Open Philology Workshop

8.-10. August 2013, Leipzig University

An Introduction to a Latin Language Toolkit

Prometheus*

* working title

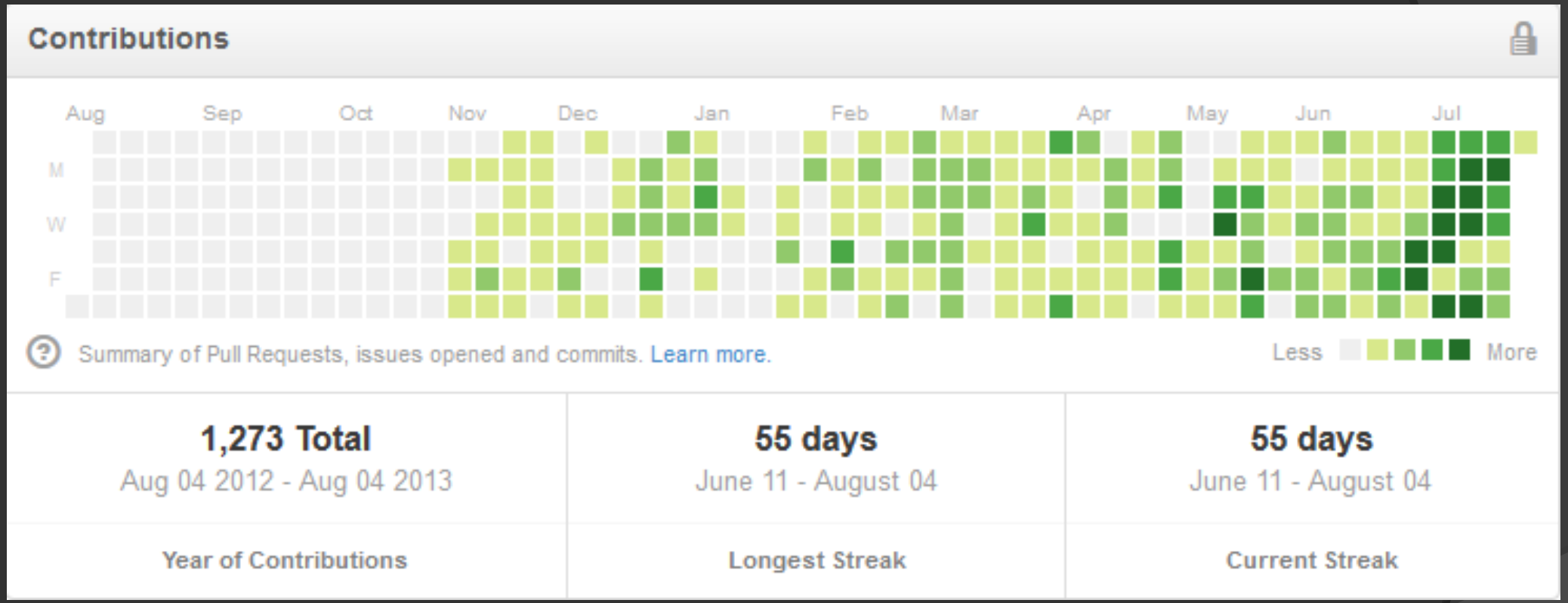
Overview

- Background
- Principles and Goals
- Understanding Latin
  - Tokenization
  - Morphological Analysis
  - Syntactical Analysis - the system's internals
  - Evaluating Caesar
  - Upcoming Improvements and Challenges
- Linking with Data-driven parsers - the system as a Post-Processing Unit
- Creating Latin
  - A framework for Latin studies
  - Dynamic e-learning tools
- The Future - cooperating with the research community

Background

- Who are we?
  - no experience in computer linguistics
  - Study of Latin Philology, Classical History, History (teacher-certificate)
  - Experience in teaching Latin
- Ideas and Motivation
  - Author specific exercises
  - Immediate feedback
- Working conditions
  - Private project
  - Built from scratch

Timeline


Contributions
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
M
W
F
Summary of Pull Requests, issues opened and commits. Learn more.
Less
More
1,273 Total
Aug 04 2012 - Aug 04 2013
Year of Contributions
55 days
June 11 - August 04
Longest Streak
55 days
June 11 - August 04
Current Streak

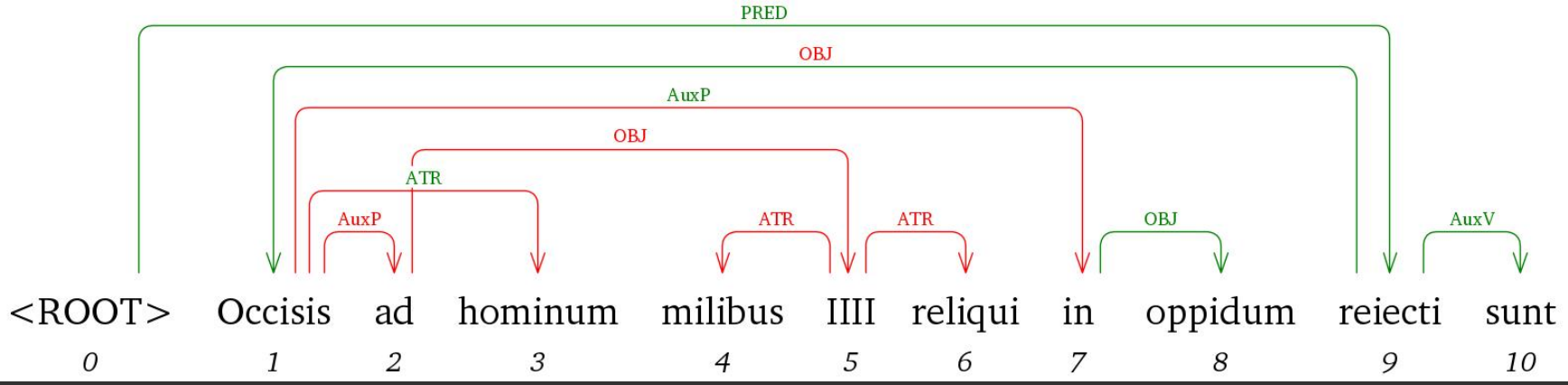
Principles

- Texts are written to be understood
- No special training required
- Semantics are not deciding
- Non-projectivity is nothing special

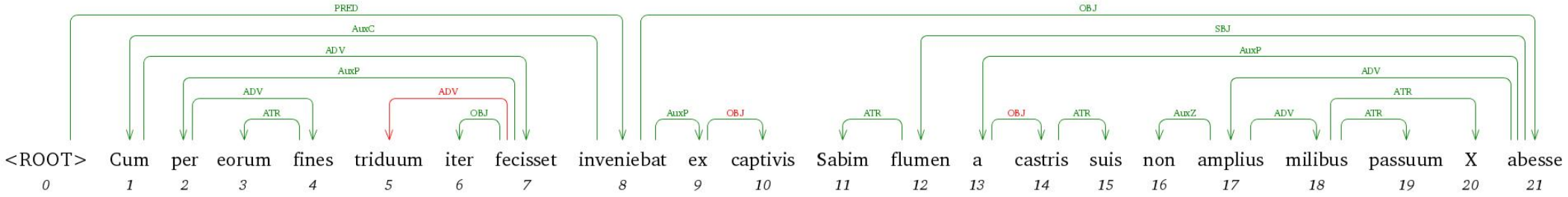
Requirements

- Two-way-engine human style
- Modular, extendable and flexible
- Easy recalibration
- No results are better than malformed results

malformed result:



acceptable result:



Requirements

- Two-way-engine human style
- Modular, extendable and flexible
- Easy recalibration
- No results are better than malformed results
- One language done well
- “real-time” computation speed

=> object oriented programming

Understanding Latin
-
the parsing sequence

Tokenization

Example sentence (*Caes. BG 2,9,5*):

Ubi neutri transeundi initium faciunt, secundiore equitum proelio nostris Caesar suos in castra reduxit.

„As neither side started to cross (the river),

Caesar led his troops back to the camp, after a cavalry battle, which favoured our men.”

Tokenization

Example sentence (*Caes. BG 2,9,5*):

Ubi neutri transeundi initium faciunt , secundiore equitum proelio nostris Caesar suos in castra reduxit .

Morphology

- Analysis of words
  - Morphem-specification (stem, thematic-vowels, tempus/modus signs/marker, ending)
- Database look-up
  - Stem dictionary

Database entries

- „castra“ will return 4 forms:
  - nom. pl. n. – O. Decl
  - acc. pl. n. – O. Decl
  - voc. pl. n. – O. Decl
  - Imperative, pr. 2. p. sg. – A. Conj.

Morphology

- Analysis of words
  - Morphem-specification (stem, thematic-vowels, tempus/modus signs/marker, ending)
- Database look-up
- Phonology and metrical info
- Didactic aspects (patterns!)
  - **ama – ba – t**

Syntax – process overview

- Clause finding process
- Clause analysis
- Clause rating
- Clause merging

Clause Finding

Ubi neutri transeundi initium **faciunt ,** secundiore equitum **proelio** nostris Caesar suos in **castra reduxit**.

Safe Subjunction / Verb / Comma

Unsafe Verb

Clause Finding

Clause 1 (subordinate):

(Ubi neutri transeundi initium **faciunt ,)**

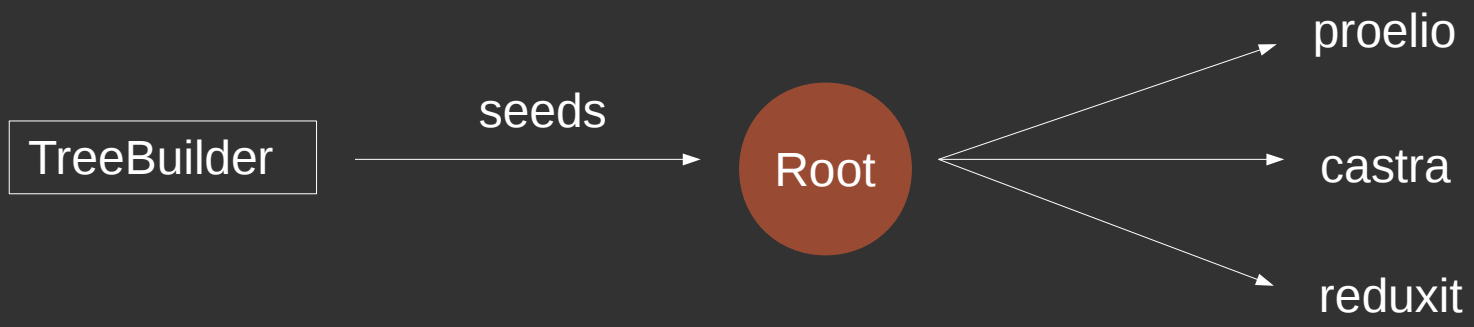
Clause 2 (main):

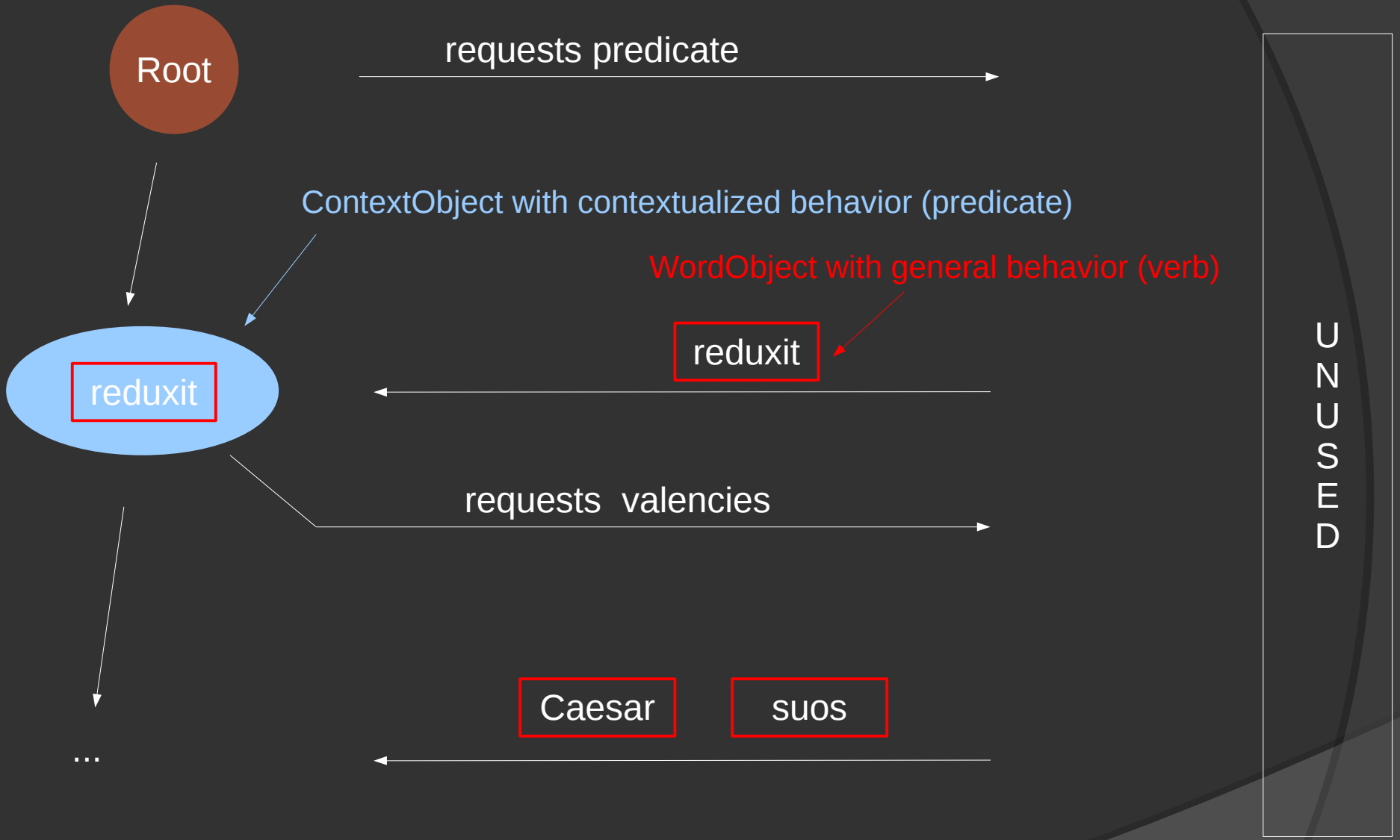
(secundiore equitum **proelio** nostris Caesar suos in **castra reduxit**. **)**

Safe Subjunction / Verb / Comma

Unsafe Verb

Clause analysis (tree-building)




Root
requests predicate
ContextObject with contextualized behavior (predicate)
WordObject with general behavior (verb)
reduxit
reduxit
requests valencies
Caesar
suos
...
UNUSED

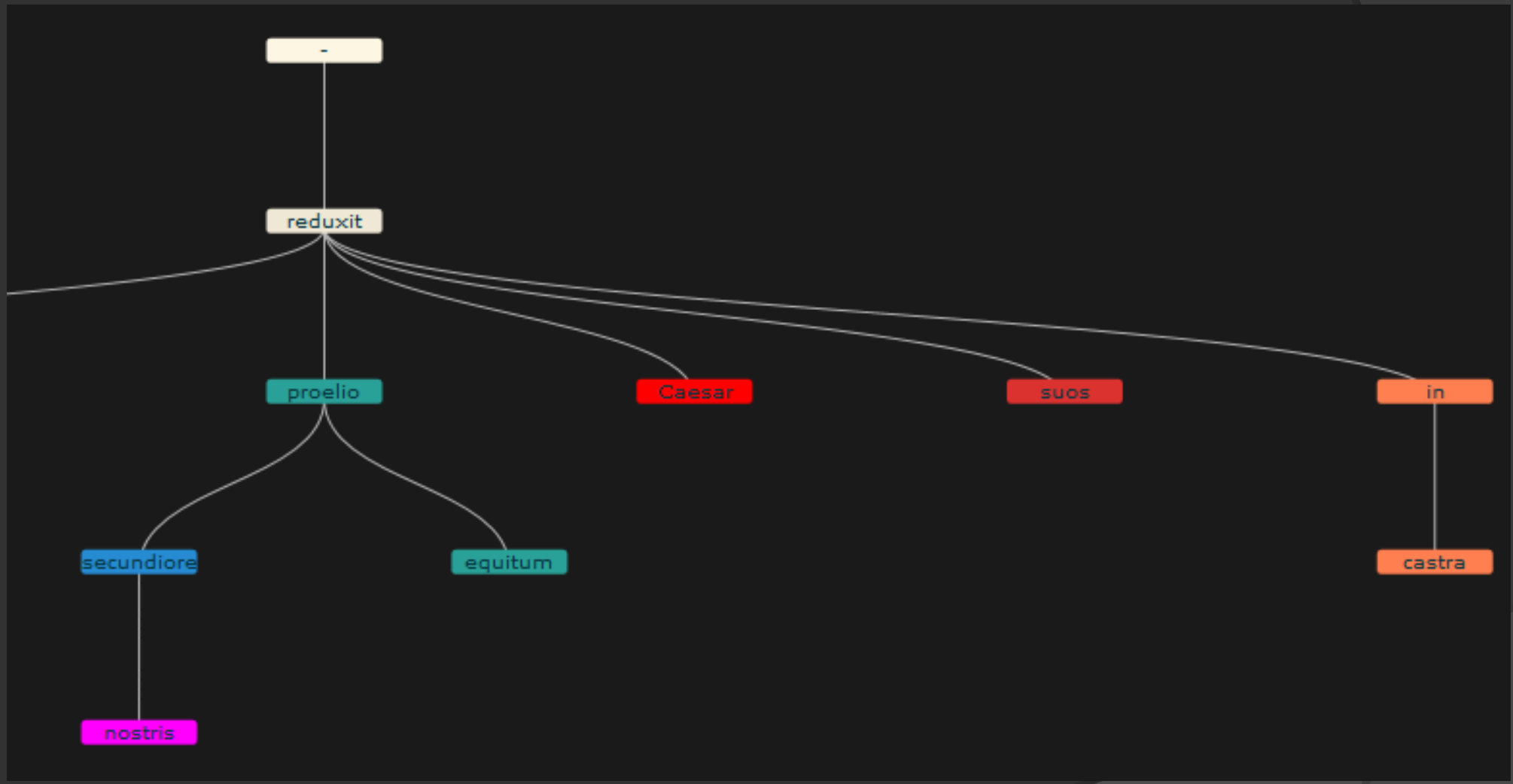
Constraint Behavior

amicus militis salutat.
(„The soldier‘s friend“)

amicus miles Caesaris pugnat.
(„Caesar‘s friendly soldier“)

SBJ / ATR (congruence) / ATR (genetive)

Full clause leaf



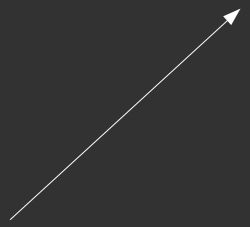
Ratings

Et id quidem nemini video Graecorum adhuc contigisse
134

Awards				Penalties			
50	for	1x	nominalized neutrum pl	-20	for	1x	has neutrum subject
30	for	1x	lone infinitive is regens	-70	for	1x	adjective is nominalized
25	for	1x	subj acc				
25	for	1x	close to regens				
20	for	1x	root encapsulation				
20	for	1x	object in valency group				
20	for	1x	is nested				
19	for	1x	root dependent				
10	for	1x	indirect object				
5	for	1x	animated				

Clause Merging

secundiore equitum proelio nostris Caesar suos in castra reduxit



Ubi neutri transeundi initium faciunt

Problems

Eo cum de improviso celeriusque omnium opinione venisset, Remi, qui proximi Galliae ex Belgis sunt, ad eum legatos Iccium et Andebrogium, primos civitatis, miserunt, qui dicerent se suaque omnia in fidem atque potestatem populi Romani permittere, neque se cum reliquis Belgis consensisse neque contra populum Romanum coniurasse, paratosque esse et obsides dare et imperata facere et oppidis recipere et frumento ceterisque rebus iuvare;

Permittere => consensisse => coniurasse => paratos esse => dare => facere => recipere => iuvare

Evaluation

Caesar LDT annotation*:

LAS	UAS	LA
84,7 %	91,0 %	89,7 %

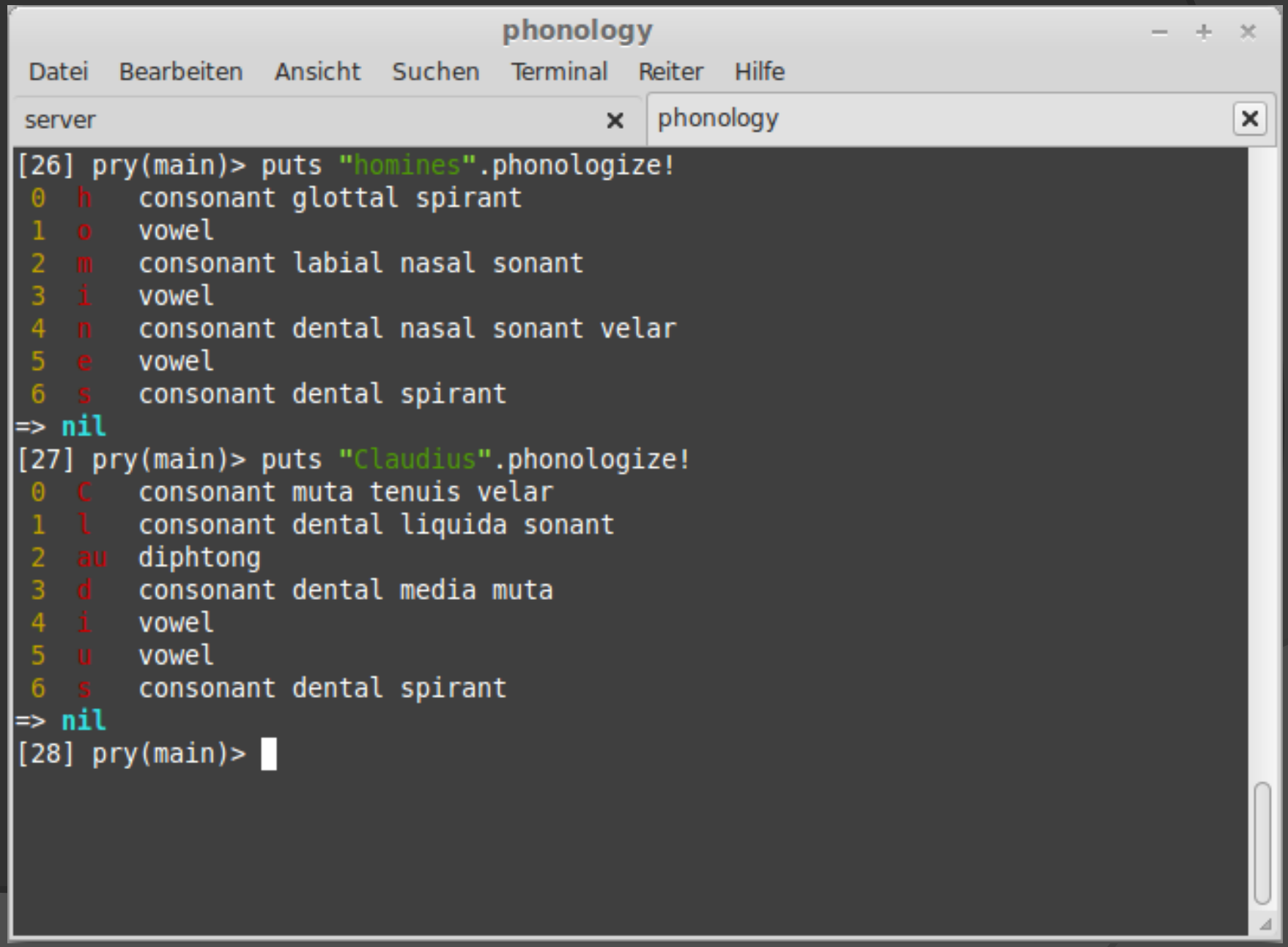
**preliminary results (85% of all tokens, without punctuation)*

40% of the parsed sentences at 100 %

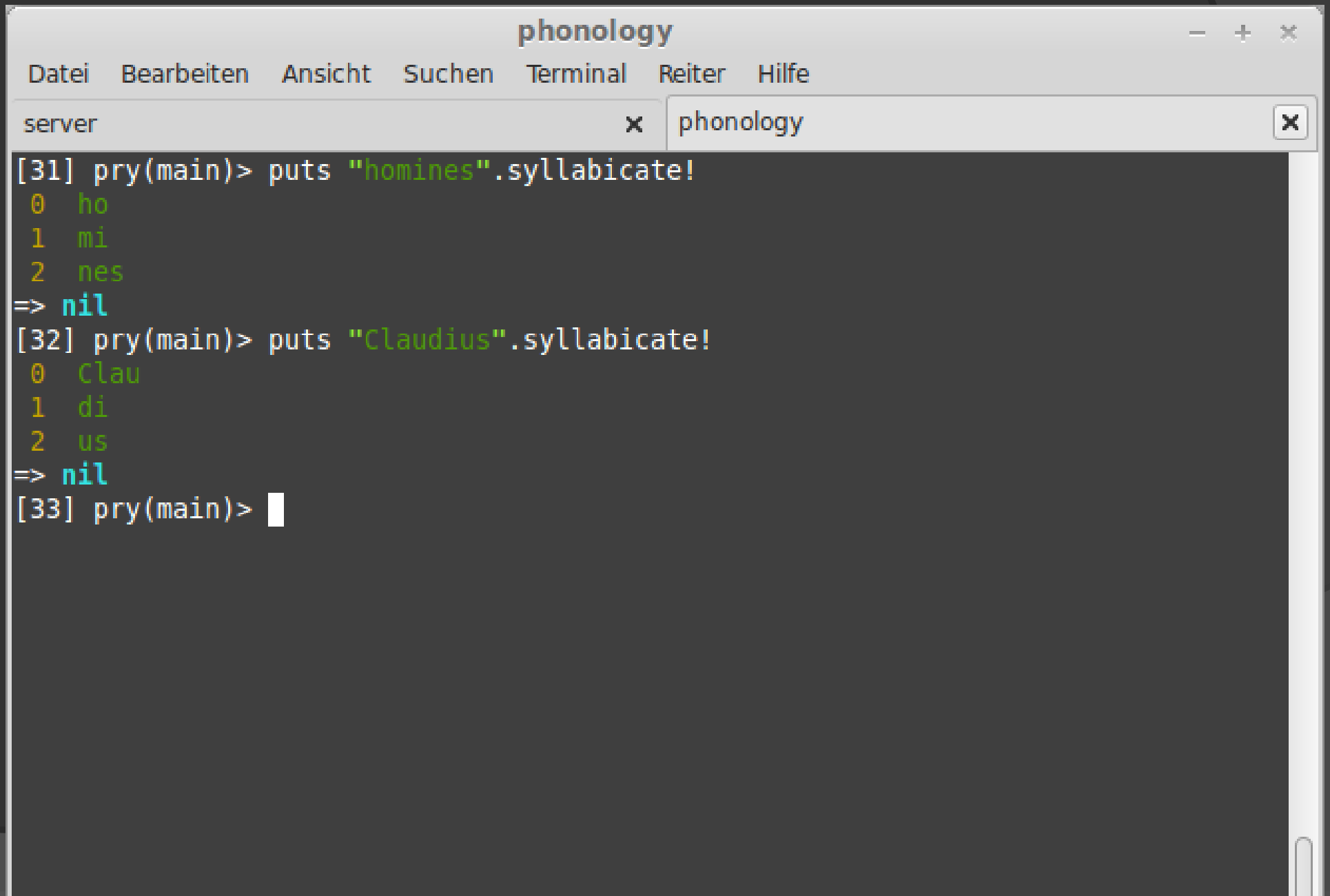
Challenges & Improvements

- Poetry

Basic-Tools: Phonology



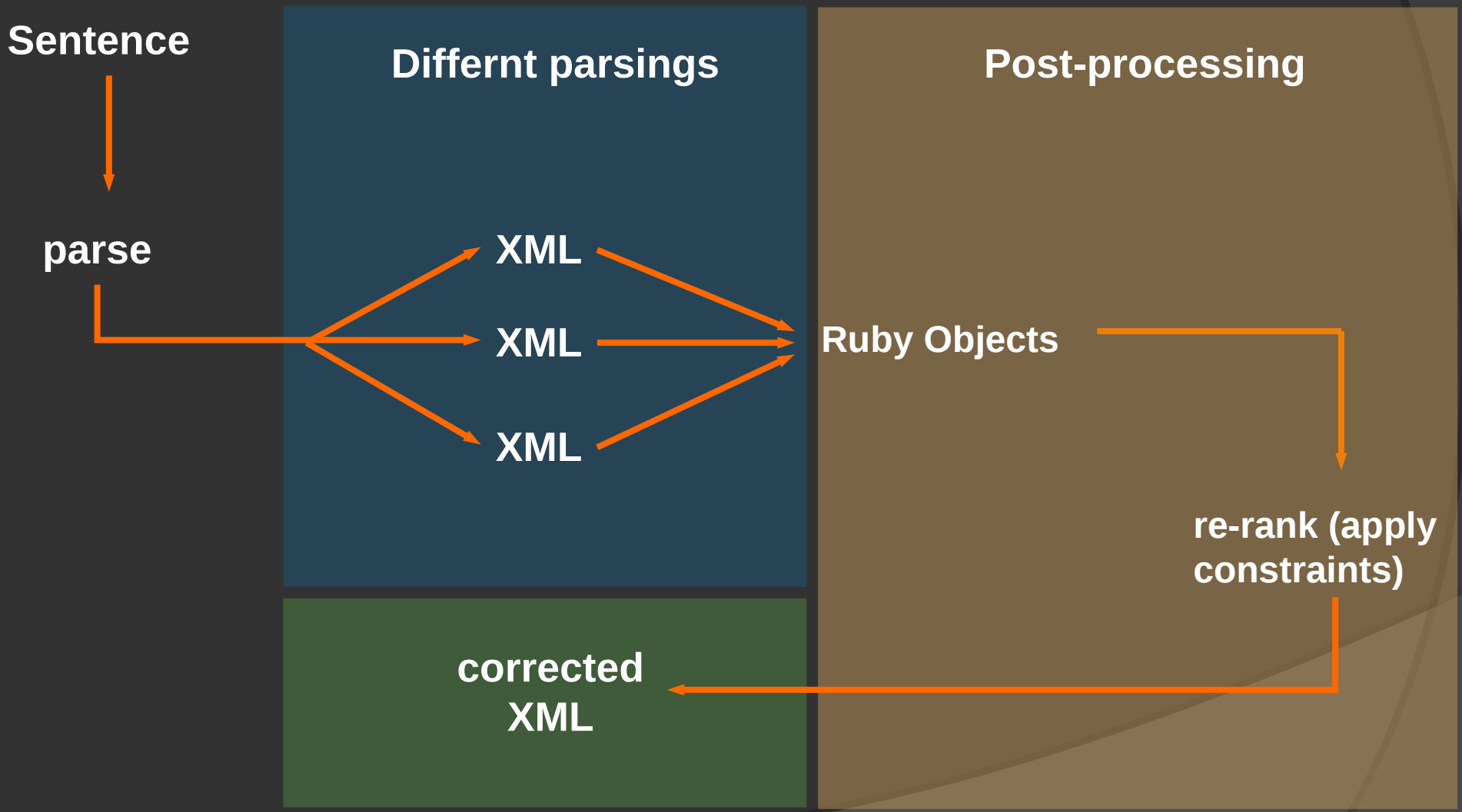
Basic-Tool: Syllables



Challenges & Improvements

- Poetry
- Semantics
- Dictionary/Database

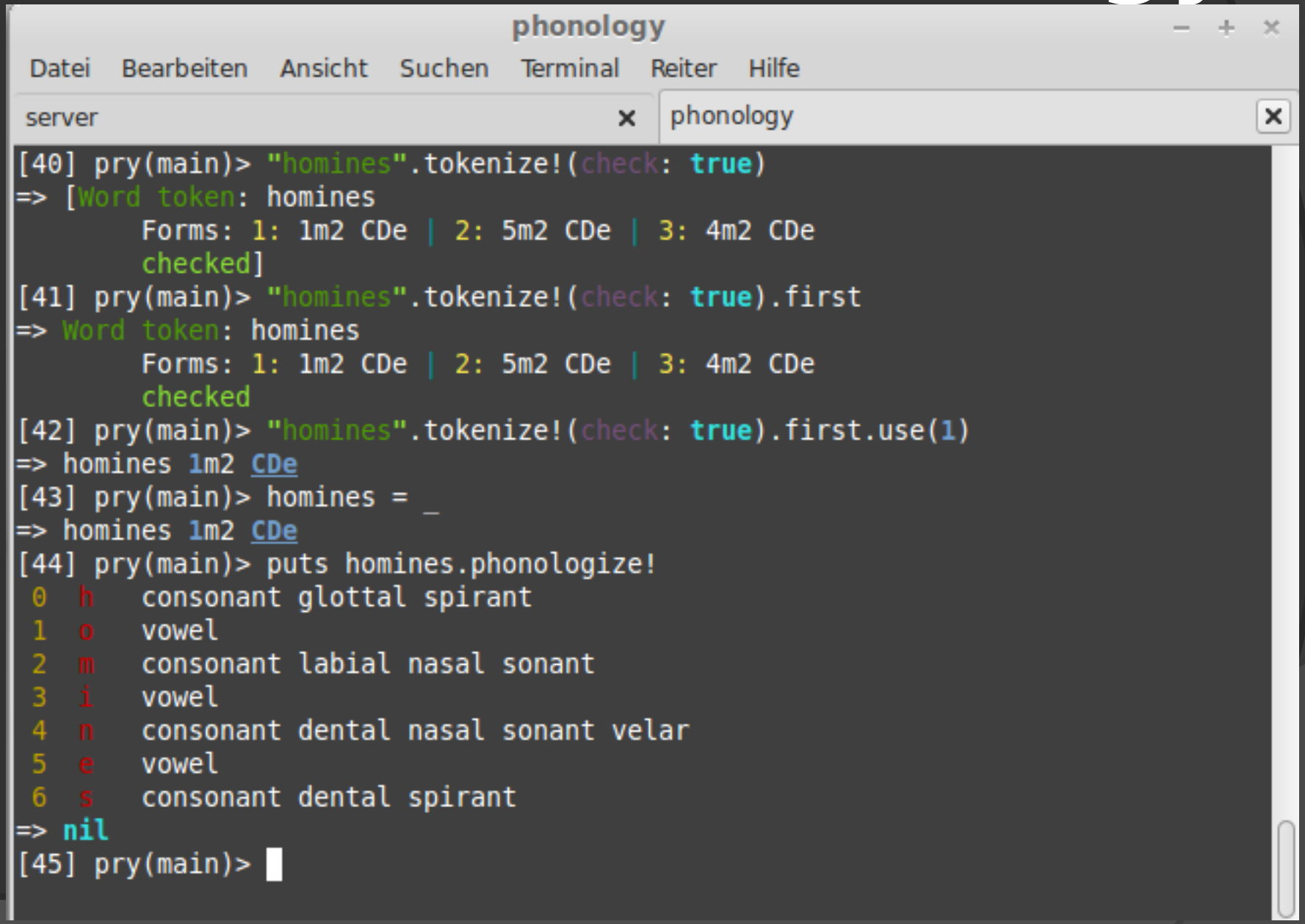
Post-Processing
Data-driven parsings



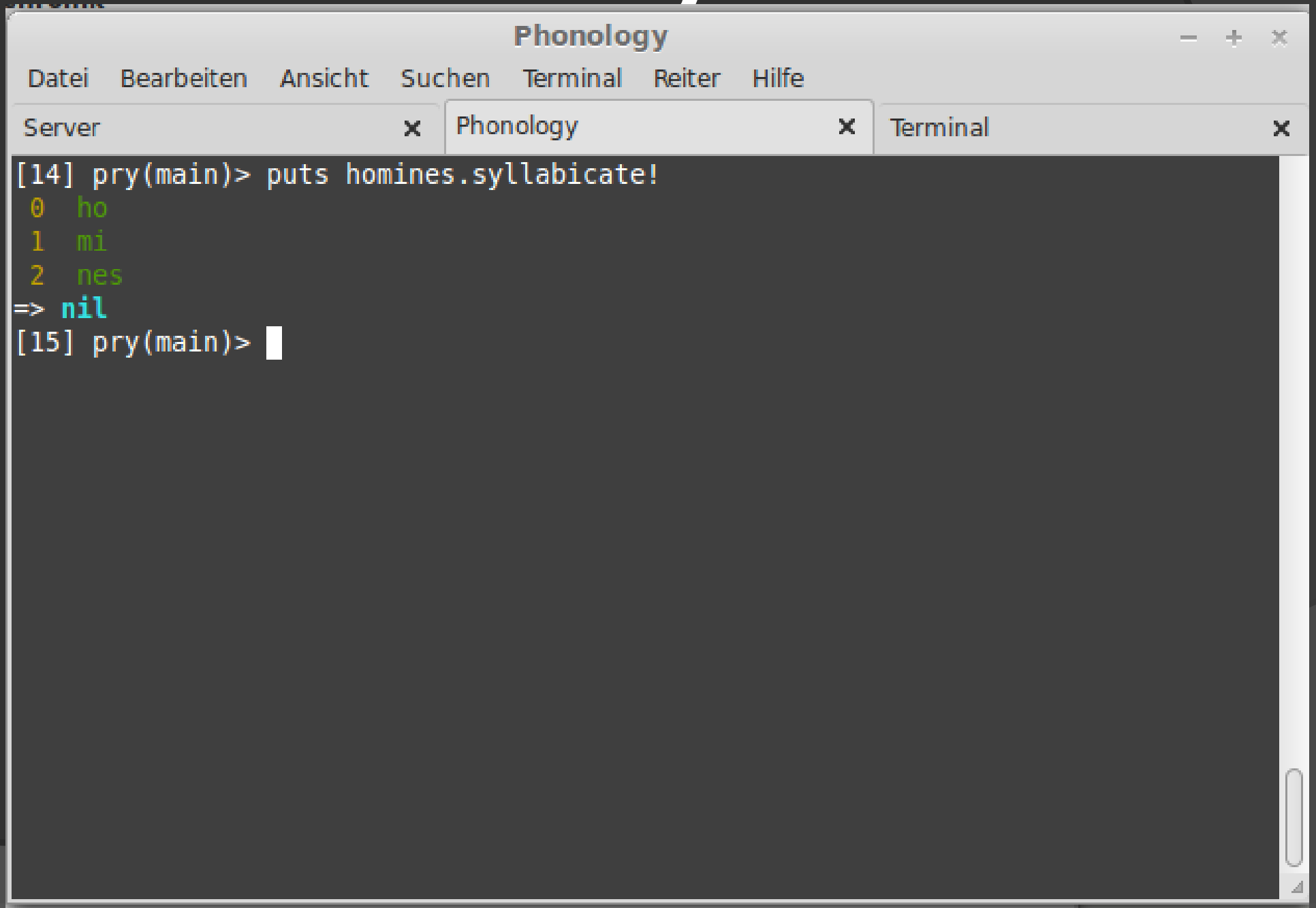
Thanks for your patience and attention!

to discuss…

Basic-Tools: Phonology



Basic-Tool: Syllables



Clause Finding

Caes. BG 2,34

…quem cum legione una miserat ad Venetos, Venellos, Osismos, Coriosolitas, Esuvios, Aulercos, Redones…

Parsing Poetry

Ovid, Meta., 10,529

Capta viri forma non iam Cythereia curat | litora

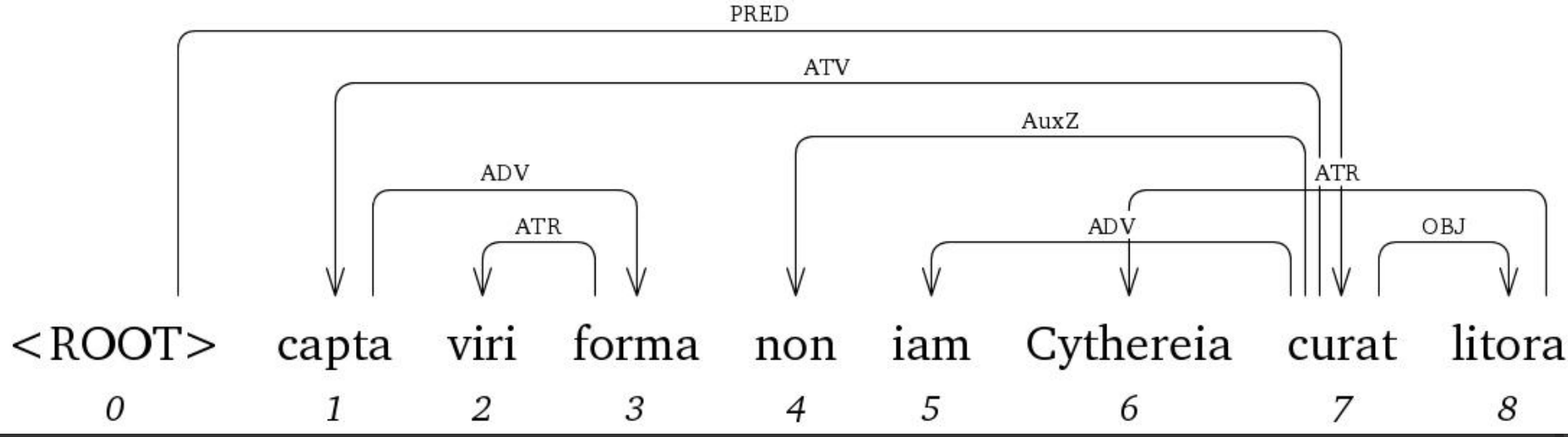
1: 1f1 PPP
2: 1n2 PPP
3: 4n2 PPP
4: 5n2 PPP
5: 5f1 PPP
6: 6f1 PPP
7: 1n2 PPP
8: 4n2 PPP
9: 5n2 PPP

1: 1f1 ADe
2: 5f1 ADe
3: 6f1 ADe
4: 1f1 AOD
5: 5f1 AOD
6: 6f1 AOD
7: 1n2 AOD
8: 4n2 AOD
9: 5n2 AOD
10: 2,1.i.a.pr ACo

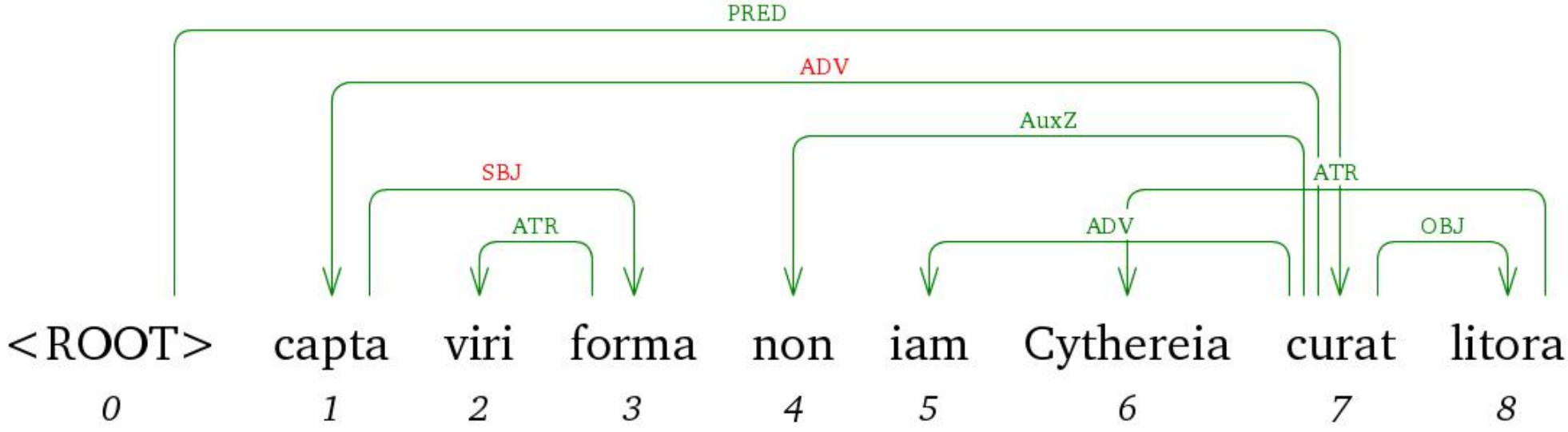
1: 1f1 AOD
2: 5f1 AOD
3: 6f1 AOD
4: 1n2 AOD
5: 4n2 AOD
6: 5n2 AOD

1: 1n2 CDe
2: 5n2 CDe
3: 4n2 CDe

Gold-Standard:



Result:



Parsing Poetry

Cāptă vĭrī fōrmā nōn iām Cŭthĕrēĭă cūrăt | lītŏră,

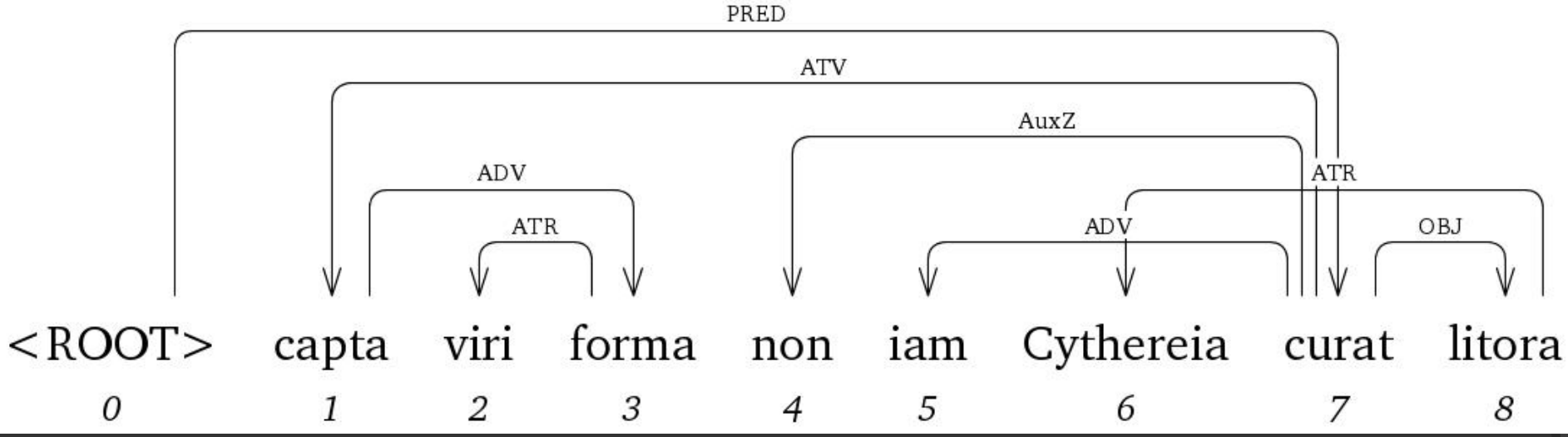
1: 1f1 PPP
2: 1n2 PPP
3: 4n2 PPP
4: 5n2 PPP
5: 5f1 PPP
~~6: 6f1 PPP~~
7: 1n2 PPP
8: 4n2 PPP
9: 5n2 PPP

~~1: 1f1 ADe~~
~~2: 5f1 ADe~~
3: 6f1 ADe
~~4: 1f1 AOD~~
~~5: 5f1 AOD~~
6: 6f1 AOD
~~7: 1n2 AOD~~
~~8: 4n2 AOD~~
~~9: 5n2 AOD~~
10: 2,1.i.a.pr ACo

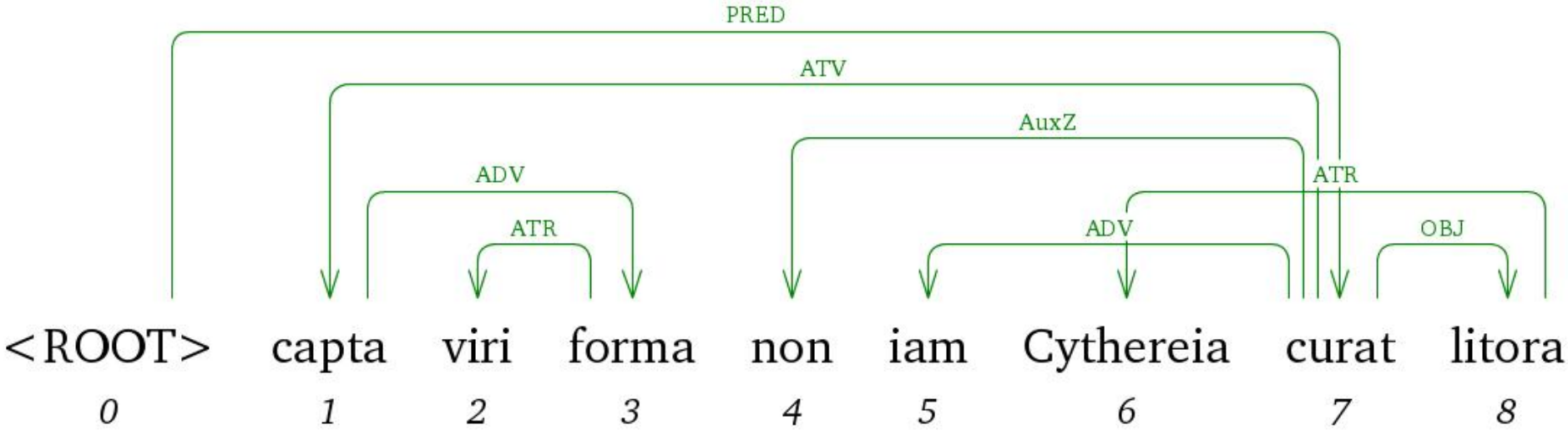
1: 1f1 AOD
2: 5f1 AOD
~~3: 6f1 AOD~~
4: 1n2 AOD
5: 4n2 AOD
6: 5n2 AOD

1: 1n2 CDe
2: 5n2 CDe
3: 4n2 CDe

Gold-Standard:



New result:



Performance/computational speed

Cum esset Caesar in citeriore Gallia in hibernis ita, uti supra demonstravimus, crebri ad eum rumores adferebantur litterisque item Labieni certior fiebat omnes Belgas, quam tertiam esse Galliae partem dixeramus, contra populum Romanum coniurare obsidesque inter se dare.

45 tokens => 7.3 seconds

Dat negotium Senonibus reliquisque Gallis, qui finitimi Belgis erant, uti ea, quae apud eos gerantur, cognoscant seque de his rebus certiorem faciant. **29 tokens => 1.7 seconds**

His nuntiis litterisque commotus Caesar duas legiones in citeriore Gallia novas conscripsit et inita aestate in ulteriorem Galliam, qui deduceret, Q. Pedium legatum misit.

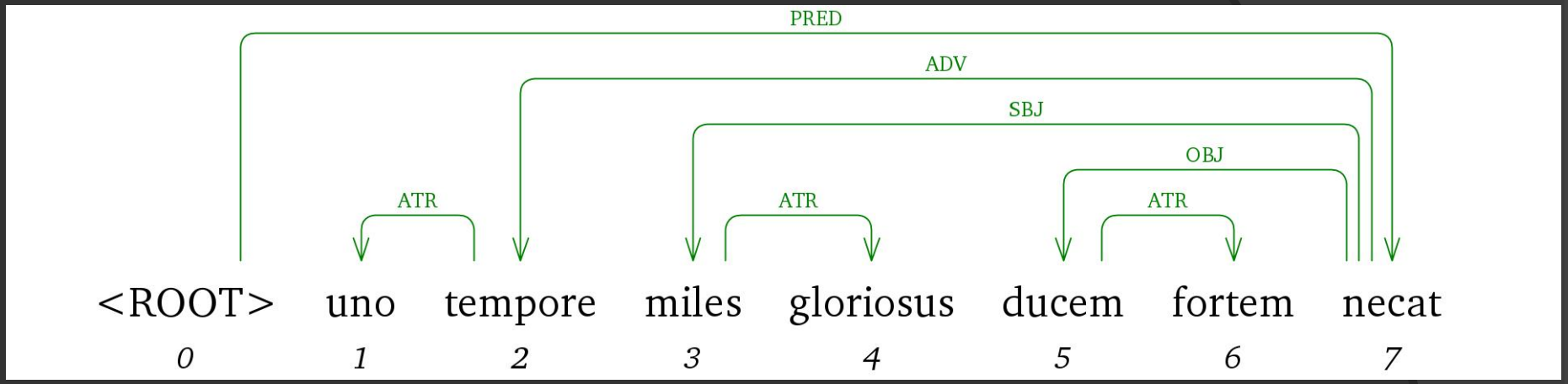
28 tokens => 4.4 seconds

Performance/computational speed

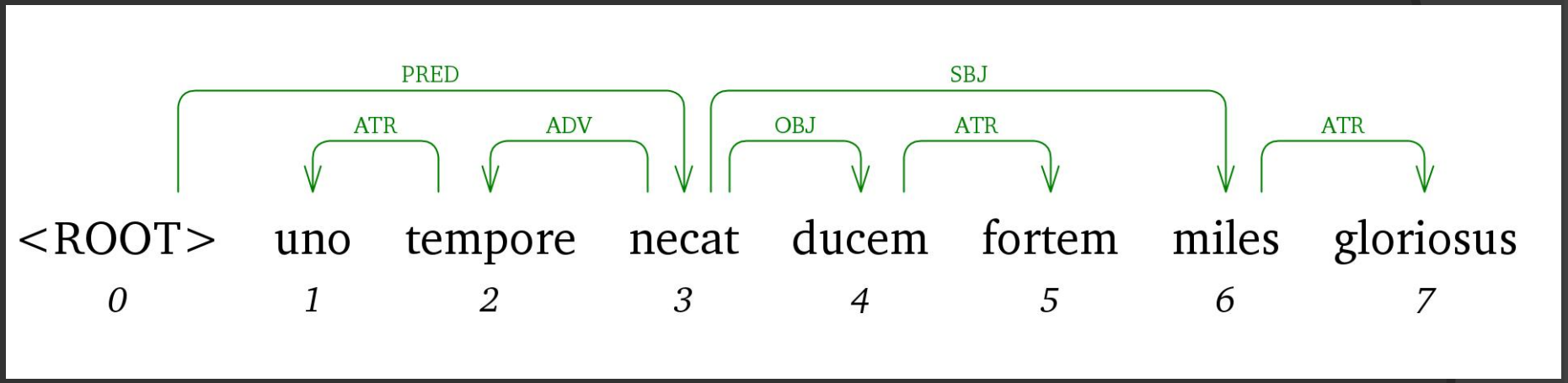
- Intel i5-3570k (3.4GHz), 8GB RAM (4 cores)
- All Caesar-tokens (~ 1400):

  2 min. 20 sec
- ~ 500 tokens per minute

SOV:



OSV:



Non-projectivity:

SOV: