Hell Awaits: Building a Universal Dependencies Treebank for Dante Alighieri's *Comedy*

Claudia Corbetta*
Università di Bergamo-Pavia

Marco Passarotti†
Università Cattolica del Sacro Cuore

Flavio Massimiliano Cecchini**
KU Leuven

Giovanni Moretti†
Università Cattolica del Sacro Cuore

In this paper, we describe the creation of a treebank for Dante's Comedy in Universal Dependencies, the first syntactically annotated text for Old Italian following a dependency-based paradigm. We detail the phase of treebanking the first part of the Comedy, the Inferno, and we discuss some annotation issues, specifically ellipses and comparative structures. Then, we perform an evaluation of automated dependency parsing with models trained on the currently available annotated portion of the text.[1]

1. Introduction

Over the past two decades, there has been a growing convergence between the world of corpora for ancient languages and the scholarly community working in the area of technologies for Natural Language Processing (NLP). Because of the absence of native speakers, dealing with ancient languages means lacking the possibility of introspective analysis or field inquiries. The only empirical evidence historical linguists can engage with is confined to old texts, many of which are fortunately digitally available today. Enhancing these data sources with meta-linguistic annotation provides scholars with enriched data to support their investigations. Moreover, building annotated sets of textual data for an ancient language following *de facto* standards is a way to make these old texts compatible with several ready-made NLP tools, as well as to make them comparable with annotated corpora for other (modern) languages.

* Università degli studi di Bergamo, via Salvecchio 19, 24129 Bergamo, Italy. Università di Pavia, corso Strada Nuova 65, 27100 Pavia, Italy. E-mail: claudia.corbetta@unibg.it
** KU Leuven, Erasmushuis, Blijde-Inkomststraat 21, 3000 Leuven, Belgium. E-mail: flaviomassimiliano.cecchini@kuleuven.be
† Università Cattolica del Sacro Cuore, largo A. Gemelli 1, 20123 Milan, Italy. E-mail: {marco.passarotti,giovanni.moretti}@unicatt.it

1 This article is an extended version of a paper by the same authors, *Highway to Hell. Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy* (Corbetta et al. 2023). This paper is the result of the collaboration between the four authors. For the specific concerns of the Italian academic attribution system: Claudia Corbetta is responsible for Sections 2.1, 2.2, 4, 4.1, 4.1.1, 4.1.2, 4.2; Flavio Cecchini is responsible for Section 5; Marco Passarotti is responsible for Sections 1 and 6; Giovanni Moretti developed the script for the conversion; Giovanni Moretti and Flavio Cecchini developed the scripts for the Evaluation task. Sections 3, 4.2.1 and 4.2.2 were collaboratively written by Flavio Cecchini and Claudia Corbetta. Sections 2, 3.1 and 4.1.3 were collaboratively written by Marco Passarotti and Claudia Corbetta.

Universal Dependencies[2] (de Marneffe et al. 2021) is an annotation framework initiated in 2015 which aims to provide a universal formalism for dependency-based syntactic annotation, for facilitating cross-linguistic comparison. Currently, the project boasts 283 treebanks for 161 languages,[3] including historical languages such as Ancient Greek, Latin, Old French, Akkadian and Classical Chinese. With regard to the Italian language, Universal Dependencies includes 10 treebanks,[4] covering a diverse range of genres,[5] amounting to 879 657 tokens and 37 871 sentences. This paper details the process of developing a treebank in Universal Dependencies out of Dante's *Comedy*, starting from the annotation of the *Inferno*, the first out of the three parts (*cantiche*) of the work. The motivation for this is the current absence of any dependency-based treebank for Old Italian.[6] Besides providing the scholarly community of historical linguistics with a valuable resource, we create gold data that can be used for the supervised training and testing of stochastic NLP tools.

This paper is organised as follows: in Section 2, we introduce Old Italian and the resources available for this language, with a specific focus on the DanteSearch corpus. In Section 3, the Universal Dependencies project is introduced in general and with regard to the Italian language. In Section 4, we detail the creation of the treebank, starting from the *Inferno*. In Section 5, we describe training and evaluation of a number of models for parsing. Section 6 concludes the paper by summarizing our findings and sketching future work.

2. Old Italian

In earlier stages of linguistic research, scholars noted similarities between Old and Modern Italian.[7] This was especially evident when compared to the evolution of other Romance languages like French, where differences between old and modern varieties are more pronounced (Dardano 2013). However, numerous studies now recognise and emphasise the distinction between Old and Modern Italian (Dardano and Frenguelli 2004), particularly from a syntactic perspective (Tesi 2004).

The *Grammatica dell'italiano antico* (GIA; 'Grammar of Old Italian') (Salvi and Renzi 2010) defines Old Italian as the language spoken in Florence during the 13th century and the early 14th century. The authors of the GIA justify their choice of selecting Florentine texts (later expanded to texts from all Tuscany) on the basis of the abundant documentation of vernacular *scripta* in Florence, driven also by the diligence and productivity of the Florentine scribes. However, it should be noted that there are numerous (written) varieties that characterise Medieval Italy, albeit on a lesser scale when it comes to documentation and written evidences.

Regardless of whether Old Italian should be strictly limited to the Tuscan area or can also encompass non-Tuscan varieties, the significance and influence of Tuscan on the evolution of the Italian language is undeniable. Therefore, while choosing an Old Italian

2 https://universaldependencies.org
3 Version 2.14, released on May 15, 2024 (Zeman, Nivre, and others 2024).
4 We specify that the Italian treebanks consist of 9 Modern Italian treebanks and 1 Old Italian treebank.
5 For Modern Italian, the treebanks include texts from various genres such as legal, news, wiki, nonfiction, government legal, social, learner essays, and grammar examples. In contrast, the Old Italian treebank consists of poetry texts. Notably, no poetry texts have been included in the Modern Italian treebanks thus far.
6 Whereas, with regard to Dante Alighieri, his works in Latin are already part of Universal Dependencies, see (Cecchini et al. 2020; Passarotti et al. 2022).
7 As exemplified by a statement by (Ascoli 1882–1885, p. 124), cf. (Tomasin 2019, ch. VI).

text for a treebank in Universal Dependencies, given the importance of Florentine, it seems significant to select a Tuscan text, specifically a Florentine one, namely the *Comedy* by Dante Alighieri. He was born in Florence in 1265 and he is legitimately considered one of the greatest poets and writers of the Middle Ages. His most important work is the *Comedy*, which was written between 1308 and 1320, and is crucial to Italian literature, due to its historical (and still continuing) success among readers.

The decision of Dante to write the *Comedy* in the Florentine vernacular represents a pivotal moment in the history of Italian literature and language, as it contributed to the spread and elevation of the vernacular to a literary language (Manni 2013).

Together with the undeniable significance of the text, the availability of a digital resource, DanteSearch (Tavoni 2011), containing all of Dante's works enhanced with a number of fundamental layers of annotation further supports our decision to choose the *Comedy* as the text for the first treebank of Old Italian in Universal Dependencies.

2.1 Resources for Old Italian

There is quite a substantial amount of texts and lexical resources in digital format available for Old Italian. Among them, the *Opera del Vocabolario Italiano* corpus[8] (OVI) contains Old Italian texts dating before the 15th century and is one of the major corpora, containing 3 443 texts of Old Italian for a total of 30 176 628 word occurrences.

Closely related to the the historical dictionary of Old Italian built by OVI is the *Tesoro della Lingua Italiana delle Origini* corpus (TLIO) (Beltrami 2003), which collects 3 173 texts for a total of 23 685 634 word occurrences. Additionally, there are corpora that cover a wider temporal span, such as the *Morfologia dell'Italiano in Diacronia* corpus (MIDIA) (D'Achille and Grossmann 2017), a lemmatised and morphologically annotated collection of Italian texts from the 13th century up to the first half of the 20th century, and the *Corpus di Italiano Scritto* (CODIT) (Micheli 2022), a diachronic corpus of Italian that covers the period from the 13th century until 1947.

However, although a preliminary effort has been made towards the creation of a digital corpus of Old Italian with respect to the quotations reported in the *Grande dizionario della lingua italiana* (Favaro et al. 2022),[9] no dependency-based syntactic annotation of Old Italian texts is currently available.

2.2 DanteSearch

Among the resources available for Old Italian, DanteSearch (DS) (Tavoni 2011) is an annotated corpus containing all of Dante Alighieri's works. These include both his Latin texts, namely *De vulgari eloquentia*, *Eclogues*, *Epistles*, *Monarchia*, *Questio de Aqua et Terra*, and his vernacular texts, i. e. *Rhymes*, *Vita Nova*, *Convivio*, *Detto d'Amore* and *Comedy*. The resource has been developed at the University of Pisa and consists of a set of XML files,[10] based on the TEI[11] guidelines, providing both textual data and linguistic annotation. All of Dante's works provided by DS are tokenised, lemmatised and enhanced with grammatical annotations about parts of speech and morphosyntactic features. Moreover, *Comedy*, *Convivio*, and *Rhymes* present a clause-based syntactic annotation, which

8 http://www.ovi.cnr.it/Il-Corpus-Testuale.html
9 The work by Favaro consists in a conversion from an XML source file to the CoNLL-U format adopted by Universal Dependencies, for tokenization, lemmatization, and morphological annotation.
10 Downloadable from https://dantesearch.dantenetwork.it.
11 https://tei-c.org

distinguishes main and subordinate clauses, the latter being assigned a label for their function, such as "declarative", "temporal", and "relative" (Tavoni 2022).

Concerning the *Comedy*, the text included in DS is based on Petrocchi's edition (Alighieri 1994) and is recorded in two separate XML files: one file provides the grammatical layer of annotation, while the other contains a clause-based layer of syntactic annotation (Gigli 2015). However, the clause-based syntactic annotation used by DS is not fully compatible with other frameworks, such as the one adopted by Universal Dependencies, currently considered the *de facto* standard for dependency-based syntactically annotated corpora.

3. Universal Dependencies

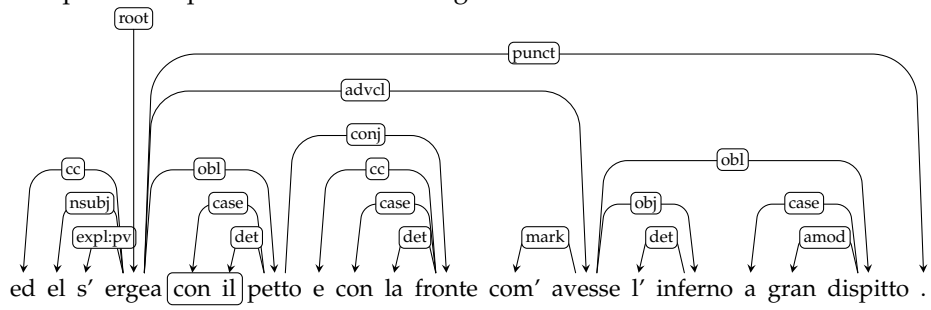
Universal Dependencies (UD) is a framework for cross-linguistic annotation of lemmas, parts of speech, morphological features and syntactic dependencies, aiming to "offer a linguistic representation that is useful for morphosyntactic research, semantic interpretation, and for practical natural language processing across different human languages" (de Marneffe et al. 2021, §2, p. 256). Being an open community, the treebanks and languages represented in UD's collection are steadily growing. Taking into account the latest release (v2.14), as mentioned in 1, UD's dataset comprises 283 treebanks representing 161 languages, encompassing both contemporary and historical languages from different language families.

In UD's framework, linguistic annotation is based on the distinction between three phrasal units (de Marneffe et al. 2021, §2.1.2): *nominals*, primarily referring to entities; *clauses*, expressing events; and *modifiers*, conveying attributes of events or entities. To annotate the relations involving these phrasal units, UD adopts a dependency grammar perspective, wherein each phrase has a head and its other elements, which can themselves be phrases, depend on that head; phrases then appear nested into a hierarchical structure (de Marneffe et al. 2021, §2.1.1, pp. 256-257). Relations take place directly between (syntactic) words,[12] with no representation of intermediate constituents. A dependency relation is "a binary asymmetrical relation" (de Marneffe et al. 2021, §2.1.1, p. 257). From a formal, mathematical point of view, the syntactic structure of a sentence is represented by means of a *syntactic tree*, i. e. an acyclic, directed, rooted graph with a linear ordering on its nodes (Havelka 2007), in which nodes correspond to syntactic words, and edges between a head and its dependents are tagged for their corresponding syntactic relations. UD's dependency-based annotation scheme is predicate-centered, with the head of the sentence's main predicate serving as the tree's root. Moreover, in UD's formalism, function words have to depend on the content words they are related to. This is not the case for all dependency-based schemes, like, for instance, for the analytical layer of annotation of the Prague Dependency Treebank for Czech (PDT), where e. g. conjunctions govern conjuncts and adpositions are the heads of adpositional phrases, i. e. noun phrases introduced by adpositions.[13]

12 In UD, a distinction between "token" and "syntactic word" is made: while "token" refers to an orthographic unit of segmentation, "syntactic word" refers to the actual level of analysis as it appears in the syntactic tree. These two levels often, but not always, coincide: for example, in Italian the token *nel* 'in the' would be analysed into the syntactic words *in* 'in' (an adposition) and *il* 'the' (a determiner), each bearing its own morphosyntactic annotation. In this sense, we note that also punctuation marks are subsumed under syntactic words, even if they do not represent lexical elements. For more details, see `https://universaldependencies.org/u/overview/tokenization.html`.

13 `https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html`

We report a sample annotated following UD's formalism[14] from our treebank:



Example 1
ed el s'ergea col petto e con la fronte / com'avesse l'inferno a gran dispitto.
'and up he rose -his forehead and his chest- / as if he had tremendous scorn for Hell.'[15]
– `Canto10-356` (*Inferno* X vv. 35–36)

The sentence, taken from *Canto* X (vv. 35-36), describes Farinata degli Uberti, a prominent figure in Florence's political history and a character in Dante's *Inferno*. Dante portrays him standing proudly from a grave where he is confined as a punishment for his heresy. Despite his condemned position, Dante depicts Farinata as a fierce man, characterised by his steady and resolute posture even within the desolate confines of Hell.

We describe the UD syntactic tree of the sentence by moving from the root of the tree to its leaves. Since dependency grammar follows a predicate-centric approach, the root of the tree (indicated with dependency relation `root`) is the verbal predicate of the main clause, which in this example is *ergea*, 'rose up'. From the root, other dependency relations branch out.

Starting from the left, we observe that the relations depending on the root include the 'coordinating conjunction' (`cc`), which connects the conjunction *ed* 'and' with the verb; the 'nominal subject' relation (`nsubj`), which identifies the subject of the sentence, expressed by the pronoun *el* 'he'; the 'expletive' relation (`expl`), which captures the relation between the reflexive clitic *si* (*s'* in the text) and the verb. The `expl` relation is marked with the specific subtype (`:pv`), used for reflexive clitics attached to inherently reflexive verbs.

Moving to the right side of the root, the dependency relations directly depending on the root include the 'oblique relation' (`obl`), the 'adverbial clause modifier' relation (`advcl`), and the 'punctuation' (`punct`).

Starting with the oblique relation, `obl` is used to express a non-core argument or adjunct depending from a verb, an adverb or an adjective. In our example, the noun *petto* 'chest' takes the oblique relation as it is a content word expressing a non-core meaning, specifically describing the manner in which Farinata rose up. Dependent on the content word *petto* are two function words: the preposition *con* 'with', marked by a case dependency relation (`case`), and the masculine article or determiner *il* 'the', marked by a determiner relation (`det`). As shown in Example 1, the text features the combined form 'col' (*con* 'with' + *il* 'the'). In this case, a single orthographic token, *col*,

14 `https://universaldependencies.org/guidelines.html`
15 The English translations of the examples from the *Comedy* are by Allen Mandelbaum, available at: `https://digitaldante.columbia.edu/dante/divine-comedy/`.

corresponds to multiple syntactic words, namely *con* and *il*. UD treats such cases as multiword tokens,[16] splitting the combined form into its component syntactic words, each of which receives its own annotation.

Dependent on the content word *petto* is another content word, *fronte* 'forehead'. This noun, along with *petto*, conveys the manner in which the character is standing. Consequently, the second noun (*fronte*) takes the 'conjunct' relation (`conj`), indicating a coordination relationship between the two elements. While coordinate structures are generally symmetrical, in UD, the first element serves as the parent of all subsequent coordinated elements. The conjunct is introduced by the coordinating conjunction *e* 'and', which depends on the second conjunct, namely *fronte*. The word *fronte* serves as the parent of two other function words: namely the preposition *con* 'with', which exhibits a case marking relation (`case`), and the feminine determiner *la* 'the', marked by a determiner relation (`det`). Unlike previous conjunct (*col petto*), where the preposition and the determiner were treated as a multiword token (*col* > *con il*), in this context they appear as distinct words.

Another relation that depends on the root of the tree is the 'adverbial clause modifier' relation (`advcl`), which signals a clause that modifies a verb (or another kind of predicate) not as a core complement but as a modifier. This relation is used for clauses expressing various meanings (e.g., temporal, causal, final, among others), including hypothetical comparatives,[17] as illustrated in the present example. Being predicate-centric, the head of the adverbial clause is a verb *avesse* 'had', on which all other words in the clause depend. The adverb *come* 'as', which introduces the adverbial clause, depends on the verb *avesse* with a marker relation (`mark`). This relation is used specifically for the marker that indicates a clause is subordinate to another.

The content words *inferno* 'Hell' and *dispitto* 'scorn' also depend on the verbal head of the adverbial clause. The former takes an 'object' relation (`obj`), serving as the second core argument of the verb (following the subject),[18] while the latter is an oblique argument, functioning as a non-core argument of the verb *avesse*. The noun *inferno* has a masculine determiner *l'* 'the' as its dependent, with the `det` relation. In contrast, the noun *dispitto* 'scorn' takes the `obl` relation and has a preposition *a* 'to' and an adjective *gran* 'tremendous' (lit. 'big') as its dependents. The preposition *a* is linked by the `case` relation, while the adjective *gran* is connected through the adjectival modifier relation (`amod`), which indicates adjectives modifying nouns or pronouns.

Finally, the last node that depends on the root of the clause is the full stop. The dependency relation (`punct`) is used for all punctuation marks.

We briefly mention the fact that UD also displays a "deeper" level of annotation, called *Enhanced Dependencies*, that aims to "[make] some of the implicit relations between words more explicit, and [augment] some of the dependency labels to facilitate the disambiguation of types of arguments and modifiers".[19] Enhanced Dependencies are not mandatory for a treebank, and they are designed to handle specific cases, such as ellipsis, i.e., cases of omitted phrases in a sentence (see 4.2.1), marking the arguments of

16 Refer to `https://universaldependencies.org/u/overview/tokenization.html`.

17 Refer to DanteSearch classification at `https://dantesearch.dantenetwork.it` and refer to (Gigli 2004).

18 It should be noted that Old Italian, like Modern Italian, is a pro-drop language, meaning that the subject can be unexpressed.

19 `https://universaldependencies.org/u/overview/enhanced-syntax.html`

passive verbs with a specific sub-relation (such as `pass` 'passive' and `agent` 'agency'), specifying the co-reference in relative clauses, among others.[20]

3.1 The Italian language in Universal Dependencies

As far as Italian is concerned, the current status of UD encompasses 10 treebanks,[21] which represent diverse genres and styles. Table 1 lists all the Modern Italian treebanks in UD v2.14, arranged by size, along with the number of syntactic words, text genre, and a reference. Among Modern Italian treebanks, ISDT, VIT, ParTUT, PoSTWITA, and PUD are the result of a conversion from a previous annotation schema, whereas ParlaMint, TWITTIRO, VALICO-UD, and MarkIT are automatically annotated from pre-existing texts/corpora and later manually corrected. The references in Table 1 provide details on these conversions and annotations of each treebanks.

Table 1
Modern Italian UD treebanks (in UD 2.14).

Treebank	Syntactic words	Genre	Reference
ISDT	298K	legal, news, wiki	(Bosco, Montemagni, and Simi 2013)
VIT	280K	news, non fiction	(Tonelli, Delmonte, and Bristot 2008)
ParTUT	55K	legal, news, wiki	(Sanguinetti and Bosco 2015)
ParlaMint	20K	government legal	(Agnoloni et al. 2022)
TWITTIRO	29K	social	(Cignarella, Bosco, and Rosso 2019)
VALICO-UD	6K	learner-essays	(Di Nuovo et al. 2022)
PoSTWITA	124K	social	(Sanguinetti et al. 2018)
MarkIT	40K	grammar-examples	(Paccosi and Palmero Aprosio 2022)
PUD	23K	news, wiki	(McDonald et al. 2013)

Prior to the release (from v2.13) of the Italian-Old treebank,[22] no Old Italian text had been integrated into UD: the inclusion of the *Comedy* by Dante Alighieri thus marked the inaugural step for this Italian variety into the UD project. We note that, even though several languages have separate repositories for the modern spoken language and any of its older varieties, as it happens for Old French with PROFITEROLE[23] (Prévost et al. 2023) versus contemporary French with, among others, GSD[24] (Guillaume, de Marneffe, and Perrier 2019), we include Old Italian among the other Italian treebanks. This decision stems foremost from the lack of a dedicated ISO code[25] (required by UD) for Old Italian, a state of affairs which relates to the ongoing debate about the continuity between older and modern varieties of Italian (cf. Section 2). For now, this leads us to subsume Italian-Old under the family of `ita` treebanks.

4. Treebanking Dante's *Comedy*: the *Inferno*

Dante's *Comedy* is composed of three parts, called *cantiche*, which are *Inferno* 'Hell', *Purgatorio* 'Purgatory' and *Paradiso* 'Heaven'. These *cantiche* are divided respectively

20 Refer to (Nivre et al. 2018) for an introduction to Enhanced Dependencies.
21 We include the Italian-Old treebank in the count of Italian treebanks.
22 https://github.com/UniversalDependencies/UD_Italian-Old
23 https://github.com/UniversalDependencies/UD_Old_French-PROFITEROLE
24 https://github.com/UniversalDependencies/UD_French-GSD
25 https://iso639-3.sil.org/about

into 34, 33 and 33 subsections called *canti*. As of now, only the annotation of *Inferno* and *Purgatorio* has been completed and is available through UD (from v2.14). This Section details the process of annotating the *Inferno* according to UD's formalism for the first release of Italian-Old.

4.1 From DanteSearch to Universal Dependencies

We perform a conversion from the grammatical XML-TEI file of the *Inferno* provided by DS, consisting of 33 416 tokens out of a total of 99 390 (excluding punctuation marks) for the whole *Comedy*, to the CoNLL-U format adopted by UD's treebanks.

Before proceeding with the description of the conversion process, we first briefly introduce the two formats.

4.1.1 DanteSearch format: XML-TEI

The XML-TEI format adopted by DS is an XML that adheres to the guidelines established by the Text Encoding Initiative (TEI), which is a consortium dedicated to the development and maintenance of standards for the representation of texts in digital form.[26] In the grammatical XML-TEI of *Inferno*, information regarding each word's lemma, grammatical category (*catg*), and form is encoded.

Below, we show the annotation of a portion of the sentence from Example 1:

Example 1
Inferno, X, v. 35
ed el s'ergea col petto
'and up he rose - his chest'

in the grammatical XML-TEI format used by DS:

```
<LM lemma="ello" catg="pp3mslso">el</LM>
<LM lemma="si" catg="pf3ypr">s'</LM>
<LM lemma="ergere" catg="vi+2iis3">ergea</LM> <LM1>
<LM lemma="con" catg="epakm">col</LM>
<LM lemma="il" catg="rdms">col</LM> </LM1>
<LM lemma="petto" catg="sm2ms">petto</LM>
```

As shown in the XML-TEI snipped, each word form is the content of the <LM> tag, its lemma (lemma=), and its grammatical category (catg=), which is encoded as an alphanumerical string.[27] In the case of multiword tokens such as *col* 'with the', the XML-TEI format splits the multiword token into its components, providing separate tokens for each word with their respective lemmas, while retaining the multiword token as the form. For example, for the multiword token *col* 'with the', XML-TEI provides two distinct single tokens: one with the lemma *con* 'with' and the other with the lemma *il* 'the'. For both tokens, the form remains 'col'.

26 For further information, see `https://tei-c.org`.
27 Refer to (Tavoni 2011) for further information.

4.1.2 UD format: CoNNL-U

CoNLL-U is a format with tab-separated values where lines contain the annotation of syntactic words into 10 fields.[28] The fields are organised as follows:

- **ID**: an integer index for the word. In the case of multiword tokens, it is represented by a range.
- **FORM**: the word form.
- **LEMMA**: the lemma of the word.
- **UPOS**: the Universal part-of-speech tag,[29] which marks the core part-of-speech categories.
- **XPOS**: a field for language-specific part-of-speech tag or morphological features.
- **FEATS**: this field contains morphological features, which can either belong to a universal feature inventory or be language-specific. The features are expressed in a 'Name=Value' format, with each feature separated by a '|'.
- **HEAD**: the ID number of the syntactic head of the current word (if the word has no head, i.e. it is the root of the syntactic tree, it takes the value 0).
- **DEPREL**: the UD relation (henceforth deprel) to the head (if the word has no head, the deprel is `root`).
- **DEPS**: a field that reports the enhanced dependency graph.[30]
- **MISC**: a field for any additional annotation.

In Table 2, we present the CoNLL-U of the same sentence provided in the XML-TEI format:

Table 2
CoNLL-U file of a portion of the sentence of *Inferno*, X, v. 35.

Id	Form	Lemma	UPOS	XPOS	Feats	Head	Deprel	Deps	Misc
1	el	ello	PRON	pp3mslso	Gender=Masc(...)	3	nsubj	_	X, 35
2	s'	si	PRON	pf3ypr	Clitic=Yes(...)	3	expl:pv	_	X, 35
3	ergea	ergere	VERB	vi+2iis3	Aspect=Imp(...)	0	root	_	X, 35
4–5	col	_	_	_	_	_	_	_	X, 35
4	con	con	ADP	_	_	6	case	_	X, 35
5	il	il	DET	_	Definite=Def(...)	6	det	_	X, 35
6	petto	petto	NOUN	sm2ms	Gender=Masc(...)	3	obl	_	X, 35

In accordance with the CoNLL-U format, in the Italian_Old treebank information is provided for all the 10 fields, except for the DEPS field, which remains empty as we do not address enhanced dependencies. The XPOS field is populated with the DS grammatical category to preserve the original annotation from DS within the CoNLL-U format. In the MISC field, we include details related to the position of the word, namely

28 See `https://universaldependencies.org/format.html`. Whenever a value of a field is either not relevant or not annotated, the value "_" is used.
29 See `https://universaldependencies.org/u/pos/index.html`.
30 See `https://universaldependencies.org/u/overview/enhanced-syntax.html`.

the number of the *Canto* and the verse. This choice is made to ensure that the position of a word in the poem is always available, as word placement is crucial for a poem in rhyme and meter. Regarding the treatment of multiword tokens, we insert a line with the corresponding range (e.g., 4-5 *col* 'with the'), followed by the normalised individual syntactic words, each with their respective information.

4.1.3 Conversion

The conversion from the DS XML-TEI file to CoNLL-U focuses on the forms, lemmas, parts of speech, and morphological features of the tokens. However, in the CoNLL-U file we do not report the syntactic annotation contained in the XML syntactic file of DS (cf. Section 2.2), due to its incompatibility with the word-based syntactic analysis in UD (de Marneffe et al. 2021, §2.2).

Table 3
Table of conversion for tags related to articles. Some tags represent mutually exclusive alternatives.

DS tags	Description	UD tags
r	article	DET + PronType=Art
+ d	definite	+ Definite=Def
+ i	indefinite	+ Definite=Ind
+ m	masculine	+ Gender=Masc
+ f	feminine	+ Gender=Fem
+ s	singular	+ Number=Sing
+ p	plural	+ Number=Plur

The conversion of grammatical tags (expressed with `catg="..."`) is performed on a 1:1 basis (DS:UD), whenever possible. Different criteria for the assignment of parts of speech and morphological tags between the two annotation styles are managed case by case. For instance, DS alternately assigns the tag for "pronouns" (p) or "adjectives" (a) to possessives such as *mio* 'my', while in UD we always tag them as "determiners" (DET). Some morphological tags of DS cannot be considered in our conversion process, since they pertain to a different level of annotation according to UD's standards. For example, in its morphological analysis DS includes information about the valency of predicates annotating transitivity on verbs. Similar tags are not included in our conversion to UD, since this information can be retrieved from the syntactic layer of annotation, and Italian does not possess specific markings for valency. The conversion is performed automatically using specific conversion tables. Table 3 shows an example of mapping of morphosyntactic tags between DS and UD in the specific case of articles, a subclass of determiners (DET). Additionally, Table 4 illustrates the conversion of the masculine singular article *il* ('the'), showing the input (i.e., the annotation of the article in DS), and its output (i.e., the annotation of the same article in UD) after conversion.

Table 4
Table of conversion for the masculine singular article *il* 'the'.

article	DS (input)	UD (output)	
		UPOS	Feats
il 'the'	rdms	DET	Definite=Def|Gender=Masc|Number=Sing|PronType=Art

As shown in Table 4, the alphanumeric string from DS is used to derive the universal PoS tag and the morphological features adopted in UD. For instance, the part of speech and the morphological features obtained from the conversion of the article include details such as category (DET, indicating 'determiner'), definiteness, gender, number, and pronominal type (Definite=Def|Gender=Masc|Number=Sing|PronType=Art).[31]

With regard to tokenization and lemmatization, in a few cases we modify the criteria followed by DS to fit the ones of UD. Specifically, this applies to the tokenization and lemmatization of what are referred to as *locuzioni* 'locutions' in DS, i. e. , sets of two or more words arranged in a fixed sequence (Serianni and Castelvecchi 1991, p. 491), such as *mentre che* 'while' and *davanti a* 'in front of'. In DS, such multiword expressions are analysed as single tokens, while the UD's annotation schema requires that the words they are composed of are considered separately and analysed individually. As a consequence, for locutions we employ distinct tokenization, lemmatization, and part-of-speech taggings in contrast to DS, as shown in Table 5 with regard to the following example:

Example 2
noi udiremo e parleremo a voi, ***mentre che*** *'l vento, come fa, ci tace*
'will please us, too, to hear and speak with you, now **while** the wind is silent, in this place'
– Canto5-175 (*Inferno* V vv. 95–96)

Table 5
Annotations for the locution *mentre che* 'while'.

	DS	UD	
no. tokens	1	2	
lemma(s)	*mentre che*	*mentre*	*che*
tag(s)	clst	ADV	SCONJ

Here, the DS tag sequence clst stands for a subordinating conjunction (cs) used in a locution (l) within a temporal clause (t), while the UD part-of-speech tags ADV and SCONJ stand respectively for 'adverb' and 'subordinating conjunction'.

Modifications of lemmatization and part-of-speech taggings are required also for multiword proper nouns, which are lemmatised under a unique lemma in DS, in contrast to what happens in UD. Table 6 exemplifies this by means of the multiword proper name *Filippo Argenti*. Here, the DS tag n stands for 'onomastics', while the UD tag PROPN stands for 'proper noun'.

Further, we also need to adjust the lemmatization of articles. In DS, there are separate lemmas *la/una* and *il/uno* for the definite/indefinite feminine and masculine articles respectively, whereas, following the convention of most UD Italian treebanks, we lemmatise both under the respective masculine forms.

31 Where the value 'Def' stands for definite, 'Masc' for masculine, 'Sing' for singular, and 'Art' for article. For more details about UD features, see https://universaldependencies.org/u/feat/index.html.

Table 6
Annotations for the personal name *Filippo Argenti*.

	DS	UD	
no. tokens	1	2	
lemma(s)	*Filippo Argenti*	*Filippo*	*Argenti*
tag(s)	n	PROPN	PROPN

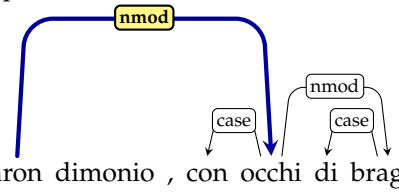
4.2 Issues of syntactic annotation

We perform the syntactic annotation of the *Inferno* manually[32] using *ConlluEditor* (Heinecke 2019) and with the support of a few critical commentaries on the work, namely those by Chiavacci Leonardi (Alighieri 2005) and Inglese (Alighieri 2007). The syntactic annotation is performed by a single annotator with expertise in Italian studies. Following UD's guidelines, annotation is carried out at sentence level; we base sentence splitting on full stops and question or exclamation marks followed by an uppercase letter, according to Petrocchi's edition of the *Comedy* recorded in DS (Alighieri 1994). The total number of sentences in the *Inferno* is 1 228, for a total of 41 367 syntactic words.

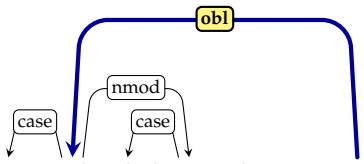
While annotating the *Inferno* according to UD's formalism, we encounter several issues that require us to take specific decisions. For instance, we refer to sentences where more than one syntactic annotation is possible, either due to linguistic ambiguity or/and to differing interpretations of the text by various editors. In such cases, we adhere to the interpretation that is supported by the majority of the editors. We provide an example of linguistic ambiguity and choice of syntactic annotation:

Example 3
Caron dimonio, con occhi di bragia loro accennando, tutte le raccoglie;
'The demon Charon, with his eyes like embers, by signaling to them, has all embark;'
– Canto3-104 (*Inferno* III vv. 109–110)

The preceding verses of *Canto* III introduce the demon Caronte, the ferryman of damned souls, who is responsible for ferrying the damned across the river Acheronte to Hell. Caronte is depicted as collecting all the souls (*tutte le raccoglie* 'has all embark', lit. 'he gathers them all') by observing them and issuing his commands (*loro accennando* 'by signaling to them'). The ambiguity of this sentence arises from the attachment of the prepositional phrase *con occhi di bragia* 'with his eyes like embers', which could either modify the noun *Caronte* (Case 1) or the verb *accennando* 'signaling' (Case 2). We report the two possible subtrees of the sentence in Example 3:



32 Annotating pre-parsed data has been ruled out after evaluating the accuracy of the UDPipe model (Straka, Hajič, and Straková 2016) trained on the largest UD treebank of Italian (ISDT) and tested on the first three *canti* of *Inferno*: its LAS score is 63,52% (see Section 5).



In Case 1, the prepositional phrase *con occhi di bragia* functions as a nominal modifier (`nmod`) and depends on the noun *Caronte*, indicating that the embers' eyes are a characteristic feature of the ferryman. In contrast, in Case 2, the prepositional phrase is interpreted as an oblique (`obl`) that depends on the verb *accennando*. In this interpretation, the embers' eyes characterise the action by which *Caronte* gestures to the souls. We have chosen to adopt the interpretation of the verse that is supported by the majority of editors (including Chiavacci Leonardi), specifically the one presented in Case 2.[33]

Another challenging task encountered in the annotation process is the treatment of two linguistic phenomena: ellipsis and comparative structures. These phenomena are extensively discussed within the UD community due to their complexity in annotation.[34] Given the high occurrence of such phenomena,[35] likely due to the poetic genre of the text, we discuss how we handle these structures specifically.

4.2.1 Ellipses

The term *ellipsis* refers to the omission of words or phrases that can be inferred from the context of a sentence or utterance.[36] Ellipsis (Merchant 2019)

> represents a situation where the usual form/meaning mappings, the algorithms, structures, rules, and constraints that in non-elliptical sentences allow us to map sounds and gestures onto their corresponding meanings, break down.

While annotating the *Inferno*, we come across several cases of ellipses, including nominal ellipses, which represent (Saab 2019, p. 526)

> different types of anaphoric phenomena involving a gap within the internal structure of the nominal phrase

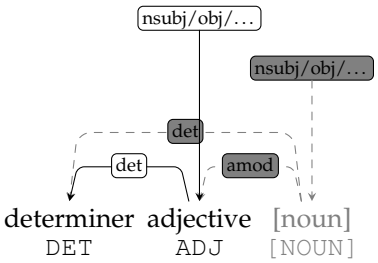
and predicate ellipses (Lobke and Harwood 2019, p. 504),

> a type of ellipsis that leaves the main predicate of the clause unpronounced, most often together with one or more of its internal arguments or (low) adjuncts.

In the matter of nominal ellipses, we follow the solution of *promotion*, as outlined in UD's guidelines.[37] Promotion involves the selection of an element inside the elliptical phrase

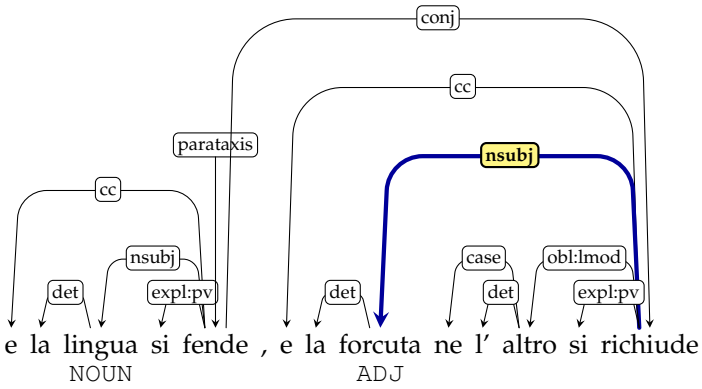
33 For an insight into the specific verse, refer to (Alighieri 2005, p. 54; 58).
34 For comparative structures, refer to `https://universaldependencies.org/workgroups/comparatives.html`, and for ellipsis, see `https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis`.
35 In *Inferno*, the number of elliptical structures annotated with the dependency relation `orphan`, which is used to annotate specific cases of ellipsis, is 156, while the number of adverbial clauses marked as comparative (`advcl:cmp`) is 272, out of 1 268 occurrences of adverbial clauses without a specific subtype, `advcl`.
36 See (Merchant 2019) for an introduction to the topic.
37 `https://universaldependencies.org/u/overview/specific-syntax.html\#ellipsis`

unit to make it take the place of the omitted element in the syntactic tree, following a specific hierarchy: `amod` 'adjectival modifier' > `nummod` 'numeric modifier' > `det` 'determiner' > `nmod` 'nominal modifier' > `case` 'case marking'. This implies that the promoted element inherits the dependency relation of the omitted head, and becomes itself the head of the other elements in the phrase, which keep their dependency relations. For instance, in the syntactic tree for a phrase of the type '*determiner* `DET` + *adjective* `ADJ` [+ *noun* `NOUN`]' with an elliptical noun, the adjective is promoted to head, assuming the dependency relation of the elided noun, and it also takes the determiner as a dependent with relation `det`.



We report an example of promotion extracted from our corpus:[38]

Example 4
e la lingua (...) si fende, e la ***forcuta*** *ne l'altro si richiude*
'his tongue (...) now cleaves; the other's tongue, which had been forked, now closes up'
– `Canto25-898` (*Inferno* XXV vv. 133–135)

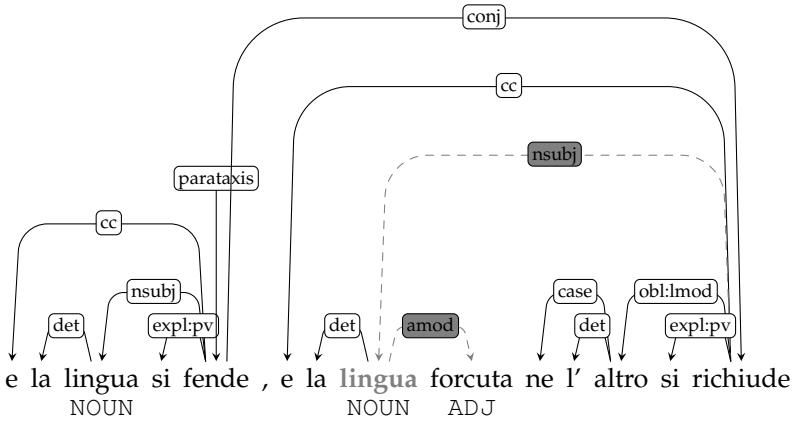


As shown in the dependency tree, here the adjective (`ADJ`) *forcuta* 'forked' depends on an omitted noun (`NOUN`) *lingua* 'tongue', as suggested by the definite article (`DET`) *la* 'the', hinting to the introduction of a new phrase and indicating that the adjective *forcuta* 'forked' refers to a different tongue from the one previously mentioned. In fact, the text describes the metamorphosis of a human into a serpent and vice versa: specifically, the first tongue, the one that 'now cleaves', belongs to the human, whereas the second one (implied by the ellipsis), 'which had been forked' and 'now closes up', is the tongue of the animal. In this case, we promote the `ADJ` *forcuta*, having the precedence in the

38 The boldfaces in the samples are added by the authors and indicate the promoted elements.

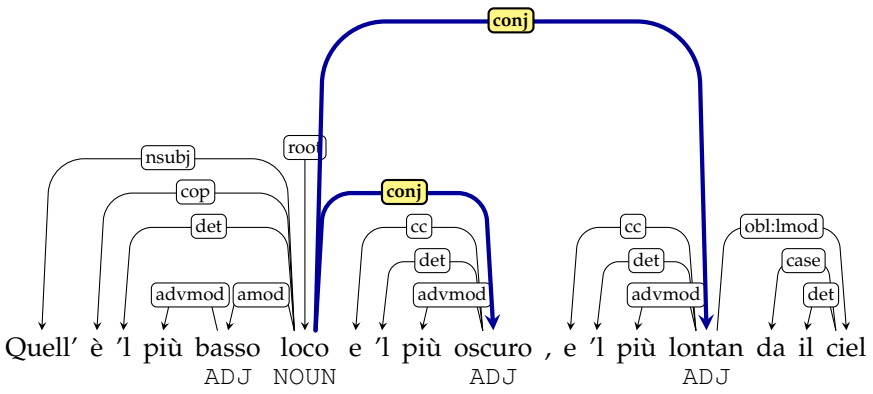
phrase with respect to *la* as a content word, to head of the nominal phrase, and assign it the dependency relation `nsubj` 'nominal subject' that would have been of the expected noun, instead of what would have been `amod` 'adjectival modifier'. The article *la* now depends on *forcuta*, but its relation `det` does not change.
For clarity, in Example 2.1, we provide the syntactic tree for the sentence as if the omitted noun *lingua* were present.

Example 4.1
e la lingua (. . .) si fende, e la **lingua** *forcuta ne l'altro si richiude*
'his tongue (. . .) now cleaves; the other's tongue, which had been forked, now closes up'
– reconstructed sentence of (*Inferno* XXV vv. 133–135)



As illustrated in the syntactic tree, if the omitted noun *lingua* (in bold in the text) had been present in the sentence, promotion would not have occurred. In this case, the (bold) noun *lingua* would assume the `nsubj` relation, while the adjective *forcuta* would function as an `amod` modifying the noun *lingua*.
Another example of promotion comes from:

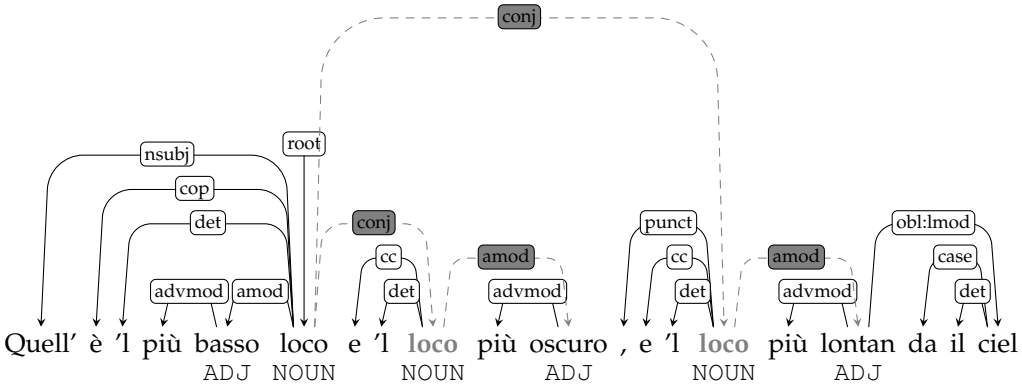
Example 5
Quell'è 'l più basso loco e 'l più ***oscuro****, e 'l più* ***lontan*** *dal ciel*
'That is the deepest place and the darkest place, the farthest from the heaven'
– `Canto9-316` (*Inferno* IX vv. 28–29)



Similarly as for the previous example, here the adjectives (ADJ) *oscuro* 'dark' and *lontan* 'far' depend on the omitted NOUN *loco* 'place', as shown by the repetition of the DET *'l* 'the' which defines the noun. In this case as well, we promote the content words ADJ *oscuro* and *lontan* to heads of their respective coordinate phrases using the pseudo-dependency relation conj 'conjunct' (cf. Section 3). We cannot consider *oscuro* and *lontan* as two conjuncts of the first adjective *basso*.

We report in Example 3.1 the reconstructed syntactic tree without the nominal ellipsis.

Example 5.1
Quell'è 'l più basso loco e 'l **loco** *più oscuro, e 'l* **loco** *più lontan dal ciel*
'That is the deepest place and the darkest place, the farthest from the heaven'
– reconstructed sentence of (*Inferno* IX vv. 28–29)



Not omitting the elliptical nouns *loco* 'place' (in bold in the text) would result in the establishment of two amod syntactic relations between the adjectives *oscuro* 'dark' and *lontan* 'far', and their respective (elided) nouns. Moreover, the two conjunctions *e* 'and' and the two articles *'l* 'the' would depend on the elided nouns *loco* rather than being attached to the promoted adjective, as in the case in the original sentence.

Adopting the promotion mechanism without at the same time employing enhanced dependencies (see Section 3), and specifically without expressing the omitted element by means of an empty (or null) node, leads to a loss of information regarding elliptical phenomena. Even though the syntactic relation of the head is passed on to the promoted element, there is no explicit indication that an ellipsis is involved.[39] Hence, the retrieval of instances of nominal ellipses requires considerable effort, as it calls for the reconstruction of the expected underlying "standard" structure from which the elliptical one is derived.

This process involves crossing morphosyntactic information to identify the elliptical element(s). For instance, examining a discrepancy between the part of speech of the promoted node and its dependency relation can serve as an effective strategy. Specifically, when the promoted node displays a dependency relation that is "unconventional" for its corresponding part of speech, cf. (de Marneffe et al. 2021, §2.1.2); e. g. when an ADJ has a nominal dependency relation such as nsubj which should be fulfilled by a

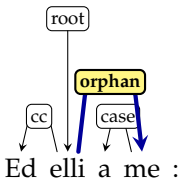
39 The complexity of dealing with ellipsis has also been analysed for user-generated texts from the web and social media, see (Sanguinetti et al. 2020).

nominal phrase unit (i. e. a noun NOUN/PROPN or a pronoun PRON), this may suggest the presence of an ellipsis. This is exactly the case in our first example, where the ADJ *forcuta* 'forked' (a modifier) has to take the nsubj dependency relation.

Nevertheless, this mechanism is not always reliable, as there may be cases, as demonstrated by our second example, where the omitted element cannot be directly inferred from a mismatch between part of speech and dependency relation. In fact, in this case the dependency relation of the promoted node, conj 'conjunct', is compatible both with the part of speech of the promoted node, ADJ, and the one of the elliptical element, NOUN (it is actually compatible with any part of speech by its own nature). Similar cases require more complex, generalised queries, over entire subtrees and not only single relations, to be extracted, such as searching for coordinations where the part of speech of the first conjunct is of a different category than that of the second conjunct.

When dealing with predicate ellipsis, we employ the dependency relation orphan, as required by UD's guidelines. The orphan relation is used in cases where the promotion mechanism would yield a misleading dependency relation, particularly in instances where the elided element serves as the head of the predicate of the sentence. Indeed, in scenarios where e. g. a VERB is omitted, the promotion mechanism might select an argument to be promoted to head of the sentence, thus resulting in an unnatural attachments with respect to the other predicate's arguments. For instance, we might encounter cases where a nominal subject is promoted to head, with another nominal serving now as its object with relation obj: but the head of a nominal phrase cannot have an object as its dependent, as it does not have an argument structure.[40] An example of the use of the orphan relation is provided below:

Example 6
Ed elli a me:
'And he (said) to me:'
– Canto3-91 (*Inferno* III v. 76)

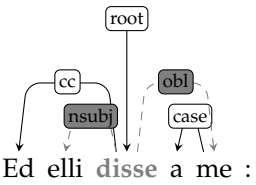


where the predicate of the sentence, namely the *verbum dicendi* (i. e. a speech verb, such as *tell*, *say*, *answer*), is omitted. This structure is extremely common to introduce a reported speech.[41] As shown in Example 4, the omission of the predicate requires the promotion of *elli* 'he' to root of the tree (root), and the annotation of the phrase *a me* 'to me' as an orphan relation.
As with the previous cases of ellipsis, we provide in Example 4.1 a reconstructed version of the sentence without ellipsis.

Example 6.1
Ed elli **disse** *a me:*
'And he said to me:'

40 At least not in the same way as a verb.
41 We refer to (Cofano 2003) for an insight of the 'silence' in the *Comedy*.

– reconstructed sentence of (*Inferno* III v. 76)



The presence of the omitted predicate *disse* 'said' gives rise to two distinct syntactic relations: the subject relation `nsubj`), carried out by *elli* 'he', and the oblique relation (`obl`), by *me* 'me'.

If no enhanced dependencies (here in particular the use of empty nodes) are involved, adopting the `orphan` relation leads to the opacification of the actual syntactic structure, as the underlying `nsubj` and `obl` relations are replaced and thus obscured. Therefore, while, on the one hand, the relation `orphan`, if present, facilitates the easy retrieval of elliptical structures, on the other hand it also leads to deficient structures, that can only be unraveled through enhanced dependencies. We note that ellipsis notoriously represents a challenge for most dependency-based annotation formalisms, and thus also in UD's framework: looking at other treebanks, we find cases similar to our previous examples, as when syntactic annotation and/or conversions between different formalisms might fail to recognise ellipsis as such, especially in the case of nominal ellipsis, cf. the discussion in (Cecchini et al. 2018, §2.2.2) for Medieval Latin, or when the absence of context makes syntactic interpretation ambiguous and unclear, as e. g. discussed in (Sanguinetti et al. 2023, §5.1) for user-generated data. Further, in Section 4.2.2 about comparative clauses we display constructions where ellipsis appears systematically, possibly influencing annotation choices at the syntactic level.

In order to tackle the challenge of extracting ellipses from a treebank annotated with basic dependencies, we propose the introduction of a specific subtype, say `ellp`, in UD's schema, specifically for ellipses treated via promotion. This would allow for the rapid and accurate identification of instances of ellipsis, which are not explicitly captured by the current annotation strategies and which might otherwise be overlooked. It would be similar in spirit to other "accessory" subtypes highlighting particularly marked syntactic structures, such as `outer`[42] or (for Latin) `abs` for `advcl`, 'ablativus absolutus'.[43] This "transversal" subtype would also streamline the enrichment of treebanks with enhanced dependencies, which remain the most complete approach to handling such cases.

4.2.2 Comparative clauses

In the *Inferno*, we find a multifarious typology of comparative clauses, ranging from full-fledged sentences which can be even longer than their matrix clauses, to others extremely reduced to just a few words and possibly elliptic. In the light of this and of the long-lasting discussion on the treatment of comparative clauses in UD,[44] we annotate such comparative constructions uniformly by labeling their heads with the

42 `https://universaldependencies.org/u/dep/nsubj-outer.html`
43 `https://universaldependencies.org/la/dep/advcl-abs.html`
44 Cf. the output of the work group on comparatives:
`https://universaldependencies.org/workgroups/comparatives.html`.

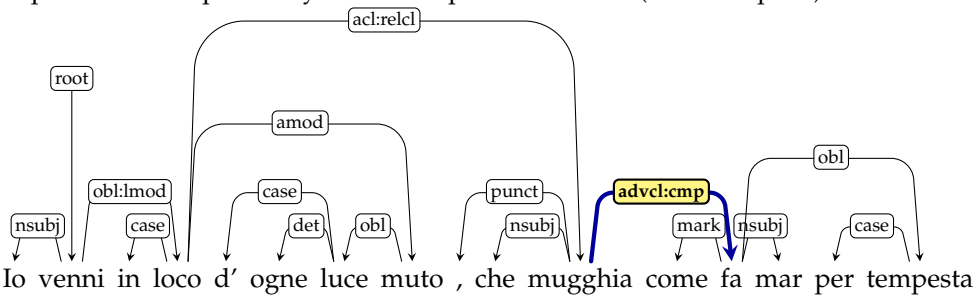
clausal dependency relation `advcl` 'adverbial clause modifier', specified for the subtype `cmp` 'comparative'.[45]

We report an example of comparative clause:

Example 7
Io venni in loco d'ogne luce muto, / che mugghia come fa mar per tempesta
'I reached a place where every light is muted, / which bellows like the sea beneath a tempest,'
– `Canto5-161`(*Inferno* V v. 28-29)

The verses come from *Canto* V, where the souls of the lustful are punished for their sins. This is the famous *Canto* of the lovers Paolo and Francesca. Dante describes the place as devoid of light, *d'ogne luce muto* 'where every light is muted', using synesthesia, in which the sense of vision (*luci* 'light') is unusually associated with the sense of sound (*muto* 'muted'), setting the stage for the forthcoming comparative description based on hearing. In fact, the noise of the place is immediately compared to a stormy sea through the comparative structure *come fa mar per tempesta* 'like the sea beneath a tempest' (lit. 'like the sea does in a storm').

We provide the dependency tree of the previous verses (see Example 7):



The head of the comparative clause *come fa mar per tempesta* 'like the sea beneath a tempest' is the verb *fa* '(it) make(s)', which depends on the verb of the relative clause (`acl:relcl`) *mugghia* 'bellows' as `advcl`, functioning as an adverbial clause modifier with the specific subtype, `:cmp`, highlighting its comparative function. The other words in the clause depend on the head of the comparative, namely *come* 'like' as the marker (`mark`), *mar* ('sea') as the subject (`nsubj`), and *tempesta* ('storm') as the oblique (`obl`), with the adposition *per* ('for') depending on *tempesta* as a case marker (`case`).

Comparative structures may also appear with an elided verb. The boundary between a comparative elliptical clause and a nominal phrase used comparatively is not always easy to define, and this issue has also been discussed in the UD annotation guidelines.[46]

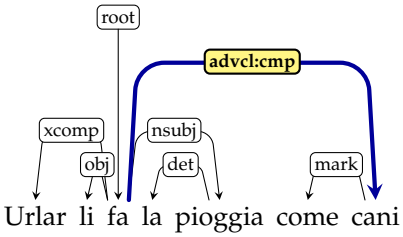
We decide to annotate also nominal phrases used as comparatives as `advcl:cmp`. This decision, was made to: i) maintain uniformity in the annotation of comparative structures, which are pervasive in a poetic text like the *Comedy*, and ii) avoid losing comparative nominal phrases among all other nominal phrases. In fact, in UD v2.14, the subtype for comparative structures for Italian, `:cmp`, is only accepted for adverbial clauses, not for nominal phrases. Therefore, to ensure that comparative information is

45 Cf. documentation at `https://universaldependencies.org/la/dep/advcl-cmp.html` (for Latin, which is currently the most exhaustive one in absence of a universal documentation page).
46 Refer to `https://universaldependencies.org/workgroups/comparatives.html`.

preserved, we treat such cases as adverbial clauses with an elided verb, marking them as `advcl:cmp`.

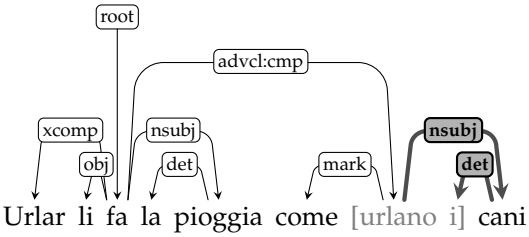
Example 8 illustrates a comparative nominal phrase.

Example 8
Urlar li fa la pioggia ***come cani***
'That downpour makes the sinners howl **like dogs**'
– `Canto6-199` (*Inferno* VI v. 19)



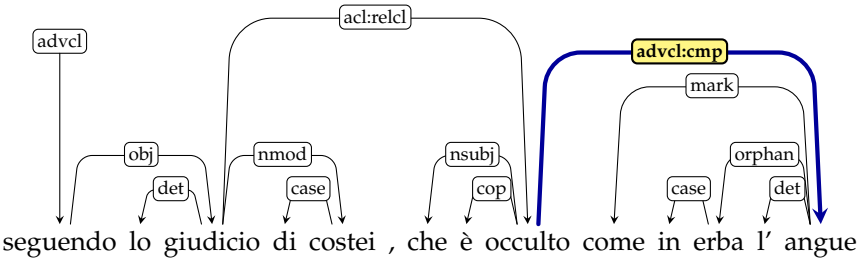
We consider *come cani* as a comparative clause with an elliptical predicate, namely:

Example 8.1
Urlar li fa la pioggia ***come [urlano i] cani***
'That downpour makes the sinners howl **like dogs [howl]**'



The decision to consider *come cani* 'like dogs' as an elliptical comparative clause is made to assimilate similar structures to cases like the one reported below:

Example 9
seguendo lo giudicio di costei, che è occulto ***come in erba l' angue***
'obeying the decision she has given, which, **like a serpent in the grass**, is hidden'
– `Canto7-249` (*Inferno* VII vv. 83–84)



Here the ellipsis of the predicate *è occulto* 'is hidden' (see 4.2.1) is signalised by the `orphan` relation of the nominal phrase introduced by a preposition *in erba* 'in the grass',

indicating the place where *l'angue* 'the serpent' is hidden (*è occulto*). We analyse the oblique phrase *in erba* 'in the grass' as a direct dependent of the omitted predicate 'is hidden', rather than as a nominal modifier (nmod) of the noun *angue* 'serpent'. In fact, we believe that the *lectio facilior* is the one presenting the locative argument as dependent on the predicate and this is also supported by the fact that this verse is a Virgilian reference (Vergil, *Bucolics* III 93: *latet anguis in herba* 'the snake lurks in the grass'), as cited by various commentators, among which Chimenz (Alighieri 1962).

For the upcoming release, we will propose and discuss with the Italian UD community the introduction of subtype :cmp for nominal phrases in Italian language as well, in order to harmonise the annotation of such constructions and avoid losing information about the comparative.

5. Evaluation

We use the manually annotated *Inferno* to train models with UDPipe 1[47] (Straka and Straková 2017) and to assess their performances in view of employing them for parsing the *Paradiso*, so as to facilitate its subsequent manual annotation.[48] In our evaluation framework, we employ a cross-validation based on 10%/90% splits of the data: each test set consists of approximately 4 137 out of 41 367 tokens and 123 out of 1 228 sentences, while train sets of approximately 37 230 tokens and 1 105 sentences. The evaluation of the models' accuracies is performed by measuring Labeled Attachment Score (LAS), i. e. the ratio of tokens "scored" for which the system correctly predicted the head and the dependency label, and Unlabeled Attachment Score (UAS), i. e. assessing whether the output has the correct head (Buchholz and Marsi 2006).

The training and evaluation process is based on one eleven- and one tenfold partition of the data, for a total of 11+10 iterations: the first partition patterns upon the original division into *canti*, with batches of 3 consecutive *canti*[49] assigned to the test set and the remaining 31[50] forming the training set (cf. Table 7); the second partition is obtained by a fully random selection of sentences.[51]

Moreover, evaluation is carried out according to two scenarios: one (+Morph) in which lemmas, parts of speech and morphological features are given, and one (−Morph) in which every annotation level has to be tagged from scratch.[52]

The accuracy of each model is calculated using eval.py,[53] an evaluation script provided by the UD project. As shown in Table 8, evaluations conducted on the random partition result into slightly higher average accuracy scores than those based on triplets[54] of consecutive *canti*: in the +Morph scenario, a difference of 0,16% is observed for UAS, whereas in the opposite −Morph scenario the improvement is more marked,

47 https://github.com/ufal/udpipe

48 We acknowledge that doing tests within a single *cantica* may not guarantee the same performances when compared to other *cantiche*.

49 We actually note that, since the number of *canti*, 34, is not divisible by 3, one *canto* would be left out, and is instead aggregated to the last batch, which then consists of 4 consecutive *canti* (31, 32, 33, 34).

50 Or 30; see fn. 49.

51 Refer to the GitHub page https://github.com/ClaudiaCorbe/Inferno_treebank for the data and detailed statistics on the partitions.

52 Corresponding respectively to -parse and -tag -parse options for UDPipe; see https://ufal.mff.cuni.cz/udpipe/1/users-manual, §3.6.

53 https://github.com/UniversalDependencies/tools/blob/master/eval.py

54 Or a quadruplet; see fn. 49.

Table 7
Statistics for the partition of the dataset into blocks of 3 (or 4) *canti*, with absolute values and percentage on the total for the respective category.

Split	Tokens	Syntactic words	Sentences
1-2-3	3 479 (0,086)	3 561 (0,086)	110 (0,090)
4-5-6	3 378 (0,084)	3 460 (0,084)	120 (0,098)
7-8-9	3 392 (0,084)	3 469 (0,084)	114 (0,093)
10-11-12	3 389 (0,084)	3 477 (0,084)	102 (0,083)
13-14-15	3 575 (0,089)	3 664 (0,089)	110 (0,090)
16-17-18	3 446 (0,085)	3 525 (0,085)	106 (0,086)
19-20-21	3 444 (0,085)	3 522 (0,085)	115 (0,094)
22-23-24	3 883 (0,096)	3 964 (0,096)	115 (0,094)
25-26-27	3 664 (0,091)	3 754 (0,091)	94 (0,077)
28-29-30	3 678 (0,091)	3 781 (0,091)	94 (0,077)
31-32-33-34	5 058 (0,125)	5 190 (0,125)	148 (0,121)

Table 8
Averages and standard deviations of accuracy metrics.

Partition	**Scenario**	**Avg. UAS**	**Avg. LAS**
random	`+Morph`	81,95±0,94%	77,07±1,00%
consecutive	`+Morph`	81,79±1,38%	77,09±1,34%
random	`-Morph`	75,32±0,91%	67,97±0,80%
consecutive	`-Morph`	74,90±1,37%	67,71±1,17%

but still minor, at 0,42% for UAS and 0,26% for LAS. The only exception regards LAS in the `+Morph` scenario, though the difference of 0,02% encountered there is negligible. Consistently with our expectations, we also observe that parsing performed with prior assignment of the other annotation levels produces better results compared to the case where the parser has to handle all annotation levels simultaneously. Specifically, in the `+Morph` scenario the average of models trained on the random partition exhibits an improvement of 6,63% for UAS and 9,10% for LAS, and, similarly, models trained on consecutive *canti* show an improvement of 6,89% for UAS and 9,38% for LAS.

We can conclude that, on the one hand, sampling the dataset randomly or by selecting consecutive parts of the text does not seem to significantly affect performances, and this could point to the fact that, at least in this *cantica*, morphosyntactic phenomena are uniformly distributed across the text, as also standard deviation is very small. On the other hand, LAS and UAS metrics improve significantly when the text is already enriched with linguistic annotation. This allows us to have positive expectations with regard to the parsing of the *Paradiso*, a *cantica* for which lemmatization and morphosyntactic taggings are inherited from the conversion from DS.

6. Conclusions and future perspectives

Building a treebank in UD for Dante's *Comedy* is the first step towards incorporating Old Italian among the languages of UD. This paper describes the development of the first part of this treebank, which consists of the first *cantica* of the *Comedy*, the *Inferno*.

We also present the results of an experiment of supervised automated dependency parsing using data from the *Inferno* both as training and test sets. We run this experiment to understand to what extent the process of syntactic annotation of the *Comedy*, which has been performed so far fully manually, can benefit from the results of the application of an NLP tool. Although the accuracy rates reported in the paper are fairly good (≈77% LAS), in the near future we will have to evaluate how and to what extent they will drop once a model trained and evaluated on the *Inferno* is applied to a different *cantica*. Should the accuracy rates drop heavily, even such a negative result might prove helpful in pointing out syntactic differences among the three *cantiche*. Moreover, the use of other parsers, based on different algorithms and resources (like embeddings), might lead to better and, most importantly, diverging results and errors.

As for annotation issues, we suggest to introduce a specific subtype, which we defined as `ellp`, in UD's documentation, so as to properly and more readily identify cases of ellipses, as they are not explicitly captured by the current standard annotation strategies, namely promotion and the use of the relation `orphan`: the former does not signal the presence of ellipsis, while the latter obscures the real dependency relations which are replaced by it. While adopting a subtype like `ellp` would make it possible to collect cases of ellipses, their resolution remains up to the implementation of so-called enhanced dependencies, which is a further, deeper, partly independent annotation layer enriching the text with information beyond the shallow morphosyntactic level, such as coreference or underlying structures.

We plan to use trained models to pre-parse the text in order to expedite the work by manually checking the pre-parsed annotation of *Paradiso.* Additionally, we intend to apply error detection processes, like those described in (Dickinson 2005), to retrieve possible mistakes or inconsistencies in syntactic annotation. Finally, we plan to augment the treebank of Dante's *Comedy* with enhanced dependencies, once the basic syntactic annotation of the entire work will be completed.

References

Agnoloni, Tommaso, Roberto Bartolini, Francesca Frontini, Simonetta Montemagni, Carlo Marchetti, Valeria Quochi, Manuela Ruisi, and Giulia Venturi. 2022. Making Italian parliamentary records machine-actionable: The construction of the ParlaMint-IT corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 117–124, Marseille, France, June.

Alighieri, Dante. 1962. *La Divina Commedia*. UTET Torino. Siro Amedeo Chimenz, editor.

Alighieri, Dante. 1994. *La Commedia secondo l'antica vulgata voll.* i–iv. Number 7 in Edizione nazionale delle Opere di Dante Alighieri a cura della Società Dantesca Italiana. Le Lettere, Florence, Italy. Giorgio Petrocchi, editor.

Alighieri, Dante. 2005. *Inferno*. Number 613 in Oscar classici. Arnoldo Mondadori, Milan, Italy. Anna Maria Chiavacci Leonardi, editor.

Alighieri, Dante. 2007. *Commedia. Inferno*. Number 1 in Opere. Carocci, Rome, Italy. Guglielmo Inglese, editor.

Ascoli, Graziadio Isaia. 1882–1885. L'Italia dialettale. *Archivio glottologico italiano*, VIII:98–128.

Beltrami, Pietro Giovanni. 2003. Il Tesoro della Lingua Italiana delle Origini (TLIO). In Nicoletta Maraschio, Teresa Poggi Salani, Marina Bongi, and Maria Palmerini, editors, *Italia linguistica anno mille. Italia linguistica anno duemila. Atti del XXXIV Congresso Internazionale di Studi della Società di Linguistica Italiana, Firenze 19–21 ottobre 2000*, number 45 in Società di linguistica

italiana. Bulzoni, Rome, Italy, pages 695–698.
Bosco, Cristina, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper, editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August. Association for Computational Linguistics.
Buchholz, Sabine and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In Lluís Màrquez and Dan Klein, editors, *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, USA, June. Association for Computational Linguistics (ACL).
Cecchini, Flavio Massimiliano, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, editors, *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium, November. The Association for Computational Linguistics (ACL).
Cecchini, Flavio Massimiliano, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020, Bologna, Italy, March 1–3 2021)*, pages 99–105, Online. Associazione italiana di linguistica computazionale (AILC), Accademia University Press.
Cignarella, Alessandra Teresa, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRO-UD: An Italian Twitter treebank in universal dependencies. In *Proceedings of the fifth international conference on dependency linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France, August. ACL.
Cofano, Domenico. 2003. *La retorica del silenzio nella Divina Commedia*. Palomar.
Corbetta, Claudia, Marco Passarotti, Flavio Massimiliano Cecchini, and Giovanni Moretti. 2023. Highway to Hell. Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy. In Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini, and Nicole Novielli, editors, *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, Venice, Italy, November-December. CEUR Workshop Proceedings (ceur-whs.org).
Dardano, Maurizio, editor. 2013. *Sintassi dell'italiano antico*. Lingue e Letterature Carocci. Carocci, Rome, Italy.
Dardano, Maurizio and Gianluca Frenguelli, editors. 2004. *SintAnt. La sintassi dell'italiano antico*. Aracne, Rome, Italy.
de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, July.
Di Nuovo, Elisa, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. 2022. VALICO-UD: Treebanking an Italian Learner Corpus in Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 8(1).
Dickinson, Markus. 2005. *Error Detection and Correction in Annotated Corpora*. Ph.D. thesis, The Ohio State University.
D'Achille, Paolo and Maria Grossmann, editors. 2017. *Per la storia della formazione delle parole in italiano*. Quaderni della Rassegna. Franco Cesati, Florence, Italy, eighth edition.
Favaro, Manuel, Elisa Guadagnini, Eva Sassolini, Marco Biffi, and Simonetta Montemagni. 2022. Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (*lt4hala*)*, pages 94–100, Marseille, France, June. European Language Resources Association (ELRA).
Gigli, Sara. 2004. *Codifica sintattica della Commedia dantesca*. Ph.D. thesis, Università di Pisa.
Gigli, Sara. 2015. La codifica sintattica della Commedia di Dante. In Marta D'Amico, editor, *Sintassi dell'italiano antico e sintassi di Dante. Atti del seminario di studi (Pisa 15/16 ottobre 2011)*. Felici, Ghezzano (PI), Italy, pages 81–95.
Guillaume, Bruno, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies [Conversion and Improvement of Universal Dependencies French corpora]. *Traitement automatique des langues*, 60(2):71–95.
Havelka, Jiří. 2007. *Mathematical Properties of Dependency Trees and their Application to Natural Language Syntax*. Ph.D. thesis, Univerzita Karlova – Matematicko-fyzikální fakulta, Prague,

Czech Republic, June.
Heinecke, Johannes. 2019. ConlluEditor: a fully graphical editor for Universal Dependencies treebank files. In Alexandre Rademaker and Francis Tyers, editors, *Proceedings of the Third Workshop on Universal Dependencies (*udw*, SyntaxFest 2019)*, pages 87–93, Paris, France, August. Association for Computational Linguistics (ACL).
Lobke, Aelbrecht and William Harwood. 2019. Predicate Ellipsis. In Jeroen van Craenenbroek and Tanja Temmerman, editors, *The Oxford Handbook of Ellipsis*, Oxford Handbooks. Oxford University Press, Oxford, UK, chapter 21, pages 504–525.
Manni, Paola. 2013. *La lingua di Dante*. Le vie della civiltà. il Mulino, Bologna, Italy.
McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August.
Merchant, Jason. 2019. Ellipsis: A survey of analytical approaches. In Jeroen van Craenenbroek and Tanja Temmerman, editors, *The Oxford Handbook of Ellipsis*, Oxford Handbooks. Oxford University Press, Oxford, UK, chapter 2.
Micheli, Maria Silvia. 2022. CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field. *Corpus*, 23.
Nivre, Joakim, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018. Enhancing Universal Dependency treebanks: A case study. In Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, editors, *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 102–107, Brussels, Belgium, November. Association for Computational Linguistics.
Paccosi, Teresa and Alessio Palmero Aprosio. 2022. It Is MarkIT That Is New: An Italian Treebank of Marked Constructions. In *Proceedings of CLiC-it 2021 Italian Conference on Computational Linguistics*, Milan, Italy, June.
Passarotti, Marco, Flavio Massimiliano Cecchini, Rachele Sprugnoli, and Giovanni Moretti. 2022. UDante. *Studi Danteschi*, LXXXVI:309–338.
Prévost, Sophie, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev, and Serge Heiden. 2023. Profiterole: un corpus morpho-syntaxique et syntaxique de français médiéval. *Corpus*, (25).
Saab, Andrés. 2019. Nominal Ellipsis. In Jeroen van Craenenbroek and Tanja Temmerman, editors, *The Oxford Handbook of Ellipsis*, Oxford Handbooks. Oxford University Press, Oxford, UK, chapter 22, pages 526–561.
Salvi, Giampaolo and Lorenzo Renzi, editors. 2010. *Grammatica dell'Italiano Antico*. il Mulino, Bologna, Italy.
Sanguinetti, Manuela and Cristina Bosco. 2015. PartTUT: The Turin University Parallel Treebank. In Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi, editors, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, volume 589 of *Studies in Computational Intelligence*. Springer, Cham (ZG), Switzerland, pages 51–69.
Sanguinetti, Manuela, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2023. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, 57(2):493–544, June.
Sanguinetti, Manuela, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May.
Sanguinetti, Manuela, Lauren Cassidy, Cristina Bosco, Özlem Çetinoglu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking User-Generated Content: a UD Based Overview of Guidelines, Corpora and Unified Recommendations. *CoRR*, abs/2011.02063.
Serianni, Luca and Alberto Castelvecchi. 1991. *Grammatica Italiana*. Universitaria. UTET Università, Turin, Italy, second edition.
Straka, Milan, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asunción Moreno, Jan Odijk Odijk, and Stelios

Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Straka, Milan and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Jan Hajič and Dan Zeman, editors, *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, BC, Canada, August. Association for Computational Linguistics (ACL).

Tavoni, Mirko. 2011. DanteSearch: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica. In Anna Cerbo, Roberto Mondola, Aleksandra Žabjek, and Ciro Di Fiore, editors, *Lectura Dantis 2002-2009. Omaggio a Vincenzo Placella per i suoi settanta anni*, volume 2 (2004–2005). Il Torcoliere – Officine Grafico-Editoriali di Ateneo, Naples, Italy, pages 583–608.

Tavoni, Mirko. 2022. Allestimento, fruizione e prospettive di DanteSearch. In Emanuela Cresti and Massimo Moneglia, editors, *Corpora e Studi Linguistici. Atti del* liv *Congresso della Società di Linguistica Italiana (Online, 8-10 settembre 2021)*, number 6 in nuova serie. Officinaventuno, Milan, Italy, pages 255–273.

Tesi, Riccardo. 2004. Parametri sintattici per la definizione di "Italiano antico". In Maurizio Dardano and Gianluca Frenguelli, editors, *SintAnt. La sintassi dell'italiano antico*. Aracne, Rome, Italy, pages 425–444.

Tomasin, Lorenzo. 2019. *Il caos e l'ordine*. Piccola Biblioteca Einaudi. Giulio Einaudi, Turin, Italy.

Tonelli, Sara, Rodolfo Delmonte, and Antonella Bristot. 2008. Enriching the Venice Italian Treebank with Dependency and Grammatical Relations. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Zeman, Daniel, Joakim Nivre, et al. 2024. Universal Dependencies 2.14. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.