




INFORMATION EXTRACTION USING NATURAL LANGUAGE PROCESSING



Cvetana Krstev
University of Belgrade, Faculty of Philology



Information Retrieval and/vs. Natural Language Processing



So close yet so far



Outline of the talk

- Views on Information Retrieval (IR) and Natural Language Processing (NLP)
- IR and NLP in Serbia
- Language Resources (LT) in the core of NLP
 - at University of Belgrade (4 representative resources)
- LR and NLP for Information Retrieval and Information Extraction (IE)
 - at University of Belgrade (4 representative applications)

Wikipedia

- **Information retrieval**

- Information retrieval (IR) is **the activity** of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing.

- **Natural Language Processing**

- Natural language processing is **a field** of **computer science**, **artificial intelligence**, and **computational linguistics** concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve: natural language understanding, enabling computers to derive meaning from human or natural language input; and others involve natural language generation.

Experts

- **Information Retrieval**

- As an academic **field of study**, Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collection (usually stored on computers).
- C. D. Manning, P. Raghavan, H. Schutze, “Introduction to Information Retrieval”, Cambridge University Press, 2008

- **Natural Language Processing**

- The term ‘Natural Language Processing’ (NLP) is normally used to describe the **function** of software or hardware components in computer system which analyze or synthesize spoken or written language. ‘Natural Language Understanding’ is associated with more ambitious goal of having a computer system actually comprehend natural language as a human being might.
- P. Jackson and I. Moulinier, „Natural Language Processing for Online Applications – Text retrieval, Extraction and Categorization, John Benjamins Publishing Co., 2007

ACM classification

- Information systems
 - **Information retrieval**
 - Document representation, Information retrieval query processing, Users and interactive retrieval, Retrieval models and ranking (**Language models**), Retrieval tasks and goals (**Information extraction**), Evaluation of retrieval results, Search engine architectures and scalability, Specialized information retrieval
- Computing methodologies
 - Artificial Intelligence
 - **Natural Language Processing**
 - **Information extraction**, Machine translation, Discourse, Dialogue and pragmatics, Natural language generation, Speech recognition, Lexical semantics, Phonology / morphology, Language resources

28 classification of Information Science (A Knowledge Map of Information Sciences, one of four Critical Delphi studies conducted in 2003-2005)

- **Information retrieval:**

- First level (2), under Information technology (2), Information processing (2), Information Systems, Electronic information systems & services, Systems & products, Data organization & retrieval, Knowledge Organization, Information and Knowledge, Retrieval and use of information & knowledge, Information Use Process, Activities, Metrics, evaluation & research, Information Science Research, Disciplines, Information studies, Technological disciplines, Processes, entities & institutions in information work, Communication, Purpose, other groups

- **Natural Language Processing:**

- as Natural language for searching, Automatic processing of language

- **Artificial Intelligence:**

- under Information science epistemology, Concepts (2), Cognition, Information technology (3), Information processing, storing & communication processes, Utilization, Retrieval and use of Information & Knowledge, Supporting disciplines, Related disciplines & tools

- **Common:**

- Information technologies, Retrieval and use of Information & Knowledge

META-net White Paper Series

Languages in the Digital Age – Serbian (1)

Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	3	5	5	5	4	4
Parsing (shallow or deep syntactic analysis)	1	2	5	3	2	2	2
Sentence Semantics (WSD, argument structure, semantic roles)	0	0	0	0	0	0	0
Text Semantics (coreference resolution, context, pragmatics, inference)	0	0	0	0	0	0	0
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0	0	0	0	0	0	0
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	3	1	3	3	2	2	3
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	1	2	2	2	3	2	3
Language Generation (sentence generation, report generation, text generation)	0	0	0	0	0	0	0
Summarization, Question Answering, advanced Information Access Technologies	1	1	0	1	0	1	1
Machine Translation	1	1	0	1	0	1	1
Speech Recognition	2	2	1	1	1	1	0
Speech Synthesis	2	2	4	4	5	5	1
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0

Grades:

- Quantity
- Availability
- Quality
- Coverage
- Maturity
- Sustainability
- Adaptability

META-net White Paper Series

Languages in the Digital Age – Serbian (2)

Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	2	4	2	4	4	4	4
Syntax-Corpora (treebanks, dependency banks)	0	0	0	0	0	0	0
Semantics-Corpora	0	0	0	0	0	0	0
Discourse-Corpora	0	0	0	0	0	0	0
Parallel Corpora, Translation Memories	3	3	3	2	2	2	3
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	1	2	4	4	3	3	3
Multimedia and multimodal data (text data combined with audio/video)	1	2	2	1	2	1	2
Language Models	1	3	2	3	2	2	3
Lexicons, Terminologies	2	3	4	4	3	3	3
Grammars	1	1	0	1	0	1	1
Thesauri, WordNets	2	4	3	2	4	2	4
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	1	1	0	1	0	1	1

Grades:

- Quantity
- Availability
- Quality
- Coverage
- Maturity
- Sustainability
- Adaptability

LANGUAGE RESUOURCES



1. Reference corpora – The Corpus of Contemporary Serbian (SrpKor)

- The Corpus of contemporary Serbian, SrpKor, consists of 4,925 texts.
- Total size of SrpKor is 118,767,279 words.
- It is lemmatized and PoS tagged using TreeTagger.
- SrpKor texts consist of:
 - fiction written by Serbian authors in 20th and 21th century (10,191,092 words),
 - various scientific texts from various domains (both humanities and sciences) (3,542,169 words),
 - legislative texts (6,874,318 words) and
 - general texts (98,159,700 words).
 - daily news
 - texts in journals and magazines
 - internet portal texts
 - agency news 1995-96,
 - newspaper feuilletons.

Use of SrpKor

- Web Interface (IMS Open Corpus Workbench) and efficient query processor CQP
 - Used by humans, mostly linguists and philologists
- For applications
 - Training language models (bigrams, trigrams,...)
 - Bag-of-Words (BOW) for general texts (general domain)

2. Parallel corpora – aligned corpora

- Multilingual / bilingual corpora (one of languages is Serbian);
- Texts are segmented – paragraphs, **sentences, phrases**, words – and aligned;
 - **English/Serbian**
 - English source texts translated into Serbian, and vice versa, and English and Serbian translations of texts originally written in language X.
 - The texts belong to various domains: fiction, general news, scientific journals, web journalism, health, law, education, movie sub-titles.
 - The alignment was performed on the sub-sentential level.
 - The size of the corpus is 5,078,280 words (2,672,911 in the English part, 2,405,369 in the Serbian part).
 - **French/Serbian**
 - French or Serbian source literary and newspaper texts and their translations.
 - The size of the corpus is 59,425 aligned segments and 1,948,679 words (1,063,564 in the French part, 885,115 in the Serbian part).

Illustration of an aligned text – a human view

Aligned text - Jane Austen - Pride and Prejudice - 5 chapters	
English	Serbian
n1 : It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.	n1 : Opsxte je poznata istina da je bogatom neozxenenom cyoveku xzena neophodna.
n2 : However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters.	n2 : Ma kako da su malo poznata osećanja i gledišta takvog cyoveka prilikom nxegova prvog dolaska u neko susedstvo, ta je istina tako duboko ukorenjena u svesti susednih porodica, da se on smatra punopravnom svojinom ove ili one nxihove kcxeri.
n3 : "My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"	n3 : Dragi moj Benete - recyue mu jednog dana nxegova xzena - jesi li cyuo da je Nederfildski park najzad iznajmljen?
n4 : Mr. Bennet replied that he had not.	n4 : Benet odgovori da nije cyuo.
n5 : "But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."	n5 : Pa eto, jeste - izjavi ona - gospodxa Long je malo pre bila kod mene i ispricjala mi sve o tome.
n6 : Mr. Bennet made no answer.	n6 : Benet nisxta ne odgovori.
n7 : "Do you not want to know who has taken it?" cried his wife impatiently.	n7 : Zar ne xzelisx da znasx ko ga je uzeo? - nestrplxivo upita nxegova xzena.
n8 : "_You_" want to tell me, and I have no objection to hearing it."	n8 : Ti xzelisx da mi kazxesx, te nemam nisxta protiv da to cyujem.
n9 : This was invitation enough.	n9 : Ovo je bilo dovolxno ohrabrenxe.
n10 : "Why, my dear, you must know, Mrs. Long says that Netherfield is taken by a young man of large fortune from the north of England; that he came down on Monday in a chaise and four to see the place, and was so much delighted with it, that he agreed with Mr. Morris immediately; that he is to take possession before	n11 n12 : Gospodxa Long kazxe da je Nederfild zakupio vrlo bogat mladix iz severne Engleske; da je dolazio u ponedelxak u cyetvoroprezxnim kocyijama da vidi to mesto i bio toliko ocyaran nxime da se odmah pogodio s gospodinom Morisom. On cxe se doseliti pre Miholxdana, a neke nxegove sluge bicxe

Use of parallel corpora

- Web Interface (IMS Open Corpus Workbench) and efficient query processor CQP
- Bibliša – the aligned text digital library (more later)
 - Used by humans, mostly translators, linguists and philologists
- For applications
 - Machine translation
 - multi- / bi-lingual terminology extraction

Use of parallel texts – a query

socijaln[a-z]* osiguranj[a-z]* *social insurence*

help of occupational health

2614234: 3 . da nastoje da postupno podignu sisten [g osiguranja](#) na viši nivo ;

EN: 3 . to endeavour t

2614207: 2 . da održe onom koji je potreban

EN: 2 . to maintain the ssary for the ratification of the Euro

2614325: b) dodeljiv prema zakonodavstvu sredstvima kao što je akumulacija osig anja ili perioda zaposlenosti koji su ostvareni

EN: b) the granting , of each of the Parties s the accumulation of insurance or employment periods completed under the legislation

2781426: Prava iz [<socijalnog osiguranja>](#) za građane koji nisu obuhvaćeni obaveznim socijalnim osiguranjem uređuju se zakonom .

EN: Social security right for those citizens who are not covered by the obligatory social insurance scheme shall be regulated by law

2344131: U međuvremenu , republikanci su postali ekstremni kakvi nisu bili pre tri generacije ; uporedite totalnu opoziciju kakva je Obamu dočekala u ekonomskim poslovima sa činjenicom da je većina republikanaca u Kongresu 1935 . podržala Ruzveltov najvažniji zakon , zakon o [<socijalnom osiguranju>](#) .

EN: Meanwhile , Republicans have become extremists in a way they weren ' t three generations ago ; contrast the total opposition Obama has faced on economic issues with the fact that most Republicans in Congress voted for , not against , FDR ' s crowning achievement , the Social Security Act of 1935

2614298: a) jednak tretman državljana drugih država ugovornica sa tretmanom sopstvenih državljana u pogledu prava na koje proističu iz zakonodavstva o [<socijalnom osiguranju>](#) , bez obzira na akcije koje zaštićena lica mogu da preduzmu izn

EN: a) equal treatment with their own nationals of the nationals of other Parties in respect of social security rights , including legislation , whatever movements the persons protected may undertake between the territories of the Parties

2105867: S druge strane , 6 . 330 penzionera već primaju svoje penzije iz Hrvatske , od kojih 85 odsto preko Komercijalne banke . [12] Poseban problem je što Sporazum o [<socijalnom osiguranju>](#) ne predviđa naknadu za neizmirene penzije u poslednjih 12 godina .

EN: On the other hand , 6 . 330 pensioners are already receiving their Croatian pensions , of which 85 % via the Commercial Bank . [19] Particular problem is that the Agreement on social security does not envisage the compensation for unpaid pensions during past 12 years

2072443: U Srbiji je 2003 . donet Zakon o penzijskom i [<socijalnom osiguranju>](#) (SI . glasnik RS , br . 32 / 03) .

EN: In 2003 a new Act on Pension and Social Insurance (SI . glasnik RS , No . 34 / 03) was passed in Serbia

Socijalnog osiguranja

Socijalnom osiguranju

Solution to the problem of various inflectional forms

Parallel corpora – an application view (TMX – XML)

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<!--tmx SYSTEM ..test.-->
<tmx version="1.3">
  <header creationtool="MoramoMuSmislitiIme" creationtoolversion="1.3" segtype="sentence" datatype="plaintext" o-tmf=
    "MoramoMuSmislitiIme 1.0" adminlang="EN" srclang="EN" />
  <body>
    <tu>
      <prop type="Domain">Fajlovi: NBS Prvi - NBS Drugi</prop>
      <tuv lang="EN" creationid="n1 " creationdate="20040101T000000Z">
        <seg>It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a
          wife. </seg>
      </tuv>
      <tuv lang="SR" creationid="n1 " creationdate="20040101T000000Z">
        <seg>Opsjte je poznata istina da je bogatom neozxenxenom cyoveku zxena neophodna. </seg>
      </tuv>
    </tu>
    <tu>
      <prop type="Domain">Fajlovi: NBS Prvi - NBS Drugi</prop>
      <tuv lang="EN" creationid="n2 " creationdate="20040101T000000Z">
        <seg>However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth
          is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or
          other of their daughters. </seg>
      </tuv>
      <tuv lang="SR" creationid="n2 " creationdate="20040101T000000Z">
        <seg>Ma kako da su malo poznata osecxanxa i gledisxta takvog cyoveka prilikom nxegova prvog dolaska u neko susedstvo,
          ta je istina tako duboko ukorenxena u svesti susednih porodica, da se on smatra punopravnom svojinom ove ili one
          nxihove kcxeri. </seg>
      </tuv>
    </tu>
    <tu>
      <prop type="Domain">Fajlovi: NBS Prvi - NBS Drugi</prop>
      <tuv lang="EN" creationid="n3 " creationdate="20040101T000000Z">
        <seg>"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?" </seg>
      </tuv>
```

3. Morphological e-dictionaries of Serbian (SrpMD)

- Serbian language is from the Slavic family – very rich inflection and derivation
- Dictionaries are in the format known as LADL – introduced for French by **Maurice Gross**, linguist and computer linguist from **Laboratoire d'automatique documentaire et linguistique**, Paris, France
- E-dictionaries in this format exist for many languages
- Basically, the format consists of two dictionaries:
 - **DELAS** – produced ‘manually’ to describe properties of lemmas; it is used to produce automatically:
 - **DELAF** – a dictionary of word forms with their morphological properties.

An example from SrpMD – for **žena** ‘woman’, **čovjek** ‘man’ and **ljudi** ‘men’

◦ **žena, N601+Hum**

- žena, žena.N:fp2v:fs1v
- žene, žena.N:fp5v:fp4v:fp1v:fw2v:fs2v
- ženi, žena.N:fs7v:fs3v
- ženu, žena.N:fs4v
- ženo, žena.N:fs5v
- ženom, žena.N:fs6v
- ženama, žena.N:fp7v:fp6v:fp3v

◦ **čovjek, N1010+Hum+Ek**

- čovek, čovek.N:ms1v
- čoveka, čovek.N:ms4v:ms2v
- čoveku, čovek.N:ms7v:ms3v
- čoveče, čovek.N:ms5v
- čovekom, čovek.N:ms6v

◦ **ljudi, N3067+Hum**

- ljudima, ljudi.N:mp7v:mp6v:mp3v
- ljude, ljudi.N:mp4v
- ljudi, ljudi.N:mp5v:mp2v:mp1v

Multiword expressions (or units) in SrpMD

- Various definitions:
 - “Idiosyncratic interpretations that cross word boundaries (spaces)”
 - “MWEs describe different but related phenomena, including fixed or semi-fixed phrases, compounds, support verbs, idioms, phrasal verbs, collocations, etc. “
 - “MWEs are decomposable into multiple simplex words and lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic.”
- SrpMD covers following MWEs
 - contiguous
 - Various Part-Of-Speech except verbs
- Size of SrpMD
 - simple word lemmas – 140.000; corresponding number of morphological forms > 5mil.
 - MWU lemmas – 18.000

An example from SrpMD – MWE

radnički pokret ‘workers’ movement’

- **radnički(radnički.A2:adms1g) pokret(pokret.N1:ms1q),N_AXN+PolProcess**
 - radnički pokret,radnički pokret.N:ms1q:ms4q:ms5q.N:mp1q:mp5q
 - radničkoga pokreta,radnički pokret.N:ms2q
 - radničkog pokreta,radnički pokret.N:ms2q
 - radničkomu pokretu,radnički pokret.N:ms3q:ms7q
 - radničkome pokretu,radnički pokret.N:ms3q::ms7q
 - radničkom pokretu,radnički pokret.N:ms3q:ms7q
 - radničkim pokretom,radnički pokret.N:ms6q
 - radničkih pokreta,radnički pokret.N:mp2q
 - radničkima pokretima,radnički pokret.N:mp3q:mp6q:mp7q
 - radničkim pokretima,radnički pokret.N:mp3q:mp6q:mp7q
 - radničke pokrete,radnički pokret.N:mp4q
 - radnička pokreta,radnički pokret.N:mw2q

Use of SrpMD

- **Unitex** corpus processing system
 - By humans, mostly linguists (computer linguists, corpus linguists, etc.), philologists, (formal) grammar developers
- By applications:
 - Tokenization, sentence boundaries;
 - POS tagging, lemmatization;
 - Syntax grammars for shallow parsing;
 - Text annotation

4. WordNet - What Wikipedia says about it

- **WordNet** is a [lexical database](#) for the [English language](#).
- It groups [English words](#) into sets of [synonyms](#) called [synsets](#), provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of [dictionary](#) and [thesaurus](#).
- While it is accessible to human users via a web browser, its primary use is in automatic [text analysis](#) and [artificial intelligence](#) applications.
- The [database](#) and [software](#) tools have been released under a [BSD style license](#) and are freely available for download from the WordNet website. Both the lexicographic data (*lexicographer files*) and the compiler (*called grind*) for producing the distributed database are available.

Synset – the basic unit of WordNet

- **Synset** – synonym set;
- A synset is a representation of a concept – a definition is added only to facilitate development and usage;
- „These synonym sets (synsets) do not explain what the concepts are; they merely signify that the concepts exist.“
- Synsets are connected with various relations: hypernymy/hyponymy, meronymy, antonymy, etc.
- **Hypernymy/hyponymy** is a basic relation – more general concepts/more specific concepts
 - Hierarchical structure

A wordform – concept relation

- This relation is many-to-many
- Example:
 - {**board**, **plank**} - def: a stout length of sawn timber; made in a wide variety of sizes and used for many purposes
 - {**board**, **table**} - def: food or meals in general; usage: „she sets a fine table“; „room and board“
- A concept can be lexicalized by several word forms (one concept – two word forms, *board* and *plank*)
- A word form can be used for lexicalization of several concepts (one word form – *board* or *table*– can be used for two and many more concepts)

An example from Princeton Wordnet

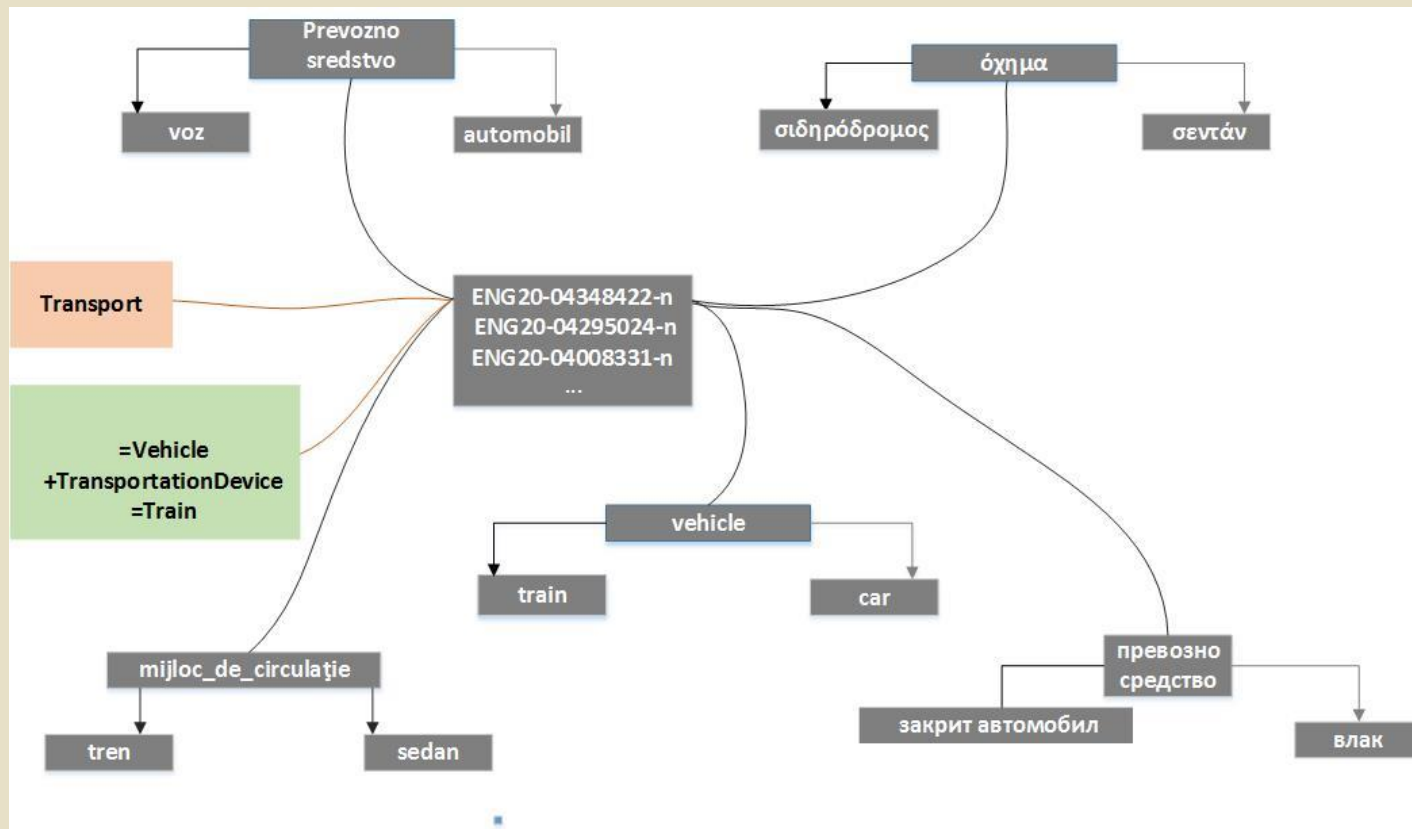
- Example of one noun synset:
- Synset
 - {dog, domestic_dog, Canis_familiaris}
- Definition
 - a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds;
- Usage
 - "the dog barked all night"

Dog – upward hierarchy

```
{entity}
  {physical_entity}
    {object, physical_object}
      {whole, unit}
        {living_thing, animate_thing}
          {organism, being}
            {animal, animate_being, beast, brute, creature, fauna}
              {chordate}
                {vertebrate, craniate}
                  {mammal, mammalian}
                    {placental, placental_mammal, eutherian, eutherian_mammal}
                      {carnivore}
                        {canine, canid}
                          {dog, domestic_dog, Canis_familiaris}
```

The diagram illustrates the upward hierarchy of a dog. It starts with the most specific classification, {dog, domestic_dog, Canis_familiaris}, and moves up through various taxonomic levels. Two red arrows highlight specific paths: one from {chordate} to {domestic_animal, domesticated_animal}, and another from {canine, canid} to {domestic_animal, domesticated_animal}.

Multilinguality – concepts (synsets are connected via Interlingual Index)

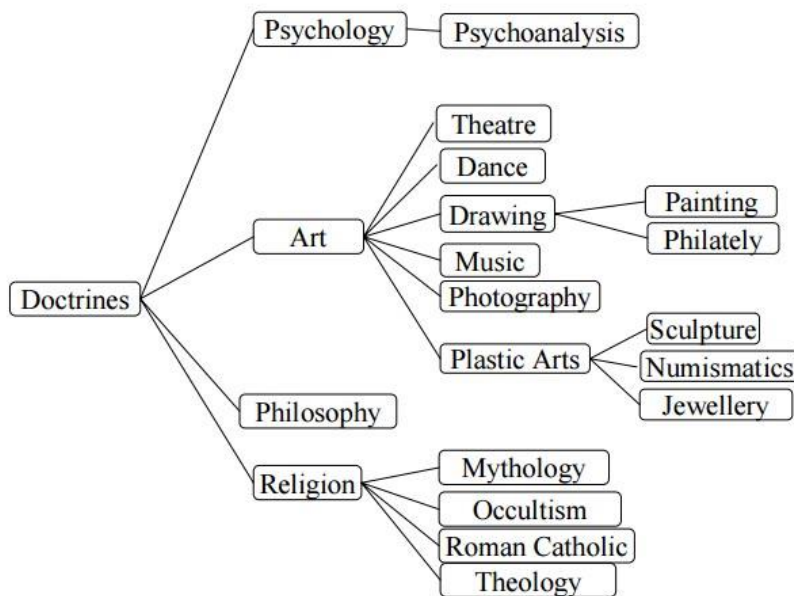


Wordnet enhancement - Domain Hierarchy

- The WordNet Domains Hierarchy (**WDH**) is a language-independent resource composed of 164, hierarchically organized, domain labels (e.g. Architecture, Sport, Medicine).
- WordNet Domains is a lexical resource developed at ITCirst where each WordNet synset is annotated with one or more domain labels selected from a domain hierarchy which was specifically created to this purpose.
- The first version of the WDH was composed of 164 domain labels selected starting from the subject field codes used in current dictionaries, and the subject codes contained in **the Dewey Decimal Classification (DDC)**, a general knowledge organization tool which is the most widely used taxonomy for library organization purposes.
- More info: <http://wndomains.fbk.eu/index.html>

One of the five main trees in the WordNet Domains original hierarchy

Other 4 trees are:
free_time
applied_science
pure_science
social_science



The label FACTOTUM was assigned in case all other labels could not be assigned.

SUMO – The Suggested Upper Merged Ontology (SUMO)

- An ontology is a set of definitions in a formal language for terms describing the world.
- An Upper Ontology is an attempt to capture the most general and reusable terms and definitions.
- **SUMO:**
 - 1000 terms, 4000 axioms (assertions), 750 rules;
 - Mapped by hand to all of WordNet 1.6;
 - then ported to newer versions
 - Associated domain ontologies totaling 20,000 terms and 60,000 axioms;
 - Free
 - SUMO is owned by IEEE but basically public domain
 - Domain ontologies are released under GNU
 - www.ontologyportal.org

Relations between SUMO concepts and Wordnet Synsets

- Synonymy
 - {battle, conflict, fight, engagement} -> SUMO Battle= (Domain: history)
- Subordination
 - {naval_battle} -> SUMO Battle+ (Domain: history)
- Instance
 - {Trafalgar, battle_of_Trafalgar} -> SUMO Battle@ (Domain:history)
- Less straightforward
 - {writer, author} -> SUMO authors= (Domain: literature)
 - {dramatist, playwright} -> SUMO Position+ (Domain: literature)
 - {poet} -> SUMO authors+ (Domain: literature)
 - {Brecht, Bertolt_Brecht} -> SUMO Man@ (Domain:literature)

SentiWordNet

- **SentiWordNet** is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications.
- SentiWordnet is the result of automatically annotating all WORDNET synsets according to their degrees of positivity, negativity, and neutrality.
- Each synset s is associated to three numerical scores $Pos(s)$, $Neg(s)$, and $Obj(s)$ which indicate how positive, negative, and “objective” (i.e., neutral) the terms contained in the synset are.
- Each of the three scores ranges in the interval $[0.0, 1.0]$, and their sum is 1.0 for each synset.

SentiWordNet

- Different senses of the same term may have different opinion-related properties.
- Example for the adjective *estimable* from SentiWordNet 1.0:
- {computable, estimable} def: may be computed or estimated Pos=0, Neg=0, Obj=1.
- {estimable} def: deserving of respect or high regard Pos=0.75, Neg=0.0, Obj=0.25.

Serbian Wordnet - SrpWN

- Developed in the scope of BalkaNet project; today aligned with Princeton Wordnet 3.0
- The size of Serbian WN – 22,000 synsets
- Connected to domains, SUMO, SentiNET
- SWN web tool for developers and users
 - Used by humans, linguists (computer linguists, psycholinguists, cognitive linguists), lexicographers, etc.
- By applications
 - For semantic annotation
 - Word sense disambiguation
 - Improvement of web search
 - Anaphora and co-reference resolution
 - Summarizers
 - Question answering

SWN – a user view

prehrambeni proizvodi ‘foodstuff’

Semantički resursi srpskog jezika WordNet RetFig Niste prijavljeni Prijava

ENG30-07566340-n Traži ☒ sadrži ☐ počinje sa ☐ tačna fraza

☒ Literal ☐ Def ☐ Usage ☐ Domain

Ukupno nađeno: 1 sinset

ID: **ENG30-07566340-n** POS: n BCS: 3 0.000 0.000 Dusk
08.04.2003 Approved: yes PWN XML

Literals: **prehrambeni proizvodi**
Definition: *Supstancija koja se može koristiti kao hrana ili pripremiti da bi se koristila kao hrana.*

▼ - Relations... hypernym-> ENG30-00021265-n, hranljiva materija
hypernym: 2
ENG30-00021265-n hranljiva materija;
ENG30-00019613-n materija; supstanca; supstancija; tvar;

▼ - Relations... hyponym-> ENG30-07755089-n, kakao
▼ - Relations... hyponym-> ENG30-07840804-n, jaje
▼ - Relations... hyponym-> ENG30-07843775-n, mlečni proizvod
▼ - Relations... hyponym-> ENG30-07809096-n, sastojak hrane
▼ - Relations... hyponym-> ENG30-07802417-n, žito
▼ - Relations... hyponym-> ENG30-07569106-n, brašno
▼ - Relations... hyponym-> ENG30-07673397-n, jestivo ulje
SUMO: Food +
DOMAIN: gastronomy

7566340 Search ☐ Word ☒ Sinset ID

☒ Tree View

Number of Nouns: 1

ID {7566340} Sense {{foodstuff, food_product}: a substance that can be used or prepared for use as food} SWN

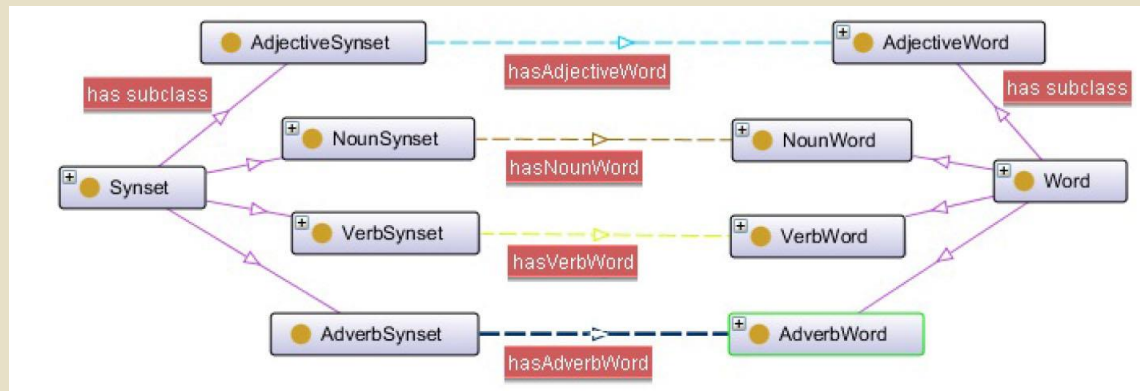
▼ - Relations...

SWN – an application view – sindikalni pokret ‘syndicalism’

◦ <SYNSET>
 <ID>ENG30-08321621-n</ID>
 <POS>n</POS>
 <SYNONYM>
 <LITERAL>sindikalizam<SENSE>1</SENSE><LNOTE></LNOTE></LITERAL>
 <LITERAL>sindikalni pokret<SENSE>1</SENSE><LNOTE>Empty</LNOTE></LITERAL>
 </SYNONYM>
 <DEF>pokret koji se zalaže za zaštitu radničkih prava u odnosima sa poslodavcima i državom</DEF>
 <ILR>ENG30-08472335-n<TYPE>hypernym</TYPE></ILR>
 <ILR>ENG30-09791816-n<TYPE>derived</TYPE></ILR>
 <NL>yes</NL>
 <STAMP>2/15/2017 1:24:00 PM cvetanak</STAMP>
 <SUMO>PoliticalOrganization<TYPE>+</TYPE></SUMO>
 <SENTIMENT>
 <POSITIVE>0.00000</POSITIVE>
 <NEGATIVE>0.00000</NEGATIVE>
 </SENTIMENT>
 <DOMAIN>politics</DOMAIN>
</SYNSET>

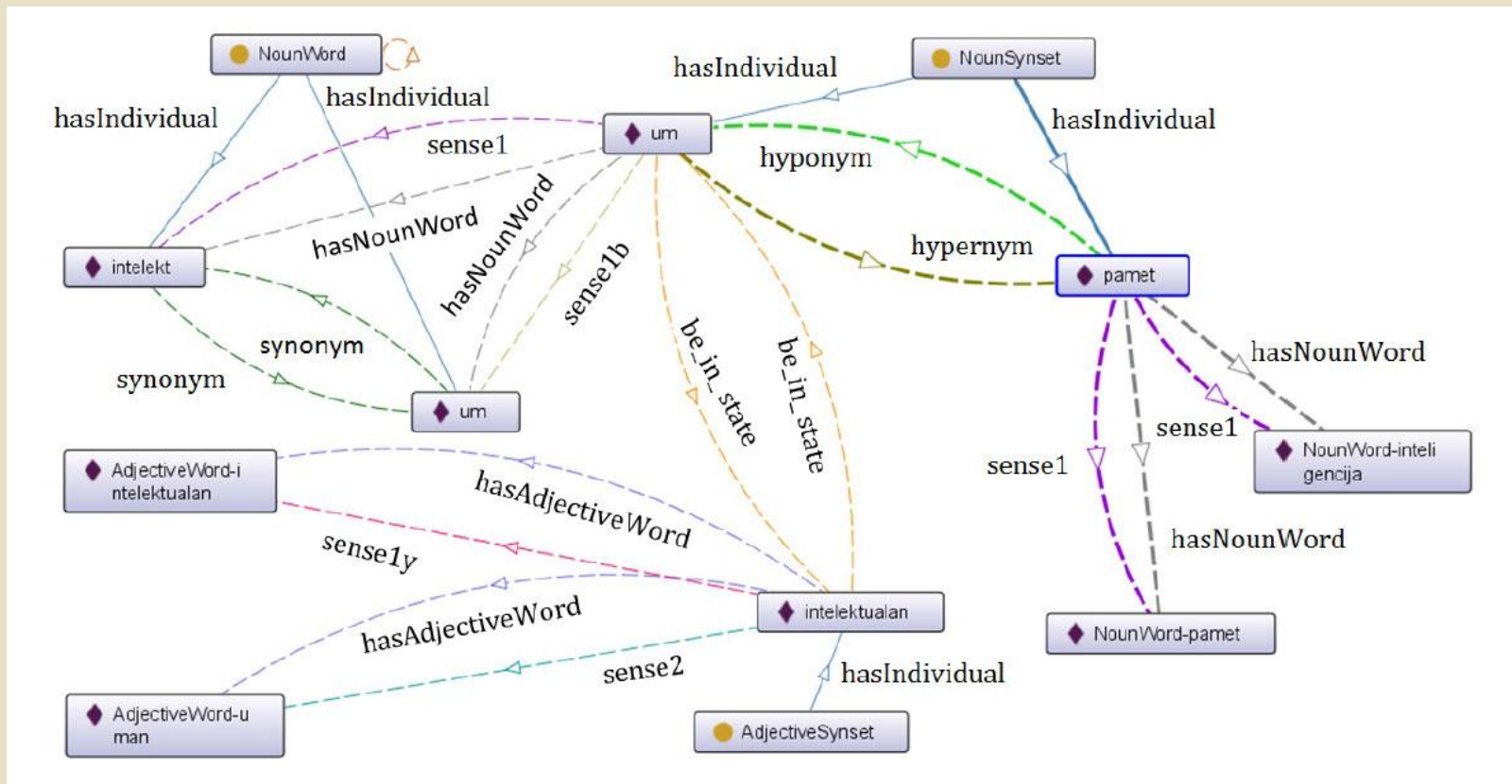
Turning Wordnet into ontologies


- RDF/OWL representation of a BalkaNet WN format
- Two-level taxonomy, with two main classes – Synset class and Word class.
- Specialized ontologies were derived from it: domain and application



- Application of SWN ontology and its derivatives: semantic annotation, sentiment analysis, and text classification

A NounSynset individual in SWN ontology – um ‘mind, intellect’





NLP LANGUAGE RESOURCES FOR INFORMATION RETRIEVAL AND INFORMATION EXTRACTION (IE)



1. Named entity recognition

- **Named entity recognizers** identify proper names in documents, and may also classify these proper names as to whether they designate people, places, companies, organizations, and the like.
- In the sentence:
 - **Italy**'s business world was rocked by the announcement **last Thursday** that **Mr. Verdi** would leave his job as vice-president of **Music Masters of Milan, Inc** to become operations director of **Arthur Andersen**.
- 'Italy' would be identified as a place, 'last Thursday' as a date, 'Verdi' as a person, 'Music Masters of Milan, Inc' and 'Arthur Andersen' as companies.
- Some would consider recognition of 'Milan' as a place, and identifying 'Arthur Andersen' as a person as an error in this context.

NER system for Serbian (SrpNER)

- Entities that are tagged belong to classes:
 - **Person names** (full names and distinguished person names) their titles, roles and functions, if present, preceding or following them;
 - **Geopolitical names** – countries and settlements – **geographic names** – water bodies and oronyms.
 - **Organization names** – including names of political parties.
 - **Number expressions – monetary, measurements, count, percentage**
 - **Time expressions – dates, times of day, periods and frequencies**, absolute and relative

General resources used for the Serbian NER

- A rule-based approach supported by lexical resources;
- Comprehensive morphological e-dictionaries of Serbian in DELA/DELAF format (simple words, Multi-word names), including:
 - general lexica,
 - geographic names,
 - personal names,
 - encyclopedic knowledge (in development).
- Dictionary entries are provided with elaborate semantic markers.
- Use of local grammars to specify the context
 - For rejecting false recognitions
 - For accepting false rejections

Examples of Dictionary Entries

- Geographic names:
 - **Dunav,N+NProp+Top+Hyd** (*Danube* is a proper name, geographic notion, hydronym)
 - **Egejsko more,N+NProp+Top+Hyd** (*The Aegean Sea*)
- Geopolitical names:
 - **Solun,N+NProp+Top+Gr** (*Thessaloniki* – a proper name, city)
 - **Helenska republika,N+NProp+Top+Dr** (*the Hellenic Republic* – a proper name, country)
- Organizations:
 - **Atinska novinska agencija,N+NProp+Org+Acr=ANA** (*The Athens News Agency*)
- Person names:
 - **Venizelos,N+NProp+Hum+Last+Cel** (*Venizelos* – a last name of a famous person)
 - **Riga od Fere,N+NProp+Hum+Last+Cel** (*Rigas Feraios* – a full name of a famous person)

Evaluation results

	OK		NOK		MISS		Token/ Type	Precision		Recall		F-measure	
	Token	Type	Token	Type	Token	Type		Token	Type	Token	Type	Token	Type
TIME-P	8	8	0	0	0	0	1.0	1.00	1.00	1.00	1.00	1.00	1.00
TOP-C	3,768	353	19	8	18	15	10.3	0.99	0.98	1.00	0.96	1.00	0.97
CURR	313	230	3	3	4	4	1.4	0.99	0.99	0.99	0.98	0.99	0.99
DATE-M	506	369	10	10	3	3	1.4	0.98	0.97	0.99	0.99	0.99	0.98
TOP-H	24	12	1	1	0	0	2.0	0.96	0.92	1.00	1.00	0.98	0.96
MEASURE	281	136	4	4	8	8	2.0	0.99	0.97	0.97	0.94	0.98	0.96
DATE-P	104	84	2	2	3	3	1.2	0.98	0.98	0.97	0.97	0.98	0.97
TOP-S	3,660	666	49	30	163	112	4.9	0.99	0.96	0.96	0.86	0.97	0.90
TIME-M	71	66	4	4	1	1	1.1	0.95	0.94	0.99	0.99	0.97	0.96
NAME	3,344	1,539	94	77	532	347	2.1	0.97	0.95	0.86	0.82	0.91	0.88
TOP-W	16	13	0	0	5	3	1.3	1.00	1.00	0.76	0.76	0.88	0.88
TOTAL	12,095	3,476	186	139	737	496	3.2	0.98	0.96	0.94	0.94	0.96	0.96
RS	2,335	1,260	140	68	407	347	1.7	0.94	0.95	0.85	0.85	0.92	0.92

F1-measure
Token=0.96
Type=0.92

SrpNER in use – text de-identification

- Especially important in narrative clinical (medical) texts;
- These texts contain accurate and comprehensive clinical data valuable as a vital resource for secondary uses;
- But they also include many items of patient identifying information – **Protected Health Information**;
- De-identification permits sharing absent explicit content for secondary research;
- Our system is based on NER - automatic recognition of particular phrases in text (persons, organizations, locations, dates, etc.)
- It also has to take care about specifics of clinical narrative texts: fragmented and incomplete utterances, lack of punctuation marks and formatting, many spelling and grammatical errors, domain specific terminology and abbreviations, large number of eponyms and other non-PHI erroneously categorized as PHI

De-iden

All male names → Fred Kremenko 'Fred Flinstone'

All female names → Vilma Kremenko 'Wilma Flinstone'

All Geo-locations → Kamengrad 'Bedrock'

Vaš broj **Posl. Br. Ki 250/08**

Naš broj **33/06**

OPŠTINSKI SUD Istražni sudija G-đa **Rada Anđelić-**

Vašom naredbom zatražili ste od **Komisije lekara veštaka Medicinskog fakulteta Univerziteta u Nišu** sudsko medicinsko veštačenje u predmetu **Ki 250/08** na ... koje je dana **20.03.2008. god.** zadobio oštećeni **Marković Ivan** iz **Dragačeva....**

Numeric data

Person names

Organisation

Geographic names

Vaš broj **<number PHI="yes">XXXX</number>**

Naš **<number PHI="yes">XXXX</number>**

<org PHI="yes">SUD</org> <pers><role>Istražni sudija gospođa</role> **<persName.full**

PHI="yes">Vilma Kremenko</persName.full></pers>

Vašom naredbom zatražili ste od

<org PHI="yes">Komisije</org>

<org PHI="yes">fakulteta</org>

<org PHI="yes">Univerziteta</org> sudsko

medicinsko veštačenje u predmetu

<number PHI="yes">XXXX</number> na ... koje je dana

<date PHI="yes">25.08.2008.</date> zadobio oštećeni

<persName.full PHI="yes">Barni

Kamenko</persName.full> iz **<top.gr**

PHI="yes">Kamengrada</top.gr>....

Performance Measures

PHI	Precision	Recall	F-measure
pers	0.88	0.97	0.92
top	0.99	0.94	0.96
org	0.99	0.93	0.96
num	0.93	0.78	0.85
date	0.98	0.99	0.98
address	0.97	0.85	0.91
Total	0.94	0.94	0.94

For this task Recall is crucial

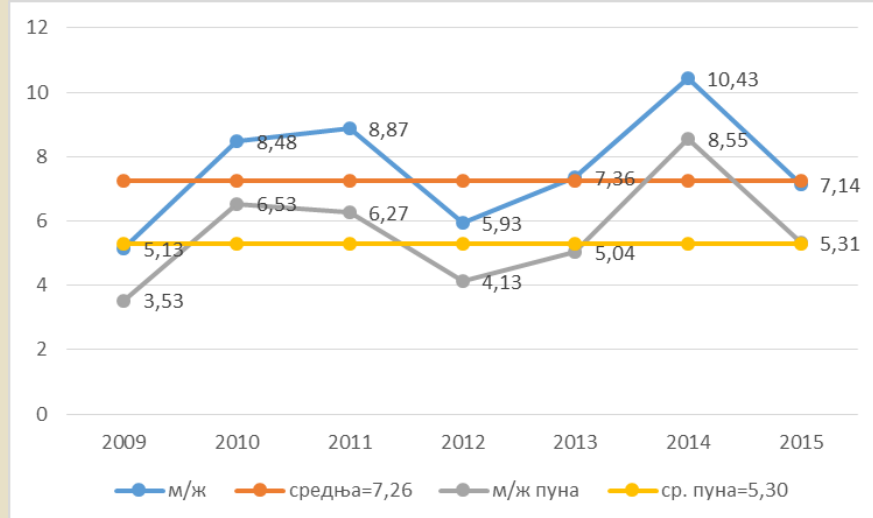
SrpNER in use –representation of women (and men) in newspaper texts

- NER is used to extract personal names from newspaper text;
- Additional information is extracted – which of these names refer to women and which refer to man
- A corpus of selected news from 5 different sources collected during period 2009-2015 was used; in addition a small corpus of articles from one weekly magazine for women.

	Man	Women
5 sources	9825	1352
Woman magazine	424	137

A ratio between names referring to man and women

- Blue line – a ratio between all **M** names and all **F** names
 - Orange line - Average value 7,14 (one **F** name comes for each 7 **M** names)
- Gray line – a ration between all full **M** names (a name and a surname) and all full **F** names
 - Yellow line – Average value 5,31 (one full **F** name comes for each 5-6 **M** names)
- In 2014 women were not interesting for newspapers (one **F** name came for each 10-11 **M** names)
- Man are more often referred to by a surname only (because they are mentioned more often?)



The most frequent functions (roles) of men and women

Man	
Function	Frequency
president	519
minister	261
prime minister	185
director	143
vice-president	131
leader	74
secretary	63
attorney	37
advisor	29
spokesman	23

Women	
Function	Frequency
president	46
representative	35
minister	32
spokesman	24
Vice-president	22
director	21
chancellor	13
secretary	13
depute	12
professor	7
wife	7

2 Full-text search

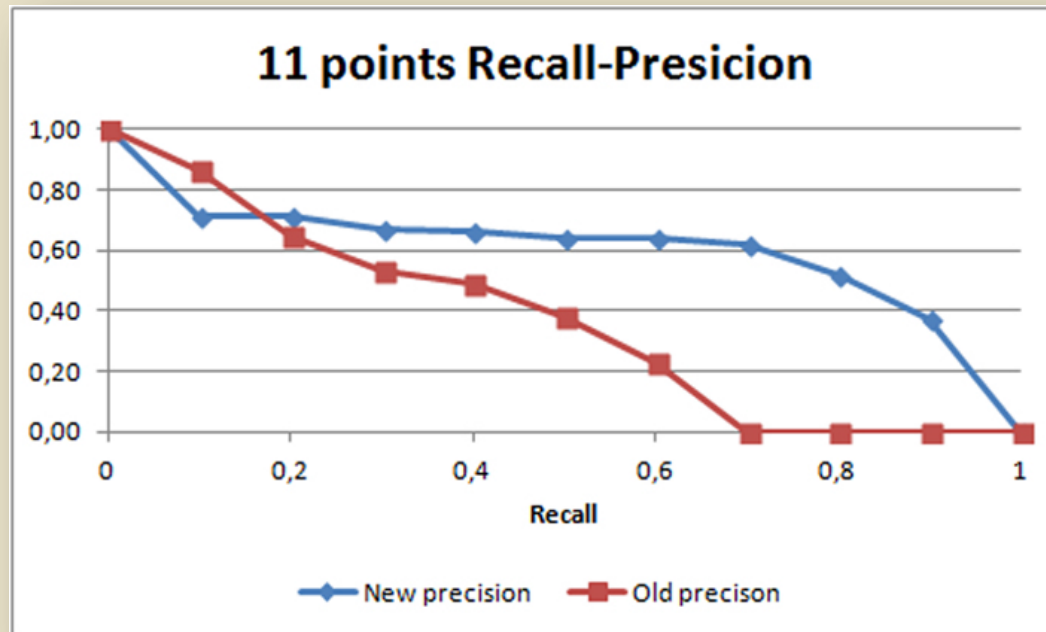
- One of the problems of full text search in Serbian is its rich morphology;
- Keywords are in first person singular, while in texts they take different inflectional forms;
- Normalization of morphological forms for document indexing and query processing is necessary;
- For full-text search we use Morphological e-dictionaries for Serbian (simple and MWUs), and SrpNER for named entity recognition

Application of full-text search: **FoDiB** - geological projects documentation Serbia

- FoDiB - geological projects documentation with structured descriptions of over 4,900 national geological projects from 1956 to the present day;
- Architecture of the new system:
 - **BOW** - representation of the document by a set of ungrammatical words (nouns, adjectives, adverbs and acronyms) followed by their frequencies
 - Text is lemmatized and lemmas (simple and multi-word) are extracted and their frequencies are calculated;
 - NEs are recognized and tagged (approximately 4 NEs per document);
 - In this approach 12,204 simple lemmas (with 450,418 occurrences) and 271 MWUs (with 6,525 occurrences) were extracted;
 - Term weights were associated to all index terms.

Evaluation results

- Comparative graph of the relationship between precision and recall - interpolated average precision for 11 levels of recall 0.0, 0.1, 0.2, ..., 0.9 (for a sample of queries)
- Precision of the old system is significantly better among first-ranked documents
- Recall is better with the new system



Application of full-text search: **Bibliša**, a bilingual digital library

- Bibliša – a digital library of parallel texts (journals, project documentation, etc.)
- Texts are in TMX format; all texts are supported by metadata;
- Search is supported by:
 - For inflection expansion, Morphological e-dictionaries for Serbian (simple and MWUs),
 - For semantic expansion:
 - Serbian Wordnet;
 - Various domain dictionaries and thesauri (librarianship, information sciences, mining, geology);
- Vebran – Web service for query expansion

The aligned collection of Bibliša



Number of documents 509; number of sentences 101670



A query – machine translation

- Automatic query expansion:
 - **Bilingual expansion** – mašinsko prevođenje
 - **Semantic expansion** – automatsko prevođenje
 - **Inflection expansion** – mašinskog prevođenja, mašinskom prevođenju, automatska prevoda,...

The screenshot shows the 'BIBLIŠA: ALIGNED COLLECTION SEARCH TOOL' interface. At the top, there's a navigation bar with links: Home, Metadata browse, Metadata search, Mongo search, Manage data, Help, Tutorial, and About. A user is logged in as 'cvetana' with a 'Log out' link. The main section is titled 'WELCOME TO ALIGNED TEXT COLLECTION SEARCH TOOL!'. It features a search bar with the keyword 'machine translation', a language dropdown set to 'en', and a text collection dropdown set to 'All'. Below this is a table for expanding the query. The table has three columns: 'Synonyms', 'en', and 'sr'. The 'en' column contains the original query 'machine translation, MT' and its expanded forms: 'machine translation', 'machine translation', 'machine translation', 'machine translation', and 'machine translation'. The 'sr' column contains the corresponding translations: 'mašinsko prevođenje', 'automatsko prevođenje, mašinsko prevođenje', and empty cells for the other rows. To the left of the table, there are checkboxes for various expansion methods: WordNet, Dictionary of Librarianship, Biblimir, GeoISSTerm, RudOnto, and Termi. At the bottom, there are checkboxes for 'Include Hypernyms' and 'Include Hyponyms', a 'Preview and modify terms for query' button, a 'Database' selector (MarkLogic, MongoDB), a 'Match query' selector (EN&SR, EN, SR), and a 'Morphological query expansion' checkbox. Finally, there are 'Obtain concordances' and 'New Search' buttons.

Synonyms	en	sr
<input checked="" type="checkbox"/> WordNet...	machine translation, MT	mašinsko prevođenje
<input checked="" type="checkbox"/> Dictionary of Librarianship...	machine translation	automatsko prevođenje, mašinsko prevođenje
<input checked="" type="checkbox"/> Biblimir...		
<input checked="" type="checkbox"/> GeoISSTerm...		
<input checked="" type="checkbox"/> RudOnto...	machine translation	
<input checked="" type="checkbox"/> Termi...	machine translation	

Query results

<p>1.2008.1/2.3 metadata</p>	<p>translation of news texts and parliamentary proceedings have developed to reach a high quality for practical use.</p>	<p>pretrage, automatskog prevođenja novinskih tekstova i skupštinskih zapisnika dostigle su visok kvalitet pogodan za praktičnu upotrebu.</p>
<p>Vitas et al., 2012, ID: 9.2012.1.4 metadata</p>	<p>n263 This corpus was used for tagger training, as well as for experiments in alignment at the word level and in automatic translation.</p>	<p>n263 Ovaj korpus je poslužio za obučavanje tagera i za eksperimente u poravnavanju na nivou reči i u automatskom prevođenju.</p>
<p>Vitas et al., 2012, ID: 9.2012.1.3 metadata</p>	<p>n227 Internet users and providers of Web content can also profit from language technology in less obvious ways, e.g., if it is used to automatically translate Web content from one language into another.</p>	<p>n227 Koristi koje korisnici interneta i dobavljači sadržaja na vebu mogu da imaju od jezičkih tehnologija možda su manje očigledne, na primer, u automatskom prevođenju veb sadržaja sa jednog jezika na drugi.</p>
<p>Vitas et al., 2012, ID: 9.2012.1.5 metadata</p>	<p>n23 In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.</p>	<p>n23 Posebno, ona se fokusira na sprovođenje najsavremenijih istraživanja u automatskom prevođenju, prikupljanju podataka i organizovanju jezičkih resursa za potrebe evaluacije, sastavljanje inventara alata i metoda i organizovanje radionica i obuka za članove zajednice.</p>
<p>Vitas et al., 2012, ID: 9.2012.5 metadata</p>	<p>n23 In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.</p>	<p>n23 Posebno, ona se fokusira na sprovođenje najsavremenijih istraživanja u automatskom prevođenju, prikupljanju podataka i organizovanju jezičkih resursa za potrebe evaluacije, sastavljanje inventara alata i metoda i organizovanje radionica i obuka za članove zajednice.</p>
<p>Karagiozov et al., 2012, vol. XIII:1, ID: 1.2012.1.2 metadata</p>	<p>For the MT engine of the ATLAS system we decided on a hybrid architecture combining EBMT and SMT at word-based level (no syntactic trees will be used).</p>	<p>U slučaju mehanizma za mašinsko prevođenje u sistemu ATLAS, odlučili smo se za hibridnu arhitekturu, u kojoj se na nivou reči kombinuju mašinsko prevođenje zasnovano na primerima i statističko mašinsko prevođenje (neće biti korišćena sintaksička drveća).</p>
<p>Vitas et al., 2012, ID: 9.2012.1.4 metadata</p>	<p>n184 Evaluation campaigns help compare the quality of MT systems, the different approaches and the status of the systems for different language pairs.</p>	<p>n184 Akcije za procenjivanje omogućavaju da se poredе kvalitet sistema za mašinsko prevođenje, različiti pristupi, kao i status sistema za mašinsko prevođenje za različite jezičke parove.</p>
<p>Popović, 2010, vol. XI:2, ID: 1.2010.2.2 metadata</p>	<p>During 1990-1994 period in the DARPA MTEval initiative, Jelinek's machine translation system that was statistically-oriented and with which the language structure had no importance in the beginning, CANDIDE, has never shown better performance neither from the older symbolic-oriented</p>	<p>Tokom perioda 1990-- 1994. u okviru DARPAMTEvalinicijative, Jelinekov sistem za mašinsko prevođenje koji je bio statistički orijentisan i kome jezička struktura nije igrala bitnu ulogu, CANDIDE, nije se pokazao boljim ni od starijeg simboličkog SYSTRAN-a, niti od u tom smislu "opreznog" Babelica-34</p>

Evaluation

initial query
in Serbian

keyword	expansion	type of exp.	No. of sentences	FP
rudnik		Orig	33	0
	rudnika, rudniku, rudnicima,...	Morph	246	0
	površinski kop, okno	Sem	48	0
		SemMorph	331	0
	mine, open pit, surface mine, colliery, pit	SemOtherLng	784	46
		All	837	46
povreda		Orig	153	116
	povrede, povredi, povredom, povredu, povredama	Morph	433	347
	ozleda, šteta, zlo	Sem	190	147
		SemMorph	681	584
	damage, harm, hurt, injury, scathe, trauma	SemOtherLng	630	541
		All	975	134
rotorni bager		Orig	3	0
		Morph	6	0
	rotorni bager, glodar	Sem	3	0
		SemMorph	8	0
	BWE, bucket wheel excavator	SemOtherLng	17	0
		All	19	0
informacioni sistem		Orig	68	3
		Morph	189	4
		Sem	68	3
		SemMorph	189	4
	data system, information system	SemOtherLng	209	3
		All	218	4

Serbian	Average of Recall	Average of Precision
Orig	0.11	0.81
Sem	0.13	0.81
Morph	0.43	0.84
SemMorph	0.50	0.80
SemOtherLn	0.80	0.71
All	1.00	0.83

3 Terminology extraction

- Terminology from various domains consists mainly from Multi-Word Terms (MWT);
- Most of these MWTs belong to few syntactic structures, e.g. in Serbian;
 - The most common structure is a noun (**N**) preceded by an adjective (**A**), which agrees with a noun in gender, number, case and animateness (**AXN** class).
 - The second most common is a noun (**N**) followed by a Noun in the genitive case (**Ng**);
- We have selected 12 most common structures (4 with 2 components, 4 with 3 components, 2 with 4 components, and 2 with 5 components), for them we developed local grammars that rely on SrpMD that:
 - recognize potential terms with high recall and precision;
 - produce lemmas of extracted terms (e.g. forms in the nominative singular) that are given to evaluators, and after evaluation prepared for automatic inclusion in SrpMD.

Motivation for rule-based extraction and lemmatization

◦ MWT candidates are extracted from texts in various inflected forms

◦ Example:

- električna energija
- električne energije
- električnoj energiji
- električnu energiju



električna energija
'electric power'

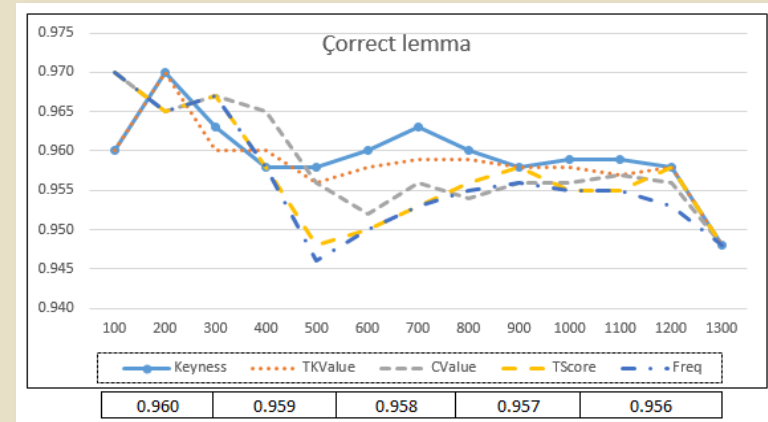
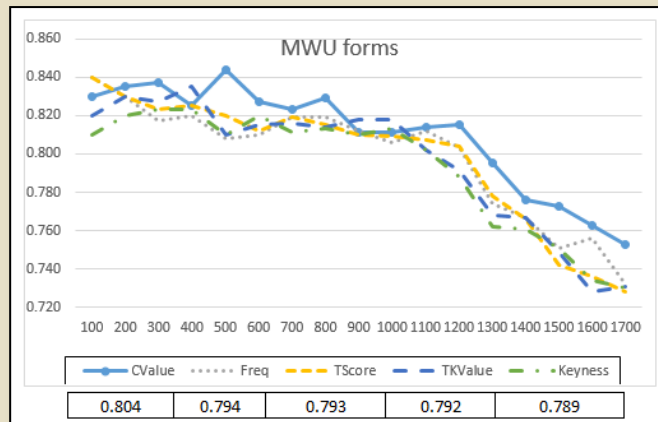
1. Do we want evaluators to evaluate all different inflected forms?
2. Each particular form might not be frequent enough and so not included as a candidate, while all forms belonging to one lemma might have a more significant frequency.
3. Do we want to include all extracted forms into a term base, when this list might not be

DELAC	električna(električni A2:aefs1g) energija(energija N600:fs1q),NC_AYN
DELACF	električnoj energiji,električna energija.N:fs7q
	električne energije,električna energija.N:fp1q
	električnih energija,električna energija.N:fp2q
	električnim energijama,električna energija.N:fp3q

Impact on the whole process

Stat. extractor confused	Stat. extractor satisfied	Stat. extractor satisfied
Evaluator confused	Evaluator confused	Evaluator satisfied
Dictionary prod. confused	Dictionary prod. confused	Dictionary prod. satisfied
magnetskog polja	magnetski polje	magnetsko polje 'magnetic field'
vučne sile	vučni sila	vučna sila 'tractive force'
radnih uslova	radni uslov	radni uslovi 'working conditions'
površinskih voda	površinski voda	površinske vode 'surface water'
Inflected forms	Simple Word Lemma	Multi-Word Lemma

Evaluation results (1)



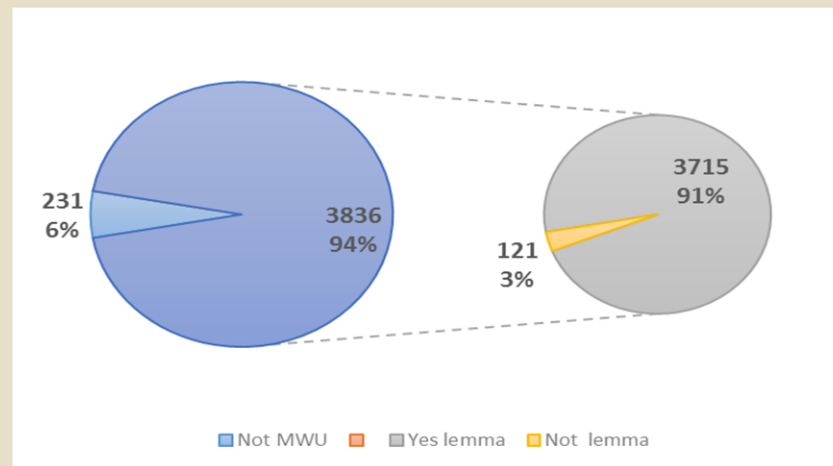
The precision of retrieval was calculated for two tasks for groups of hundreds ranked by measures:

Frequency, C-Value, T-Score, and Keyness:

- checking for each lemma whether they actually represent a MWU, and
- verifying for each proposed lemma whether it is a correct lemma

Mean average precision given at the bottom shows that all measures gave comparable results.

Evaluation results (2)



Out of 4067 distinct forms, **3836 (94%)** were evaluated as proper MWUs and **231 (6%)** were removed as not being proper MWUs.

Among proper MWUs there were **3715 (97%)** with a correct lemma and **121 (3%)** with an incorrect lemma.

4 Sentiment analysis

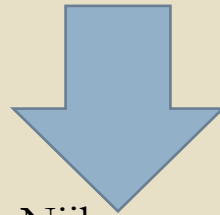
- Emotions in texts:
 - **opinion classification** or **subjectivity identification**, which classifies texts to those that carry opinionated content and those that have factual content;
 - **sentiment classification** or **polarity identification** which classifies opinionated texts to those with positive content and those with negative content;
 - Determining the **strength** or **intensity** of the sentiment polarity
 - by identifying the sentiment of a text as weak or strong positive or negative;
 - by classification into predefined emotion classes;
 - by calculating the sentiment grading.
 - Sentiment classification or polarity identification can be treated in three different ways:
 - Document-level sentiment analysis;
 - Sentence-level sentiment analysis;
 - Attribute-level sentiment analysis.

Sentiment Analysis Framework for Serbian - **SAFOS**

- A hybrid method that uses a sentiment lexicon and Serbian WordNet (SWN) synsets assigned with sentiment polarity scores in the process of feature selection.
 - Sentiment lexicon and lexicon derived from SWN were morphologically expanded using SrpMD.
- SAFOS uses a feature reduction method for document-level sentiment polarity classification using maximum entropy modeling.
- It is based on mapping of a large number of related feature candidates (sentiment words, phrases and their inflectional forms) to a few concepts and using them as features.
- As training and test sets we used websites that already have emotionally classified texts. As there are several news portals in Serbian having information categorized by topics, we took into account those covering one of two topics: **crna hronika** ‘bad news’ and **dobre vesti** ‘good news’.
- Additional test set – film reviews.

Use of sentiment lexicons

- Words and phrases: **ljubavna priča** 'love story', **ljubavna veza** 'love affaire', **ljubav** 'love' are marked as positive.
- Ta **ljubavna priča** započela je letos. Njihovom **ljubavnom vezom** se svi bave. Ta **ljubav** traje i danas. 'That love story has begun last summer. Everybody is dealing with their love affair. That love continues today.'



- Ta **LexiconPOS** započela je letos. Njihovom **LexiconPOS** se svi bave. Ta **LexiconPOS** traje i danas. 'That LexiconPOS has begun last summer. Everybody is dealing with their LexiconPOS. That LexiconPOS continues today.'

Evaluation - 10-fold cross validation metrics (without and with mapping)

Feature set	precision	recall	F1	accuracy
$Cf_1(U)$	0.925	0.944	0.934	0.950
$Cf_2(U + mapping)$	0.937	0.950	0.943	0.956
$Cf_1(U + SWL)$	0.922	0.943	0.932	0.948
$Cf_2(U + SWL + mapping)$	0.934	0.946	0.940	0.954
$Cf_1(U + B)$	0.889	0.914	0.910	0.909
$Cf_2(U + B + mapping)$	0.904	0.919	0.912	0.914
$Cf_1(U + B + SWL)$	0.919	0.943	0.931	0.946
$Cf_2(U + B + SWL + mapping)$	0.929	0.951	0.940	0.950
$Cf_1(U + B + T)$	0.779	0.825	0.800	0.787
$Cf_2(U + B + T + mapping)$	0.783	0.820	0.800	0.788
$Cf_1(U + B + T + SWL)$	0.899	0.921	0.910	0.920
$Cf_2(U + B + T + SWL + mapping)$	0.908	0.924	0.916	0.923

U- unigrams, B- bigrams, T – trigrams, SWL – stop-word list

Evaluation – film news validation metrics (without and with mapping)

Feature set	precision	recall	F1	accuracy
$Cf_1(U)$	0.708	0.821	0.760	0.729
$Cf_2(U + mapping)$	0.727	0.824	0.772	0.746
$Cf_1(U + SWL)$	0.710	0.824	0.763	0.732
$Cf_2(U + SWL + mapping)$	0.730	0.821	0.773	0.747
$Cf_1(U + B)$	0.727	0.880	0.796	0.764
$Cf_2(U + B + mapping)$	0.752	0.887	0.814	0.788
$Cf_1(U + B + SWL)$	0.736	0.855	0.791	0.764
$Cf_2(U + B + SWL + mapping)$	0.767	0.865	0.813	0.792
$Cf_1(U + B + T)$	0.720	0.814	0.764	0.737
$Cf_2(U + B + T + mapping)$	0.745	0.824	0.782	0.760
$Cf_1(U + B + T + SWL)$	0.736	0.841	0.785	0.759
$Cf_2(U + B + T + SWL + mapping)$	0.768	0.850	0.807	0.787

U- unigrams, B- bigrams, T – trigrams, SWL – stop-word list



THANK YOU FOR
YOUR
ATTENTION

