Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD

Federica Gamba and Daniel Zeman
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia
gamba,zeman@ufal.mff.cuni.cz

Abstract
This paper presents the harmonisation process carried out on the five treebanks available for Latin in Universal Dependencies, with the aim of eliminating the discrepancies in their annotation styles. Indeed, this is the first issue to be addressed when parsing Latin, as significant drops in parsing accuracy on different Latin treebanks have been repeatedly observed. Latin syntactic variability surely accounts for this, but parsing results are as well affected by divergent annotation choices. By analysing where annotations differ, we propose a Python-based alignment of the five UD treebanks. Consequently, the impact of annotation choices on accuracy scores is assessed by performing parsing experiments with UDPipe and Stanza.

1 Introduction
A significant number of resources is available for Latin. With respect to syntax, notable are the five treebanks in Universal Dependencies\(^1\) (de Marneffe et al., 2021), which represent a remarkable amount of data. Here is an overview:

- **Index Thomisticus Treebank** (ITTB) (Passarotti, 2019): encompassing texts by Thomas Aquinas (1225–1274) and other authors related to Thomas, it represents an example of philosophical Medieval Latin. It is the largest of the Latin treebanks.

- **Late Latin Charter Treebank** (LLCT) (Cechchini et al., 2020b): it consists of Early Medieval (VIII–IX century) Latin charters written in Tuscany, Italy, all representing the legal/documentary genre.

- **Perseus** (Bamman and Crane, 2011): it includes some of the most representative Classical Latin texts (e.g., by Augustus, Cicero, Vergil, Propertius, Sallust, Tacitus) of different genres. It is the smallest treebank in terms of number of tokens.


- **UDante** (Cechchini et al., 2020a): it includes literary texts (letters, treatises, poetry) by Dante Alighieri, corresponding to literary Medieval Latin (XIV century).

The treebanks highly differ in terms of included texts and size (see Table 1), as well as in annotation. Indeed, despite the five treebanks all following the UD annotation guidelines, some differences in the annotation scheme persist. Specifically, the treebanks have been annotated by different teams and in different moments of the development of UD guidelines, resulting in different annotation choices. Thus, despite the remarkable effort made by the UD project, divergences can still be observed at all annotation levels, from word segmentation to

<table>
<thead>
<tr>
<th>Treebank</th>
<th>train sents</th>
<th>train words</th>
<th>dev sents</th>
<th>dev words</th>
<th>test sents</th>
<th>test words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITTB</td>
<td>22,775</td>
<td>390,785</td>
<td>2,101</td>
<td>29,888</td>
<td>2,101</td>
<td>29,842</td>
</tr>
<tr>
<td>LLCT</td>
<td>7,289</td>
<td>194,143</td>
<td>850</td>
<td>24,189</td>
<td>884</td>
<td>24,079</td>
</tr>
<tr>
<td>Perseus</td>
<td>1,334</td>
<td>18,184</td>
<td>0</td>
<td>0</td>
<td>939</td>
<td>10,954</td>
</tr>
<tr>
<td>PROIEL</td>
<td>15,917</td>
<td>172,133</td>
<td>1,234</td>
<td>13,939</td>
<td>1,260</td>
<td>14,091</td>
</tr>
<tr>
<td>UDante</td>
<td>926</td>
<td>30,441</td>
<td>376</td>
<td>11,611</td>
<td>419</td>
<td>13,451</td>
</tr>
</tbody>
</table>

Table 1: Size of UD Latin treebanks in v2.10.
lemmatisation, POS tags, morphology, and syntactic relations. In the present work we focus on the syntactic annotation. Our interventions mainly concern dependency relations, but comparable work will be needed also for lemmas and POS tags.

This study aims to syntactically harmonise the five Latin treebanks, as well as to assess the impact of different annotation choices on parsing accuracy. Section 2 motivates the present study. Section 3 presents an overview of the alignment process, while in Section 4 the harmonising interventions are highlighted in more detail. Section 5 reports the parsing scores on the aligned treebanks, demonstrating the impact of diverse annotations on parsing. Finally, Section 6 presents the conclusions and future research directions.

2 Related Work and Motivation

Parsing accuracy scores on Latin texts drop significantly when a model is applied to data that differ from those it was trained on. The issue is of course more general and concerns out-of-domain data, but with respect to Latin it is strongly intertwined with the issue of its syntactic variability. Indeed, spread over a span of more than two millennia and all across an area that corresponds to Europe, the Latin language has undergone a number of significant changes, which affected the syntactic layer as well. To be able to investigate genuine syntactic diversity, first we have to ask how much the observed drop in parsing performance is due to divergent annotation styles. A deeper understanding, and possibly levelling of such divergences would allow to isolate the impact of annotation choices and highlight intra-linguistic syntactic variability.

Such syntactic diversity, leading to lower parsing accuracies, has been repeatedly noted. For instance, Passarotti and Ruffolo (2010) and Ponti and Passarotti (2016) observed how performances drop when a model is employed to parse out-of-domain data, while Passarotti and Dell’Orletta (2010) dealt with the need of adapting a pre-existing parser to the specific processing of Medieval Latin. The issue of Latin variability has also been addressed in the EvaLatin campaigns (Sprugnoli et al., 2020; Sprugnoli et al., 2022), devoted to the evaluation of NLP tools for Latin. On the other hand, the issue of inconsistent annotations is not unprecedented. Methods for inconsistency detection in treebanks have been proposed e.g. by Dickinson and Meurers (2003), Volokh and Neumann (2011), Ambati et al. (2011), de Marneffe et al. (2017), Aggarwal and Zeman (2020), and Aggarwal and Alzetta (2021).

With respect to Latin, a huge effort towards harmonisation has been made by the LiLa project3. Within the framework of Linguistic Linked Open Data, LiLa seeks to overcome the different lemmatisation criteria through a pivotal use of lemmas and hypolemmas in a knowledge base.

3 Alignment Process

For the alignment process we decide to model our interventions on the 2.10 version of the UDante treebank, which was released in May 2022. This choice is motivated by several factors:

- UDante is the only Latin treebank that has been annotated directly in UD, rather than being converted from another framework; conversion errors are thus ruled out.
- It is the newest Latin treebank in UD, meaning that it follows the latest version of the UD guidelines.
- It is developed by the same team as the other non-neglected Latin treebanks (ITTB and LLCT); this team has also defined the UD guidelines for Latin.5

For all these reasons, UDante should be the Latin treebank most conforming to the current UD guidelines. Hence when aligning the annotation decisions in individual treebanks, we try to push them towards those of UDante. This should not be understood as pushing the language towards that of the genre, geographical location or historical period of UDante. Changes that we do are about annotation guidelines, and while some of them may address phenomena that are not present in all varieties of

3See https://lila-erc.eu/.
4As of the latest UD release, 2.10 (May 2022) Neglected is a technical label of the UD infrastructure, assigned to treebanks after three years since the oldest validation error. See the UD Validation Report at http://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/validation-report.pl.
5See https://universaldependencies.org/la/index.html#documentation.
Latin, the guidelines would not be different for
different varieties.

As mentioned in Section 2, the interventions
mainly focus on dependency relations, yet not ex-
clusively: spotted conversion and random errors
are corrected, as well as some inconsistencies in
terms of lemmatisation and POS tags.

As a starting point of our alignment process, we
choose the treebanks’ train, dev and test sets as
available in their UD GitHub dev branch as of
August 30th, 2022. The treebanks are then aligned
through Python scripts, specifically designed for
each treebank. To manipulate data we exploit
Udapi (Popel et al., 2017), a framework provid-
ing an application programming interface for UD
data. Our scripts are openly available on GitHub,6
together with the aligned treebanks. Moreover, we
are ready to contribute the harmonised treebanks
to the official UD releases.

4 Treebank Investigation

An overview of the current state and our modifica-
tions of the treebanks is presented in the following
subsections. Further information can be retrieved
directly from the scripts available in GitHub.

4.1 Tokenisation

Although we focus on syntactic relations, some of
our interventions affect other annotation levels as
well. Some issues can be found already at the level
of tokenisation. For instance, a form like nobiscum
‘with us’, composed of the pronoun nobis ‘us’ and
the postponed, enclitic adposition cum ‘with’, is of-
ten not properly split in a multi-word token, but it is
considered as a unique token. However, this entails
losing the value of the preposition cum. Occur-
cences are found in ITTB, LLCT and Perseus. In
ITTB and LLCT, such instances (although rare) are
attached as obl; in Perseus, the advmod relation
is assigned. We thus split these tokens, by assign-
ing an obl relation to the pronoun, and annotating
cum as its case marker.

Negative conjunctions like neque and nec ‘and
not’ can be problematic, as happens in Perseus,
where they are currently split and inverted. See e.g.
et nemo poterat in caelo que ne in terra que ne sub-
tus terram aperire librum que ne respicere illum
‘And nobody in heaven, nor in earth, neither under
the earth, could open the book, neither to look at it’
(Bible, Rev. 5.3). This tokenisation does not corre-
spond to the original text (neque...neque...neque),
and is probably an erroneous result of the conver-
sion from the original data.

Moreover, across the treebanks (except for
LLCT and UDante) some instances are found
where the abbreviation dot is not separated from
the abbreviated form: e.g., C. Rufus in Perseus, Kal.
lan. in PROIEL. We thus split those occurrences
into two distinct tokens.7

4.2 POS tags

Some interventions concerning POS tags are
needed, especially as they often affect the choice
of the dependency relation. A critical point in all
the four treebanks (with the exception of UDante)
is represented by discourse adverbs like enim, ig-
itur, itaque (‘indeed, therefore’), that do not con-
stitute true adverbs but rather discourse elements
reinforcing the deployment of the sentence. Often
annotated as adverbs (ADV, advmod), they are
corrected in PARTs with discourse deprel. The
line between these two POS tags is often not clearly
drawn, and the case of o, used to address a recipient
in vocative case, proves it as well: mainly tagged
as ADV in ITTB, Perseus and PROIEL, it has been
reannotated as PART. No instances of o are found
in LLCT, due to the genre of the corpus.

A general harmonisation of determiners (DET, det)
is performed on all treebanks by defining a
lexical list of determiners, modeled on those oc-
curring in UDante. While being a shared issue,
this is particularly relevant for Perseus. Indeed, the
Perseus-employed tagset does not include some,
quite important, tags. It is the case of AUX, DET
and PART. PROPN, although officially used, is
often missing. The absence of the DET tag is ex-
tremely relevant, given its widespread distribution
over Latin texts. Through the lexical list, as well
as through morphological accordance with parent
node and after re-annotating the many determiners
originally attached as amod or nmod, we assign
the correct POS tag and relations.

The AUX for auxiliaries was not employed ei-
ther; it is now assigned to occurrences of sum ‘to be’
with deprel cop, aux or aux:pass. We also
retrieve proper nouns in a very trivial way, by locat-
ing capitalised nouns, since it is needed to correct

---

6https://github.com/fjambe/
Latin-variability (commit 303acc5).

7Tokenization of abbreviations is not unified UD-wide. In
some languages the guideline is to keep the abbreviation
with its punctuation as one token, while in others, including Latin,
the punctuation should be separated.
some dependencies. Indeed, proper nouns represent a very critical point in Perseus annotation, also due to the ample variety of different combinations of nouns and proper nouns.\(^8\) We restore correct dependencies and assign the appropriate dependency label: flat for a PROPN depending on a NOUN, flat:name to different components of a same proper noun.

In Perseus and PROIEL, we try to replace the X tag (unknown word) with the appropriate one. Some subordinating conjunctions, currently tagged as ADV, are corrected to SCONJ.

### 4.3 Syntax

As already mentioned, our main interventions concern dependency relations. In this regard, we replace expl:pass deprel—either with obj or obl according to the grammatical case of the word form—as it is not employed in UDante. Consider for instance aliter se habet intellectus divinus, atque aliter intellectus noster (lit. ‘otherwise itself has intellect divine, and otherwise intellect our’) ‘there is a difference between the divine intellect and ours’ (SCG 1, XXII, 5): expl:pass (habet, se) is reannotated as obj (habet, se).

Compound numerals like viginti quattuor ‘twenty-four’ display various annotations in the original treebanks, representing one of the most diverging phenomena. In LLCT, the numbers are connected as compound with the first number as head (compound(viginti, quattuor)). Other treebanks use different relations: in Perseus, the numbers are connected using nummod, and in PROIEL, fixed (the first number is the head in all cases). In accordance to UD guidelines, all these dependencies are reannotated as flat (i.e. flat(viginti, quattuor)).\(^9\)

Indirect objects (iobj) often occur in Perseus and PROIEL. They are replaced with obl:arg in the latter, and with obl, or obl:arg if in dative form, in the former. Indeed, despite the label being the same, its use in the two treebanks is not completely identical.

In ITTB some prepositions depending on the wrong head, namely on a token that precedes in

\(^8\)E.g., Tarquinio Prisco, Q Titrium Sabiniun legatum, L Valerio Flacco et C. Pomptino praetoribus, Aemilio Pipio imperatore.

\(^9\)Except for cases where a coordinating conjunction is present: viginti et quattuor is coordination, hence conj(viginti, quattuor); cc(quattuor, et).

the word order\(^10\) are reassigned to their correct head, which is identified based on dependency relations and POS tags. For instance, in Voluntas autem non ex necessitate furtur in ea quae sunt ad finem (lit. ‘will but not by necessity lead to those that are for a goal’) ‘the will is not necessarily directed to the means’ (Summa Contra Gentiles, 1, LXXXI, 2) the parent node of both ex and necessitate was non. We restore the correct dependencies, resulting in case(necessitate, ex) and obl(furtur, necessitate).

Interestingly, PROIEL contains the dep relation (intended for cases where a more precise dependency type cannot be determined). Through POS tags and morphology, we replace it with a more appropriate one,\(^11\) as illustrated in Figures 1 and 2.

Often problematic across treebanks, and in many different ways, the advmod deprel needs a closer inspection. In general, the dependency is improperly assigned to many non-adverbial instances. In ITTB, an interesting case is provided by bibliographic references, e.g. dicitur enim hebr. 3-1 ‘it is said in the letter to the Hebrews, 3-1’ ‘it is said in the letter to the Hebrews, 3-1’. The specification of the relevant Bible’s book sometimes depends as advmod on its parent node, i.e., the predicate dicitur in the proposed example; we convert it into obl, since it is a nominal form. In Perseus, we solve the issue of non-adverbal advmod through different criteria: lexical ones,

\(^10\)Postpositions are very rare in Latin.

\(^11\)For more detailed information, see the harmonisation script on Github.

\(^12\)The most accurate dependency label would be obl:lmod. However, as it is difficult to assign this subtype automatically, and subtypes are ignored in current parsing scores, we just assign obl.
e.g., to all tokens with lemma *autem* ‘but’ deeprel cc is assigned; morphological ones, e.g., if a substantive has Case=Loc, Dat or Voc, it is attached as obli:arg or vocative respectively; and POS criteria, e.g., if a token is tagged SCONJ, it receives the mark relation. The same issue is found also in PROIEL. For instance, *hic a mortuis resurrexit* ‘he is risen from the dead’ (Jerome’s Vulgate, Mark 6) is once annotated as follows: advmod(resurrexit, mortuis), case(mortuis, a). We thus try to restore similar occurrences of obliques, and other dependencies wrongly considered adverbial.

Another example of incorrect advmod relations is provided by calendar expressions, often found in PROIEL data. Consider, for instance, the sentence *pater nobis decessit a.d. iv Kal. Dec* ‘Our father died on November 28th’ (Cicero, Epistulae ad Atticum, 1, 6). Before the alignment, *a.d.* (ante diem ‘before the day’) and *iv* are not properly lemmatised, as their lemmas are respectively calendar and expression, they have no morphological features, and each token of the whole phrase, including *Kal.* and *Dec.*, is tagged as ADV. The relation between the date and its parent is advmod. The annotation is not even internally consistent: occurrences where tokens are not split, e.g. *Kal.Decembr* (lemmatised as calendar.expression and tagged ADV) can be found. The annotation of abbreviated dates in UD should reflect how the date would be pronounced (Zeman, 2021). However, cases like *Kal.Dec* are not straightforward, as they could be expanded in two possible ways—leading to two different analyses. The month can be either understood as an ADJ which takes a plural feminine form to agree with *kalendae/nonae/idus* (e.g., *Kalendae Decembres*), or a genitive singular (*Kalendae Decembris*). In cases where this is impossible to disambiguate, we take the first as the default reading. As far as possible, we try to align these occurrences and replace shallow labels like calendar.expression, as well as to assign correct dependencies (Figures 3 and 4).

In terms of coordination, the main intervention concerns reattaching conjunctions to the second conjunct instead of the first one; it is applied to Perseus and PROIEL. This is a significant change between UD v1 and v2 (Nivre et al., 2020), showing that the conversion of these treebanks to UD v2 was not perfect. Moreover, in Perseus parataxis is often found to be employed for coordination, and is corrected into conj.

A significant intervention in ITTB and LLCT applies to constructions involving the copula *sum* ‘to be’ and a prepositional phrase (often, but not exclusively, with locative meaning). In many such cases the copula occurs as the head, while the prepositional phrase depends on it as obli. Following the UD guidelines, we reverse the hierarchy by making the oblique the head and the copula its cop dependant. An example from ITTB: *successio autem propter motum aliquem est* (lit. ‘succession however because of movement some is’) ‘succession results from change of some kind’ (SCG 1, XCIX, 6): obli(est, motum) is reannotated as cop(motum, est), and all the dependents of the former head (est), e.g. the subject *successio*, are reattached to the new one (motum).

Comparative clauses are often problematic across the Latin treebanks, perhaps with the exception of ITTB, where our interventions are mostly limited to subtyping the advcl relation to :cmp.
for standards of comparison, as in *ut supra ostensum est* ‘as we have proved above’. In Perseus and PROIEL, and less in LLCT, various incorrect annotation patterns can be spotted. An example from PROIEL is provided in Figures 5 and 6. In PROIEL, relative clauses present some issues as well. See for instance *ea quae sunt his similia* ‘those things that are similar to these’ (Cicero, *De officis*, 1, 17): *similia* should depend on *ea* as *acl:relcl*, whereas it occurs as *appos*.

An unusual annotation pattern, observed in PROIEL with respect to adverbial clauses, is exemplified by the sentence hereafter: *postea quam agros et cultum et copias Gallorum homines feri ac barbari traductos plures* ‘after that these wild and savage men had become enamored of the lands and the refinement and the abundance of the Gauls, more were brought over’ (Cesar, *De bello Gallico*, 1.31). The parent node of *adamassent*, predicate of the adverbial clause, should be the *root traductos*, and its *deprel* *advcl*, while the subordinating conjunction *quam* ‘that’ should be its child node with *mark* dependency relation. However, in the original annotation we observe *fixed(quam, adamassent)* and *advcl(traductos, quam)*.

In some cases, dependency relations are lacking subtypes. Although the current parsing evaluation does not take them into account (see Section 5), we still believe that it is useful to unify them, also in view of more detailed work in the future. Therefore, for adverbs we identify a list of locative and temporal adverbs, and mark them with the *lmod* and *tmod* subtypes. This applies to all the five treebanks, UDante included. Indeed, in UDante locative and temporal adverbs (*advmod*) are already marked; yet, since in some cases the subtypes are missing, we assign them using the lexical list. Similarly, in the other four treebanks relative clauses and absolute ablatives respectively receive the subtypes (mostly ADV); negation counted separately.

<table>
<thead>
<tr>
<th>abbr</th>
<th>ITTB</th>
<th>LLCT</th>
<th>Perseus</th>
<th>PROIEL</th>
<th>UDante</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>advcl:abs</em></td>
<td>521</td>
<td>2019</td>
<td>163</td>
<td>1088</td>
<td>-</td>
<td>split dot and abbreviated word; in PROIEL, removed dot as punctuation is missing</td>
</tr>
<tr>
<td><em>advcl:cmp</em></td>
<td>2582</td>
<td>621</td>
<td>59</td>
<td>821</td>
<td>-</td>
<td>added subtype to absolute ablatives</td>
</tr>
<tr>
<td><em>advmod:indm</em></td>
<td>2505</td>
<td>1224</td>
<td>56</td>
<td>581</td>
<td>27</td>
<td>added subtype</td>
</tr>
<tr>
<td><em>advmod:deg</em></td>
<td>-</td>
<td>624</td>
<td>274</td>
<td>2691</td>
<td>-</td>
<td>added subtype to negation</td>
</tr>
<tr>
<td><em>advmod:tm</em></td>
<td>-</td>
<td>386</td>
<td>231</td>
<td>1099</td>
<td>77</td>
<td>added subtype</td>
</tr>
<tr>
<td><em>AUX</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>assigned AUX tag</td>
</tr>
<tr>
<td><em>aux-pass-periph</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>578</td>
<td>-</td>
<td>intervention on date/calendar expression; can refer to both label and dependency</td>
</tr>
<tr>
<td><em>dep</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>47</td>
<td>replaced dep with more appropriate label</td>
</tr>
<tr>
<td><em>DET</em></td>
<td>1206</td>
<td>53</td>
<td>2557</td>
<td>14225</td>
<td>-</td>
<td>assigned DET tag; most often, det entailed</td>
</tr>
<tr>
<td><em>exp:pass</em></td>
<td>335</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>replaced with <em>obj</em></td>
</tr>
<tr>
<td><em>flat-for-names</em></td>
<td>-</td>
<td>-</td>
<td>82</td>
<td>202</td>
<td>-</td>
<td>assigned flat if appropriate</td>
</tr>
<tr>
<td><em>incorrect-advmod</em></td>
<td>115</td>
<td>48</td>
<td>2086</td>
<td>1030</td>
<td>-</td>
<td>corrected advmod if assigned to non-adverbials</td>
</tr>
<tr>
<td><em>inversion-sum</em></td>
<td>2843</td>
<td>162</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>inverted head-dependent in copular constructions (both dependencies and labels)</td>
</tr>
<tr>
<td><em>inverted-prep</em></td>
<td>248</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>reattached prepositions depending on preceding node</td>
</tr>
<tr>
<td><em>iobj</em></td>
<td>-</td>
<td>-</td>
<td>491</td>
<td>5870</td>
<td>-</td>
<td>replace iobj with <em>obj</em> arg; obj used inappropriately (in Perseus) included</td>
</tr>
<tr>
<td><em>j-i</em></td>
<td>-</td>
<td>-</td>
<td>345</td>
<td>-</td>
<td>-</td>
<td>substituted j with i to normalise lemmas</td>
</tr>
<tr>
<td><em>mwt</em></td>
<td>44</td>
<td>28</td>
<td>20</td>
<td>60</td>
<td>-</td>
<td>split a token into multi-word token</td>
</tr>
<tr>
<td><em>nc</em></td>
<td>-</td>
<td>-</td>
<td>55</td>
<td>-</td>
<td>-</td>
<td>corrected <em>c ne → ne c</em></td>
</tr>
<tr>
<td><em>nsubj:pass</em></td>
<td>2</td>
<td>-</td>
<td>428</td>
<td>338</td>
<td>27</td>
<td>added subtype to subjects of passive verbs</td>
</tr>
<tr>
<td><em>num</em></td>
<td>60</td>
<td>61</td>
<td>29</td>
<td>40</td>
<td>-</td>
<td>corrected numerals; mostly label, sometimes also dependency</td>
</tr>
<tr>
<td><em>parataxis-to-conj</em></td>
<td>-</td>
<td>-</td>
<td>159</td>
<td>-</td>
<td>-</td>
<td>parataxis used for coordination is replaced with <em>conj</em></td>
</tr>
<tr>
<td><em>PART</em></td>
<td>7198</td>
<td>203</td>
<td>179</td>
<td>2254</td>
<td>10</td>
<td>assigned PART tag instead of incorrect ones (mostly ADV); negation counted separately</td>
</tr>
</tbody>
</table>

Table 2: Count of harmonising interventions.
and abs, if missing. In Perseus and PROIEL, the same applies to negations, which are assigned the advmod:neg dependency relation.

4.4 Summary

The investigation reveals recurring issues which are spread across all treebanks (Table 2), although differing in various ways. The most widespread issues are the tmod and lmod relation subtypes, as well as comparative clauses.

However, more interventions are needed in Perseus and PROIEL than in the other three treebanks. Indeed, the degree of accordance with the UD guidelines is definitely lower in the Perseus treebank—perhaps unsurprisingly, as it has not been updated since its initial conversion to UD v2 in 2017. PROIEL’s condition resembles that of Perseus, including the status of neglected in the latest release (May 2022).

Only minor modifications are needed in UDante, which comes as no surprise, as this treebank was selected as the reference point for the whole harmonisation process. Overall, the main divergence between UDante and the other treebanks lies in relation subtypes. Indeed, UDante employs a range of subtypes that is not shared by the other treebanks, and that would be problematic if the parsing evaluation process included subtypes; since it is currently not the case (Section 5), we choose not to focus on this specific issue.

5 Impact on Parsing

Afterwards, we try to assess the impact that a harmonised annotation of the five treebanks has on parsing accuracy. In order to achieve this, with both UDPipe and Stanza we retrain a model for every aligned treebank. We then test the obtained models on each of the treebanks; Tables 4 and 6 summarise the scores, in terms of Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) (Buchholz and Marsi, 2006), obtained with models trained with UDPipe and Stanza respectively. To measure accuracy, we employ the Python evaluation script designed for the CoNLL 2018 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., 2018). As mentioned earlier, the script takes into account only main dependency types, without considering subtypes. This reflects our current needs; nevertheless, the present treebank alignment is only the first stage of a larger harmonisation effort, and additional evaluation criteria (including relation subtypes) can be introduced in the future.

To demonstrate the effect of harmonisation, we also present LAS and UAS scores of models trained on pre-harmonisation data (Tables 3 and 5), again with UDPipe and Stanza. Such models are trained and tested on master data of Universal Dependencies 2.10, officially released in May 2022.

Both series of models, pre- and post-alignment, are trained with the same settings. With respect to UDPipe, version 1.2 is used; we employ pretrained fastText embeddings (Grave et al., 2018) and optimised training hyperparameters as described for reproducible training by Straka and Straková (2019), within the publication of UD 2.5 models for UDPipe. Since optimised hyperparameters are available only for ITTB, Perseus and PROIEL, for LLCT and UDante we experiment with different options and select the best ones. As for pre-alignment models for Stanza, we employ the ITTB, Perseus and PROIEL models made available by the Stanza team and pretrained on UD 2.8, since those treebanks did not change afterwards, as reported in their change log. We train pre-alignment models for LLCT and UDante, as well as all post-alignment ones, with default parameters and fastText embeddings.

Some conclusions can be drawn from the comparison of the tables. With UDPipe, the interventions prove effective in most cases, as models trained on harmonised treebanks reach higher scores than the pre-alignment ones. This holds true especially with respect to results on Perseus and PROIEL; indeed, each of the post-alignment models gains higher scores on these two treebanks. The improvement is substantial (up to +9% with more than one model), and confirms once more the absolute relevance of a truly universal annotation style. Higher impact on Perseus and PROIEL is expected, given their previous condition (Section 4).

Analogously, the models trained on harmonised

---

14 Additional information can be found at the website of the project: [https://github.com/UniversalDependencies/tools/blob/master/eval.py](https://github.com/UniversalDependencies/tools/blob/master/eval.py).

15 It would also be problematic if a parser were trained jointly on concatenated Latin treebanks.


17 LLCT: learning_rate=0.02, transition_system=swap, transition_oracle=static_lazy, structured_interval=8.

18 UDante: learning_rate=0.01, transition_system=projective, transition_oracle=dynamic, structured_interval=8.
Perseus and PROIEL achieve better scores on every of the five treebanks. Peaks are represented by LLCT parsed with a Perseus model (around +17% both in LAS and UAS). As for PROIEL, the increases are slightly lower, yet still substantial. Consider, for instance, the performance of a PROIEL post-alignment model on Perseus test data: an improvement of +11 percentage points is assessed with respect to LAS.

The model trained on aligned UDante proves to gain higher scores on almost every treebank, with more substantial increases on Perseus and PROIEL. This is mostly due to the alignment interventions on the other treebanks than on UDante itself, as the harmonisation process was minimal on UDante data. The increase observed when a UDante model is employed to parse UDante test data could be probably caused by divergences between release 2.10 of UDante, which the model in Table 3 was trained on, and UDante dev data, used as the basis for the alignment.

ITTB and LLCT models show a less consistent behaviour, performing sometimes better (i.e. on Perseus and PROIEL), sometimes marginally worse (e.g. ITTB model on ITTB and LLCT test data). A closer analysis of the parser outputs, despite not providing a precise explanation for the parser behaviour, reveals that the harmonisation can be further enhanced. For instance, it emerges that the harmonisation of copular constructions, as discussed in Subsection 4.3, did not catch all occurrences and the wrong original annotation survives in some sentences. Such coexistence of pre- and post-harmonisation annotations, and thus a lower degree of consistency, may partially explain the observed decrease in parsing accuracy.

The general trend of improved scores can be observed also when models are trained with Stanza. Yet, the increase is not as considerable as when UDPipe is employed.

However, Tables 3, 5, 4 and 6 also highlight how the treebank annotation alignment does not solve the issue discussed in Section 2: the drop is still significant when data are parsed with models trained on a different treebank. Moreover, the absolute scores presented depend also on the size of training data, which varies substantially across the treebanks (see Table 1), Perseus being particularly small.

6 Conclusion and Future Work

The annotation alignment proposed in the present paper confirms the relevance of a shared and universal annotation scheme. Thus, although the Universal Dependencies project already represents an outstanding milestone, the effort needed in this direction is still remarkable, and two-fold: on the one hand, treebanks should be constantly updated to the latest UD guidelines, as they keep developing to-wards a more consistent annotation formalism. On the other hand, different research teams working on the same language should collaboratively define shared guidelines and adopt the same approach in annotation, so that Universal Dependencies can grow more and more universal.

Many future directions can be envisaged for this study. The alignment needs to be further inves-

<table>
<thead>
<tr>
<th>ittb.udp</th>
<th>llct.udp</th>
<th>perseus.udp</th>
<th>proiel.udp</th>
<th>udante.udp</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAS</td>
<td>UAS</td>
<td>LAS</td>
<td>UAS</td>
<td>LAS</td>
</tr>
<tr>
<td>ITTB</td>
<td>84.51%</td>
<td>86.23%</td>
<td>44.25%</td>
<td>52.16%</td>
</tr>
<tr>
<td>LLCT</td>
<td>44.22%</td>
<td>50.16%</td>
<td>93.02%</td>
<td>93.85%</td>
</tr>
<tr>
<td>Perseus</td>
<td>33.28%</td>
<td>44.21%</td>
<td>39.85%</td>
<td>48.71%</td>
</tr>
<tr>
<td>PROIEL</td>
<td>39.10%</td>
<td>50.86%</td>
<td>43.16%</td>
<td>53.08%</td>
</tr>
<tr>
<td>UDante</td>
<td>50.78%</td>
<td>58.51%</td>
<td>36.95%</td>
<td>45.78%</td>
</tr>
</tbody>
</table>

Table 3: UDPipe scores before treebank alignment. Columns correspond to trained models, rows to test data.
This work was supported by the Grant No. 20-16819X (LUSyD) of the Czech Science Foundation (GAČR) and by the GAUK project “Syntactic parsing of Latin texts – dealing with linguistic diversity”.    

<table>
<thead>
<tr>
<th></th>
<th>ittb.mdl</th>
<th>llct.mdl</th>
<th>perseus.mdl</th>
<th>proiel.mdl</th>
<th>udante.mdl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LAS</td>
<td>UAS</td>
<td>LAS</td>
<td>UAS</td>
<td>LAS</td>
</tr>
<tr>
<td>ITTB</td>
<td>89.16%</td>
<td>91.26%</td>
<td>47.27%</td>
<td>60.00%</td>
<td>45.99%</td>
</tr>
<tr>
<td>LLCT</td>
<td>47.57%</td>
<td>58.79%</td>
<td>94.56%</td>
<td>95.78%</td>
<td>29.38%</td>
</tr>
<tr>
<td>Perseus</td>
<td>51.31%</td>
<td>65.56%</td>
<td>34.33%</td>
<td>49.73%</td>
<td>61.65%</td>
</tr>
<tr>
<td>PROIEL</td>
<td>54.53%</td>
<td>68.10%</td>
<td>40.70%</td>
<td>56.06%</td>
<td>48.25%</td>
</tr>
<tr>
<td>UDante</td>
<td>57.07%</td>
<td>68.44%</td>
<td>39.16%</td>
<td>52.88%</td>
<td>32.09%</td>
</tr>
</tbody>
</table>

Table 5: Stanza scores before treebank alignment. Columns correspond to trained models, rows to test data.

<table>
<thead>
<tr>
<th></th>
<th>ittb.mdl</th>
<th>llct.mdl</th>
<th>perseus.mdl</th>
<th>proiel.mdl</th>
<th>udante.mdl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LAS</td>
<td>UAS</td>
<td>LAS</td>
<td>UAS</td>
<td>LAS</td>
</tr>
<tr>
<td>ITTB</td>
<td>88.60%</td>
<td>90.55%</td>
<td>45.63%</td>
<td>58.74%</td>
<td>50.55%</td>
</tr>
<tr>
<td>LLCT</td>
<td>40.84%</td>
<td>52.66%</td>
<td>94.61%</td>
<td>95.81%</td>
<td>37.82%</td>
</tr>
<tr>
<td>Perseus</td>
<td>57.68%</td>
<td>67.85%</td>
<td>40.80%</td>
<td>53.88%</td>
<td>58.41%</td>
</tr>
<tr>
<td>PROIEL</td>
<td>62.34%</td>
<td>71.27%</td>
<td>46.76%</td>
<td>59.92%</td>
<td>55.03%</td>
</tr>
<tr>
<td>UDante</td>
<td>56.62%</td>
<td>67.27%</td>
<td>39.67%</td>
<td>52.97%</td>
<td>39.53%</td>
</tr>
</tbody>
</table>

Table 6: Stanza scores after treebank alignment. Columns correspond to trained models, rows to test data.


Acknowledgements

This work was supported by the Grant No. 20-16819X (LUSyD) of the Czech Science Foundation (GAČR) and by the GAUK project “Syntactic parsing of Latin texts – dealing with linguistic diversity”.

References

Conference, Marseille, France. European Language Resources Association.


