated results are obtained. As Google tries to be too smart it assumes that an occurrence of 's' in Serbian text can be replaced by 'š'. Thus, when searching for *strasna nedelja* 'Passion Week' the unrelated results for *strašna nedelja* 'horrible week' or 'horrible Sunday' are obtained as well.

2. Typical problems when retrieving documents using a web search engine

1. In general, when the concept the query relates to is lexicalized by one or more multi-word terms in a highly inflective language, the search engines are faced with a problem they are practically unable to cope with. For example, let us consider that we wish to search the web for the information on *beli luk* 'garlic'. When searching with the two constituent keywords *beli* 'white' AND *luk* ‘onion’ the search engine would typically return an irrelevant document based on the following content:

> Sastojci za 10 porcija: 3 glavice crnog **luka**, 1 šoljica ulja, 1/2 čaša **belog** vina, 1 čaša soka od paradajza
>
> (The ingredients for 10 portions: 3 onions, 1 cup of oil, ½ glass of white wine, 1 glass of tomato juice.)

This false retrieval occurs because two constituents of the multi-word term are treated separately, and neither nearness conditions nor grammatical agreement conditions are taken into account, which reduces precision. Conversely, if a literal search is performed as with “*beli luk*” then inflected forms of this multi-word term are not taken into account, and this reduces recall. In this case the aforementioned irrelevant document would be omitted, but so would be many relevant results, for instance

> Gambori u maslacu sa belim lukom
>
> (Shrimps on butter with garlic (in the instrumental case))

2. The simple keyword search is based on the lexical realization of a concept and not on the concept itself. Thus, it does not take into account the synonyms, unless the user himself remembers to include them in the search, for instance by adding the Serbian synonym *češnjak* to *beli luk*, which would improve recall. Even more relevant results could be obtained if the search is further expanded with the Latin name *Allium sativum* which many users probably would not even know. This is, however, the simplest conceptual expansion of a query. A more sophisticated would be a web query on Amerindian languages (*amerindijanski* in Serbian). The user issuing such a query is most probably not looking for the occurrences of the exact term with its possible synonyms – *indijanski* and *amerindski* – in all inflectional forms (*amerindijanskog*, *indijanskog*, *amerindskog*, etc.), but also for the occurrences of the specific languages belonging to that language class, for instance, *atakapa*, *mozan*, *tupi-gvarani* and many others that are derivationally unrelated to the original keyword, thus making any stemmer useless.

3. In some cases the user may wish to perform a bilingual search in order to find documents on the chosen subject in two languages, e.g. English and Serbian. In the case of *garlic* the appropriate query should be composed of the keywords *beli luk*, *češnjak*, *Allium sativum*, and *garlic*. It is not to be expected that a common user would normally possess the knowledge necessary to expand a query in this way.

3. The lexical resources used

In order to achieve an optimal balance between recall and precision in retrieving documents from the web we have developed WS4QE (Work Station for Query Expansion) which uses various language resources we have developed for Serbian (Krstev et al., 2008). These resources include morphological e-dictionaries and finite state transducers, which offer the possibilities for solving the problem of flections in queries, and electronic thesauri, ontologies and wordnets which offer various possibilities for automatic or semi-automatic refinement of queries by adding new words to the set of words initially specified by the user.

1. *Morphological dictionaries* of simple words and compounds in the so called LADL format (Courtois et al., 1990) basically consist of lemmas accompanied with inflectional class codes which enables a precise production of all inflectional forms. The Serbian morphological dictionary of simple words contains 117,000 lemmas which yields the production of approximately 1,400,000 different lexical words. More than 85,000 simple lemmas belong to general lexica, while the remaining 32,000 lemmas represent various kinds of simple proper names. The Serbian morphological dictionary of compounds contains approximately 2,700 lemmas (yielding more than 60,000 different forms) and it is being constantly upgrading.

2. Inflectional *finite state transducers* (FST) for the inflection of both simple and compound words have been developed for the Unitex system (http://www-igm.univ-mlv.fr/~unitex/). It is important to stress that WS4QE does not rely only on a simple list of word forms for Serbian simple and compounds words, but on the inflectional transducers as well. This enables a more elaborate query expansion that can significantly improve retrieval performances. For instance, if a query is performed with the keyword *beli luk*, three inflectional transducers are used: one for inflection of the adjective *beli* ‘white’, one for inflection of the noun *luk* ‘onion’ and one for the compound as whole which takes care of agreement conditions. These transducers expand the query *beli luk* into

> *beli luk* AND *belim lukom* AND *beli lukovi* AND *belih lukova* AND *belima lukovima* AND *belim lukovima* AND *bele lukove* AND *bela luka* AND *beloga luka* AND *belog luka* AND *belome luku* AND *belom luku*

Due to the third inflectional transducer this query expands into only 12 combinations of an adjective form and a noun form, instead of 216 possible combinations, thus disabling false retrieval such as: *Tako, posmatrano sa dna vidika, izgleda kao da iz širokih* ***lukova belog*** *mosta teče i razliva se ne samo zelena Drina*… ‘Thus, from a bottom view, it appears that not only green Drina flows and spills over under the wide **arcs** of the **white** bridge…’

3. Wordnets in XML format are used for query expansion with related words as well as for bilingual searches. The Serbian and English lexicalizations of the same (or similar) concepts in the Serbian wordnet (SWN - conceived within the Balkanet project (Tufiş, 2004), and presently encompassing 14.593 synsets) and the Princeton wordnet which is publicly available are connected via the Interlingual index (ILI) (Vossen, 1998).

4. In a similar way queries can be expanded by *Prolex*, a multilingual database of proper names which represents the implementation of an elaborate four-layered ontology of proper names (Krstev, et al., 2005) organized around a conceptual proper name that represents the same concept in different languages. For instance, *Prolex* establishes the meronymy relation between concepts ‘New York’ and ‘United States of America’, and automatically between their Serbian equivalents *Njujork* and *Sjedinjene Američke Države*. Various other relations are implemented as well.

4. The system options

Our system for query expansion allows the user to decide how his query will be expanded by choosing one or several of the offered options:

1. Alternate alphabet usage – for instance, the user can submit a keyword in Latin alphabet: *štrajk* ‘strike’ which will be expanded automatically by adding the keyword in Cyrillic: *штрајк*.

2. The inclusion of inflectional forms, for instance, *štrajk, štrajka, štrajkovi, ...* The inflection is done by Unitex procedures that use morphological dictionaries and inflectional FSTs for Serbian. The inflection works both for simple words and compounds.

3. The addition of synonyms – for instance, the synonym *obustava rada* ‘work stoppage’ can be added to the keyword *štrajk*. Synonyms are added on basis of the Serbian Wordnet (SWN). All the other relations included in SWN can also be used for the query expansion, for instance the keyword *solarni sistem* ‘solar system’ can be expanded by *Merkur, Venera, Zemlja, Mars*, etc. if meronymy is used for query expansion.

4. The expansion of proper names using Prolex which offers to the user the option of adding proper name aliases, its synonyms, but also other proper names which are semantically related to the initial proper name through holonym and meronym relations. Thus a query with the word *Engleska* ‘England’ can be expanded with *Englez* ‘Englishman’, *Engleskinja*, ‘English woman’ but also with *Albion*.

5. The inflection of free phrases by predicting their syntactic structure. Our presumption is that many free phrases used for search will have the same syntactic structure as a compound, and that the inflectional transducers for compounds that we have already developed can be applied to inflect them correctly. Our further presumption is that in many cases this structure can be predicted on the basis of morphological and syntactic features of the phrase components. These features can be obtained from the morphological e-dictionaries that are at our disposal during the query expansion process. The prediction of the phrase structure is also based on the frequencies of compound structures that we have obtained from our existing dictionary of compounds. This analysis shows that, not surprisingly, the most frequent structure for compounds with two components is adjective+noun, followed by the compounds with the structure X+noun, where X means “a word form that does not inflect within the compound”. For compounds with three components the most frequent structure is noun+X+X. Data on frequencies can help in deciding which structure should be attributed to a free phrase when several options exist according to e-dictionaries. A nice example is the phrase *Republika Francuska* which, according to the dictionaries can be analyzed as a phrase of the form noun+noun or noun+adjective. Since the latter structure is not very frequent in Serbian, the former is chosen that is also the correct one. In this particular case the latter solution would not yield erroneous results either since for query expansion we need only correctly inflected forms and not grammatical categories.

6. In Serbian many compounds have a structure in which some of its components do not inflect (like X+noun or noun+X+X). When identifying the structure of a free phrase it may sometimes be difficult to decide which components inflect and which don’t. One simple rule would be that word forms that are unknown (i.e. that do not have a corresponding entry in our e-dictionaries) do not inflect. It would yield correct examples in some cases (for instance, in *šper ploče* ‘plywood’ *šper* does not inflect and it is not in our e-dictionaries since it is not a valid Serbian word). In some other situations the prediction would be incorrect, as for *Telecom Srbija* ‘Telecom Serbia’ where *Telecom* is an unknown word but it inflects (e.g. the dative form is *Telecomu Srbije*). More sophisticated rules are also used to detect the components that do not inflect, one of them being “if the word that follows a noun is possibly a preposition and the next word is in the grammatical case that is required by that preposition, neither of the word forms following the noun will inflect”. This rule would correctly determine that the free phrase *kamatne stope na dinarsku štednju* ‘interest rates on savings in dinars[1] ’ has the form adjective+noun+X+X+X due to the fact that the adjective form *dinarsku* is in the accusative case that is required by the preposition *na*.

7. In order to test our system we have used a log file of one of Serbian professional journals that deals with economic issues. The journal’s web site is supported by a search engine that enables its readers to retrieve information from journal’s archive. The used log file thus gives a good insight in users’ queries. Many of the multi word queries are of no interest since they represent simple lists of key words, for instance *Beograd, Gradska čistoća, privatizacija* ‘Belgrade, City Waste Disposal, privatization’. It is not expected that the user would be interested for inflections of such a list as a whole. Some phrases, as we have expected, had a structure not yet found among compounds, such as adjective+noun+conjunction+noun in *Beogradski vodovod i kanalizacija* ‘Belgrade water supply and sewage system’. For many free phrases, especially those with fewer components, the structure was correctly detected and their inflected forms produced, e.g. *smrznuto voće i povrće* ‘frozen fruits and vegetables’. As a by-product, the analysis of the log file detected some compounds that were not yet in the dictionary of compounds and which were subsequently added to it (the most frequent one being *kursna lista* ‘the exchange rate list’). In order to be able to correctly inflect more free

[1] Dinar is Serbian currency

phrases we have produced some new inflectional transducers as for the structure adjective+conjunction+adjective+noun in *ekonomska i monetarna unija* ‘economic and monetary union’

8. The bilingual search – for instance, to the keyword *štrajk* and its Serbian synonym *obustava rada* a corresponding English set of synonyms can be added: {strike, work stoppage}. The bilingual search is, however, done separately and the results are presented in two columns.

5. Technical implementation

The developed web application receives the user query, and subsequently uses the local web service WS4QE to expand the query and forward it to the Google search engine using the *Google AJAX Search API*. Google AJAX Search API is a Java script library which enables the embedding of Google searches into personal web pages or web applications. This library is composed of simple web objects which perform “inline” search using numerous Google services (Web Search, Local Search, Video Search, Blog Search, News Search and Book SearchNew!). We have embedded a simple, dynamic search box and the search results are displayed within our own web pages for different types of query expansions, depending on the resources and type of expansion. Web service WS4QE uses classes from .NET dll components developed within WS4LR (WorkStation for Lexical Resources) (Krstev et al., 2006), which enable the usage of lexical resources for query expansion.

The web service returns the required information in XML form, which is being received and converted to appropriate application structures (string, array, table,...). Some of the typical calls are: getObliciLeme(lema), which retrieves all inflective forms of a lemma, getSinonimiWN_WithFlex(lema) which retrieves all wordnet synonyms with inflective forms, getSinonimiWN_NoFlex(lema) which retrieves all wordnet synonyms without inflective forms, getProlexTable(rec, jezikSearch, Inflect, ExpandWith) which retrieves all chosen proper name expansions according to the request specified by the user.

We will now illustrate some of WS4QE features related to query expansion. Figure 1 depicts the home page of WS4QE where the left hand side shows the menu with the functions offered and the right side the login part. Besides query expansion, WS4QE also offers functions for manipulation of aligned texts and wordnet management, as listed in the menu, but we will leave here these functions aside and concentrate on query expansion.

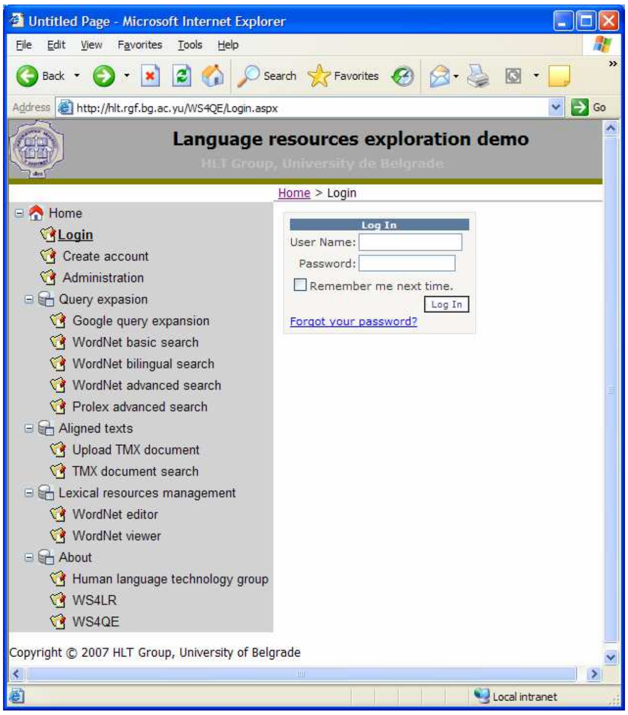
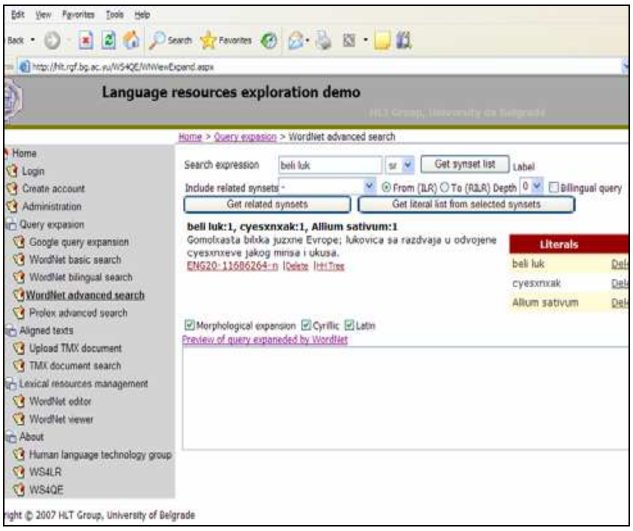


Figure 1. WS4QE home page

The user can choose from several options for query expansion, the wordnet advanced search being the most complex. Figure 2 shows the page for this type of search with the word *beli luk* in Latin alphabet chosen as the initial search string. As semantic expansion was chosen, the appropriate synset was retrieved and two other synonyms for *beli luk*, namely *češnjak* (as ‘cyesxnxak’ in the Aurora[2] code) and *Allium sativum* appeared in the list of words that can be used for composing the query. However, given that one of the synonyms is a Latin word, it was estimated that its introduction in the query would generate a great number of irrelevant documents in languages other than Serbian, so the options for removing some of the synonymous words was used and the word list was reduced to two Serbian words: *beli luk* and *češnjak*. In this particular case morphological expansion was omitted, and the query is further expanded only by including both chosen words in Cyrillic (Figure 3).

[2] For reasons of flexibility letters specific for the Serbian language ć, č, š ,ž ,đ, dž, lj and nj, are internally coded as cx, cy, sx, zx, dx, dy, lx and nx, respectively)


Edit View Favorites Tools Help
Back Search Favorites
http://hlt.rgf.bg.ac.yu/WS4QE/WNViewExpand.aspx
Language resources exploration demo
HLT Group, University of Belgrade
Home > Query expasion > WordNet advanced search
Home
Login
Create account
Administration
Query expasion
Google query expansion
WordNet basic search
WordNet bilingual search
WordNet advanced search
Prolex advanced search
Aligned texts
Upload TMX document
TMX document search
Lexical resources management
WordNet editor
WordNet viewer
About
Human language technology group
WS4LR
WS4QE
Get synset list
Label
Get related synsets
Get literal list from selected synsets
Literals
Dele
Dele
Dele
ENG20-11686264-n |Delete |H Tree
Preview of query expaneded by WordNet
ight © 2007 HLT Group, University of Belgrade

Search expression: beli luk, sr

Include related synsets: - ◉ From (ILR) ○ To (R&LR) Depth 0 □ Bilingual query

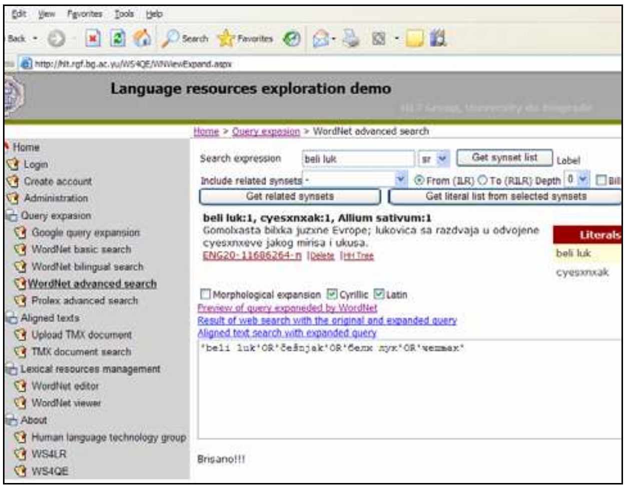
beli luk:1, cyesxnxak:1, Allium sativum:1
Gomolxasta bilxka juzxne Evrope; lukovica sa razdvaja u odvojene cyesxnxeve jakog mirisa i ukusa.

Literals
beli luk
cyesxnxak
Allium sativum

☑ Morphological expansion ☑ Cyrillic ☑ Latin

Figure 2. Semantic expansion of a query

The query, now composed of two Latin and two Cyrillic strings was then submitted by WS4QE to Google and, as a result, a total of 92,700 documents were obtained. The same query submitted directly to Google with only the initial string *beli luk* returned a total of 54,900. Thus the expanded expansion, without the morphological expansion almost doubled the number of documents obtained. It could, however, be argued that this does not necessarily mean that all obtained documents are relevant.


Edit View Favorites Tools Help
Back Search Favorites
http://hlt.rgf.bg.ac.yu/WS4QE/WNViewExpand.aspx
Language resources exploration demo
HLT Group, University of Belgrade
Home > Query expasion > WordNet advanced search
Home
Login
Create account
Administration
Query expasion
Google query expansion
WordNet basic search
WordNet bilingual search
WordNet advanced search
Prolex advanced search
Aligned texts
Upload TMX document
TMX document search
Lexical resources management
WordNet editor
WordNet viewer
About
Human language technology group
WS4LR
WS4QE
Get synset list
Label
Get related synsets
Get literal list from selected synsets
Literals
ENG20-11686264-n |Delete |H Tree
Preview of query expaneded by WordNet
Result of web search with the original and expanded query
Aligned text search with expanded query

Search expression: beli luk, sr

Include related synsets: - ◉ From (ILR) ○ To (RILR) Depth 0 □ Bil

beli luk:1, cyesxnxak:1, Allium sativum:1
Gomolxasta bilxka juzxne Evrope; lukovica sa razdvaja u odvojene cyesxnxeve jakog mirisa i ukusa.

Literals
beli luk
cyesxnxak

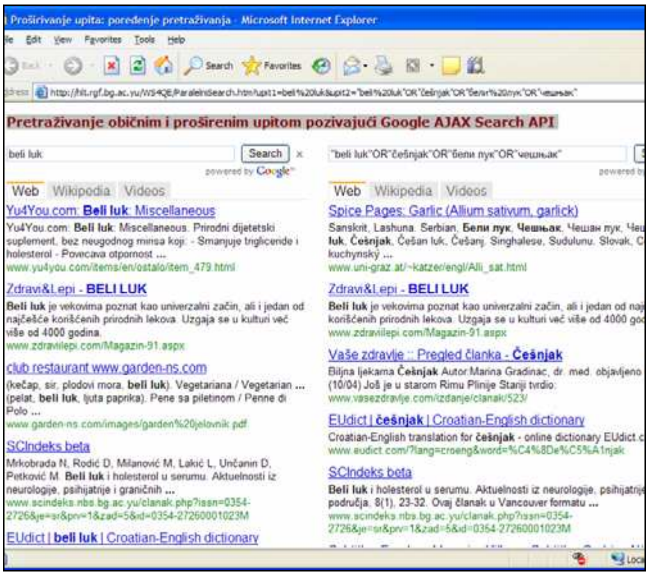
□ Morphological expansion ☑ Cyrillic ☑ Latin

```
'beli luk'OR'češnjak'OR'бели лук'OR'чешњак'
```

Brisano!!!

Figure 3. Finalized query to be submitted to Google

A thorough inspection of all documents was not performed, for obvious reasons, but it is safe to say that it is most unlikely that any of the documents obtained is irrelevant because both words used are specific in that they are neither homonymous nor polysemous. Part of the results is depicted in Figure 4. On the left hand side results obtained by the direct, unexpanded query are given, while the right hand side shows the results of the expanded query.


Proširivanje upita: poređenje pretraživanja - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hlt.rgf.bg.ac.yu/WS4QE/ParalelnSearch.htm?upit1=beli%20luk&upit2="beli%20luk"OR"češnjak"OR"бели%20лук"OR"чешњак"
Search
powered by Google
Web Wikipedia Videos
Web Wikipedia Videos
Local

Pretraživanje običnim i proširenim upitom pozivajući Google AJAX Search API

beli luk

Yu4You.com: **Beli luk**: Miscellaneous
Yu4You.com: **Beli luk**: Miscellaneous. Prirodni dijetetski suplement, bez neugodnog mirisa koji: - Smanjuje trigliceride i holesterol - Povećava otpornost ...
www.yu4you.com/items/en/ostalo/item_479.html

Zdravi&Lepi - **BELI LUK**
Beli luk je vekovima poznat kao univerzalni začin, ali i jedan od najčešće korišćenih prirodnih lekova. Uzgaja se u kulturi već više od 4000 godina.
www.zdravilepi.com/Magazin-91.aspx

club restaurant www.garden-ns.com
(kečap, sir, plodovi mora, **beli luk**). Vegetariana / Vegetarian ... (pelat, **beli luk**, ljuta paprika). Pene sa piletinom / Penne di Polo ...
www.garden-ns.com/images/garden%20jelovnik.pdf

SCIndeks beta
Mrkobrada N, Rodić D, Milanović M, Lakić L, Unčanin D, Petković M. **Beli luk** i holesterol u serumu. Aktuelnosti iz neurologije, psihijatrije i graničnih ...
www.scindeks.nbs.bg.ac.yu/clanak.php?issn=0354-2726&je=sr&prv=1&zad=5&id=0354-27260001023M

EUdict | **beli luk** | Croatian-English dictionary

"beli luk"OR"češnjak"OR"бели лук"OR"чешњак"

Spice Pages: Garlic (Allium sativum, garlick)
Sanskrit, Lashuna. Serbian, **Бели лук**, **Чешњак**, Чешан лук, Чеш **luk**, **Češnjak**, Češan luk, Češanj. Singhalese, Sudulunu. Slovak, C kuchynský ...
www.uni-graz.at/~katzer/engl/Alli_sat.html

Zdravi&Lepi - **BELI LUK**
Beli luk je vekovima poznat kao univerzalni začin, ali i jedan od naj korišćenih prirodnih lekova. Uzgaja se u kulturi već više od 4000 go
www.zdravilepi.com/Magazin-91.aspx

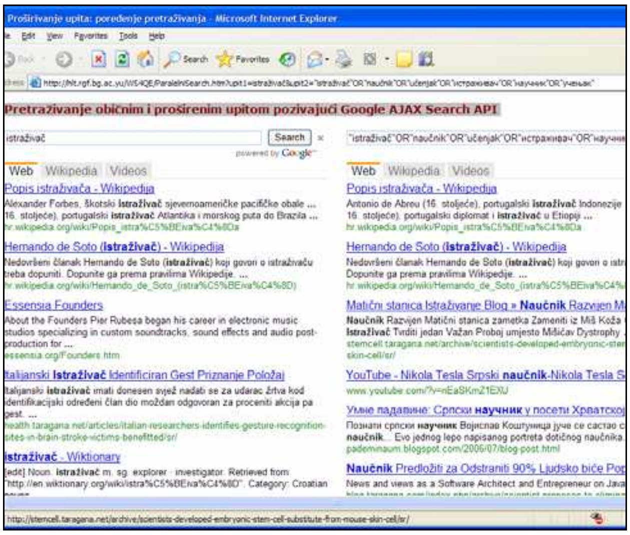
Vaše zdravlje :: Pregled članka - **Češnjak**
Biljna ljekarna **Češnjak** Autor Marina Gradinac, dr. med. objavljeno (10/04) Još je u starom Rimu Plinije Stariji tvrdio:
www.vasezdravlje.com/izdanje/clanak/523/

EUdict | **češnjak** | Croatian-English dictionary
Croatian-English translation for **češnjak** - online dictionary EUdict c
www.eudict.com/?lang=croeng&word=%C4%8De%C5%A1njak

SCIndeks beta
Beli luk i holesterol u serumu. Aktuelnosti iz neurologije, psihijatrije područja, 8(1), 23-32. Ovaj članak u Vancouver formatu ...
www.scindeks.nbs.bg.ac.yu/clanak.php?issn=0354-2726&je=sr&prv=1&zad=5&id=0354-27260001023M

Figure 4. Results for expanded query for 'beli luk'

For illustration purposes, two additional queries were performed using the word *istraživač* 'researcher'. Since the word *istraživač* has no synonyms in Serbian wordnet, semantic expansion was performed by including the words from the hypernym of *istraživač*, namely *naučnik* and *učenjak* 'scientist'. The query was further expanded by including all words in Cyrillic alphabet, morphological expansion once more omitted. The result of the expanded query was a total of 160,000 documents as opposed to 66,600 obtained by the unexpanded query (Figure 5). The expanded query once again doubled the number of documents obtained. Finally, a second query was performed for the word *istraživač*. This time a morphological expansion was performed and the semantic expansion omitted, but the extension to Cyrillic alphabet remained. As a result 285,000 documents were obtained, which means that the recall has been quadrupled. Thus we may conclude that a considerable increase of recall was obtained in all three examples.


Proširivanje upita: poređenje pretraživanja - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hlt.rgf.bg.ac.yu/WS4QE/ParalelnSearch.htm?upit1=istraživač&upit2="istraživač"OR"naučnik"OR"učenjak"OR"истраживач"OR"научник"OR"учењак"
Search
powered by Google
Web Wikipedia Videos
Web Wikipedia Videos
http://stemcell.taragana.net/archive/scientists-developed-embryonic-stem-cell-substitute-from-mouse-skin-cell/sr/

Pretraživanje običnim i proširenim upitom pozivajući Google AJAX Search API

istraživač

Popis istraživača - Wikipedija
Alexander Forbes, škotski **istraživač** sjevernoameričke pacifičke obale ... 16. stoljeće), portugalski **istraživač** Atlantika i morskog puta do Brazila ...
hr.wikipedia.org/wiki/Popis_istra%C5%BEiva%C4%8Da

Hernando de Soto (**istraživač**) - Wikipedija
Nedovršeni članak Hernando de Soto (**istraživač**) koji govori o istraživaču treba dopuniti. Dopunite ga prema pravilima Wikipedije. ...
hr.wikipedia.org/wiki/Hernando_de_Soto_(istra%C5%BEiva%C4%8D)

Essensia Founders
About the Founders Pier Rubesa began his career in electronic music studios specializing in custom soundtracks, sound effects and audio post-production for ...
essensia.org/Founders.htm

Italijanski **Istraživač** Identificiran Gest Priznanje Položaj
Italijanski **istraživač** imati donesen svjež nadati se za udarac žrtva kod identifikacijski određeni član dio možđan odgovoran za proceniti akcija pa gest. ...
health.taragana.net/articles/italian-researchers-identifies-gesture-recognition-sites-in-brain-stroke-victims-benefitted/sr/

istraživač - Wiktionary
[edit] Noun. **istraživač** m. sg. explorer · investigator. Retrieved from "http://en.wiktionary.org/wiki/istra%C5%BEiva%C4%8D". Category: Croatian

"istraživač"OR"naučnik"OR"učenjak"OR"истраживач"OR"научни

Popis istraživača - Wikipedija
Antonio de Abreu (16. stoljeće), portugalski **istraživač** Indonezije 16. stoljeće), portugalski diplomat i **istraživač** u Etiopiji ...
hr.wikipedia.org/wiki/Popis_istra%C5%BEiva%C4%8Da

Hernando de Soto (**istraživač**) - Wikipedija
Nedovršeni članak Hernando de Soto (**istraživač**) koji govori o istr Dopunite ga prema pravilima Wikipedije. ...
hr.wikipedia.org/wiki/Hernando_de_Soto_(istra%C5%BEiva%C4%

Matični stanica Istraživanje Blog » **Naučnik** Razvijen M
Naučnik Razvijen Matični stanica zametka Zameniti iz Miš Koža **Istraživač** Tvrditi jedan Važan Proboj umjesto Mišićav Dystrophy
stemcell.taragana.net/archive/scientists-developed-embryonic-ster skin-cell/sr/

YouTube - Nikola Tesla Srpski **naučnik**-Nikola Tesla S
www.youtube.com/?v=nEaSKmZ1EXU

Умне падавине: Српски **научник** у посети Хрватској
Познати српски **научник** Војислав Коштуница јуче се састао с **naučnik**... Evo jednog lepo napisanog portreta dotičnog naučnika.
pademnaum.blogspot.com/2006/07/blog-post.html

Naučnik Predložiti za Odstraniti 90% Ljudsko biće Pop
News and views as a Software Architect and Entrepreneur on Java

Figure 5. Results for expanded query for 'istraživač'

6. Conclusion

Given the rapidly growing number of documents on the web, the formulation of queries that are submitted to web search engines has become an increasingly sensitive matter. Queries often need to be 'fine tuned' in order to obtain an optimal balance between recall and precision. Lexical resources can be put to the aid of the user by offering him/her various possibilities of query expansion, with the ultimate aim of obtaining a better balanced query. We believe that the approach we have outlined in this paper purports this thesis.

Needless to say, lexical resources are invaluable for many other tasks, and some of them can already be performed using the tool that we have described here in the context of query expansion. Our further endeavors will hence be twofold. On the one hand, we shall continue do develop our lexical resources, focusing in the next stage on dictionaries of compounds. On the other hand, we will strive to broaden the scope of tasks that can be solved with our tools.

The existence of reliable lexical resources is already indispensable, but their importance, along with the tools for handling them, can only grow in the future.

7. References

Courtois, Blandine; Max Silberztein (eds.) (1990). *Dictionnaires électroniques du français*. Langue française 87. Paris: Larousse

http://www-igm.univ-mlv.fr/~unitex/

Krstev, C., et al., (2008). Resources and Methods in the Morphosyntactic Processing of Serbo-Croatian, In *Formal Description of Slavic Languages: The Fifth Conference*, Leipzig 2003, Zybatow, Gerhild et al. (eds.), Peter Lang: Frankfurt am Main, pp. 3-17...

Krstev, C., Stanković, R., Vitas, D., Obradović, I. (2006). *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

Krstev, C., Vitas, D., Maurel, D., Tran, M. (2005). *Multilingual Ontology of Proper Names. In Proc. of Second Language & Technology Conference*, Poznań, Poland, April 21-23, Wydawnictwo Poznańskie Sp. z o.o, Poznań

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

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