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) (Cecchini 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.
- **PROIEL** (Haug and Jøhndal, 2008): it contains most of the Vulgate New Testament translations, and selections from Caesar's *De bello Gallico*, Cicero's *Epistulae ad Atticum*, Palladius' *Opus Agriculturae* and the first book of Cicero's *De officiis* (examples of Classical Latin, yet representing different genres).
- **UDante** (Cecchini et al., 2020a): it includes literary texts (letters, treatises, poetry) by Dante Alighieri, corresponding to literary Medieval Latin (XIV century).

		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	15,917	1,234	1,260
	words	172,133	13,939	14,091
UDante	sents	926	376	419
	words	30,441	11,611	13,451

Table 1: Size of UD Latin treebanks in v2.10.

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

[1] See https://universaldependencies.org/.

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.[2]

[2]So far EvaLatin has been focusing on lemmatisation, morphological analysis and POS tagging; in the future, EvaLatin campaigns will probably extend the cross-time and cross-genre sub-tasks to syntactic diversity (Sprugnoli et al., 2022).

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 project[3] (Passarotti et al., 2020). 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[4] 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

[3]See https://lila-erc.eu/.

[4]As 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.

[5]See 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 exclusively: 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 providing 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 modifications 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 often 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*. Occurrences 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 assigning 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*** *subtus 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 correspond to the original text (*neque...neque...neque*), and is probably an erroneous result of the conversion 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. Ian.* 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*, *igitur*, *itaque* ('indeed, therefore'), that do not constitute 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 occurring 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 extremely 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 either; 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 locating capitalised nouns, since it is needed to correct

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

[7] Tokenization 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 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 fertur 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(fertur, necessitate)`.

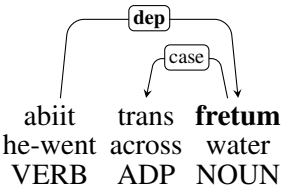

Figure 1: Example of a `dep` dependency (en. 'he departed to the other side').

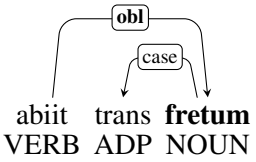

Figure 2: Result of the harmonisation process.[12]

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 biblical references, e.g. *dicitur enim hebr. 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-adverbial `advmod` through different criteria: lexical ones,

[8] E.g., *Tarquinio Prisco*, *Q Titurium Sabinum legatum*, *L. Valerio Flacco et C. Pomptino praetoribus*, *Aemilio Papo imperatore*.

[9] Except for cases where a coordinating conjunction is present: *viginti et quattuor* is coordination, hence `conj(viginti, quattuor); cc(quattuor, et)`.

[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`.

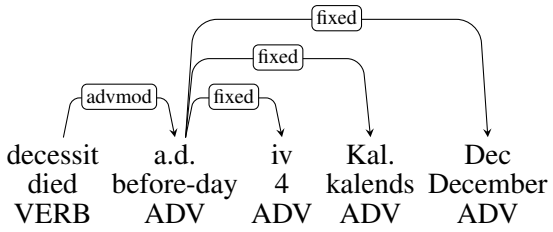

Figure 3: Annotation in UD 2.10.

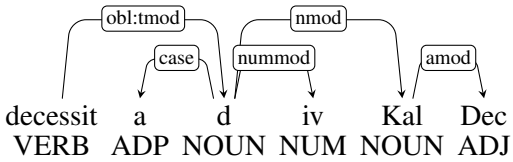

Figure 4: Result of the harmonisation process.

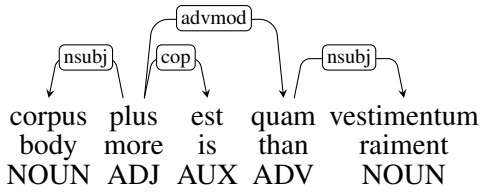

Figure 5: Annotation in UD 2.10.

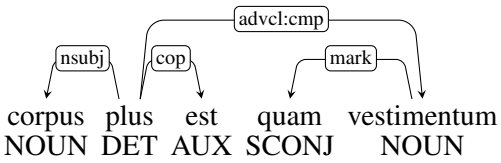

Figure 6: Result of the harmonisation process.

e.g., to all tokens with lemma *autem* 'but' deprel `cc` is assigned; morphological ones, e.g., if a substantive has `Case=Loc`, `Dat` or `Voc`, it is attached as `obl`, `obl: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 `obl`iques, 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 Decembr**es***), or a genitive singular (*Kalendae Decembr**is***). 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 `obl`. 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): `obl(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`

	ITTB	LLCT	Perseus	PROIEL	UDante	notes
abbr	1302	-	24	107	-	split dot and abbreviated word; in PROIEL, removed dot as punctuation is missing
advcl:abs	521	2019	163	1088	-	added subtype to absolute ablatives
advcl:cmp	2582	621	59	821	-	corrected deprel for comparative clauses (often, dependencies as well)
advmod:lmod	2505	1224	56	581	27	added subtype
advmod:neg	-	624	274	2691	-	added subtype to negation
advmod:tmod	-	386	231	1099	77	added subtype
AUX	-	-	366	-	-	assigned AUX tag
aux-pass-periph	-	-	14	283	-	added subtype to periphrastic passive
dates	-	-	-	578	-	intervention on date/calendar expression; can refer to both label and dependency
dep	-	-	-	47	-	replaced `dep` with more appropriate label
DET	1206	53	2557	14225	-	assigned DET tag; most often, `det` entailed
expl:pass	335	-	-	-	-	replaced with `obj`/`obl`
flat-for-names	-	-	82	202	-	assigned `flat` (`flat:name` if appropriate) to PROPNs
incorrect-advmod	115	48	2086	1030	-	corrected `advmod` if assigned to non-adverbials
inversion-sum	2843	162	-	-	-	inverted head-dependent in copular constructions (both dependencies and labels)
inverted-prep	248	-	-	-	-	reattached prepositions depending on preceding node
iobj	-	-	491	5870	-	replace `iobj` with `obj`/`obl:arg`; `obj` used inappropriately (in Perseus) included
j-i	-	-	345	-	-	substituted *j* with *i* to normalise lemmas
mwt	44	28	20	60	-	split a token into multi-word token
nec	-	-	55	-	-	corrected *c ne* →*ne c*
nsubj:pass	2	-	428	338	27	added subtype to subjects of passive verbs
num	60	61	29	40	-	corrected numerals; mostly label, sometimes also dependency
parataxis-to-conj	-	-	159	-	-	`parataxis` used for coordination is replaced with `conj`
PART	7198	203	179	2254	10	assigned PART tag instead of incorrect ones (mostly ADV); negation counted separately

Table 2: Count of harmonising interventions.

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* ***adamassent*** *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' (Caesar, *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[13] 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 sub-

[13]The list is not intended to be exhaustive in the present stage of the research.

types `relcl` 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;[14] 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[15] 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[16] (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.[17] As for pre-alignment models for Stanza, we employ the ITTB, Perseus and PROIEL models made available[18] 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] It would also be problematic if a parser were trained jointly on concatenated Latin treebanks.

[15] Available at https://github.com/UniversalDependencies/tools/blob/master/eval.py.

[16] Available at https://fasttext.cc/docs/en/crawl-vectors.html.

[17] LLCT: learning_rate=0.02, transition_system=swap, transition_oracle=static_lazy, structured_interval=8.
UDante: learning_rate=0.01, transition_system=projective, transition_oracle=dynamic, structured_interval=8.

[18] At https://stanfordnlp.github.io/stanza/available_models.html.

	ittb.udp		llct.udp		perseus.udp		proiel.udp		udante.udp	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
ITTB	**84.51**%	**86.23**%	44.25%	52.16%	29.54%	40.56%	30.54%	45.43%	59.93%	65.77%
LLCT	44.22%	50.16%	**93.02**%	**93.85**%	28.92%	37.44%	40.37%	52.10%	45.57%	53.42%
Perseus	33.28%	44.21%	39.85%	48.71%	**61.80**%	**67.18**%	38.93%	55.16%	35.64%	45.79%
PROIEL	39.10%	50.86%	43.16%	53.08%	41.52%	52.36%	**73.51**%	**77.45**%	39.43%	48.62%
UDante	50.78%	58.51%	36.95%	45.78%	22.44%	32.41%	26.72%	40.41%	**50.81**%	**57.32**%

Table 3: UDPipe scores before treebank alignment. Columns correspond to trained models, rows to test data.

	ittb.udp		llct.udp		perseus.udp		proiel.udp		udante.udp	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
ITTB	**83.83**%	**85.51**%	43.80%	51.45%	43.17%	53.12%	40.46%	51.33%	61.68%	67.39%
LLCT	43.12%	48.55%	**93.11**%	**93.88**%	47.31%	54.13%	46.69%	55.23%	41.56%	49.05%
Perseus	42.73%	53.54%	48.69%	55.24%	**63.80**%	**68.38**%	49.98%	59.25%	43.59%	54.23%
PROIEL	46.77%	55.39%	50.37%	57.48%	53.11%	59.88%	**75.78**%	**78.87**%	46.13%	55.15%
UDante	53.06%	59.95%	38.51%	46.69%	35.59%	45.64%	30.72%	44.11%	**54.50**%	**61.02**%

Table 4: UDPipe scores after treebank alignment. Columns correspond to trained models, rows to test data.

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,[19] 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,[20] 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 towards 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-

[19] LLCT represents an exception.

[20] See the example from ITTB: *Successio autem propter motum aliquem est.*

	ittb.mdl		llct.mdl		perseus.mdl		proiel.mdl		udante.mdl	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
ITTB	**89.16**%	**91.26**%	47.27%	60.00%	45.99%	59.32%	44.49%	60.37%	60.80%	70.37%
LLCT	47.57%	58.79%	**94.56**%	**95.78**%	29.38%	46.17%	38.34%	51.77%	41.96%	53.54%
Perseus	51.31%	65.56%	34.33%	49.73%	**61.65**%	**71.35**%	45.19%	61.89%	44.26%	59.71%
PROIEL	54.53%	68.10%	40.70%	56.06%	48.25%	65.42%	**79.80**%	**84.17**%	44.83%	57.75%
UDante	57.07%	68.44%	39.16%	52.88%	32.09%	48.42%	37.21%	50.32%	**56.84**%	**66.12**%

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

	ittb.mdl		llct.mdl		perseus.mdl		proiel.mdl		udante.mdl	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
ITTB	**88.60**%	**90.55**%	45.63%	58.74%	50.55%	61.47%	51.16%	60.72%	63.78%	72.96%
LLCT	40.84%	52.66%	**94.61**%	**95.81**%	37.82%	47.50%	40.97%	53.24%	43.64%	56.09%
Perseus	57.68%	67.85%	40.80%	53.88%	**58.41**%	**68.22**%	47.30%	58.68%	52.98%	64.06%
PROIEL	62.34%	71.27%	46.76%	59.92%	55.03%	65.25%	**80.57**%	**84.36**%	52.61%	63.91%
UDante	56.62%	67.27%	39.67%	52.97%	39.53%	52.98%	41.27%	52.41%	**57.92**%	**67.60**%

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

tigated, not only at the level of tokenisation and dependency relations, but also with respect to lemmatisation, POS tagging and morphological features. In the near future, we plan to test some error detection methods in order to locate annotation inconsistencies within and among the five treebanks and intervene on them. See Section 2 for some preliminary references.

Moreover, we intend to carry out an error analysis of automatically parsed treebanks, so as to identify some error trends, and possibly compare parsing errors before and after treebank alignment.

Once the treebanks follow a more uniform annotation style, it will be possible and appropriate to investigate the actual linguistic differences causing performance drops when models trained on one treebank are applied to another. Possible directions for this future work include an analysis of genre diversity, a closer examination of different types of employed embeddings, and exploitation of Latin BERT (Bamman and Burns, 2020). The results could lead to the definition of strategies to overcome the issue of Latin syntactic variability.

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

Akshay Aggarwal and Chiara Alzetta. 2021. Atypical or underrepresented? A pilot study on small treebanks. In *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Akshay Aggarwal and Daniel Zeman. 2020. Estimating POS annotation consistency of different treebanks in a language. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 93–110, Düsseldorf, Germany. Association for Computational Linguistics.

Bharat Ram Ambati, Rahul Agarwal, Mridul Gupta, Samar Husain, and Dipti Misra Sharma. 2011. Error detection for treebank validation. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 23–30, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin Dependency Treebanks. In *Language technology for cultural heritage*, pages 79–98. Springer.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Flavio M Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020a. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In *Seventh Italian Conference on Computational Linguistics*, pages 1–7. CEUR Workshop Proceedings.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020b. A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages. In *Proceedings of the 12th Language Resources and Evaluation*

Conference, Marseille, France. European Language Resources Association.

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the Universal Dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 108–115, Pisa,Italy. Linköping University Electronic Press.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Markus Dickinson and W Detmar Meurers. 2003. Detecting inconsistencies in treebanks. In *Proceedings of TLT*, volume 3, pages 45–56.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dag TT 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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10:299–320.

Marco Passarotti and Felice Dell'Orletta. 2010. Improvements in parsing the index Thomisticus treebank. revision, combination and a feature model for medieval Latin. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa knowledge base of linguistic resources for Latin. *Linguistic Studies and Essays*, 58(1):177–212.

Marco Passarotti and Paolo Ruffolo. 2010. Parsing the Index Thomisticus Treebank. Some Preliminary Results. In *15th International Colloquium on Latin Linguistics*, pages 714—725. Innsbrucker Beiträge zur Sprachwissenschaft.

Edoardo Maria Ponti and Marco Passarotti. 2016. Differentia compositionem facit. a slower-paced and reliable parser for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 683–688, Portorož, Slovenia. European Language Resources Association (ELRA).

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. 2019. Universal dependencies 2.5 models for UDPipe (2019-12-06). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Alexander Volokh and Günter Neumann. 2011. Automatic detection and correction of errors in dependency treebanks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 346–350, Portland, Oregon, USA. Association for Computational Linguistics.

Daniel Zeman. 2021. Date and time in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 173–193, Sofia, Bulgaria. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.