Language processing pipelines for knowledge technologies

Marko Tadić
University of Zagreb
Faculty of Humanities and Social Sciences
Department of Linguistics
marko.tadic@ffzg.hr

6th ESWC Summer School
The Rise of Data Science
CAAS, Dubrovnik, 2016-09-06

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

Is such a dialogue possible today?

- HAL is an artificial agent capable to apply the advanced NLP and AI techiques
  - speech and language
  - recognition and generation
  - understanding
  - information retrieval and extraction
  - reasoning
  - visual processing (lips reading, paralinguistic systems used in face-to-face interaction)
- Was Arhur Clarke too optimistic with 2001?

What do we need for such a dialogue?

- “There is no data like more data!”
- Big Data revolution
  - only with the rise in use of “unstructured data”
  - texts are considered as “unstructured data”
  - linguists say: “NO! A lot of structure in text”
    - text structure (headings, paragraphs, emphasis...)
    - structure of language expressions used in text
- ancient example: document retrieval
  - TF/IDF used to measure document relevancy for a given query
  - computational methods alone came to a ceiling

Basic concepts

- Computational Linguistics (CL)
- Natural Language Processing (NLP)
- Language Technologies (LT)
- difference between CL & NLP
  - linguists: Computational Linguistics
    - using computers in lingustic description (modelling)
    - aim: better description of language facts
  - informaticians: Natural Language Processing
    - using computers in processing language data
    - type of text processing
    - aim: efficiently process more (language) data with less computational resources used

Language Technologies (LT)

- linguistics = unique between humanities
  - research methods could be like in natural sciences
  - usage of scientific knowledge in making products
  - a lot of (commercial) products of LT exist already
- technology = “set of methods and procedures for processing of raw matter into products” (Croatian Lexicon, LZMK, 1996)
- raw matter and product in LT?
  - raw matter: language data
  - product: systems that enable simple(r) usage of L in computational environment

Language Technologies 2

- Language Resources (LR)
  - language data
    - raw text
    - structured (using markup Ls or predefined fixed format)
  - in the form of
    - corpora or text/document collections
    - dictionaries or lexical/terminological collections/databases

Language Technologies 3

- Language Tools (Lt)
  - programs that process LR
  - inherit the stratification at different language levels
    - morphology
    - syntax
    - semantics
    - more complex tools that combines several levels
- Language Services (LS)
  - generalized/specialized LR/Lt accessible on-line
  - usually as web services (REST/SOAP protocols)
  - used by humans and computers (APIs)

Language Technologies 4

- (commercial) products
  - checkers (spelling, grammar, style)
  - digital dictionaries (on-/off-line)
  - automatic indexing systems
  - summarisation systems
  - text to speech (TTS) & automatic speech recognition (ASR) systems
  - machine (aided) translation systems (MT, MAT)
  - computer aided language learning systems (CALL)

Corpus annotation

- corpus linguistics introduced corpus annotation on different language levels
  - in your case text annotation
- annotation
  - adding interpretations to the existing L units
  - inherent linguistic information turned into explicit data
  - becomes searchable
  - computer usability of text grows with the amount of annotation added

Corpus annotation 2

- storing of annotations
  - embedded in text (intermixed)
  - stand-off (separated)
- non-linguistic annotation
  - marking elements of the text structure: headings, titles, captions, bylines, signatures etc.
- linguistic annotation
  - takes into account language units and levels
  - segmentation: sentences, clauses, words (tokens)
  - tagging: adding of linguistic description to segmented language units

Corpus annotation: morphology

- part of speech tagging (PoS-tagging)
  - adding information about the PoS to each token in corpus, e.g. noun, verb, adjective etc.
- tagset = list of possible PoS in a given L
  - structure & cardinality: depends on L complexity and tagset design
  - initiatives for
    - Google universal PoS tagset (https://github.com/slavpetrov/universal-pos-tags) or
    - Universal Tagset (http://ckirov.github.io/UniMorph)
- taggers
  - programs that provide automatical tagging
  - precision around 98.8%

Corpus annotation: morphology 2

```
For_IF the_AT members_NN2 of_IO this_DD1
university_NN1 this_DD1 charter_NN1 enshrines_VVZ
a_AT1 victorious_JJ principle_NN1 ;_; and_CC the_AT
fruits_NN2 of_IO that_DD1 victory_NN1 can_VM
immediately_RR be_VBI seen_VVN in_II the_AT
international_JJ community_NNJ of_IO scholars_NN2
that_CST has_VHZ graduated_VVN here_RL today_RT ._.
```

```
NN1   singular common noun
NN2   plural common noun
NP    singular proper noun
NP$   genitive proper noun
PP$   possesive pronoun
RP    adverbial particle
VBD   past tense form of lexical verb
VBN   past participle of lexical verb...
```

Corpus annotation: morphology 3

- lemmatisation
  - adding information about the lemma to each token in corpus/text

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 m^2.

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

Corpus annotation: morphology 3

- lemmatisation
  - adding information about the lemma to each token in corpus/text
- lemma
  - basic word-form
  - particularly useful for inflectionally rich languages
  - helps avoiding data sparsness problem and downgrading in statistical approaches
  - precision around 98.5%
- disambiguation procedures
  - needed to resolve the homography

Corpus annotation: morphology 4

```
<seg lang="sl">
<w lemma="do">Do</w>
<w lemma="let letati leto">leta</w>
<w type=dig>2008</w>
<w lemma="se">se</w>
<w lemma="pričakovati">pričakuje</w>
<c>,</c>
<w lemma="da dati">da</w>
<w lemma="biti">bo</w>
<w lemma="odpravljen odpraviti">odpravljen</w>
<w lemma="bistven">bistveni</w>
<w lemma="del delo deti">del</w>
<w lemma="tradicionalen">tradicionalnih</w>
<w lemma="problem">problemov</w>
<w lemma="onesnaženost">onesnaženosti</w>
```

Corpus annotation: morphology 5

- morpho-syntactic tagging (MSD-tagging)
  - adding the morphosyntactic description (MSD) to each token in corpus
  - needed in parsing of languages with rich inflection
  - MSD categories and their values encode the syntactic roles, e.g. usually nouns in
    - nominatives = subjects, accusatives = direct objects, datives = indirect objects, instrumentals = instruments...
- tagsets
  - structure & cardinality: depends on L complexity and tagset design
  - precision around 92.5%

Corpus annotation: morphology 7

Njemački	njemački	Afpmpny	njemački	Afpmpvy	njemački
tužitelji	tužitelj	Ncmpn	tužitelj	Ncmpv	
kod	kod	Ncmsa--n	kod	Ncmsn	kod Spsg
Bajića	Bajić	Np-pg	Bajić	Np-sa	Bajić N
ZAGREB	Zagreb	Npmsa--n	Zagreb	Npmsn	
-	-	Z			
Glavni	glavni	Afpmpn-	glavni	Afpmpvy	glavni A
državni	državan	Afpmpn-	državan	Afpmpvy	državan A
odvjetnik	odvjetnik	Ncmsn			
trebao	trebati	Vmps-sma			
bi	biti1	Vcia2s	biti1	Vcia3p	biti1 V
ovaj	ovaj	Pd-msa--n-a-n	ovaj	Pd-msn--n-a--	
tjedan	tjedan	Ncmsa--n	tjedan	Ncmsn	
u	u	Spsa	u	Spsg	u Spsl
posjet	posjet	Ncmsa--n	posjet	Ncmsn	
primiti	primiti	Vmn			
njemačke	njemački	Afpfpay	njemački	Afpfpny	njemački
državne	državan	Afpfpa-	državan	Afpfpn-	državan A
odvjetnike	odvjetnik	Ncmpa			
koji	koji	Pi-mpn--n-a--	koji	Pi-msa--n-a-n	koji P
rade	rad	Ncmpa	rad	Ncmsv	raditi V

Corpus annotation: morphology 8

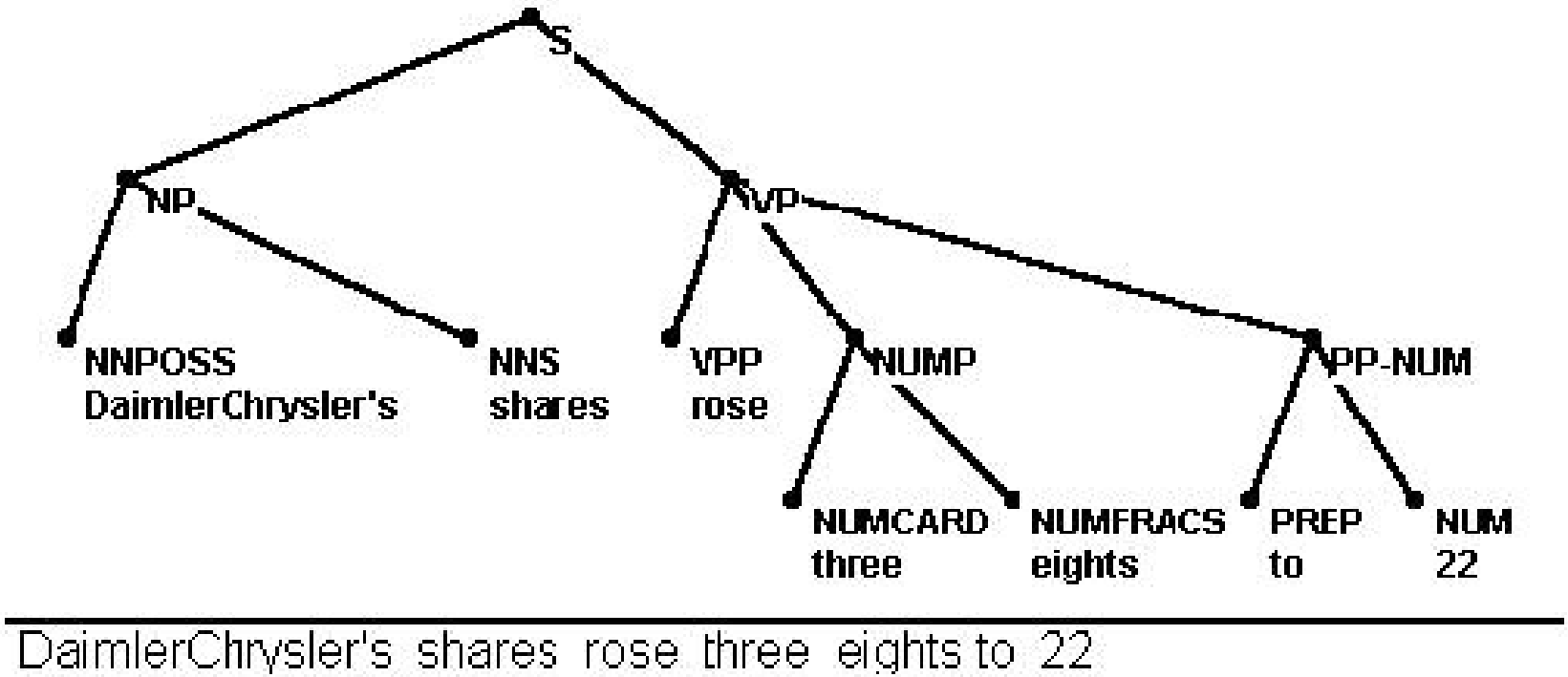
Njemački	njemački	Afpmpny
tužitelji	tužitelj	Ncmpn
kod	kod	Spsg
Bajića	Bajić	Np-sg
ZAGREB	Zagreb	Npmsn
-	-	Z
Glavni	glavni	Afpmsny
državni	državan	Afpmsny
odvjetnik	odvjetnik	Ncmsn
trebao	trebati	Vmps-sma
bi	biti1	Vcia3s
ovaj	ovaj	Pd-msn--n-a--
tjedan	tjedan	Ncmsn
u	u	Spsa
posjet	posjet	Ncmsa--n
primiti	primiti	Vmn
njemačke	njemački	Afpmpay
državne	državan	Afpmpa-
odvjetnike	odvjetnik	Ncmpa
koji	koji	Pi-mpn--n-a--
rade	raditi	Vmip3p

Corpus annotation: syntax

- analysis
  - manual: today only for training datasets
  - automatic
    - parsers = programs that syntactically analyse clauses and sentences
      - shallow/deep/robust
      - left-branching/right-branching
- tree-banks
  - corpora with inserted syntactic annotation
  - training material for stat. parsers (ML approaches)
- two main formalisms
  - constituency parsing vs. dependency parsing

Corpus annotation: syntax 2

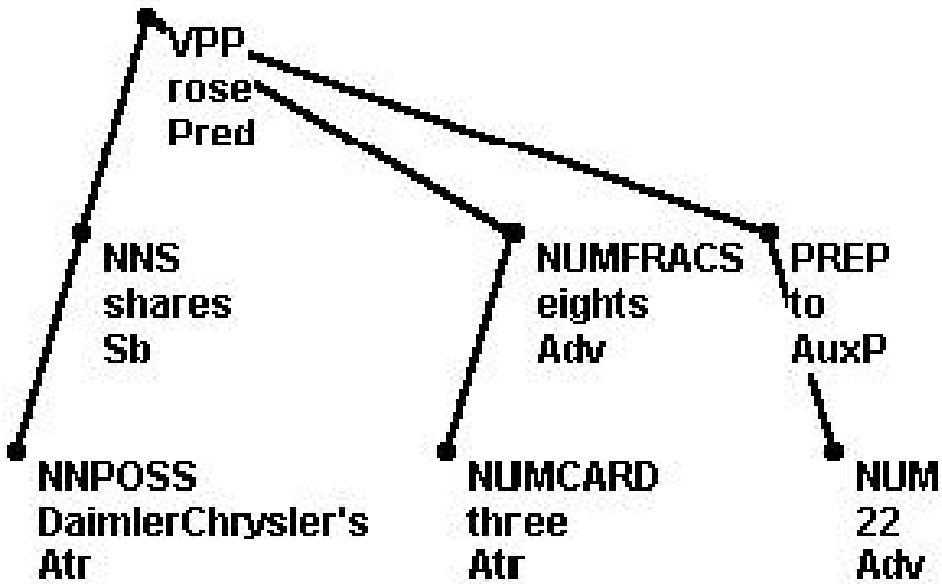
- constituency parsing



$$((\text{DaimlerChrysler's shares})_{NP}\ (\text{rose}\ (\text{three eights})_{NUMP}\ (\text{to } 22)_{PP\text{-}NUM})_{VP})_{S}$$

Corpus annotation: syntax 3

- dependency parsing

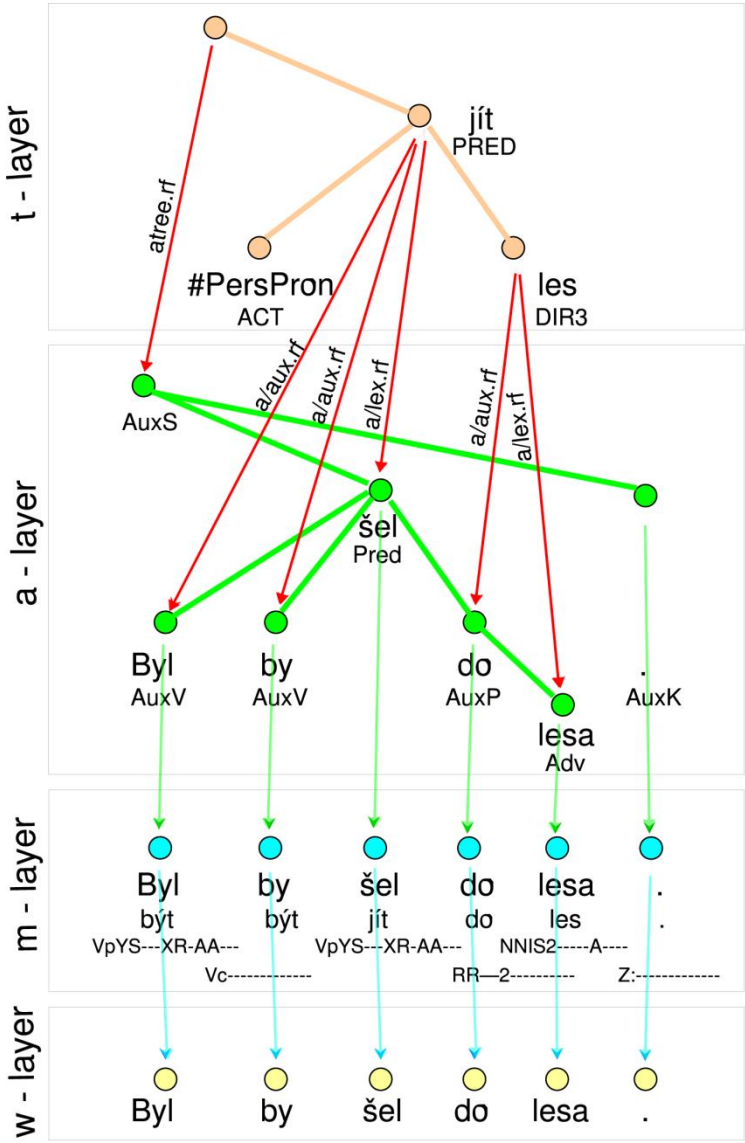


DaimlerChrysler's shares rose three eights to 22

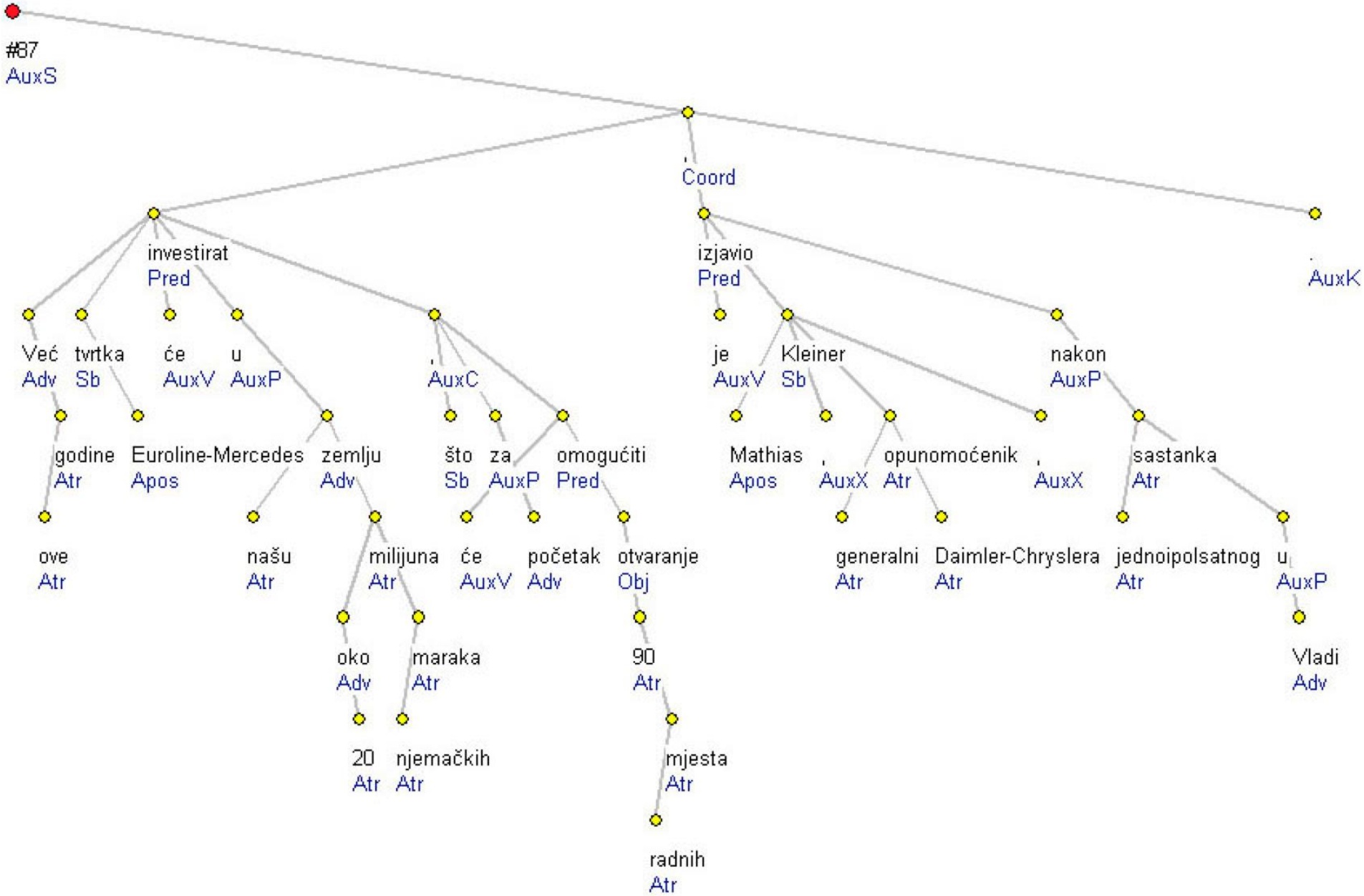
$$\mathbf{rose}_{\mathrm{Pred}}(\mathbf{shares}_{\mathrm{Sb}}(\mathbf{DaimlerChrysler's}_{\mathrm{Atr}}),\mathbf{eights}_{\mathrm{Adv}}(\mathbf{three}_{\mathrm{Atr}}),\mathbf{to}_{\mathrm{AuxP}}(\mathbf{22}_{\mathrm{Adv}}))$$

Corpus annotation: syntax 4

- surface vs. deep parsing
- e.g. Prague Tectogrammar
  - Daneš, Sgall, Hajičova, Hajič
  - Prague Dependency Treebank (PDT v2.0)
  - 2 main syntactic levels
    - analytical (a-layer)
    - tectogrammatical (t-layer, i.e. semantic roles)
- approach useful for Ls with free word order



#87 Već ove godine tvrtka Euroline-Mercedes investirat će u našu zemlju oko 20 milijuna njemačkih maraka , što će za početak omogućiti otvaranje 90 radnih mjesta , izjavio je Mathias Kleiner , generalni opunomoćenik Daimler-Chryslera , nakon jednoipolsatnog sastanka u Vladi .



Corpus annotation: syntax 6

1	Lujo	Lujo	N	N	PERSON	5	Sb	_	_
2	danas	danas	R	R	O	5	Adv	_	_
3	u	u	S	S	O	5	AuxP	_	_
4	Dubrovniku	Dubrovnik	N	N	LOCATION	3	Adv	_	_
5	predstavlja	predstavljati	V	V	O	0	Pred	_	_
6	jezične	jezičan	A	A	O	7	Atr	_	_
7	tehnologije	tehnologija	N	N	O	5	Obj	_	_
8	.	.	Z	Z	O	0	AuxK	_	_

Corpus annotation: syntax 7

- Universal dependencies initiative
  - cross-L consistent syntactic annotation
  - for 48 languages
  - facilitating
    - multilingual parser development
    - cross-L learning & cross-L retrieval
    - parsing research
  - http://universaldependencies.org

Corpus annotation: syntax 8

- local grammars (shallow parsing)
  - detection of classes of words and/or phrases
  - chunks = unrecursive clause parts
  - particular type of expressions
    - temporal
    - spatial...
  - useful for Information Extraction, text-mining…
  - could be also used for detecting
    - multiword expressions (MWEs)
    - e.g. SWE vs. MWE
      - *carrier* vs. *aircraft carrier*
      - *take* vs. *take off*

Corpus annotation: syntax 8

- NERC
  - Named Entities Recognition and Classification
  - 7 types of named entities (following MUC1997)
    - person
    - location
    - organisation
    - date
    - time
    - value (money, measurements)
    - percentage
  - more NEs added later
    - geopolitical entities (e.g. EU, NATO, EEA,...)
    - protein, gene, drug, desease, species names...

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 povećanje**180 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 priremu 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**.

Legenda:

brojčani i postotni iznosi

vremenski izrazi

imena osoba

imena lokacija

imena organizacija

Corpus annotation: syntax 9

```
<BODY>
<DIV0 type="MAIN">
<HEAD type="NA">Nagrada zagrebačkim gitaristima</HEAD>
<P><ENAMEX TYPE="ORGANIZATION">Zagrebački gitaristički kvartet</ENAMEX> osvojio
je prvu nagradu na <ENAMEX TYPE="ORGANIZATION">Međunarodnome gitarističkom
natjecanju Simone Salmaso</ENAMEX> u <ENAMEX TYPE="LOCATION">Viareggiu</ENAMEX>
u konkurenciji 14 komornih sastava (u kategoriji D). Prvo mjesto je kao solist
osvojio i član toga renomiranoga zagrebačkog sastava <ENAMEX
TYPE="PERSON">Darko Pelužan</ENAMEX> u konkurenciji 30 gitarista (u kategoriji
C). Članovi <ENAMEX TYPE="ORGANIZATION">Zagrebačkoga gitarističkog
kvarteta</ENAMEX> (koji je 1990. osnovao profesor <ENAMEX TYPE="PERSON">Ante
Čagalj</ENAMEX>, pretežno od studenata gitare) sada su još <ENAMEX
TYPE="PERSON">Mihaela Pažulinec</ENAMEX>, <ENAMEX TYPE="PERSON">Krunoslav
Pehar</ENAMEX> i <ENAMEX TYPE="PERSON">Melita Ivković</ENAMEX>. To nije prvi
put da <ENAMEX TYPE="ORGANIZATION">Zagrebački gitaristički kvartet</ENAMEX>
osvaja prvu nagradu na nekome međunarodnom natjecanju u <ENAMEX
TYPE="LOCATION">Italiji</ENAMEX>: pobijedio je i prije dvije godine u <ENAMEX
TYPE="LOCATION">Tarantu</ENAMEX> na 6. međunarodnom natjecanju <ENAMEX
TYPE="ORGANIZATION">Trofeo Kawai</ENAMEX>.</P>
<BYLINE>(<ENAMEX TYPE="ORGANIZATION">Večernji list</ENAMEX>)</BYLINE>
</DIV0>
</BODY>
```

Corpus annotation: semantics

- lexical semantics
  - annotation of word senses, i.e. the meaning that has realised with a particular token (SWE or MWE)
  - incl. synonyms/hyponyms
  - mapping of SWE/MWE to a conceptual space
    - WordNet (1990), Babelnet, Cyc-ontology, Wikipedia/DBpedia, LOD universe...
- sentential semantics
  - semantic roles labelling (SRL)
  - detecting the roles that a predicate in clause opens
    - e.g. agent, patient, beneficiary, instrument

Corpus annotation: semantics 2

```
<contextfile concordance=brown1>
<context filename=br-j30 paras=yes>
<p pnum=1>
<s snum=1>
<wf cmd=done pos=NN lemma=analysis wnsn=1 lexsn=1:04:00::>Analysis</wf>
<wf cmd=done pos=VB lemma=mean wnsn=1 lexsn=2:32:01::>means</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=evaluation wnsn=1 lexsn=1:04:00::>evaluation</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=subpart wnsn=1 lexsn=1:24:00::>subparts</wf>
<punc>,</punc>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=JJ lemma=comparative wnsn=2 lexsn=5:00:00:relative:00>comparative<
<wf cmd=done pos=NN lemma=rating wnsn=3 lexsn=1:26:00::>ratings</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=part wnsn=5 lexsn=1:09:00::>parts</wf>
<punc>,</punc>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=comprehension wnsn=1 lexsn=1:09:00::>comprehension</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=meaning wnsn=2 lexsn=1:09:00::>meaning</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=JJ lemma=isolated wnsn=3 lexsn=5:00:01:separate:00>isolated</wf>
<wf cmd=done pos=NN lemma=element wnsn=1 lexsn=1:09:00::>elements</wf>
<punc>.</punc>
</s>
<s snum=2>
```

Corpus annotation: semantics

- Wordnets
  - Princeton Wordnet
    - http://wordnet.princeton.edu
  - Global Wordnet Association
    - http://globalwordnet.org
    - list of wordnets
  - Open Multilingual Wordnet
    - http://compling.hss.ntu.edu.sg/omw/
  - Universal Wordnet
    - http://www.mpi-inf.mpg.de/yago-naga/uwn/
    - over 1.5 Mwords, 200 Ls
- Babelnet: http://babelnet.org

Employing

Definition:

An Employer employs an Employee whose Position entails that the Employee perform certain Tasks in exchange for Compensation.
I EMPLOYED him as Chief Gardener for ten years.

FEs:

Core:

Employee [Empee] — The FE Employee denotes the person who is obligated to perform some Task in order to receive Compensation.
I was EMPLOYED by an international corporation.

Employer [Emper] — The Employer is the person (or institution) that gives Compensation to an Employee.
I EMPLOYED him as Chief Gardener for ten years.

Field [Field] — The FE Field identifies the field in which the Employee is employed.
He was EMPLOYED in finance fourteen years ago.

Position [Posit] — The FE Position indicates a particular type of employment.
I'm not EMPLOYED as your waitress!

Task [Task] — The Task indicates the action/duty that the Employee is obligated to do for the Employer.
I am EMPLOYED to collect the trash.

Non-Core:

Compensation [Compense] — The Compensation is the payment that the Employee receives for performing a Task.

Contract_basis [CB] — The Contract_basis is the condition of employment with respect to permanency, hours per time period or payment arrangements.
Chrysler EMPLOYS 1000 skilled worker full-time.

Descriptor [] — This FE describes a characteristic or the entity in question.

Duration [Dur] — The FE Duration identifies the amount of time for which the Employee continues in employ.

More complex useful LT tasks

- Language Identification
  - recognizing the language of a text
- Sentiment Analysis
  - detection of the attitude of writer towards the topic s/he is conveying or in comments
- Q/A systems
  - systems that find answers on questions from a given collection of documents/facts/ontologies...
  - e.g. IBM Watson winner in the Jeopardy in 2013
- Machine Translation
  - systems that translate from one natural language into another

XLike linguistic processing pipelines

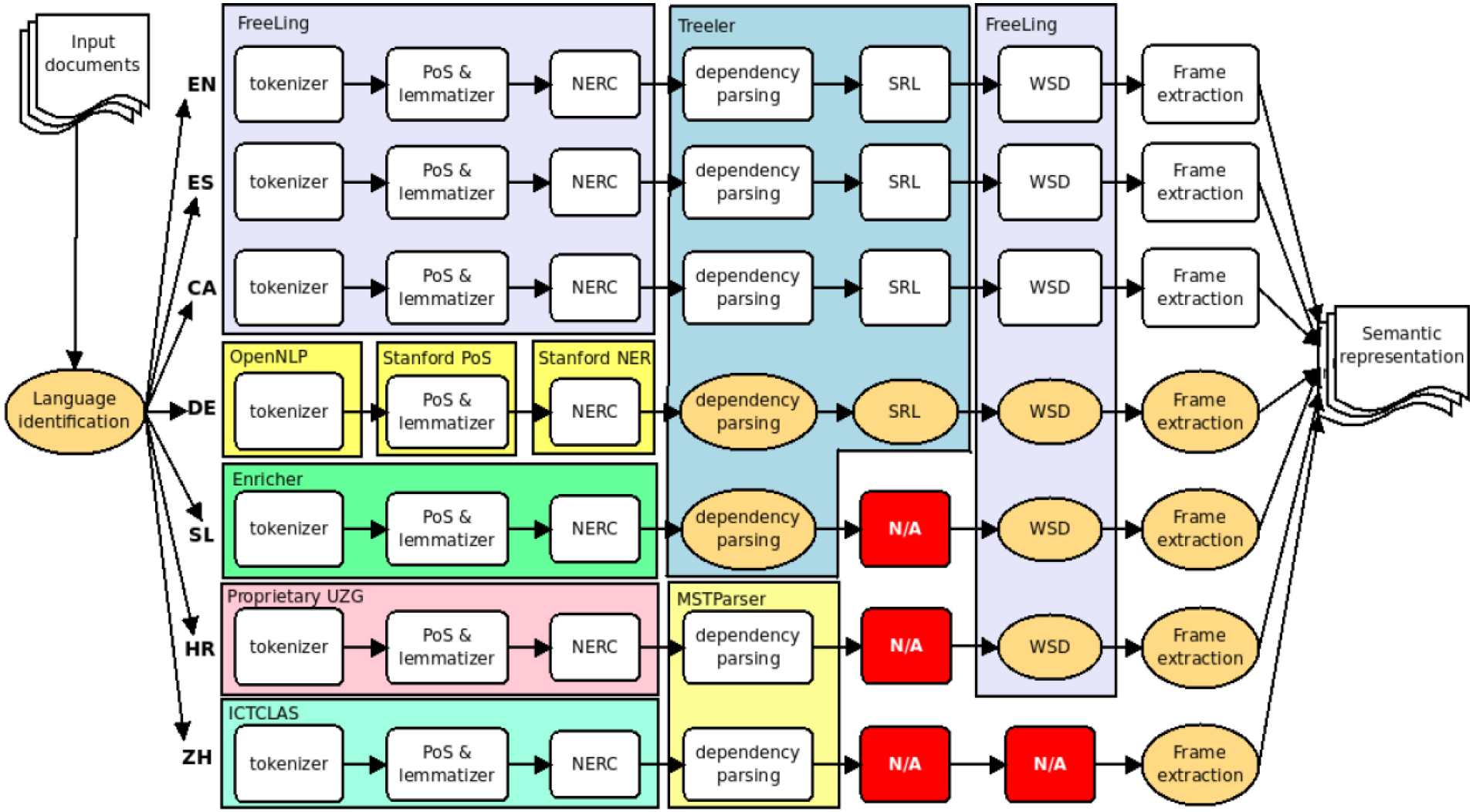
- XLike project (http://xlike.org) combined scientific capabilities and insights from several areas of science
  - computational linguistics (LT)
  - machine learning
  - text mining
  - semantic technologies

  in order to enable cross-lingual text "understanding" by machines
- goals of XLike WP2
  - develop tools to extract entities and relations found in documents
  - for multiple
    - languages, domains, language registers (standard vs. non-standard)
  - a solid LT foundation used throughout the project

XLike linguistic processing pipelines

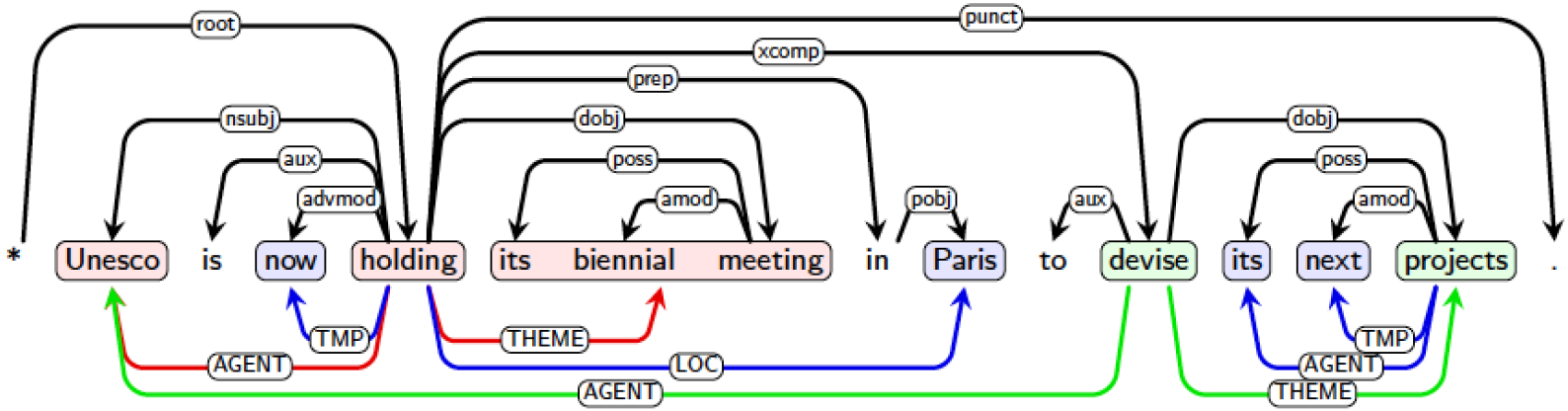
- entry module: automatic language identification
- 7 pipelines for each of XLike languages
  - sentence splitting
  - tokenization
  - lemmatization
  - POS/MSD-tagging
  - NERC
  - dependency parsing
  - semantic role labelling
- pipelines function as web services
- languages covered
  - English, Spanish, German, Chinese, Catalan, Slovene, Croatian

XLike linguistic processing pipelines



XLike linguistic processing pipelines

- starting from
  - "Unesco is now holding its biennial meeting in Paris to devise its next projects."
- at the end we want to come to



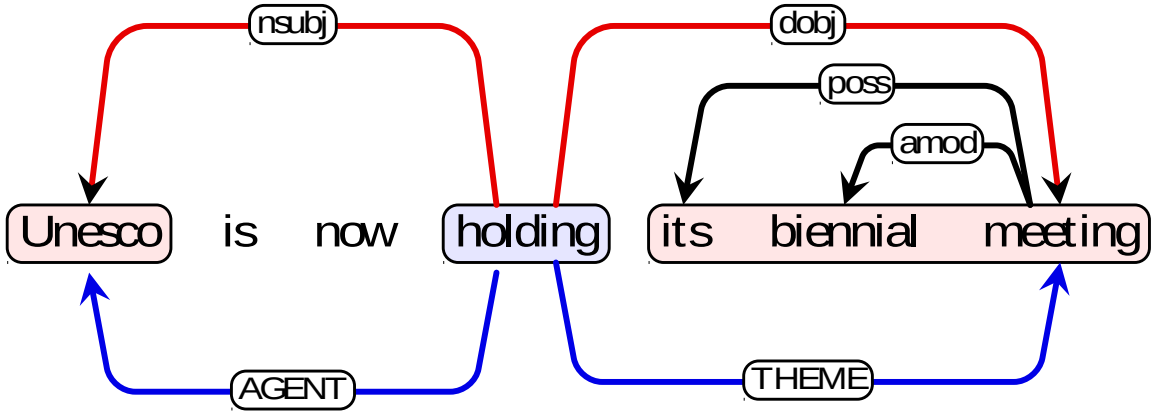
- Two propositions:

[EVENT: hold [AGENT: Unesco] [THEME: meeting]
[LOCATION: Paris] [TIME: now]]

[EVENT: devise [AGENT: Unesco] [THEME: projects]]

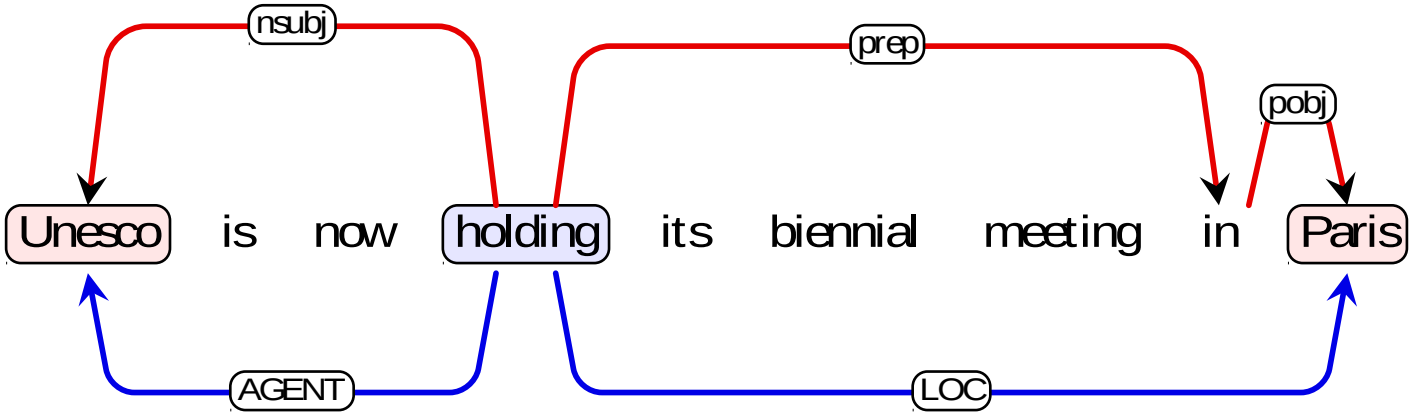
XLike linguistic processing pipelines

- five types of target extraction elements
  - Tokens (requires tokenization)
  - Lemmas (requires lemmatization and PoS tagging)
  - Syntactic Triples (requires syntactic parsing)
  - Semantic Triples (requires semantic role labeling)
  - Entity Relations (requires connecting entities)
- Dependency-based Extraction
  - Syntactic: subject-verb-object
  - Semantic: agent-predicate-theme



XLike linguistic processing pipelines

- we may be interested in detecting relations between entities (beyond subject-object)



- Syntactic/Semantic paths provide a rich, meaningful source of features to characterize such relations
- statistical classification methods can be used to label such relations

XLike linguistic processing pipelines

ID	WORD--->	LEMMA-->	POS	MSD	NE-->	DEP-->	SRL---------->				
1	Unesco	unesco	NNP	_	B-ORG	2	SBJ	_	A0	A0	_
2	is	be	VBZ	_	O	0	ROOT	_	_	_	_
3	now	now	RB	_	O	2	TMP	_	AM-TMP	_	_
4	holding	hold	VBG	_	O	2	VC	hold.04	_	_	_
5	its	its	PRP	_	O	7	NMOD	_	_	_	_
6	biennial	biennial	JJ	_	O	7	NMOD	_	_	_	_
7	meetings	meeting	NNS	_	O	4	OBJ	_	A1	_	_
8	in	in	IN	_	O	7	LOC	_	AM-LOC	_	_
9	New	new	NNP	_	B-LOC	9	NAME	_	_	_	_
10	York	york	NNP	_	I-LOC	8	PMOD	_	_	_	_
11	to	to	TO	_	O	4	PRP	_	_	_	_
12	devise	devise	VB	_	O	11	IM	devise.01	_	_	_
13	its	its	PRP	_	O	15	NMOD	_	_	_	A0
14	next	next	JJ	_	O	15	NMOD	_	_	_	AM-TMP
15	projects	project	NNS	_	O	12	OBJ	project.02	_	A1	_
16	.	.	.	_	O	2	P	_	_	_	_

- live demo
  - http://xorrai.lsi.upc.edu/xlike/demo.php
  - Unesco is now holding its biennial meeting in Paris to devise its next projects
  - Unesco u Parizu održava svoj dvogodišnji sastanak s planiranjem svojih sljedećih projekata.

```
<item>
 <sentences>
  <sentence id="1">
   <text>Unesco is now holding its biennial meetings in New York.</text>
    <tokens>
      <token pos="NP00SP0" end="6" lemma="unesco" id="1.1" start="0">Unesco</token>
      <token pos="VBZ" end="9" lemma="be" id="1.2" start="7">is</token>
      <token pos="RB" end="13" lemma="now" id="1.3" start="10">now</token>
      <token pos="VBG" end="21" lemma="hold" id="1.4" start="14">holding</token>
      <token pos="PRP$" end="25" lemma="its" id="1.5" start="22">its</token>
      <token pos="JJ" end="34" lemma="biennial" id="1.6" start="26">biennial</token>
      <token pos="NNS" end="43" lemma="meeting" id="1.7" start="35">meetings</token>
      <token pos="IN" end="46" lemma="in" id="1.8" start="44">in</token>
      <token pos="NP00G00" end="55" lemma="new_york" id="1.9" start="47">New_York</token>
      <token pos="Fp" end="56" lemma="." id="1.10" start="55">.</token>
    </tokens>
   </sentence>
 </sentences>
 <entities>
   <entity type="location" displayName="new_york" id="2">
    <mentions>
      <mention SentenceId="1" id="1.9" words="New York"/>
    </mentions>
   </entity>
   <entity type="person" displayName="organization" id="1">
    <mentions>
      <mention SentenceId="1" id="1.1" words="Unesco"/>
    </mentions>
   </entity>
  </entities>
  <relations>
    <relation subject="1.1" name="hold" object="1.9" id="3"/>
  </relations>
</item>
```

Wikifier service

- online service that in texts detects
  - entities existing in Wikipedia/DBpedia
  - maps to their IDs
  - provides links to concepts
  - covers 134 languages
- http://wikifier.org
  - e.g. http://#

Event registry service

- online service that tracks RSS newsfeeds
  - more than 100.000 news sources worldwide
  - daily amount 100.000-150.000 news articles
  - 10+ Ls
    - S-splitting
    - tokenization
    - POS-tagging
    - NERC + disambiguation (e.g. Washington)
    - date references
    - DMOZ categorization
  - cross-L service
    - compares articles in different Ls
    - computes their similarity
  - event identification, tracking, contextualization
- http://eventregistry.org

LT repositories

- CL/NLP/LT community already quite mature
  - you can find tools/services for Ls you need in existing (federations) of repositories
  - CLARIN-ERIC
    - European Research Infrastructure Consortium
      - particularly Weblicht (http://weblicht.sfs.uni-tuebingen.de)
    - http://clarin.eu
  - META-SHARE
    - http://meta-share.eu
  - European Language Resources Agency (ELRA)
    - http://www.elra.info
  - Linguistic Data Consortium (LDC)
    - https://www.ldc.upenn.edu

LT repositories

- other popular online pipelines/tool collections
  - GATE
    - http://gate.ac.uk
  - FreeLing
    - http://nlp.lsi.upc.edu/freeling
  - NLP Stanford
    - http://nlp.stanford.edu/software ...
- commercial analytics packages today include LT components
  - IBM Watson
  - Linguamatics in I2E ...

Bright future for NLP/LT

DataInformed
Big Data and Analytics in the Enterprise

Advanced Analytics | Cloud | Customer Analytics | Data Visualization | IoT | Management | Operations

eBooks | Events | Glossary | University Map | Use Cases | Webinars | Whitepapers |

Why Natural Language Processing Will Change Everything

by **Bernard Marr** | August 2, 2016 5:30 am | 2 Comments

Bernard Marr

Do you talk to your computer or smartphone? Just a few years ago, that question would have been absurd. But with advances in natural language processing, the likelihood is that you have asked your phone to send a text or search the web for something within the last day.

In fact, natural language processing (NLP) is one aspect of machine learning, big data, and artificial intelligence that has the potential to truly change everything.

In its most basic terms, natural language processing is the ability of a computer to understand natural human speech as it is spoken. It's the difference between saying, "Siri, where's the nearest coffee shop?" and, "Search coffee shops ZIP Code 80021."

For a long time, searches online had to be done by typing in strings of words combined with Boolean search terms that ended up looking and sounding nothing like a conversation. Now, however, you can type a question into Google exactly how you'd ask it to a friend, and Google can reliably provide a good answer.

The same recognition of natural language is being developed for speech. AI assistants like Siri,

Search

submit

Featured Downloads

- 10 Ways Data Preparation Can Help Excel
- Boost your Bottom Line with IT Operations Analytics
- Tap into the Power of Machine Learning
- Data Preparation and the Evolution of Analytics
- How to Overcome Challenges in Complex Data

Thank you for your attention!