Latin Morphology through the Centuries: Ensuring Consistency for Better Language Processing

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 focuses on the process of harmonising the five Latin treebanks available in Universal Dependencies with respect to morphological annotation. We propose a workflow that allows to first spot inconsistencies and missing information, in order to detect to what extent the annotations differ, and then correct the retrieved bugs, with the goal of equalising the annotation of morphological features in the treebanks and producing more consistent linguistic data. Subsequently, we present some experiments carried out with UDPipe and Stanza in order to assess the impact of such harmonisation on parsing accuracy.

1 Introduction

In Universal Dependencies (de Marneffe et al., 2021) five treebanks are available for Latin:[1] Index Thomisticus Treebank (ITTB; Passarotti, 2019), Late Latin Charter Treebank (LLCT; Cecchini et al., 2020b), Perseus (Bamman and Crane, 2011), PROIEL (Haug and Jøhndal, 2008), UDante (Cecchini et al., 2020a). These treebanks differ on multiple levels. First, they cover different domains: a shallow distinction can be made between poetry (found in Perseus and, less, in UDante) and prose (all treebanks), but it can be further specified in terms of specific genre included. For instance, ITTB encompasses philosophical texts, while LLCT consists of charters, representing an instance of documentary genre. Additionally, the history of Latin, spoken for over two millennia, entails a substantial diachronic variation, as the language gradually evolved over time. Indeed, the five Latin treebanks include data that differ substantially in this respect. Already considering the Medieval treebanks alone, we can observe how wide the covered time range is: ITTB encompasses Medieval texts dating back to XIII century, LLCT features Early Medieval charters (VIII-IX century), while Dante Alighieri's work available in UDante belongs to XIV century. In addition to that, Perseus and PROIEL include classical texts (I BC - IV AD), as well as the *Vulgate* (IV century). A level of spatial variability can be observed too; for instance, LLCT includes texts written in Tuscany, Italy, and some features typical of the Romance languages are already emerging.

In addition to the aforementioned levels of variability, and besides variation in size, Latin treebanks also differ in terms of annotation choices, in spite of the UD work towards consistency. This issue can be doubly problematic: first with respect to UD itself, as the annotation is expected to be consistent across and within languages; secondly, in light of the fact that the quality of data may affect the results of any experiment or linguistic investigation carried out on those data. Gamba and Zeman (2023), investigating parsing performances, already observe this as regards the syntactic layer of these data. Nevertheless, what has been observed with respect to syntax does not necessarily apply to morphological features as well, and the extent to which inconsistent morphological annotation affects parsing performances thus remains unclear.

For this purpose, we first propose a harmonisation of the morphological features of the five treebanks, and thereafter assess its impact on models predicting morphology, as well as syntactic parsers. Section 2 presents some related work and the motivation behind our study. Section 3 features an overview of the harmonisation process, while in Section 4 we describe the strategy designed to spot inconsistent or missing annotations. Section 5 highlights the main harmonising interventions, whose impact on parsing accuracy is assessed in Section 6. Finally, Section 7 concludes the paper and suggests future research directions.

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

2 Related Work and Motivation

Any NLP task is likely to show degraded performance when a model is applied to data that differ from training data. It has been observed several times that this issue is particularly prominent in (morpho-)syntactic parsing of Latin texts. The issue is strongly intertwined with Latin intralinguistic variability, as the language has undergone a number of significant changes by spreading over a period of more than two millennia and across Europe. In order to investigate genuine linguistic diversity, first and foremost the impact of divergent annotation styles has to be ruled out. To perform any experiment that exploits data, we need those data to be consistent. Harmonising such discrepancies would allow for the isolation of the impact that annotation choices have, so that actual intralinguistic variability can emerge and be examined.

The issue of Latin variability has been addressed in the two EvaLatin campaigns (Sprugnoli et al., 2020; Sprugnoli et al., 2022), aiming to evaluate NLP tools for Latin. In particular, EvaLatin has been focusing on lemmatisation, morphological analysis and POS tagging. However, Latin diversity has been observed several times already before, in light of the behaviour of parsing accuracy, which was far from being homogeneous. See, for instance, Passarotti and Ruffolo (2010), Ponti and Passarotti (2016), Passarotti and Dell'Orletta (2010). Several studies have also been addressing the issue of inconsistent annotations. 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) are only some of the methods that have been proposed to detect inconsistencies in treebanks. Gamba and Zeman (2023) present a harmonisation of dependency relations in Latin treebanks, yet without intervening at the level of morphological features. Their harmonisation highlighted several levels of inconsistencies and proved to lead to substantial improvements in terms of parsing accuracy. We investigate whether similar improvements can be achieved by also addressing inconsistencies in morphological annotation.

The output of the present study is two-fold:

- Producing a new version of the treebanks, harmonised at the level of morphological features, to be potentially contributed to the UD official release or to serve as an inspiration for other treebank maintainers to refine morphological annotation. Towards the latter goal, we develop a UDapi (Popel et al., 2017) block for detecting required and allowed morphological features in Latin treebanks. The Latin block was inspired by a similar block for Czech and we will contribute it to the official UDapi repository; it can be adapted to any other language by modifying the template according to language-specific features.

- Investigating the impact of harmonised morphological features in parsing, by assessing if and to what extent they affect accuracy scores. A comparison of two parsers, UDPipe (Straka et al., 2016) and Stanza (Qi et al., 2020) is proposed.

3 Overview of the Harmonisation Process

The focus of the harmonisation process is exclusively on morphological features.

We define the workflow to detect inconsistencies and missing features as follows. First, we run the UDapi block on the input data, with the goal of spotting features which are either required but missing, or not allowed. As output, the trees that feature either of these two kinds of inconsistencies are stored in a `html` file, where those bugs are prominently highlighted (see Figure 1). In light of the output `html` file, we build Python scripts that address and fix the observed bugs.

We employ the harmonised version of the five treebanks, as made available by Gamba and Zeman (2023), as input. Nevertheless, differently from what was done for syntactic harmonisation, we do not strictly follow UDante annotation. This choice is justified by the fact that we observe a considerable difference in the set of morphological features employed in UDante – predominantly – and the other treebanks (ITTB and LLCT) maintained by the same developers, i.e. the team at Università Cattolica del Sacro Cuore in Milan, Italy, as opposed to the two remaining treebanks (Perseus and PROIEL) out of the five available for Latin. We thus decide to define two levels of coherence:

- lower level (default): only information which can be considered somehow core, or more essential, is required. For instance, all pronouns must have a `PronType`, and all verbs must have `VerbForm` and `Aspect`.

- higher level: additional information, such as InflClass, is expected and allowed. This level of validation can be applied only to a subset of the Latin treebanks.

By default, the block operates at the lower level, but a parameter can be supplied to UDapi, which will trigger the more detailed features.

Morphologically harmonised treebanks and harmonisation scripts are available on GitHub,[2] while the block is available in UDapi GitHub repository. Moreover, we are ready to contribute the harmonised treebanks to the official UD release.

4 The markFeatsBugs Block

The markFeatsBugs block is structured as follows. For each UPOS tag, a set of *required* features is first defined. (Note that the official UD validator[3] has some limited ability to check *permitted* UPOS-feature combinations, but not to enforce required features.) Additional features that are permitted but not required are listed, and for each permitted feature the set of its permitted values is defined. Unlike in the official UD validator, the conditions for a feature-value to be permitted or required are not limited to whole UPOS categories. For example, the UD validator knows that the Person feature is allowed for verbs and auxiliaries; but we further restrict it to finite forms, i.e., the feature VerbForm must be present and its value must be Fin.

The set of allowed features is then expanded to include additional feature-value pairs that may be found in UDante, ITTB or LLCT (higher level of detail). Eventually, the block checks for each node whether its morphological features are permitted and if every node has all the required features. If not, invalid and missing features are explicitly marked with a transparent label allowing to easily distinguish them, and saved in the Bug attribute in the MISC column of the CoNLL-U file. It can be later used in filters and highlighted in the data. The code snippet in Script 1 provides an example, although not exhaustive, of the block section concerning verbs and auxiliaries, in compliance to what has been implemented in the treebanks among all the proposals illustrated in Cecchini (2021). Script 2 illustrates the expansion of the feature-value sets to the higher level, applicable to only three treebanks. UDante is used as reference to select those features.

```
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']
  elif node.feats['VerbForm'] ==
  ↪  'Part':
    rf.extend(['Gender', 'Number',
    ↪  'Case'])
    af['Number'] = ['Sing', 'Plur']
    ↪  if
    ↪  node.misc['TraditionalMood']
    ↪  != 'Gerundium' else ['Sing']
    af['Gender'] = ['Masc', 'Fem',
    ↪  'Neut'] if
    ↪  node.misc['TraditionalMood']
    ↪  != 'Gerundium' else ['Neut']
    af['Case'] = ['Nom', 'Gen',
    ↪  'Dat', 'Acc', 'Voc', 'Loc',
    ↪  'Abl']
    af['Degree'] = ['Abs', 'Cmp']
    if node.misc['TraditionalMood'].
    ↪  startswith('Gerundi'):
      af['Voice'] = ['Pass']
      af['Aspect'] = 'Prosp'
  elif node.feats['VerbForm'] ==
  ↪  'Conv':
    rf.extend(['Case', 'Gender',
    ↪  'Number'])
    af['Case'] = ['Abl', 'Acc']
    af['Gender'] = ['Masc']
    af['Number'] = ['Sing']
    af['Voice'] = ['Act']
  elif node.feats['VerbForm'] ==
  ↪  'Inf':
    af['Tense'].remove('Pqp')
```

Script 1: Portion of the block that partially exemplifies how morphological features are checked for the verbal system: rf stands for 'required features', af stands for 'allowed features'.

[2] https://github.com/fjambe/Latin-variability/tree/main/morpho_harmonization (commit 2d14807).

[3] https://github.com/UniversalDependencies/tools/blob/master/validate.py.

```
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']
```

Script 2: Richer, more detailed morphological features as allowed by the relevant parameter if set to 1.

The most representative example is `InflClass`,[4] which reflects the original endings of the Proto-Indo-European stems. `InflClass` has not been added everywhere in UDante, therefore – when higher-level validation is turned on – it is only considered as allowed, instead of required.

5 Harmonisation Examples

Three treebanks have been harmonised to the higher level of detail (as defined in the previous sections): LLCT, ITTB and UDante. The remaining two treebanks (Perseus and PROIEL) have been harmonised to the lower level because the high-level annotation is not available for them.

The harmonisation process derives transparently from the feature constraints in the UDapi block. It would not be helpful to discuss every constraint in detail here (and if necessary, the reader can refer directly to the source code of the block); nonetheless, we want to discuss some interesting examples regarding verbs and auxiliaries. There is a more general issue raised by Cecchini (2021), who proposes a reorganisation of Latin non-finite verbal features towards a higher degree of universality. In accordance with their proposal,[5] we reannotate all gerund and gerundive forms as participles (`VerbForm=Part`) with `Aspect=Prosp`. Traditional terminology used in grammars, i.e. gerund and gerundive, is saved in the MISC field as `TraditionalMood=Gerund` and `TraditionalMood=Gerundive` to prevent loss of information and allow linguistic research based on traditional categories. Similarly, supine forms are reannotated as `VerbForm=Conv` with `Aspect=Prosp` and `TraditionalMood=Sup`. The use of `TraditionalMood` and `TraditionalTense` is extended to finite forms as well, for the purpose of consistency and in line with UDante. As far as finite forms are concerned, auxiliaries occurring in ITTB require some intervention as well. Unlike in the other treebanks, such forms (e.g. *sum* 'they are') do not present `Aspect`, `Mood`, `Person` and `Tense`. For the sake of consistency, we annotate them with respect to those features, assigning the relevant value.

Overall, the examinations of bugs highlighted by the block confirms what has been already noted in Gamba and Zeman (2023) with respect to Perseus and PROIEL status: their level of annotation detail is remarkably lower in comparison to ITTB, LLCT and UDante. An outstanding example is provided by `PronType`, which is a key feature for pronouns and determiners. Often missing in particular in Perseus, it is systematically added during the harmonisation process.

Additionally, the block can also serve as a tool to spot isolated errors. Whenever such errors are highlighted, we proceed to correct them.

Table 1 presents a quantitative overview of the major interventions applied.

6 Impact on Parsing

To evaluate the significance of the harmonisation process of morphological features, we try to investigate its impact on parsing accuracy. Therefore, we train new models for every morphologically harmonised treebank. The models are trained on the same data, but in the first case UDPipe 1.2 is used, while for the second one we choose to employ Stanza. With both Stanza and UDPipe we train the parser model on predicted lemmas and tags. Indeed, through Stanza's `prepare_depparse_treebank.py` script,[6] the trained POS tagging model is used to retag the training data before training the parser. Similarly, for UDPipe[7] we train a parsing model that relies on lemmas, UPOS tags and features as generated by the tagger. We use pretrained fastText embeddings[8] (Grave et al., 2018) and training hyperparameters as used for syntactic harmonisation in Gamba and Zeman (2023). For UDPipe, these hyperparameters correspond to the optimised ones

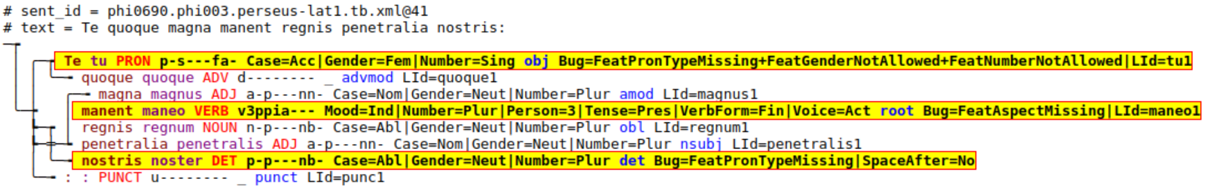
[4] https://universaldependencies.org/la/feat/InflClass.html.

[5] With the only exception of the VNoun feature, which has eventually not been introduced in UDante.

[6] https://stanfordnlp.github.io/stanza/training_and_evaluation.html.

[7] https://ufal.mff.cuni.cz/udpipe/1/users-manual#model_training_parser.

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



```
# sent_id = phi0690.phi003.perseus-lat1.tb.xml@41
# text = Te quoque magna manent regnis penetralia nostris:
Te tu PRON p-s---fa- Case=Acc|Gender=Fem|Number=Sing obj Bug=FeatPronTypeMissing+FeatGenderNotAllowed+FeatNumberNotAllowed|LId=tu1
quoque quoque ADV d-------- _ advmod LId=quoque1
magna magnus ADJ a-p---nn- Case=Nom|Gender=Neut|Number=Plur amod LId=magnus1
manent maneo VERB v3ppia--- Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act root Bug=FeatAspectMissing|LId=maneo1
regnis regnum NOUN n-p---nb- Case=Abl|Gender=Neut|Number=Plur obl LId=regnum1
penetralia penetralis ADJ a-p---nn- Case=Nom|Gender=Neut|Number=Plur nsubj LId=penetralis1
nostris noster DET p-p---nb- Case=Abl|Gender=Neut|Number=Plur det Bug=FeatPronTypeMissing|SpaceAfter=No
: : PUNCT u-------- _ punct LId=punc1
```

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

	ITTB	LLCT	Perseus	PROIEL	UDante	notes
Aspect	26,243	4,596	4,344	35,420	-	Aspect is added.
Gender_N	2,655	9,746	1,037	11,756	-	Gender is added, corrected or deleted (nominal only).
Gender_V	1,514	1,834	30	3,899	-	Gender is added, corrected or deleted (verbal only).
Gerund(ive)s	2,740	1,855	91	1,046	-	Interventions on gerunds and gerundives.
Mood	20,269	-	-	-	-	Mood is added.
Number_V	21,783	1,834	30	322	-	Number is added (verbal only).
NumForm=Word	2,029	2,415	162	1,671	142	NumForm=Word is added to numerals like *viginti* ‘twenty’.
Person	20,269	-	-	-	-	Person is added to verbs.
Person_P	-	-	1,346	15,887	-	Person in pronouns is either added, if missing, or deleted, if not relevant.
PronType	24,825	21,062	3,105	31,023	21	PronType is either added, if missing, or corrected.
Tense	51,096	10,988	1,277	9,430	-	Tense is either added, corrected or deleted.
Voice	2,591	1,855	216	1,064	-	Voice is added when missing.
Voice_NO	-	4,113	369	7,848	-	Voice is deleted when not relevant.

Table 1: Count of harmonising interventions.

made available for reproducible training by Straka and Straková (2019) when available (ITTB, Perseus and PROIEL), and to parameters inspired by those in the case of LLCT and UDante.[9]

We then evaluate the parsing model on morphologically harmonised test data for each treebank and compare results to the accuracy scores obtained with parsing models trained on data that underwent a harmonisation process only at syntax level.[10] Tables 2 and 3 report results obtained with UDPipe, in terms of Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) (Buchholz and Marsi, 2006), whereas Tables 6 and 7 presents analogous scores as obtained by the model trained with Stanza. Scores highlighted in blue denote an increase, while scores highlighted in red pinpoint decreased results. Accuracy is measured with the evaluation script[11] designed for the CoNLL 2018 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., 2018), which takes into consideration main dependency relations only and not subtypes.

First and foremost, a clarification is necessary. As explained earlier, the treebanks are not forced all to the same set of features: LLCT, ITTB and UDante have some extra features that are not found in Perseus and PROIEL. It would be possible to remove these extra features for the sake of parsing evaluation but we chose to keep them. One can thus expect somewhat worse results when applying models from one of these treebank groups to test data from the other group.

As illustrated in the tables, the results do not show any clear pattern and, overall, the improvements are neither widespread nor substantial. A closer look at the scores reveals that UDPipe shows improved accuracy scores in less than half of the cases, and in general performs worse than Stanza, with the gap being almost around 10% on average. Improvements obtained with models trained on UDPipe are never substantial and, in general, very hard to interpret. Stanza seems to allow for some additional remark. We first want to examine distinctly the two groups that correspond to the two possible values of the discussed parameter. The

[9] 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.

[10] In both cases parsing models are trained on predicted tags.

[11] https://github.com/UniversalDependencies/tools/blob/master/eval.py.

	ittb.udp		llct.udp		perseus.udp		proiel.udp		udante.udp	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
ITTB	**79.86**%	**83.11**%	38.62%	50.59%	44.16%	53.86%	45.19%	55.56%	53.51%	63.06%
LLCT	35.50%	45.63%	**91.84**%	**93.20**%	32.64%	42.66%	35.81%	47.55%	30.86%	41.31%
Perseus	44.14%	55.57%	32.60%	45.50%	**43.73**%	**57.28**%	40.36%	53.25%	42.13%	54.23%
PROIEL	49.37%	58.58%	36.67%	48.72%	45.23%	54.41%	**70.02**%	**75.16**%	41.85%	53.13%
UDante	46.89%	57.28%	31.85%	44.31%	34.51%	45.73%	35.50%	47.64%	**48.24**%	**57.99**%

Table 2: UDPipe LAS and UAS before morphological harmonisation. 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	**81.01**%	**84.05**%	39.31%	51.29%	45.22%	55.37%	44.42%	54.10%	53.66%	62.40%
LLCT	34.76%	44.55%	**91.57**%	**92.72**%	32.12%	41.00%	36.55%	48.25%	32.69%	42.44%
Perseus	42.89%	53.76%	31.52%	44.65%	**47.76**%	**57.33**%	39.99%	51.96%	41.49%	53.36%
PROIEL	49.96%	58.89%	36.84%	49.02%	45.16%	54.51%	**70.24**%	**75.59**%	41.80%	52.72%
UDante	46.31%	56.18%	31.20%	43.72%	34.20%	45.60%	35.76%	46.51%	**47.99**%	**57.44**%

Table 3: UDPipe LAS and UAS after morphological harmonisation. Columns correspond to trained models, rows to test data.

LLCT model obtains lower accuracy scores only on Perseus, which presents a more coarse-grained morphological annotation, but not on any of the treebanks belonging to the same class. A similar remark could be made about the ITTB model; the lower scores obtained on ITTB test data, despite being coloured in red, are probably not significant. Nevertheless, this reasoning does not hold true for the model trained on UDante, which incongruously performs best on Perseus and PROIEL. On the other hand, the PROIEL model is the only one showing improvements on all test data; despite not being substantial in most of the cases, a +3% increase can be observed when the model is used to parse LLCT data.

All the discussion so far concerns syntactic parsing, which is only indirectly affected by the consistency of morphological annotation. So the natural next question is about the impact of the harmonisation on prediction of morphology. Both UDPipe and Stanza predict morphological annotation together with syntax. Tables 4 and 8 show accuracy of feature prediction (percentage of correct words, whereas a word is correct if all its feature-value pairs have been predicted correctly). Each accuracy is computed before and after harmonisation, shown in the same table. Here we see a clear improvement in all experiments where a model is applied to data from different treebank; and for ITTB and PROIEL, the improved consistency led to improvement also in the in-domain experiment. The improvement is further confirmed in Tables 5 and 9, which show the MLAS scores (Zeman et al., 2018), combining morphology and syntax.

7 Conclusion and Future Work

The paper presents the harmonisation process that we carried out, with respect to morphology, on the five Latin UD treebanks. We first defined an UDapi block for Latin, listing which morphological features a token should possess. Such lists of features are defined based on UPOS tags. Subsequently, we corrected the retrieved inconsistencies – consisting in either missing or not allowed features – via Python scripts. As a result, we produced morphologically harmonised versions of the Latin treebanks that were previously harmonised syntactically (Gamba and Zeman, 2023). We contributed the script to investigate Latin features, possibly reusable by anyone working on Latin treebanks, and we described a workflow that can be replicated and applied to potentially any other language, provided that language-specific information is supplied within the template. In the second part of the paper, we presented some parsing experiments carried out with UDPipe and Stanza. By comparing syntactic attachment scores before and after morphological harmonisation, we observed the absence of a clear pattern that would allow to explain results; on the other hand, morphological accuracy clearly improved. The coexistence of a coarse-grained and a fine-grained level of consistency in annotation partially explains the outcome of the parsing experiments, that however must not discourage from pursuing an ever-growing harmonisation of linguistic resources in terms of annotation choices. Intra- and inter-resource consistency is a key factor to exploit data, whether it comes to

	ittb.udp		llct.udp		perseus.udp		proiel.udp		udante.udp	
	before	after	before	after	before	after	before	after	before	after
ITTB	**93.57**%	**93.91**%	55.41%	63.72%	53.76%	69.09%	54.50%	78.02%	62.67%	70.68%
LLCT	52.38%	60.39%	**95.89**%	**95.86**%	50.53%	60.36%	54.45%	67.45%	50.59%	58.68%
Perseus	52.54%	65.25%	46.74%	55.10%	**72.03**%	**71.11**%	69.45%	76.26%	45.12%	57.77%
PROIEL	46.47%	69.98%	45.12%	56.83%	61.16%	69.11%	**87.19**%	**88.87**%	40.35%	59.81%
UDante	58.30%	64.99%	47.47%	54.90%	44.60%	59.29%	48.30%	69.57%	**74.84**%	**74.67**%

Table 4: Comparison of UDPipe accuracy scores on morphological features. Columns correspond to trained models, rows to test data.

	ittb.udp		llct.udp		perseus.udp		proiel.udp		udante.udp	
	before	after	before	after	before	after	before	after	before	after
ITTB	**69.97**%	**71.64**%	15.10%	17.23%	15.24%	22.90%	18.68%	30.85%	25.39%	29.04%
LLCT	10.41%	13.14%	**85.76**%	**85.50**%	6.49%	11.38%	11.07%	17.04%	7.52%	8.93%
Perseus	15.37%	21.98%	8.68%	12.80%	**28.60**%	**28.89**%	23.59%	29.45%	10.70%	17.59%
PROIEL	16.14%	29.49%	10.81%	15.58%	19.00%	25.21%	**56.42**%	**58.07**%	11.47%	18.82%
UDante	18.87%	21.32%	8.62%	10.15%	8.94%	13.53%	11.97%	19.43%	**25.90**%	**25.46**%

Table 5: Comparison of UDPipe MLAS scores. 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 LAS and UAS before morphological harmonisation. 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	**88.29**%	**90.28**%	46.93%	60.21%	50.02%	60.22%	52.86%	62.13%	64.87%	72.91%
LLCT	42.18%	54.50%	**94.91**%	**96.08**%	38.10%	48.50%	42.48%	56.08%	42.43%	54.97%
Perseus	59.00%	69.00%	39.82%	53.34%	**59.43**%	**68.97**%	47.97%	59.36%	54.26%	65.17%
PROIEL	62.33%	71.27%	48.17%	61.25%	55.56%	64.81%	**81.25**%	**84.91**%	54.37%	64.41%
UDante	58.24%	68.42%	40.39%	53.84%	39.73%	52.47%	41.41%	52.74%	**57.40**%	**66.79**%

Table 7: Stanza LAS and UAS after morphological harmonisation. Columns correspond to trained models, rows to test data.

	ittb.mdl		llct.mdl		perseus.mdl		proiel.mdl		udante.mdl	
	before	after	before	after	before	after	before	after	before	after
ITTB	**95.70**%	**96.15**%	57.07%	66.19%	55.19%	72.91%	52.14%	79.97%	66.22%	75.34%
LLCT	56.92%	63.95%	**96.89**%	**96.81**%	53.53%	65.33%	57.07%	71.87%	55.73%	63.47%
Perseus	57.29%	72.49%	48.66%	57.23%	**78.02**%	**77.86**%	70.01%	79.51%	49.75%	64.63%
PROIEL	49.88%	75.90%	48.31%	60.97%	66.57%	75.95%	**90.91**%	**92.72**%	44.53%	67.10%
UDante	62.47%	69.85%	48.56%	56.32%	45.89%	63.42%	46.22%	70.64%	**79.39**%	**79.30**%

Table 8: Comparison of Stanza accuracy scores on morphological features. Columns correspond to trained models, rows to test data.

	ittb.mdl		llct.mdl		perseus.mdl		proiel.mdl		udante.mdl	
	before	after	before	after	before	after	before	after	before	after
ITTB	**78.97**%	**80.74**%	16.56%	19.07%	19.45%	27.87%	22.13%	40.05%	33.14%	39.59%
LLCT	12.22%	17.67%	**89.46**%	**90.04**%	9.12%	16.63%	15.98%	24.25%	12.59%	18.02%
Perseus	22.63%	35.20%	11.57%	16.92%	**38.86**%	**40.21**%	31.33%	38.66%	16.25%	27.29%
PROIEL	22.23%	41.32%	14.86%	22.74%	27.64%	35.92%	**68.49**%	**71.23**%	17.17%	30.61%
UDante	25.06%	29.95%	12.21%	14.77%	10.64%	17.37%	13.45%	25.40%	**35.96**%	**35.32**%

Table 9: Comparison of Stanza MLAS scores. Columns correspond to trained models, rows to test data.

linguistic research or any other application.

In light of the slight improvement that resulted in parsing accuracy from the harmonisation process, we do not plan on further developing the harmonisation of treebanks. The higher degree of consistency in treebank annotation, i.e. the availability of more homogeneous data, allows now to investigate the actual reasons for variability in parsing. Syntactic constructions evolving over time may be inspected, as well as other factors that may affect parsing results on data that differ from training data – as already problematised several times, e.g. by Passarotti and Dell'Orletta (2010), Passarotti and Ruffolo (2010), Ponti and Passarotti (2016). Variation in time is most probably expected to be a relevant factor, and it is strongly connected to two other relevant variables, i.e. space and domain. Consider, for instance, the Late Latin Charter Treebank: while featuring early medieval Latin (VIII-IX century), not as late as ITTB (XIII century) and UDante (XIV century) Latin varieties, the treebank does not include literary texts yet charters written in Tuscany, Italy. The gradual development of Latin towards Romance languages, exemplified by evolving syntactic constructions and changes in word endings, can already be observed in the treebank (Cecchini et al., 2020c). Variation in terms of genre appears to be relevant also with respect to the distinction between poetry and prose. With Latin treebanks encompassing mostly literary data, such distinction cannot be overlooked. Indeed, Latin poetry is strongly affected by prosody and metre: the sequence of short and long syllables in words, as defined by prosodic rules, together with the specific structure of the selected metre, rigidly determine possible sequences of words. As a result, the natural word order is unsettled, and the position of a word in the verse (and, hence, in the sentence) is mostly defined by the way its short and long syllables follows one another. This whole mechanism, highly affecting word order, entails a high degree of non-projectivity, and would need to be further inspected.

Acknowledgements

This work was supported by the Grant No. 20-16819X (LUSyD) of the Czech Science Foundation (GAČR).

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 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. 2021. Formae reformandae: for a reorganisation of verb form annotation in Universal Dependencies illustrated by the specific case of Latin. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 1–15, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020c. A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942, 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.

Federica Gamba and Daniel Zeman. 2023. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

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.

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 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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. 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, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. 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, 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.