Language Technologies: a happy marriage between linguistics and informatics

Marko Tadić
(marko.tadic@ffzg.hr, http://www.hnk.ffzg.hr/mt)

Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb

ECSS, Paris, 2009-10-09

Open the pod bay door, HAL!

- **Stanley Kubrick, *Space Odyssey 2001*, 1968.**

Is such a conversation possible today?

- **HAL is an artificial agent capable of advanced processing of natural language and showing "intelligent" behaviour**
- **language & speech**
  - **recognition**
  - **generation**
- **understanding**
  - **information retrieval**
  - **information extraction**
  - **"reasoning"**
- **lips reading**
  - **processing of visual paralinguistic signals in face-to-face communication**
- **Was Arthur Clarke too optimistic with year 2001?**

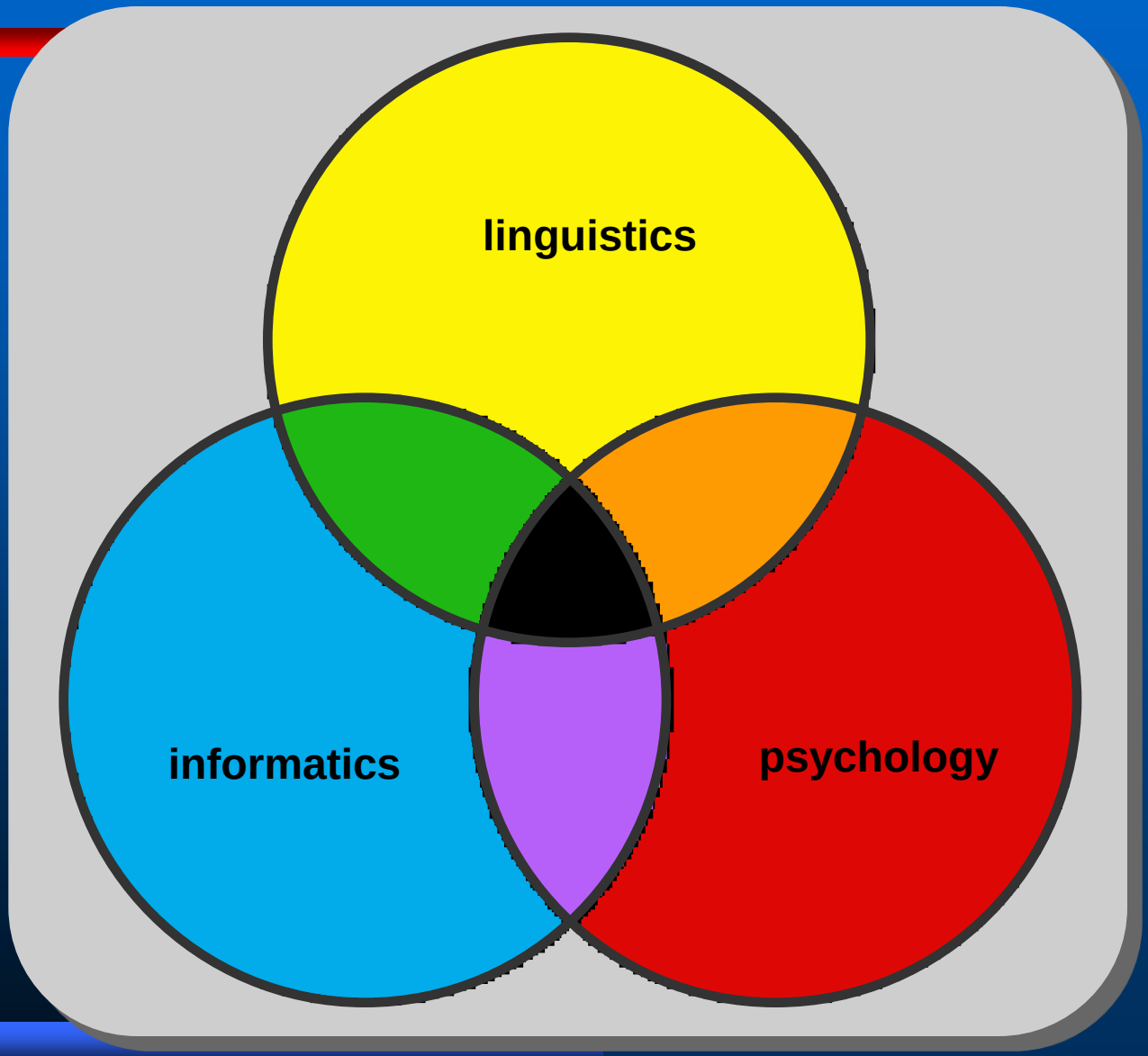
Intro 1: computational linguistics

- **term:**
  language + computer = computational treatment of natural language
  - **linguistics = pivot science**
- **computer: in many sciences today indispensible tool (physics, (bio-)chemistry, economy, traffic...)**
  - **collecting primary data (= empirical approach)**
  - **formation of secondary data and theories (= models)**
- **computational treatment of natural language interesting to**
  - **linguists**
  - **information scientists**
  - **cognitive scientists...**

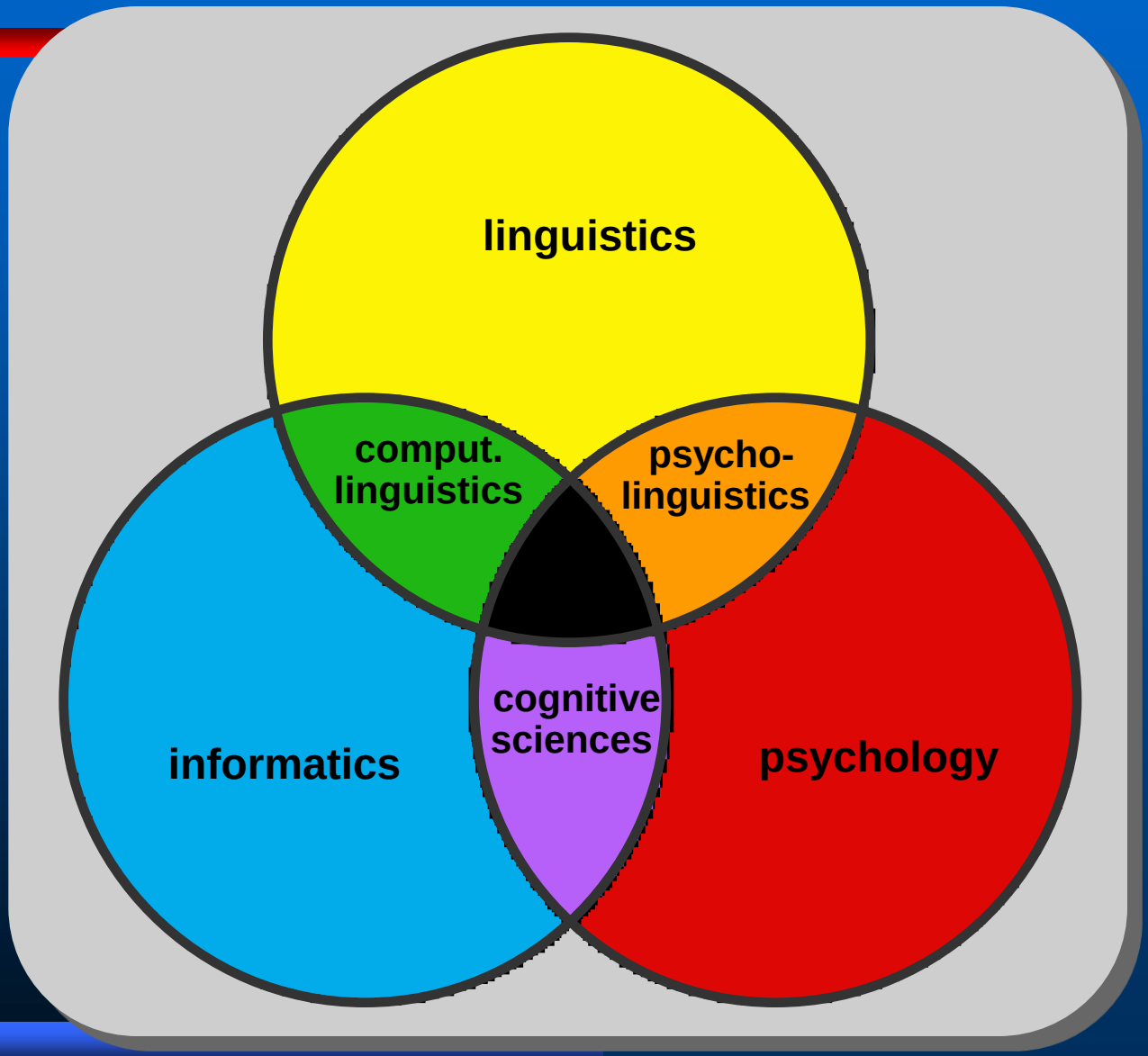
Intro 2: natural language processing

- **term 2:**
  computer + language = computational treatment of natural language
  - **informatics = pivot science**
- **difference:**
  - **linguists: *computational linguistics (CL)***
    - **computers used in linguistic description (of models of sub-systems in a certain language)**
    - **aim: high quality in description of linguistic facts**
  - **informaticians: *natural language processing (NLP)***
    - **computers used in processing of natural language data**
    - **special type of text processing (text = realisation of linguistic system)**
    - **aim: to process in an efficient manner the largest amount of data with the smallest usage of computational resources**

What is computational linguistics 1?



What is computational linguistics 2?



What is computational linguistics 3?

- linguistic discipline that corresponds with
  - information sciences
  - computing
  - psychology, i.e., cognitive sciences
- aim: description of natural language phenomena with the help of computers
- necessary conditions for CL, i.e., its research methods
  - data about language
  - programmes (tools) which are used for
    - collecting that data
    - processing that data
  - development of theoretical models of language (sub-)systems
  - development of systems that verify the models on real language

Basics of CL: two approaches

- **two fundamental approaches in CL**
- **1) theoretical CL**
  - **deals with formal theories of human knowledge necessary for language generation and understanding**
  - **cooperates with cognitive psychology, artificial intelligence, computing, mathematics, etc.**
  - **contributes to the overall knowledge of general linguistics with new findings about complexity of phenomena at particular language levels, e.g.**
    - **syntactic formalisms: HPSG, LFG…**
    - **morphological formalisms: Two-level morphology**
    - **…**

Basics of CL: two approaches 2

- **2) applied CL**
  - **deals with development and realisation of computational models of human language usage**
  - **builds the technologies that rely on theoretical CL findings**
    - **language technologies (LT)**
    - **older term: language engineering (LE)**
  - **contributes with linguistic knowledge in**
    - **human-computer communication: speech/listening and/or writing/reading interfaces**
    - **human-human communication mediated by computer:**
      - **machine translation systems (written/spoken)**
      - **document retrieval**
      - **automatic indexing**
      - **document summarisation**
      - **information extraction**
      - **spelling/grammar/style checking…**

Language Technologies 1

- linguistics = unique between humanities
  - research methods are like ones in natural sciences (empiricism)
  - usage of scientific knowledge for making products
  - a whole range of commercial products based on linguistic knowledge
- technology = “a set of methods and procedures for processing raw materials into final products” (Croatian General Lexicon, Lexicographic Institute, Zagreb, 1996)
- what is raw material, and what is a final product in LT?
  - raw material = data about language
  - final products = systems that enable the user to use his/her own natural language eas(il)y in digital environment
- LT build upon IT like CT also build on IT (ICT)
- without developed IT, LT would not be possible

Language technologies 2

- **defined in EU Framework Programme 5**
  - **predecessors (in FP3 and FP4): L. industry and L. engineering**
- **the largest individual research area in FP5:**
  - **IST = Information Society Technologies (26.3% of the whole FP5 budget = 3,900 M€)**
- **key action III of IST:**
  - **MC&T = Multimedia Content & Tools (564 M€)**
- **the largest part of MC&T:**
  - **HLT = Human Language Technologies**
    - **include also speech processing**
    - **deceased portal: HLTcentral (www.hltcentral.org)**
- **continuation in FP6: eContent**
- **in FP7: also in Research Infrastructures (RI)**

Division of LT 1

- language resources
  - corpora
  - dictionaries
- language tools
  - morphology
    - generators vs. analysers
    - POS/MSD taggers, lemmatisers
  - syntax
    - shallow/deep/robust parsers vs. generators
    - phrases detection: chunkers (NP, VP, multi-word units,…)
    - named entity recognition and classification
  - semantics
    - lexical meaning detection (synonymy/antonymy, WSD…)
    - sentence meaning detection (semantic roles: agent/patient/means…)
  - machine (aided) translation
  - computer aided language learning
  - dialog systems (Q&A…)

Division of LT 2

- **final products**
  - **checkers**
    - **spelling**
    - **grammar**
    - **style**
  - **e-dictionaries**
    - **thesauri**
    - **lexical bases (general/specialised dictionaries)**
  - **automatic indexing**
  - **document summarisation**
  - ***text-to-speech* and *speech-to-text* systems**
  - **systems for machine (aided) translation**
    - **translation memories (= parallel corpora)**
    - **limited MT (controlled languages)**
    - **simple MT (basic information detection)**
    - **HQFAMT (?, Systan), SMT (Google Translate)**
  - **systems for computer aided language learning**

Development of LT for a language 1

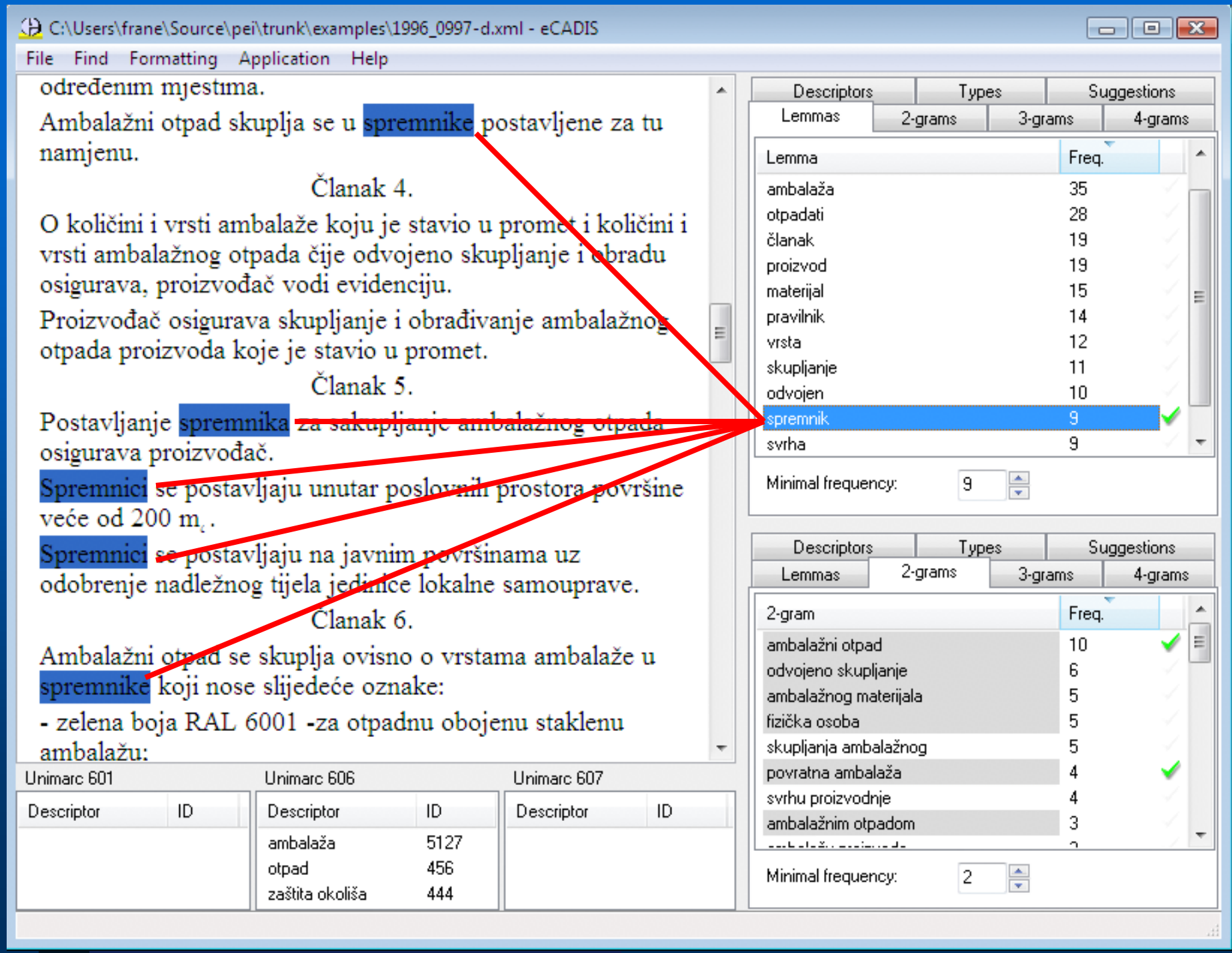
- resources and tools
  - language specific
  - development starts from the fundamental language data
- resources
  - supply the basic language data for development of
    - other resources (e.g. dictionary from a corpus)
    - language tools (e.g. spelling checker from a dictionary)
- development of LT for a language should be
  - planned
    - too expensive to be left to curiosity-driven research
    - BLARK (Basic Language Resources and tools Kit) & ELARK
  - heavily financially supported
    - industry: in linguistic communities with many speakers
    - (state) institutions: in communities with less speakers

LT helping information sciences

- **after being developed on the shoulders of IT and information sciences, LT can pay its tribute back**
  - **providing new solutions for old tasks**
  - **facilitating new tasks**
- **e.g.**
  - **document retrieval**
    - **search engines**
  - **information extraction (text-mining)**
    - **NERC**
  - **…**

Search engines

- web search engines: mostly tailored for English
- what about other languages with other structures?
  - words appearing in many word-forms (WF)
  - e.g. Croatian word “spremnik” (‘container’)
    - Nsg: spremnik Npl: spremnici
    - Gsg: spremnika Gpl: spremnika
    - Dsg: spremniku Dpl: spremnicima
    - Asg: spremnik Apl: spremnike
    - Vsg: spremniče Vpl: spremnici
    - Lsg: spremniku Lpl: spremnicima
    - Isg: spremnikom Ipl: spremnicima

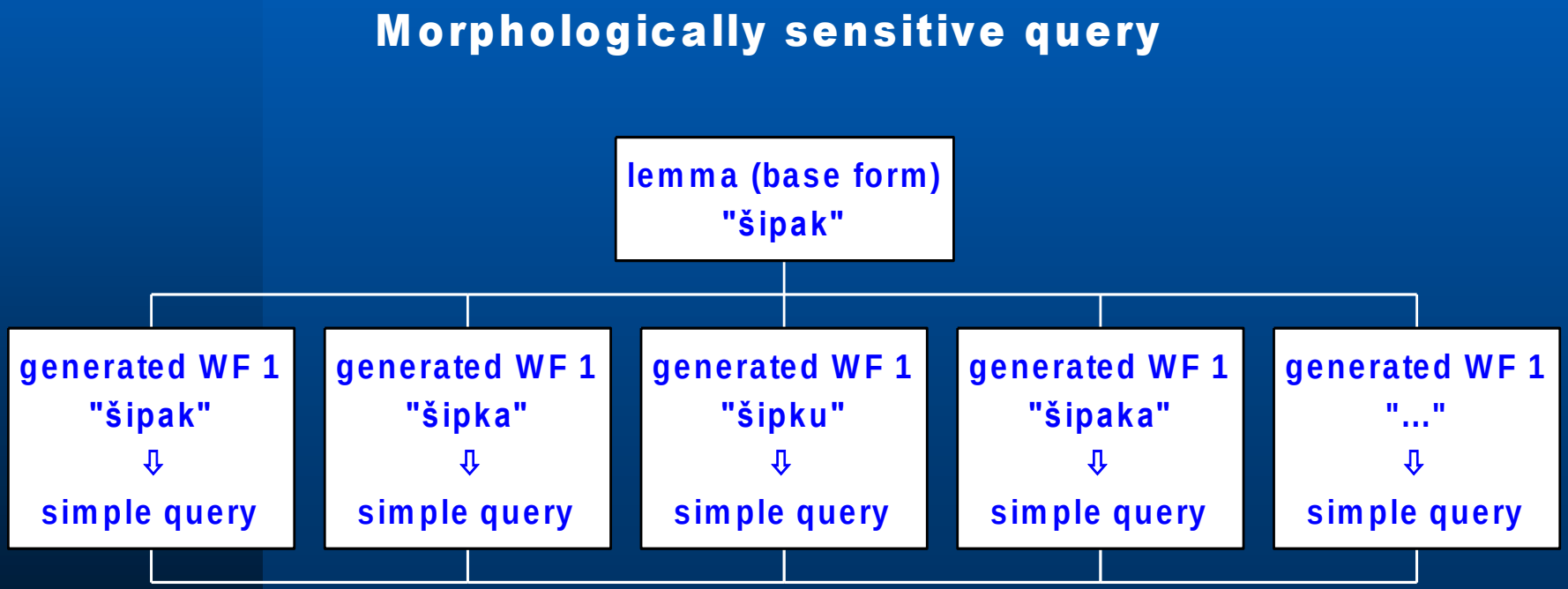

C:\Users\frane\Source\pei\trunk\examples\1996_0997-d.xml - eCADIS
File Find Formatting Application Help
određenim mjestima.
Ambalažni otpad skuplja se u spremnike postavljene za tu namjenu.
Članak 4.
O količini i vrsti ambalaže koju je stavio u promet i količini i vrsti ambalažnog otpada čije odvojeno skupljanje i obradu osigurava, proizvođač vodi evidenciju.
Proizvođač osigurava skupljanje i obrađivanje ambalažnog otpada proizvoda koje je stavio u promet.
Članak 5.
Postavljanje spremnika za sakupljanje ambalažnog otpada osigurava proizvođač.
Spremnici se postavljaju unutar poslovnih prostora površine veće od 200 m².
Spremnici se postavljaju na javnim površinama uz odobrenje nadležnog tijela jedinice lokalne samouprave.
Članak 6.
Ambalažni otpad se skuplja ovisno o vrstama ambalaže u spremnike koji nose slijedeće oznake:
- zelena boja RAL 6001 -za otpadnu obojenu staklenu ambalažu:
Descriptors Types Suggestions
Lemmas 2-grams 3-grams 4-grams
Lemma Freq.
ambalaža 35
otpadati 28
članak 19
proizvod 19
materijal 15
pravilnik 14
vrsta 12
skupljanje 11
odvojen 10
spremnik 9
svrha 9
Minimal frequency: 9
Descriptors Types Suggestions
Lemmas 2-grams 3-grams 4-grams
2-gram Freq.
ambalažni otpad 10
odvojeno skupljanje 6
ambalažnog materijala 5
fizička osoba 5
skupljanja ambalažnog 5
povratna ambalaža 4
svrhu proizvodnje 4
ambalažnim otpadom 3
Minimal frequency: 2
Unimarc 601
Descriptor ID
Unimarc 606
Descriptor ID
ambalaža 5127
otpad 456
zaštita okoliša 444
Unimarc 607
Descriptor ID

Search engines

- web search engines: mostly tailored for English
- what about other languages with other structures?
  - words appearing in many word-forms (WF)
  - e.g. Croatian word “spremnik” (‘container’)
    - Nsg: spremnik Npl: spremnici
    - Gsg: spremnika Gpl: spremnika
    - Dsg: spremniku Dpl: spremnicima
    - Asg: spremnik Apl: spremnike
    - Vsg: spremniče Vpl: spremnici
    - Lsg: spremniku Lpl: spremnicima
    - Isg: spremnikom Ipl: spremnicima
- google.hr or google.fi search: users intutively input Nsg
  - you miss all documents where your word appeared in other WFs
  - G and A more frequent than N in Croatian

Search engines 2

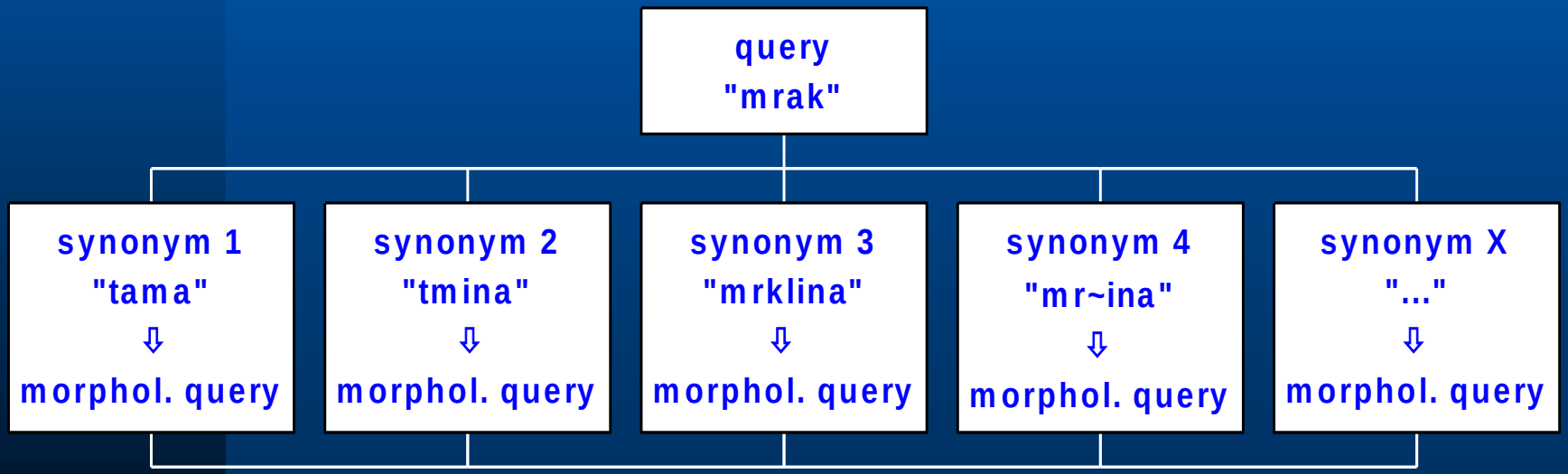
- **CL helps search engines**
  - ***document retrieval meets language technologies…***



Search engines 3

- **what do we really search for using search engines?**
  - **exact words (matching phrases)?**
  - **concepts (regardless of their exact wording)**

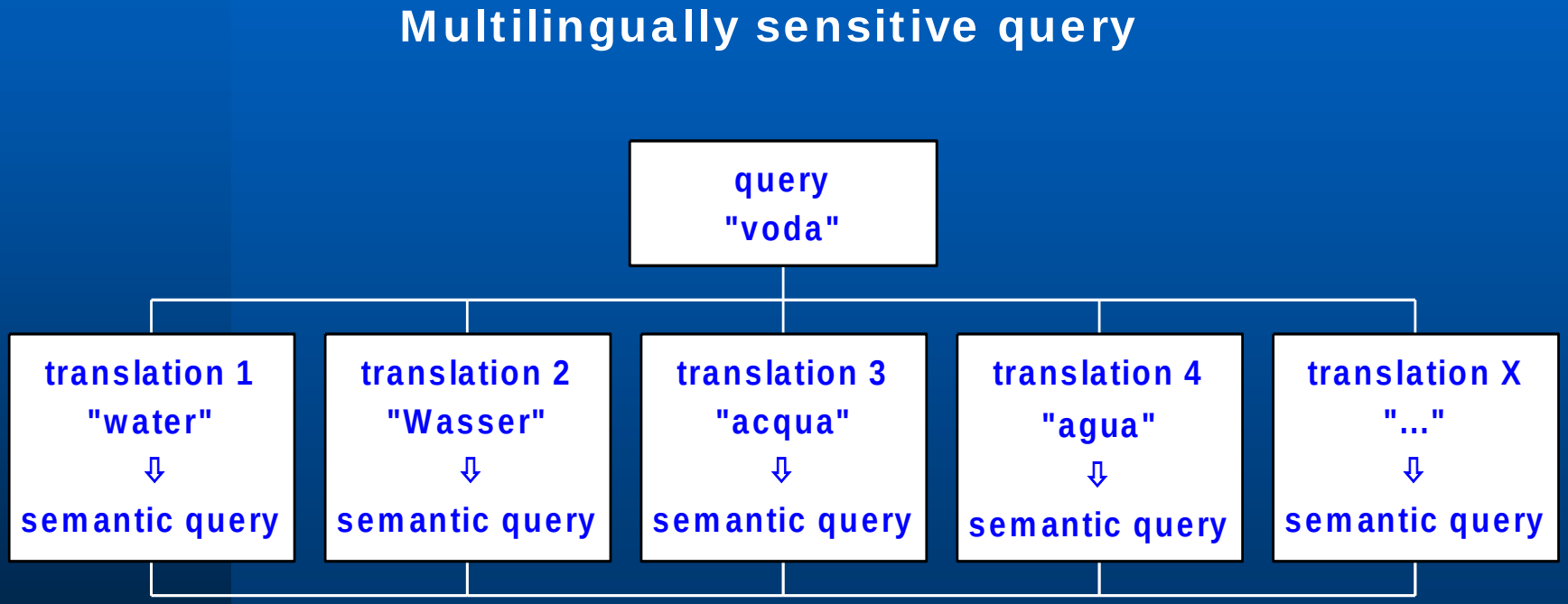
Semantically sensitive query



- **semantic networks (thesauri, WordNets, ontologies)**

Search engines 4

- **cross-linguistic querying**



- **interlingually connected wordnets (WordNet Grid)**

Document retrieval

- paradoxically: until recently the usage of linguistic knowledge (i.e., LT) in document retrieval was minimal
  - primary methods were statistical (TF/IDF…)
- today
  - robust statistical methods have reached its peek
  - knowledge about the language of the document is needed
- methods
  - linguistic pre-processing of documents
    - traditionally: dropping stop-words
    - lemmatisation or normalisation (stemming, truncating)
    - collocation detection (multi-word units in place of individual words)
  - “bag of words” replaced by structured document approach
    - retrieval sensitive to a document structure (INEX conferences)

Document retrieval 2

- **vector-space models**
  - **document collection = matrix**

	a	abonman	acidoza	adlatus	adaptacija	adorirati	aeroban	afinitet	...
doc1	15	0	0	0	0	0	0	0	
doc2	23	0	0	0	0	0	0	0	
doc3	9	0	4	0	2	0	1	0	
doc4	34	1	0	0	0	0	0	2	
...									

  - **serious problem = dimensions of matrices (e.g. 0.8 mil. x 1.3 mil.)**
  - **dimensionality reduction (e.g. for Latent Semantic Indexing…)**
- **lemmatisation**
  - **boosts statistical processing, i.e., accumulates frequencies**
  - **helps with the notorious data sparsness problem**
- **collocations**
  - **detecting MWU that express single concepts (e.g. *real estate*)**
  - **chunkers and shallow parsers needed**

Information extraction

- automatical recognition of
  - selected types of entities (named entities, events…)
  - their relations in free text
- contrary to terms used in informatics for textual documents
  - non-structured
  - semi-structured documents
- linguistic level
  - highly structured
  - carrying a lot of information

NERC

- named entity recognition and classification
- introduced by DARPA as a part of message understanding process
- competition at MUC6 (1996) and MUC7 (1998) conference
- 7 basic types of NEs
  - person
  - organisation
  - location
  - date
  - time
  - currency (+ measures)
  - percentage
- NEs carrying valuable information about the world beyond the document
  - who?, where?, when?, how much?

NERC 2

- **NERC looks simple**
  - **use a gazeteer and match it with the text**
  - **morphology?: NEs behave by the general rules of a language**
- **performance**
  - **humans: 98-99%**
  - **best systems: 94%**
- **identification of NEs**
  - **less problematic**
- **classification**
  - **complex (ambiguities: “Boston plays against Detroit”)**
  - **co-textual information important**
    - **strategies: inner and outer evidence, longest match, one meaning per discourse…**

Home

Aktualnosti
Vijesti
Crna kronika
Sport
Scena
Kultura
Gospodarstvo
Poslovni svijet
Zanimljivosti
Regije

Free Time
Kompas
Događanja
Kino
TV Vodič
Vrijeme
Lifestyle
Nedjeljni Večernji

NOVO
Vijesti bez slika
Moja karijera
e-Shop

→ Gospodarstvo → Poslovni svijet

25.01.2005 18:45

Hrvatski izvoz još na niskim razinama

90 posto tvrtki uopće ne izvozi!

Autor **Piše Josip Bohutinski**

Hrvatski izvoz napokon je prošle godine počeo rasti brže od uvoza te je, prema podacima za prvih 11 mjeseci 2004. godine, izvoz u kunama rastao 15,7 posto a uvoz 5,7 posto. Iz Hrvatske je izvezeno robe u vrijednosti nešto manjoj od 44 milijardi kuna ili 7,25 milijardi američkih dolara, dok je vrijednost uvoza bila 91,19 milijardi kuna ili više od 15 milijardi dolara.

No podaci o izvozu po glavi stanovnika upozoravaju da je hrvatski izvoz još na niskim razinama u usporedbi s drugim i sličnim zemljama. Prema podacima udruge Hrvatski izvoznici, u 2003. godini vrijednost hrvatskog izvoza po glavi stanovnika bila je samo 1106 dolara.

Koliko je je to mala vrijednost, govori podatak o slovenskom izvozu po glavi stanovnika od čak 4774 dolara. Irska na svakog svoga stanovnika izveze 22.119 dolara roba i usluga. Amerikanci, pak, po glavi stanovnika izvezu robe u vrijednosti 2360

Ostale vijesti
- Novi igrači zajedno protiv T-HT-a?
- Hrvatska ipak nije prezadužena
- Povjerenstva odvaguju veletrgovine
- Državne banke kreću u investicijsko bankarstvo
- Dubrovnik ne bi smio dominirati u promidžbi
- Fokus na Sirelu i Somboled
- Dioničari ne žele novac nego banku

Prijava
Korisničko ime
Lozinka
Prijava
→ Registrirajte se!
→ Zaboravili ste lozinku?

Kolumna

→ Nakon revolucije, slijedi evolucija

→ Devizno tržište
→ Tržište novca
→ Tržište obveznica
→ Tržište kapitala
→ Karijere

Večernji list, 2005-02-17, Gospodarstvo

Hrvatski izvoz još na niskim razinama

90 posto tvrtki uopće ne izvozi!

Autor Piše Josip Bohutinski

Hrvatski izvoz napokon je prošle godine počeo rasti brže od uvoza te je, prema podacima za prvih 11 mjeseci 2004. godine, izvoz u kunama rastao 15,7 posto a uvoz 5,7 posto. Iz Hrvatske je izvezeno robe u vrijednosti nešto manjoj od 44 milijardi kuna ili 7,25 milijardi američkih dolara, dok je vrijednost uvoza bila 91,19 milijardi kuna ili više od 15 milijardi dolara.

No podaci o izvozu po glavi stanovnika upozoravaju da je hrvatski izvoz još na niskim razinama u usporedbi s drugim i sličnim zemljama. Prema podacima udruge Hrvatski izvoznici, u 2003. godini vrijednost hrvatskog izvoza po glavi stanovnika bila je samo 1106 dolara.

Koliko je je to mala vrijednost, govori podatak o slovenskom izvozu po glavi stanovnika od čak 4774 dolara. Irska na svakog svoga stanovnika izveze 22.119 dolara roba i usluga. Amerikanci, pak, po glavi stanovnika izvezu robe u vrijednosti 2360 dolara.

No vrijednost izvoza velikih zemalja po glavi stanovnika u pravilu je manja od izvoza malih zemalja zbog velikog domaćeg tržišta koje može apsorbirati veliki dio domaće prozivodnje. To potvrđuju i podaci o izvozu po stanovniku i "malih zemalja" poput Belgije, Nizozemske i Finske.

Uz malu vrijednost izvoza po glavi stanovnika, za Hrvatsku je nepovoljan i podatak o broju domaćih tvrtki čija godišnja vrijednost izvoza premašuje milijun kuna.

Njih je samo pet posto od ukupno aktivnih poduzeća. Naime, prema podacima Hrvatskih izvoznika, od 70-ak tisuća aktivnih kompanija u Hrvatskoj, svoje proizvode i usluge na strana tržišta izvozi samo njih 6700. Pritom je izvoznika čija vrijednost izvoza premašuje milijun kuna samo 3144. Ta grupa izvoznika, prema podacima udruge Hrvatski izvoznici, ostvaruje čak 96 posto ukupnog hrvatskog izvoza.

Koliko je bitna uloga izvoznika u cjelokupnom hrvatkom gospodarstvu, potvrđuje podatak da 2688 izvoznika izdvaja 83 posto ukupne dobiti u Hrvatskoj, odnosno 16,6 od 19,9 milijardi dolara.

Upozoravajući na podatke o hrvatskom izvozu po glavi stanovnika, predsjednik Hrvatskih izvoznika Darinko Bago, prilikom prošlotjednog potpisivanja Sporazuma o suradnji s Hrvatskom bankom za obnovu i razvitak, najavio je sklapanje sličnih sporazuma s drugim udruženjima i institucijama koje mogu pridonijeti afirmaciji hrvatskog izvoza, bez kojeg, naglasio je Bago, Hrvatska nema budućnosti.

A velike zasluge za prošlogodišnji brži rast hrvatskog izvoza sigurno ima upravo HBOR i njegovi programi poticanja izvoza. Preko programa Kreditiranje priprema roba za izvoz i izvoza roba lani je odobreno 170 kredita u vrijednosti 1,25 milijardi kuna, što je čak 448 posto veći iznos nego 2003. godine kada su odobrena 52 kredita, ukupno vrijedna nešto više od 279 milijuna kuna.

I Program osiguranja izvoza zabilježio je lani veliki rast. U 2004. godini osiguran je promet od 580 milijuna kuna, što je povećanje180 posto prema prethodnoj godini, a odobreno je 357 zahtjeva, što je povećanje od 306 posto. Lani je HBOR osigurao izvoz 67 izvoznika, za razliku od 35 u 2003. godini. Od početka poslovanja HBOR je dosad isplatio 12 odšteta u iznosu 3,2 milijuna kuna, a od toga je lani četvero izvoznika dobilo odštetu od 538.000 kuna.

Predsjednik Uprave HBOR-a Anton Kovačev, potpisujući sporazum s Hrvatskim izvoznicima, rekao je da je 2004. bila godina izvoza za njegovu banku te da se nada da će ova biti izvozna za cijelu Hrvatsku, čemu bi trebao pridonijeti i sporazum o suradnji HBOR-a i HIZ-a.

Kovačev je upozorio i da rast hrvatskog izvoza lani nije isključivo rezultat brodogradnje.

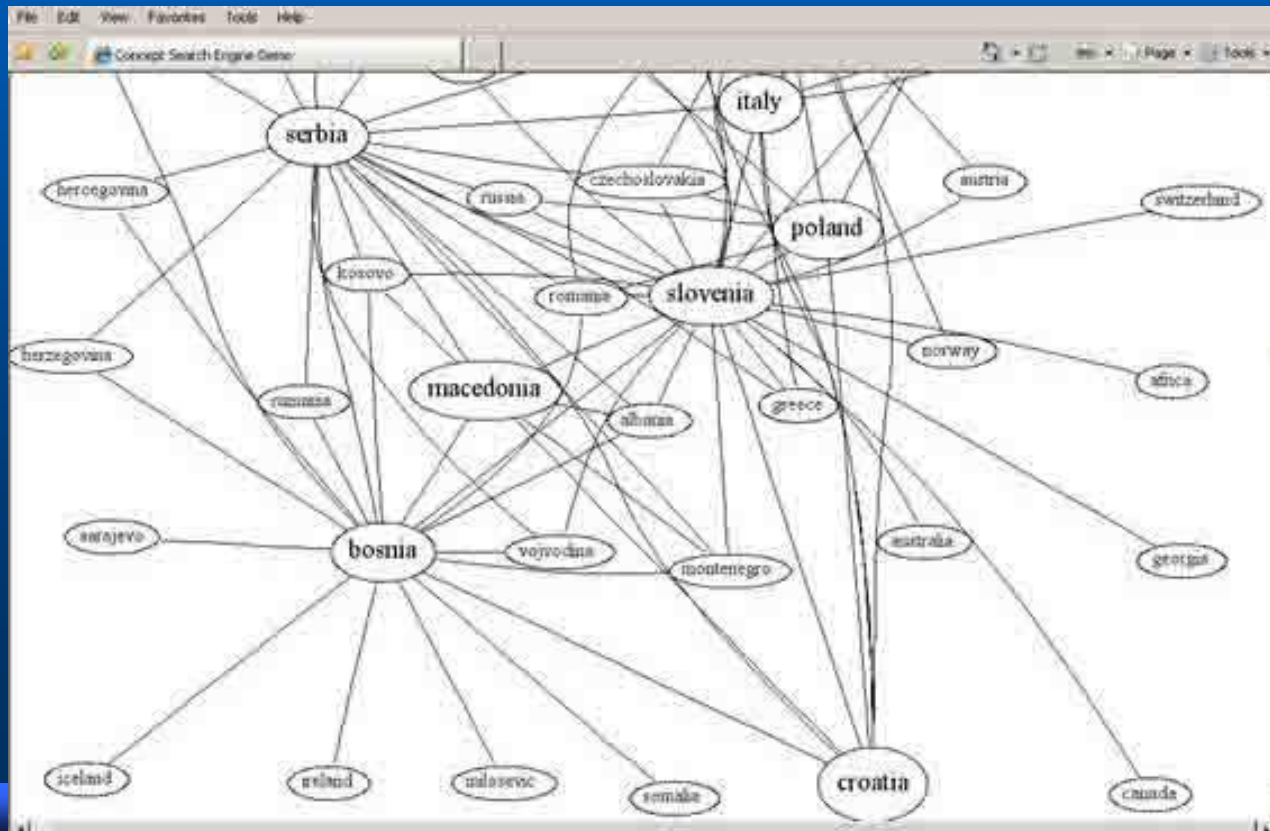
- Oko 90 posto kredita koje smo dali za pripremu roba za izvoz i izvoz roba odnosi se na prerađivačku industriju, poput prehrambene, metalske, farmaceutske i drvne industrije. A te industrije su ostvarile porast izvoza 6,5 posto, što je veći rast od prosječnog ukupnog rasta od 15,7 posto - rekao je Kovačev.

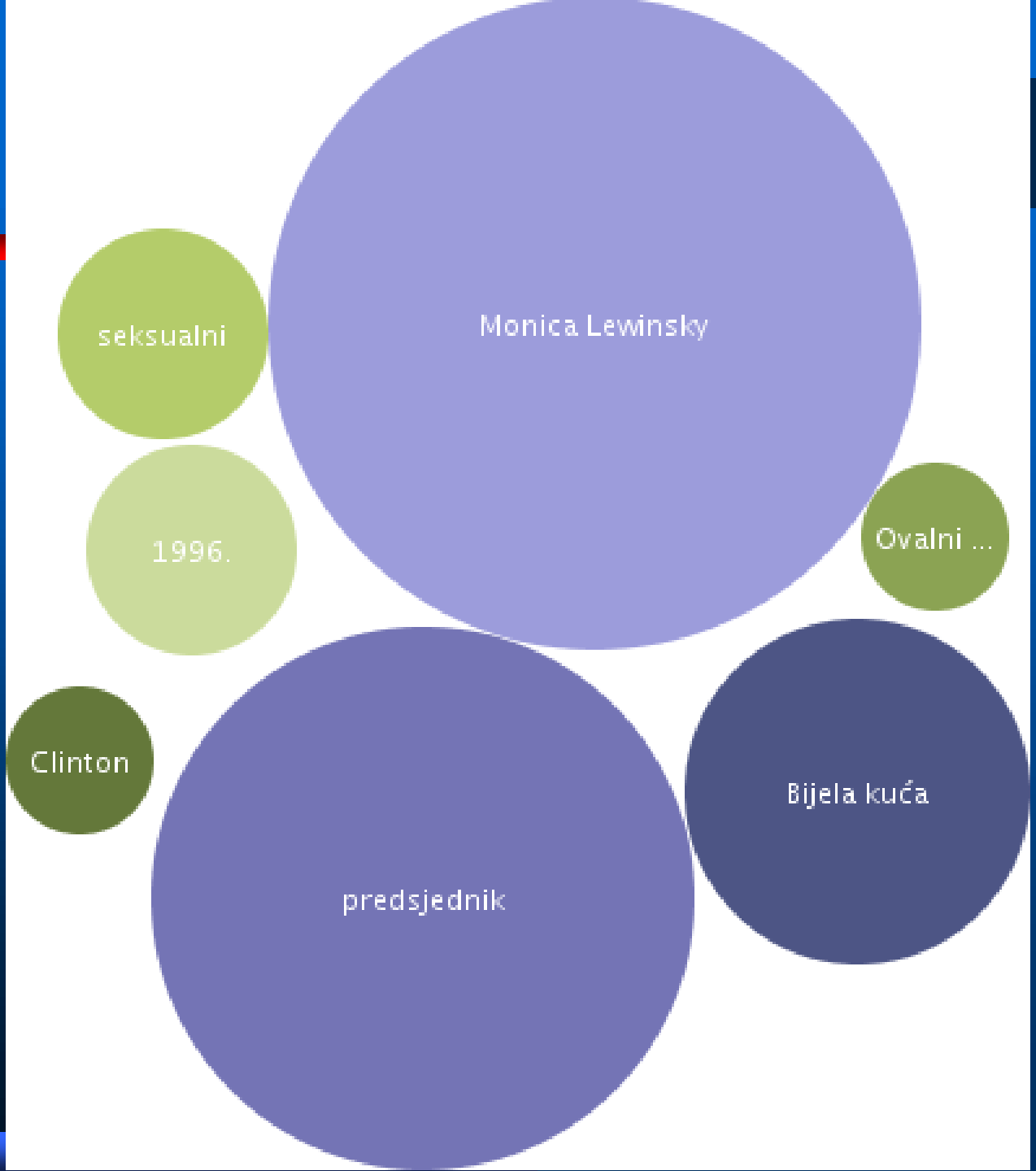
.

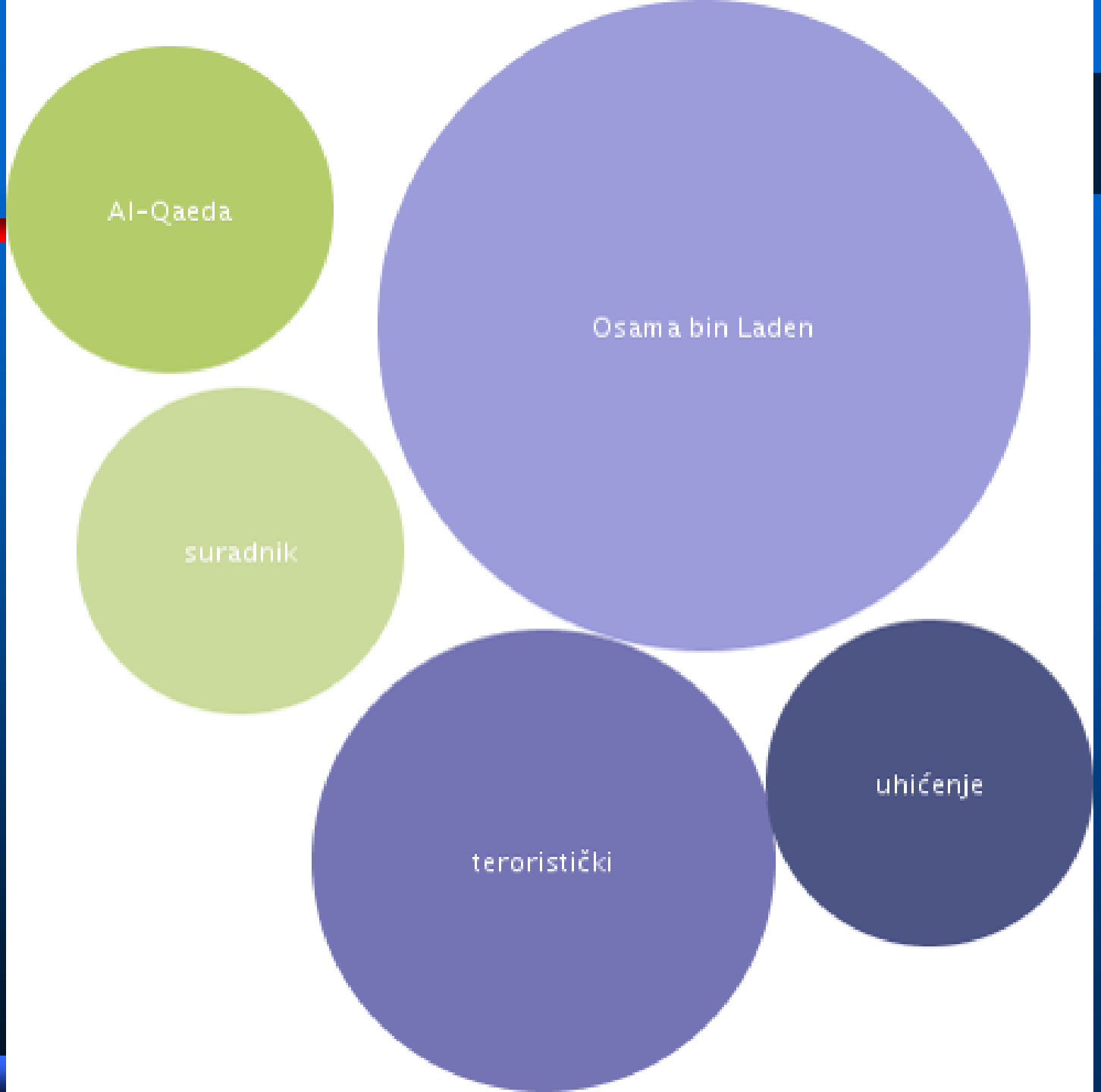
numerical and percentual values
temporal expressions
persons
locations
organizations

LT basis for knowledge technologies

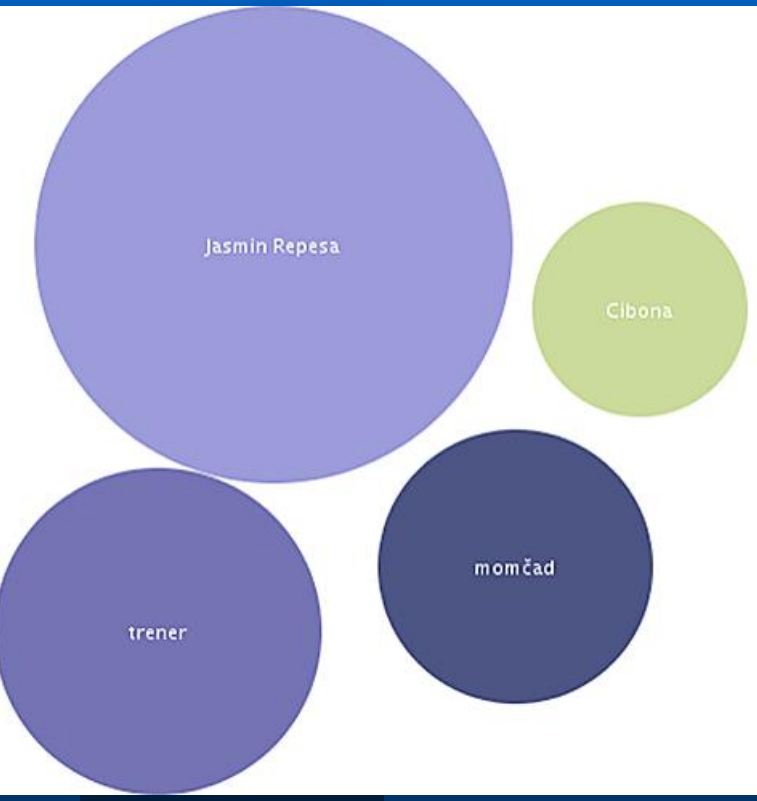
- **detection of relations between entities in**
  - **collections, documents, paragraphs, sentences, clauses**
  - **LT: sentence and clause splitters needed**
- **semantic graphs**



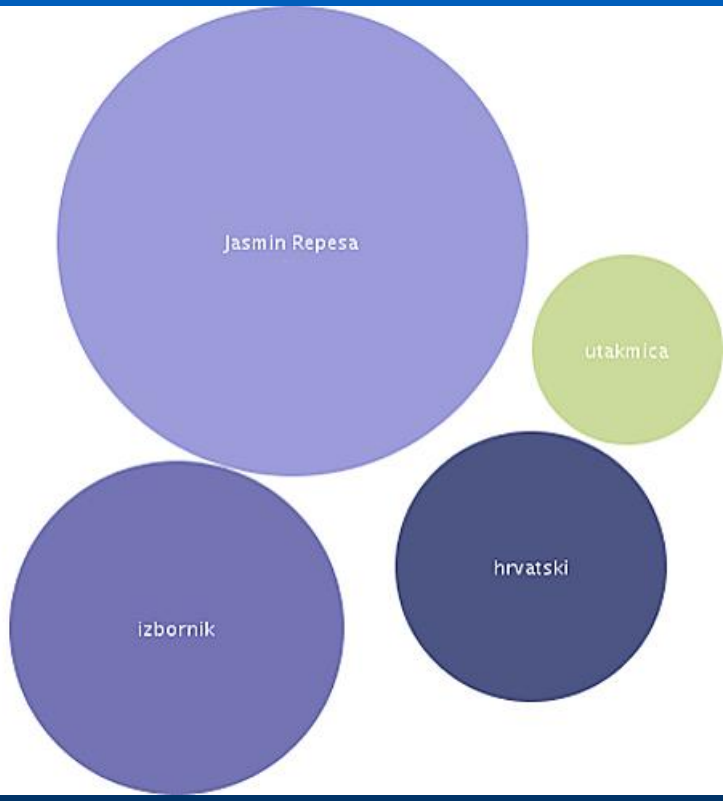

seksualni
Monica Lewinsky
1996.
Ovalni ...
Clinton
predsjednik
Bijela kuća

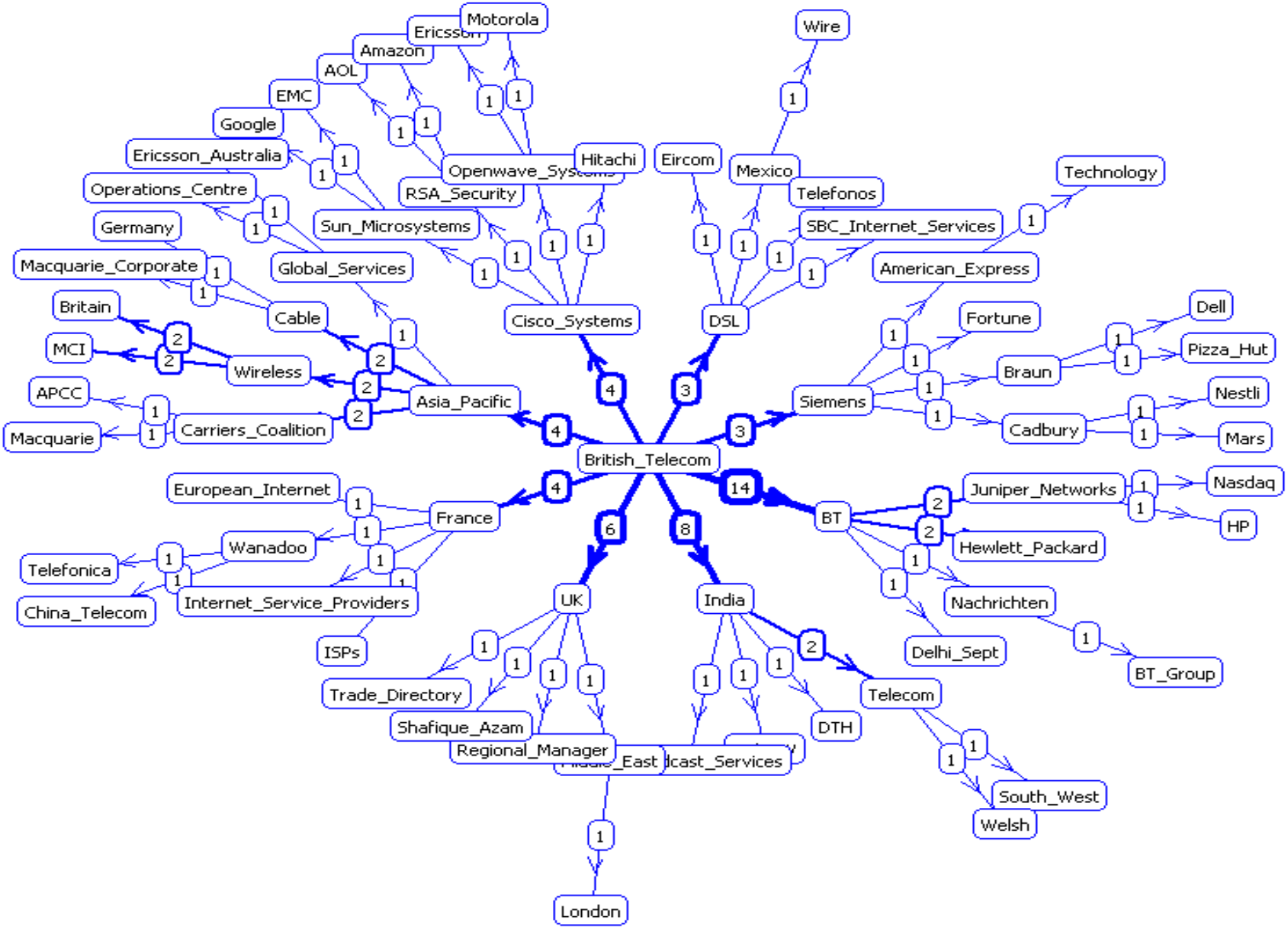

Al-Qaeda
Osama bin Laden
suradnik
uhićenje
teroristički

2001


Jasmin Repesa
Cibona
momčad
trener

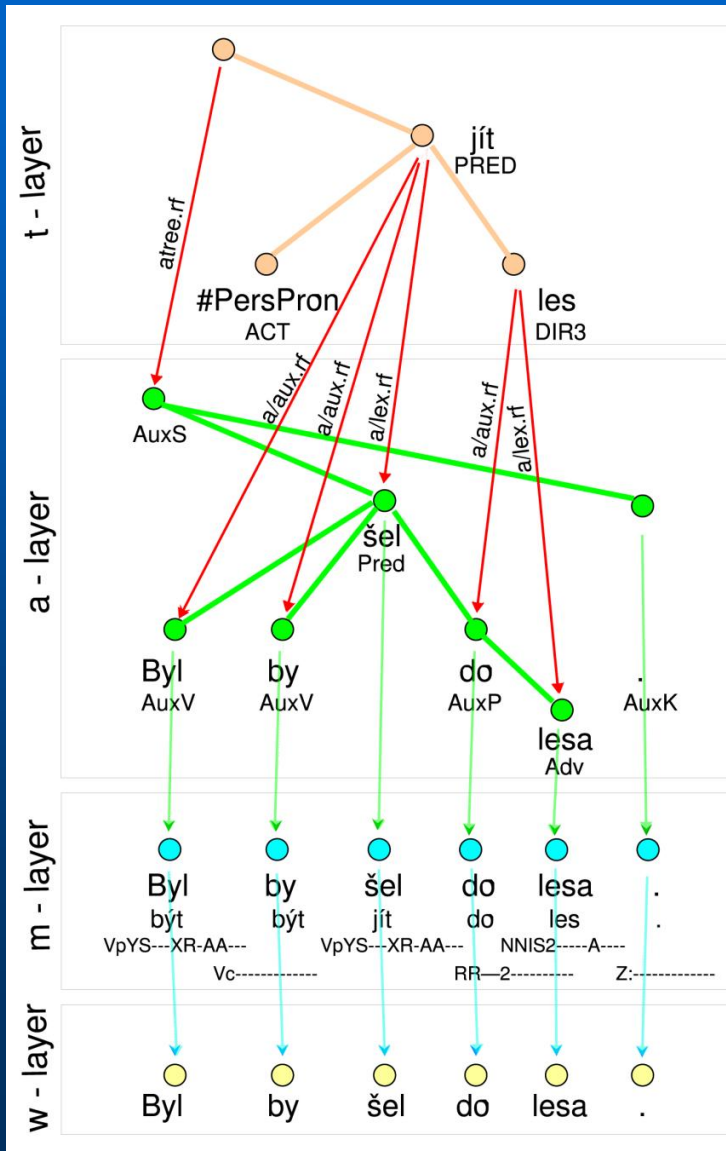
2009


Jasmin Repesa
utakmica
hrvatski
izbornik


British_Telecom
Cisco_Systems
DSL
Siemens
Asia_Pacific
Carriers_Coalition
France
Internet_Service_Providers
UK
India
BT
Telecom
Regional_Manager
Shafique_Azam
Trade_Directory
ISPs
Wanadoo
European_Internet
Telefonica
China_Telecom
Macquarie
APCC
Wireless
Britain
MCI
Cable
Macquarie_Corporate
Germany
Global_Services
Operations_Centre
Ericsson_Australia
Google
EMC
AOL
Amazon
Ericsson
Motorola
Sun_Microsystems
RSA_Security
Openwave_Systems
Hitachi
Eircom
Mexico
Wire
Telefonos
SBC_Internet_Services
American_Express
Technology
Fortune
Braun
Cadbury
Dell
Pizza_Hut
Nestli
Mars
Juniper_Networks
Nasdaq
HP
Hewlett_Packard
Nachrichten
BT_Group
Delhi_Sept
DTH
South_West
Welsh
London

LT basis for knowledge technologies

- **SVO detection**
  - fixed word order languages (en): easy
  - free word order languages (Slavic): problematic, morphology helps
- **semantic roles detection**
  - agent, patient, benefactor, instrument.
  - deep linguistic analysis
  - verb(subject,object) → V[S,O]
- **automatic ontology population**
  - RDF triples (“is a”, “is made of”, “is part of”, “is kind of”…)
  - RDFs in dbpedia
  - other languages?: cz, hu, fi, pl,…



LT as research infrastructures (RI)

- emergence of e-science paradigm
  - computationally intensive sciences
  - highly distributed network environments
  - immense data sets
  - grid computing
  - term by John Taylor, 1999
- research infrastructures
  - should enable the e-science approach
  - part of FP7: e.g. project CLARIN
- field of LT (i.e., LRT = language resources and tools)
  - mature enough to serve as research infrastructure for other sciences, particularly humanities and social sciences (HSS)