Latin Morphology through the Centuries

Ensuring Consistency for Better Language Processing

Federica Gamba and Daniel Zeman

September 8, 2023

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

Outline - Latin Treebanks in UD

Available UD data I

1. Index Thomisticus Treebank (ITTB): texts by Thomas Aquinas and related authors. Philosophical Medieval Latin, XIII century.
2. Late Latin Charter Treebank (LLCT): early Medieval Latin charters written in Tuscany, Italy, in VIII-IX centuries. Legal/documentary genre.
3. Perseus: Classical Latin texts (e.g., by Cicero, Propertius, Sallust, Tacitus, Vergil) of different genres.
4. PROIEL: Vulgate New Testament translations plus excerpts from Caesar's *Gallic War*, Cicero's *Letters to Atticus*, Palladius' *Opus Agriculturae* and the first book of Cicero's *De officiis* (classical Latin, different genres).
5. UDante: literary texts - letters, treatises, poetry - by Dante Alighieri. Literary Medieval Latin (XIV century).

Available UD data II

		train	dev	test
ITTB	sents	22,775	2,101	2,101
	words	390,785	29,888	29,842
LLCT	sents	7,289	850	884
	words	194,143	24,189	24,079
Perseus	sents	1,334	0	939
	words	18,184	0	10,954
PROIEL	sents	16,196	1,233	1,260
	words	177,558	13,917	14,091
UDante	sents	926	376	419
	words	30,441	11,611	13,451

Table 1: Size of UD Latin treebanks in v2.12.

Outline - Motivation

Latin variability

- Time span over **two millennia** (VII century BC to now).
- Wide **geographical expanse**.
- Differences entailed by literary **genre**.
  - e.g. poetry/prose, plus further distinctions: charters, letters, treatises, ...
- However, also **divergences in annotation** (despite UD).
  - different teams
  - different moments of the development of UD guidelines

↓

- Significant **drop** in parsing performances when a model is applied to data that differ from training data.

Outline - Harmonisation Overview

Harmonisation Workflow

- Focus on **morphological features** only.
- **Workflow** to detect not allowed and missing (yet required) features:
  - **UDapi** (Popel et al., 2017) **block** run on input data, i.e. treebanks from Gamba and Zeman (2023).
  - Output stored in `html` file that highlights spotted inconsistencies.
  - Difference in the set of morpho features in UDante-ITTB-LLCT *vs* Perseus-PROIEL. Hence, **two levels of coherence**:
    1. lower level (default): only core information required. E.g., all pronouns must have a `PronType`, all verbs `VerbForm` and `Aspect`.
    2. higher level: additional information, e.g. `InflClass`, expected and allowed.
- Data manipulation through Python scripts exploiting **UDapi**.

Outline - The `markFeatsBugs` Block

The markFeatsBugs Block I

```
if re.match(r'^(VERB|AUX)$', node.upos):
    rf = ['VerbForm', 'Aspect']
    af = {'VerbForm': ['Inf', 'Fin', 'Part', 'Conv'],
        'Aspect': ['Imp', 'Inch', 'Perf', 'Prosp']}
    if node.feats['VerbForm'] not in ['Part', 'Conv']:
        rf.append('Tense')
        af['Tense'] = ['Past', 'Pqp', 'Pres', 'Fut']
    if node.upos == 'VERB' or (node.upos == 'AUX' and node.lemma != 'sum'):
        rf.append('Voice')
        af['Voice'] = ['Act', 'Pass']
    if node.feats['VerbForm'] == 'Fin':
        rf.extend(['Mood', 'Person', 'Number'])
        af['Mood'] = ['Ind', 'Sub', 'Imp']
        af['Person'] = ['1', '2', '3']
        af['Number'] = ['Sing', 'Plur']
    [...]
```

The markFeatsBugs Block II

```
    elif node.feats['VerbForm'] == 'Conv':
        rf.extend(['Case', 'Gender', 'Number'])
        af['Case'] = ['Abl', 'Acc']
        af['Gender'] = ['Masc']
        af['Number'] = ['Sing']
        af['Voice'] = ['Act']

    if self.flavio:
        af['Compound'] = ['Yes']
        af['Variant'] = ['Greek']
        af['NameType'] = ['Ast', 'Cal', 'Com', 'Geo', 'Giv', 'Let', 'Lit', 'Met',
            'Nat', 'Rel', 'Sur', 'Oth']
        af['InflClass'] = ['Ind', 'IndEurA', 'IndEurE', 'IndEurI', 'IndEurO',
            'IndEurU', 'IndEurX']
```

HTML Output

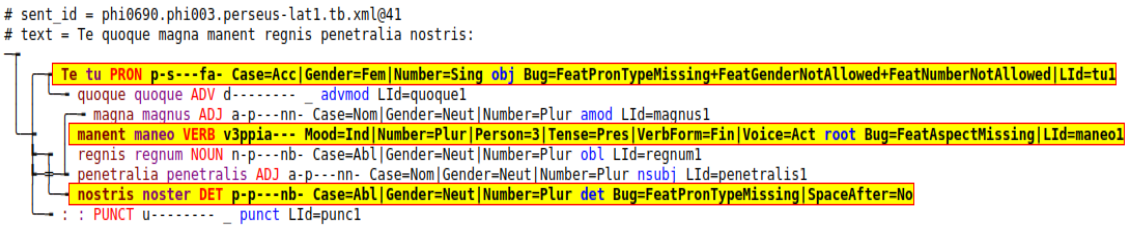

Figure 1: Example of the HTML file highlighting bugs found in the data.

An Example

Verbal system

- Reorganisation of **non-finite verbal features**, as in Cecchini (2021).
  - Gerund and gerundive forms as `VerbForm=Part` with `Aspect=Prosp` (e.g, *faciendum*, *dicendus*).
  - Supine forms as `VerbForm=Conv` with `Aspect=Prosp` (*visum*, *visu*).
  - Traditional terminology stored in MISC (e.g., `TraditionalMood=Gerund`).
- **AUXs** in ITTB: `Aspect`, `Mood`, `Person` and `Tense` missing (*sunt* 'they are').

Outline - Results

MLAS

	ittb.mdl		llct.mdl		udante.mdl	
	before	after	before	after	before	after
ITTB	**78.97**%	**80.74**%	16.56%	19.07%	33.14%	39.59%
LLCT	12.22%	17.67%	**89.46**%	**90.04**%	12.59%	18.02%
Perseus	22.63%	35.20%	11.57%	16.92%	16.25%	27.29%
PROIEL	22.23%	41.32%	14.86%	22.74%	17.17%	30.61%
UDante	25.06%	29.95%	12.21%	14.77%	**35.96**%	**35.32**%

	perseus.mdl		proiel.mdl	
	before	after	before	after
ITTB	19.45%	27.87%	22.13%	40.05%
LLCT	9.12%	16.63%	15.98%	24.25%
Perseus	**38.86**%	**40.21**%	31.33%	38.66%
PROIEL	27.64%	35.92%	**68.49**%	**71.23**%
UDante	10.64%	17.37%	13.45%	25.40%

Table 2: Comparison of Stanza MLAS scores.

Morphological Features

	ittb.mdl		llct.mdl		udante.mdl	
	before	after	before	after	before	after
ITTB	**95.70**%	**96.15**%	57.07%	66.19%	66.22%	75.34%
LLCT	56.92%	63.95%	**96.89**%	**96.81**%	55.73%	63.47%
Perseus	57.29%	72.49%	48.66%	57.23%	49.75%	64.63%
PROIEL	49.88%	75.90%	48.31%	60.97%	44.53%	67.10%
UDante	62.47%	69.85%	48.56%	56.32%	**79.39**%	**79.30**%

	perseus.mdl		proiel.mdl	
	before	after	before	after
ITTB	55.19%	72.91%	52.14%	79.97%
LLCT	53.53%	65.33%	57.07%	71.87%
Perseus	**78.02**%	**77.86**%	70.01%	79.51%
PROIEL	66.57%	75.95%	**90.91**%	**92.72**%
UDante	45.89%	63.42%	46.22%	70.64%

Table 3: Comparison of Stanza accuracy scores on morphological features.

Outline - Conclusive Remarks

Summary & What's next

- Observed impact:
  1. MLAS / morpho features: clear improvement, up to **+19%** and **+26%** respectively (ITTB model on PROIEL data).
  2. LAS / UAS: no pattern, no widespread or substantial improvements.
- Lower annotation detail in Perseus and PROIEL (cf. `PronType` missing/under-specified).
- Ready to contribute the harmonised treebanks to next UD official release.
- UDapi block contributed to the official UDapi repository; scripts and harmonised treebanks available on GitHub as well.

- No more harmonisation, yet continuous effort needed at community level.
- Higher degree of annotation consistency (i.e, more homogeneous data) allowing now to investigate the actual reasons for variability in parsing.

Thank you!