Annotating the *Divine Comedy* in Universal Dependency: A Journey through *Inferno*, *Purgatorio* and *Paradiso*.

How to (…) ellipses, and other annotation issues.

Claudia Corbetta, University of Bergamo/Pavia (Italy)

key-words: UD, annotation, Italian-Old, ellipsis.

Description of a resource related to the topics of the training school:

Italian-Old[1] is an existing Universal Dependencies (UD) (De Marneffe et al., 2021) treebank that includes Dante Alighieri's *Divine Comedy*, an Old Italian (specifically Florentine) poem composed between approximately 1306 and 1321 (Inglese, 2012).

The *Comedy* is considered a pivotal text in the history of Italian literature and language (Manni, 2013). This work comprises three *Cantiche*: *Inferno* (Hell), *Purgatorio* (Purgatory), and *Paradiso* (Heaven). Each *Cantica* is subdivided into *Canti*, culminating in a total of 100 (34 in *Inferno*, 33 in *Purgatorio*, and 33 in *Paradiso*). Speaking about numbers, the *Comedy* is composed of 3.373 sentences and 122.047 syntactic words[2]. Specifically, *Inferno* counts 1.228 sentences and 41.367 syntactic words, *Purgatorio* 1.174 sentences and 41.277 syntactic words and *Paradiso* respectively 971 and 39.403[3].

As of now, Italian-Old only contains the first *Cantica*, namely *Inferno*[4]. The syntactic annotation of *Inferno* has been conducted from scratch by me, following UD annotation scheme (Corbetta et al., 2023).

While facing the annotation process, I encountered several issues that required making specific decisions. Among them, I dealt with ellipses (Merchant, 2019), including both predicate ellipses (Lobke and Harwood, 2019), i.e., cases of omission of the predicate and, potentially, its internal arguments or adjuncts, and nominal ellipses, "involving a gap within the internal structure of the nominal phrase" (Saab, 2019, p. 526).

Following the UD annotation style, the options for annotating ellipses are promotion, i.e., the choice of an element to occupy the position of the omitted element in the syntactic tree, adhering to a predetermined hierarchy[5], and the *orphan* dependency relation, which is used to indicate that the element associated with the *orphan* relation lacks an overt dependent element within the syntactic tree. However, UD annotation (not considering Enhanced Dependency[6]) makes it difficult to retrieve and analyze ellipses. In fact, on one hand, the promotion mechanism is used without explicitly signaling the ellipsis, resulting in a loss of information regarding the presence of this phenomenon (see example i). On the other hand, the *orphan* relation signals the presence of an ellipsis, but it obscures the dependency relations of the sentence (see example ii).

Explanation how the participation in the training school will be useful for the project:

I would like to participate in the training school to discuss how to deal with cases of ellipsis (which will be the linguistic focus of my PhD), particularly how to identify instances of nominal ellipsis annotated via

---

[1] Refer to: https://github.com/UniversalDependencies/UD_Italian-Old.

[2] In UD "syntactic word" refers to the actual level of analysis in the syntactic tree. Refer to: https://universaldependencies.org/u/overview/tokenization.html.

[3] Being a resource in process of creation, the numbers are subject to variations (specifically for *Paradiso* and *Purgatorio*).

[4] *Purgatorio* may be published in the upcoming UD release (May 2024).

[5] See UD guidelines: https://universaldependencies.org/u/overviewspecific-syntax.html#ellipsis.

[6] Refer to: https://universaldependencies.org/u/overview/enhanced-syntax.html.

promotion in UD annotation (without an available enhanced dependency annotation)[7]. Additionally, I am interested in understanding how similar phenomena are treated in other languages, specifically within the poetry genre, and if there are cases of study of ellipsis in treebanks. This discussion would provide valuable insights into handling ellipsis in linguistic annotation of my resource, where ellipsis often occurs in intricate and nuanced ways.

Furthermore, the syntactic annotation of *Paradiso* is yet to be completed, and within the *Cantiche*, several annotation decisions still require harmonization in preparation for the final release of the *Comedy*. Participating in the training school offers an opportunity to discuss the annotation of various cases[8] beyond ellipsis, including the treatment of:

i) **locutions**, i.e., sets of two or more words arranged in a fixed sequence (Serianni and Castelvecchi, 2006), such as *mentre che* 'while' and *davanti a* 'in front of'. Should I a) employ the *fixed* relation or b) consider treating similar cases as independent? (For some of them, I occasionally encounter occurrences of material interspersed between the two elements).

davanti a X 'in front of X'	*davanti* (lit. 'front'); *a* (lit. 'to')
a) *davanti* ADV/ADP(?) <—(*fixed*) — *a* ADP	only *davanti* attached to the head X with *advmod/ case* (?) relation.
b) *davanti* ADV (*advmod*) // *a* ADP (*case*)	both attached to the head X.

ii) **comparative clauses**, usually occurring with ellipses, e.g., *I dannati urlavano come cani di notte,* 'The damned screamed like/as dogs in the night'; should it be treated as a comparative clause (*advcl:cmp*) or as an oblique (*obl:cmp*)?

iii) **relative clauses**, particularly the decision of whether to adopt annotation with a) an inner or b) external head, e.g., *Abbraccio chi voglio* 'I hug <u>who</u> I want':

a) *chi* ('who') is *obj* of *voglio* ('want')	inner head
b) *chi* ('who') is the *obj* of *abbraccio* ('hug')	external head

iv) **complex expressions** like *di gente in gente* 'from one person to another'. Should I annotate this structure as *flat* or consider "*di gente*" - "*in gente*" as two different *obl* deprels (both attached to the head of the clause or with the latter *gente* dependent on the first *gente*)?

v) **cases of ambiguity,** where multiple possibilities of annotation arise, due to several interpretations of the text. I acknowledge that only one interpretation is currently feasible in UD annotation, but it would be interesting to accommodate multiple interpretations, particularly in poetic texts intended to convey multiple layers of meaning. Could incorporating multiple interpretations of the same sentence pose a challenge/confusion in NLP training?

Open questions related to the project which could be addressed during the brainstorming hackathon:

I would suggest the possibility of introducing dependency relation subtypes for ellipsis. That could facilitate the rapid extraction of ellipsis sites and their contexts in treebanks that do not present an enhanced dependency annotation. Additionally, I would suggest exploring the treatment of cases i) to v), aiming to understand how they are annotated in other languages.

---

[7] I propose examining mismatches between parts of speech (POS) and their corresponding dependency relations (deprel), as illustrated in example ii), where an adjective (ADJ) is paired with a nominal subject (*nsubj*) deprel, to identify instances of ellipsis annotated with promotion. However, identifying such mismatches can be challenging, as sometimes the promoted element may receive a dependency relation that nullifies the mismatch with the POS, such as *conj* (conjunction relation), which is compatible with both nouns (NOUN) and ADJ.

[8] For reasons of space, I will not report the syntactic trees of the issues mentioned, but if needed, I can provide them for clarification and exemplification.

Short statement of the project phase:

Italian-Old is a treebank currently under development.

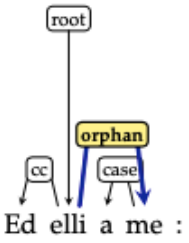
Examples:

i) ellipsis with *orphan* dependency relation: it obscures the *nsubj* deprel (*elli*) and the oblique *obl* deprel (*me*).

ii) ellipsis with promotion: it is difficult to retrieve the ellipsis.

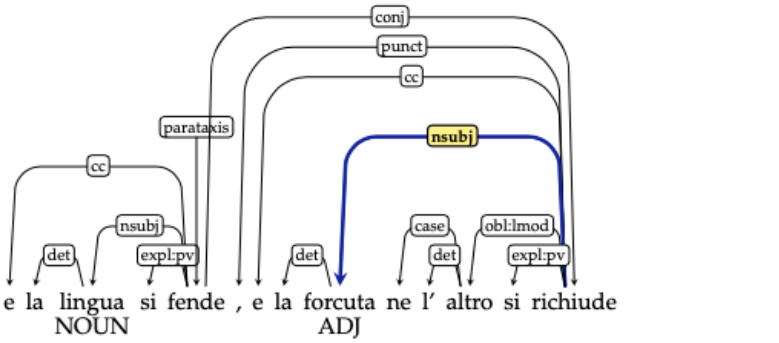
i) orphan

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



ii) promotion

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



References:

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

Corbetta, Claudia, Passarotti, Marco, Cecchini, Flavio Massimiliano and Moretti, Giovanni. 2023. Highway to hell. towards a universal dependencies treebank for Dante Alighieri's Comedy. In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30—Dec 02, 2023*, Venice, Italy, pages 1–8. CEUR-WS.

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

Inglese, Giorgio. 2012. *Dante: guida alla Divina Commedia*. Nuova edizione. Carocci.

Lobke, Aelbrecht and William Harwood. 2019. Predicate Ellipsis. In Jeroen van Craenenbroek and Tanja Temmerman, editors, *The Oxford Handbook of Ellipsis*, Oxford Handbooks. Oxford University Press, Oxford, UK, chapter 21, pages 504–525.

Manni, Paola. 2013. *La lingua di Dante*. Il mulino, Bologna.

Merchant, Jason 2019. Ellipsis: A survey of analytical approaches. In Jeroen van Craenenbroek and Tanja Temmerman, editors, *The Oxford Handbook of Ellipsis*, Oxford Handbooks, chapter 2. Oxford University Press, Oxford, UK.

Saab, Andrés. 2019. Nominal Ellipsis. In Jeroen van Craenenbroek and Tanja Temmerman, editors, *The Oxford Handbook of Ellipsis*, Oxford Handbooks. Oxford University Press, Oxford, UK, chapter 22, pages 526–561.

Serianni, Luca and Castelvecchi, Alberto. 2006. *Grammatica italiana*. Universitaria. UTET Università, Turin, Italy, second edition.