Highway to Hell. Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy

Claudia Corbetta[1,2,*], Marco Passarotti[3], Flavio Massimiliano Cecchini[3] and Giovanni Moretti[3]

[1]Università degli studi di Bergamo, via Salvecchio 19, 24129 Bergamo, Italy
[2]Università di Pavia, corso Strada Nuova 65, 27100 Pavia, Italy
[3]Università Cattolica del Sacro Cuore, largo A. Gemelli 1, 20123 Milan, Italy

Abstract

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

Keywords

Dante Alighieri, Old Italian, treebank, Universal Dependencies

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 and newly written texts, 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.

Universal Dependencies[1] (UD) [1] is an annotation framework started in 2015 which aims to provide a universal formalism for dependency-based syntactic annotation, with the goal of facilitating cross-linguistic comparison. Currently, the project boasts 245 treebanks for 141 languages,[2] including historical languages such as Ancient Greek, Latin, Old French, Akkadian and Classical Chinese. With regard to the Italian language, there are 9 UD treebanks, covering a diverse range of genres,[3] amounting to 879 657 tokens and 37 871 sentences.

This paper details the process of developing a UD treebank 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.[4] 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 organized 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, we describe the creation of the treebank, starting from the *Inferno*. In Section 4, we describe training and evaluation of a number of models for parsing. Section 5 concludes the paper by summarizing our findings and sketching future work.

2. Old Italian

Although in earlier stages of linguistic research there were claims of similarity between Old Italian, particu-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
claudia.corbetta@unibg.it (C. Corbetta); marco.passarotti@unicatt.it (M. Passarotti); flavio.cecchini@unicatt.it (F. M. Cecchini); giovanni.moretti@unicatt.it (G. Moretti)
0009-0000-7425-196X (C. Corbetta); 0000-0002-9806-7187 (M. Passarotti); 0000-0001-9029-1822 (F. M. Cecchini); 0000-0001-7188-8172 (G. Moretti)

[1]https://universaldependencies.org.

[2]UD version 2.12, May 2023 [2].

[3]Including "legal, news, wiki, nonfiction, government legal, social, learner-essays and grammar-examples". No literary texts have been included thus far.

[4]Whereas, with regard to Dante Alighieri, his works in Latin are already part of UD, see [3, 4].

larly Dante Alighieri's vernacular, and Modern Italian,[5] especially when compared to the evolution of other Romance languages like French, where differences between old and modern varieties are more pronounced [7], numerous studies have now recognized and emphasized the distinction between Old and Modern Italian [8], particularly from a syntactic perspective [9].

The *Grammatica dell'italiano antico* (GIA; 'Grammar of Old Italian') [10], 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 the Tuscan region) 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 characterize Medieval Italy, albeit in a minority when it comes to documentation and written evidence.

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 text for a UD treebank, it seems obvious to select a Tuscan text, specifically a Florentine one, namely the *Comedy* of Dante Alighieri.

Dante Alighieri 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, and relevance among scholars. 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 spreading and elevating the vernacular to a literary language [11].

Together with the undeniable significance of the text, the availability of a digital resource, DanteSearch [12], 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 UD treebank of Old Italian.

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[6] (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. Strictly related to the the historical dictionary of Old Italian built by OVI is the *Tesoro della Lingua Italiana delle Origini* corpus (TLIO) [13], which collects 3 173 texts for a total of 23 685 634 occurrences. Additionally, there are corpora that cover a wider temporal span, such as the MIDIA corpus [14], a lemmatized and morphologically annotated collection of Italian texts from the 13th century to the first half of the 20th century, and the CODIT corpus [15], a diachronic corpus of Italian that covers the period from the 13th century until 1947.

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* [16],[7] no dependency-based syntactic annotation of Old Italian texts is currently available.

2.2. DanteSearch

Among the resources available for Old Italian, DanteSearch (DS) [12] is an annotated corpus containing all of Dante Alighieri's works, including both the Latin and the vernacular texts. The resource has been developed by the University of Pisa and consists of a set of (downloadable)[8] XML files providing both textual data and linguistic annotation.

Concerning the *Comedy*, the text included in DS is based on Petrocchi's edition [17] and is recorded in two separate XML files: one file provides the grammatical layer of annotation (featuring tokens, lemmas, and tags representing both parts of speech and morphosyntactic features), while the other contains a clause-based layer of syntactic annotation [18].

The clause-based annotation of syntax distinguishes main and subordinate clauses, the latter being assigned a label for their function, such as "declarative", "temporal", and "relative" [19].

3. 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 into 34, 33 and 33 subsections called *canti*. This Section details the process of annotating the *Inferno* according to UD's formalism.

[5] As exemplified by a statement by [5, p. 124], cf. [6, ch. VI].
[6] http://www.ovi.cnr.it/Il-Corpus-Testuale.html
[7] The work by Favaro consists of a conversion from an XML source file to the CoNLL-U format adopted in UD, for tokenization, lemmatization, and morphological annotation.
[8] https://dantesearch.dantenetwork.it.

3.1. From DanteSearch to UD

In DS, the *Inferno* consists of 33 416 tokens out of a total of 99 390 (without punctuation marks).

We perform a conversion from the grammatical XML file of the *Inferno* provided by DS to the CoNLL-U format adopted by UD's treebanks.[9] The conversion focuses on tokens (i. e. forms), lemmas, parts of speech (PoS), and morphological features. However, in the CoNLL-U file we do not report the syntactic annotation contained in the XML syntactic file of DS, due to its incompatibility with the word-based UD syntactic analysis [1, §2.2].

The conversion of tags happens on a 1:1 basis (DS:UD) whenever possible. Different criteria for the assignment of PoS 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).

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 [20], such as *mentre che* 'while' and *davanti a* 'in front of'. In DS, such multiword expressions are analyzed as single tokens, while the UD annotation schema requires that the words they are composed of be analyzed individually and considered as separate tokens. As a consequence, for locutions we employ a distinct tokenization, lemmatization, and PoS tagging in contrast to DS, as shown in Table 1 with regard to the following example:[10]

> *Inferno*, V, vv. 95–96
>
> 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.'[11]

Modifications of lemmatization and PoS tagging are required also for multiword proper nouns, which are lemmatized under a unique lemma in DS in contrast to UD. Table 2 shows the example of the multiword proper name *Filippo Argenti*:[12]

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

Table 1
Example of locution *mentre che*

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

Table 2
Example of the proper noun *Filippo Argenti*

Further, we also want 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 lemmatize both under the respective masculine forms.

3.2. Syntactic annotation

We perform the syntactic annotation of the *Inferno* manually[13] using *ConlluEditor* [22] and with the support of a few critical commentaries on the work, namely those by Chiavacci Leonardi [23] and Inglese [24]. Following the UD guidelines, annotation is made 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 [17].

A sentence corresponds to a syntactic tree, i. e. an acyclic, oriented, rooted graph [25], whose nodes correspond to tokens in the text.[14] Nodes are related to each other through dependencies, i. e. hierarchical binary relations, which are labeled with a syntactic function, such as

[9] CoNLL-U is a format with tab-separated values where lines contain the annotation of tokens into 10 fields; see https://universaldependencies.org/format.html.

[10] In this example, the DS tag `clst` stands for a subordinating conjunction (`cs`) used in a locution (`l`) within a temporal clause (`t`), while the UD PoS tags ADV and SCONJ stand respectively for "adverb" and "subordinating conjunction".

[11] The English translations of the examples from the *Comedy* are by Allen Mandelbaum, available at: https://digitaldante.columbia.edu/dante/divine-comedy/.

[12] The DS tag `n` stands for "onomastics" and the UD tag PROPN stands for "proper noun".

[13] The syntactic annotation is performed by a single annotator with expertise in Italian studies. Annotating pre-parsed data has been ruled out after evaluating the accuracy of the UDPipe model [21] 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 4).

[14] 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 in the syntactic tree. These two levels often, but not always, coincide, e. g. the token *nel* 'in the' would be analyzed into the syntactic words *in* 'in' and *il* 'the', each bearing its own annotation. Refer to https://universaldependencies.org/u/overview/tokenization.html. In this paper, the term "token" will be used throughout as an equivalent to UD's "syntactic word".

nsubj for "nominal subject". Dependency-based annotation schemes are predicate-centered, with the sentence's main predicate serving as the tree's root. In UD's formalism, function words depend on the content words they modify.[15]

While annotating the *Inferno* according to the UD formalism, we encounter several issues that require taking specific decisions. In the following, we discuss the annotation of ellipses and comparative clauses.

The total number of sentences in the *Inferno* is 1 228, for a total of 41 367 tokens.

3.2.1. Ellipsis

"Ellipsis" refers to the omission of words or phrases that can be inferred from the context of a sentence or utterance.[16] While annotating the *Inferno*, we encounter several cases of ellipses, including nominal ellipses, i. e. [27, p. 526]:

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

and predicate ellipsis [28, 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 ellipsis, we follow the solution of promotion, as outlined in the UD guidelines.[17] We present here an example of nominal ellipsis (Figure 1):

> *Inferno*, IX, vv. 28–29:
>
> Quell'è 'l più basso loco e 'l più *oscuro* / e 'l più *lontan* dal ciel che tutto gira
>
> 'That is the deepest place and the darkest place, / the farthest from the heaven that girds all'

where *oscuro* 'dark' and *lontan* 'far' depend on the omitted noun (NOUN) *loco* 'place', as shown by the repetition of the definite article (DET) *'l* 'the', which modifies the noun. In this case, we promote the adjectives (ADJ) *oscuro* and *lontan* to heads of their respective coordinate clauses using the dependency relation `conj` "conjunct".

Following to the UD guidelines, we handled predicate ellipsis by using the dependency relation `orphan` (orphan relation), like in the following example:

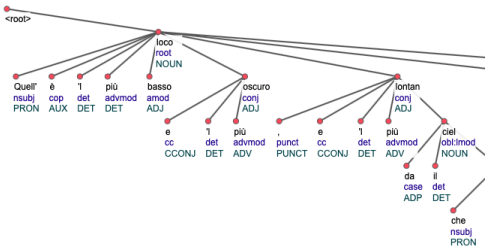

Figure 1: Nominal ellipsis (*Inf.*, IX, vv. 28–29)

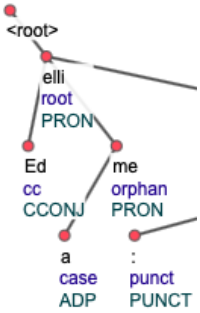

Figure 2: Predicate ellipsis (*Inf.*, III, v. 76)

> *Inferno*, III, v. 76:
>
> Ed elli a me:
>
> [And he (said) to me:]

where the predicate of the sentence, namely the *verbum dicendi*, is omitted. This structure is extremely common to introduce a reported speech. As shown in Figure 2, the omission of the predicate requires promoting the subject of the sentence, *elli* "he", to the root of the tree (`root`) and annotating the underlying oblique relation (`obl`) of the phrase *a lui* "to him" with an orphan relation (`orphan`).

Currently, the syntactic annotation in UD handles these cases of ellipsis with the promotion mechanism, which involves promoting an element to function as the omitted element in the sentence and replacing it in its dependency relation without explicitly signaling this omission, and the use of the `orphan` dependency relation, whose function is to indicate that the element subject to

[15] 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. Refer to https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html.

[16] See [26] for an introduction to the topic.

[17] Promotion involves selecting an element to take the place of the omitted element in the syntactic tree, following a specific hierarchy. Promotion is used without explicitly signaling the ellipsis. See UD guidelines: https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis.

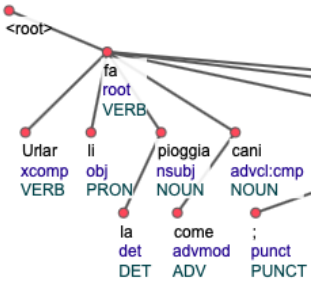

Figure 3: Comparative clause (*Inf.*, VI, v. 19)

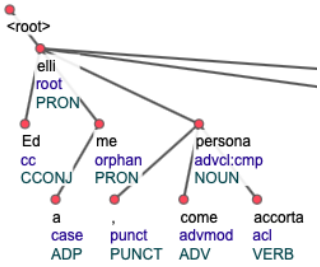

Figure 4: Secondary predication or comparative clause (*Inf.*, III, v. 13)

the `orphan` relation does not have an overt dependent element in the syntactic structure.

3.2.2. Comparative clauses

In the *Inferno*, we find a diverse usage of comparative clauses, ranging from sentences where the comparative clause is longer than the main clause it depends on, to others where comparatives consist of just a few tokens. In light of the long-lasting discussion on the treatment of comparative clauses in UD,[18] we annotate such clauses by labeling their head tokens with the dependency relation `advcl` "adverbial clause modifier" specified for the subtype `cmp` for comparative clauses.[19]

A number of issues concerns cases of clauses where our annotation, following the UD framework, parts from the interpretation provided by DS, like in the following example (Figure 3):

> *Inferno*, VI, v. 19:
>
> Urlar li fa la pioggia *come cani*
>
> 'That downpour makes the sinners howl *like dogs*'

In the annotation of DS, the portion *come cani* 'like dogs' is considered a phrase that is part of a declarative clause. Instead, we consider *come cani* as a comparative clause with an elliptical predicate, namely *Urlar li fa la pioggia come [fa urlare] i cani* 'That downpour makes the sinners howl like [it makes] dogs [howl]'.

We observe a few cases where a *come*-clause can be considered either a comparative clause or a secondary predication. In such cases, we rely on the interpretation provided by commentaries, like in the following sentence (Figure 4):

> *Inferno*, III, v. 13:
>
> Ed elli a me, *come persona accorta*:
>
> 'And he to me, *as one who comprehends*:'

In this sentence, the *come*-clause can be interpreted either as a secondary predication (therefore, annotated using the subtyped relation `advcl:pred`[20]), 'He, being a comprehensive person, answered to me', or as a comparative clause (with subtyped relation `advcl:cmp`), 'He answered to me like a comprehensive person'. In this case, we follow the interpretation of Chiavacci Leonardi [23] in considering the *come*-clause as a comparative construction.

4. Evaluation

We use the manually annotated *Inferno* to train models with UDPipe 1[21] [29] and to assess their performances in view of employing them for parsing *Purgatorio* and *Paradiso*, so as to facilitate their subsequent manual annotation.[22] In our evaluation framework, we employ a cross-validation based on 10%/90% splits of the data: each test set will then consist 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 (LAS) and Unlabeled Attachment Score (UAS) [30].

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

[18] Cf. the discussion group on comparatives in UD: https://universaldependencies.org/workgroups/comparatives.html.

[19] Cf. documentation at https://universaldependencies.org/la/dep/advcl-cmp.html (for Latin).

[20] Cf. documentation at https://universaldependencies.org/la/dep/advcl-pred.html (for Latin).

[21] https://github.com/ufal/udpipe.

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

Partition	Scenario	Avg. UAS	Avg. LAS
random	+Morph	81,95±0,94%	77,07±1%
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%

Table 3
Averages and standard deviations of accuracy metrics

canti[23] assigned to the test set and the remaining 31[24] forming the training set; the second partition is obtained by fully random selection of sentences.[25] 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.[26]

The accuracy of each model is calculated using `eval.py`,[27] an evaluation script provided by the UD project. As shown in Table 3, evaluations conducted on the random partition result into slightly higher average accuracy scores than those based on triplets[28] 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, 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 low. 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 *Purgatorio* and *Paradiso*, *cantiche* for which lemmatization and morphosyntactic taggings are inherited from the conversion from DS.

5. Conclusions and future perspectives

Building a UD treebank 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 both as training and test sets data from the *Inferno*. 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 between 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 will suggest to introduce a specific subtype, e. g. `ellp`, in UD's documentation, so as to properly identify cases of ellipses, as they are not explicitly captured by the current annotation strategies mentioned in the paper, 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 is up to the annotation of so-called enhanced dependencies, which are a kind of advanced annotation that augments dependency

[23] 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).
[24] Or 30; see fn. 23.
[25] Please refer to the GitHub page https://github.com/ClaudiaCorbe/Inferno_treebank for the data and detailed statistics on the partitions.
[26] Corresponding respectively to `–parse` and `–tag –parse` options for UDPipe; see https://ufal.mff.cuni.cz/udpipe/1/users-manual, §3.6.
[27] https://github.com/UniversalDependencies/tools/blob/master/eval.py.
[28] Or a quadruplet; see fn. 23.

labels to facilitate disambiguation.[29]

We plan to engage additional annotators with expertise in Old Italian to expedite the process of annotation of *Purgatorio* and *Paradiso*. Additionally, we intend to apply error detection processes (like, for instance, those described in [31]) to retrieve possible mistakes or inconsistencies in syntactic annotation.

Another task we intend to address is the extension of the UD documentation for Italian in order to make the validator[30] correctly deal with some peculiarities of Old Italian, like for instance enclitic adpositions (e.g., *meco* 'with me'), which require the introduction of the feature `Clitic=Yes` combined with the PoS tag ADP, currently permitted only with the PoS tag PRON.

Finally, we plan to include enhanced dependencies [31] in the UD treebank of Dante's *Comedy*, once the basic syntactic annotation of the entire work will be completed.

References

[1] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. URL: https://direct.mit.edu/coli/article/47/2/255/98516/Universal-Dependencies. doi:10.1162/coli_a_00402.

[2] D. Zeman, et alii, Universal dependencies 2.12, 2023. URL: http://hdl.handle.net/11234/1-5150, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Available at http://hdl.handle.net/11234/1-5150.

[3] F. M. Cecchini, R. Sprugnoli, G. Moretti, M. Passarotti, UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works, in: J. Monti, F. Dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020, Bologna, Italy, March 1–3 2021), Associazione italiana di linguistica computazionale (AILC), Accademia University Press, Turin, Italy, 2020, pp. 99–105. URL: http://ceur-ws.org/Vol-2769/paper_14.pdf.

[4] M. Passarotti, F. M. Cecchini, R. Sprugnoli, G. Moretti, *UDante*, Studi Danteschi LXXXVI (2022) 309–338.

[5] G. I. Ascoli, L'Italia dialettale, Archivio glottologico italiano VIII (1882–1885) 98–128. Available at https://archive.org/details/archivioglottolo08fireuoft/page/n5/mode/2up.

[6] L. Tomasin, Il caos e l'ordine, Piccola Biblioteca Einaudi, Giulio Einaudi, Turin, Italy, 2019.

[7] M. Dardano (Ed.), Sintassi dell'italiano antico, Lingue e Letterature Carocci, Carocci, Rome, Italy, 2013. URL: https://www.carocci.it/prodotto/sintassi-dellitaliano-antico.

[8] M. Dardano, G. Frenguelli (Eds.), SintAnt, Aracne, Rome, Italy, 2004.

[9] R. Tesi, Parametri sintattici per la definizione di "Italiano antico", in: M. Dardano, G. Frenguelli (Eds.), SintAnt. La sintassi dell'italiano antico, Aracne, Rome, Italy, 2004, pp. 425–444.

[10] G. Salvi, L. Renzi (Eds.), Grammatica dell'italiano antico, il Mulino, Bologna, Italy, 2010. URL: https://www.mulino.it/isbn/9788815134585.

[11] P. Manni, La lingua di Dante, Le vie della civiltà, il Mulino, Bologna, Italy, 2013.

[12] M. Tavoni, DanteSearch: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica, in: A. Cerbo, R. Mondola, A. Žabjek, C. D. Fiore (Eds.), 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, 2011, pp. 583–608.

[13] P. G. Beltrami, Il Tesoro della Lingua Italiana delle Origini (TLIO), in: N. Maraschio, T. Poggi Salani, M. Bongi, M. Palmerini (Eds.), 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, 2003, pp. 695–698. URL: https://www.torrossa.com/it/catalog/preview/2280318. doi:10.1400/28371.

[14] P. D'Achille, M. Grossmann (Eds.), Per la storia della formazione delle parole in italiano, Quaderni della Rassegna, eighth ed., Franco Cesati, Florence, Italy, 2017. URL: https://www.francocesatieditore.com/catalogo/per-la-storia-della-formazione-delle-parole-in-italiano/.

[15] M. S. Micheli, CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field, Corpus 23 (2022). URL: https://journals.openedition.org/corpus/7306. doi:10.4000/corpus.7306.

[16] M. Favaro, E. Guadagnini, E. Sassolini, M. Biffi, S. Montemagni, Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations, in: R. Sprugnoli, M. Passarotti (Eds.), Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA), European Language Resources

[29] https://universaldependencies.org/u/overview/enhanced-syntax.html.

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

[31] https://universaldependencies.org/u/overview/enhanced-syntax.html

Association (ELRA), Marseille, France, 2022, pp. 94–100. URL: https://aclanthology.org/2022.lt4hala-1.13/.

[17] D. Alighieri, 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, 1994. URL: https://www.lelettere.it/libro/9788871661483, editor: Giorgio Petrocchi.

[18] M. Tavoni, Allestimento, fruizione e prospettive di DanteSearch, in: E. Cresti, M. Moneglia (Eds.), 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, 2022, pp. 255–273. URL: https://www.societadilinguisticaitaliana.net/wp-content/uploads/2022/11/017_Tavoni_Atti_LIV_Congresso_SLI.pdf. doi:10.17469/O2106SLI000017.

[19] S. Gigli, La codifica sintattica della Commedia di Dante, in: M. D'Amico (Ed.), Sintassi dell'italiano antico e sintassi di Dante. Atti del seminario di studi (Pisa 15/16 ottobre 2011), Felici, Ghezzano (PI), Italy, 2015, pp. 81–95.

[20] L. Serianni, A. Castelvecchi, Grammatica italiana, Universitaria, second ed., UTET Università, Turin, Italy, 2006.

[21] M. Straka, J. Hajič, J. Straková, UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. O. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 4290–4297. URL: https://aclanthology.org/L16-1680.

[22] J. Heinecke, ConlluEditor: a fully graphical editor for Universal Dependencies treebank files, in: A. Rademaker, F. Tyers (Eds.), Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019), Association for Computational Linguistics (ACL), Paris, France, 2019, pp. 87–93. URL: https://aclanthology.org/W19-8010. doi:10.18653/v1/W19-8010.

[23] D. Alighieri, Inferno, number 613 in Oscar classici, Arnoldo Mondadori, Milan, Italy, 2005. Editor: Anna Maria Chiavacci Leonardi.

[24] D. Alighieri, Commedia. Inferno, number 1 in Opere, Carocci, Rome, Italy, 2007. Editor: Guglielmo Inglese.

[25] J. Havelka, Mathematical Properties of Dependency Trees and their Application to Natural Language Syntax, Ph.D. thesis, Univerzita Karlova – Matematicko-fyzikální fakulta, Prague, Czech Republic, 2007. URL: https://dspace.cuni.cz/handle/20.500.11956/12614?locale-attribute=en.

[26] J. Merchant, Ellipsis: A survey of analytical approaches, in: J. van Craenenbroek, T. Temmerman (Eds.), The Oxford Handbook of Ellipsis, Oxford Handbooks, Oxford University Press, Oxford, UK, 2019. URL: https://academic.oup.com/edited-volume/41718/chapter/353990361.

[27] A. Saab, Nominal Ellipsis, in: J. van Craenenbroek, T. Temmerman (Eds.), The Oxford Handbook of Ellipsis, Oxford Handbooks, Oxford University Press, Oxford, UK, 2019, pp. 526–561. URL: https://academic.oup.com/edited-volume/41718/chapter-abstract/353995808. doi:10.1093/oxfordhb/9780198712398.013.26.

[28] A. Lobke, W. Harwood, Predicate Ellipsis, in: J. van Craenenbroek, T. Temmerman (Eds.), The Oxford Handbook of Ellipsis, Oxford Handbooks, Oxford University Press, Oxford, UK, 2019, pp. 504–525. URL: https://academic.oup.com/edited-volume/41718/chapter-abstract/353995176.

[29] M. Straka, J. Straková, Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, in: J. Hajič, D. Zeman (Eds.), Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 2017, pp. 88–99. URL: https://aclanthology.org/K17-3009. doi:10.18653/v1/K17-3009.

[30] S. Buchholz, E. Marsi, CoNLL-X Shared Task on Multilingual Dependency Parsing, in: L. Màrquez, D. Klein (Eds.), Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), Association for Computational Linguistics (ACL), New York City, NJ, USA, 2006, pp. 149–164. URL: https://aclanthology.org/W06-2920.

[31] M. Dickinson, Error Detection and Correction in Annotated Corpora, Ph.D. thesis, The Ohio State University, 2005. URL: https://sifnos.sfs.uni-tuebingen.de/decca/publications/dickinson-dissertation.html.