What Is Better for Syntactic Parsing? A Comparison Between Supervised and Unsupervised Models on Dante and Cavalcanti

Claudia Corbetta[1,2,*], Anna Erminia Colombi[2], Giovanni Moretti[3] and Marco Passarotti[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
This paper investigates the performance of two models on Cavalcanti's *Rhymes*: a supervised neural model (Stanza) trained on the Italian-Old treebank (comprising Dante's *Divine Comedy*), and an unsupervised generative Large Language Model (LLM) accessed via the ChatGPT API (o3 version). This study highlights the crucial role of textual edition in processing historical texts, illustrating this through examples from different editions. It also presents a manual error analysis of the models' outputs, focusing on both the most frequent and the most linguistically nuanced errors.

Keywords
dependency parsing, Divine Comedy, Cavalcanti, Universal Dependencies, Stanza, LLM

1. Introduction

Syntactic studies[1] can offer valuable insights from multiple linguistic perspectives [1]. In the domain of dependency syntax, Universal Dependencies[2] [2] (henceforth UD) provides a cross-linguistic framework along with a suite of Natural Language Processing (NLP) tools for consistent and reliable comparison [2, p.256]. However, annotating dependency syntax remains both time-consuming and demanding for human annotators. In light of this, increasing attention has been devoted to the use of computational models, with both supervised and unsupervised methods, to support and potentially accelerate the annotation process.

Although previous studies [3, 4] have highlighted the limitations of unsupervised models, particularly Large Language Models (LLMs), in capturing syntactic structure, thereby reinforcing the view that manual annotation remains essential, albeit labor-intensive, we explore this task in the context of Old Italian. As a historical language variety, Old Italian[3] has not yet been the subject of experiments with unsupervised models. To date, experiments on the dependency syntax of Old Italian (more specifically, in the context of 13th-century Old Florentine poetry) have been conducted exclusively using supervised models, including UDPipe [5] and Stanza [6].[4]

To this end, we set out to test and compare the performance of an unsupervised LLM and a supervised model trained on a specific textual domain. The goal is to assess their accuracy and their potential utility in supporting the annotation process.

We conduct our test on Old Italian texts, for which a gold-standard UD-annotated treebank, (i.e., a syntactically annotated corpus) is available, namely, Italian-Old (see Section 2.1). Since this treebank includes Dante Alighieri's *Divine Comedy*, we chose to test models trained on that data on an author contemporary with Dante, namely Guido Cavalcanti.[5]

This paper is structured as follows. Section 2 describes the data used for the experiment (2.1 and 2.2), offers insights into the impact of editorial choices on annota-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
claudia.corbetta@unibg.it (C. Corbetta);
annaerminia.colombi01@universitadipavia.it (A. E. Colombi);
giovanni.moretti@unicatt.it (G. Moretti);
marco.passarotti@unicatt.it (M. Passarotti)

[1]For the specific concerns of the Italian academic attribution system: all authors are jointly responsible for Section 1 and Section 5. Claudia Corbetta is also responsible for Section 2 (2.1, 2.2 and 2.4), Section 3, Section 4 (4.1 and 4.3) and annotation; Anna Erminia Colombi is also responsible for Section 2 (2.3), Section 4 (4.2) and annotation; Marco Passarotti also is responsible for Section 3; Giovanni Moretti trained Stanza models and is responsible for the prompt. Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[2]https://universaldependencies.org/#language-u.

[3]For an insight into the definition of Old Italian, we refer to our previous work [7].

[4]For tests with UDPipe, see [8], and for a comparison between UDPipe and Stanza, refer to [9]. Finally, for an evaluation of a combined model trained on both Modern and Old Italian, see [10].

[5]Guido Cavalcanti (c. 1255–1300) was a prominent poet of the late 13th century. A contemporary and close friend of Dante Alighieri, he significantly influenced the early Italian lyric tradition and the development of vernacular literature. For insight about Cavalcanti refer to [11].

tion (2.3), and discusses annotation practices and inter-annotator agreement (2.4). Section 3 outlines the training processes (3.1 for the neural model and 3.2 for the LLM) as well as the evaluation of the models' performance (3.3). In Section 4, we provide a manual error analysis of the output of the LLM (4.1) and the supervised model (4.2), along with a discussion of the errors common to both models (4.3). Finally, Section 5 presents the conclusions.

2. Data

This Section presents the data used in the study.

2.1. Italian-Old Treebank

As of now, the Italian-Old Universal Dependencies treebank [8] is the only resource within the UD framework that provides annotation for Old Italian. It consists of a dependency-annotated corpus of Dante Alighieri's *Divine Comedy*.[6] The Italian-Old treebank is an openly available resource based on the DanteSearch corpus [13], enriched with dependency syntax annotation and adapted to conform to the UD guidelines. The linguistic annotation of the corpus is encoded in CoNLL-U format,[7] and includes tokenization, lemmatization, morphological annotation (covering both part-of-speech and morphological features), and dependency syntax. Moreover, since the annotation is word-level, additional metadata indicating the position of each word, namely, the verse and the *Canto*, are also provided.

2.2. Cavalcanti's *Rhymes*

For the present work, we selected the *Rhymes* of Guido Cavalcanti edited by Ercole Rivalta [14]. We chose this particular edition because the texts are publicly available online,[8] which allowed us to extract them via web scraping. We then cleaned the text by removing editorial notes related to manuscript sources and the editor's commentary.

Rivalta's edition of the *Rhymes* divides the poems into three chronological groups: those composed before 1290, those of uncertain date, and those composed after 1290. The corpus includes a total of 63 poems, distributed as follows: 25 in the first group, 23 in the second, and 15 in the third. Various metrical forms are attested among the poems, ranging from ballads and sonnets to *canzoni* and similar lyric types.[9] Moreover, alongside the classical poems, some rhymes belong to the "sonetti di risposta" (stilnovist reply sonnets addressed to other poets) and thus differ in their compositional purpose.

It is important to recall that, when working with texts transmitted through manuscript traditions, editorial choices play a critical role. This is also the case for the *Rhymes*, whose textual variants and structure depend heavily on editorial interpretation, as it is discussed in Subsection 2.3. By selecting the edition of the *Rhymes* edited by Rivalta, we adhere to his reconstruction of the texts, along with all the implications that such a choice entails.[10]

2.3. Impact of Editorial Choices on Annotation

As mentioned in 2.2, the decision to adopt Rivalta's edition for the selection of poems necessarily entails a reflection on editorial differences. Accordingly, three examples will be presented below, in which the comparison with the edition edited by Roberto Rea and Giorgio Inglese [19], chosen because it is one of the most recent critical edition with commentary, highlights stylistic and interpretative differences that are reflected in the lexical and syntactic analysis.

Firstly, in verse 14 of Sonnet 28, the editions differ as follows:[11]

Example I - Sonn. 28 v. 14
*chè **morte** 'l porta in man* (Rivalta ed.)
*ch'**e' morto** 'l porta in man* (Rea ed.) (whose heart Death [...] carve into the man's gravestone)

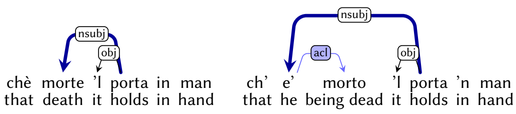

Figure 1: Dependency tree of Sonn. 28 v. 14 Rivalta and Rea ed., respectively.

[6]The critical edition of the *Divine Comedy* used while building the treebank is Petrocchi (1994) [12].
[7]CoNLL-U is a tab-separated format in which each line encodes the annotation of a syntactic word across 10 fields. For further details, see https://universaldependencies.org/format.html. When a field is not applicable or remains unannotated, the placeholder "_" is used.
[8]See https://it.wikisource.org/wiki/Rime_(Cavalcanti).
[9]Refer to [15] for a detail on metrical forms.
[10]Rivalta's edition is not the most up-to-date edition of the *Rhymes*, which have since been the subject of further scholarly studies and have led to other editions curated by other scholars (such as Contini [16], Ciccuto [17], De Robertis [18], and Rea and Inglese [19]). However, as mentioned above, the choice of Rivalta's edition was motivated by practical reasons. While leaving a comparison with these more recent editions to future work, we emphasize the need for up-to-date digital editions to be freely and openly available online.
[11]For the translation of Cavalcanti's poems, we rely on [20]. This translation does not always adhere closely to the original Italian text. Consequently, we provide our own glosses where appropriate.

In this case, the main divergence is exegetical: in Rivalta's version, *morte* (death), personified, is the subject (`nsubj`) who holds the object (`obj`) *'l* (it) in her hand. By contrast, in Rea's edition, the personified *morte* is replaced by an actual agent, *e'* (he), who is *morto* (dead). While the logical relation, namely, that there is a subject who "holds" something in their hand, remains consistent across both versions, Rea's interpretation, as shown in the Figure 1, introduces an additional token. This addition occurs since *morte* is replaced by the nominal modifier *morto*, in the tree `acl` (adnominal clause) and the modified referent *e'* must therefore be explicitly included. From a strictly syntactic perspective, this difference results in a clearly distinct sentence subject, as shown in Example I, and in the presence of an additional token, which inevitably yields a slightly more complex syntactic structure.[12]

Secondly, in verse 7 of Sonnet 10, we observe a case in which the difference between editions affects only the token count, without any consequences for interpretation:

> ***Perchè*** *Sospiri e Dolor mi pigliaro* (Rivalta ed.)
> ***Per che*** *Sospiri e Dolor mi pigliaro* (Rea ed.)
> (and were delighted to hear my sighs and groans)

In Rivalta's edition, the causal conjunction appears as *perché* (because), which is a single token. In Rea's edition [19, p.62], the same causal meaning is conveyed through the two-word form *per che*, resulting in two token. This variation affects not only the token count, but also the structure of the tree, as it requires an additional syntactic dependency and dependency relation label. Moreover, frequent variation in tokenization may lead to measurable differences in other types of metrics (such as Type/Token Ratio analysis) depending on the reference edition adopted.

Finally, the third case examines an example that differs from the first two previously discussed. In this instance, the editorial variation does not result in a different number of token, but rather in the presence of lemmas with different meanings:

> *Veder poteste quando vui* ***scontrai*** (Rivalta ed.)
> *Veder poteste quando v'****inscontrai*** (Rea ed)
> (one seldom sees him as if in flesh)

As can be observed, the difference in verse 1 of sonnet 35 between the two editions concerns the use of *scontrai* versus *inscontrai*. When these token are lemmatized, they result in different lemmas: *scontrai* is the conjugated form of *scontrare* (to collide), whereas *inscontrai* corresponds to *incontrare* (to meet). As seen previously, even though this variation does not pose a syntactic problem in this case, the use of two different forms can impact other levels of linguistic analysis and obviously yield different results depending on which edition was used as the basis for annotation.

In conclusion, editorial choices play a crucial role, as they can influence tokenization, interpretation, and lemmatization, as demonstrated by the three examples discussed above.

2.4. Manual Annotation and Inter-Annotator Agreement

Once the reference edition was established, we manually annotated 22 sonnets out of the 63 poems included in Rivalta's edition of Cavalcanti's Rhymes (2135 token, excluding punctuation marks). We chose to focus exclusively on sonnets, as this metrical form is the most frequently attested throughout the collection. The manual annotation was carried out by two expert annotators.

Among the 44 sonnets[13] in the corpus, we selected 22, distributing them across the three periods identified in Rivalta's edition: 6 sonnets from the first period, 7 from the second, and 6 from the third (totaling 19 sonnets[14]). In addition, 3 sonnets were annotated by both annotators to calculate inter-annotator agreement.

To evaluate consistency between the two annotators and alignment with the annotation style of the Italian-Old treebank, we also performed inter-annotator agreement on three *Canti* of the *Divine Comedy*. The selected *Canti* were the 13th of each *Cantica*, namely *Inferno*, *Purgatorio* and *Paradiso*, in order to account for potential stylistic variation across the three parts of the poem (3296 token excluding punctuation marks).

Table 1 reports the inter-annotator agreement results for both the three 13th *Canti* of the *Divine Comedy* (Dante) and the three sonnets by Cavalcanti. Inter-annotator agreement was assessed using Fleiss' kappa [21], a statistical measure for evaluating the reliability of agreement between multiple annotators.

Table 1
Table of inter-annotator agreement

	Dante	Cavalcanti
edges	0.95	0.94
labels	0.92	0.98

[12]For reasons of space and clarity, we report only the dependency relations discussed in the text. To view the full trees, please refer to the GitHub page: https://github.com/CIRCSE/CavalcantiRepository.git.

[13]This total includes one "ritornellato" sonnet (i.e., a sonnet composed of more than the canonical 14 lines and featuring a specific metrical pattern; for details, refer to [15, p. 284]) and one "rinterzato" sonnet (i.e., a composition structured as a duplex sonnet; see [15, p. 283]).

[14]The 19 sonnets were split between the two annotators: 10 for one and 9 for the other.

The overall inter-annotator agreement is very high for both authors. For Cavalcanti, Fleiss' kappa on dependency edges reaches 0.94, and the agreement on dependency labels assigned to correctly matched edges is even higher, at 0.98. For Dante, the values are similarly strong, with 0.95 for edges and 0.92 for labels. These results indicate a high level of consistency between annotators and confirm the overall reliability of the annotation process across both corpora.

3. Training and Models Performances

In this experiment, we evaluate the parsing performance of two different models in order to identify the most suitable for this task. Specifically, we assess the performance of (i) a Stanza retrained model [6], a neural network model with a parsing-specific architecture specifically trained on Old Italian data, namely Italian-Old and henceforth referred to as **Dante model**, and (ii) a zero-shot generative LLM accessed via the ChatGPT API (o3 version), [22], henceforth referred to as **LLM**. For this experiment, Stanza was selected as the supervised model, as it has demonstrated superior performance compared to UDPipe (see [9]). As for the unsupervised model, we opted to begin by testing the ChatGPT API, while leaving the evaluation of bidirectional LLMs, such as BERT-based models like UDify [23] and U-DepPLLaMA [24], for future work.[15]

3.1. Dante Model

We train a Stanza model [6] using Italian-Old data and use it to parse Cavalcanti's *Rhymes*.[16] We conduct two types of experiments: one in which the model performs full annotation from scratch, namely, tokenization, lemmatization, part-of-speech tagging, and syntactic parsing (hereafter labeled All) and one in which it performs only syntactic analysis (hereafter labeled OS, for Only Syntax), with the other annotation layers pre-supplied.

We evaluate only syntactic metrics,[17] specifically the Unlabeled Attachment Score (UAS) and the Label Attachment Score (LAS).[18] Table 2 reports the model's performance. Specifically, the scores in the Dante column reflect the performance of Dante model on a *Divine Comedy* test set. These scores derive from Stanza's automatic internal evaluation on that test split and pertain to complete annotation from scratch, rather than being limited to syntactic analysis alone.[19] In contrast, the scores in the Cavalcanti columns, both for full parsing and syntax-only (OS), reflect the evaluation of the sonnets annotated by Dante model, compared against the gold-standard annotations produced by the two annotators (as described in Section 2.4).

Table 2
Dante Model performances on Dante and Cavalcanti's texts

	Dante All	Cavalcanti All	Cavalcanti (OS)
UAS	80.13	82.59	85.26
LAS	74.29	75.24	78.28

Interestingly, the performance of the Dante-trained model on Cavalcanti's sonnets is higher (+2.46 for UAS and +0.95 for LAS) than the internal evaluation conducted by Stanza on the *Comedy* data. These results seem to suggest good portability of Dante model, even when applied to texts by a different author, though from the same period and literary context. Clearly, this is only a preliminary test, and further research is required, especially on texts that move away from the poetic genre.

3.2. LLM

In light of the growing prominence of LLMs, we investigate whether ChatGPT can produce results comparable to those of Dante model we trained.

To this end, we tested the ChatGPT API using a tailored prompting technique. We report the prompt in Appendix A. More specifically, we prompt ChatGPT to generate the UD annotation of a sonnet by providing, in a first setting, the raw text as input, and in a second setting, the gold-standard CoNLL-U file with syntactic annotations removed. Using the "assistant" role, we first provide the model with a gold-standard annotated sonnet as an example. We then ask it to perform the same task, producing a CoNLL-U formatted annotation, for a different sonnet. We set the temperature to a minimum value (0.05) and set top_p[20] to 1 in order to make the model as deterministic as possible.

Since our aim is to compare the performance of the LLM with that of Dante model, we tested the LLM in two settings, mirroring the evaluation setup used for Dante model: (i) generating the full CoNLL-U file from scratch,

[15]While preliminary results suggest that such models yield an improvement in syntactic accuracy for Italian (see [23] and [24]), it is worth noting that they have not yet been tested on Old Italian data.

[16]The dataset comprises 122 000 token and is divided into training, development, and test sets with an 80-10-10 split.

[17]To perform evaluation we use `eval.py` script, available at https://github.com/UniversalDependencies/tools/blob/master/eval.py.

[18]Refer to [25] for details on the evaluation metrics.

[19]These internal evaluation scores are also consistent with those obtained in a similar experiment [10], in which an Old Italian model (trained on a small amount of data) was tested in-domain, yielding to similar scores (UAS 82.24 and LAS 75.86).

[20]The top_p value corresponds to nucleus sampling: for high values of p, the model selects from a small subset of the vocabulary, the nucleus, which contains the majority of the probability mass [26, p. 5].

that is, producing tokenization, lemmatization, PoS tagging, morphological feature attribution, and syntactic annotation (using the first setting, with raw text as input); and (ii) filling only the syntactic fields, based on gold-standard tokenization and morphological information (corresponding to the second setting).

The first experiment did not yield satisfactory results, as the output failed to conform to the CoNLL-U format (both in terms of column structure and syntactic annotation) and frequently produced cycles[21] in the dependency trees. Even though we repeated the experiment multiple times, testing different sonnets and explicitly instructed the model to avoid cycles in the syntactic structure, its performance remained consistently poor across all runs, producing CoNLL-U outputs that were ultimately unusable, forcing us to discard this approach.

The second scenario, namely, instructing the LLM to focus exclusively on the syntactic fields while retrieving tokenization, lemmatization, and morphological annotation from the gold data, produced usable outputs, although some errors were still occasionally present (see Subsection 4.1). We report the results in Subsection 3.3.

3.3. Comparing Dante Model and LLM

To compare the two models, we evaluate their syntactic performance on eight randomly selected sonnets[22], specifically, sonnets 9, 10, 28, 29, 30, 31, 35 and 54.[23]

The corresponding UAS and LAS scores are reported in Table 3, showing the performance of the LLM for the syntactic task (under the LLM column) and that of Dante's model in the syntax-only setting (D_OS). Although, as mentioned in Subsection 3.2, we were unable to fully evaluate the LLM across all annotation levels, we nonetheless report the scores of Dante model when performing full annotation (D_All) for reference. The Diff column indicates the difference in performance between D_OS and the LLM.

Notably, the LLM achieves relatively high performance (see average row (av.) in Table 3), demonstrating its potential for syntactic annotation. However, its results remain below those of the neural model (Dante model), which was specifically trained on data closely aligned with the target material. As shown in Table 3, these findings suggest that, at least in this experiment and at the current stage of development, at the current stage of development, the neural model trained on a domain-relevant corpora offers a more reliable solution for high-quality linguistic annotation.

[21] A cycle in a dependency tree occurs when a word ends up depending on itself, violating the tree structure and rendering the annotation invalid.

[22] The number of sonnets was limited to eight due to funding constraints, as the ChatGPT API is a paid service.

[23] Refer to Appendix A for the titles of the selected sonnets.

Table 3
Dante Model and LLM's Performances and Differences

Son		LLM	D_All	D_OS	Diff_LLM_OS
10	UAS	68.22	80.77	89.72	21.50
	LAS	58.88	73.08	83.18	24.30
28	UAS	75.21	85.95	87.60	12.39
	LAS	68.60	77.69	80.99	12.39
29	UAS	74.77	73.58	74.77	0.00
	LAS	70.09	69.81	71.03	0.94
30	UAS	75.42	86.44	88.14	12.72
	LAS	66.95	76.27	81.36	14.41
31	UAS	65.85	80.49	82.93	17.08
	LAS	56.91	73.17	76.42	19.51
54	UAS	76.64	89.62	88.79	12.15
	LAS	70.09	82.08	82.24	12.15
9	UAS	74.79	84.75	86.55	11.76
	LAS	65.55	80.51	83.19	17.64
35	UAS	81.73	90.35	90.38	9.03
	LAS	80.77	89.42	89.42	8.65
av.	UAS	74.08	83.99	86.11	12.07
	LAS	67.23	77.75	80.97	13.74

In light of the obtained scores, we conducted a sample-based manual error analysis, which is described in Section 4.

4. Manual Errors Analysis

In this Section , we analyse the models' errors to provide useful insights for potential improvements and further considerations. We conduct a manual error analysis on the three sonnets exhibiting the most significant discrepancies in scores, namely, sonnets 10 and 31, which performed poorly with the LLM (see Subsection 4.1), and sonnet 29, whose score was comparable for both the LLM and Dante model (see Subsection 4.2). For the sake of clarity, we provide both sonnets along with their translations in Appendix A.

4.1. LLM Errors

The two sonnets that received the lowest scores from the LLM are sonnet 10 and sonnet 31.

Interestingly, a detailed inspection of Sonnet 10, the one that received the lowest score, reveals that one of the model's errors was the incorrect identification of the sentence's `root`. In this case, the LLM mistakenly analyses a subordinate clause as the main clause, assigning the verb *vider* (saw) in example II the status of `root` instead of identifying it as the head of an adverbial clause (`advcl`). This misassignment results in a series of incorrect syntactic attachments throughout the sentence. To illustrate this, we present two syntactic trees: the gold-standard tree (Figure 3) and the one produced by the LLM (Figure

2). Incorrect attachments in the LLM output are highlighted in red, while incorrectly assigned dependency labels are shown in yellow.

Example II - Sonn. 10 v. 10
Quando mi vider, tutti con pietanza/ dissermi
(Their somber welcome to me I still shudder at)

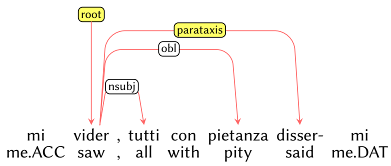

Figure 2: LLM syntactic tree of Sonn. 10 v.10

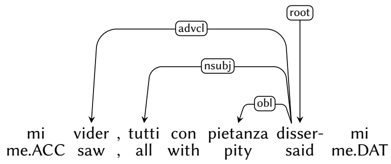

Figure 3: Gold syntactic tree of Sonn. 10 v.10

As evident from the comparison of the trees, assigning the `root` to an incorrect token triggers a domino effect, resulting in both attachment errors (e.g., the subject relation (`nsubj`) of *tutti* (all) and the oblique relation (`obl`) of *pietanza* (pity), which are incorrectly attached to *vider* (saw) instead of *disser* (said) and labeling errors (e.g., *disser* (said) is mistakenly labeled as `parataxis` instead of being correctly identified as the `root`). Interestingly, when this sentence is removed and the evaluation is repeated, the averages increase to UAS 75.31 and LAS 65.43, which are more or less in line with the overall average (refer to Table 3).

In line with the previously observed error, the analysis of the second worst-annotated sonnet reveals another issue concerning `root` attribution. Specifically, the LLM erroneously assigns `root` status to the auxiliary verb *avere* (to have) (*ài* in the text), which is inconsistent with the UD formalism[24]. Figure 4 shows the erroneous output generated by the LLM, alongside the gold annotation, as produced by both the Dante model and the human annotator.[25]

[24]Refer to: https://universaldependencies.org/u/dep/aux_.

[25]Dante model correctly annotates the tree, identifying *piena* (full) as the `root` and correctly attaching its dependents. The only mistake made by the model is the labeling of the auxiliary *ài* (to have), which is incorrectly assigned the label `cop` (copula) instead of the correct label `aux` (auxiliary).

Example III - Sonn. 31 v. 1
Tu m'ài sì piena di dolor la mente
(you have filled my mind with so much sorrow)

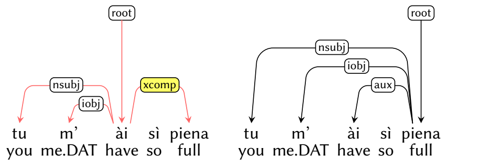

Figure 4: LLM and gold dependency trees of Sonn. 31 v.1

More specifically, in such structures, the `root` dependency relation should be assigned to *piena* (full), as correctly handled by Dante model. By misassigning the `root`, and incorrectly labeling the actual `root` as `xcomp` (open clausal complement), the model also produces erroneous attachments. For example, the subject *tu* (you) and the pronoun *mi* (me), marked as `iobj` (indirect object), are correctly labeled but incorrectly attached.

Alongside the auxiliary *avere* (to have) discussed in Example III, the auxiliary *essere* (to be) also proves problematic. In particular, the LLM incorrectly assigns head status also to the auxiliary *essere*, *è* in the text, (is), as shown in Example IV and Figure 5. As previously noted, in the UD formalism, auxiliaries are treated as leaf nodes rather than heads. By contrast, Dante model performs correctly, assigning the auxiliary the `cop` (copula) dependency relation and attaching it as a dependent of the noun *vita* (life).

Example IV - Sonn. 31 v. 9
Io vo come colui ch'è fuor di vita
(I am unmanned and have to wander through the world like an intricate figure)

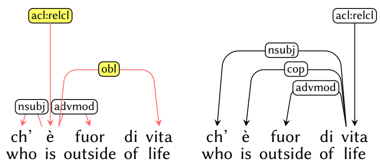

Figure 5: LLM and gold dependency trees of Sonn. 31 v.9

The incorrect identification of the clause head, assigned to the copula *è* (is), leads to a structurally inconsistent analysis, accompanied by further errors. For instance, the noun *vita* (life) is incorrectly labeled as `obl` (oblique), instead of being identified as the head of a relative clause (`acl:relcl`), a role mistakenly assigned to the copula *è* (is). In addition to the incorrect labeling, the subject *che* (who) (marked as `nsubj`) and the adverb *fuori* (out) (`advmod`) are also wrongly attached.

In sum, this close examination of the two lowest-scoring sonnets highlights how a single error, especially one involving a crucial dependency relation from which other subtrees depend, such as the incorrect identification of the `root`, can compromise both annotation quality and evaluation scores. Auxiliaries also pose a particular challenge for the LLM, as they are often incorrectly annotated as heads rather than leaves.

4.2. Dante Model Errors

During the evaluation of the models' performance, Sonnet 29 emerges as a particularly problematic case, in which both Dante models (All and OS) perform poorly, showing no substantial improvement over the LLM's performance.

Upon comparing the gold annotation with the output of the Dante-All model, it becomes immediately evident that one of the main reasons behind this unexpectedly low performance lies in an incorrect token split of the word *angosciosi* (anguished), at verse 5:

> Examples V - Sonn. 29 v. 5
> *angosciosi diletti miei sospiri ... giriano*
> *angoscio si diletti miei sospiri ... giriano*
> (my sighs would not only subside but turn into hosannas of praise [...] gives)

angosciosi	*diletti*	*miei*	*sospiri*	...	*giriano*	
34	35	36	37	...	60	
angoscio	*si*	*diletti*	*miei*	*sospiri*	...	*giriano*
34	35	36	37	38	...	61

Table 4
Tokenization and ID alignment in Sonn. 29, v. 5 in the gold annotation and the Dante-All model, respectively.

The Dante-All model splits the word into two distinct token: *angoscio* and *si*, interpreting *si* (originally part of the adjective *angosciosi*) as a reflexive clitic. As a result, it is annotated as a separate token. As already observed for the LLM in 3.2, this error brings to light a range of phenomena that, while not entirely incorrect, are influenced by issues stemming from improper tokenization. In fact the addition of a token leads to a misalignment of token indices (i.e., the *id* field in the CoNLL-U format; see footnote 3), such that even when syntactic dependencies are correctly labeled and attached to the correct token, they are still considered incorrect due to the wrong numbering with respect to the gold standard (see Table 4).

Indeed, this error can only be observed in the Dante-All model, which is required to perform all annotation tasks independently, including tokenization. Despite this specific tokenization issue, there are also errors common to both models (Dante-All and Dante-OS). One such error is the incorrect identification of the `root`, as shown in Figures 7 and 8, in contrast to the gold standard shown in Figure 6. This error is reported here due to its significant impact on the overall sentence structure. More specifically, as indicated in the gold annotation (Figure 6), the `root` of the sentence is *giriano* (wander), token 60. Nevertheless, both Dante models fail to recognize it, selecting instead *diletti* (delights), token 35, as the `root`.

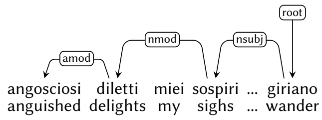

Figure 6: Gold dependency tree of Sonn. 29 v. 5

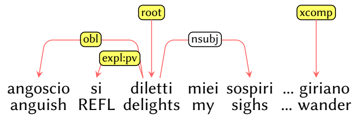

Figure 7: Dante-All dependency tree of Sonn. 29 v. 5

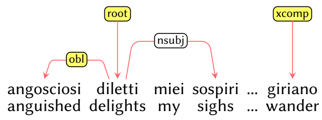

Figure 8: Dante-OS dependency tree of Sonn. 29 v. 5

Following this error, subsequent dependency relations are misassigned. The adjective *angosciosi* is annotated as an oblique (`obl`) dependent of *diletti*. In the Dante-All model, this is further complicated by incorrect tokenization, which results in *si* being erroneously split off and assigned the `expl:pv` (expletive: pronominal) relation.

The noun *sospiri*, although correctly labeled as a subject (`nsubj`), is attached to the incorrect `root` (*diletti*) rather than to *giriano* (the gold `root`). Finally, *giriano* itself is annotated as an open clausal complement (`xcomp`) depending on another token.

We hypothesize that this error may be caused by the length of the sentence, in which the token identified as the `root` appears in the 60th position, thus making the parsing of the sentence more complex. However, further experiments in this direction are needed to confirm this hypothesis.

4.3. Errors of Both Models

While performing the manual error analysis, we identified a set of specific errors shared by both models (Dante(-

All and -OS) and and LLM), highlighting a common difficulty in analysing such constructions. Notably, these errors both arise from and reflect the complexity of annotating these texts. We report two cases.

In Sonnet 10, both models fail to correctly annotate a construction that poses difficulties even for human annotators. This issue is found specifically in verses 10–11.

> Example VI - Sonn. 10 vv. 10-11:
> *in una parte là 'v'i' trovai gente /*
> ***che** ciascun si doleva d'Amor forte.*
> (to a place where noblemen gathered who also suffered in the thrall of Love.)

In v. 11, the *che* (who) is a specific instance of the "neutrum relative pronoun *che*" [27, p.193], used to replace a pronoun governed by a preposition. This construction was typical of familiar Tuscan and is also attested in other texts of the same period [27, pp.193-194]. As noted by the editors Rea and Inglese in their commentary on this verse [19, p.62], this instance of *che* represents a case of preposition elision, namely, standing in for the form *dei quali*, (of whom), which functions as a nominal modifier of the actual subject *ciascun* (each) (*ciascun dei quali*) (each of whom). The gold annotation is reported in Figure 9.

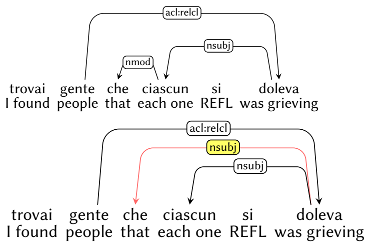

Figure 9: Gold dependency tree and both model outputs for Sonnet 9, vv. 10-11, respectively.

Both models misanalyse *che*, incorrectly identifying it as the subject of a relative clause. At the same time, they correctly assign the `nsubj` relation to *ciascun* (each); however, this leads to an inconsistent syntactic analysis, resulting in a spurious double subject, as shown in Figure 9.

Another error made by both models that is worth commenting on, as it reflects a stylistic peculiarity, is the one reported in Example VII (Sonnet 10, v. 13):

> Example VII - Sonn. 10 v. 13:
> *Fatto sè di tal servente*
> (You are a denizen of the kingdom or the dungeon)

This verse presents a fronting of the genitive complement *di tal* (of such), as noted by Rea and Inglese [19, p.63], along with an ellipsis of the noun *donna* (woman), that is, *di tal (donna)*. As a result, the preposition *di* (of) should be attached as a case marker (`case`) to *tal*, and the entire phrase *di tal* should be analysed and annotated as a nominal modifier (`nmod`) of the noun *servente* (servant), yielding the interpretation *servente di una tale donna* (servant of such a woman). This full phrase functions as an open clausal complement (`xcomp`) of the main verb *fatto* (made). We report the correct syntactic tree in Figure 10.

In this case, both models incorrectly annotate *di* and *tal* as modifiers of the noun *servente*, assigning them the labels `case` and `det` (determiner), as shown in Figure 10. Neither model captures the ellipsis of *donna*, and both attach *di* (of) and *tal* (such) directly to *servente*. However, this annotation leads to a different interpretation: *di tale servente* (of such a servant). Based on this incorrect interpretation, *servente* is assigned the syntactic role of oblique (`obl`), and is correctly attached to the verb *fatto*.

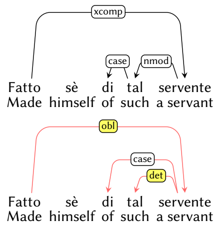

Figure 10: Gold dependency tree and both model outputs for Sonnet 10, verse 13, respectively.

These two examples, both incorrectly annotated by the models, underscore the importance of human revision in cases involving stylistic varieties that current models fail to adequately capture.

5. Conclusion and Future Work

This study has evaluated the syntactic parsing performance of two distinct models on a corpus of Old Italian poetry: Stanza, a supervised neural model trained on a domain-specific corpus, and an unsupervised autoregressive large language model accessed via the ChatGPT API (o3 version). The results show that while the LLM demonstrates acceptable accuracy scores, the Stanza-based model trained on the Italian-Old treebank consistently outperforms it in syntactic annotation tasks, particularly when provided with gold-standard tokenization and morphology. Although this experiment involved only a single LLM , the results suggest that using a neural model trained on domain-specific data to pre-parse texts as a basis for human annotation could be advantageous. Moreover, while this may appear self-evident, the experiment reinforces the importance of involving expert annotators, especially in cases like those shown in this study, where stylistic nuances escape the models. Future directions include conducting experiments with other LLMs, such as the aforementioned UDify [23] and U-DepPLLaMA [24], as well as testing custom transformer-based pipelines using MaChAmp [28] and Trankit [29]. Manual error analysis was crucial in highlighting the challenges encountered by both models, revealing not only more mechanical errors, such as incorrect `root` identification often accompanied by mistaken head and dependency assignments of the dependent nodes, but also more specific issues, particularly in handling poetic constructions, ellipses, and auxiliary verbs. While in the first scenario some workarounds can be found to mitigate such errors (for example, by adopting rule-based integrations to guide model performance), when it comes to more specific errors, human skill and expertise become decisive.

Another crucial element that emerged from this study is the influence of editorial choices on syntactic annotation. As demonstrated in Section 2.3, different editions of the same text can result in substantial variation in tokenization, lemma interpretation, and syntactic structure. This observation underscores the need for openly accessible, high-quality digital editions, particularly for historical texts, which often lack standardized resources.

Future work will extend the evaluation to include other critical editions of Cavalcanti's *Rhymes* to analyse stylistic distance and to assess model performance across distinct editorial variants. Moreover, further investigations will also aim to explore syntactic variation across different poetic genres and authors within the same historical period, as well as to examine prose texts in order to assess whether significant syntactic differences emerge.

References

[1] W. Croft, Radical Construction Grammar: Syntactic Theory in Typological Perspective, Oxford University Press, Oxford, 2001. URL: https://doi.org/10.1093/acprof:oso/9780198299554.001.0001. doi:10.1093/acprof:oso/9780198299554.001.0001, online edition, Oxford Academic, 1 Sept. 2007.

[2] M.-C. De Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308.

[3] A. Ezquerro, C. Gómez-Rodríguez, D. Vilares, Better Benchmarking LLMs for Zero-Shot Dependency Parsing, in: R. Johansson, S. Stymne (Eds.), Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), University of Tartu Library, Tallinn, Estonia, 2025, pp. 121–135. URL: https://aclanthology.org/2025.nodalida-1.13/.

[4] B. Lin, X. Zhou, B. Tang, X. Gong, S. Li, ChatGPT is a Potential Zero-Shot Dependency Parser, 2023. URL: https://arxiv.org/abs/2310.16654. arXiv:2310.16654.

[5] M. Straka, J. Straková, Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, in: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 88–99. URL: http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

[6] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[7] C. Corbetta, M. C. Passarotti, F. M. Cecchini, G. Moretti, Hell Awaits: Building a Universal Dependencies Treebank for Dante Alighieri's Comedy, Italian Journal of Computational Linguistics (IJCOL) 11 (2025) 21–46. URL: https://hdl.handle.net/10807/319736. doi:10.17454/IJCOL111.02.

[8] C. Corbetta, M. Passarotti, F. M. Cecchini, G. Moretti, Highway to Hell. Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 154–161. URL: https://aclanthology.org/2023.clicit-1.20/.

[9] C. Corbetta, M. Passarotti, G. Moretti, The Rise and Fall of Dependency Parsing in Dante Alighieri's Divine Comedy, in: Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024, 2024, pp. 50–56.

[10] C. Corbetta, G. Moretti, M. Passarotti, Join Together? Combining Data to Parse Italian Texts (2024).

[11] M. Marti, Storia dello Stil Nuovo, Collezione di Studi e Testi, Milella, Lecce, 1972.

[12] 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.

[13] M. Tavoni, DanteSearch: il Corpus delle Opere Volgari e Latine di Dante Lemmatizzate con Marcatura Grammaticale e Sintattica, in: Lectura Dantis

2002-2009. Omaggio a Vincenzo Placella per i suoi settanta anni, volume 2, Università degli Studi di Napoli "L'Orientale", Il Torcoliere-Officine, 2012, pp. 583–608.

[14] G. Cavalcanti, [Le Rime di Guido Cavalcanti], 1st. ed., N. Zanichelli, 1902. Editor: Ercole Rivalta.

[15] P. Beltrami, et al., La Metrica Italiana, il Mulino, 1991.

[16] G. Contini (Ed.), Rime, Poeti del Duecento, vol. II, Ricciardi, Milano; Napoli, 1960. URL: https://catalog.hathitrust.org/Record/101904107, 1ª ed. 1957; 2ª ed. 1970.

[17] M. Ciccuto (Ed.), Rime, 5 ed., Rizzoli, Milano, 1998. Introduzione di Maria Corti.

[18] D. D. Robertis (Ed.), Rime, Einaudi, Torino, 1986. Edizione critica.

[19] R. Rea, G. Inglese (Eds.), Rime, Carocci, Roma, 2011. Edizione commentata.

[20] G. Cavalcanti, The Metabolism of Desire: The Poetry of Guido Cavalcanti, AU Press, Edmonton, Alberta, 2012. Print.

[21] J. L. Fleiss, Measuring Nominal Scale Agreement among many Raters, Psychological Bulletin 76 (1971) 378–382. doi:10.1037/h0031619.

[22] OpenAI, GPT-4 Technical Report, https://openai.com/research/gpt-4, 2023. URL: https://doi.org/10.48550/arXiv.2303.08774, accessed via the ChatGPT API.

[23] D. Kondratyuk, M. Straka, 75 Languages, 1 Model: Parsing Universal Dependencies Universally, arXiv preprint arXiv:1904.02099 (2019).

[24] C. D. Hromei, R. Basili, U-DepPLLAMA: Universal Dependency Parsing via Auto-regressive Large Language, IJCoL-Italian Journal of Computational Linguistics vol. 10, n. 1 June 2024 (2024) 21.

[25] 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.

[26] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The Curious Case of Neural Text Degeneration, arXiv preprint arXiv:1904.09751 (2019).

[27] G. Rohlfs, Grammatica Storica della Lingua Italiana e dei suoi Dialetti: Morfologia, Turin: Einaudi (1968).

[28] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP, in: D. Gkatzia, D. Seddah (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 176–197. URL: https://aclanthology.org/2021.eacl-demos.22/. doi:10.18653/v1/2021.eacl-demos.22.

[29] M. V. Nguyen, V. D. Lai, A. Pouran Ben Veyseh, T. H. Nguyen, Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing, in: D. Gkatzia, D. Seddah (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 80–90. URL: https://aclanthology.org/2021.eacl-demos.10/. doi:10.18653/v1/2021.eacl-demos.10.

A. Appendix

Sonnet 31[26]

1 *Tu m'ài sì piena di dolor la mente*
che l'anima si briga di partire,
e li sospir che manda il cor dolente
mostrano a li occhi che non pon soffrire.

5 *Amor, che lo tuo grande valor sente,*
dice: — mi duol che ti convien morire
per questa fera donna, che neente
par che pietade di te voglia udire.

9 *Io vo come colui ch'è fuor di vita,*
che pare, a chi lo sguarda, c'omo sia
fatto di rame o di pietra o di legno,

12 *che sè conduca sol per maestria,*
e porti ne lo core una ferita
che sia, com'egli è morto, aperto segno.

1 You have filled my mind with so much sorrow
that the soul itself is assaulted and tries to flee.
Heartsick, my body sighs from my bones' marrow
and I've reached my limit, as anyone can see.

5 Even Love is sympathetic and says,
"It is hard that the cruel lady for whom you pine
gives you no pitiful glance or comforting phrase.
This was never a part of my design."

9 I am unmanned and have to wander through
the world like an intricate figure of wood or brass
produced by some toymaker to amuse.

12 Strangers who pause to stare at me as I pass
can't tell that I suffer and haven't a clue
that I am dead, a victim of her abuse.

Sonnet 10

1 *Li miei foll'occhi, che prima guardaro*
vostra figura piena di valore,
fuor quei che di voi, donna, m'acusaro
nel fero loco, ove ten corte amore.

5 *E mantenente avanti lui mostraro*
ch'io era fatto vostro servidore;
perchè sospiri e dolor mi pigliaro
8vedendo che temenza avea lo core.

9 *Menarmi tosto senza riposanza*
in una parte, là 'v'i' trovai gente
che ciascun si doleva d'amor forte.

12 *Quando mi vider, tutti con pietanza*
dissermi: — fatto se' di tal servente
che mai non dei sperare altro che morte. —

1 It was my reckless eyes that first beheld
your ineffable worth and condemned me to
live in that wasteland over which the bold
master, Love, holds court as tyrants do.

5 They welcomed me there, a new captive, a slave,
and were delighted to hear my sighs and groans.
We are taught as little boys to try to be brave,
but I felt an icy fear deep in my bones.

9 They led me to a place where noblemen
gathered who also suffered in the thrall
of Love. Their somber welcome to me I

12 still shudder at: "You are a denizen
of the kingdom or the dungeon that holds us all
and from which there is no escape until you die."

Sonnet 29

1 *Se mercè fosse amica a' miei disiri*
e 'l movimento suo fosse dal core
di questa bella donna e 'l suo valore
mostrasse la virtute a' miei martiri,

5 *d'angosciosi diletti miei sospiri,*
che nascon de la mente ov'è amore
e vanno sol ragionando dolore
e non trovan persona che li miri,

9 *giriano a gli occhi con tanta vertute*
che 'l forte e duro lagrimar che fanno
ritornerebbe in allegrezza e 'n gioia.

12 *Ma sì è al cor dolente tanta noia*
e a l'anima trista è tanto danno
che per disdegno uom non dà lor salute.

1 If luck could look with favour on my desire
and if it came from my lady's heart with the power
to encourage me and let me thrive in a shower
of hope from heaven that knows how to admire

5 devotion of any kind, I do believe
my sighs would not only subside but turn into
hosannas of praise as grey brightens to blue
when angry weather grants us a reprieve.

9 My squalls of tears would cease and joy at last
would be what gives my eyes their special shine,
each teardrop like a jewel delighting to be

12 dug from the earth to glitter and be free ...
But it hasn't happened, and the pain that has been mine
deforms me so that I am an oucast.

[26] See footnote 7 for details on the translation edition.

Correspondence between sonnet numbering and the first verse of each sonnet:

The poems composed before 1290	
1	Certe mie rime a te mandar volendo
2	Dante ai poeti
3	Vedeste, al mio parere, ogni valore
4	Se vedi amore assai ti prego, Dante
5	Avete 'n vo' li fiori e la verdura
6	Chi è questa che ven ch' ogn'om la mira
7	Beltà di donna di piagente core
9	Io vidi li occhi dove Amor si mise
10	Li miei folli occhi che prima guardaro
13	Dante a Guido
14	S'io fossi quelli che d'amor fu degno
16	Dante, un sospiro messagier del core
17	Sonetto dell'Orlandi
19	La bella donna dove amor si mostra
20	Guido Orlandi a Guido
21	Amore e monna Lagia e Guido ed io
23	L'Orlandi a Guido
24	Di vil matera mi conven parlare
25	L'Orlandi a Guido
The poems of uncertain date	
26	Un amoroso sguardo spirituale
27	Voi che per li occhi mi passaste 'l core
28	Perchè non furo a me gli occhi dispenti
29	Se mercè fosse amica a' miei disiri
30	L'anima mia vilment'è sbigotita
31	Tu m'ài sì piena di dolor la mente
32	S'io prego questa donna che pietade
33	Io temo che la mia disaventura
34	Certo non è de lo 'ntelletto accolto
35	Veder poteste quando vui scontrai
36	De! spiriti miei, quando mi vedete
37	Pe' gli occhi fere un spirito sottile
38	A me stesso di me pietate vene
39	Gianni Alfani a Guido
The poems composed after 1290	
49	Io vengo il giorno a te infinite volte
50	Una figura de la Donna mia
51	Guido Orlandi a Guido
52	Una giovane donna di Tolosa
54	O tu che porti ne li occhi sovente
55	Donna mia non vedestu colui
56	Noi sian le triste penne isbigotite
57	Novelle ti so dire, odi, Nerone
59	Farinata degli Uberti a Guido
60	Se non ti caggia la tua Santalena
61	Guata, Manetto, quella scrignotuzza

Prompt LLM: The prompt was given in Italian; the original version is shown in italics, with the English translation provided in parentheses.
role user: *dato il testo in formato txt, produci un'annotazione completa secondo lo standard di Universal Dependencies. Produci un file in formato CoNLL-U, assicurandoti che abbia 10 colonne.* (Given a plain text file (.txt), produce a complete annotation according to the Universal Dependencies standard. Generate a CoNLL-U formatted file, ensuring that it includes all 10 required columns.)
role user: *produci l'annotazione eseguendo i task di tokenizzazione, lemmatizzazione, part of speech tagging, morphological features e dependency parsing.* (Perform the annotation by carrying out the tasks of tokenization, lemmatization, part-of-speech tagging, morphological feature annotation, and dependency parsing.)
role user content: "file: "file_raw (raw sonnet)"
role user assistant: "file: "file_gold (gold sonnet)"
role user content: *assicurati che non ci siano dei cicli nella sintassi e che ogni frase abbia soltanto una root.* (Ensure that the syntactic structure contains no cycles and that each sentence has exactly one root.)
role user content: *assicurati anche che tutti i nodi dell'albero sintattico siano raggiungibili fra di loro.* (Also ensure that all nodes in the syntactic tree are mutually reachable.')
role user content: *procedi con l'altro file. Assicurati di annotare la colonna 7 e 8 e di avere 10 colonne per linea.* (Proceed with the other file. Make sure to note columns 7 and 8 and to have 10 columns per line.)

B. Online Resources

- Italian-Old,
- Cavalcanti Repository.