Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 933–942
Marseille, 11–16 May 2020

A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages

Flavio Massimiliano Cecchini°, Timo Korkiakangas*, Marco Passarotti°
°CIRCSE, Università Cattolica del Sacro Cuore; *Helsingin yliopisto
°Largo Gemelli 1, 20155 Milan, Italy; *Yliopistonkatu 3, 00014 Helsinki, Finland
°{flavio.cecchini,marco.passarotti}@unicatt.it; *timo.korkiakangas@helsinki.fi

Abstract

The present work introduces a new Latin treebank that follows the Universal Dependencies (UD) annotation standard. The treebank is obtained from the automated conversion of the Late Latin Charter Treebank 2 (LLCT2), originally in the Prague Dependency Treebank (PDT) style. As this treebank consists of Early Medieval legal documents, its language variety differs considerably from both the Classical and Medieval learned varieties prevalent in the other currently available UD Latin treebanks. Consequently, besides significant phenomena from the perspective of diachronic linguistics, this treebank also poses several challenging technical issues for the current and future syntactic annotation of Latin in the UD framework. Some of the most relevant cases are discussed in depth, with comparisons between the original PDT and the resulting UD annotations. Additionally, an overview of the UD-style structure of the treebank is given, and some diachronic aspects of the transition from Latin to Romance languages are highlighted.

Keywords: Latin, treebank, Universal Dependencies

1. Introduction

1.1. Universal Dependencies

Universal Dependencies (UD) is an open-access and collaborative project that aims to provide "a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages"[1] (Nivre et al., 2016). The current release (2.5 as of February, 2020) of UD includes 157 treebanks over 90 languages, among them several ancient languages (such as Akkadian and Gothic) and historical phases of modern ones (such as Old French or Classical Chinese).

Currently, Latin features the most data (ca. 582 000 tokens) and the most treebanks of all the ancient languages of UD (three). The Latin treebanks are: PROIEL (Haug and Jøhndal, 2008), which includes the entire New Testament in Latin and texts from the Classical and Late era (ca. 200 000 tokens); the Latin Dependency Treebank (LDT) by the Perseus Digital Library (Bamman and Crane, 2006), comprised of a small selection of texts by Classical authors (ca. 29 000 tokens); and the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2019), based on works written by Thomas Aquinas in the XIII century (353 000 tokens). This testifies to the huge diachronic, diatopic and diamesic variety of Latin: initially the language of Ancient Rome, it was for a long time the European *lingua franca* of arts, sciences and law, including official documents; it still is the official language of the Catholic Church (Cecchini et al., forthcoming). Because of this variety, no single treebank alone can sufficiently represent the entire Latinity.[2] The present UD conversion the Late Latin Charter Treebank 2 (LLCT2), a treebank of Latin from the VIIIth and IXth centuries thus adds another *tessera* to the mosaic: the LLCT is the first treebank of non-Classical and non-literary Latin.

1.2. The Late Latin Charter Treebanks

The LLCT2 is the second part of the Late Latin Charter Treebank (LLCT). It was built between 2016 and 2018 as a chronological extension of the LLCT1[3] and contains 257 918 tokens from 521 early Medieval Latin original documents (*cartulae* 'charters') written in Tuscany, Italy, between A. D. 774 and 897. The LLCT2 charters were mainly written by professional notaries to record private transactions, such as the selling, exchange, and renting out of property. The LLCT2 text has originally been published in four out-of-copyright editions in the XIXth and the early XXth centuries.

The language of the charters is a variety of Latin that differs from both Classical and Medieval Latin standards in terms of lexicon, spelling, morphology and syntax. Charters are useful resources for historical Latin linguistics because they shed light on the transition between Latin and Italo-Romance that was taking place in those centuries. However, charter Latin is highly formulaic, which means that prefabricated formulae were exploited to guarantee the legal validity of the charter. Due to this repetition, the type/token ratio in the LLCT2 is only 0.04. Since formulae draw on a centuries-old legal Latin tradition, due caution must be observed when charter Latin is used to draw linguistic conclusions on the spoken language of the time.

This paper describes some of the most important annotation and conceptual challenges encountered during the conver-

[1] From: http://universaldependencies.org/

[2] For instance, (Ponti and Passarotti, 2016) shows a dramatic decrease of accuracy rates of a dependency parsing pipeline trained on the IT-TB when applied to texts of the Classical era taken from the LDT.

[3] The LLCT1 is available at https://zenodo.org/record/1197357#.XbmDlmZS9EY in the Prague Markup Language (Hana and Štěpánek, 2012) format (online documentation at http://ufal.mff.cuni.cz/jazz/pml/doc/pml_doc.pdf).

sion of the LLCT2 into the UD framework and is organised as follows: Section 2. describes the original treebank and sketches the process of conversion into UD; Section 3. goes into the detail of how some syntactic and linguistic phenomena are tackled during the conversion process, and Section 4. gives an overview of the major formal differences between the two versions of the LLCT2. Finally, Section 5. concludes the paper and points to future work.

2. The Structure and Conversion Process of the LLCT2

In this section we briefly present the annotation approach of the original LLCT2 treebank and discuss the technicalities of the UD conversion process.

2.1. The Original Annotation of the LLCT2

The original annotation of the LLCT2 mainly follows the LDT and IT-TB styles, which are based on (Bamman et al., 2007, henceforth *Guidelines*). The latter rests on the annotation style for the analytical layer of the Prague Dependency Treebank (PDT) presented in (Hajič et al., 2017). The LLCT2 was first automatically annotated and then manually corrected. However, the *Guidelines* only apply to Latin that follows regular Classical Latin grammar.[4] Therefore, the non-Classical syntactic (and morphological) features had to be annotated following an additional set of rules described in (Korkiakangas and Passarotti, 2011), some of which are more closely detailed in Section 3.

The morphological annotation in the LLCT2 is based on the categories present in Classical Latin grammar. Since the LLCT is meant to be used to study the evolution of Latin that had taken place between Classical and Early Medieval Latin, the non-classical word forms attested in LLCT are given the morphological analyses of their equivalents in Classical Latin. For example, in the LLCT phrase *debeam reddere soledo uno* 'I have to pay one gold coin', the form *soledo* 'gold coin' (the direct object of the verb *reddere* 'to pay') is analysed as a masculine singular accusative like its Classical ancestor *solidum*, although it is questionable whether the Latin of the VIIIth and IXth centuries still had a morphologically expressed *o*-stem case. The masculine singular ending *-o* of the Romance type (cf. Italian *soldo* 'coin; pay') is likely to have derived from the accusative case form, hence its analysis as such (Zamboni, 2007, §3.1.6). A somewhat more complex issue is gender change: *offertas* 'offerings, tithe' has a seemingly feminine plural ending *-as*, but the form is annotated as a neuter plural because the original Classical Latin-like form *offerta* was a neuter plural. It is known that Latin neuter plurals became feminine in Romance languages (Väänänen, 2012, §215), but if *offertas* were annotated as a feminine, the possibility to track the underlying diachronic change would be lost.

[4]The Latin variety used by Thomas Aquinas in the IT-TB, albeit Medieval, is quite formal and tends to follow the morphology and syntax of Classical Latin.

No consensus prevails about the grammatical categories of the spoken Latin of the time. Therefore, the evolutionary principle that exploits Classical Latin grammatical categories seems to be the only workable way to annotate non-Classical Latin. In sum, the annotation style of the LLCT2 incorporates a diachronic component and, consequently, the annotation of the LLCT2 does not claim to be a synchronic representation of any specific stage of Latin. This is at the same time an obstacle and an added value for typological comparisons, since it provides a collection of data belonging to a late and very specific variety of Latin, and yet is diachronically related to the Classical one. A prospective user of the UD version of the LLCT2 has to be aware of the peculiarities of its annotation.

2.2. Conversion into UD and the UD Style

The scripts behind the conversion process are written in the Perl language[5] and function as modules of Treex's architecture.[6] The scripts were developed as part of the HamleDT (**ha**rmonised **m**ulti-**l**anguag**e** **d**ependency **t**reebank) project (Zeman et al., 2012; Zeman et al., 2014; Rosa et al., 2014) and can be broken down into two sections: the *harmonisation* phase and the UD *conversion proper*. Since its latest version (3.0), HamleDT has been using the UD syntactic annotation style,[7] and so the conversion process has been readapted for UD.[8]

"Harmonisation" here means the adjustment of a treebank so as to more closely conform to the PDT morphosyntactic annotation style originally used by HamleDT (versions 1.0 and 2.0) that serves as the basis for the current UD conversion proper, which rewrites morphological features and intervenes on the structure of the syntactic trees in the second phase. Since the LLCT2 is very close to the PDT standard to start with (see Section 2.1.), in our case the harmonisation is mainly used to encode parts of speech (POS) and morphology by means of the Interset tagset (Zeman, 2008)[9] used by HamleDT and UD, and whose formal aspects and terminology are inspired by (Petrov et al., 2011) and described in (Sylak-Glassman, 2016; Zeman, 2018). In this phase, we also make some preliminary structural and annotation modifications in view of the UD re-interpretation of some linguistic phenomena (a part of which is illustrated in Section 3.).

Finally, the conversion proper is carried out by the main script and the output in the UD style is returned according to the CoNLL-U format, i. e. as a plain-text, csv-like file

[5]`https://www.perl.org/`

[6]Treex is a modular software system in Perl for Natural Language Processing. It is described in (Popel and Žabokrtský, 2010) and is available online at `http://ufal.mff.cuni.cz/treex`.

[7]See `https://ufal.mff.cuni.cz/hamledt/news`.

[8]The conversion process sketched here closely follows the one used for the UD conversion of the IT-TB, which is described in (Cecchini et al., 2018).

[9]See `https://ufal.mff.cuni.cz/interset`. For details on the tagset, see also `https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset:features`.

that follows a revised version of the CoNLL-X format (Buchholz and Marsi, 2006).[10] The harmonisation and conversion scripts can be downloaded from the Github pages of the Treex and HamleDT projects. The UD project also provides a validator script[11] in Python[12] to ensure that the final result complies with the general UD principles. The differences between the UD principles and the PDT style of the original LLCT2 form the basis of the discussion in Sections 3. and 4.

2.3. Notation

In the following, we refer to labels used for syntactic relations as uppercase *afun*s (**a**nalytical **fun**ctions) when dealing with the PDT style of the original LLCT2 treebank,[13], and as lowercase *deprel*s (**dep**endency **rel**ations) with regard to the UD style.[14] When citing an LLCT2 sentence, we use "s2019" for the sentence number 2019, and "*a. o.*" ("and others") if the same formula also recurs in other sentences.

3. Challenging Syntactic Structures

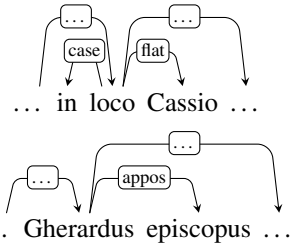
The differences between the annotation styles of LLCT2 and UD pose a challenge to any automated conversion process. The most marked difference is that in the LLCT2 conjunctions, prepositions and copulas govern phrases, while UD favours dependencies between content words so that function words tend to end up as leaves of the syntactic tree.[15] This has consequences on the resulting syntactic trees (see Section 4.), especially on the rendition of often recurring syntactic constructions like co-ordinations and appositions. Appositional constructions in particular entail many linguistically and technically complex aspects, and they are tackled in Section 3.1. Moreover, the Latin of the LLCT2 features a few syntactic constructions that are absent, or only rarely attested, in Classical Latin, and that have been varyingly, if at all, treated in other Latin treebanks. In the following, we describe some of these constructions, along with how we deal with them in the UD framework. Such interpretations have, as a broader consequence, the constructive revision and improvement of the current UD annotation schemes for Latin, which still do not have official guidelines.

3.1. Apposition-like Constructions

Due to the formal and legal nature of the LLCT2, the notaries regularly strove to specify all the contracting parties as precisely as possible. This involves the use of proper names with kinship terms, such as *filius* 'son' and *germanus* 'brother', or titles, such as *imperator* 'emperor' and *archiepiscopus* 'archbishop', but also other expressions of qualification, profession, social status and the like. Proper names of locations are also often specified in their nature by generic terms, such as *locus* 'place', *rivus* 'creek' or *civitas* 'city': e. g. *civitate Papia* '[in] the city of Pavia' (s7133), *loco Cassio* '[in] the place [called] Cassium' (s3910 a. o.).

On a syntactic level, such specifications give rise to *appositions*. Broadly speaking, the Latin apposition is defined as a noun phrase acting as the co-referential complement of another noun phrase by means of case agreement and simple juxtaposition, with very few exceptional cases of discontinuity (Longrée, 1990; Spevak, 2014, Ch. 4).[16] Of particular interest are the many appositions (15 414 occurrences out of a total of 16 760 appositional constructions) where at least one of the two participating elements is a proper name, like the above-mentioned *loco Cassio* or *Gherardus episcopus* 'Gherardus, the bishop'/'bishop Gherardus' (s5715 a. o.). In such cases, the position of the specifying element in relation to the proper noun is the criterion by which we can distinguish so-called *close* appositions from *free* ones (Longrée, 1990; Spevak, 2014, Ch. 4).

A close apposition implies a stronger bond between its constituents than a free one, which on the contrary is more similar to a predicative construction than to an attributive one. While the sequence *noun phrase + proper noun* (as in *loco Cassio*) can undoubtedly be considered a close apposition where the first element acts as its (weak) head, more doubts arise whether the sequence *proper noun + noun phrase* (as in *Gherardus episcopus*) is to be held as a close or free apposition (Longrée, 1990). We argue that, in the LLCT2, the latter pattern is always a free, non-restrictive apposition with the proper noun as its head, and as such we assign it the UD deprel `appos`, as opposed to `flat` for close appositions (see Section 3.1.1.).[17] In the examples above:



[10] `https://universaldependencies.org/format.html`

[11] `https://universaldependencies.org/release_checklist.html#validation`

[12] `https://www.python.org/`

[13] See `https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/` for details about the single relations.

[14] The complete UD guidelines are available at `https://universaldependencies.org/guidelines.html`.

[15] For more details, refer to `http://universaldependencies.org/u/overview/syntax.html`.

[16] We notice here that some constructions traditionally labeled as "appositions", where a term is specified later in the sentence by a phrase introduced by an adverbial element like *scilicet* or *id est* 'that is', actually belong to the realm of co-ordination, and their introductory elements are to be treated as (explicative) conjunctions. Uncertainties in their treatment in UD are discussed in (Cecchini et al., 2018, §2.2.3).

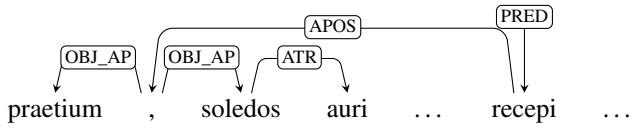
[17] Cf. the discussion at `https://github.com/UniversalDependencies/docs/issues/536#issuecomment-375030335`.

In the original LLCT2 annotation, appositional constructions of this kind are considered simple chains of noun phrases where the head noun is modified by another noun via the *afun* ATR. This does not, however, apply to non-classical constructions that indicate price or quantity, e. g. s6:

> *recepi . . . praetium in praefinito, auri soledos numero quinque*
> 'I received . . . the price as agreed, five gold coins',

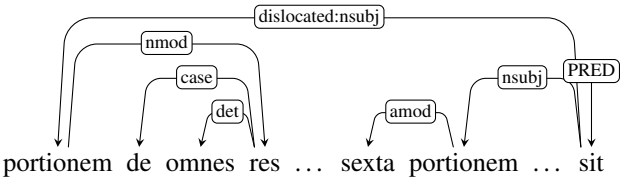
which are analyzed in the LLCT2 as appositive constructions and marked with APOS (*praetium* 'price' and *soledos* 'gold coins' depend on the intervening comma tagged as APOS).

This use of APOS represents the standard way of treating appositions in the *Guidelines* (Bamman et al., 2007, p. 28): the two noun phrases are made dependent on the comma as their "connecting element" and are labelled with the same syntactic function in the sentence, while the suffix _AP marks them as part of this construction:



In the appositive constructions of the LLCT2, the two or more juxtaposed items can be rather distant from each other: in s560 *omnem rebus substantie mee . . . [23 intervening words] . . . tam coltum quam et desertum*, the quality of *rebus* 'things, property' is specified by *tam coltum quam et desertum* 'both cultivated and deserted land' after lengthy subordinate clauses.

Another peculiarity of the LLCT2 is that APOS is also used with reprisals of noun phrases. This phenomenon is sometimes caused by an interruption of the writer's train of thought, or by his wish to ensure that a specific noun phrase remains topical, i. e. under discussion, as is the case in s20 *portionem de omne res eius, quot est sexta portionem de ex omnibus rebus meis . . . [53 intervening words] . . . de omnia, ut dicxi, sexta portionem . . . sit . . .* '(a) part of all his property, which is the sixth part of all my property . . . as I said, the sixth part of everything . . . be . . . '. Here, the latter *portionem* 'portion, part' is even picked up by *ut dicxi* 'as I said'. This phenomenon is related to the fact that formulaic sentences are often long and that the legal validity of the document depends on its correct wording. In addition, the LLCT2 abounds with sentences that contain various combinations of appositions and reprisals.
In our UD treatment of the LLCT2, we use the *deprel* dislocated, along with the already discussed apos, to represent the structure of such sentences. This is currently the best way to deal with reprisals of this kind, in which we can distinguish between a "dislocated" element, i. e. the element that is more peripheral with respect to the predicate and is usually also the longer and more complex one, and an "anaphoric" element, i. e. an abbreviated repetition of the former which is more "core" with respect to the predicate. Both elements are made dependent on the predicate, but while only the latter receives its core *deprel*, the former is tagged as dislocated with an appropriate subrelation. In our example:



We identify 834 such cases (i. e. affecting the 9.24% of sentences in the corpus) in the LLCT2. Among them, we report 7 cases in which the UD relations required by the dislocated and by the anaphoric elements are asymmetric, such as in s357, where *quidquid . . . remanserit* 'whatever . . . is left' is clausal and gets the *deprel* ccomp, while *meam portionem* 'my part' (dislocated) would be assigned obj. Since in this and in all other such cases one of the two elements is a so-called *free relative* clause (Grosu, 2003), the need for a coherent interpretation of this phenomenon in UD will require further analyses and, in the light of the appositional constructions at hand, might encourage a rethinking of part of the overall annotation strategy for Latin.

Note that our UD conversion currently does not include 121 sentences of the original corpus wherein the use of the *afun* APOS is reputed to be problematic. This applies in particular to sentences where 1. the APOS relation is not binary, i. e. there are more than two elements marked by the _AP suffix, and where 2. the APOS appears as the root. Further investigation is needed to assess how these anomalous structures are to be analysed in the UD framework, and if their annotation in the LLCT2 should actually be modified.

3.1.1. Close versus Free Appositions

Table 1 presents figures concerning the above discussed appositions with proper nouns. For this analysis, only non co-ordinated appositions are examined. Our first observation is that noun phrases postposed to proper nouns almost exclusively concern terms related to kinship and social status, and, consequently, to human beings. On the contrary, all the toponymal specifications are preposed, except for a few postposed occurrences (*civitas* 'city' twice, and *ecclesia* 'church' and *plebs* 'parish' once each). Even then, there is at least one preposed occurrence, as *civitate Papia*, s7133, vs. *Papia civitate*, s1513). The only cases in which human proper nouns are regularly postposed are either when they follow a first- or second-person pronoun (*ego*, *tu*, *nos* 'we', *vos* 'you (pl.)'), or when they are part of a fixed expression (*domnus*). The only term showing unstable behaviour is the multifaceted and very generic term *vir*, which is always further specified by an epithet, a fact we will explain below.

Lemma	Occurrences as first element	Occurrences as second element
ego 'I'	4069	0
tu 'you (sing.)'	825	0
domnus 'lord'	695	0
locus 'place'	585	0
fluvius 'creek'	40	0
finis 'region'	40	0
civitas 'city'	39	2
vir 'man'	21	13
filius 'son'	139	1226
homo 'man (person)'	1	29
germanus 'brother'	1	42
episcopus 'bishop'	1	788
presbyter 'priest'	0	1393
notarius 'notary'	0	1294
clericus 'cleric'	0	762
imperator 'emperor'	0	414
augustus 'august (emperor)'	0	389
diaconus 'deacon'	0	188
rex 'king'	0	177
scabinus 'alderman'	0	144
subdiaconus 'sub-deacon'	0	134
archipresbyter 'archpriest'	0	89
archidiaconus 'archdeacon'	0	73
advocatus 'counsellor'	0	72
habitator 'dweller'	0	60
massarius 'bailiff'	0	59
rector 'rector'	0	58
comes 'count'	0	43
ancilla 'maidservant'	0	31
abbatissa 'abbess'	0	29

Table 1: The 30 most frequent lemmas of tokens appearing in at least 25 appositional constructions with proper names, ranked by absolute frequency with respect to their use as the first element, and where the second element is not in the genitive case (this is needed to distinguish noun phrases that function as attributes from actual appositions).

These patterns can be explained by means of a generic definiteness criterion. Although this does not apply particularly well to legal contracts where various specifications are often used to unambiguously identify the contracting parties, in general, a name is sufficient to unambiguously identify a person, while further qualifications added as postposed appositions only serve to better describe their background, without being necessary. Instead, personal pronouns can only denote one specific individual in the context in which they are uttered. Indeed, we only observe the pattern *personal pronoun + proper noun* and never the inverse: it is the proper noun here that acts as extra information.
Consequently, we attest chains of appositions in this 'hierarchical' order, as in (s2406 a. o.) *ego Petrus episcopus* 'I, Petrus, the bishop' (as signature). Another sign that we are in the presence of free appositions, and also a motivation for their postposition, is their expandability with no functional changes, which signals a greater independence from its head, as in s7689:

> *. . . tu Daiprando diacono, rectorem adque custodem ecclesie beati sancti Petri sita foras civitatem ista Lucense ubi dicitur Maiore . . . dedisti. . .*
> '. . . you, Daiprandus, deacon, rector and keeper of the church of blessed Saint Peter, which lies outside this city of Lucca, which is called *Maior* . . . gave. . . '

One reason for the prevalence of close appositions with place names might be the stronger equivalence between the nature of a place and the place itself (represented by its name). While a person called Petrus might be a *gastaldus*, a *presbiter*, a *schabinus* or even the *apostolus*, and change his social status or occupation as time goes by, the status of the *fluvius Cassina* is practically immutable. In fact, while the syntactic head might be detected in *fluvius* 'creek' on the basis of agreement or similar factors (Spevak, 2014, Ch. 4, §3.4), semantically it is the apposition *Cassina* that allows the unambiguous identification of this landmark. The expression might be substituted by *is fluvius* 'that creek', but not by *fluvius* alone (Longrée, 1990). This actual 'headlessness' is managed in UD by using the deprel `flat`. This also clarifies why *domnus* or *vir beatissimus* 'a most blessed man' are always preposed while *imperator* 'emperor' or *vir excellentissimus* 'a most excellent man' are postposed: the former ones are seen as innate qualities of dominance or sanctity of the person, *domnus* having become too generic with respect to other titles of nobility and *vir beatissimus* being very close to *sanctus* (which is also always preposed), while the latter ones pertain more to the described person's social position. In fact, the rank of a *domnus* is nearly always further specified by *imperator* or a similar title, whereas *vir excellentissimus* might also be translated as 'gentleman'. This mirrors the different placements of adjectives according to their "qualifying" or "determining" function (Iovino, 2012, Ch. 2). A similar interpretation can be given to the constructions of the type *nomine Ermilinda* 'by the name of Ermilinda' (s140) or *numero quatraginta* 'in the number of forty' (s914), which we similarly analyse with `flat`.[18]

In all other cases where we encounter a nominal phrase modifying another nominal phrase (where the former is usually in the genitive or introduced by a preposition), we deem it to have a simple (i. e. non co-referential) attributive function; hence the application of the deprel `nmod`.

3.2. Auxiliary Verbs of Active Tenses

In Classical Latin, the only verb considered an auxiliary is *sum/esse* 'to be'.[19] It is used exclusively for passive tenses bearing a perfect aspect, e. g. *amatus sum* 'I was loved' vs. *amabar* 'I was being loved' (Barbieri, 1995, §142–144, 147–148), (Greenough and Allen, 2006, §184–188), and passive periphrases, e. g. *amandum est* 'it is to be loved/everybody has to love' (Barbieri, 1995, §156, 167), (Greenough and Allen, 2006, §193–196).[20] Among other

[18] This fact is corroborated by the observation that the second element can always be formally and functionally labelled as an ablative, like *nomine* or *numero*, and hence respect the criterion of case agreement between the two elements that are in apposition.

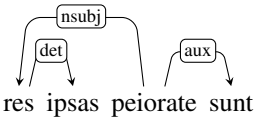
[19] We note that *eo/ire* 'to go' in its present passive infinitive form *iri* is used, together with a verb's supine in *-um* (bearing a final valence), similarly to an auxiliary in the periphrastic construction of the future passive infinitive, e. g. *ductum iri* 'to [be going to] be conducted' (Barbieri, 1995, §163),(Greenough and Allen, 2006, §509). However, this use is marginal with respect to the sheer mass of formations using *sum/esse*.

[20] Although it might be argued that *sum/esse* is used as a copula here.

structural changes that eventually have given shape to modern Romance languages, Late Latin saw the emergence of new periphrastically expressed active tenses based again on *sum/esse*, but also on the auxiliary *habeo/habēre* 'to have', accompanied by a perfect participle. Their origin is probably to be sought in the resultative use of a subject or an object predicative, respectively (Väänänen, 2012, §298, 300). For this reason, in these tense forms the perfect participle originally agreed with the subject or the object, respectively, while in modern Romance languages grammaticalization has been fully accomplished and agreement can be lost, especially in conjunction with the outcomes of the auxiliary *habeo/habēre*.[21] In the LLCT2, s5792, which belongs to a lawsuit from A. D. 871, may show two occurrences of these new formations:

> *non adimpletum* ***abetis***
> 'you (pl.) have not fulfilled [your duties]'
> *res ipsas pegiorate* ***sunt***
> 'things have worsened'

Here, *adimpletum* (from *adimpleo/adimplēre* 'to fulfil') is a generic singular neuter accusative without an explicit object (*abetis* is a spelling variant of *habetis* 'you (pl.) have'), and *pegiorate* (a spelling variant of postclassical *peioratae*) is a feminine plural, like *res* 'things'. Moreover, here the voice of *peiorate sunt* may be active and not passive, as would have been expected in Classical Latin from such a construction; this interpretation of *pegiorate sunt* is also possible because the form occurs parallel to the active verb *adimpletum habetis*, and the Classical counterparts would be *adimplevistis* and *peioraverunt*, respectively. In both cases we use the *deprel* `aux` and not `aux:pass`, an unprecedented fact with respect to Classical Latin as well as to the learned Medieval Latin of the IT-TB (Cecchini et al., 2018). E. g., the UD tree structure for the second example is:



While we can easily see that the auxiliary *habeo* occurs 14 times in the LLCT2, the use of *sum* in active tenses is *a priori* indistinguishable from its use in passive ones, so that manual disambiguation will be necessary to assess the extent of this phenomenon.

3.3. New and Old Conjunctions

Throughout its history, Latin has had a constant tendency to create new conjunctions from adverbialised forms of interrogative-relative pronouns or determiners, e. g. the conjunction *cum* 'when' from the accusative masculine singular form *quom* of *qui* 'who/which/that' (Palmer, 1988; Vineis, 2005). We can observe that this process is still active in the Late Latin of the charters. An adverb with a relative aspect lies in a syntactically grey area, so to speak, in that it tends to function without referring to an (explicit) antecedent. If its antecedent can be interpreted to be a preceding clause, the relative adverb takes on the role of a clause connector and is gradually grammaticalized as a conjunction. This phenomenon is actually very common among European languages and beyond (Kortmann, 1997).

In the LLCT2, the most frequent of such adverbs is *qualiter*, 'in what way', formed from the interrogative-relative determiner *qualis* 'of what kind', but clearly used to introduce a conditional/comparative adverbial subordinate clause. *Qualiter* is always found at the beginning of a clause with no corresponding referent in the main clause, similarly to the conjunctional *ut* 'as, like'. It is frequently used in formulaic expressions, such as *qualiter [supra/superius/...] legitur* 'as it can be read above' (179 out of 522 total occurrences of *qualiter*), or more freely, as in s34:

> *...dispensare ipsa res mea secundum Deo* ***qualiter*** *melius potueritis pro anima mea remedium*
> '...to spend my property in accordance with God, in the best way you (pl.) can, for the salvation of my soul'

The LLCT2 also has other words traditionally labelled as adverbs which behave like conjunctions (and are often coordinated among them, as in *comodo et qualiter* 'in what sort and manner', 12 occurrences): *quando/quandoque* 'when', *quantum* 'how much', *quatenus* 'how far', *quit* (= *quid*) 'that', *comodo/como/quomodo* 'in what manner', *ubicumque/ubicunque* 'wherever', *unde* 'wherefore'. Some of them survive with an identical function in modern Romance languages, like Italian *come* 'how' (formerly *como*), ultimately from *quomodo*. We label them with the UD part of speech `SCONJ`[22] and not `ADV`, and they consequently receive the *deprel* `mark`.

3.4. Anacolutha and Gaps

The LLCT2 charters are preserved as originals, which means that their language has not been emended during a long manuscript tradition, contrary to what has happened to literary Latin texts. Accordingly, the LLCT2 also features original misspellings, grammatical mistakes and gaps, unlike any existing Latin treebank. Any annotation style of such a language variety has to manage these mistakes and lacunae in some way.

Contrary to what is suggested by the *Guidelines*, the *afun* `ExD` is not used to mark ellipsis in the LLCT2 (cf. Section 3.5.). Instead, `ExD` is used for three other purposes in the LLCT2:

1. to mark the node closest to the root in sentences that do not contain a verbal predicate because they are mere nominal phrases, e. g. *signum + manus Pertualdi clerici* 'sign + of the hand of Pertualdus the cleric' (s12);

[21] For example, in Italian there still is agreement in e. g. ***lui*** *è venut**o*** / ***lei*** *è venut**a*** 'he/she has come' (it. *è* < lat. *est* 'he/she/it is'), but it is no longer present in e. g. *ho mangiat**o*** ***la*** *mela* 'I ate the apple' (it. *ho* < lat. *habeo* 'I have'), where *mela* 'apple' is feminine, but *-o* is the masculine ending.

[22] Until the current, i. e. 2.5, version, UD does not allow a closer specification of conjunction types, such as "relative".

2. to mark phrases that remain disconnected from the overall syntactic structure of a sentence for various reasons; a typical example of which is *id est* 'that is' (s159 a. o.). Another example is (s20)

   > *volo* ***ut*** *... filius meus Rachiprandu ...,* ***ut*** *portionem de omne res eius ...* ***ut*** *sit in potestatem Deusdedi presbiteri*
   > 'I want **that** ... my son Rachiprandus ..., **that** a part of all his possessions ... **that** it be in the possession of Deusdedit priest',

   where only the first subordinating conjunction *ut* 'that' is functional, whereas the latter two *ut* were added by mistake, probably caused by the distance between the subordinating verb *volo* 'I want' and the subordinated verb *sit* '(that) it be';

3. to mark anacolutha, i. e. sentences that show an irreconcilable discontinuity in their syntactic structure, e. g. *constat me Sanitulum ... vendere et tradere praevideo* 'I, Sanitulus, am determined ... I arrange for the selling and trading' (s2 a. o.), with two unconnected predicate verbs (*constat* and *praevideo*) that are seemingly derived from two different formulae.

For these reasons, the outcome of the *afun* ExD in the UD conversion is quite diversified, as will be shown in Section 4.

Material gaps and fragmentary passages that occur in the text are also taken into account in the LLCT2. If the POS of a missing or fragmentary token can be deduced with a high degree of probability, but without certainty about the exact lexical item, it is marked in square brackets as an artificial placeholder token and lemmatised as *missing^token*, e. g. + *Ego David filio [Propn] rogatus [--]* '+ I, David, son of [Propn], having been asked [--]' (s180), where a generic [Propn] stands for the proper noun expected in this context. If a gap of one or more words cannot be restored at all, like the last part of the above example, it is marked with an artificial token [--], which is attached either to the node on which it is most likely dependent, or, in the case of uncertain dependency, to the sentence root. This kind of restoration is possible thanks due to the highly formulaic documentary language, where it is often obvious what type of word is missing (Korkiakangas and Lassila, 2013). In our UD conversion, such *missing^token*s maintain their forms and dependencies and bear their corresponding UD morphological and syntactic analysis.

3.5. Ellipses

Another phenomenon indicated by special placeholders is ellipsis. Ellipsis is usually disambiguated in the LLCT2 by adding an artificial token, similar to the one used for gaps, with the POS of the elided element in square brackets and with the lemma *missing^token*. This is typical, for example, of comparative clauses, like s485:

> *set tibi obedire et servire debeat, sicut filius [verb] patri*
> 'but he/she must obey and serve you, like the son [obeys and serves] the father'.

While ellipsis is intentional by definition, the same kind of placeholder is also used to introduce unintentionally omitted tokens if their POS can be reliably reconstructed. For example, it is probable that the participle *regnante* 'reigning' is missing from the opening of s1524 a. o.:

> *+ In nomine Patris et Filii et Spiritus sancti [participle] Carolus serenissimus augustus*
> '+ In the name of the Father and the Son and the Holy Spirit [participle] Carolus, most serene emperor'

Here, the scribe seems to have forgotten the participial head required by the formula. As it is not clear whether the scribe meant to write exactly *regnante* 'reigning', the restoration only indicates the POS. Such constructions differ from anacolutha (see Section 3.4.) in that they do not seem to result from the overlap of two competing syntactic structures or formulae, but from a sheer lapse of concentration by the scribe. However, it is sometimes hard to draw the line between different cases, because one cannot be sure whether a construction in conflict with the standard grammar of Classical Latin is an unintentional anacoluthon, a lapsus, or an intentional, albeit non-standard syntactic choice. For example, in s31,

> *manifestu sum ego Aufuso presbitero filio quondam Gualfridi ... offero Deo et tibi ecclesia sancti Fridiani ...*
> 'I, Aufusus priest, son of the late Gualfridus, have shown up here ... I offer to God and you, the church of St. Fridianus, ...',

one should expect to find the subordinating conjunction *quia* 'because; that', which is normally placed between the predicates *manifestu sum* 'I have shown up' and *offero* 'I offer'. However, it is not obvious whether *quia* was omitted by negligence or because the scribe considered the latter predicate somehow subordinate to the previous one; in the LLCT2, predicates in such a relation are connected by the *afun* ExD.

As with anacolutha and gaps, the corresponding tokens in the UD conversion maintain the forms and lemmas as in the LLCT2 and bear the corresponding morphological and syntactic UD tags.

4. Annotation Statistics

As mentioned in Section 3., applying the UD annotation principles to the LLCT2 has a significant impact on the structures of PDT-style syntactic trees. In particular, since function words (corresponding to *afun*s starting with Aux, cf. Table 3) no longer mediate between content words (e. g. between a predicate and one of its oblique arguments), trees are tendentiously shallower in UD, as shown in Table 2, and their depth is more stable: even if the

median is the same, variance is nearly halved. This makes parsing these trees for particular syntactic constructions easier, quicker and less convoluted, especially with regard to more complex structures for which the PDT style poses a problem. This is in line with the operational objectives of the UD project, and is also confirmed by the different frequency of so-called *non-projective* constructions or *discontinuities* (Hajičová et al., 2004) in UD when confronted with the PDT style.

	Depth of trees			Non-projectivities		
	μ	M	V	μ	M	V
LLCT2-PDT	5.47	5.0	6.16	2.35	2.0	3.72
LLCT2-UD	4.74	5.0	3.27	2.16	1.0	3.52

Table 2: Means (μ), medians (M) and variances (V) of tree depths (i. e. maximum number of nodes in a path from the root to a leaf node) and non-projective (i. e. discontinuous) arrangements of nodes for the PDT and UD versions of the LLCT2.

The data for non-projective arrangements in Table 2 are computed only relative to the trees where at least one such construction appears, which amount to 2765 in the PDT version of the LLCT2, against 2582 in the UD conversion. We note here that our current UD conversion consists of 9023 sentences as compared to the 9246 of the original LLCT2, because some had to be excluded for further analysis (cf. Section 3.1.1.). In any case we observe a slight independent decrease in non-projective occurrences (a positive note for the training of syntactic parsers), which again is in line with UD objectives. Their persistence, however, can be interpreted as an inherent presence of syntactic discontinuities in Latin in general, and in the formulaic language of the charters in particular.

As discussed in Sections 3.1. and 3.4., constructions that in the original LLCT2 are mostly dealt with only with the two *afun*s `APOS` and `ExD` are interpreted in the UD by means of many more syntactic relations, expressed by a variety of *deprel*s, as seen in Table 3 (where appositions are represented by the suffix `_AP` instead of `APOS`). The *deprel* `parataxis`, used when two clauses are joined with no apparent or explicit syntactic connection, is not particularly frequent, and thus not in the Table, but it is significant for the treatment of anacolutha and ellipses (see Sections 3.4. and 3.5.). Altogether, the UD standard seems more capable of describing the heterogenous array of non-standard constructions of the LLCT2 than the PDT style. Indeed, the relative lack of representative means within the latter had the underspecification of some *afun*s as a logical consequence.

Lastly, Table 4 further highlights the greater representativeness of UD with respect to parts of speech. Notably, this conversion of the LLCT2 systematically introduces the determinants (`DET`), which in the PDT style are mostly labelled as adjectives (`a`) or pronouns (`p`), a fact which can also be inferred from the presence of the *deprel* `det` as a possible interpretation of the universal *afun* `ATR` in Table 3.

PDT-style *afun*s	UD *deprel*s
`ADV`	`obl`, `advmod`, `advcl`, `conj`, `nsubj`
`APOS`	`punct`
`ATR`	`nmod`, `det`, `appos`, `amod`, `acl`, `conj`
`ATV`	`advcl`, `xcomp`
`AtvV`	`xcomp`, `conj`
`AuxC`	`mark`
`AuxP`	`case`
`AuxV`	`aux`
`AuxY`	`cc`, `case`, `fixed`, `mark`
`AuxZ`	`advmod`
`COORD`	`cc`, `punct`
`ExD`	`root`, `fixed`, `orphan`, `reparandum`
`OBJ`	`obl`, `obj`, `conj`, `xcomp`, `ccomp`
`OCOMP`	`xcomp`, `conj`
`PRED`	`root`, `conj`, `cop`
`Pnom`	`xcomp`, `acl`, `root`, `conj`
`ROOT`	`punct`
`SBJ`	`nsubj`, `conj`
`_AP`	`dislocated`, `obj`, `cc`, `conj`
`_CO`	`conj`, `obl`

Table 3: Correspondences between *afun*s and *deprel*s: only the most frequent (> 5% of corresponding occurrences) *deprel*s are shown. *Deprel*s are ordered for relative frequency, but percentages are not shown for visual clarity.

PROIEL POS-tags	UD POS-tags
`a`	`ADJ`: 50.85 , `DET`: 43.43 , `NUM`: 5.71
`c`	`CCONJ`: 78.77 , `SCONJ`: 20.44
`d`	`ADV`: 86.12 , `PART`: 6.64 , `SCONJ`: 5.83
`e`	`INTJ`: 100
`m`	`NUM`: 100
`n`	`NOUN`: 99.93
`p`	`PRON`: 65.80 , `DET`: 33.70
`Propn`	`PROPN`: 100.0
`Punc`	`PUNCT`: 99.20
`r`	`ADP`: 100
`t`	`VERB`: 100
`v`	`VERB`: 87.56 , `AUX`: 12.44

Table 4: Correspondences between POS-tags in the PROIEL and UD styles; only the most frequent (> 5% of corresponding occurrences) UD POS-tags are shown, together with their respective relative occurrences in %.

5. Conclusion

Along with introducing a new Latin treebank annotated according to the Universal Dependencies framework, the present work focuses on the technical and linguistic challenges that are posed by an automated conversion process into UD from a different annotation standard. Besides significantly expanding the coverage of Latin in UD, the distinctive features of the Late Latin of the charters, especially with respect to "standardised" Classical

Latin, have brought forth the reassessment and revision of many aspects of the annotation strategy for Latin in general, of which some of the most relevant instances have been presented in Section 3. The consequences of these reinterpretations will hopefully expand beyond the present corpus, as our plan is to compile a set of new comprehensive guidelines for Latin annotation in UD with the ulterior aim of enriching extant and future Latin treebanks and making them more consistent.

As the conversion of the LLCT2 is a particularly complex process, we still consider it to be a work in progress, albeit in a very advanced state. The current version will need some more refinements before being officially included in the UD collection, probably as soon as in version 2.6 (scheduled for May 2020).

6. Bibliographical References

Bamman, D. and Crane, G. (2006). The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 67–78, Prague, Czech Republic. Univerzita Karlova.

Bamman, D., Crane, G., Passarotti, M., and Raynaud, S. (2007). *Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3)*. Tufts Published Scholarship. Tufts University's Digital Collections and Archives, Medford, MA, USA. Permanent URL: `http://hdl.handle.net/10427/42683`.

Barbieri, G. (1995). *Nuovo corso di lingua latina*. Lœscher Editore, Turin, Italy.

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In Lluís Màrquez et al., editors, *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York, NY, USA, June. Association for Computational Linguistics (ACL).

Cecchini, F. M., Passarotti, M., Marongiu, P., and Zeman, D. (2018). Challenges in Converting the *Index Thomisticus* Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (*UDW *2018) at* EMNLP *2018*, pages 27–36, Bruxelles, Belgium. Special Interest Group on linguistic DATa and corpus-based approaches to NLP (SIGDAT), ACL.

Cecchini, F. M., Franzini, G. H., and Passarotti, M. C. (forthcoming). Verba Bestiae: How Latin Conquered Heavy Metal. In Riitta Valijärvi, et al., editors, *Multilingual Metal: Sociocultural, Linguistic and Literary Perspectives on Heavy Metal Lyrics*, Emerald Studies in Metal Music and Culture. Emerald group publishing, Bingley, UK.

Greenough, J. B. and Allen, J. H. (2006). *Allen and Greenough's new Latin grammar*. Dover publications, Mineola, NY, USA.

Grosu, A. (2003). A Unified Theory of 'Standard' and 'Transparent' Free Relatives. *Natural Language & Linguistic Theory*, 21(2):247–331, May.

Hajič, J., Hajičová, E., Mikulová, M., and Mírovský, J. (2017). Prague Dependency Treebank. In Nancy Ide et al., editors, *Handbook of Linguistic Annotation*, pages 555–594. Springer, Dordrecht, Netherlands.

Hajičová, E., Havelka, J., Sgall, P., Veselá, K., and Zeman, D. (2004). Issues of Projectivity in the Prague Dependency Treebank. *Prague Bulletin of Mathematical Linguistics*, 81. Available at `https://ufal.mff.cuni.cz/pbml/81/hajicova-et-al.pdf`.

Hana, J. and Štěpánek, J. (2012). Prague Markup Language Framework. In Nancy Ide et al., editors, *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 12–21, Jeju, Republic of Korea, July. Association for Computational Linguistics (ACL).

Haug, D. T. T. and Jøhndal, M. (2008). Creating a parallel treebank of the old Indo-European Bible translations. In Caroline Sporleder et al., editors, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, Marrakesh, Morocco. European Language Resources Association (ELRA).

Iovino, R. (2012). *La sintassi delle espressioni nominali latine*. Ph.D. thesis, Università Ca' Foscari, Venice, Italy.

Korkiakangas, T. and Lassila, M. (2013). Abbreviations, fragmentary words, formulaic language: Treebanking medieval charter material. In Francesco Mambrini, et al., editors, *Proceedings of The Third Workshop on Annotation of Corpora for Research in the Humanities (*ACRH-3*)*, pages 61–72, Sofia, Bulgaria, December. Bulgarian Academy of Sciences.

Korkiakangas, T. and Passarotti, M. (2011). Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics (*JLCL*)*, 26(2):105–116, november. In Francesco Mambrini, et al., editors, *Annotation of Corpora for Research in the Humanities: Proceedings of the* ACRH *Workshop, Heidelberg, 5. Jan. 2012*.

Kortmann, B. (1997). *Adverbial Subordination*, volume 18 of *Empirical Approaches to Language Typology*. Mouton de Gruyter, Berlin, Germany – New York, NY, USA.

Longrée, D. (1990). À propos du concept d'« apposition » : les constructions *rex Ancus* et *urbs Roma*. *L'Information Grammaticale*, 45:8–13.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Palmer, L. R. (1988). *The Latin language*. Oklahoma University Press, Norman, OK, USA. Reprint.

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology*, volume 10 of *Age of Access? Grundfragen der Informationsgesellschaft*, pages 299–320. De Gruyter Saur, Munich, Germany, August. Open access at `https://www.degruyter.com/viewbooktoc/product/502894`.

Petrov, S., Das, D., and McDonald, R. (2011). A Universal Part-of-Speech Tagset. *ArXiv e-prints*. arXiv:1104.2086 at https://arxiv.org/abs/1104.2086.

Ponti, E. M. and Passarotti, M. (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 2016)*, pages 683–688, Portorož, Slovenia. European Language Resources Association (ELRA).

Popel, M. and Žabokrtský, Z. (2010). TectoMT: modular NLP framework. In Hrafn Loftsson, et al., editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin – Heidelberg, Germany, August. Springer.

Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., and Žabokrtský, Z. (2014). HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (*LREC *2014)*, pages 2334–2341, Reykjavík, Iceland, may. European Language Resources Association (ELRA).

Spevak, O. (2014). *The Noun Phrase in Classical Latin Prose*, volume 21 of *Amsterdam Studies in Classical Philology*. Brill, Leiden, Netherlands – Boston, MA, USA.

Sylak-Glassman, J., (2016). *The Composition and Use of the Universal Morphological Feature Schema (Uni-Morph Schema)*. Johns Hopkins University, June. Working draft, v. 2.

Väänänen, V. (2012). *Introduction au latin vulgaire*. Librairie Klincksieck – Série linguistique. Klincksieck, Paris, France. Reprint.

Vineis, E. (2005). *Il latino*. Introduzioni (Linguistica). Il Mulino, Bologna, Italy. Reprint from: Anna Giacalone Ramat et al., editors. (1993). *Le lingue indoeuropee*, pages 289–348. Il Mulino, Bologna, Italy.

Zamboni, A. (2007). *Alle origini dell'italiano*, volume 213 of *Università (Linguistica)*. Carocci, Rome, Italy, second edition.

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., , Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). HamleDT: To Parse or Not to Parse? In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (*LREC *2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L. R., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2014). HamleDT: Harmonized Multi-Language Dependency Treebank. *Language Resources and Evaluation*, 48(4):601–637.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In Nicoletta Calzolari, et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakesh, Morocco, May. European Language Resources Association (ELRA). Published as CD. Available at http://ufal.mff.cuni.cz/~zeman/publikace/2008-02/tagdrivers-marrakech-styl-lrec.pdf.

Zeman, D. (2018). *The world of tokens, tags and trees*, volume 19 of *Studies in Computational and Theoretical Linguistics*. Ústav formální a aplikované lingvistiky (ÚFAL), Prague, Czech Republic, first edition.