

Cursul 13  
Noțiuni de  
prelucrare a limbajului natural

# Domeniul

- *Lingvistica computațională (LC)* – osatura teoretică
  - *computational linguistics*
- *Prelucrarea limbajului natural* – zona aplicativă
  - *natural language processing*
  - tehnologia limbajului natural, *natural language technology*
  - tehnologia limbajului uman, *human language technology*

# Tehnologia limbajului natural

- Limbajul vorbit
- Limbajul scris
- Limbajul în corelație cu alte modalități de expresie (multimodalitate)

# Tehnologiile limbajului vorbit

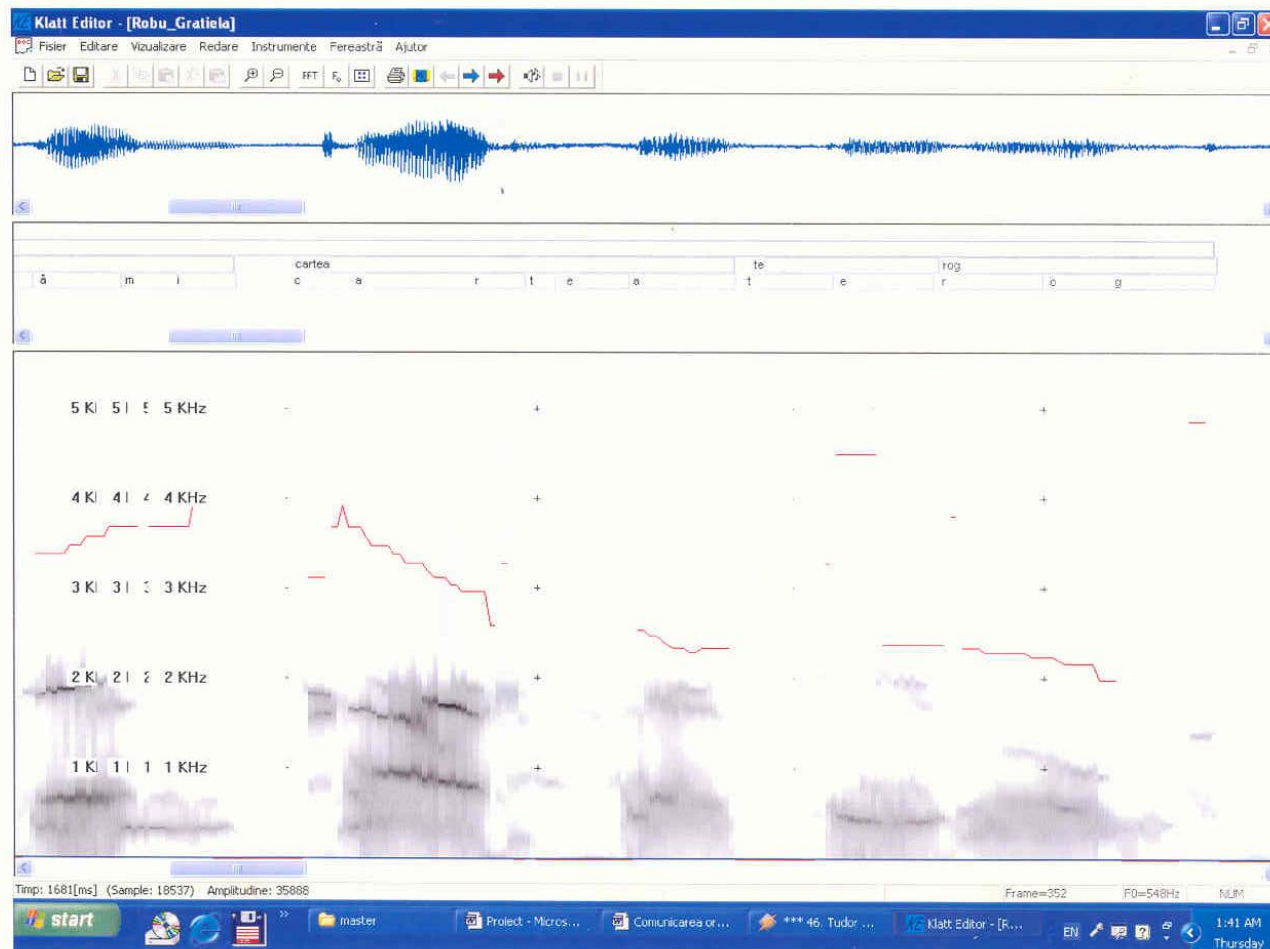
- Interpretarea vocii
  - reprezentarea semnalelor vocale
  - recunoașterea vorbirii
  - analiza prozodiei
  - recunoașterea vorbitorului
- Sinteza vocală

# Prelucrarea semnalului sonor

Robu Grațela-Andreea

Masterat Lingvistică Computațională – anul al II-lea

Facultatea de Informatică, Univ. „Al. I. Cuza”

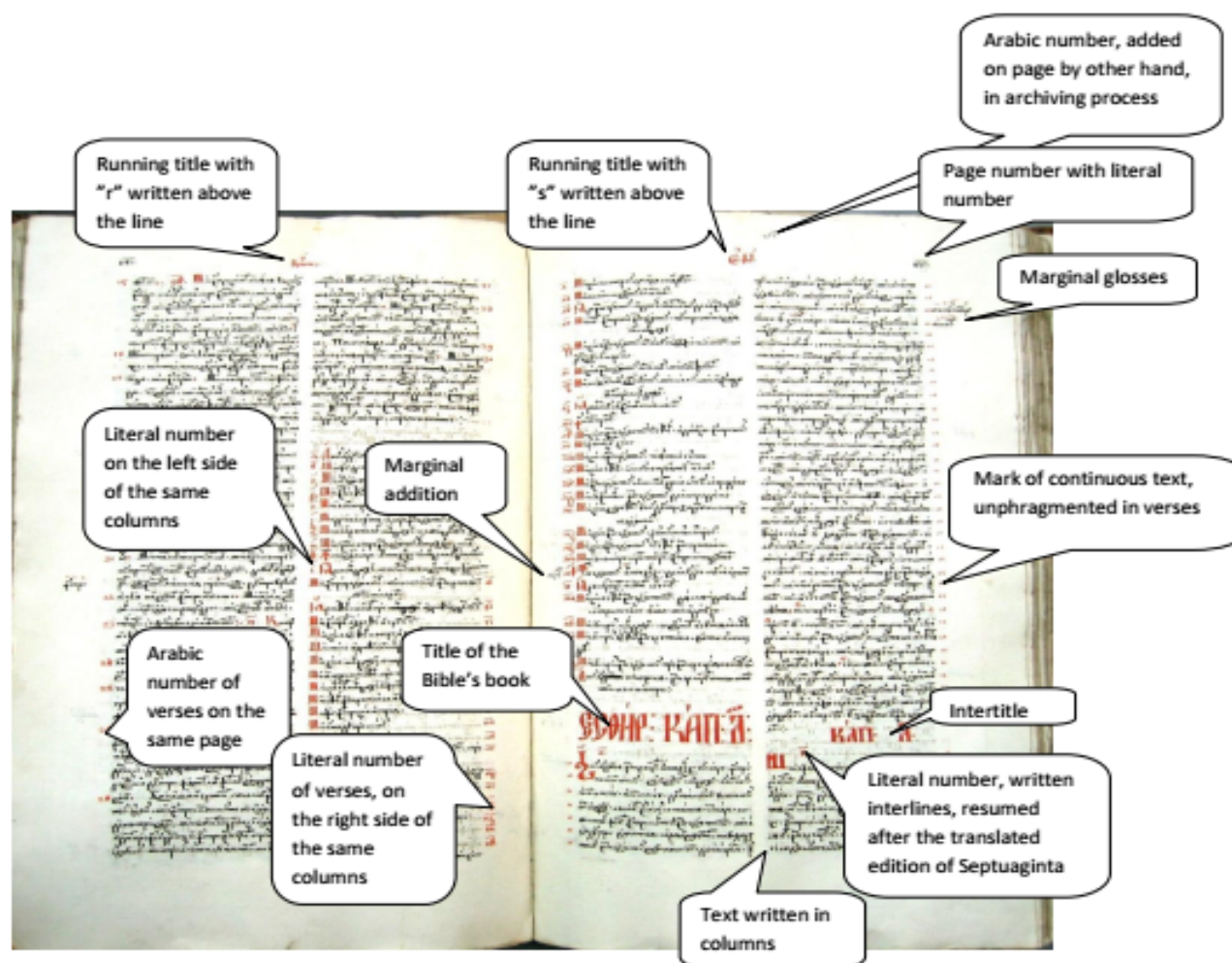


Prin bunăvoința Acad. H.N.Teodorescu

# Tehnologiile limbajului scris

- Tehnologii primare
  - Analiza imaginii documentelor
  - Recunoașterea caracterelor de tipar
  - Recunoașterea caracterelor de mână
    - *Optical Character Recognition* (OCR)

# Ms. 45 BAR Cluj-Napoca, second half of 17<sup>th</sup> century



Different types of writing in the revised copy of Nicolae Milescu's translation of *Septuaginta*, p. 412-413

# Tehnologiile limbajului scris

- Analiza și înțelegerea limbajului
  - prelucrări sub-sintactice
    - unitățile lexicale
    - granițele de frază
    - granițele de propoziții
    - partea de vorbire și marca morfologică
    - lema
    - numele de entități
    - grupurile (nominale, verbale, prepoziționale etc.) și atracțiile lexicale (colocații)



# Fraze

Comisia juridică a Camerei Deputaților a votat luni împotriva începerii urmăririi penale a ministrului demisionar al Fondurilor Europene, Rovana Plumb, după ce aceasta a fost audiată, alături de avocat, aproximativ o oră de către deputații juriști. | Rovana Plumb s-a declarat, din nou, la ieșirea de la audierile din comisia juridică, nevinovată de acuzațiile aduse de procurorii anticorupție. |

# Propoziții

Pe 22 septembrie, DNA a anunțat că | vicepremierul Sevil Shhaideh este suspectă de | săvârșirea infracțiunii de abuz în serviciu | când era secretar de stat la MDRAP, într-un dosar conform căruia, în 2013, | prin acțiunea concertată a unor persoane cu funcții publice, | părți din Insula Belina și Brațul Pavel au trecut ilegal din proprietatea statului în proprietatea județului Teleorman și în administrarea CJ Teleorman, | pentru ca, doar la câteva zile, să fie închiriate tot ilegal unei firme private. |

# Unități lexicale

Pe 22 septembrie , DNA a anunțat că vicepremierul  
Sevil Shhaideh este suspectă de săvârșirea  
infracțiunii de abuz în serviciu când era secretar de  
stat la MDRAP , într-un dosar conform căruia, în  
2013, prin acțiunea concertată a unor persoane cu  
funcții publice, părți din Insula Belina și Brațul Pavel  
au trecut ilegal din proprietatea statului în  
proprietatea județului Teleorman și în administrarea  
CJ Teleorman, pentru ca, doar la câteva zile, să fie  
închiriate tot ilegal unei firme private.



persoană

## Entități cu nume

Pe 21 septembrie, DNA a anunțat că vicepremierul **Sevil Shhaideh** este suspectă de săvârșirea infracțiunii de abuz în serviciu când era secretar de stat la MDRAP, într-un dosar conform căruia, în 2013, prin acțiunea concertată a unor persoane cu funcții publice, părți din Insula Belina și Brațul Pavel au trecut ilegal din proprietatea statului în proprietatea județului Teleorman și în administrarea CJ Teleorman, pentru ca, doar la câteva zile, să fie închiriate tot ilegal unei firme private.

dată

## Entități cu nume

Pe **22 septembrie**, DNA a anunțat că vicepremierul Sevil Shhaideh este suspectă de săvârșirea infracțiunii de abuz în serviciu când era secretar de stat la MDRAP, într-un dosar conform căruia, în **2013**, prin acțiunea concertată a unor persoane cu funcții publice, părți din Insula Belina și Brațul Pavel au trecut ilegal din proprietatea statului în proprietatea județului Teleorman și în administrarea CJ Teleorman, pentru ca, doar **la câteva zile**, să fie închiriate tot ilegal unei firme private.

reper  
temporal

instituție

## Entități cu nume

Pe 22 septembrie, **DNA** a anunțat că vicepremierul Sevil Shhaideh este suspectă de săvârșirea infracțiunii de abuz în serviciu când era secretar de stat la **MDRAP**, într-un dosar conform căruia, în 2013, prin acțiunea concertată a unor persoane cu funcții publice, părți din Insula Belina și Brațul Pavel au trecut ilegal din proprietatea statului în proprietatea județului Teleorman și în administrarea **CJ Teleorman**, pentru ca, doar la câteva zile, să fie închiriate tot ilegal unei firme private.

locații  
geografice

## Entități cu nume

Pe 22 septembrie, DNA a anunțat că vicepremierul Sevil Shhaideh este suspectă de săvârșirea infracțiunii de abuz în serviciu când era secretar de stat la MDRAP, într-un dosar conform căruia, în 2013, prin acțiunea concertată a unor persoane cu funcții publice, părți din **Insula Belina** și **Brațul Pavel** au trecut ilegal din proprietatea statului în proprietatea **județului Teleorman** și în administrarea CJ **Teleorman**, pentru ca, doar la câteva zile, să fie închiriate tot ilegal unei firme private.

# Lema și partea de vorbire

Solicitat – solicita – vb

să – să – conj

comenteze – comenta – vb

un – un – art.nehot.

editorial – editorial – sb

recent – recent – adj

...



# Adnotarea morfologică

- English

```
0
1  He  he  subj:>2    @SUBJ PRON
2  did  do  v-ch:>4    @+FAUXV V
3  not  not  neg:>2    @ADVL NEG-PART
4  know know main:>0  @-FMAINV V
5  her  she  subj:>6    @OBJ PRON
6  name name obj:>4@-FMAINV V
```

- Romanian

```
<TOK ID="TOK478" root="Nu" pv="Particle" Type="negation">Nu</TOK>
<TOK ID="TOK479" root="ști" pv="Verb" Type="main" Mood="indic."
  Tense="imperfect" Person="third" Number="singular">știa</TOK>
<TOK ID="TOK480" root="cum" pv="Adverb" type="int_rel">cum</TOK>
<TOK ID="TOK481" root="ei" pv="Pronoun" Type="pers" Person="third"
  Gender="feminine" Number="singular" Case="accusative">o</TOK>
<TOK ID="TOK482" root="chema" pv="Verb" Type="main" Mood="indic."
  Tense="present" Person="third">cheamă</TOK>
```

# Grupuri nominale

Solicitat să comenteze [un editorial recent al lui [Dinu Patriciu]], în [care] [acesta] preciza că nu crede în [social-liberalism] și să aprecieze dacă, astfel, a dat [o lovitură de [image]] [USL], [Antonescu] a spus că nu știe dacă [Patriciu] s-a referit la [USL].

# Adnotare la grupuri nominale

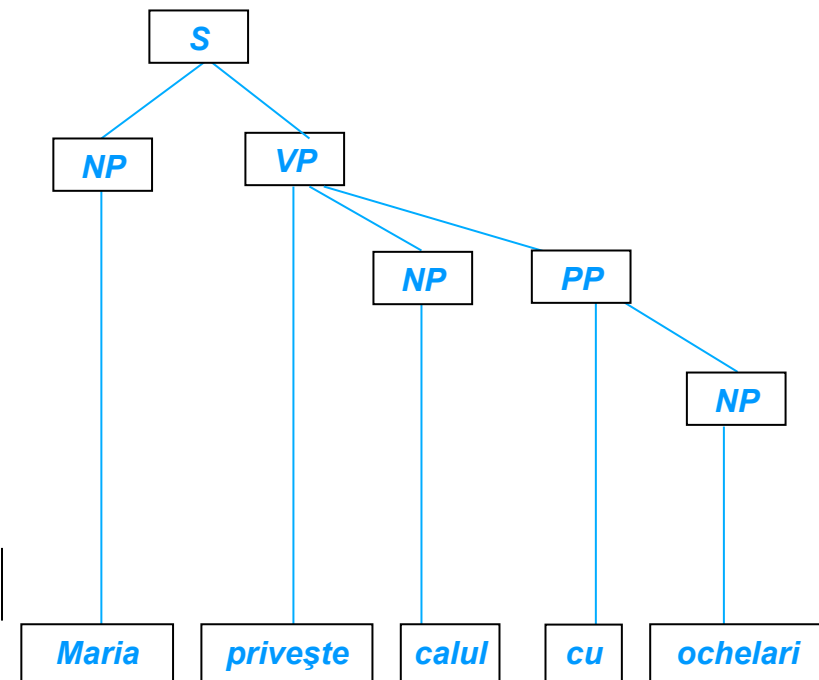
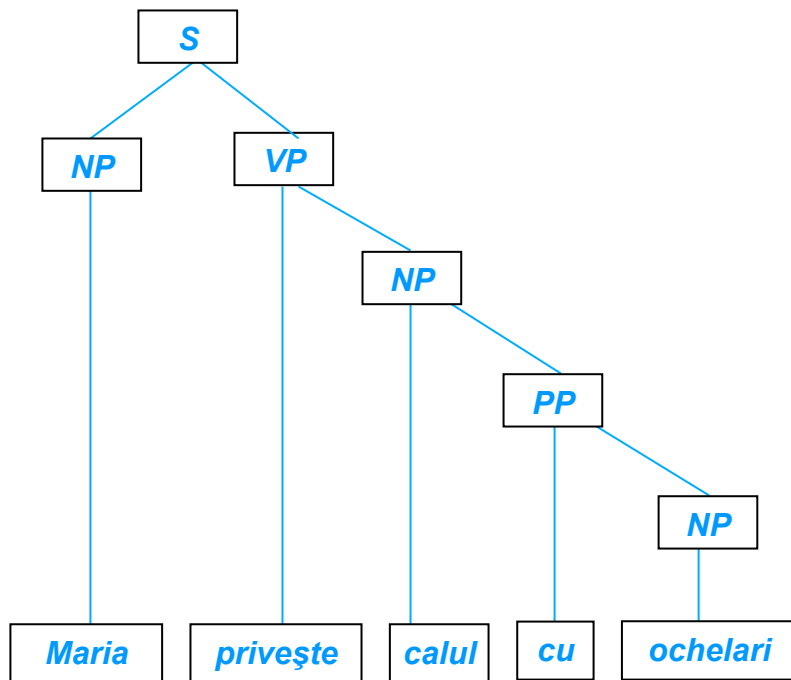
```
<NP ID="NP903" HEADID="W3190" VERBPOS="W3191">
  <W ID="W3190" POS="PRON" NUM="SG" GENDER="M" ROLE="SUBJ"      LEMMA="he"
    LINK="W3191" LINKTYPE="subj">He</W> </NP>
  <W ID="W3191" POS="V" ROLE="+FAUXV" LEMMA="do" LINK="W3193" LINKTYPE="v-ch">did</
    W>
  <W ID="W3192" POS="NEG-PART" ROLE="ADVL" LEMMA="not" LINK="W3191"
    LINKTYPE="neg">not</W>
  <W ID="W3193" POS="V" ROLE="-FMAINV" LEMMA="know" LINK="W3189"
    LINKTYPE="main">know</W>
<NP ID="NP1188" HEADID="W3195">
  <NP ID="NP904" HEADID="W3194" VERBPOS="W3189">
    <W ID="W3194" POS="PRON" NUM="SG" GENDER="F"                ROLE="OBJ"
      LEMMA="she" LINK="W3195" LINKTYPE="subj">her</W> </NP>
    <W ID="W3195" POS="V" ROLE="-FMAINV" LEMMA="name" LINK="W3193"
      LINKTYPE="obj">name</W> </NP>
```

# Tehnologiile limbajului scris

- Analiza și înțelegerea limbajului
  - prelucrări sintactice
    - formalisme gramaticale
    - parsarea → structura sintactică a frazei

# Ambiguități sintactice

*Maria privește calul cu ochelari.*

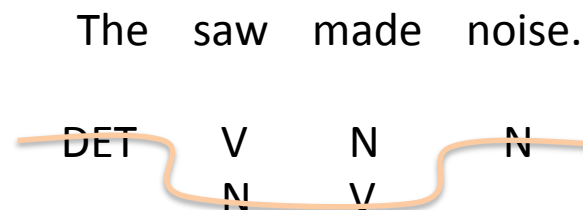


# Instrumente de bază în PLN

- **Tokenizer:** determină granițele unităților lexicale
  - intrare: text (șir de caractere)
  - ieșire: `<tok id="...">cuvânt</tok>`
  - cum: prin expresii regulate

# Instrumente de bază în PLN


- **POS-Tagger**: etichetare la parte de vorbire (dezambiguizare morfosintactică)
  - intrare: `<tok id="...">cuvânt</tok>`
  - ieșire: `<tok id="..." POS="...">cuvânt</tok>`
  - cum: exploatând frecvențele de apariție a anumitor secvențe de părți de vorbire => optimizare globală a secvențelor de etichete



# Instrumente de bază în PLN

- **Lematizator:** determină forma de bază a cuvintelor
  - intrare: `<tok id="..." POS="...">word</tok>`
  - ieșire: `<tok id="..." POS="..." lemma="...">word</tok>`
  - cum: pe baza unui dicționar de leme și exploatând frecvențe de apariție a secvențelor de leme => optimizare globală

The saw made noise.



the saw made noise

the see make noise



# Instrumente de bază în PLN

- **NP-Chunker**: detectează grupuri nominale
  - intrare: secvențe de elemente <tok>
  - ieșire: <np id="...">...</np>
  - cum: aplicând expresii regulate

# Instrumente de bază în PLN

- **NER (name entity recogniser):** recunoaște și clasifică nume de entități
  - intrare: text
  - ieșire: `<ne id="..." type = "...">...</ne>`
  - cum: pe bază de expresii regulate și liste foarte mari de nume de entități specializate pe limbi (gazeteers)

# Tehnologiile limbajului scris

- Analiza și înțelegerea limbajului
  - Prelucrări semantice și de discurs
    - dezambiguizare semantică ➔ sensurile cuvintelor
    - determinarea rolurilor semantice ale verbelor
    - structura retorică a discursului și dialogului
    - rezoluția anaforelor
    - rezumarea textelor

# Lanțuri coreferențiale

Winston was just taking his place in one of the middle rows when two people whom he knew by sight, but had never spoken to, came unexpectedly into the room. One of them was a girl whom he often passed in the corridors. He did not know her name, but he knew that she worked in the Fiction Department.

# Lanțuri coreferențiale

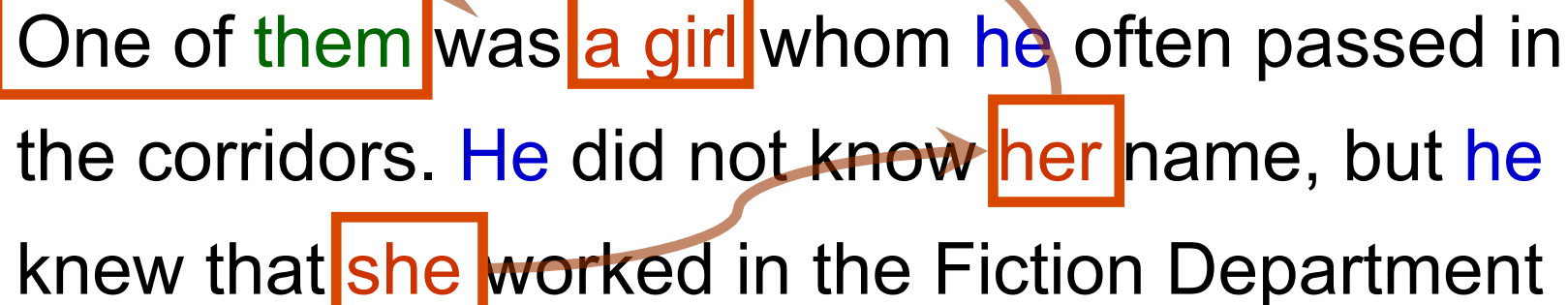
Winston was just taking his place in one of the middle rows when two people whom he knew by sight, but had never spoken to, came unexpectedly into the room.

One of them was a girl whom he often passed in the corridors. He did not know her name, but he knew that she worked in the Fiction Department.

# Lanțuri coreferențiale

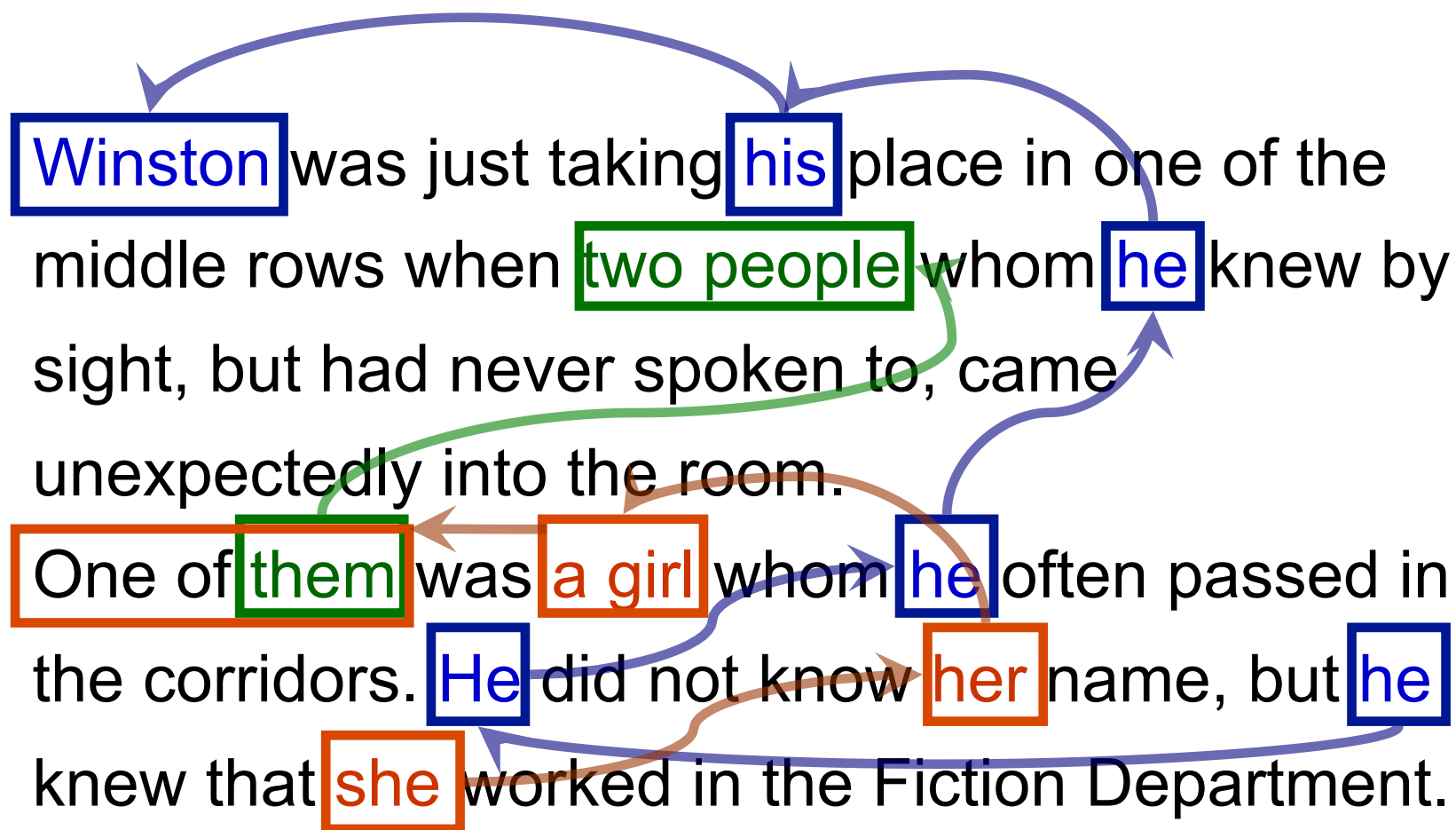
Winston was just taking his place in one of the middle rows when two people whom he knew by sight, but had never spoken to, came unexpectedly into the room.

One of them was a girl whom he often passed in the corridors. He did not know her name, but he knew that she worked in the Fiction Department



The diagram illustrates coreference chains in the second paragraph. It features three orange boxes: one around 'them', one around 'a girl', and one around 'her'. A curved arrow points from 'them' to 'a girl', and another curved arrow points from 'her' to 'a girl'. Additionally, a straight arrow points from 'he' in 'He did not know her name' to 'he' in 'he often passed in the corridors'.

# Lanțuri coreferențiale

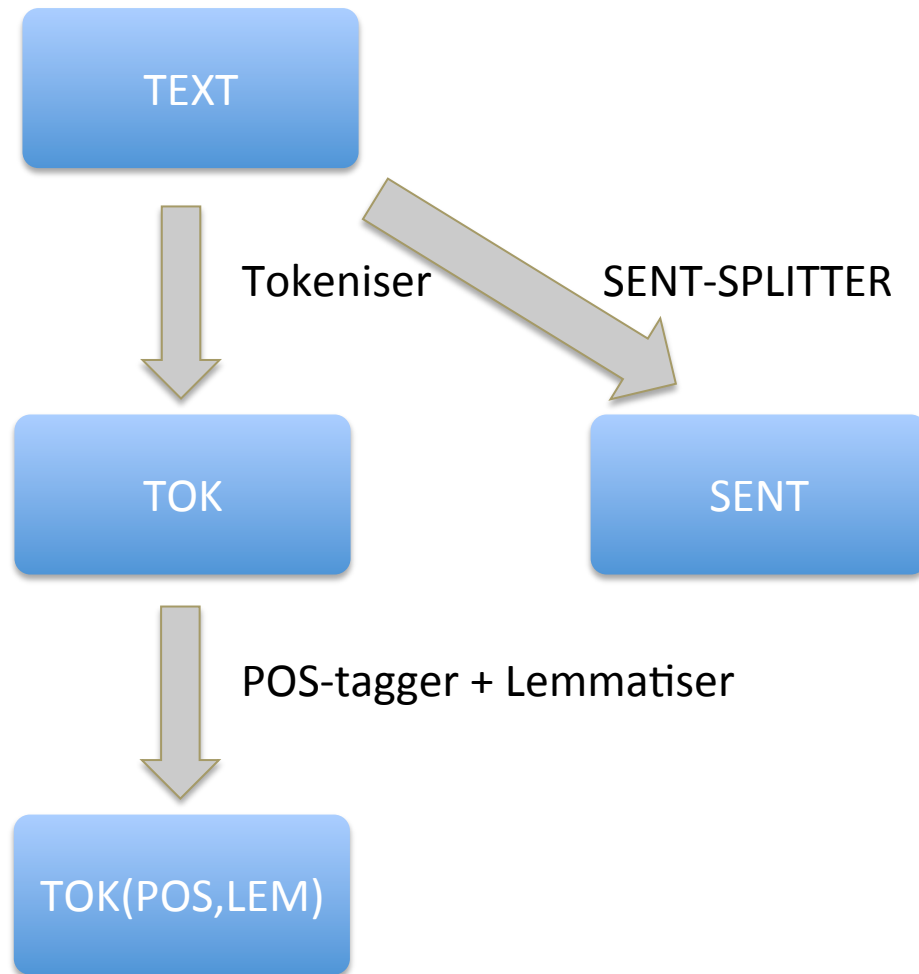


## Cuvintele își precizează sensul în context

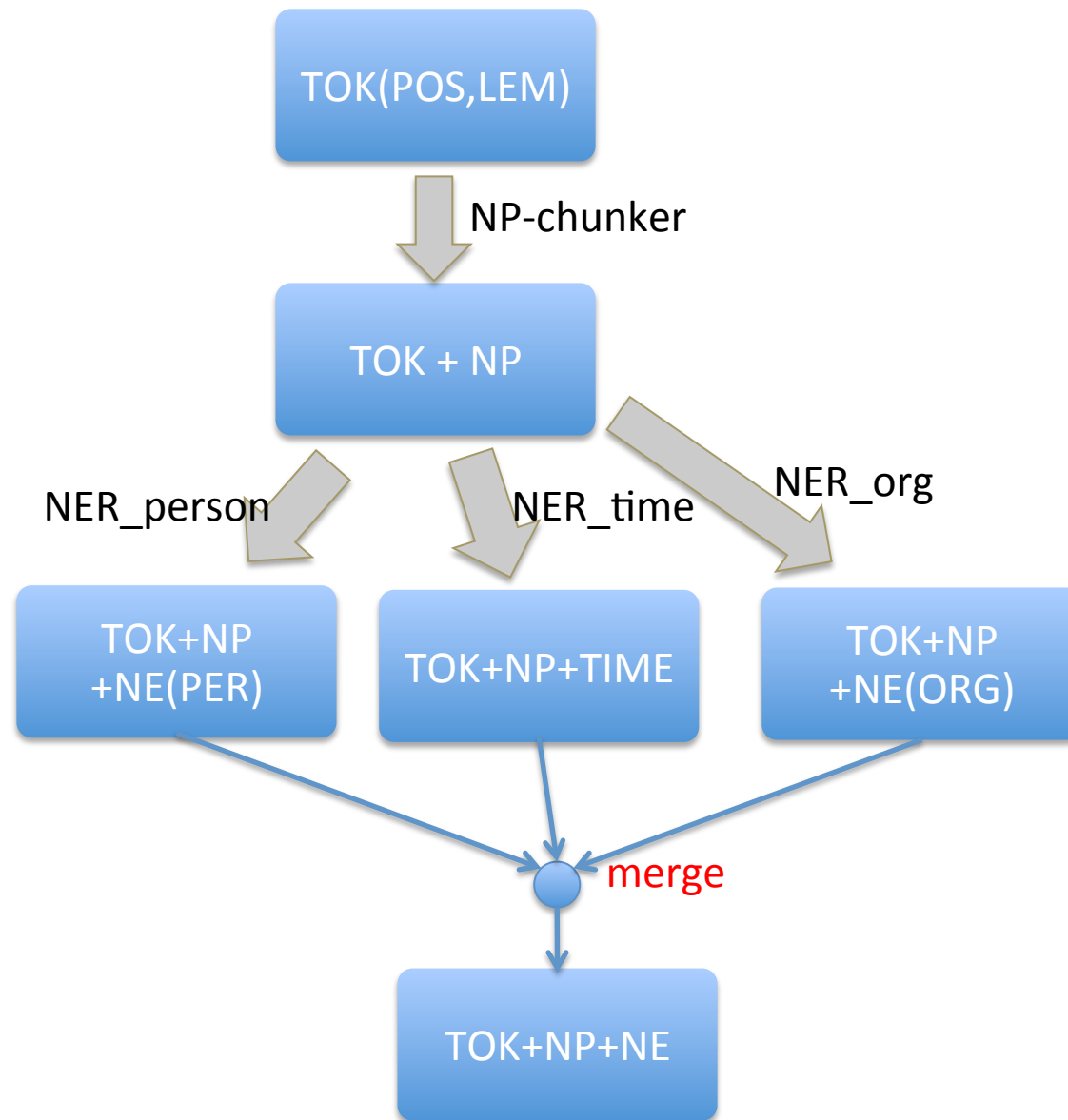
- *Ion se prinse în horă cu o fată cu cosițe lungi.*
- *Când fată iapa ta?*
- *Mă mai dau o dată pe pârtia roșie.*
- *I-am dat una peste mână.*
- *Maria a dat cartea înapoi.*
- *M-am scos...*
- *Mi-am scos măseaua de minte.*



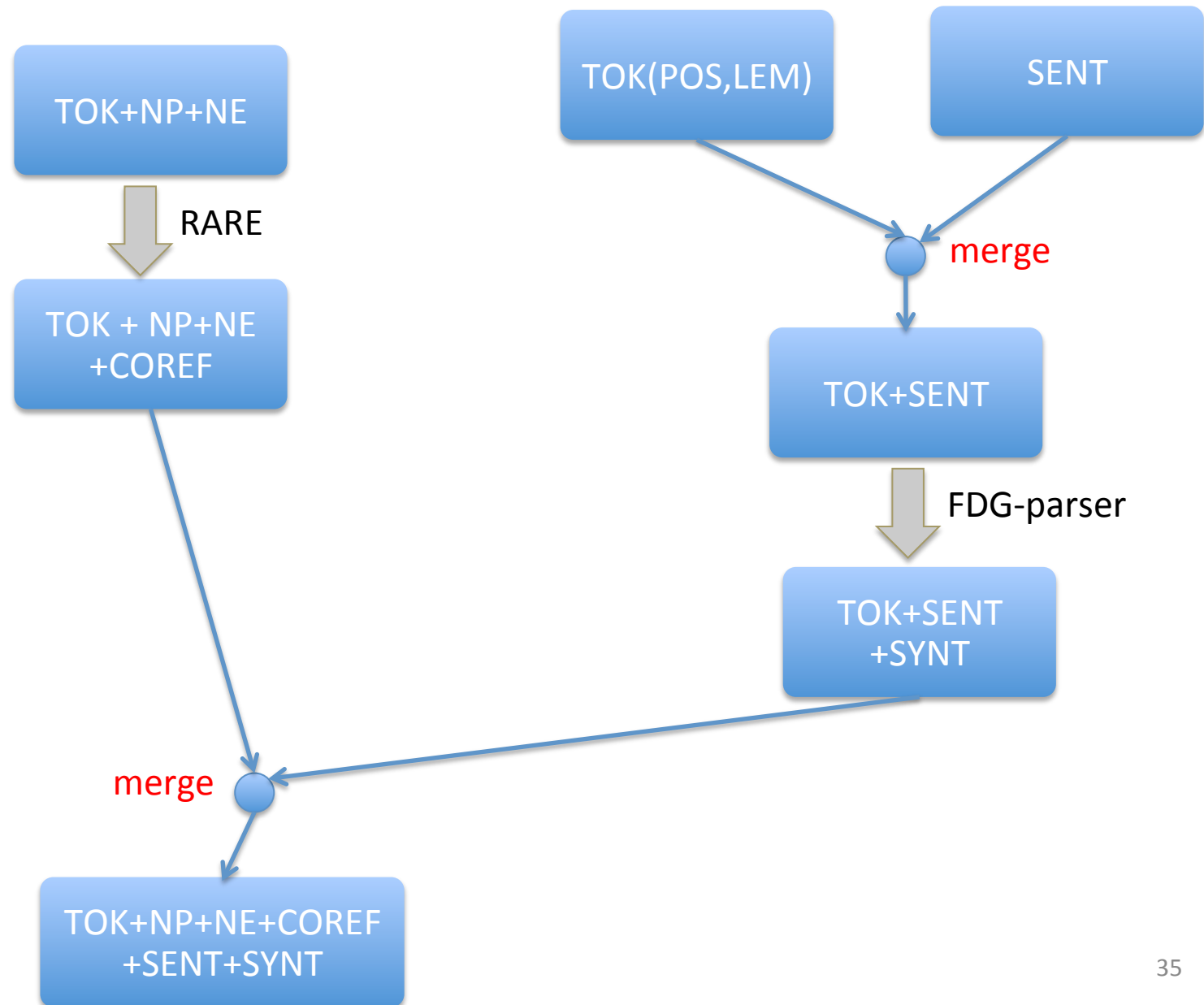
# Pre-processing



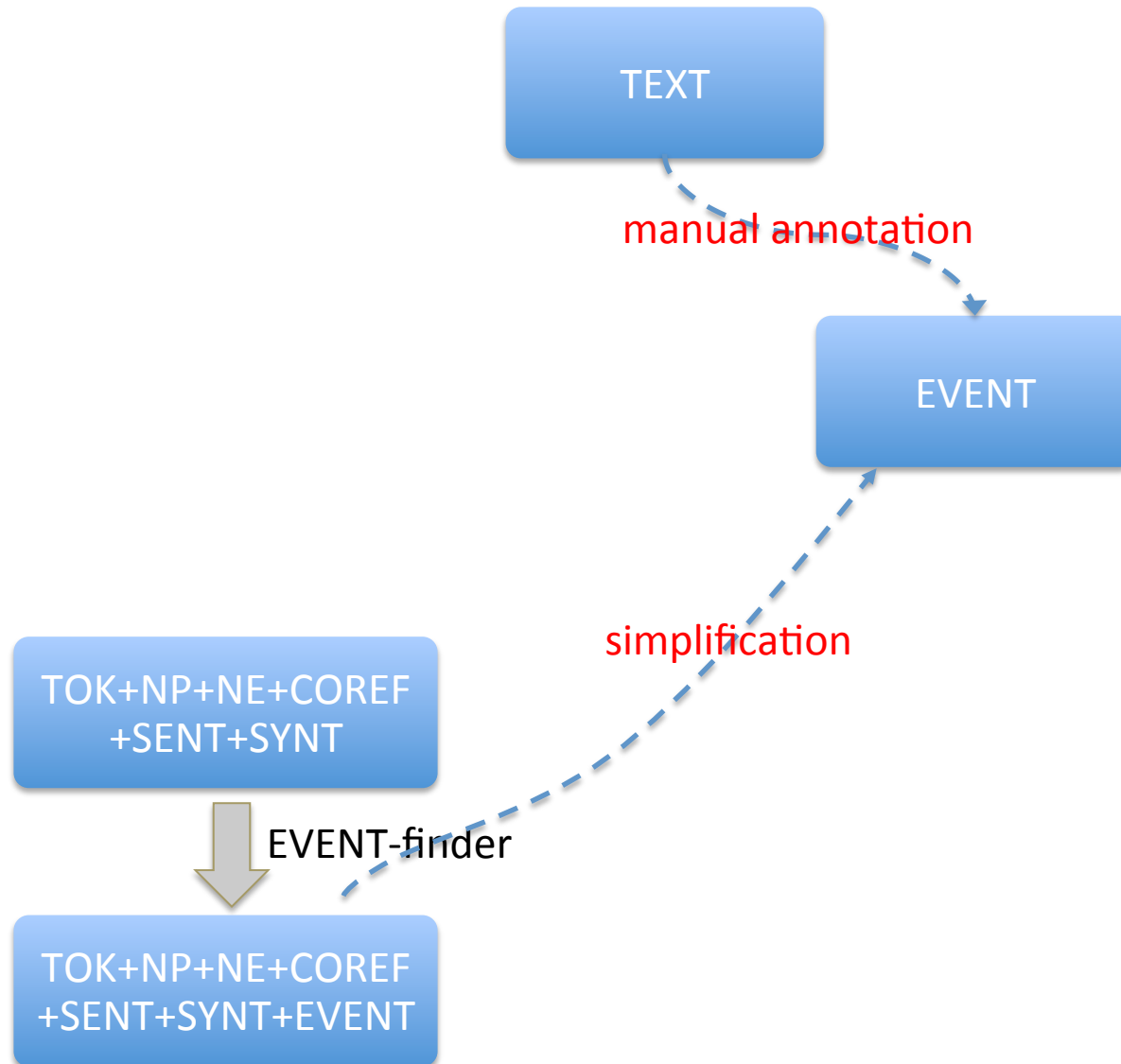
# NP-chunking, NER



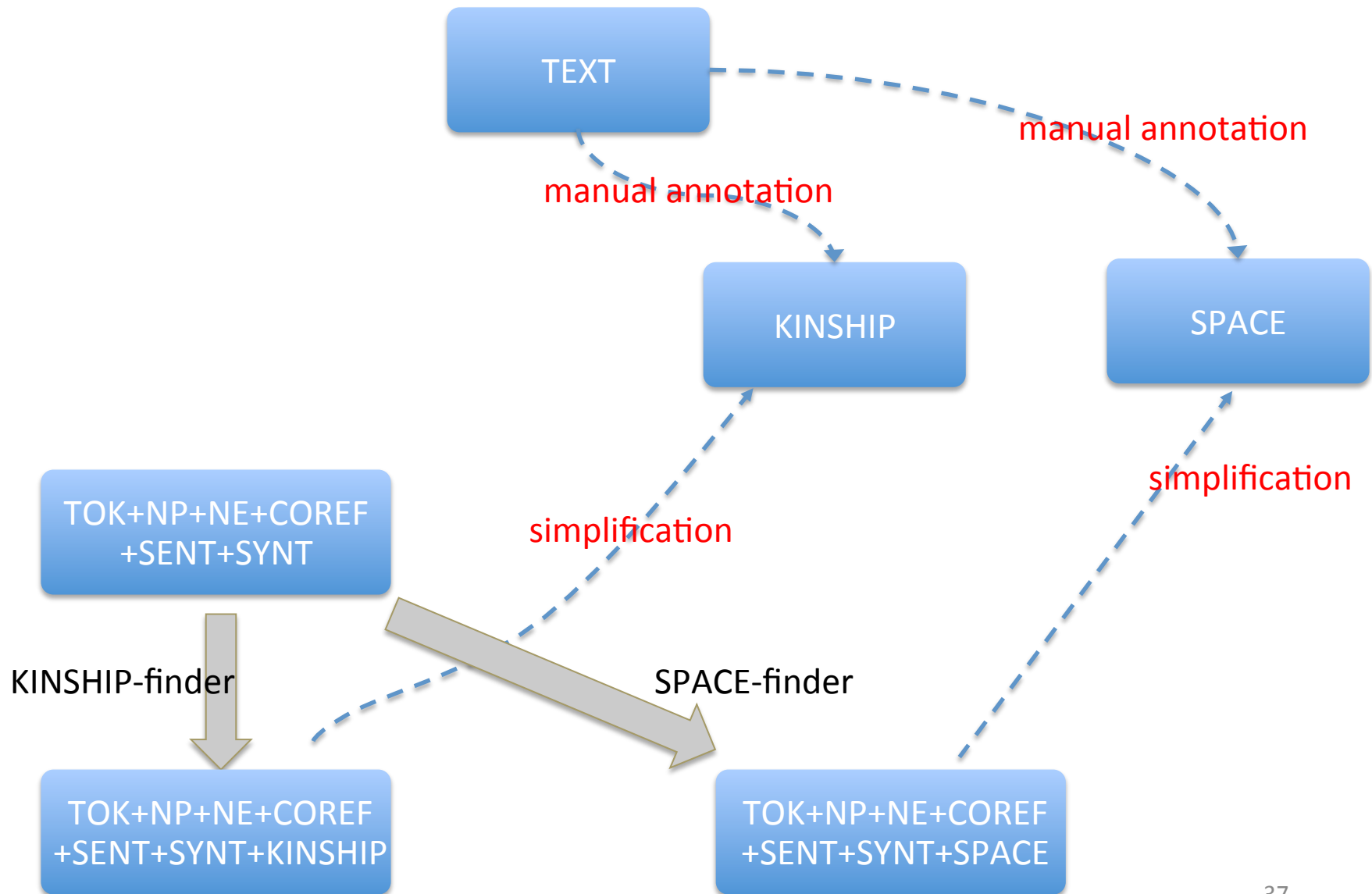
# Coreference, syntactic parsing



# Events



# Kinship and Space



# Relații de rudenie: exemplu

- *Las că cu tine mă răfuiesc după, îi scăpă printre dinți omului ei **Donca, nevasta călugărului zbanghiu Zuicu**, care-l adusesese la el acasă pe Ion și pe președinte.*

Apoziție: Per-X, Rel (atrib) Per-Y<sub>gen</sub>, =>

marriage(X:person[sex:?], Y:person[sex:?])

marriage(Donca:person[sex:f], Zuicu:person[sex:m])

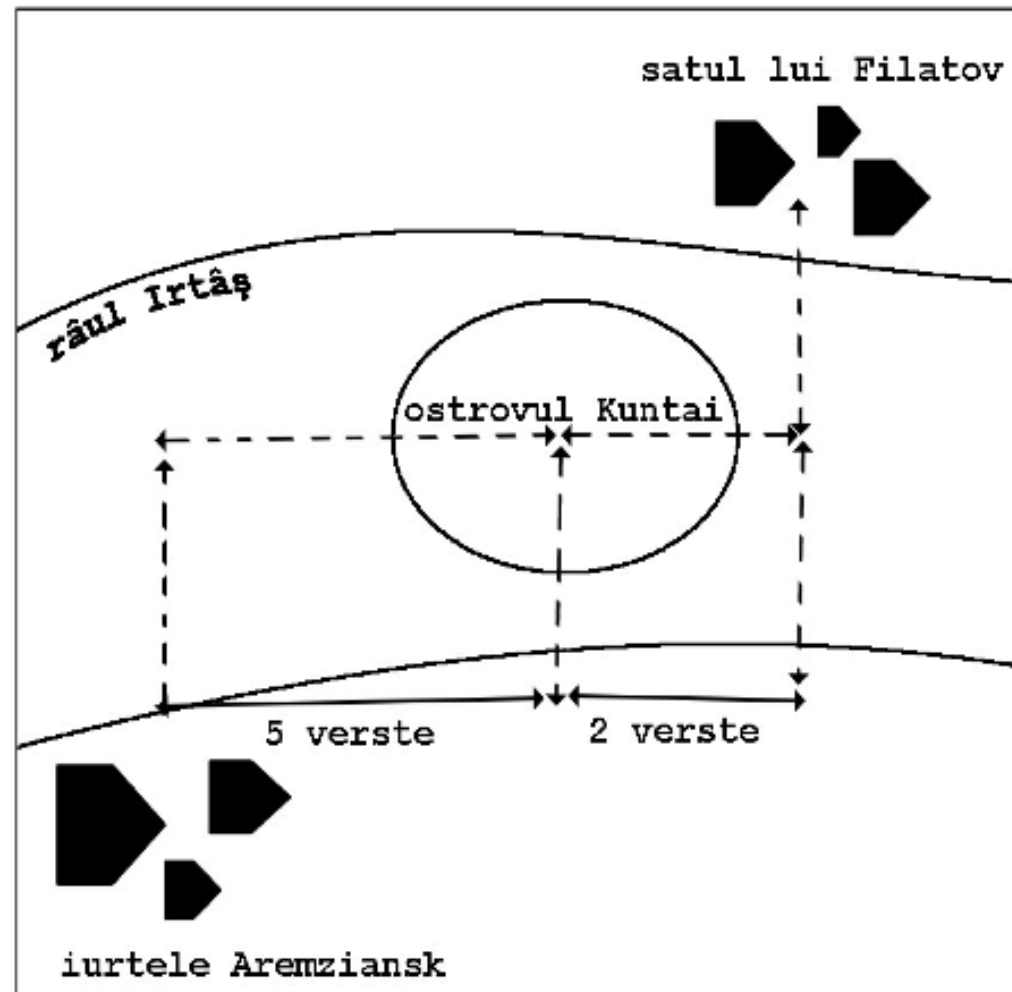
# Relații de rudenie: exemplu

- *Vreme de patruzeci de ani viața **Ellei Rubinstein<sub>1</sub>** fusese ca o apă stătătoare... **Soțul ei<sub>1</sub>, David**, era un dentist de succes...*

Apoziție: Rel Per- $X_{\text{pron,gen}}$ , Per-Y, =>  
marriage(antecedent(X):*person*[sex:?],  
Y:*person*[sex:?])  
marriage(Ella Rubinstein:*person*[sex:f],  
David:*person*[sex:m])

# Relații spațiale: exemplu

*La cinci verste de iurtele Aremziansk, în mijlocul râului Irtâș, se află ostrovul Kuntai. Satul lui Filatov se află pe malul stâng la două verste de ostrov.*





# Events happen in time

*Când a intrat în cameră<sub>e1</sub>, Ion a aprins lumina<sub>e2</sub>.*

*După cinci minute a ieșit<sub>e3</sub>. La ieșire a stins  
lumina<sub>e4</sub>.*

# Events happen in time

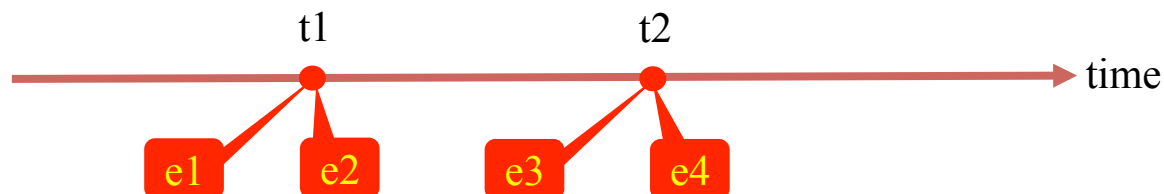
*Când a intrat în cameră<sub>e1</sub>, Ion a aprins lumina<sub>e2</sub>.*

*După cinci minute a ieșit<sub>e3</sub>. La ieșire a stins lumina<sub>e4</sub>.*

Two types of temporal expressions:

- instants...

$e1:t1 / e2:t1 / e3:t2=t1+5\text{min} / e4:t2$



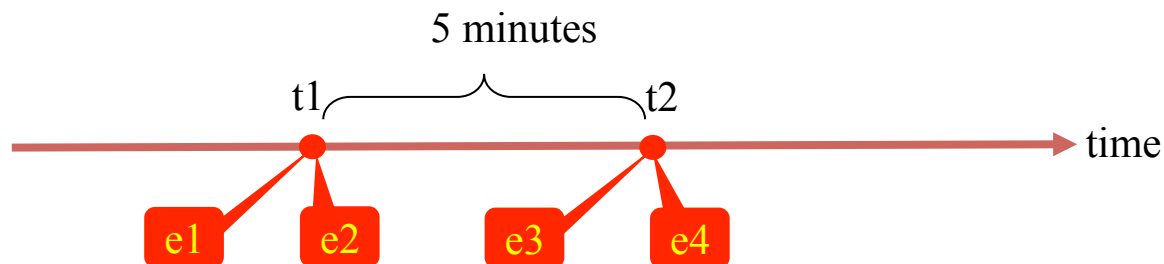
# Events happen in time

*Când a intrat în cameră<sub>e1</sub>, Ion a aprins lumina<sub>e2</sub>.*

*După **cinci minute** a ieșit<sub>e3</sub>. La ieșire a stins  
lumina<sub>e4</sub>.*

Two types of temporal expressions:

- ...and intervals:

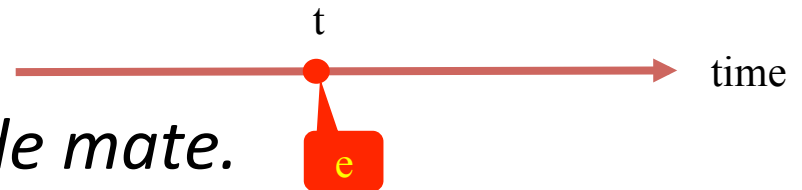


# Events can be...

- Instantaneous:

*Ion a ieșit din cameră.*

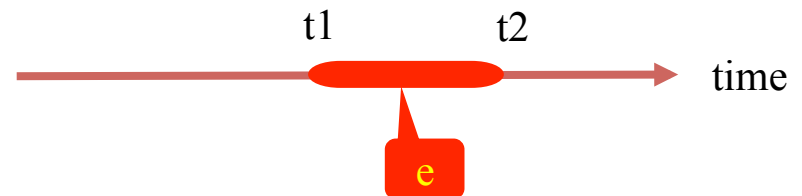
*Maria s-a întâlnit cu proful de mate.*



- Take time:

*Ion a citit toată seara.*

*Afară plouă.*



# Signals for temporal relations

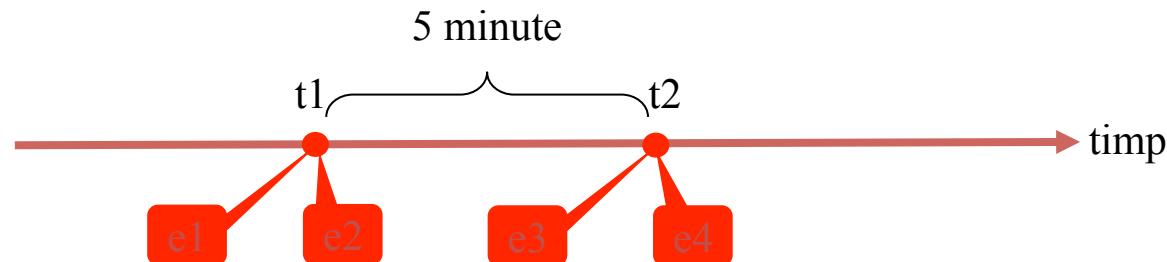
*Când* a intrat în cameră<sub>e1</sub>, Ion a aprins lumina<sub>e2</sub>.

*După cinci minute* a ieșit<sub>e3</sub>. *La ieșire* a stins  
lumina<sub>e4</sub>.

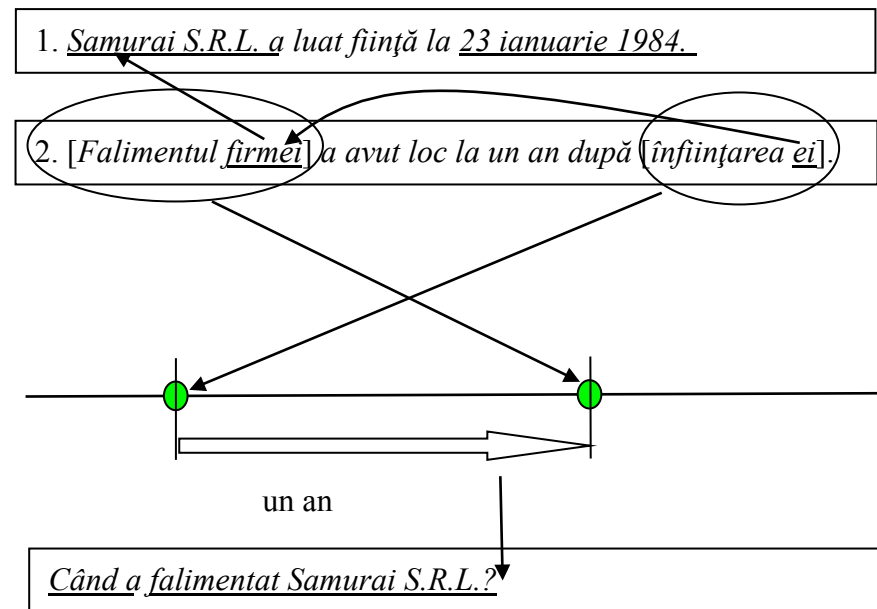
când  $e_i, e_j \rightarrow t(e_i) = t(e_j)$

$e_i$ . După  $\langle \text{interval} \rangle e_j \rightarrow t(e_j) = t(e_i) + \langle \text{interval} \rangle$

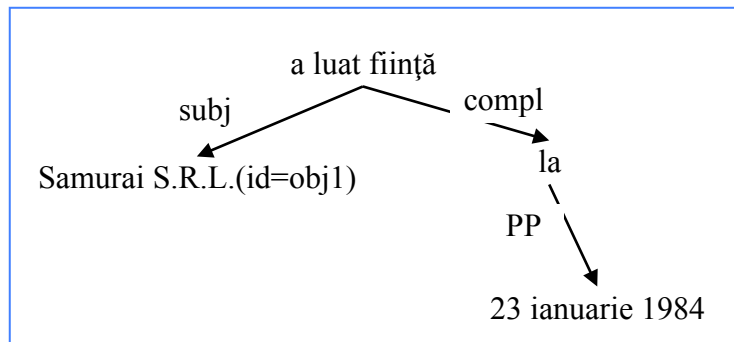
La  $\langle \text{reference}(e_i) \rangle e_j \rightarrow t(e_i) = t(e_j)$



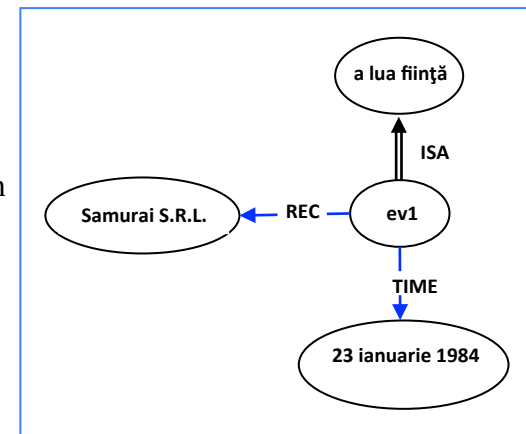
# Type of reasoning where time matters



# Processing statements

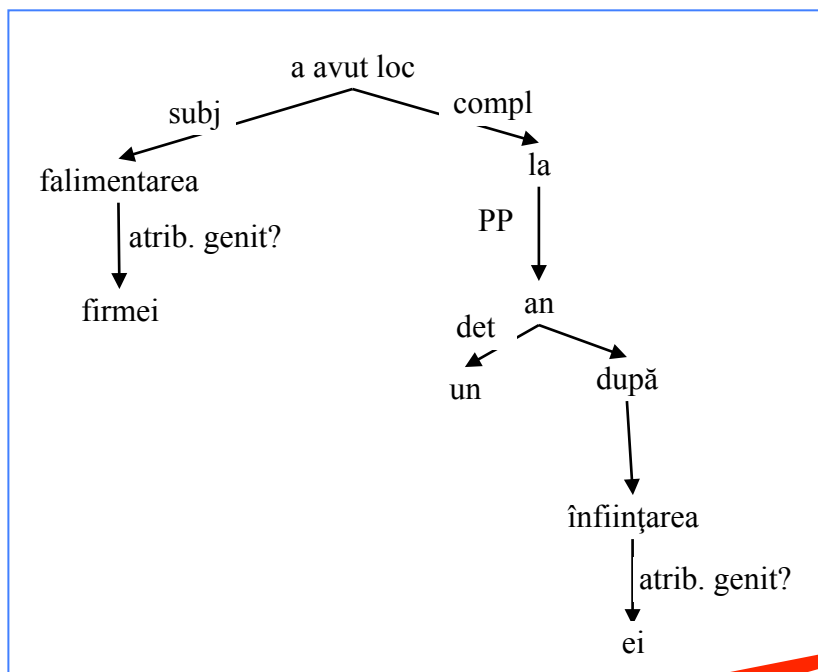


evenimential  
representation

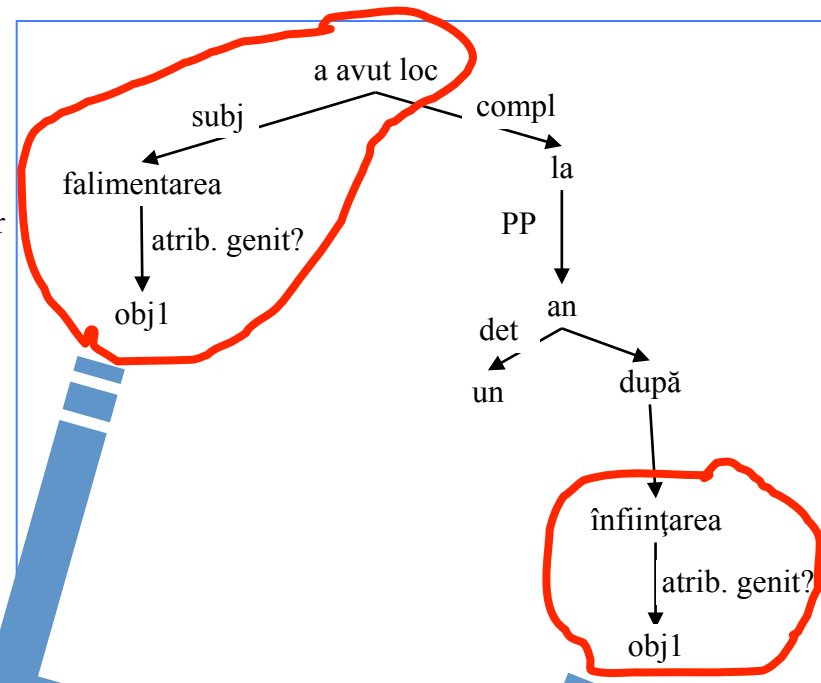


<object ID="obj1" ISA="companie" NAME="Samurai S.R.L."/>

<event ID="ev1" ISA="a\_lua\_ființă" REC="obj1" TIME="23.01.1984"/>



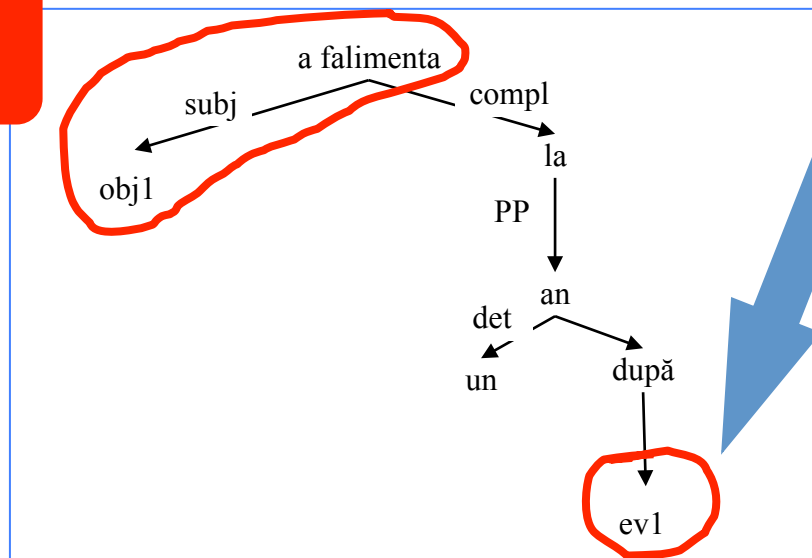
rezoluția  
anaforelor



simplificări



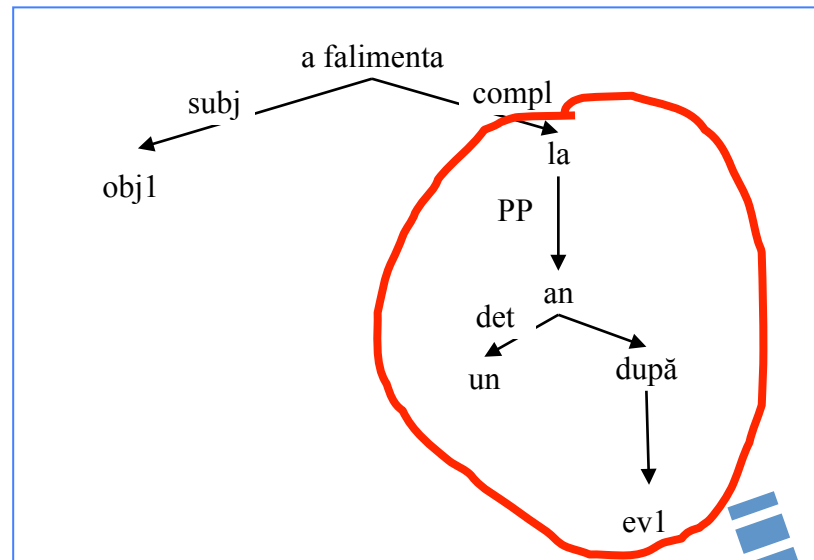
**dacă are\_loc falimentarea lui  
X atunci X falimentează**



**referință  
anaforică la un  
eveniment deja  
menționat**



# Processing statements



**temporal  
expression  
anchored in  
another event**

↓  
evenimential  
representations

<event ID="ev2" ISA="a\_falimenta" REC="obj1" TIME="timex1"/>

<timex ID="timex1" TYPE="after" REF="ev1" DUR="1" UNIT="year"/>

# Computing time

<object ID="obj1" ISA="companie" NAME="Samurai S.R.L."/>

<event ID="ev1" ISA="a\_lua\_fiintă" REC="obj1" TIME="23.01.1984"/>

<event ID="ev2" ISA="a\_falimenta" REC="obj1" TIME="timex1"/>

<timex ID="timex1" TYPE="after" REF="ev1" DUR="1" UNIT="year"/>



<event ID="ev2" ISA="a\_falimenta" REC="obj1" TIME="23.01.1985"/>

# Corpusul 'QuoVadis'



# Un corpus de entități și relații semantice

- Tipuri de entități:
  - persoane
  - zei
  - grupuri de persoane și zei
  - părți fizice
- Relații semantice exprimate între aceste tipuri de entități



# Entități

- Personaje (*Marcus Vinicius, Ligia*), grupuri (*creștinii, soldații*);
- La nivelul textului: grupuri nominale (*tânărul patrician, fiul consulului*);

- Entități incluse:

[*Te*]<sub>1</sub> [*iubesc; REALISATION=INCLUDED*]<sub>2</sub>, *Marcus!*

- Expresii referențiale imbricate:

[*fiica [lui Aulus]*]<sub>2</sub><sub>1</sub>



# Tipuri de relații

- Anaforice
- Semantice
  - rudenie
  - afective
  - sociale



# Relații anaforice

- *coref*
- *coref-interpret*
- *member-of, has-as-member* (inverse)
- *isa, class-of* (inverse)
- *part-of, has-as-part* (inverse)
- *subgroup-of, has-as-subgroup* (inverse)
- *has-name, name-of* (inverse)



1:[Acteea]... 2:[tânăra libertă]... => [2] coref [1]

1:[mâna 2:[lui] dreaptă] => [1] part-of [2]

# Relații de rudenie

- *parent-of*
- *child-of* (inverse of *parent-of*)
- *grandparent-of* and *grandchild-of* (inverse)
- *sibling* (symmetrical)
- *ant-uncle-of, nephew-of* (inverse relation)
- *cousin-of* (symmetrical)
- *spouse-of* (symmetrical)
- *unknown*



1:[celui de-al doilea soț 2:[al Popeii]] => [1] spouse-of [2]

1:[sora lui 2:[Petronius]] => [1] sibling-of [2]



# Relații sociale

- *superior-of*
- *inferior-of*
- *in cooperation-with*
- *colleague-of*
- *in competition-with*
- *opposite-to*



Eliberând- 1:[o], 2:[Nero]... => [2] superior-of [1]

1:[Tânărul] luptase sub comanda 2:[lui Corbulon] =>  
[1] inferior-of [2]

# Relații afective

- *love*
- *loved-by*
- *hate*
- *hated by*
- *upset*
- *friendship*
- *worship*
- *anger*



Pe 1:[Vinicus] îl cuprinse o mânie năprasnică împotriva  
2:[împăratului] și împotriva 3:[Acteii] => [1] anger  
[2], [1] anger [3]

# Adnotarea

<ENTITY ID="E8" TYPE="PERSON">  
<W id="28" LEMMA="Marcus">Marcus</W>  
<W id="29" LEMMA="Vinicius">Vinicius</W>  
</ENTITY>  
<W id="30" LEMMA="fi">era</W>  
<KINSHIP ID="KIN57" FROM="E12" TO="E11" TRIGGER="31"  
TYPE="child-of">  
<ENTITY ID="E12" TYPE="PERSON">  
<W id="31" LEMMA="fiu">fiu</W>  
<KINSHIP ID="KIN53" FROM="E11" TO="E10" TRIGGER="32"  
TYPE="sibling-of">  
<ENTITY ID="E11" TYPE="PERSON">  
<W id="32" LEMMA="soră">surorii</W>  
<ENTITY ID="E10" TYPE="PERSON">  
<W id="33" LEMMA="său">sale</W>  
</ENTITY>  
<W id="34" LEMMA="mai">mai</W>  
<W id="35" LEMMA="mare">mari</W>  
</ENTITY>  
</KINSHIP>  
</ENTITY>  
</KINSHIP>  
<W id="36" LEMMA=",">,</W>  
<KINSHIP ID="KIN59" FROM="E13" TO="E15" TRIGGER="44"  
TYPE="spouse-of">  
<ENTITY ID="E13" TYPE="PERSON">  
<W id="37" LEMMA="care">care</W>  
</ENTITY>  
<W id="38" LEMMA=",">,</W>  
<W id="39" LEMMA="cu">cu</W>  
<W id="40" LEMMA="an">ani</W>  
<W id="41" LEMMA="în\_urmă">în urmă</W>  
<W id="42" LEMMA=",">,</W>  
<W id="43" LEMMA="sine">se</W>

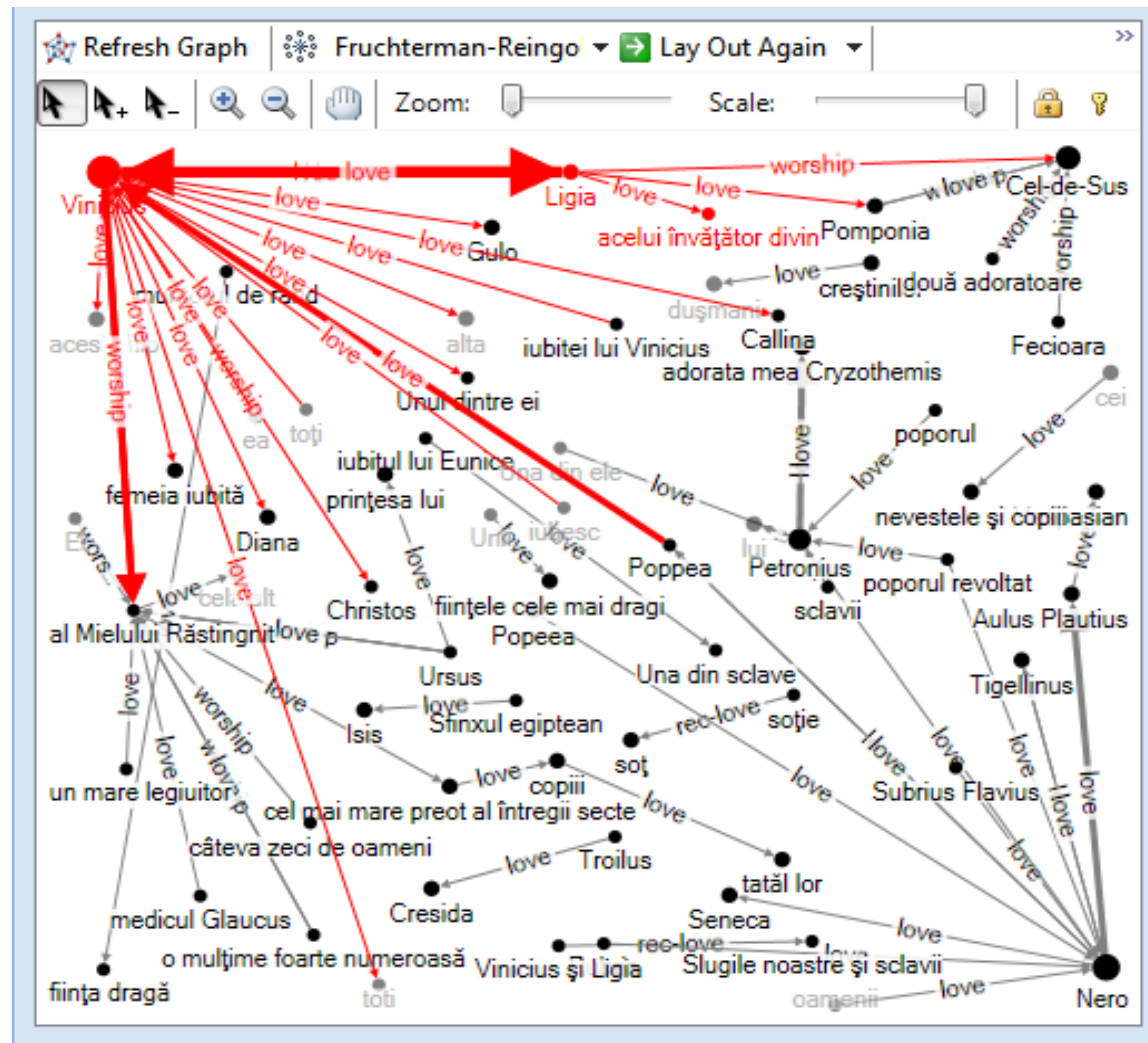
<W id="44" LEMMA="căsători">căsătorise</W>  
<W id="45" LEMMA="cu">cu</W>  
<KINSHIP ID="KIN61" FROM="E15" TO="E14" TRIGGER="46"  
TYPE="parent-of">  
<ENTITY ID="E15" TYPE="PERSON">  
<W id="46" LEMMA="tată">tată</W>  
<ENTITY ID="E14" TYPE="PERSON">  
<W id="47" LEMMA="acesta">acestui</W>  
</ENTITY>  
</ENTITY>  
</KINSHIP>  
</KINSHIP>  
<SOCIAL ID="SOC9" FROM="E17" TO="E16" TRIGGER="49"  
TYPE="inferior-of">  
<ENTITY ID="E17" TYPE="PERSON">  
<W id="49" LEMMA="consul">consul</W>  
<W id="50" LEMMA="pe">pe</W>  
<W id="51" LEMMA="vreme">vremea</W>  
<W id="52" LEMMA="el">lui</W>  
<ENTITY ID="E16" TYPE="PERSON">  
<W id="53" LEMMA="Tiberiu">Tiberiu</W>  
</ENTITY>  
</ENTITY>  
</SOCIAL>  
<W id="54" LEMMA=".">.</W>  
  
<REFERENTIAL ID="REF37" FROM="E12" TO="E8" TYPE="coref" /  
REFERENTIAL>  
<REFERENTIAL ID="REF38" FROM="E13" TO="E11" TYPE="coref" /  
REFERENTIAL>  
<REFERENTIAL ID="REF39" FROM="E14" TO="E8" TYPE="coref" /  
REFERENTIAL>  
<REFERENTIAL ID="REF40" FROM="E17" TO="E15" TYPE="class-  
of" /REFERENTIAL>

# Statistici asupra corpusului

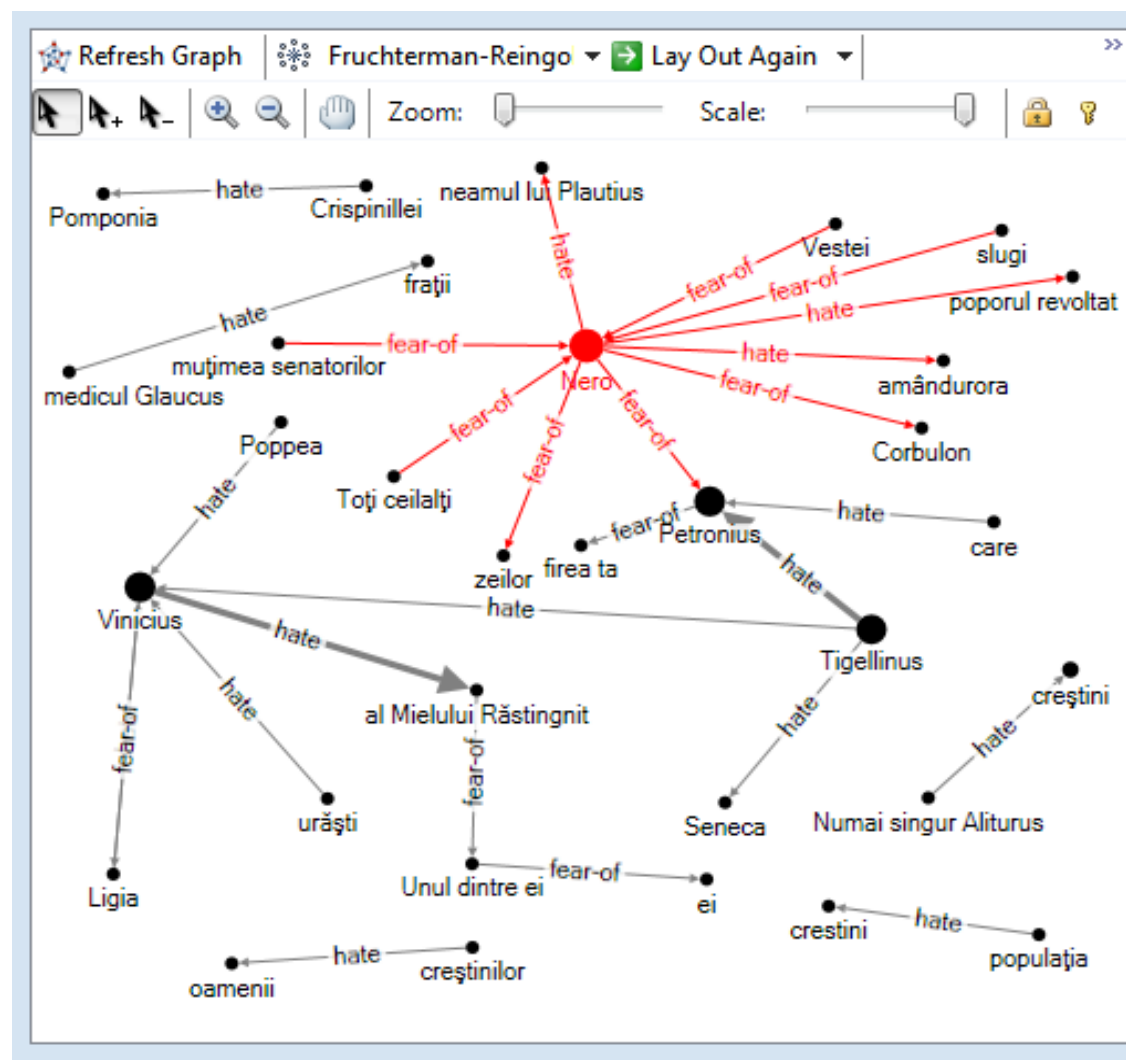
- 7.281 fraze
- 146.822 cuvinte și semne de punctuație
- 24.636 mențiuni de entități
- 22.301 relații referențiale
- 755 relații AKS (**A**ffective + **K**inship + **S**ocial)
- 752 triggere



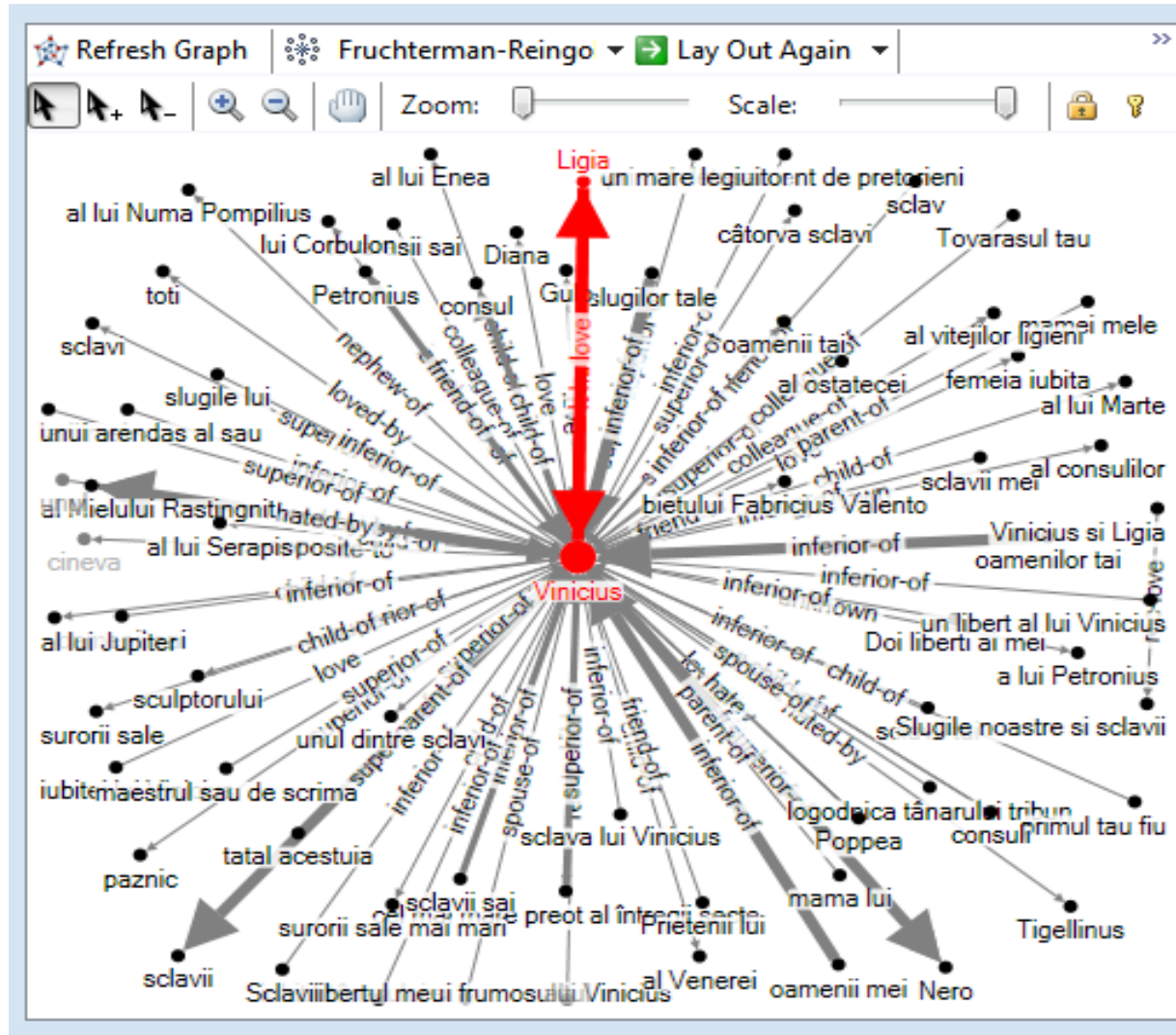
## Example: relațiile *love* și *worship*



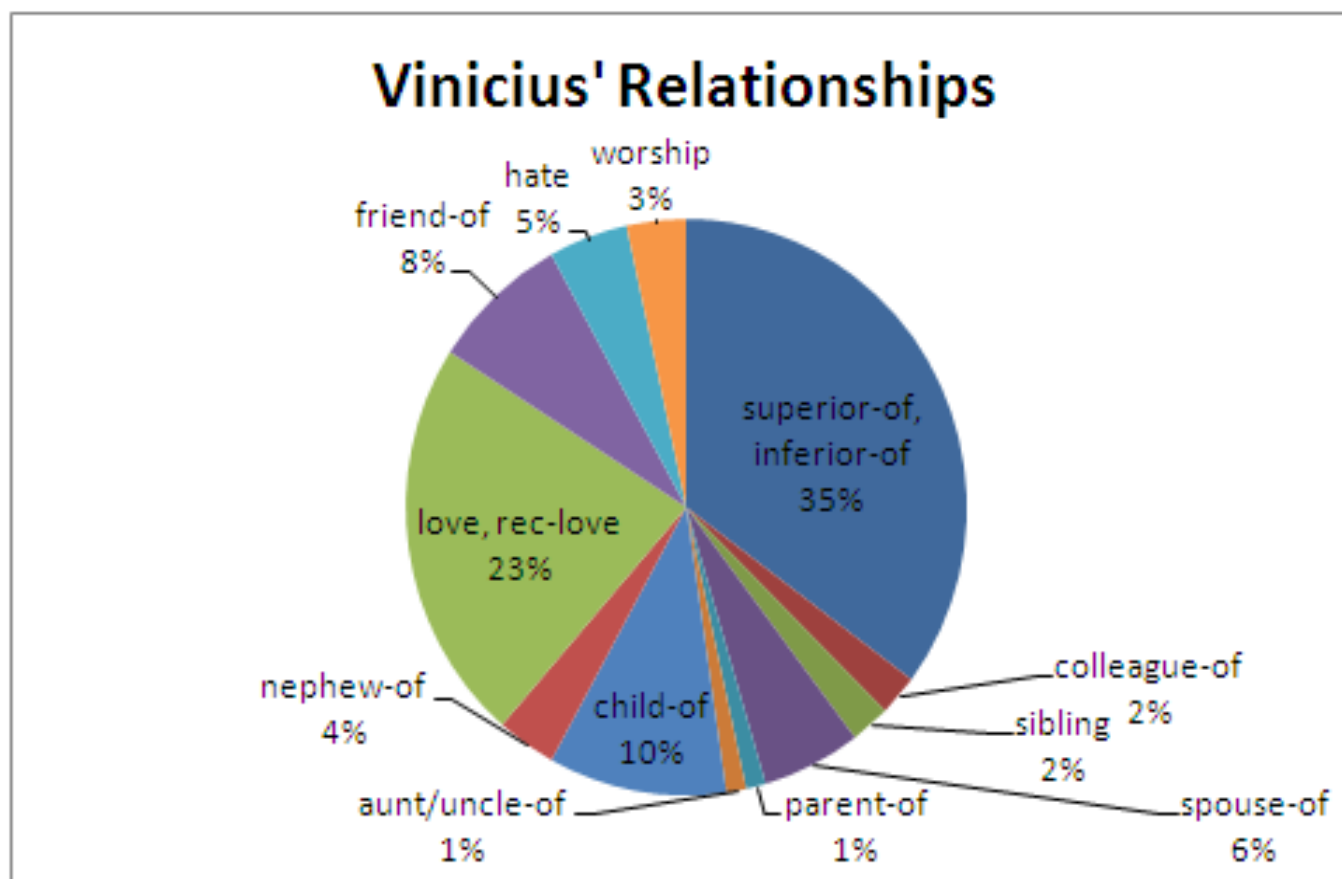
# Relațiile afective *fear-of* și *hate*



# Relațiile lui Vinicius cu alte personaje



# Distribuția relațiilor semantice în care este implicat personajul Vinicius





# *Linguistics Linked Open Data (LLOD)*

- Dezvoltarea de tehnici care vor permite **descifrarea conținutului semantic al textelor**
  - rezumate (generale, parțiale, focalizate pe personaje),
  - linii narative (e.g. evoluția sentimentelor dintre Vinicius și Ligia)
  - conexiuni statice între entități (e.g. arbori genealogici),
  - statistici asupra entităților (e.g. sentimentele majoritare ale creștinilor comparate cu cele ale romanilor)

# *Linguistics Linked Open Data (LLOD)*

- Generarea de ontologii din colecții de tratate
  - aplicații care “citesc” tratatele unui domeniu și formalizează conceptele și instanțele acestora
- Căutare documentară inteligentă
  - asistenți personalizați ai activității de cercetare