

PIANO
TRENTINO
TRILINGUE

8 9 10
MARZO
2018

FESTIVAL DELLE LINGUE ROVERETO

150 WORKSHOP
80 ENTI PARTECIPANTI



Linguistica computazionale e lingue classiche

Ovvero cosa c'entra Tommaso d'Aquino con
Google

Marco Passarotti
Centro di ricerca CIRCSE
Università Cattolica del Sacro Cuore (Milano)

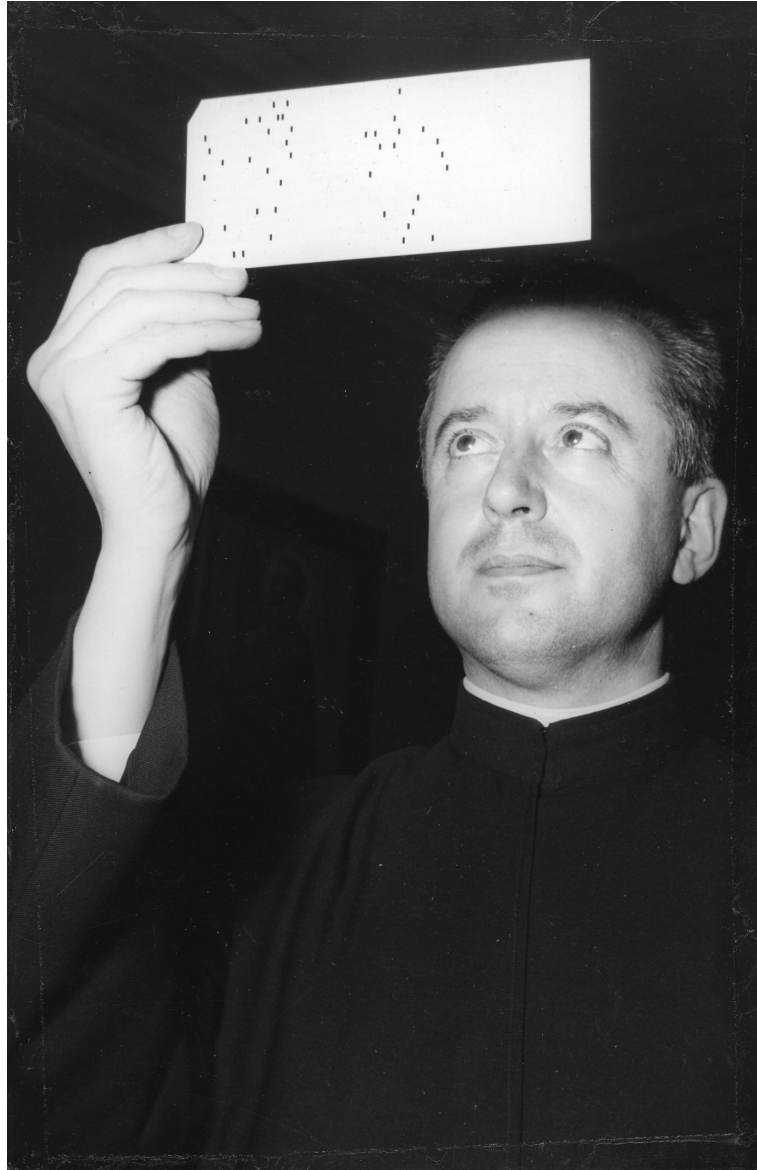
Cosa è (e cosa non è) la linguistica computazionale

- La linguistica computazionale si concentra sullo sviluppo di formalismi descrittivi del funzionamento di una lingua naturale, tali che si possano trasformare in programmi eseguibili dai computer (*Wikipedia*)
- Da non confondersi con
 - informatica umanistica
 - Digital Humanities

Cosa fa la linguistica computazionale

- sviluppa formalismi descrittivi del funzionamento del linguaggio naturale, trasferibili in programmi eseguibili dai computer per fini di TAL (NLP)
- applica strumenti di TAL per interrogare e analizzare grandi masse di dati linguistici (scritti e orali; audio e video)
- importanti ricadute applicative: T9, Google (semantic web), speech-to-text, ...

padre Roberto Busa SJ



Leggiamo Busa

L'Analisi linguistica nell'evoluzione mondiale dei mezzi d'informazione

in "Almanacco Letterario Bompiani 1962"

Milano 1962, pp. 103-107.

L'avvento dell'automazione

Un mostro per l'umanesimo

[...] è intervenuto il mostro della notte, il tecnicismo trionfante, con la sua ultima creazione: l'automazione.

*Qualcuno ha rabbrivito, pensandola come un crudo e duro bulldozer che procede ruggendo, schiacciando e stracciando i fiori. Tra questi, **vittima** delicata e gentile, l'umanesimo.*



[...] two groups, comparable in intelligence, identical in race, not grossly different in social origin, earning about the same incomes, who ha[ve] almost ceased to communicate at all. (Snow, Two Cultures, 1959)

Che tipo di mostro?

Nuove domande per gli umanisti

[...] il domani è già qui: il futuro è già cominciato [...] gli uomini dell'automazione hanno cominciato a [...] rivolgere ai filologi e ai grammatici, occupati nei campi a scegliere fior da fiore, domande di questa natura:

- *quanti sono in russo i verbi attivi transitivi e quanti quelli attivi intransitivi? quanti sono in inglese? [...] → Treebanks*
- *la mi vuol raggruppare tutte le parole del vocabolario secondo le varie categorie morfologiche e grammaticali? → PoS-tagging & Morphological-tagging*
- *mi dica tutte le parole che si possono omettere, e quando, così da accorciare un testo senza scapito della sua espressività → Text Summarization*

Troppo poco umanesimo!

È successo un fatto clamoroso: la macchina che ci ha resi consapevoli che nessun umanista possiede la sua lingua così da saper dare una risposta a simili domande.

La macchina [...] ha documentato che di umanesimo, di quello serio e sistematico, ce n'è ancora troppo poco. I fatti economici esigono oggi un incremento qualitativo delle scienze grammaticali e lessicali: come una necessità del loro sviluppo vitale.

Grandi speranze

MARCH 28, 1954

15

CYBERNETICS - A NEW SCIENCE

VISITING JESUIT LECTURES AT ROYAL UNIVERSITY

FATHER Robert Busa, S.J., Professor of Philosophy at the Pontifical Faculty of Philosophy "Aloisianum," Gallarate, Italy, lectured twice in the Aula Magna of the Royal University during a short visit to Malta which came to an end on Friday.

THE SUNDAY TIMES OF MALTA, JULY 29, 1956 11

ELECTRONIC BRAIN WILL "VIVISECT" THOMISTIC TEXT

From A Correspondent

FATHER Roberto Busa, the Italian Jesuit whose connections with Malta consist in his being the nephew of Captain and Mrs. O.F. Gollcher of Mdina, and his visit to and lectures in Malta in March 1954, is in the news again.

As máquinas
da "era electrónica"
ao serviço da cultura
humanística



O PADRE ROBERTO BUSA, inventor de um novo processo de investigação erudita, passa amanhã em Lisboa

Qualche immagine (1)

L 20 61 CIBUS **CIBUS** **358 033 041**

OPUSCULUM HYMNUS STROPHA VERSUS

V E R S U S

SCHEDA PRIMA

OPUSCULUM HYMNUS STROPHA VERSUS

V E R S U S

SCHEDA VERSUUM

VERBUM

LEMMA

VERBUM

OPUSCULUM HYMNUS STROPHA VERSUS

V E R B U M

SCHEDA VERBORUM

LEMMA

VERBUM

OPUSCULUM HYMNUS STROPHA VERSUS

V E R B U M

SCHEDA VERBORUM A CONTRARIO ORD.

VERBORUM A CONTRARIO ORD.

LEMMA

VERBUM

ANALISI FILOLOGICA

Qualche immagine (2)



Simulatore: <http://stefansinclair.name/punchcard/>

Qualche immagine (3)



Qualche immagine (4)



Risorse Linguistiche

Corpora

Cosa sono le RL?

“Language resources are collections of data which primarily document communicative acts of humans by some form of recording and/or descriptions, both directly as in corpora, or at higher levels of abstraction in lexicons and ontologies”

(International Standard for Language Engineering, ISLE)

- Componenti essenziali della LC per il TAL
- Machine-readable form
- Dati: testi scritti, registrazioni audio/video
- Strumenti per acquisizione, sviluppo, gestione, interrogazione dei dati delle RL

Quali sono le RL?

- Corpora
 - rappresentativi di una lingua, o specialistici
 - raw corpora, annotati
 - mono-/multilingue (paralleli-allineati); scritto/parlato/misti; sincronici/diacronici; completi/parziali; statici/dinamici...
- Lessici e dizionari
 - generalisti, di dominio (“terminologie”)
 - annotati (e.g. valenziali), di frequenza...
- Ontologie (WordNet):
 - rappresentazione formale ed esplicita di un concetto
 - gerarchia di concetti (rappresentati lessicalmente) organizzati per sussunzione (is_a), sinonimia, antinomia, meronimia...
- Grammatiche (formali e non)

Annotazione

Pratica di aggiungere a un corpus
diversi tipi di metadati

- Editoriali (relazione corpus – fonte originale): aggiunte/ omissioni, correzioni, normalizzazioni
- Descrittivi (informazione classificatoria): informazioni bibliografiche, autore, anno, lingua...
- Amministrativi (informazione documentaria): descrizione dei file e del tagset, disponibilità del corpus, stato di revisione dei dati, produttore...
- Analitici (interpretazione e analisi delle componenti del corpus): caratteristiche strutturali (e.g. divisione del testo), citazioni, parole straniere, named-entity recognition (nomi propri, date, luoghi...), informazione linguistica (“corpus annotation”)
 - Leech (2004): “Corpus annotation is the practice of adding interpretative linguistic information to a corpus”

Livelli di annotazione linguistica

- Testo in formato elettronico: tokenizzazione
- (spoken corpora) Annotazione fonetica e prosodica (accenti, intonazione...)
- Analisi morfologica e lemmatizzazione
 - PoS tagging
- Analisi sintattica
- Information structure (TR)
- Anaphora/ellipsis resolution
- Annotazione pragmatica-retorica (forza illocutoria)
- Annotazione semantica: SRL, WSD
- Annotazione multimodale del parlato: gesti, espressioni facciali, contesto percepibile...

Risorse
Le treebank

Cos'è una treebank?

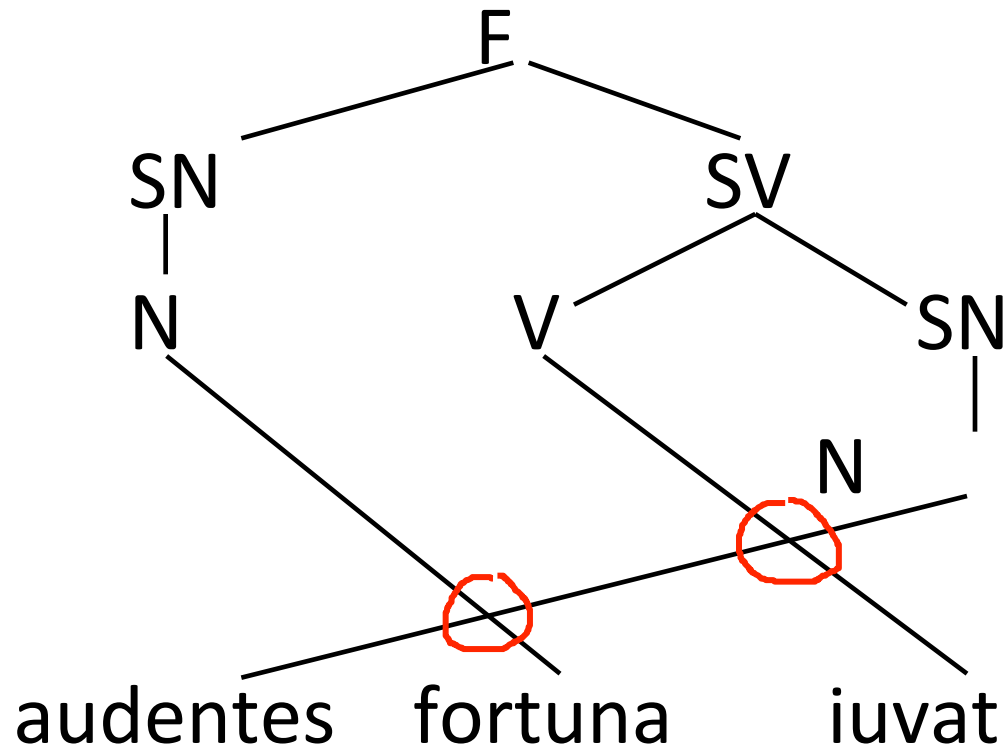
- Corpus annotato a livello sintattico
- (di solito) include:
 - Lemmatizzazione (disambiguata)
 - Tratti morfologici (disambiguati)
 - Analisi sintattica

Quali ‘semi’ per gli alberi?

- Grammatiche a costituenti (PSG):
 - Parole, PoS, Sintagmi, Start Symbols
 - Set inclusion: categorizzazione (es.: parola, N, SN, F)
- Grammatiche a dipendenze (DG):
 - Solo parole (elementi terminali)
 - I nodi lessicali sono collegati attraverso relazioni binarie e gerarchiche (‘dipendenze’)
 - Nessuna relazione lineare (orizzontale), o ciclica
 - Nessuna marcatura dell’ordine superficiale delle parole
 - Adatte alla descrizione di lingue con morfologia ricca (“free-word-order”): olandese (Alpino), italiano (TUT), ceco (PDT), latino (IT-TB; LDT; PROIEL)...

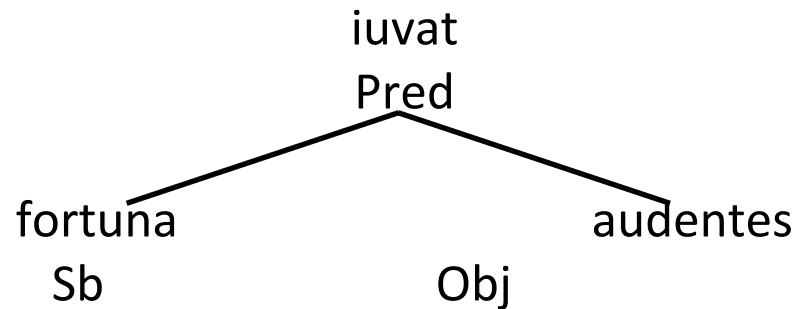
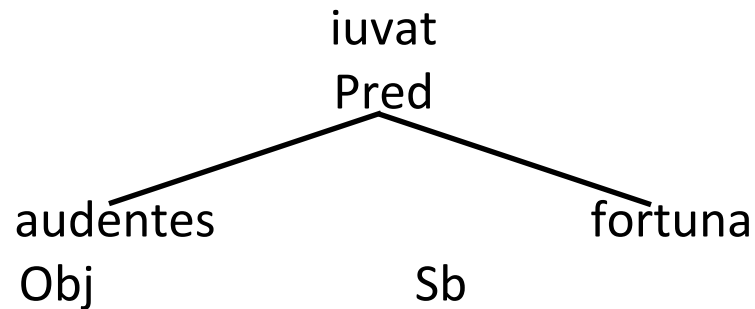
Esempio (PSG)

Audentes fortuna iuvat



Esempio (DG)

Audentes fortuna iuvat



Le Treebank per il latino

Index Thomisticus Treebank

- CIRCSE, Università Cattolica del Sacro Cuore (Milano)
- basata sull'*Index Thomisticus* (Busa): Tommaso d'Aquino
- annotazione sintattica: ca. 18,000 frasi (ca. **360,000** nodi)
- annotazione semantico-pragmatica: ca. 2,000 frasi (26,900 nodi)

Latin Dependency Treebank

- Perseus Digital Library (Boston, Lipsia)
- autori di latino classico
- annotazione sintattica: 3,473 frasi (ca. **53,000** nodi)
- annotazione semantico-pragmatica (<IT-TB): intera *De coniuratione Catilinae* (ca. 700 frasi; ca. 9.000 nodi) + estratti di Cesare e Cicerone (100 frasi l'uno)

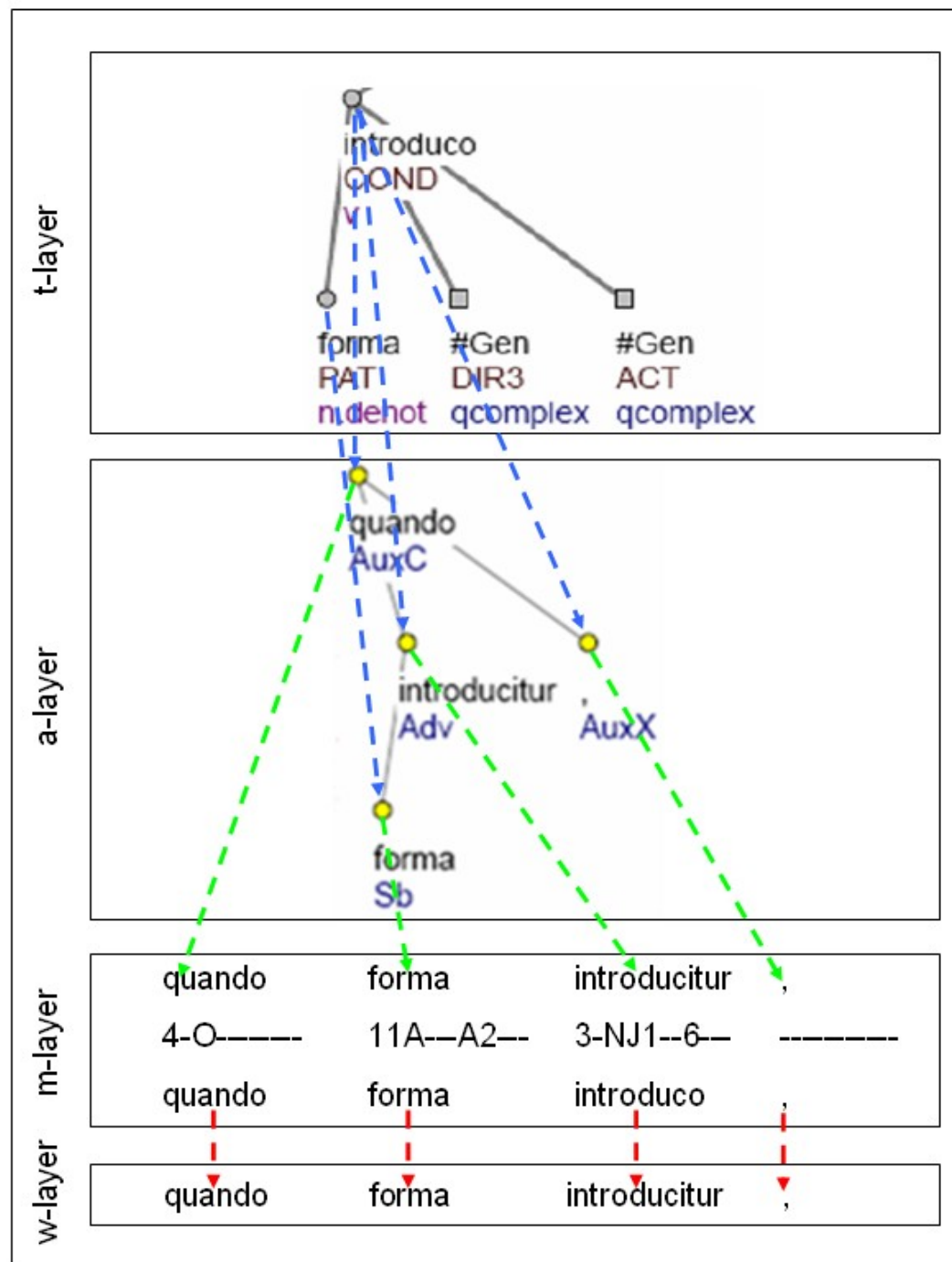
PROIEL corpus

- Università di Oslo
- autori di latino classico e NT (+ testi documentari dell'VIII secolo)
- annotazione sintattica (e semantica): ca. **190,000** nodi

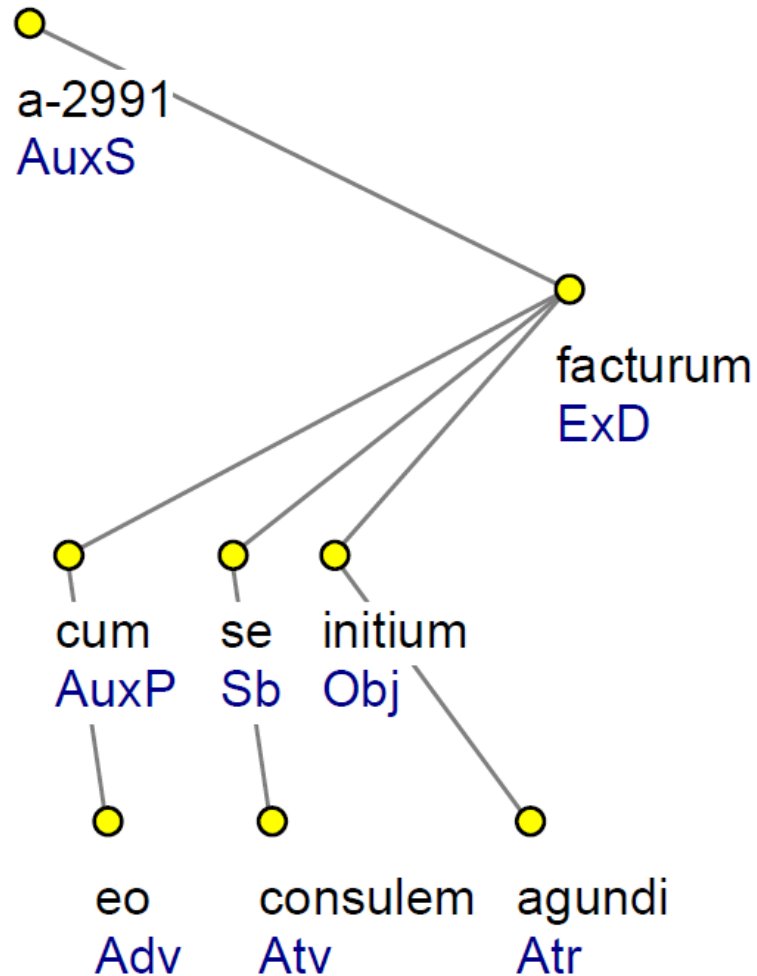
Criteri di annotazione

IT-TB & LDT

- **Teoria di riferimento:** *Functional Generative Description*
 - «La langue est un système de moyens d'expression appropriés à un but» (Circolo Linguistico di Praga, 1929)
 - «[...] treating language as form-meaning composite» (Daneš, 1968)
- **Livelli di annotazione:**
 - *W* (word): raw text
 - *M* (morfologico): analisi morfologica e lemmatizzazione
 - *A* (analitico): analisi sintattica → FORMA (*IT-VaLex*)
 - *T* (tectogrammaticale): analisi semantico-pragmatica → SIGNIFICATO (*Latin Vallex*)
- **Alberi a dipendenze:**
 - «L'acte syntagmatique fondamental [...] est la prédication» (Circolo Linguistico di Praga, 1929)

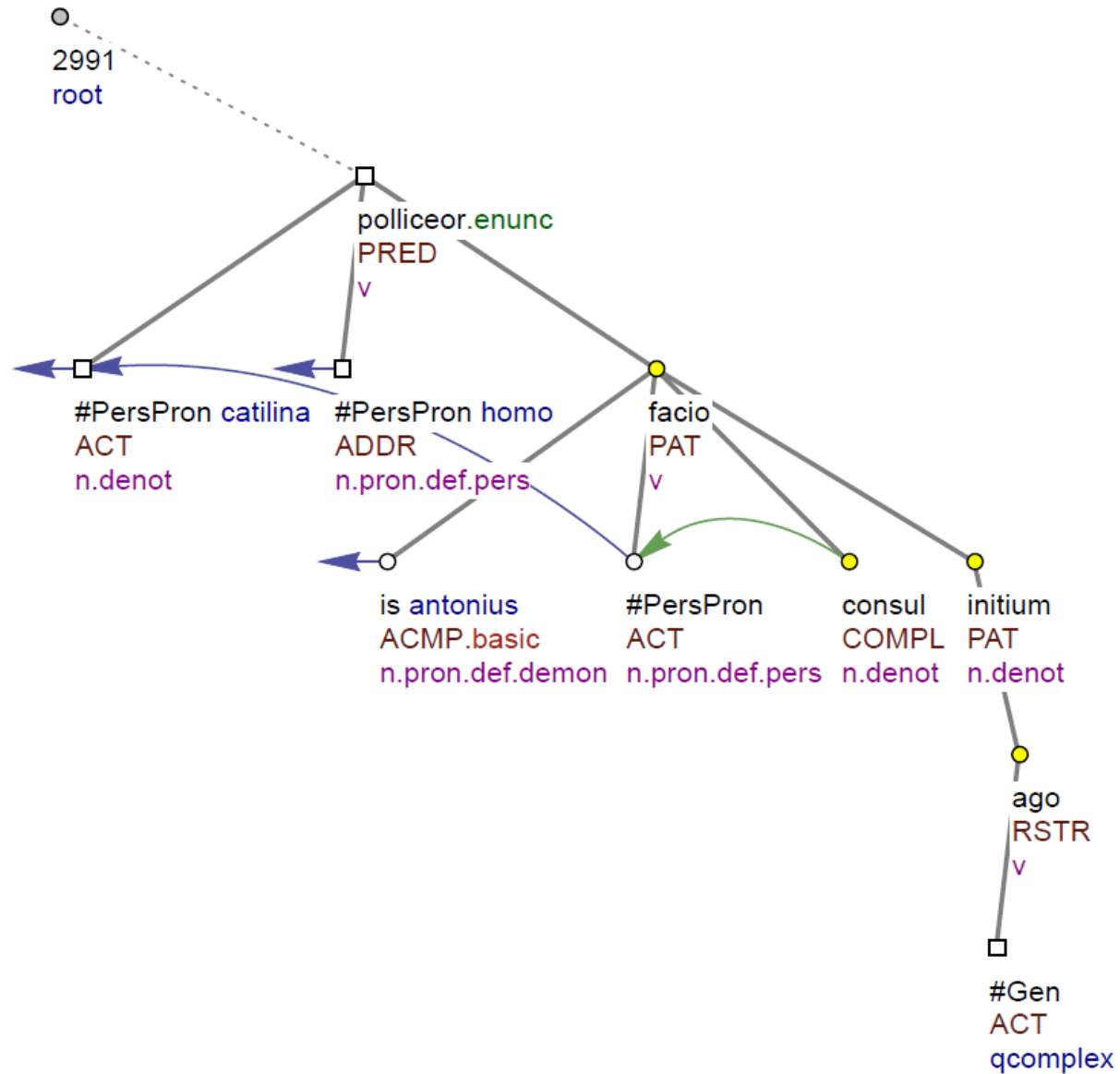


Albero analitico



cum eo se consulem initium agundi
facturum

Albero tectogrammaticale



cum eo se consulem initium agundi facturum

Universal Dependencies (UDs)

- <http://universaldependencies.org/>
- “cross-linguistically consistent treebank annotation for many languages” (ca. 100 v. 2.1)
- distribuita in LINDAT-CLARIN

Interrogare le treebank

- UDs via PML-TQ presso Lindat:
 - <http://lindat.mff.cuni.cz/services/pmltq/#!/home>
- Esempio di query (UDs treebank: IT-TB)
 - Cosa è *forma*? `a-node $n1 := [ conll/cpos = 'NOUN', child a-node $n2 := [ conll/dep1 = 'cop'], a-node $n3 := [ conll/dep1 = 'nsubj', lemma = 'forma'] ]`
`>> for $n1.lemma give $1,count() sort by $2 desc, $1`

Risorse
I lessici

Il lessico di
sottocategorizzazione sintattica
IT-VaLex

<http://itreebank.marginalia.it/itvalex/>

IT-VaLex

- distinzione tra argomenti e satelliti nel livello analitico
- lessico indotto automaticamente dalla IT-TB (dinamico)
- associa a ogni verbo tutte le sue sottocategorizzazioni argomentali sintattiche rappresentate testualmente (nella IT-TB) da un elemento lessicale: *valenza sintattico-testuale*
- dimensione: 1.001 verbi (49.616 occorrenze)

Interrogare *IT-VaLex*

Domanda e risposta

<http://itreebank.marginalia.it/itvalex/>

Scheda Query (ricerche complesse libere)

- verbi all'attivo con almeno due argomenti:
 1. Obj, preverbale (BV), all'accusativo, retto dalla preposizione *ad*
 2. Sb, postverbale (AV), espresso da una proposizione introdotta dalla congiunzione *quod*

Il lessico di valenza *Latin* *Vallex*

<http://itreebank.marginalia.it/vallex>

Dimensioni e origine dei dati

- 1,373 entrate lessicali (verbi, nomi, aggettivi, avverbi)
- 3,406 frame valenziali

Corpus-based (annotation-driven):

- *Index Thomisticus* Treebank: ca. 2,000 frasi (SCG)
- Latin Dependency Treebank: ca. 900 frasi
 - Cesare: 100 frasi (*De bello gallico*)
 - Cicerone: 100 frasi (*In Catilinam*)
 - Sallustio: 701 frasi (intera *De coniuratione Catilinae*)

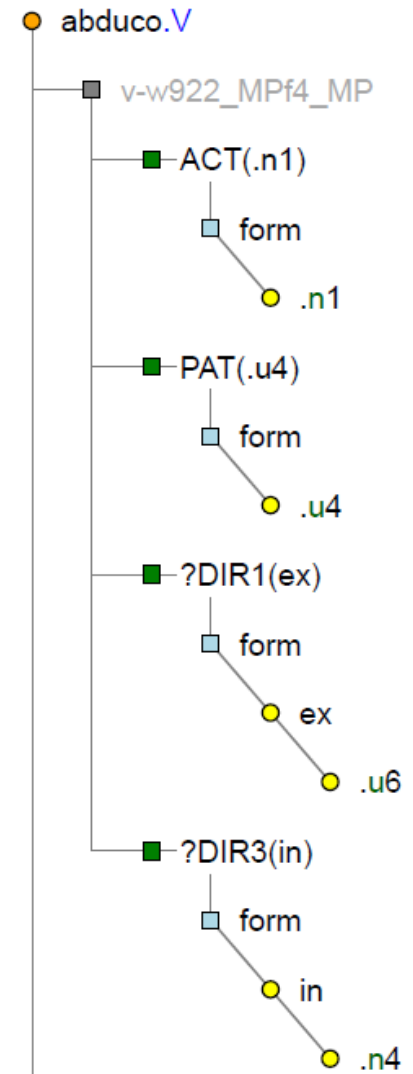
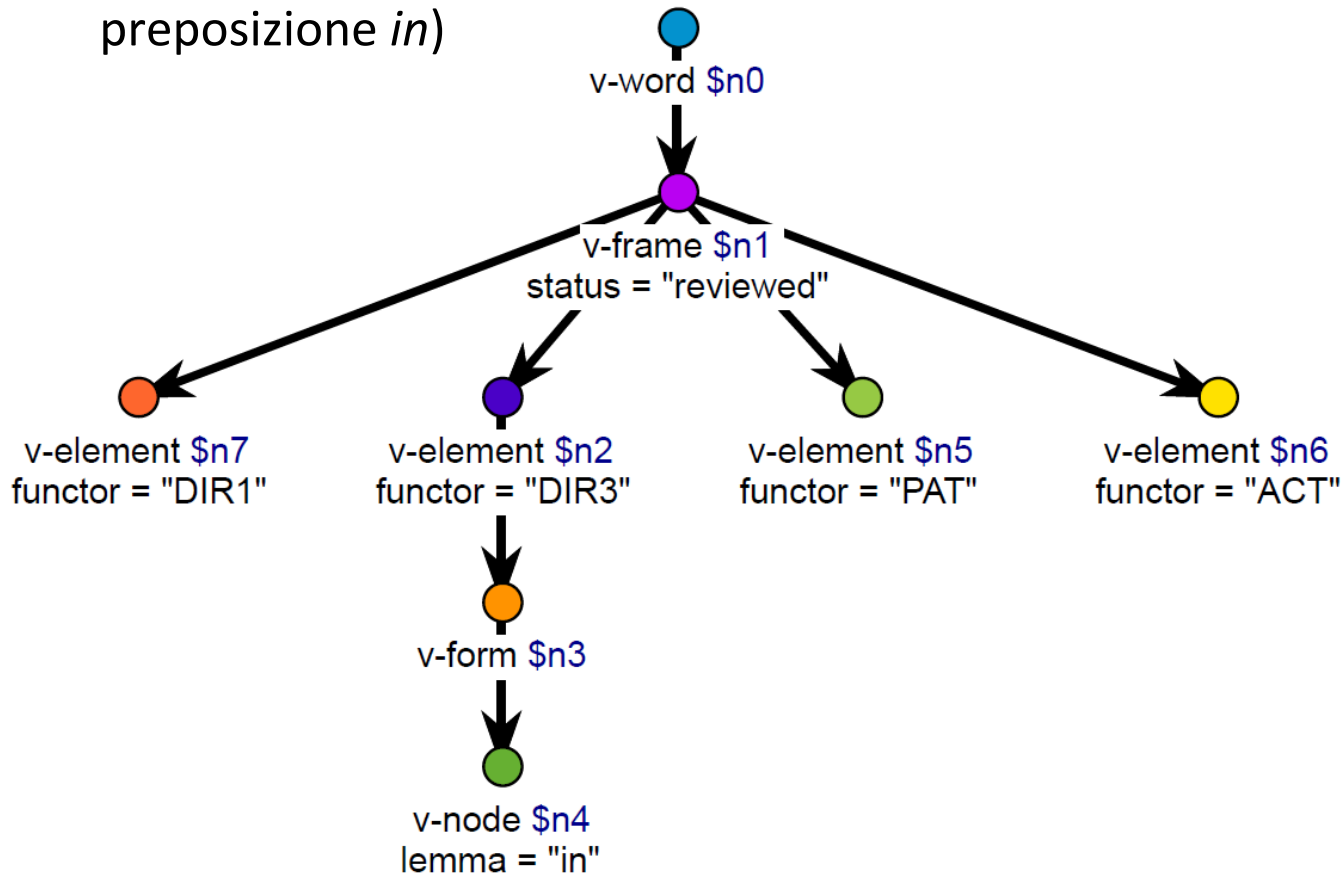
Intuition-based (per 'equilibrio'):

163 entrate, a copertura dei primi 1.000 lemmi valenziali in Delatte et alii (1981). *Dictionnaire Frequentiel et index inverse de la langue latine*

Interrogare *Latin Vallex*

Domanda (...e un po' di risposta)

Entrate lessicali con almeno un frame valenziale
tetra-argomentale: Agente, Paziente, Direzione-da,
Direzione-a (espressa con un SP introdotto dalla
preposizione *in*)



Il lessico morfologico derivazionale WFL

<http://wfl.marginalia.it/>

Cosa è WFL

- *Word Formation Latin*
- Base lessicale: ca. 70.000 lemmi di LEMLAT
- Morfotassi vs. diacronia
- Stato della risorsa (in corso):
 - 508 famiglie morfo-derivazionali
 - 137 regole derivazionali applicate (prefissali / suffissali / converse / composizionali)
 - 19.196 relazioni
 - 8.421 lemmi in input
 - 18.827 lemmi in output

Interrogare WFL

Domande e risposte

<http://wfl.marginalia.it/>

- Scheda By WFR
 - Aggettivi denominali (N-To-A) con il suffisso *-al-*
- Scheda By Affix
 - Lemmi formati con il prefisso *in-* (negazione)
- Scheda By PoS
 - Aggettivi denominali della prima classe (A1) prodotti per suffissazione (WFR: suffixes only) da un nome femminile della prima declinazione (N1,f)
- Scheda By Lemma
 - La famiglia morfo-derivazionale del lemma *forma*

Strumenti di TAL

Lemmatizzazione e morfologia
flessionale

LEMLAT

www.lemLAT3.eu

- Analizzatore morfologico e lemmatizzatore
- LEMLAT 3.0 → CIRCSE + ILC-CNR
- Base lessicale:
 - Georges and Georges, 1913-1918; Glare, 1982; Gradenwitz, 1904 → **43.432 lemmi**
 - Forcellini (1771/1940; Onomasticon) → **26.250 lemmi**

I Competitor di Lemlat

- Whitaker's *Words*
- *Morpheus* → *Parsley* (2013)
- PROIEL Latin Morphology System
- *LatMor*

	Caesar		Nepos		Godfrey	
all	type	token	type	token	type	token
PROIEL	70.0	51.6	69.4	47.9	63.1	50.6
Parsley	89.5	95.2	90.0	94.3	86.7	91.7
Words	90.5	96.6	88.1	93.3	93.0	95.4
Morpheus	92.5	93.8	89.0	92.7	87.6	92.7
LEMLAT	<u>98.2</u>	99.0	<u>98.1</u>	99.1	91.0	94.9
LatMor	97.5	<u>99.1</u>	<u>98.1</u>	<u>99.2</u>	<u>96.4</u>	<u>97.5</u>

Springmann et al. (2016, p. 389)

Quanto bene funziona?

Lemlat 3.0 applicato al *Thesaurus Formarum Totius Latinatis* (Tombeur, 1998):

- 400,886 / 554,826 forme lemmatizzate da Lemlat (72.25%)
- 153,447 forme non analizzate da Lemlat:
 - Sequenze di lettere (e.g. *aaa*)
 - numeri (e.g. *CCC*)
 - forme rarissime (e.g. *aaliza*, 1 occorrenza in TFTL)

Numero delle occorrenze in TFTL delle forme analizzate:

- Totale in TFTL: 62,922,781 occurrences
- Totale analizzate da Lemlat: 61,881,702 occurrences
- Lemlat analizza il **98.345%** delle occorrenze nei testi di TFTL

```

type the wordform you wish to analyze <type end to exit>>rosarum
*****

Form=rosarum

Input      wordform : rosarum
Analized wordform : rosarum

=====ANALISI 1=====

SEGMENTAZIONE:  ros -arum

-----codici morf. 1-----
--gfp--

Caso:      Genitivo
Genere:    Femminile
Numero:    Plurale
=====LEMMMA =====
rosa                      N1      r0795 f
-----codici morfologici-----
NcA

      PoS:      Nome
      Type:     Comune
      Categoria flessiva:      I decl

=====ANALISI 2=====

SEGMENTAZIONE:  ros -arum

-----codici morf. 1-----
k4gfp--

Modo verbale:  Participio pass/dep
Tempo verbale: Perfetto
Caso:      Genitivo
Genere:    Femminile
Numero:    Plurale

LEMMI:
=====LEMMMA 1: IPO =====
rosus                      N1/2  r1470 *
=====LEMMMA 2: IPER=====
rodo                       U3      r1470
-----codici morfologici-----
UmH

      PoS:      Verbo
      Type:     Main
      Categoria flessiva:      III coniug

type the wordform you wish to analyze <type end to exit>>

```

n_id	gen	dcm	si	smv	spf	les	codles
r0795	f					ros	n1

n_id	gen	dcm	si	smv	spf	les	codles
r1470		v				rod	v3r
r1470						ros	v7s
r1470	*					ros	n41
r1470	*	i				ros	n6p1
r1470	*	i				rosur	n6p2

segment	comp_cod	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10
arum	n1	n	c	a	-	-	g	f	p	-	-
arum	n1	n	c	a	-	-	g	m	p	-	-
arum	n1e	n	c	a	-	-	g	f	p	-	-
arum	n1e	n	c	a	-	-	g	m	p	-	-
arum	n6	a	*	-	-	-	g	f	p	-	1
arum	n6p	p	*	-	-	-	g	f	p	-	-
arum	n6s	a	*	-	-	-	g	f	p	-	3
arum	n6p1	v	m	*	k	4	g	f	p	-	-
arum	n6p2	v	m	*	j	3	g	f	p	-	-
arum	n6g	v	m	*	r	-	g	f	p	-	-
arum	n6p3	p	*	-	-	-	g	f	p	-	-

Usare Lemlat

Form=homicidam

Input wordform : homicidam

Analysed wordform : homicidam

=====ANALYSIS=====

SEGMENTATION: homicid -am

-----morphological feats-----

--ams--

Case: Accusative

Gender: Masculine

Number: Singular

=====LEMMA=====

homicida N1 h0511 m

-----morphological feats-----

NcA

PoS: Noun

Type: Common

Inflexional Category: I decl

-----derivational info-----

IS DERIVED: YES

-----rule id: 200-----

Lexical Basis:

homo

N3B h0518 m NcC

caedo

U3 c0083 UmH

Derivational Type: Compounding

Derivational Category: N+U=N

Usare Lemlat

```
A>castigabilem
*****

Form=castigabilem

Input      wordform : castigabilem
Analysed   wordform : castigabilem

=====ANALYSIS=====

SEGMENTATION:  castigabil -em

-----morphological feats 1 -----
--ams-1

Case:  Accusative
Gender: Masculine
Number: Singular
Degree: Positive

-----morphological feats 2 -----
--afs-1

Case:  Accusative
Gender: Feminine
Number: Singular
Degree: Positive

=====LEMMA=====
castigabilis          N3A  c0772 *
-----morphological feats-----
Af-

PoS:  Adjective
Type:  Qualificative

-----derivational info-----
IS DERIVED: YES

-----rule id: 38-----
Lexical Basis:
      castigo          U1      c0776      UmF
Derivational Type: Derivation_Suffix
Derivational Category: U-To-A
Affix: bil
```

Strumenti di TAL

PoS Tagging e Parsing

PoS Tagging e Parsing

Dati: *Index Thomisticus* Treebank

Training set: 236.531 nodi (13.770 frasi)

Test set: 17,353 nodi (1.180 frasi)

- Data-driven PoS Tagging:
 - HMM-based HunPos tagger (Halácsy et al., 2007) → TnT
 - Valutazione:
 - **98.57**: CPoS
 - **98.21**: CPoS + PoS
 - **93.82**: CPoS + PoS + tratti morfologici
 - Vedi anche: 2 parameter file di TreeTagger per il latino
- Data-driven Dependency Parsing
 - 3 parser: DeSR (shift-reduce), MTGB (graph-based), Joint (shift-reduce)
 - Tecniche (Ponti & Passarotti, 2016):
 - Feature Model ad hoc
 - Reverse Revision Parsing
 - Linear Tree Combination
 - Valutazione: **86.5** (LAS); 90.97 (UAS)

...13 secoli vorranno pur dire qualcosa

Author	Genre	LAS	UAS
Caesar	prose	28.2	42.1
Cicero	prose	27.6	41,1
Jerome	prose	35.8	51.2
Ovid	poetry	23.9	37.7
Petronius	prose	35.9	51.3
Propertius	poetry	25.8	40.5
Sallust	prose	27.6	41.3
Vergil	poetry	24.9	39.6

Strumenti di TAL

Mettiamo tutto insieme!

UDPipe

<http://ufal.mff.cuni.cz/udpipe>

UDPipe

<http://ufal.mff.cuni.cz/udpipe>

- pipeline addestrabile per tokenizzazione, tagging, lemmatizzazione e dependency parsing di file CoNLL-U
- modelli addestrati sono forniti per (quasi) tutte le treebank
UD: <http://lindat.mff.cuni.cz/services/udpipe/api/models>
- online demo:
<http://lindat.mff.cuni.cz/services/udpipe/run.php>

UDPipe

<http://lindat.mff.cuni.cz/services/udpipe/run.php>

- Modello: latin-itt

- «Manifestum enim est quod per peccatum praecedens fit quaedam inordinatio in affectu humano, ut facilius postmodum ad peccatum inclinetur.» (Tommaso d'Aquino, *SCG*, cap. 141, n. 9)



→ LAS: 91.48

- **MA:** «Saepe et multum hoc mecum cogitavi, bonine an mali plus attulerit hominibus et civitatibus copia dicendi ac summum eloquentiae studium.» (Cicerone, *De inventione*, I, 1)



→ LAS: 47.62

- **MA:** «et post dies sex adsumit Iesus Petrum et Iacobum et Iohannem et ducit illos in montem excelsum seorsum solos et transfiguratus est coram ipsis.» (*Evangelium secundum Marcum*, 9)



→ LAS: 45.84

- ...e poi dicono che Cesare è facile!

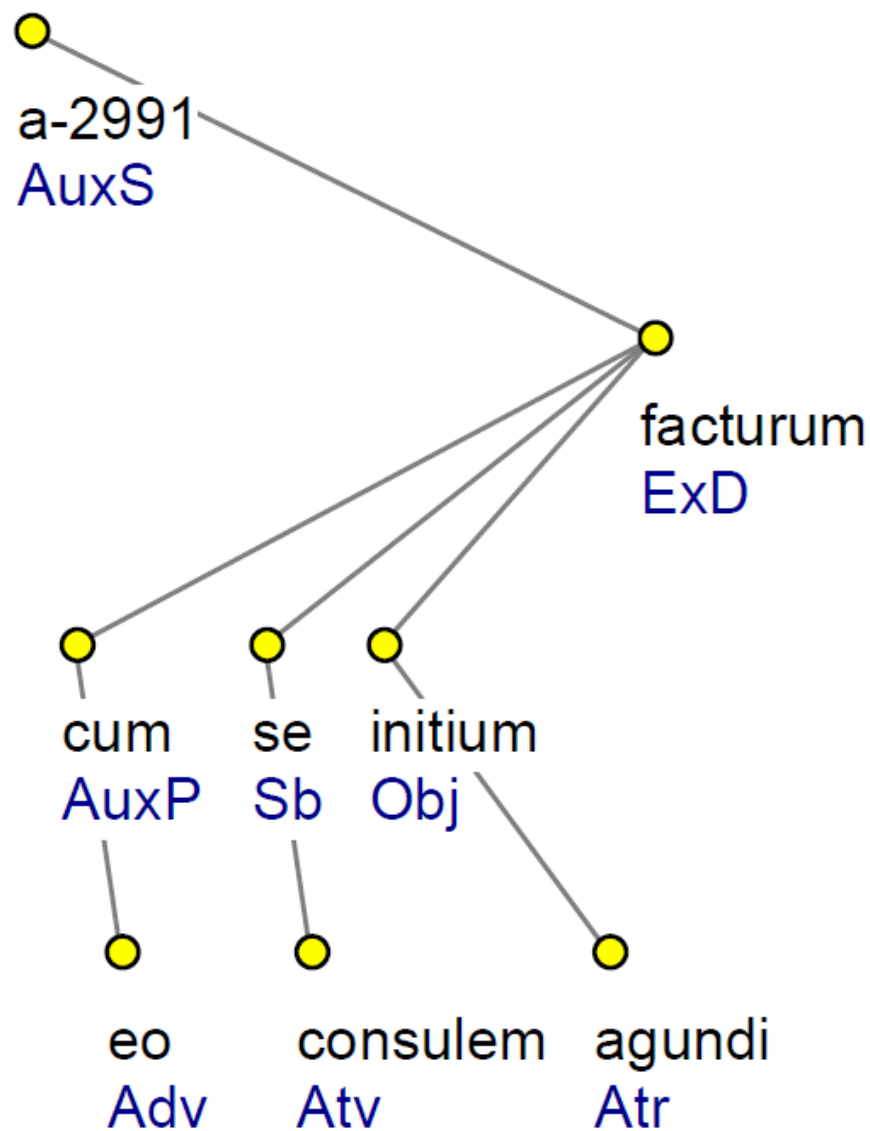
- «Apud Helvetios longe nobilissimus fuit et ditissimus Orgetorix.» (Cesare, *Bellum Gallicum*)



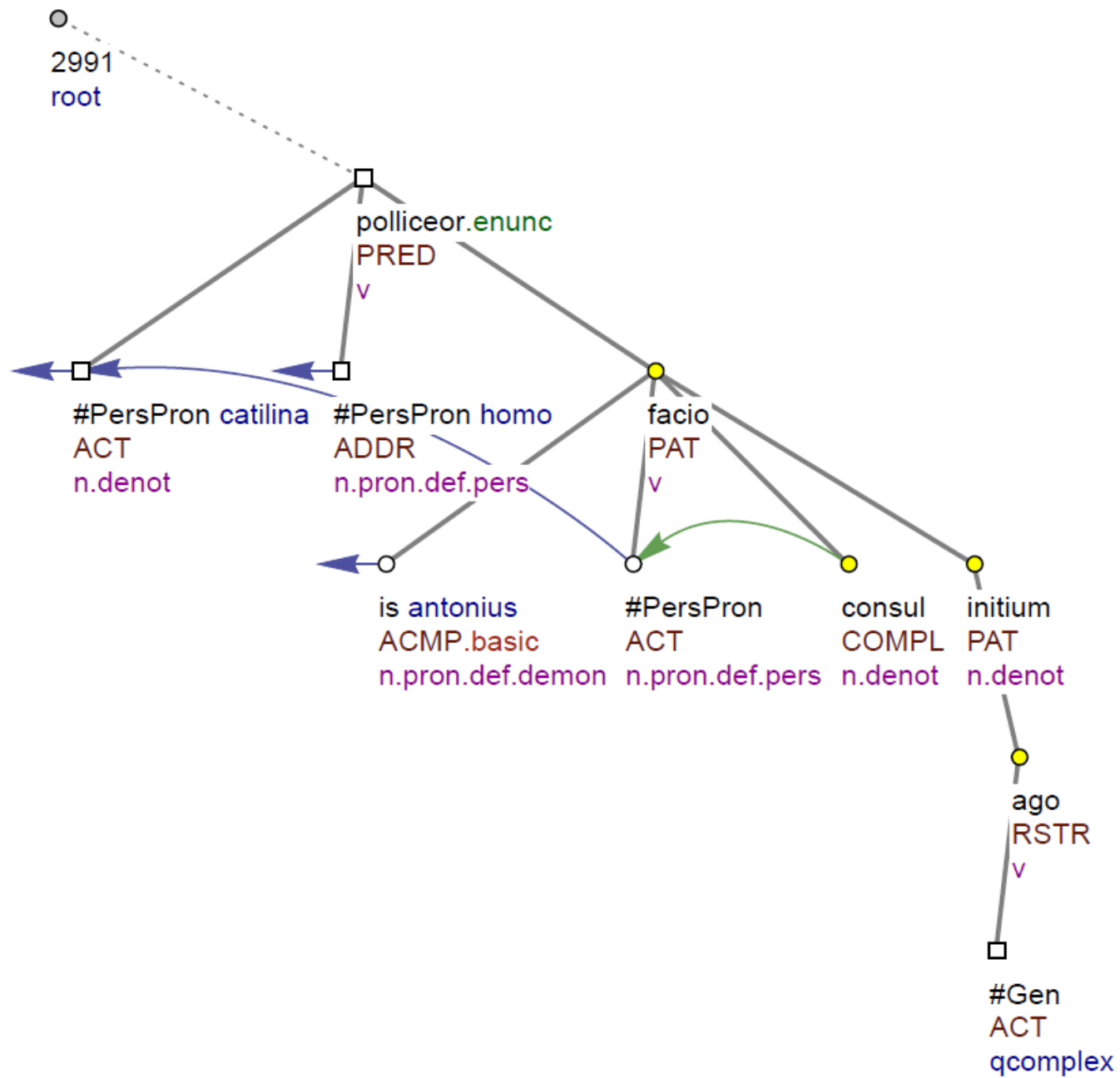
→ LAS: 25

Un esempio d'uso di una
treebank per analisi
letteraria

Bellum Catilinae



cum eo se consulem initium agundi
facturum



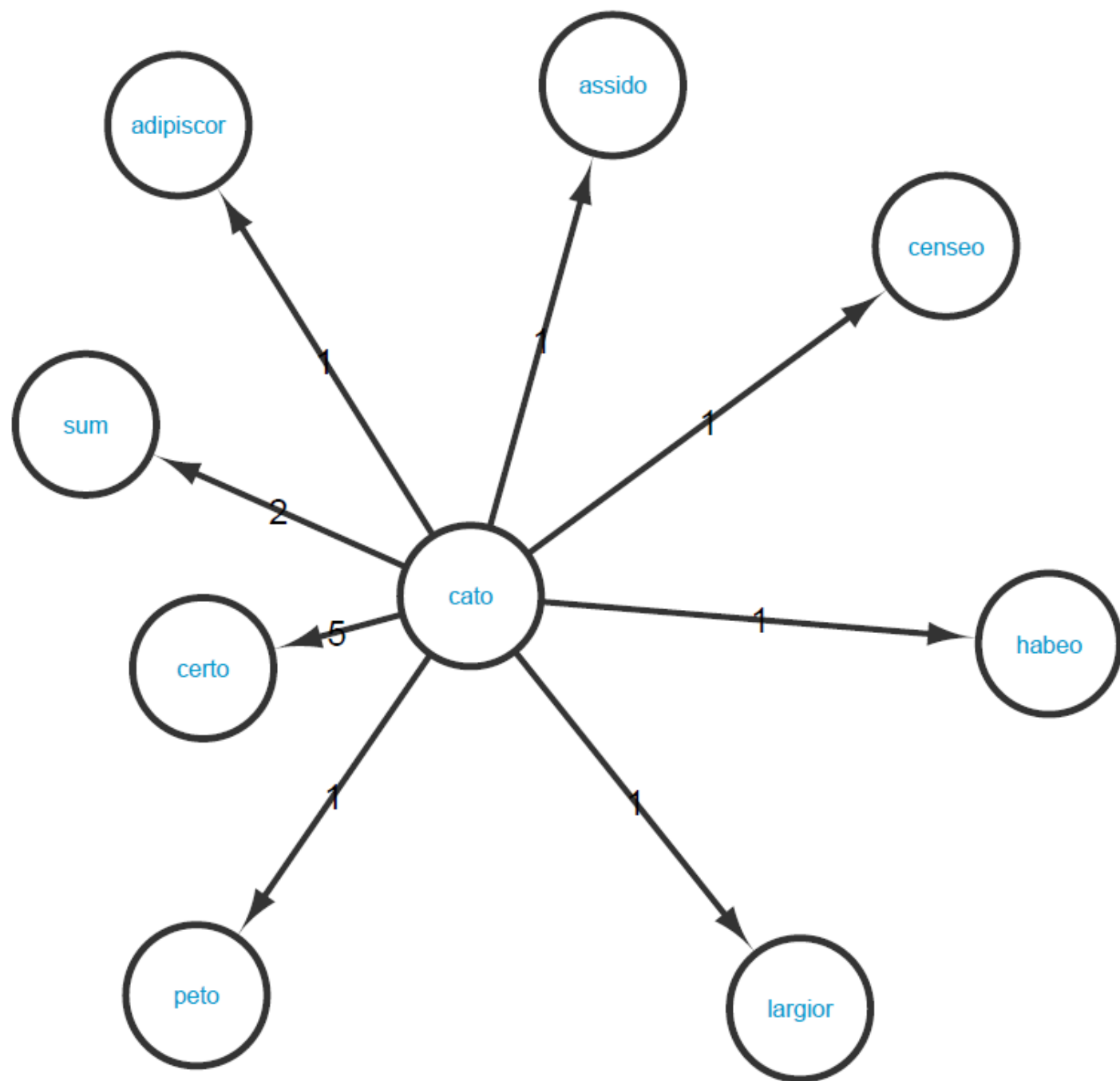
cum eo se consulem initium agundi facturum

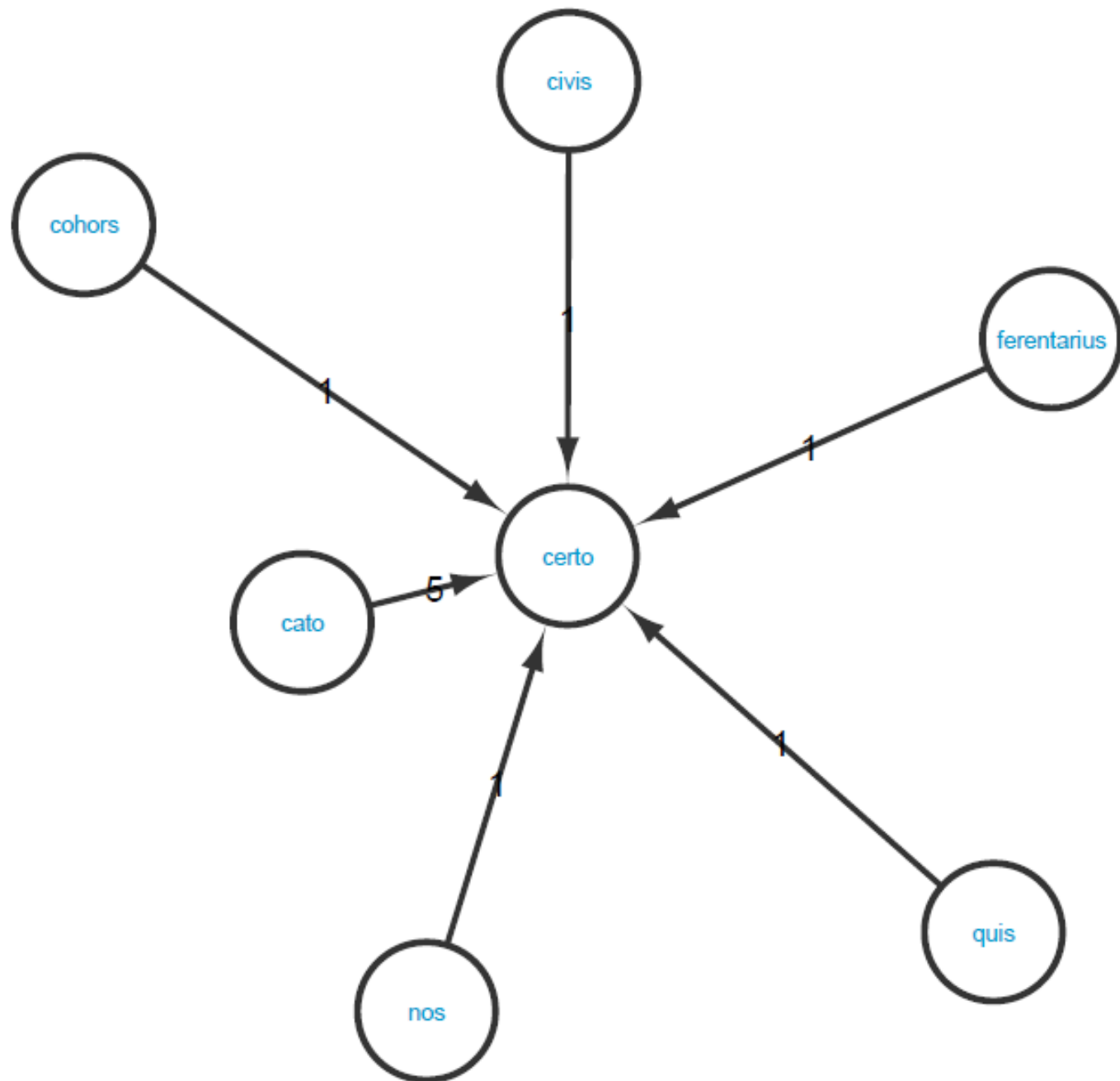
**Who does what
in *Bellum Catilinae*?**

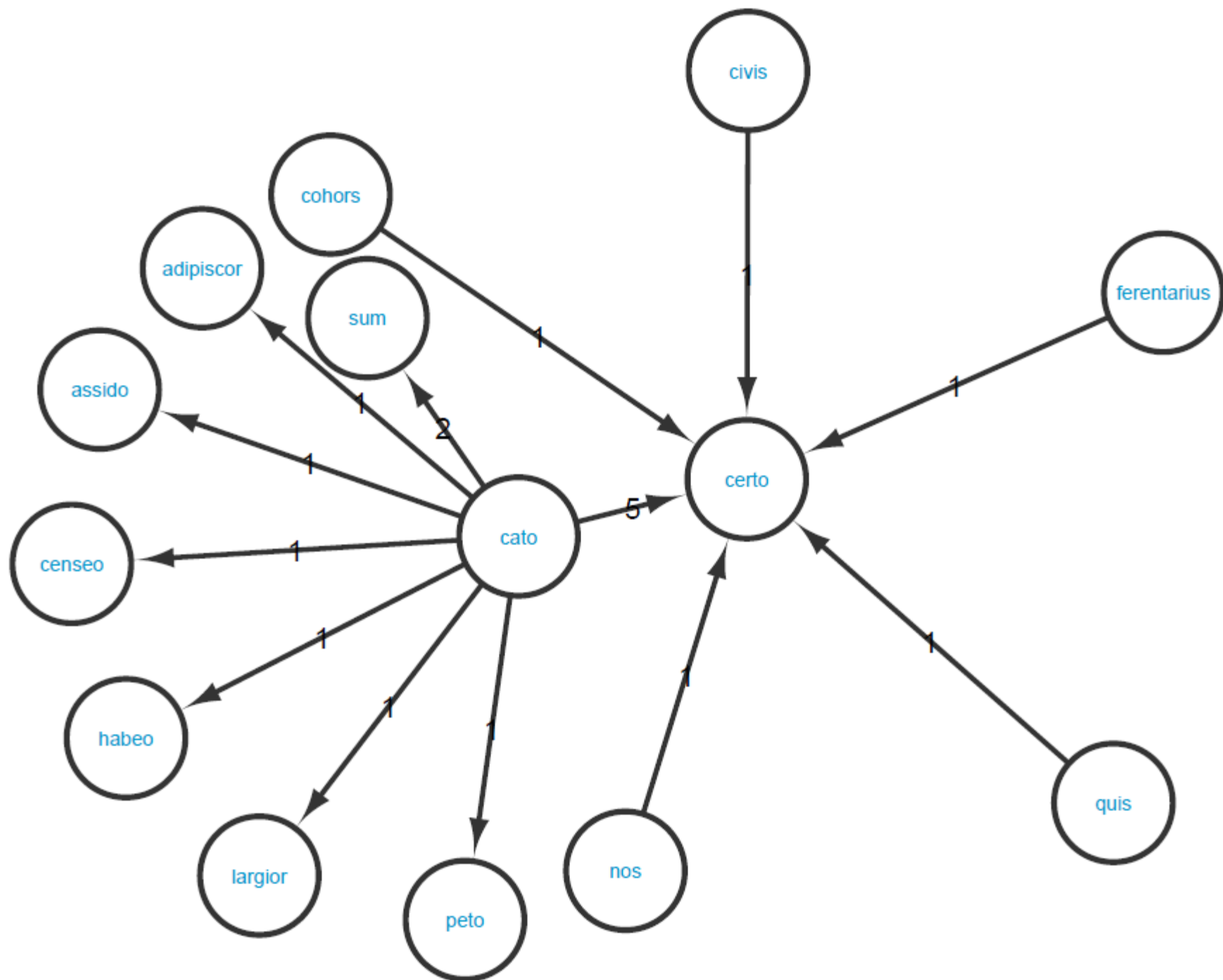
A Tecto-based Network of Actors and Actions

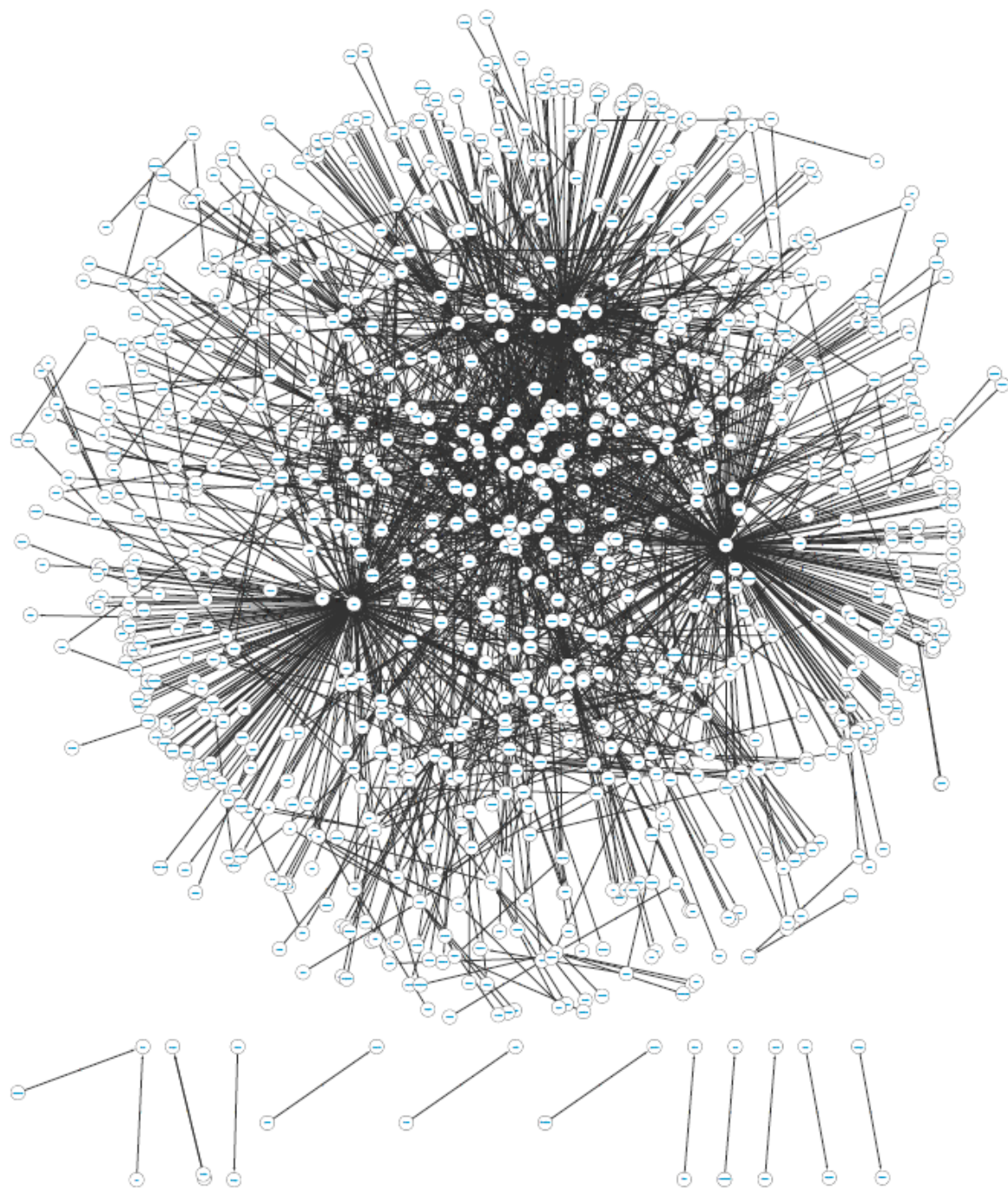
Network: $G(V, E)$

- Vertices: Actors and Actions
 - *Actors*: the human or non-human originator of the event, the bearer of the event or a quality/property, the experiencer or possessor (Mikulová, et al., 2006, p. 461)
 - *Actions*: verbs → [future work] nominalizations
- Edges (directed): dependency relations between Actors and Actions in TGTs









«Traditional» Main Characters

CATILINA

ACT of 133 verbs
216 occurrences

CICERO

ACT of 33 verbs
40 occurrences

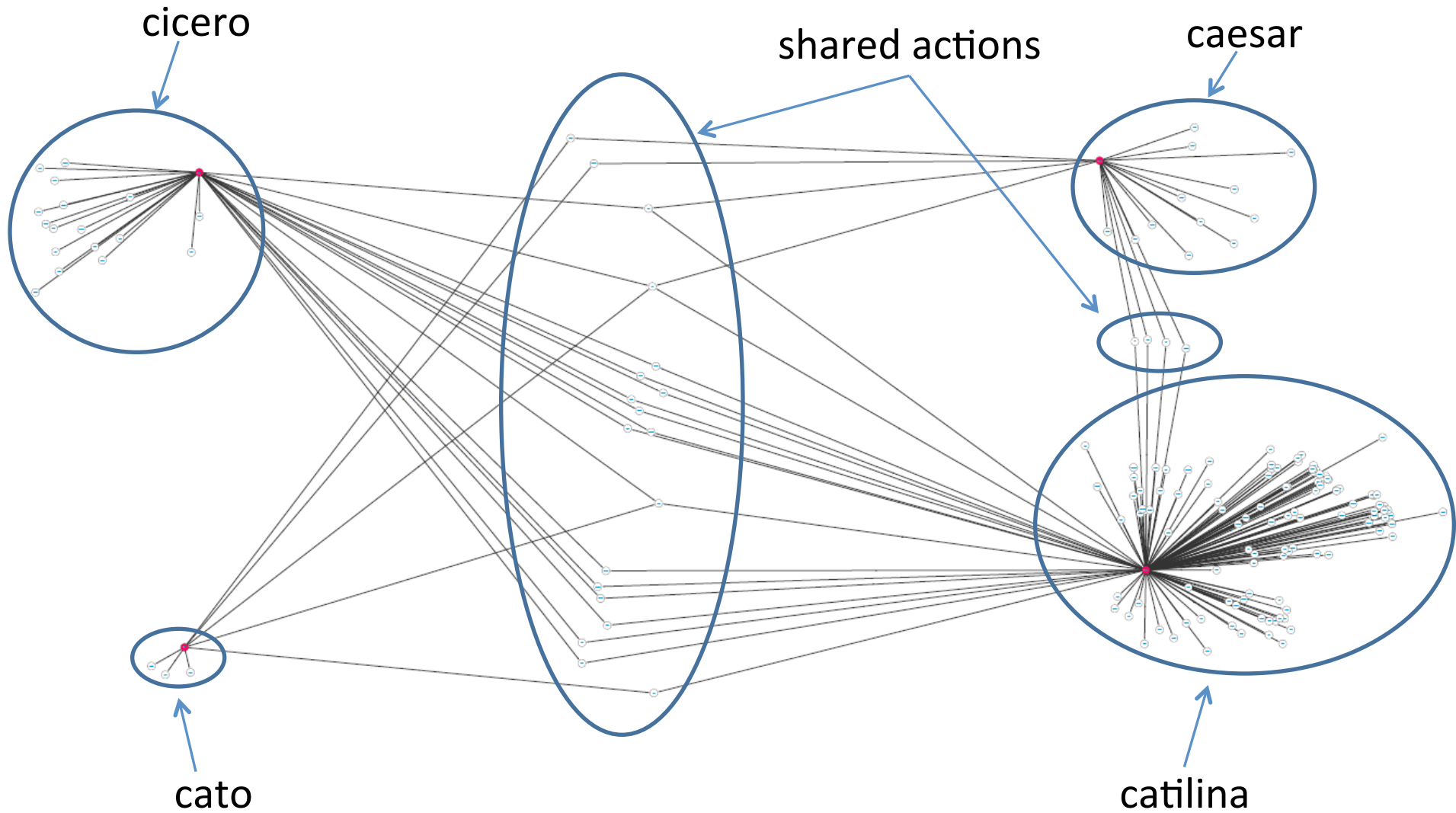
CAESAR

ACT of 20 verbs
22 occurrences

CATO

ACT of 8 verbs
13 occurrences

«Traditional» Main Actors



Catilina Only

25 most Frequent Actions



What parat?

arma
incendium
insidiae
interficio
mitto
multitudo
regnum

...a bad guy!

**Who acts similarly
(and differently)
in *Bellum Catilinae*?**

Highest Outdegree (Main Actors)

LEMMA	OUTDEGREE
catilina	133
cicero	33
homo	32
res	24
petreius	20
lentulus	20
consul	20
caesar	20
populus	19
curius	19
vulturcius	18
vir	18
animus	18

Clustering

- the process of organizing objects (*observations*) into groups (*clusters*) whose members are similar in some way
- a *cluster* is a collection of objects which are “similar” to each other and are “dissimilar” to the objects belonging to other clusters

Clustering the Actors

- R enviroment for statistical computing
 - Packages: tm, proxy
- Agglomerative hierarchical clustering:
 - compare Actors by Actions
 - compute the amount of shared & non-shared Actions
 - compare the distribution of shared and not shared Actions by relative frequency

This means that

Actors that share a high number of Actions with similar distribution are clustered together

Result

