Multilingual Vocabularies in Open Access: Semantic Network WordNet

INFORUM 2016: 22nd Annual Conference on Professional Information Resources, Prague 24–25 May 2016

MSc Sanja Antonic
antonic@unilib.bg.ac.rs

MSc Oja Krinulovic
okrinulovic@unilib.bg.ac.rs

MSc Mile Stijepović
stijepovic@unilib.bg.ac.rs

University of Belgrade
University library "Svetozar Markovic"
www.unilib.bg.ac.rs

Abstract: WordNet, known as Princeton WordNet (PWN), is a lexico-semantic network which was created at Princeton University in 1985. It has been developing constantly. Princeton WordNet contains information about nouns, verbs, adjectives and adverbs in English. They are grouped into sets of synonyms – synsets, and every synset represents a particular concept. The main semantic relation among synsets is synonymy and there are also hyponymy, antonymy, meronymy etc. The next phase was multulinguality. The project EuroWordNet (EWN) was built for Dutch, Italian, Spanish, German, French, Czech, Estonian and English. It had the new component named Inter-Lingual-Index (ILI) which connects the synsets in different languages with synsets in PWN. The next multilingual project was BalkaNet started in September 2001 and finished in August 2004. The main goal was to align Balkan languages to PWN: Bulgarian, Greek, Romanian, Serbian, Turkish and to extend the Czech wordnet, which was part of the EuroWordNet project. When the BalkaNet project finished, Serbian Wordnet (SWN) continued to develop according to innovations in Princeton WordNet (versions 2.0 and 3.0). Moreover, Serbian WordNet (SWN), is one of around 150 wordnets in multiple languages, which are members of The Global WordNet Organization. This organization supports open access and connects all these wordnets to WordNet (Princeton or others that are linked to PWN). Finally, Open Multilingual WordNet combines wordnets in open access, data from Wiktionary, the product of Wikimedia, and the Unicode Common Locale Data Repository.

Keywords: Open Access, Wordnet, Multulinguality, Serbian Wordnet, SWN.

Wordnet is defined on web site Prinston Wordnet (PWN) as :” WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.”

EuroWordNet   (EWN) added a new perspective to the Princeton WordNet (PWN) , namely “multilinguality”.. New project (EWN), started in March 1996 and ended in June 1999. EuroWordNet was build for Dutch, Italian, Spanish, German, French, Czech, Estonian and English languages at the beginning It was absolutely and very significant new feature and there were some changes need to be done.

EuroWordNet had the following objectives:

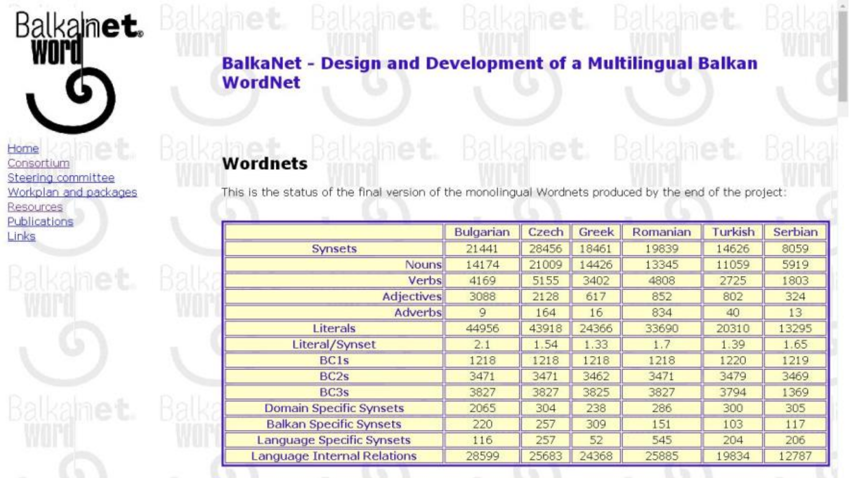
1) to create a multilingual database;

2) to maintain language-specific relations in the wordnets;

3) to achieve maximal compatibility across the different resources;

4) to build the wordnets relatively independently (re)-using existing resources;

BalkaNet was an EC funded project   started in September 2001 and finished in August 2004. It aimed at developing aligned wordnets for the following Balkan languages: Bulgarian, Greek, Romanian, Serbian, Turkish and to extend the Czech wordnet previously developed in the EuroWordNet project. The BalkaNet partners decided to use concepts from other languages (mainly English in Princeton WordNet) that are not lexicalized in their particular languages.



BalkaNet - Design and Development of a Multilingual Balkan WordNet

Wordnets

This is the status of the final version of the monolingual Wordnets produced by the end of the project:

	Bulgarian	Czech	Greek	Romanian	Turkish	Serbian
Synsets	21441	28456	18461	19839	14626	8059
Nouns	14174	21009	14426	13345	11059	5919
Verbs	4169	5155	3402	4808	2725	1803
Adjectives	3088	2128	617	852	802	324
Adverbs	9	164	16	834	40	13
Literals	44956	43918	24366	33690	20310	13295
Literal/Synset	2.1	1.54	1.33	1.7	1.39	1.65
BC1s	1218	1218	1218	1218	1220	1219
BC2s	3471	3471	3462	3471	3479	3469
BC3s	3827	3827	3825	3827	3794	1369
Domain Specific Synsets	2065	304	238	286	300	305
Balkan Specific Synsets	220	257	309	151	103	117
Language Specific Synsets	116	257	52	545	204	206
Language Internal Relations	28599	25683	24368	25885	19834	12787

Figure1. BalkaNet

Serbian wordnet (SWN) was built on the basis of Princeton WordNet v. 2.0 and synchronized at the end with the Princeton WordNet v. 3.0. SWN nowadays contain 21 877 synsets which are connected by 60 476 semantic-lexical relations.

In April 2016 statistic results was following

POS	number
nouns	17 922
verbs	2 209
adjectives	1 622
adverbs	124

Table 1. Number od POS (Part of Speech) in Serbian Wordnet

Statitic data about relation literal –synset:

1 literal in 11982 sinsets (it means that 11 982 synsets are described with only 1 literal, etc.)

4 literals in 585 synsets

2 literals in 6950 synsets

5 literals in 201 synsets

3 literals in 2042 synsets

7 literals in 38 synsets

6 literals in 64 synsets

10 literals in 2 synsets

8 literals in 10 synsets

9 literals in 2 sinsets

13 literals in 1 synsets (stoznaci da je samo 1 sinsetopisansa 13 literala)

Number of different relation :

hypernym 20035

hyponym 20768

eng_derivative 2992

near_antonym 1115

subevent 80

category_domain 1040

verb_group 185

also_see 220

causes 66

holo_part 1856

mero_portion 2052

holo_member 3910

holo_portion 223

mero_member 3914

specifiedBy 71

usage_domain 17

be_in_state 288

specificOf 72

similar_to 258

derived 681

particle 10

derived-vn 3

derived-gender 38

derived-pos 45

region_domain 149

substanceHolonym 5

substanceMeronym 8

Hypernym 54

Hyponym 22

SubstanceMeronym 2

RegionDomain 2

TopicDomain 1

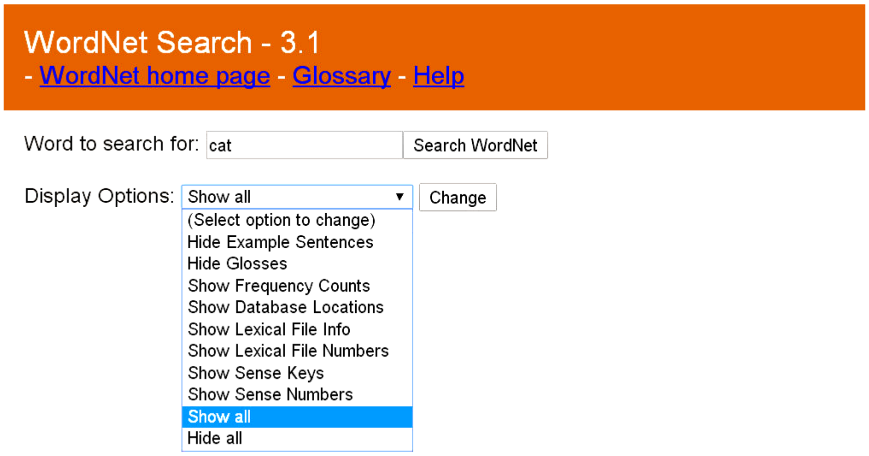
InstanceHyponym 1

Entailment 2

partMeronym 72

Multilingual vocabularies in Open Access or more precisely multilinguality in main topic in this paper we will try to demonstrate on very simple word “cat”.

In Prinston Wordnet it looks



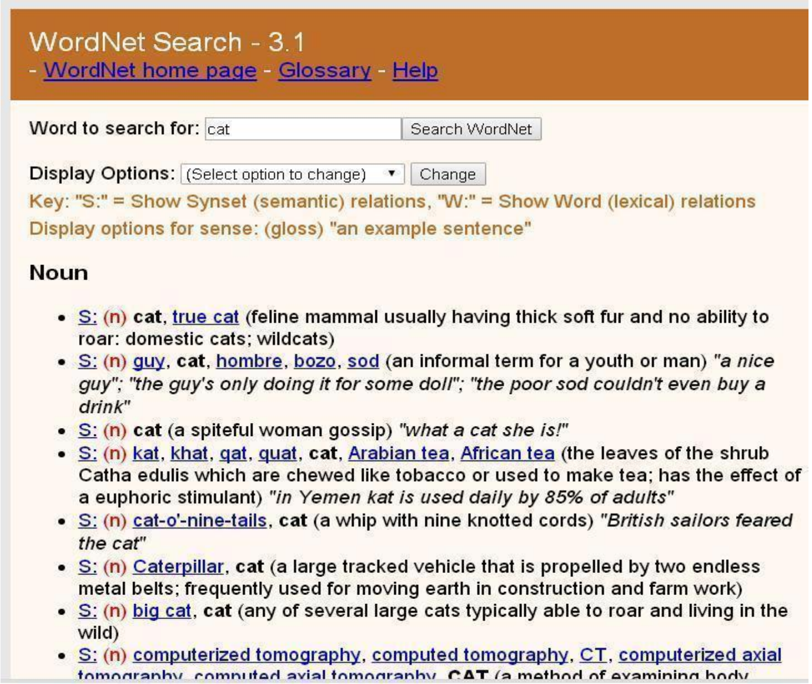
WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: cat

Display Options: Show all

Figure 2. Prinston Wordnet - searching



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: cat Search WordNet

Display Options: (Select option to change) ▾ Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **cat**, true cat (feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats)
- S: (n) guy, **cat**, hombre, bozo, sod (an informal term for a youth or man) *"a nice guy"; "the guy's only doing it for some doll"; "the poor sod couldn't even buy a drink"*
- S: (n) **cat** (a spiteful woman gossip) *"what a cat she is!"*
- S: (n) kat, khat, qat, quat, **cat**, Arabian tea, African tea (the leaves of the shrub Catha edulis which are chewed like tobacco or used to make tea; has the effect of a euphoric stimulant) *"in Yemen kat is used daily by 85% of adults"*
- S: (n) cat-o'-nine-tails, **cat** (a whip with nine knotted cords) *"British sailors feared the cat"*
- S: (n) Caterpillar, **cat** (a large tracked vehicle that is propelled by two endless metal belts; frequently used for moving earth in construction and farm work)
- S: (n) big cat, **cat** (any of several large cats typically able to roar and living in the wild)
- S: (n) computerized tomography, computed tomography, CT, computerized axial tomography, computed axial tomography, **CAT** (a method of examining body

Figure 3. Prinston Wordnet- results

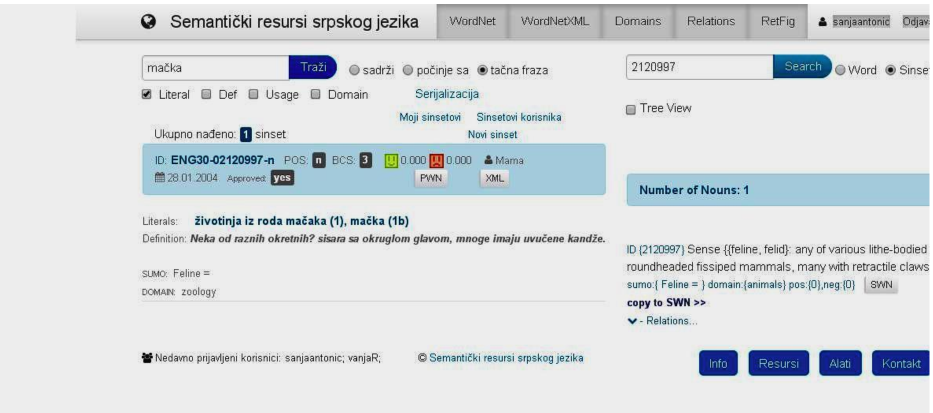
For realization of multilinguality is very important The Global WordNet Association is: “ free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world.” We have to emphysize that The Global WordNet Association (GWA) builds on the results of Princeton WordNet (PWN) and EuroWordNet (EWN). GWA posts links to resources that follow the wordnet design, which includes links to WordNet (Princeton or others that are linked to PWN) and WN structure (minimally: synset, hyponymy).

There are significant number of different languages from all over the world and most of them are in freely available. We could mentioned just a few like a Bengali, Bulgarian, Czech, Chinese, Hebrew, Latin, Serbian, Turkish, Swedish, Irish and many others.

For example, The Czech WordNet was developed by the Centre of Natural Language Processing at the Faculty of Informatics, Masaryk University, Czech Republic. The old version is in Open Access, but new not yet. The Czech WordNet contains 28,201 word senses (synsets). Every synset encodes the equivalence relation between several literals (at least one is present), having a unique meaning (specified in the SENSE tag value), belonging to one and the same part of speech (specified in the POS tag value), and expressing the same lexical meaning. Each Czech synset is related to the corresponding synset in the Princeton WordNet 2.0.

But also it is much easier to form association as , for example, IndoWordNet which develop multilingual wordnets in Hindi, Assamese, Bengali, Bodo, Gujarati, Kannada, Kashmiri, Konkani, Malayalam, Meitei, Marathi, Nepali, Sanskrit, Tamil, Telugu, Punjabi, Urdu, etc.

And finaly we demonstrate English word cat from PWN in other languages.



mačka

○ sadrži ○ počinje sa ● tačna fraza

☑ Literal □ Def □ Usage □ Domain

Ukupno nađeno: 1 sinset

ID: ENG30-02120997-n POS: n BCS: 3 0.000 0.000 Mama

28.01.2004 Approved: yes

Literals: životinja iz roda mačaka (1), mačka (1b)

Definition: ***Neka od raznih okretnih? sisara sa okruglom glavom, mnoge imaju uvučene kandže.***

SUMO: Feline =

DOMAIN: zoology

Nedavno prijavljeni korisnici: sanjaantonic; vanjaR;

© Semantički resursi srpskog jezika

2120997

○ Word ● Sinse

□ Tree View

Number of Nouns: 1

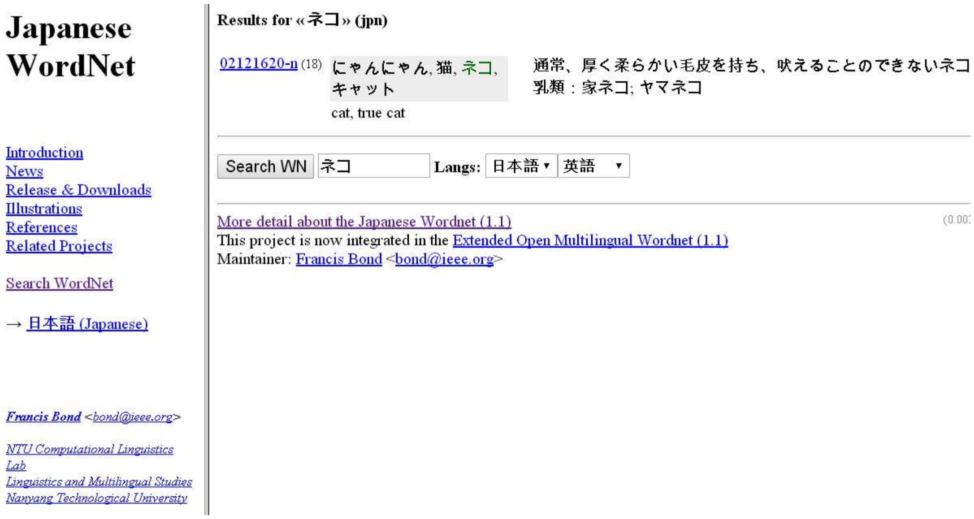
ID {2120997} Sense {{feline, felid}: any of various lithe-bodied roundheaded fissiped mammals, many with retractile claws

sumo:{ Feline = } domain:{animals} pos:{0},neg:{0}

copy to SWN >>

- Relations...

Figure 4 . Serbian Wordnet



Francis Bond <*bond@ieee.org*>

NTU Computational Linguistics Lab
Linguistics and Multilingual Studies
Nanyang Technological University

Results for «ネコ» (jpn)

02121620-n (18) にゃんにゃん, 猫, ネコ, キャット
cat, true cat

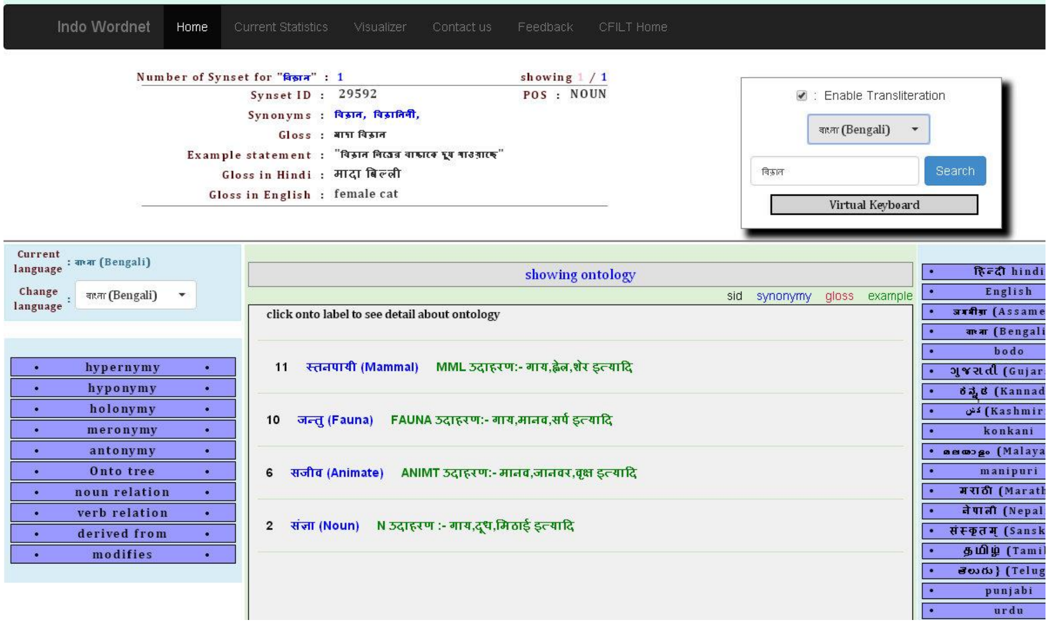
通常、厚く柔らかい毛皮を持ち、吠えることのできないネコ乳類：家ネコ; ヤマネコ

ネコ

More detail about the Japanese Wordnet (1.1)
This project is now integrated in the Extended Open Multilingual Wordnet (1.1)
Maintainer: Francis Bond <bond@ieee.org>

(0.00

Figure 5. Japanese Wordnet



Number of Synset for "বিড়াল" : 1 showing 1 / 1

Synset ID : 29592 POS : NOUN

Synonyms : বিড়াল, বিড়ালিনী,

Gloss : মাদা বিড়াল

Example statement : "বিড়াল নিজের বাচ্চাকে দুধ খাওয়াচ্ছে"

Gloss in Hindi : मादा बिल्ली

Gloss in English : female cat

Current language : বাংলা (Bengali)

Change language : বাংলা (Bengali)

showing ontology

sid synonymy gloss example

click onto label to see detail about ontology

11 स्तनपायी (Mammal) MML उदाहरण:- गाय,ह्वेल,शेर इत्यादि

10 जन्तु (Fauna) FAUNA उदाहरण:- गाय,मानव,सर्प इत्यादि

6 सजीव (Animate) ANIMT उदाहरण:- मानव,जानवर,वृक्ष इत्यादि

2 संज्ञा (Noun) N उदाहरण :- गाय,दूध,मिठाई इत्यादि

Figure 6. Hindu Wordnet

Amazing variety of numerous languages will be expressed and interconnected via many multilingual online freely available dictionaries. There are many ideas, efforts, enthusiasm , projects and very serious work in academic institutions. One of impressive results by Bond and Foster who created an open multilingual wordnet with over 26 languages. It is made by combining wordnets with open licences, data from the Unicode Common Locale Data Repository and Wiktionary. Overall there are over 2 million senses for 117,659 concepts, using over 1.4 million words in hundreds of languages. It demonstrated the ability to automatically identify many matching senses in Wiktionary and WordNet based on the similarity of monolingual features. Their study combines monolingual features with the disambiguating power of multiple languages. In a future, we can expect a lot achievements in multilinguality, very interesting and significant field of science.

References:

Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu (accessed 13.04.2016)

Christiane Fellbaum. 2010. *WordNet*. Springer, Netherlands.

Piek Vossen. 1998. *Introduction to EuroWordNet*. Computers and the Humanities 32(2-3):73-89.

Miller G., R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography, Vol 3, No.4, 235-244.

George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Fellbaum, C. 2012. WordNet. The Encyclopedia of Applied Linguistics.

Stamou, S., Oflazer, K., Pala, K., Christodoulakis, D., Tufis, D., Koeva, S., Totkov, G., Totkov, D., & Grigoriadou, M. (2002). BALKANET: A Multilingual Semantic Network for Balkan Languages. In *Proceedings of the 1st International Global WordNet Conference*. Mysore, India.

Tufis, D., Cristea, D., & Stamou, S. (2004). BalkaNet: Aims, methods, results and perspectives. a general overview. Romanian Journal of Information science and technology, 7(1-2), 9-43.

Krstev, Cvetana, et al. "Using textual and lexical resources in developing serbian wordnet." Romanian Journal of Information Science and Technology 7.1-2 (2004): 147-161.

Krstev, C., Đorđević, B., Antonić, S., Ivković-Berček, N., Zorica, Z., Crnogorac, V., & Macura, Lj. (2008). Cooperative Work in Further Development of Serbian WordNet. *INFOtheca 9(1-2)*, 59a–78a.

Mladenović, M., Mitrović, J., & Krstev, C. (2014). Developing and Maintaining a WordNet: Procedures and Tools. In H. Orav, C Fellbaum & P Vossan (Eds.), *Proceedings of Seventh Global WordNet Conference 2014*, 55–62. University of Tartu, Tartu, Estonia.

Stanković, R., & Obradović, I. (2009). An Integrated Environment for Management and Exploitation of Linguistic Resources. *Proceedings of the International Multiconference on Computer Science and Information Technology, Computational Linguistics - Applications Workshop, CLA'09*, pp. 287–294. Mragowo, Poland. .

Pala, K., & Smrž, P. (2004). Building czech wordnet. Romanian Journal of Information Science and Technology, 7(2-3), 79-88.

The Global WordNet Organization http://globalwordnet.org/ (accessed 11.03.2016)

Bond, F., & Paik, K. (2012). A survey of wordnets and their licenses. Small, 8(4), 5