«Are you Afraid of Ghosts?» A Proposal for Busting Predicate Ellipsis in Universal Dependencies

Claudia Corbetta
Università di Bergamo-Pavia
via Salvecchio 19,
24129 Bergamo, Italy.
claudia.corbetta@unibg.it

Federica Iurescia and **Marco Passarotti**
Università Cattolica del Sacro Cuore
CIRCSE Research Centre
Largo Gemelli, 1, 20123 Milan, Italy
{federica.iurescia,marco.passarotti}@unicatt.it

Abstract

This paper addresses the representation of ellipsis in dependency syntax, proposing both a theoretical and a practical workflow for its analysis and annotation in treebanks, following the state-of-the-art Universal Dependencies framework. We discuss the challenges of annotating ellipsis, with a focus on predicate ellipsis and its representation in dependency treebanks, and emphasize the importance of accounting for such phenomena for syntactic analysis and machine learning applications. We present a case study based on the Italian-Old treebank, demonstrating the applicability of the proposed workflows and invite the community to participate in this initiative with their own languages.[1]

1 Introduction

A widely acknowledged principle in science is that the manner in which we choose to represent reality significantly shapes the nature of the material we aim to structure and interpret (Kuhn, 1962). This is particularly true in the analysis of ellipsis, a syntactic phenomenon that represents the omission of linguistic material in a sentence (Merchant, 1999), where the choice of the model to represent and encode (missing) linguistic information has a crucial impact on both its study and interpretation.

Studying a phenomenon that represents, by its nature, an absence is a challenging task. Merchant (Merchant, 2018, p.25) draws a compelling comparison of ellipsis with a black hole, stating that "detecting and arguing for such "missing" structures is analogous to searching for and determining the properties of a black hole: one can tell it's there only by its effects on surrounding material". Due to this inherently elusive nature, since ellipsis is, by definition, silent in the data, an additional challenge arises in representing it within syntactically annotated corpora (treebanks), which are designed, among other purposes, to support the representation and queryability of syntactic phenomena.

In the present paper, we address the representation of ellipsis in dependency syntax by providing both a theoretical and a practical workflow for analyzing and representing it in treebanks, in accordance with the state-of-the-art dependency framework, Universal Dependencies (De Marneffe et al., 2021).

The paper is structured as follows: in Section 2, we provide a definition of ellipsis and outline key concepts. Section 3 offers an overview of ellipsis within syntactic theory, with particular attention on dependency frameworks. It also presents how ellipsis is addressed in various dependency treebanks, with a specific focus on Universal Dependencies in Subsection 3.1. In Section 4, we discuss the importance of annotating ellipsis. Sections 5 and 6 introduce the theoretical and practical workflows, respectively, along with the challenges they entail. Finally, Section 7 presents a case-study based on the Italian-Old treebank, and Section 8 concludes the paper.

2 What is Ellipsis?

Ellipsis has been defined as an asymmetry between meaning and form in an expression (Van Craenenbroeck and Temmerman, 2018).

In this Section, we will address some key concepts about ellipsis, in order to provide terminology and insights of the phenomenon. When speaking of ellipsis, we should consider the following aspects:

- **elided site**: the elided site refers to the position of the ellipsis, namely it represents the gap in the sentence where the linguistic mate-

[1]This paper is the result of a collaboration among all three authors. In accordance with the requirements of the Italian academic attribution system, Claudia Corbetta is responsible for the entire paper. Claudia Corbetta, Federica Iurescia, and Marco Passarotti are specifically responsible for Section 6.

Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025), pages 52–63
August 28-29, 2025 ©2025 Association for Computational Linguistics

rial is omitted. It is usually represented with "___" in the sentence, whereas in the syntactic structure it can be represented with an empty node.

- **remnants**: the remnants are the "survivors" in an elliptical sentence (Ortega-Santos et al., 2014, p. 55). This term refers to the linguistic material that is not elided in a clause that presents an ellipsis.
- **elided material**: it refers to the linguistic material that is omitted, and therefore that undergoes to ellipsis. In the literature, it is usually represented inside square brackets [].[2]
- **antecedent**: the antecedent is the linguistic material that leads the speaker/reader to understand and correctly process the ellipsis. It can be explicitly expressed in the sentence, or in the text, but it can also be inferred from world knowledge, i.e., not explicitly stated, but still recoverable by the listener/reader. In the literature, the term antecedent is always used in a broader way, not obligatory referring to an element that precedes the ellipsis site. In fact, the antecedent can also follow the ellipsis site, thus making more appropriate the term "postcedent" (McShane, 2005, p. 14).[3]
- **identity condition**: the identity condition has been debated by several scholars. It refers to the identity between the antecedent and the elided material. We will show in Section 6, that it is not always the case.

In Example 1, we provide an instance of ellipsis, highlighting all the aspects shown above:

Example 1
"I wish all happy holidays, and moreso,
___ peace on earth."[4]

The ellipsis site is marked with the "___". The remnants are "and", "moreso", "peace on earth". The antecedent that lead as to solve the ellipsis is composed by the subject "I", the verb "wish", the beneficiary "all", and, in this case, the elided material respects the identity condition, being a copy of the antecedent.[5] Therefore, the sentence with the elided material expressed will be as follows:

"I wish all happy holidays, and moreso,
[I wish all] peace on earth."

3 Ellipsis in Syntax and in Treebanks

From a **theoretical syntactic perspective**, studies on ellipsis have been conducted mainly within constituency frameworks (Ross, 1969; Merchant, 2001; Kennedy, 2003), which significantly outnumber those grounded in dependency syntax. Constituency-based analysis of ellipsis tends to provide a broad classification of ellipsis types,[6] primarily grounded in the notion of constituent movement within the syntactic structure.

However, within the dependency framework, ellipsis has not been extensively analyzed, positioning it as a relatively underexplored syntactic phenomenon. Among the main studies on ellipsis from a dependency perspective, Osbourne (Osborne, 2019) offers a key contribution, namely the identification of the *catena*, a syntactic unit different from the constituent. His analysis and classification of ellipsis build and provide justification of licensing different type of ellipsis. Ellipsis has also been defined as "unrealized words" by Hudson (Hudson, 2010), referring to covert words that lack pronunciation or spelling, with their unique distinction from overt ones being their inaudibility.

Regarding the representation of **ellipsis in treebanks**, approaches vary depending on the formalism adopted. In constituency-based treebanks, such as the Penn Treebank (PTB) (Marcus et al., 1993) and the BulTreeBank (Osenova and Simov, 2003), which follows the Head-Driven Phrase Structure Grammar (HPSG) formalism (Pollard and Sag, 1994), ellipses are explicitly annotated through the use of empty nodes.

Conversely, in dependency treebanks, the representation of ellipsis poses a greater challenge. This

[2]We will not address the question of whether the elided material exists at a cognitive level. For a general overview of how psycholinguistics addresses questions concerning the representation of sentences involving ellipsis, see Phillips and Parker (2014). However, it is clear that, in order to analyze ellipsis, it must be made explicit.

[3]We will align with the literature (McShane, 2005, p.14) and use the term antecedent as a macro-term that is not connected with its position with respect to the ellipsis site.

[4]This sentence (ENG_20041111_173500-0051) is taken from the UD_English-EWT. See: https://github.com/UniversalDependencies/UD_English-EWT.

[5]Within the UD framework, such an example is treated as ellipsis and annotated with the orphan relation (see Section 3.1 and 3.2 for further discussion of orphan). The presence of the adverb *moreso* provides clear evidence of a missing verb, which would constitute its syntactic head.

[6]For an overview on ellipsis classification, see Van Craenenbroeck and Temmerman (2018).

is primarily due to the principle that, in dependency-based annotation, the number of nodes in the tree corresponds exactly to the number of tokens in the sentence, thereby precluding the use of empty nodes. For instance, the Prague Dependency Treebank (PDT) (Hajič et al., 2017) addresses ellipsis explicitly through a dedicated attribute and requires its reconstruction at other annotation layers.[7] In the following Subsections 3.1 and 3.2, we address in detail the formalism adopted by the state-of-the-art dependency framework, namely Universal Dependencies.

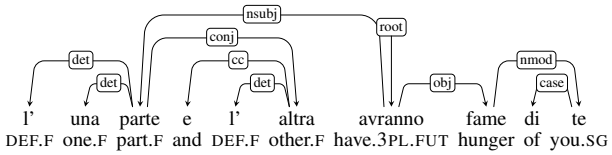
3.1 Ellipsis in Basic Universal Dependencies

When it comes to syntactic dependency resources, the state-of-the-art framework is Universal Dependencies (henceforth UD) (De Marneffe et al., 2021), which currently includes 319 treebanks covering 179 languages.[8]

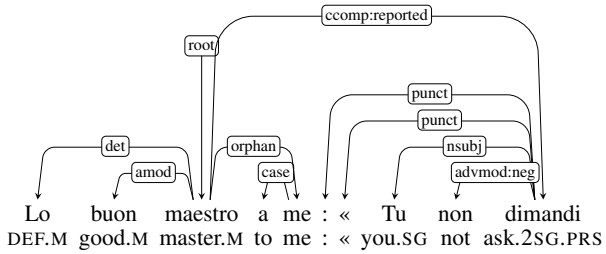
UD provides two levels of syntactic annotation: a basic layer and an enhanced one, called Enhanced Dependencies. The basic annotation includes only overt words (i.e., non-null nodes) and therefore does not allow for empty nodes.[9] As a result, ellipsis is not explicitly represented in the basic layer. Instead, when annotating elliptical structures, the UD guidelines recommend either adopting a promotion strategy,[10] if the syntactic structure remains grammatically well-formed, or using the dependency relation `orphan` when promotion would result in an ungrammatical configuration.

Example 2 illustrates the use of the promotion strategy in an Old Italian sentence from the Italian-Old treebank, whereas Example 3 provides an instance of the `orphan` relation:

Example 2 - *Inf.* XV, vv. 71-72
l'una parte e l'altra avranno fame/ di te
"one party and the other will be hungry/ for you"[11]



Example 3 - *Inf.* IV, v. 31
Lo buon maestro a me:/ «Tu non dimandi (...)
"To me the Master good:/ "Thou dost not ask (...)"



In Example 2, the node *altra* "other" is promoted to the head of the conjunct `conj`, which involves a case of nominal ellipsis, with the noun *parte*, "part" being elided. By contrast, in Example 3, applying the promotion strategy would have resulted in an ungrammatical syntactic structure. Therefore, the noun *maestro* "master" is promoted to the root of the sentence, and the prepositional phrase *a me* "to me", which functions as an oblique, is attached to the nominal root using the `orphan` relation.

As the evidence shows, neither of the proposed solutions directly addresses the annotation of ellipsis. Instead, they offer workarounds that attempt to accommodate elliptical constructions within the annotation scheme and without the construction of empty nodes.

3.2 Ellipsis in Enhanced Universal Dependencies

However, empty nodes are permitted in **Enhanced Dependencies** (henceforth EUD). EUD is an extension of basic UD, designed to make «some of the implicit relations between words more explicit».[12] Among these explicit relations, ellipsis is also addressed. Specifically, as outlined in the guidelines,[13] predicate ellipsis permits the insertion of an empty node, allowing for the restoration of syntactic relations that would otherwise be lost in the annotation. Concerning the annotation of the empty node, information on form, lemma, and

[7] See Mikulová (2014) for further discussion of ellipsis in the PDT.

[8] These numbers refer to version 2.16. See: https://universaldependencies.org.

[9] In this work, we will use the terms "empty nodes" rather than "null nodes". Although these terms are often used interchangeably in the UD guidelines, in linguistic literature null node is frequently associated with valency-related phenomena, such as null subjects or objects. Accordingly, the term "empty node" is adopted here, as it more accurately reflects the syntactic nature of the structure.

[10] See the guidelines: https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis.

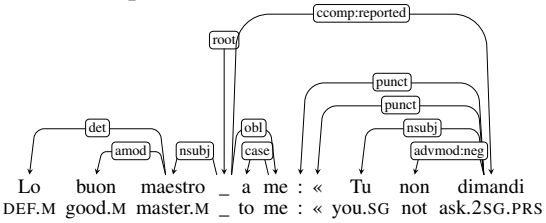
[11] The English translations of the examples from the Comedy are by Allen Mandelbaum, available at: https://digitaldante.columbia.edu/dante/divine-comedy/.

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

[13] https://universaldependencies.org/u/overview/enhanced-syntax.html#ellipsis.

UPOS, XPOS and features is optionally "copied from the overt occurrence of the predicate".[14]

We present the enhanced annotation of the sentence in Example 3.[15]



The graph represents the empty node, visually indicated by the underscore "_" in the sentence. Meanwhile, the noun *maestro* "master" is now attached to the root with the subject relation (`nsubj`), and the prepositional phrase and the complement clause depend on the root as an oblique (`obl`) and a clausal complement (`ccomp:reported`), respectively.

Even though EUD is crucial for analyzing phenomena like ellipsis, unfortunately, there are few treebanks that provide enhanced annotation,[16] resulting in a scarcity of gold-standard data for the phenomena specifically addressed by EUD.

In the following Sections 4, 5, and 6, we will introduce and justify the need for gold-annotated data for ellipsis, and propose possible methods to address this gap. More specifically, in Section 4, we will describe the challenges in automatically enhancing ellipsis. In Section 5, given the unexplored nature of the topic from a corpus-based perspective, we will suggest a general approach to investigate ellipsis in this context, also discussing the challenges encountered in doing so. In Section 6, we will specify the approach within the UD framework to provide a method for enriching the annotation of this valuable dataset.

4 Why do we Need Ellipsis?

As mentioned in Section 3.1, only a small amount of enhanced treebanks is available. Providing an enhanced version is not mandatory in UD, and in most cases, the effort required to develop and maintain a basic treebank is already substantial enough to relegate the manual creation of an enhanced layer to a lower priority. Additionally, the task of automatically enhancing a basic UD treebank is particularly challenging, especially when it comes to generating empty nodes to account for ellipsis (Simi et al., 2018; Droganova et al., 2018).

Moreover, attempts to automatically parse ellipsis in dependency treebanks using large language models (LLMs) have not yielded satisfactory results (Ćavar et al., 2024a,b), which has led to the development of corpora specifically focused on ellipsis.[17]

While it is clear that, at the current state of the art, a fully automatic reconstruction of ellipsis is not yet feasible, it is still possible to support the human annotation process by designing workflows that facilitate the manual annotation of ellipsis.

Before describing the procedures that assist manual annotation, in Section 5, we will briefly outline the general steps that are essential for dealing with ellipsis. Specifically, in this exploratory study, we focus on predicate ellipsis.[18]

5 "Ghostbusting" Ellipsis: A Theoretical Approach

Before moving on to a practical proposal for how to address ellipsis in UD (see Section 6), we will first outline the general workflow required when dealing with ellipsis, namely with something that is absent from the data. We divide this workflow into three steps: detection, identification, and reconstruction. It is important to clarify that the proposed workflow - along with the order in which the steps are presented - is intended as a methodological proposal for handling ellipsis in annotated data. In other words, it does not address how ellipsis is processed in the mind of the reader or speaker - particularly with respect to the reconstruction phase - which is the subject of specific psycholinguistic studies (Hofmeister, 2007), but lies beyond the scope of the present work. Each step is described in detail below:

- **detection**: this step refers to the detection of the ellipsis site, that is, recognizing the presence of an ellipsis in the text. While for certain linguistic phenomena the detection task is relatively straightforward, this is not the case for ellipsis, which involves an omission. In

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

[15] In Appendix A, we report the full tree with both the basic and the enhanced annotation (See Example 3.1).

[16] To date, only 19 treebanks include all types of enhancements. Moreover, EUD treebanks are difficult to quantify, as they vary in the specific enhancements they incorporate.

[17] An example is the Hoosier Corpora: https://github.com/dcavar/hoosierellipsiscorpus.

[18] We leave for future works the management of other types of ellipsis, such as nominal ellipsis.

fact, determining whether something is missing requires careful consideration, in order to avoid overgeneralizing ellipsis and identifying ellipsis sites where there are none. Particular attention should be paid to drawing a clear boundary between ellipsis and coordination. This issue will be further discussed in Section 6.

- **identification**: once the presence of an ellipsis has been detected, the next step is to concretize it by retrieving its antecedent, which may be found in the same sentence of the ellipsis site, elsewhere in the text, or inferred from world knowledge. This task is closely related to coreference and anaphora resolution,[19] as it involves identifying (or understanding) the linguistic material that supports the interpretation of the ellipsis - that is, its antecedent.

- **reconstruction**: the final step in dealing with ellipsis is the reconstruction of the elided material (or, in terms of syntactic structures, the empty nodes), and, if possible, the annotation of its linguistic information (Section 7.2). It is important to emphasize that this process is conceived as a practical task for ellipsis retrieval and annotation, in line with the principle that the more information we have in annotation, the better it is.

However, each of these steps raises questions and presents challenges that deserve to be addressed and discussed. We will list them in their respective order, presenting the problems in forms of questions:

- **where should the ellipsis site be placed?** The first challenge in the detection process involves the position of the ellipsis site. Determining the position of the ellipsis site can become problematic, especially when languages with a relatively free word order are concerned. We will discuss a possible solution to this issue in Section 7.

- **how to identify the antecedent?** As mentioned in Section 2, the antecedent is not always present in the sentence where the ellipsis occurs. Sometimes, it may not even be present in the text at all. While recognizing an antecedent within the text is a challenging task (especially for a machine), identifying it from world knowledge is even more difficult, raising the question of whether and how it is possible to circumscribe it.

- **what, if anything, should we reconstruct?** Since the aim is to make ellipsis explicit in order to analyze and recognize it, it is evident that reconstruction plays a crucial role in the task. However, it also presents significant challenges. The main risk is creating a cemetery of empty nodes, where the reconstruction of missing elements is exceedingly arbitrary. Therefore, it is essential to carefully consider the extent of reconstruction of the empty node, both in terms of how much we should reconstruct (i.e., where the antecedent ends) and what information about the reconstructed node should be reported in the empty node.

We will suggest possible solutions to some of these issues in the following Sections.

6 "Ghostbusting" Ellipsis: A Practical Workflow to Deal with Ellipsis in (E)UD

Building upon the considerations in Section 5, we provide in this Section a practical workflow for addressing ellipsis in (E)UD. This Section is structured as follows: Subsection 6.1 outlines a method for querying ellipsis in basic treebanks, while Subsection 6.2 presents a new proposal for annotating ellipsis in EUD.

6.1 Detection: How to Find Something that Is Not There

The initial step in implementing ellipsis in an enhanced annotated treebank is the **retrieval** of ellipsis instances from the data. Even though, given the current state-of-the-art NLP tools, detecting ellipsis in treebanks remains primarily a task that must be carried out manually, it is still possible to develop methods that facilitate and accelerate the detection and extraction process.

This method was developed and tested on the Italian-Old treebank (see Section 7), which documents an Old Italian poem. However, the proposed method is generalizable and can be applied to other languages as well, with possible minor adjustments (See 6.2, footnote 25).

The retrieval of ellipsis in basic UD treebanks involves two steps.

[19] See, in this regard, the analysis by Hankamer and Sag (1976) on ellipsis as a form of surface anaphora.

- The first, straightforward step is to search for the **orphan** dependency relation, as this label is specifically employed in cases of (predicate) ellipsis.

- The second step focuses on identifying instances annotated using the **promotion** strategy (see Subsection 3.1),[20] where the dependent remnant of the ellipsis is promoted and assigned the dependency relation of the elided node. This second strategy relies on detecting a **mismatch** between the morphological annotation, namely the Part-of-Speech (henceforth PoS) of the remnant and the syntactic function it inherits, one that is not prototypical for its PoS. For instance, in the case of nominal ellipsis involving the promotion of an adjective, ellipsis can be identified by querying for adjectives (ADJ) that bear dependency relations typically associated with nouns (NOUN), such as subject (`nsubj`) or object (`obj`).

While the first step offers a direct and effective method for retrieving some instances of (predicate) ellipsis,[21] the second strategy aims to retrieve additional (and not explicitly annotated) cases that *may* involve ellipsis. Naturally, a manual inspection remains necessary in order to filter and evaluate genuine instances of ellipsis, distinguishing them from possible false positives.

Similar queries can be performed using available tools such as Udapi (Popel et al., 2017), GrewMatch (Bonfante et al., 2018), or ArboratorGrew (Guibon et al., 2020), among others.

In Section 7, we will provide a practical example of retrieving predicate ellipsis in the Italian-Old treebank, applying both of these strategies.

6.2 Identification and Reconstruction: Proposal for Common Annotation for Predicate Ellipsis

Once ellipsis has been detected, the next step is to mark it as such. In EUD, this involves creating an empty node and restoring the dependency relations that were lost in basic UD (see Subsection 3.1). However, as discussed in Section 4, the challenges in handling ellipsis and empty nodes have led to the lack of in-depth enhanced guidelines, specifically for ellipsis annotation.

Accordingly, this section aims to offer proposals to address proper and consistent ellipsis annotation in the UD framework, building upon the theoretical considerations outlined and addressing the issues raised in in Section 5. Specifically, we will address each issue raised in Section 5, and provide a possible solution:

- **where should the ellipsis site be placed?** we pursue an approach based on parallelism,[22] which involves mirroring the order of the phrases present in the antecedent. In cases where the antecedent is not present, we suggest following the canonical word order of the sentence.

  For instance, in the Example 1 "I wish all happy holidays, and moreso, peace on earth", the ellipsis site mirrors the order of the sentence with the antecedent: it precedes the object ("peace on earth") ("I wish all happy holidays, and moreso, [*empty node*] peace on earth"). More complex cases will be discussed in Section 7;

- **how to identify the antecedent?** As mentioned in Section 2, the identification of the antecedent is a crucial step in the task of processing ellipsis. Since this work aims to provide an annotation of ellipsis, tracking the antecedent (when present) is essential, as it provides valuable information for analyzing the phenomenon. In line with this consideration, we propose annotating the antecedent explicitly. More specifically, we suggest using the Misc column[23] to indicate whether the **antecedent** is present, or not (`Antec=Yes`; `Antec=No`) and its position in the text, identified by a **unique_number** (`AntecPosit=[unique_number]`). The unique_number has been introduced to identify nodes independently of their position in the text, since, as shown in Section 2, the

[20] For an alternative proposal for annotating ellipsis in basic UD, see the abstract at the following site: https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:23_how_to_ellipsis_a_proposal_.pdf

[21] Note that the `orphan` relation does not account for all cases of predicate ellipsis. It only highlights instances where a predicate is omitted and at least two remnants are present. Instances where the predicate is omitted but only one remnant survives are not marked with `orphan`; instead, they are annotated through promotion and thus may be overlooked.

[22] For the notion of parallelism in ellipsis: Phillips and Parker (2014, p.79)

[23] The Misc column is the last column of the CoNLL-U format used for annotation: https://universaldependencies.org/format.html.

antecedent can also be in a different sentence. It is displayed as a numeric value in the Misc column, and its ordering is not limited to the sentence but extends across the entire text/treebank. In A, we provide examples of its usage for *Italian-Old*. In cases of node splitting or duplication, the split node will receive a decimal number to avoid modifying annotation that have already been completed. For completeness of information and to reflect the reasoning adopted by the annotator, in cases where the antecedent is absent in the text, the AntecPosit value is annotated as "wk", indicating "world knowledge") (`AntecPosit=wk`). For instance, the Misc field of the empty node in the sentence that involves ellipsis will be the following: `Antec=Yes|AntecPosit=2`. This means that the antecedent is present (`Antec=Yes`), and that its position has the unique_token number 2 (`AntecPosit=2`).

- **what, if anything, should we reconstruct?** In an effort to balance the principle that more information enhances analysis, with the understanding that this is a preliminary step pending further refinement—and in the spirit of gathering community feedback and cross-linguistic analysis cases[24]—we suggest, at this initial stage, reconstructing the elided material, including—when permitted by the context (refer to Example 6 Subsection 7.2 for a specific issue)—its form, lemma, UPOS, features, head, and dependency relation. However, at the current stage of this study, we have decided not to address the reconstruction of arguments of the elided predicate,[25] even in the case of complex predicates, such as in an expression like "to make shield of". In such cases, only the predicate "make" is reconstructed. The decision to limit the reconstruction to verbs—while excluding arguments and complex predicates—is motivated, among others, by the difficulty of deriving valency information from UD annotation, given that the distinction between arguments and adjuncts is not explicitly encoded (De Marneffe et al., 2021, p. 13). We will provide examples in Section 7.

7 A Case-Study: Predicate Ellipsis in Italian-Old Treebank. Some Preliminary results and considerations

In this Section, we present examples of the enhancement of **predicate ellipsis** in a portion of the Italian-Old treebank, following the method described in Section 6.

Italian-Old[26] is a native UD treebank containing the *Divine Comedy* by Dante Alighieri, an Old Italian poem written between approximately 1306 and 1321. The poem is divided into three *Cantiche*: *Inferno* (Hell), *Purgatorio* (Purgatory), and *Paradiso* (Heaven). The first *Cantica*, *Inferno*, was manually annotated from scratch with respect to syntax. In contrast, the other two *Cantiche* were pre-parsed using a model trained on *Inferno* data and subsequently manually corrected (Corbetta et al., 2023).

The enhancement discussed here focuses on the first *Cantica*, *Inferno*, which comprises 33,416 tokens (excluding punctuation marks). In the following Subsections, we first describe the extraction of predicate ellipsis (Subsection 7.1), and then (Subsection 7.2) report on some noteworthy cases of enhancement in light of the proposals presented in Subsection 6.2.

7.1 Extraction

As a treebank natively annotated in UD, *Inferno* encodes ellipsis according to the UD guidelines, using the `orphan` relation and the promotion mechanism (see Section 3.1). To extract instances of predicate ellipsis, we queried for occurrences of the `orphan` relation and for all instances in which a non-verbal node (based on PoS) serves as the head of a dependency relation typically associated with verbal predicates, namely, nodes functioning as clause heads. We selected the following dependency relations: `root`, `parataxis`, `advcl`, `acl`, `ccomp`, and `csubj`, including their subtypes, if present.[27]

[24] As highlighted in Section 8, this work also aims at encouraging participation on this topic across different languages and to gather supporting evidence accordingly, with the goal of enriching the proposal and making it as language-independent as possible. To this end, we have undertaken preliminary experimentation with Latin treebanks.

[25] This choice contrasts with the approach adopted in the PDT style, which also reconstructs arguments https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch06s12s01.html#elipsa1.1).

[26] https://github.com/UniversalDependencies/UD_Italian-Old.

[27] We excluded `advcl:pred` and `xcomp` from the selection, as they are used for secondary predication, which we did not consider as instances of ellipsis. Moreover, we will not address `conj` in the current discussion, as it represents a broad topic that cannot be fully addressed within the scope of this paper.

Each case was then manually inspected to determine whether the retrieved instances genuinely represented ellipsis. With only a few exceptions, identified as annotation errors, all the examples retrieved by the query can be interpreted as valid instances of ellipsis.

In A, we report Table 1, representing the queries, and Table 2,[28] which reports the number of occurrences distributed across the selected deprels.

While a thorough analysis of the various types of ellipsis identified across the cases falls outside the scope of this study, it remains a goal for future research, as mentioned in 8.

7.2 Specific cases

When addressing the reconstruction of certain instances of ellipsis, we encounter several of the issues outlined in Subsection 6.2. In what follows, we illustrate some problematic cases related to the reconstruction process and the position of the reconstructed material (7.2.1) and the retrieval of the antecedent (7.2.2), and we describe the strategies we adopt to address them.

7.2.1 Reconstruction and position

Regarding the position of the ellipsis site and the material to be reconstructed, we present the following Example 4:

> **Example 4** - *Inf.* II, vv. 88-90
> *Temer si dee di sole quelle cose/ c'hanno potenza di fare altrui male;/ de l'altre no, ché non son paurose.*
> "Of those things only should one be afraid/ Which have the power of doing others harm;/ Of the rest, no; because they are not fearful."

In this example, the elliptical sentence is *de l'altre no* ("of the rest, no") and the antecedent is *Temer si dee* ("one should be afraid").

Even though in Example 4 the antecedent includes the lexical infinitive verb (*temer* "be afraid"), its modal verb (*dee* "should") and the reflexive clitic (*si* "one"), we reconstruct only a single node—specifically, the one corresponding to the content word—in line with the prioritization of content words over function words in UD.[29] This choice, however, does not preclude indicating the full antecedent information, which is conveyed by the unique_token attribute.[30] In A, we report the syntactic tree with enhanced dependency.

7.2.2 Antecedent retrieval

Another complex example involving the identification of the antecedent is reported in Section 3.1 (repeated here for clarity):

> **Example 5** - *Inf.* IV, vv. 31-32
> *Lo buon maestro a me: «Tu non dimandi/ che spiriti son questi che tu vedi?*
> "To me the Master good: "Thou dost not ask/ What spirits these, which thou beholdest, are?"

In this example, the ellipsis concerns the main clause *Lo buon maestro a me:* and consists in the omission of a *verbum dicendi*, that is, a verb whose meaning conveys an act of speaking. In this case, no explicit antecedent is retrievable either from the sentence itself or from the preceding context, as the previous dialogic exchange is also introduced through an elliptical construction (Example 6):

> **Example 6** - *Inf.* IV, vv. 19-21
> *Ed elli a me: «L'angoscia de le genti/ che son qua giù, nel viso mi dipigne/ quella pietà che tu per tema senti.*
> "And he to me: "The anguish of the people/ Who are below here in my face depicts/ That pity which for terror thou hast taken."

Example 6, however, does have a clear antecedent in vv. 16–17: *E io, che del color mi fui accorto,/ dissi:* "And I, who of his colour was aware,/ Said:".

Unlike the ellipsis in Example 6, where the antecedent can be retrieved just a few verses earlier (in v. 17), Example 5 lacks an antecedent in the immediate context and must therefore be interpreted through world knowledge. We report the annotation of Example 5 and 6 in A. In Example 6, both the form and lemma are reconstructed, as the

[28] The asterisk (*) following a deprel indicates that its subtypes were also included in the count. More specifically, `advcl` was queried along with `advcl:cmp`, `csubj` with `csubj:pass`, `acl` with `acl:relcl` and `ccomp` with `ccomp:reported`.

[29] https://universaldependencies.org/u/overview/syntax.html#the-primacy-of-content-words

[30] We note that the elided sentence also contains an instance of nominal ellipsis, namely *de l'altre [cose] no* (lit. "of the other [things] not"). However, since the present study focuses on predicate ellipsis, we do not reconstruct it.

antecedent is explicitly present. In contrast, in Example 5, the form is not specified; only the lemma is provided, as it is inferred.

8 Conclusion and Future Work

In this paper, we focus on a specific syntactic phenomenon, ellipsis, that has been analyzed across various frameworks, though to a lesser extent within the dependency-based tradition.

The inherent difficulty of capturing the nature of an omission in the text is compounded by the challenges of representing it graphically in syntactic corpora, namely treebanks. Within the state-of-the-art dependency framework (UD), ellipsis appears to be only marginally addressed, largely due to the annotation choices made at the basic level and the complexity involved in automatically retrieving such structures.

To address the lack of gold-annotated data, which are crucial for both linguistic analysis and machine learning, in this work we experimented on predicate ellipsis, as a preliminary step towards the ultimate goal of developing language-independent guidelines for the treatment of ellipsis in Universal Dependencies. Given the complexity of the topic, in this paper we have narrowed the scope to predicate ellipsis, with the aim of extending the analysis and annotation to all types of ellipsis in future works. For this paper, we focus on two complementary workflows: a theoretical one, centered on the analysis of ellipsis, and a practical one, aimed at providing a comprehensive annotation of predicate ellipsis in EUD. Examples of reconstructions, along with practical annotation cases, are presented in the final Section with respect to the UD treebank Italian-Old. An enhanced annotation of predicate ellipsis will be provided in the next release of the treebank Italian-Old. Given the inherently non-lexicalized nature of ellipsis, the proposed workflow can be extended to other languages through the use of morpho-syntactic annotation.
Concerning future work, a data-driven classification of predicate ellipsis in Italian-Old treebank is envisaged, one that emerges directly from the data and provides examples of ellipsis in context.

Additionally, we plan to develop a rule-based script, designed to be as language-independent as possible, to support the semi-automatic enhancement of predicate ellipsis. This step does not aim at a fully automatic resolution of ellipsis, which is still far from happening, but rather at facilitating manual annotation, which remains necessary. We welcome and encourage contributions from other treebank maintainers—or from researchers interested in exploring ellipsis—who wish to enrich their treebanks with this type of information.

Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). We would like to thank Daniel Zeman, Petya Osenova, and Simov Kiril for their valuable feedback throughout this research. Any remaining errors are solely the responsibility of the authors.

References

Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. 2018. *Application of Graph Rewriting to Natural Language Processing*. John Wiley & Sons.

Damir Ćavar, Ludovic Mompelat, and Muhammad Abdo. 2024a. The typology of ellipsis: a corpus for linguistic analysis and machine learning applications. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 46–54.

Damir Ćavar, Zoran Tiganj, Ludovic Veta Mompelat, and Billy Dickson. 2024b. Computing ellipsis constructions: Comparing classical nlp and llm approaches. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 217–226.

Claudia Corbetta, Marco Passarotti, Flavio Massimiliano Cecchini, and Giovanni Moretti. 2023. Highway to hell. towards a universal dependencies treebank for dante alighieri's comedy.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Kira Droganova, Filip Ginter, Jenna Kanerva, and Daniel Zeman. 2018. Mind the gap: Data enrichment in dependency parsing of elliptical constructions. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 47–54.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovskỳ. 2017. Handbook on linguistic annotation, chapter prague dependency treebank.

Jorge Hankamer and Ivan Sag. 1976. Deep and surface anaphora. *Linguistic inquiry*, 7(3):391–428.

Philip Hofmeister. 2007. Memory retrieval effects on filler-gap procession. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

Richard Hudson. 2010. *An introduction to word grammar*. Cambridge University Press.

Christopher Kennedy. 2003. Ellipsis and syntactic representation. In *The interfaces: Deriving and interpreting omitted structures*, pages 29–53. John Benjamins Publishing Company.

Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Marjorie J McShane. 2005. *A theory of ellipsis*. Oxford University Press.

Jason Merchant. 1999. *The syntax of silence: Sluicing, islands, and identity in ellipsis*. University of California, Santa Cruz.

Jason Merchant. 2001. *The syntax of silence: Sluicing, islands, and the theory of ellipsis*. Oxford University Press.

Jason Merchant. 2018. Ellipsis: A survey of analytical approaches. In Jeroen van Craenenbroeck and Tanja Temmerman, editors, *The Oxford Handbook of Ellipsis*, Oxford Handbooks. Oxford University Press.

Marie Mikulová. 2014. Semantic representation of ellipsis in the prague dependency treebanks. In *Proceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014)*, pages 125–138.

Iván Ortega-Santos, Masaya Yoshida, and Chizuru Nakao. 2014. On ellipsis structures involving a wh-remnant and a non-wh-remnant simultaneously. *Lingua*, 138:55–85.

Timothy Osborne. 2019. Ellipsis. In *A Dependency Grammar of English*, pages 349–378. John Benjamins Publishing Company.

Petya Osenova and Kiril Simov. 2003. The bulgarian hpsg treebank: Specialization of the annotation scheme. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories; Växjö, Sweden*.

Colin Phillips and Dan Parker. 2014. The psycholinguistics of ellipsis. *Lingua*, 151:78–95.

Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Martin Popel, Zdeněk Žabokrtskỳ, and Martin Vojtek. 2017. Udapi: Universal api for universal dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101.

John Robert Ross. 1969. Guess who? In *Proceedings from the annual meeting of the chicago linguistic society*, volume 5, pages 252–286. Chicago Linguistic Society.

Maria Simi, Simonetta Montemagni, et al. 2018. Bootstrapping enhanced universal dependencies for italian. In *CEUR WORKSHOP PROCEEDINGS*, volume 2253. CEUR-WS.

Jeroen Van Craenenbroeck and Tanja Temmerman. 2018. *The Oxford handbook of ellipsis*. Oxford University Press.

A Appendix

1) **`Orphan` detection query**: pattern { N1 -[`orphan`]->N2 }
2) **Promotion detection queries**: pattern {N1 -[deprel*]-> N2} without {N2 [upos="VERB"] } without {N2-[`cop`]->X} without {N2-[`orphan`]->X} without { N2 [upos="AUX"] } deprel* = `root`, `parataxis`, `advcl`, `advcl:cmp`, `acl`, `acl:relcl`, `ccomp`, `ccomp:reported`, `csubj`, and `csubj:pass`.

Table 1: Queries for Predicate Ellipsis

deprel	**occurrences**
`orphan`	124
`root`	36
`parataxis`	12
`advcl*`	95
`ccomp*`	19
`acl*`	0
`csubj*`	0

Table 2: Occurrences of Predicate Ellipsis in *Inferno*

Figure 1: Enhanced tree of Example 3

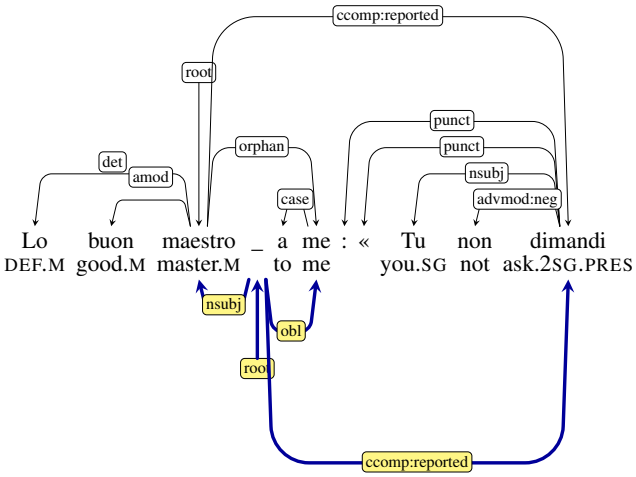


Figure 2: Enhanced tree of Example 4: antecedent in the same sentence. *For reasons of space, we do not report the unique_number (UN) for each word.

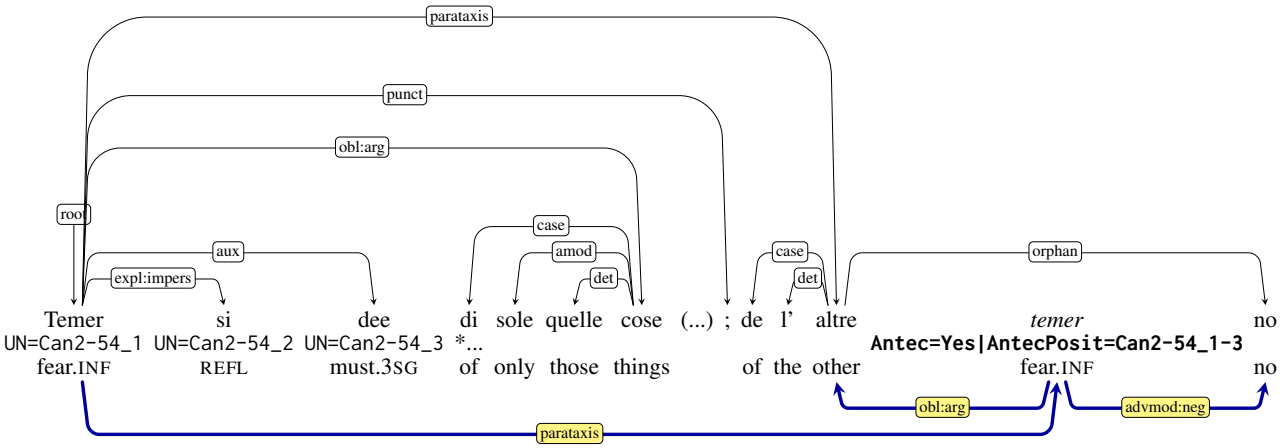


Figure 3: Enhanced tree of Example 5: antecedent not overtly present in the text.

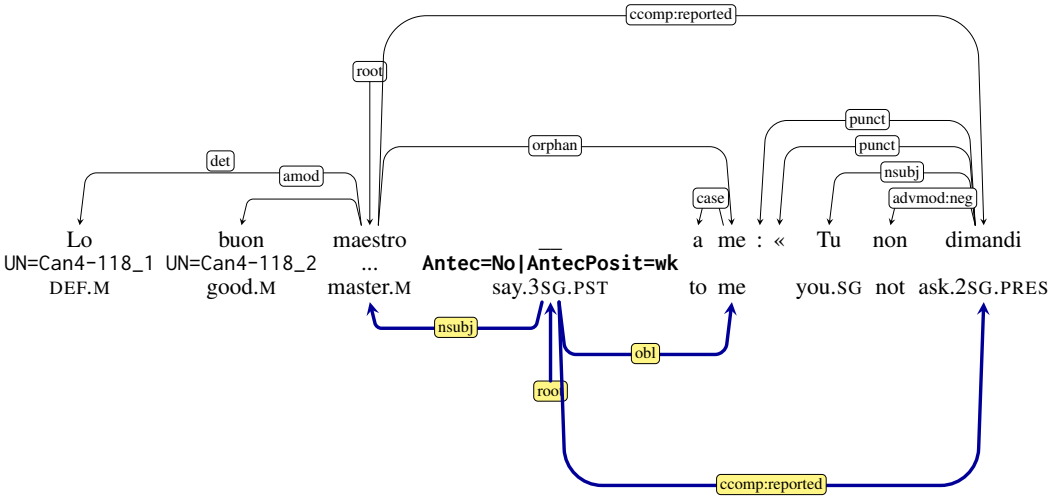


Figure 4: Enhanced tree of Example 6: antecedent not in the same sentence.