Challenges in converting the *Index Thomisticus* Treebank into Universal Dependencies

Flavio Massimiliano Cecchini[1], Marco Passarotti[1], Paola Marongiu[2], Dan Zeman[3]

[1]CIRCSE, Università Cattolica del Sacro Cuore (Milan)

[2]Università degli Studi di Pavia

[3]Charles University in Prague, Institute of Formal and Applied Linguistics (ÚFAL)

Universal Dependencies Workshop 2018 (UDW 2018),
EMNLP 2018, Bruxelles, Belgium

Latin Treebanks in Universal Dependencies

- **PROIEL** (Haug & Jøhndal 2008)
  New Testament and Classical works (Caesar, Cicero, Palladius): 18,400 sentences, 199,958 tokens
- **Perseus Latin Dependency Treebank, LDT** (Bamman & Crane 2006)
  Parts of the *Vulgata* by Jerome and of Classical works (Caesar, Cicero, Ovid, Vergil, …): 2,273 sentences and 29,138 tokens
- ***Index Thomisticus* Treebank, IT-TB** (Passarotti 2011)
  Annotation of works by Thomas Aquinas. Divided into two non-overlapping parts:
  - concordances of the lemma *forma* in *Scriptum super Sententiis Magistri Petri Lombardi (SsS)*, *Summa contra Gentiles (ScG)* and *Summa Theologiae*: 3,290 sentences and 75,489 tokens
  - the first three books of the *ScG*: 17,721 sentences and 277,547 tokens

  Total: 21,011 sentences and 353,036 tokens

Three different non-UD standards!

↓

Common guidelines established for homogeneous UD annotation/conversion

Conversion process 1. Harmonisation to Prague Dependency Treebank (PDT) style with Treex (Perl)

Parts of Speech: Treating IT-TB-specific annotation

ISSUE: The IT uses a coarse, “tripartite” PoS-tagging.

IT-TB	UD	Examples
1 = nominals	ADJ, NOUN, NUM, PRON, PROPN	*humanus*, *anima*, *quatuor*, *quis*, *Augustinus*
2 = verbal nominal inflections (e.g. participles)	VERB	*dico*: non-finite forms like *dictus*, *dicentes*,…
3 = verbs	VERB, AUX	*dico*: finite forms like *dicitur*, *dicunt*,…; *sum* as copula: *est beatus*,…
4 = invariables	ADP, ADV, CCONJ, PART, SCONJ, X, (DET)	*in*, *substantialiter/enim*, *et*, *amen*, *quod*, *litt.*, (*ly*)
- = punctuation	PUNCT	.

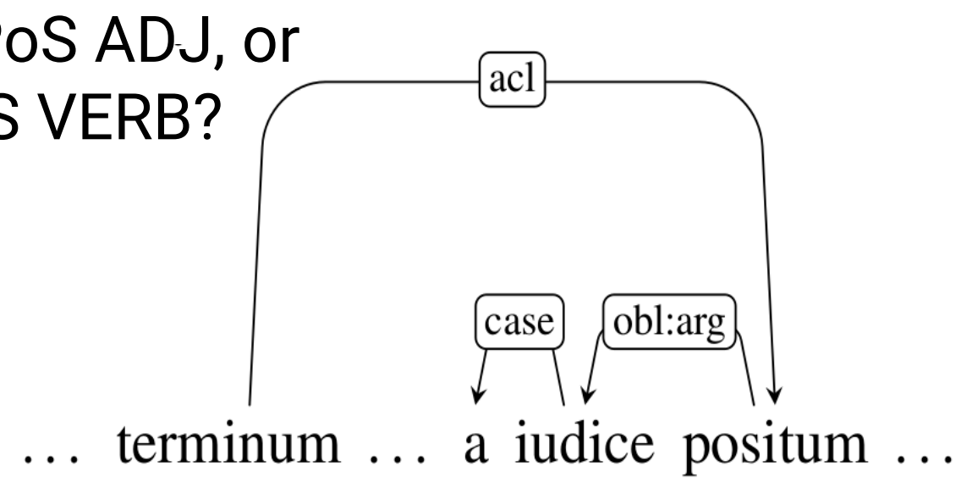
SOLUTION: List of hard-coded lemmas for adjectives, nouns, pronouns and proper nouns; *ad hoc* labeling rules for invariable lemmas.

Verbal nouns: Keeping syntactic relations

Should verbal nouns (participles, gerunds, gerundives, supines) be treated as verbs or nominals (e.g. adjectives)?

*Transgredi autem terminum hunc a iudice **positum**, non est secundum se malum...* – *ScG* Lib. III, Cap. CXXIX
‘But to pass over a boundary line **set up** by a judge is not essentially evil...’

positum (perfect participle of *pono* ‘to set’):
- lemma *positus* and PoS ADJ, or
- lemma *pono* and PoS VERB?



Tracing verbal nouns back to their verbal roots allows one to keep the core/oblique distinction between their complements.

Adverbs: Implementing common guidelines

Derived adverbs in the IT receive the corresponding adjective or verb as lemma:

- *continue, continuius, continuissime*
  → *continuus* (ADJ)
- *abundanter, abundantius, abundantissime*
  → *abundo* (VERB)

In the common guidelines, their lemma has to be the positive degree of the adverb:

- *continue, continuius, continuissime*
  → *continue* (ADV)
- *abundanter, abundantius, abundantissime*
  → *abundanter* (ADV)

The IT-TB features an “adverbial case” which helps to correct the lemmatisation of adverbs.

`OComp`: Script technicalities

The IT-TB has an additional syntactic relation with respect to the PDT: `OComp`

*Posuerunt mundum **aeternum**.* – *ScG* Lib. II, Cap. XXXVIII
‘They asserted the world’s eternity.’

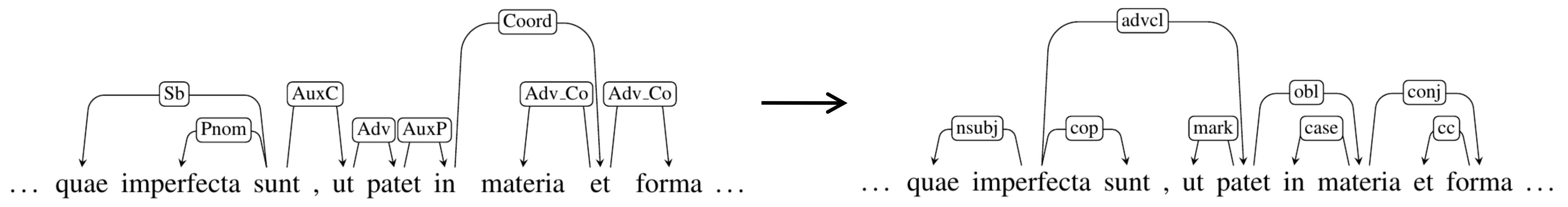
It corresponds to *deprel* `xcomp` in UD. It is not retrievable if treated as `Obj` during conversion.

The relation `OComp` has to be explicitly marked in the XML file as a special case.

Conversion process 2. Conversion proper to UD standard

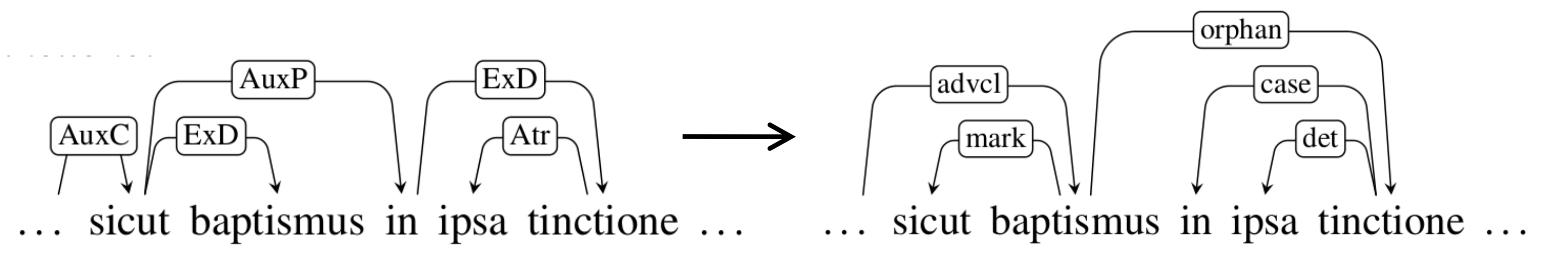
The major difference between the PDT and UD styles is the treatment of coordinations, copulas, subordinate clauses and prepositional phrases.

... quae imperfecta sunt, ut patet in materia et forma...
– *SsS* Lib. I, Dist. III, Qu. II, Art. II
‘...which are imperfect, as it clearly appears in matter and form...’


Sb
Pnom
AuxC
Adv
AuxP
Coord
Adv_Co
Adv_Co
… quae imperfecta sunt , ut patet in materia et forma …
advcl
nsubj
cop
mark
obl
case
conj
cc
… quae imperfecta sunt , ut patet in materia et forma …

Elliptical constructions (The IT-TB presents only cases of verbal ellipses)

In illis autem sacramentis quae perficiuntur in usu materiae, sicut baptismus [perficitur] in ipsa tinctione. . . – *SsS* Lib. IV, Dist. VII, Qu. I, Art. III
‘In those sacraments, however, which are accomplished through the use of matter, like baptism [is accomplished] through the submersion itself. . . ’

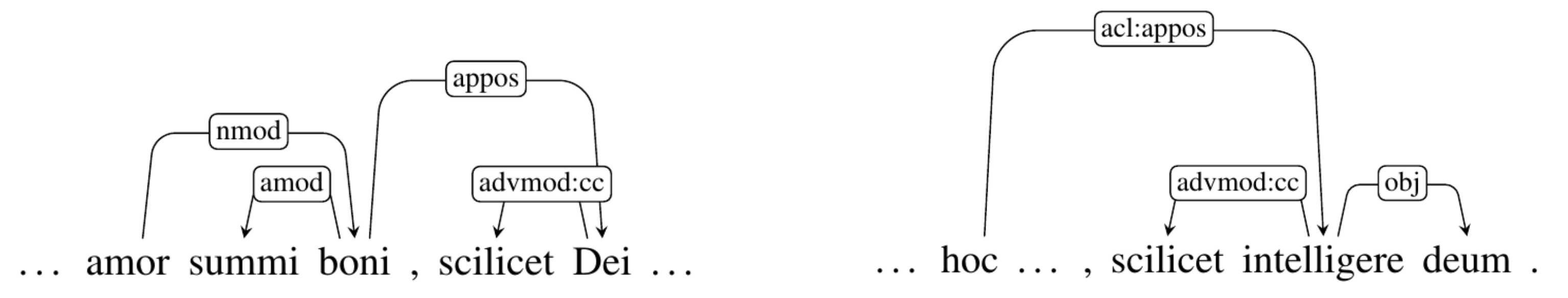

AuxC
ExD
AuxP
ExD
Atr
… sicut baptismus in ipsa tinctione …
advcl
mark
orphan
case
det
… sicut baptismus in ipsa tinctione …

One of the `ExD` siblings has to be promoted to head of the elliptical clause:
→ Cross-checking parts of speech and grammatical cases allows one to select the most probable node

Appositions – Complex restructuring of subtrees

. . . amor summi boni, scilicet Dei. . .
– *ScG* Lib. III, Cap. CXVI
’. . . the love of the highest good, namely, God. . .’

... hoc ... scilicet intelligere deum.
– *ScG* Lib. III, Cap. XXV
‘... this ... namely, to understand God.’


nmod
amod
appos
advmod:cc
… amor summi boni , scilicet Dei …
acl:appos
advmod:cc
obj
… hoc … , scilicet intelligere deum .

- The connecting particle *scilicet* ‘that is’ acts neither as an adverbial modifier nor as a co-ordinating conjunction
- Non-nominal elements can also appear in appositional constructions:
  → Dependency relations extended with `cc` and the new `appos` subtype

Evaluation

Manually-annotated gold standard built from 994 sentences randomly extracted from the corpus.

	LAS	LA	UAS	PoS	Lemma
Old	84.8%	87.9%	94.2%	95.5%	95.2%
New	97.0%	98.0%	98.3%	97.4%	99.8%

LAS: labeled attachment score
LA: label accuracy
UAS: unlabeled attachment score
PoS: accuracy of part-of-speech tagging
Lemma: Accuracy of lemma assignment

References

- David Bamman and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 67–78, Prague, Czech Republic. Univerzita Karlova.
- Dag T. T. Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, Marrakesh, Morocco. European Language Resources Association (ELRA).
- Paola Marongiu. 2018. Challenges in Converting the *Index Thomisticus* Treebank into Universal Dependencies. Master's degree thesis, Università degli Studi di Pavia, September 2018.
- Marco Passarotti. 2011. Language resources. The state of the art of Latin and the *Index Thomisticus* treebank project. In *Corpus anciens et Bases de données*, number 2 in ALIENTO. Échanges sapientiels en Méditerranée, pages 301–320, Nancy, France. Presses universitaires de Nancy.

https://itreebank.marginalia.it **https://lila-erc.eu**

This project has received funding from the European Research Council (ERC) European Union’s Horizon 2020 Research and Innovation programme under grant agreement No. 769994.

Poster designed by Greta Franzini