Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09748-6

PROJECT NOTE

Parlamint-it: an 18-karat UD treebank of Italian parliamentary speeches

Chiara Alzetta[1] · **Simonetta Montemagni**[1] · **Marta Sartor**[1] · **Giulia Venturi**[1]

Accepted: 30 April 2024

Abstract
The paper presents ParlaMint-It, a new treebank of Italian parliamentary debates, linguistically annotated based on the Universal Dependencies (UD) framework. The resource comprises 20,460 tokens and represents a hybrid language variety that is underrepresented in the UD initiative. ParlaMint-It results from a manual revision process that relies on a semi-automatic methodology able to identify sentences that are most likely to contain inconsistencies and recurrent error patterns generated by the automatic annotation. Such a method made the revision process faster and more efficient than revising the entire treebank. In addition, it allowed the identification and correction of annotation errors resulting from linguistic constructions inconsistently represented in UD treebanks and from characteristics specific to parliamentary speeches. Hence, the treebank is deemed as an 18-karat resource, since, although not fully manually revised, it is a valuable resource for researchers working on Italian language processing tasks.

Keywords Universal dependencies treebanks · Annotation revision · Italian parliamentary debates · Linguistic annotation

✉ Chiara Alzetta
chiara.alzetta@ilc.cnr.it

Simonetta Montemagni
simonetta.montemagni@ilc.cnr.it

Marta Sartor
marta.sartor@ilc.cnr.it

Giulia Venturi
giulia.venturi@ilc.cnr.it

[1] Istituto di Linguistica Computazionale "A. Zampolli", (ILC-CNR) ItaliaNLP Lab, via G. Moruzzi 1, Pisa, Italy

Published online: 06 July 2024

1 Introduction

Over the last years, the Universal Dependency (UD) initiative (de Marneffe et al., 2021)[1] has made significant contributions to Natural Language Processing (NLP) research. These contributions range from more accurate and robust approaches for cross-lingual dependency parsing (Kondratyuk & Straka, 2019) to wider cross-linguistic analyses, with a particular focus on language typology studies (Croft et al., 2017). The universal representation of the syntactic structure of languages, resulting in multiple cross-linguistically consistent treebanks for many languages and language varieties, is what mostly fostered those contributions.

Since the UD project was launched, the number of available annotated corpora has constantly grown, with a significant increase in both the number of languages included and the number of treebanks available for each language. In fact, while the first release in 2015 consisted of 10 treebanks for 10 different languages, the latest version available as of January 2024 (i.e., version 2.13) includes 259 treebanks for 148 different languages.

Despite such a varied and rich scenario concerning the typologies of languages considered, the project still faces challenges in accurately representing the nuances of different genres of text. Currently, UD treebanks collection contains 18 diverse typologies of genres, including news, fiction, scientific texts, etc., even if, according to the analysis conducted by Nivre et al. (2020), these typologies "are neither mutually exclusive nor based on homogeneous criteria". It follows that the distribution of genres is scarcely consistent across treebanks, except for genres that are clearly defined according to strong cross-lingual creation guidelines. In fact, by relying on the genre labels reported by each treebank development team in the official documentation, we can notice that around 60% of all treebanks fall into four genera, i.e. news, non-fiction, fiction, and Wikipedia. On the contrary, genres characterized by either higher variance in terms of lexical and (morpho-)syntactic features, such as spoken (6.36%), blog (4.07%) or web (3.18%), or highly specialized, such as legal (5.16%), government (1.59%) or medical (1.49%), are much more underrepresented. In addition, the categorization into one or more genre(s) is mostly available at the treebank level rather than at the sentence level thus posing a challenge for developing accurate and reliable NLP models that can handle different genres of text effectively (Müller-Eberstein et al., 2021b, a).

Starting from these premises, in this contribution, we present ParlaMint-It, a new treebank of transcriptions of Italian parliamentary debates, extracted from the official transcripts of the Italian Senate sessions, which have been linguistically annotated based on the UD framework. Parliamentary debates represent a unique variety of language use, which Nencioni et al. (1976) identifies as 'spoken-written', i.e. a variety characterized by a hybrid nature featuring a co-occurrence of traits typical of both written and spoken language. Namely, on the one hand, they contain several normative references (e.g. *article 5 of law n. 184, paragraph 2, states [...]*) that make the transcriptions more similar to a written legal text; on the other hand, they are characterized by traits specific to the spontaneous speech, such as rhetorical questions in interrog-

[1] https://universaldependencies.org/.

ative forms to convey illocutionary force to an assertion, interruptions, ellipses, and discourse markers (Ili, 2015). Such a language variety is quite underrepresented in the UD repository and labeled either as 'government' or 'legal' genre. However, these classifications exhibit considerable heterogeneity in UD, encompassing diverse text types such as patent applications (e.g., the Chinese PatentChar treebank[2]) and public administration texts, as in the CAC (Hladká et al., 2008) and CLTT (Kríž et al., 2016) treebanks of Czech, and the Irish IDT treebank (Lynn & Foster, 2016). Parliamentary speeches, to the best of our knowledge, are documented in only a few treebanks, namely Icelandic IcePaHC (Arnardóttir et al., 2020), Finnish TDT (Pyysalo et al., 2015), and ParTUT (Sanguinetti & Bosco, 2015) available for English, French, and Italian. It's worth noting that, until UD version 2.13, for these three languages, ParTUT stood as the only treebank that contains, among other genera, also parliamentary speeches (covering 47.11% of the total amount of sentences in ParTUT).

The newly introduced ParlaMint-It treebank, consisting of parliamentary speeches, is thus intended to make a valuable contribution towards enriching the UD project, particularly concerning such an underrepresented textual genre. It stems from the ParlaMint initiative (Erjavec et al., 2022) which produces comparable national parliamentary corpora for 17 European countries enriched by metadata about the mandates, sessions, speakers, and their political party affiliations etc., and automatically annotated for named entities and UD morpho-syntactic information. To ensure the correctness of the linguistic annotation of ParlaMint-It, the resource underwent a validation process aimed at identifying and correcting annotation errors and inconsistencies, which are known to naturally occur in automatic parsing outputs (Fort et al., 2012). Manually correcting these errors can be time-consuming and expensive, and previous research has proposed methods for automating the process (see, e.g., Dickinson et al. (2003); Dickinson and Meurers (2005); Boyd et al. (2008); Marneffe et al. (2017); Ambati et al. (2011); Volokh et al. (2011)). To identify and correct errors in ParlaMint-It, we employed a semi-automatic approach for detecting erroneously annotated dependency relations in treebanks that relies on an algorithm for ranking dependencies by reliability, LISCA (Dell'Orletta et al., 2013a). This method was chosen for its ability to effectively restrict the search space for errors and to identify patterns of recurrent errors that are generated by a parser (Alzetta et al., 2017).

In the remainder of the paper, after a discussion of the internal composition of the ParlaMint-It treebank (Sect. 2), in Sect. 3 we introduce the revision and correction methodology we adopted. This methodology yields an '18-karat gold' resource, i.e. a high-quality morpho-syntactic treebank that adheres to UD principles and schema partially manually corrected for annotation errors. Finally, Sect. 4 describes the results of the revision process that makes the ParlaMint-It treebank a valuable resource for researchers and practitioners working on Italian language processing tasks.

[2] https://github.com/UniversalDependencies/UD_Chinese-PatentChar

2 The ParlaMint-it corpus

The ParlaMint-It corpus is part of a larger multilingual collection of parliamentary transcripts built in the context of the ParlaMint project (Erjavec et al., 2022).[3] The project originated from the widespread interest in the compilation of corpora of national parliamentary proceedings as witnessed by the numerous initiatives organized on the topic, such as the dedicated CLARIN Resource Family of Parliamentary Corpora[4] or the ParlaCLARIN LREC Workshop series (Fišer et al., 2018, 2020, 2022). The project's main goal consists of producing a uniformly encoded, comparable, and linguistically annotated set of corpora for multiple languages reflecting the discussions, perspectives, and judgments of the European parliaments on the societal impact of the COVID-19 pandemic. To this end, each monolingual corpus is articulated in two sub-corpora, namely a *COVID sub-corpus* containing speeches after November 2019 and a *pre-COVID sub-corpus* including speech before the pandemic period.

ParlaMint version 2.1 contains 17 corpora with 16 main languages, it comprises 5 million speech transcripts and almost half a billion words. It is freely distributed[5] and fully marked according to the Parlamint XML schemas,[6] which specialize the ParlaMint TEI ODD schema (Erjavec & Pančur, 2019) as they are specifically developed for the project to better validate the ParlaMint corpora. Notably, the schema allows for the annotation of both speakers' metadata (e.g. political parties and groups, gender, the government in charge) and linguistic information compliant with the Universal Dependency formalism.

ParlaMint-It is a sub-part of the Italian section of the ParlaMint corpus which originally includes a total of around 26 million words extracted from the stenographic verbatim transcriptions of the plenary sessions of the Senate, i.e. the assembly of the upper house of the Italian Parliament (Agnoloni et al., 2022). As shown in Table 1, it is composed of 20,460 tokens (corresponding to 701 sentences) and it is internally articulated in two sections resembling the original composition of the Italian section of the ParlaMint corpus. Namely, the *2020* section includes a selection of debates extracted from Senate sessions held during the COVID-19 pandemic period, i.e. November 2019 - December 2020, while the *2015* one contains part of a 2015 session corresponding to an excerpt of the pre-COVID period ranging from March 2013 to October 2019. The debates feature 26 distinct speakers, with 15 contributors in the *2015* section and 11 in the *2020* section, providing a diverse range of voices within these distinct temporal contexts. The linguistically annotated corpus has been manually revised adopting the methodology described in Sect. 3 to identify potential labeling errors caused by the automatic linguistic annotation process and validated following the procedure presented at the beginning of Sect. 4. The methodology and the final evaluation allowed modifying 48.36% of the 701 sentences, i.e. 339 sentences, due to erroneous annotations which are differently distributed across the 2015 and 2020 sub-sections.

[3] https://www.clarin.eu/parlamint

[4] https://www.clarin.eu/resource-families/parliamentary-corpora

[5] https://www.clarin.si/repository/xmlui/handle/11356/1432

[6] https://github.com/clarin-eric/ParlaMint/blob/main/Schema/README.md

Table 1 ParlaMint-It statistics: the size of the whole treebank (first line) and of the 2015 and 2020 sections in terms of tokens and sentences, and the number of sentences that underwent manual modifications due to automatic linguistic annotation errors

	Tokens	Sentences	Modified sentences
ParlaMint-It	20,460	701	339 (48.36%)
2015	10,829	359	196 (54.60%)
2020	9,631	342	143 (41.81%)

A further peculiarity of the corpus is represented by the diverse topics accounted for during the parliamentary sessions. Specifically, in the 2015 section, the members of the Senate debated about the appropriateness of equating adoption and foster care institutions, while in the 2020 one, they discussed the prison riot that happened in March 2020 calling for better anti-COVID measures. The two topics have a different impact on the revision process as discussed in what follows.

As a final remark, note that each sentence of the ParlaMint-It corpus reports a unique identifier that allows the mapping of sentences to the original Italian section of the ParlaMint corpus and its associated metadata, thus enabling the analysis of contextual information for each speaker's statement.

3 Methodology

The approach devised to obtain the UD ParlaMint-It corpus comprises three main steps, as displayed in Fig. 1.

The initial step entails the collection of the sub-set of 701 sentences extracted from the Italian section of the ParlaMint corpus, representative of the *COVID* (2015) and of the *pre-COVID sub-corpus* (2020). As described by Agnoloni et al. (2022), all sentences were automatically annotated using the Stanza neural pipeline[7] which is reported to achieve state-of-the-art or competitive performance for different languages (Qi et al., 2020). Specifically, for Italian, Stanza demonstrated robust performance on the Italian Stanford Dependency Treebank (ISDT) with UAS=93.29, LAS=91.61, CLAS=87.32.[8] These results suggest the model's reliability in parsing standard Italian texts. Among the different models available for the Italian language, we chose `italian-isdt-ud-2.5`, trained on ISDT, since it represents the biggest UD Treebank for Italian covering different textual genres (Bosco et al., 2013). However, it's worth noting that Stanza's performance on UD treebanks containing parliamentary speeches indicates some variability in handling this specific language variety. LAS scores for English ParTUT: 88.36, French ParTUT: 89.58, Italian ParTUT: 91.47, IcePaHC: 82.85, and TDT: 88.31 highlight potential challenges in processing parliamentary speech data.

The second phase of the methodology covers the manual revision of the automatic parses produced during the first phase. The revision of ParlaMint-It relies on

[7] https://stanfordnlp.github.io/stanza/index

[8] https://stanfordnlp.github.io/stanza/performance

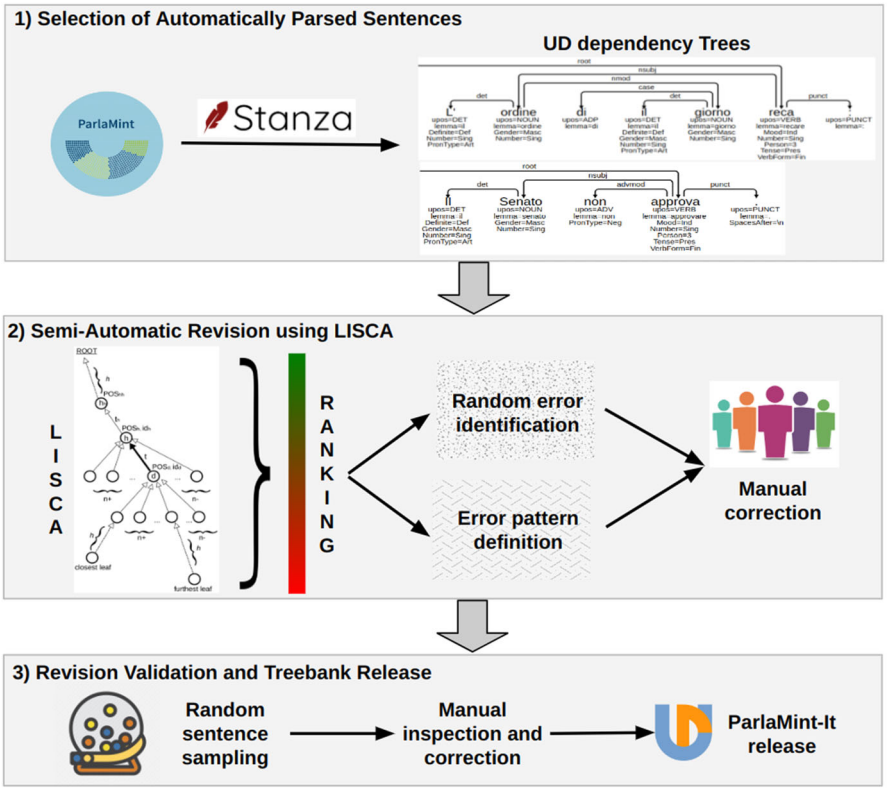


Fig. 1 Overview of the methodology workflow

the assumption that annotation errors can be either random or systematic (Agrawal et al., 2013): the former are heterogeneous, while systematic errors can be identified by searching for recurrent error patterns. To identify patterns of systematic and recurring erroneous dependency relations in the automatically annotated ParlaMint-It corpus we adopted and specialized the methodology first introduced by Alzetta et al. (2017), which leverages the LISCA (*Linguistically-driven Selection of Correct Arcs*) algorithm (Dell'Orletta et al., 2013a). As illustrated in Sect. 3.1, the LISCA algorithm measures the reliability of automatically generated dependency relations taking into account a wide range of factors. As displayed in the second step of Fig. 1, these factors include both the local properties of each dependency relation produced by Stanza and the linguistic context in which they occur. The algorithm produces a score that quantifies the reliability of individual dependency relations, which can be used to rank the dependencies based on their likelihood of being correct.

An expert in Linguistics, familiar with the UD morphosyntactic annotation, inspects the ranking, paying particular attention to relations showing a lower plausibility score (i.e., roughly, the lower third of the ranking, depicted in bright red in the ranking of Fig. 1). The choice to employ LISCA for error identification stems from the effectiveness of this approach in isolating errors efficiently, as demonstrated in prior works

(Dell'Orletta et al., 2013b; Alzetta et al., 2017, 2020). These studies have illustrated that the LISCA ranking is particularly effective for identifying instances of annotation errors and marked genre-specific constructions. Indeed, the expert, by focusing on the dependency relations obtaining the lowest scores, can identify instances of recurring errors in the annotations and define filtering heuristics that model the error pattern, thus enabling the retrieval of all matching cases in the entire corpus.

These filtering heuristics, described in detail in Sect. 3.2, operate by identifying the sentences where the targeted errors occur through Python scripts executed on the CoNLL-U output of Stanza. By isolating the annotations that most likely report an error, this approach, which combines LISCA and the heuristics, ensures that the error identification process is not only thorough but also capable of revealing patterns that might be overlooked in a smaller, manually analyzed random sample.

The revision process was performed by one of the authors of the paper who has a background in linguistics and a strong mastery of the UD guidelines, and it was followed by a discussion among a group of researchers with expertise in UD treebank construction and maintenance, including the authors of this paper, to address any problematic cases that arose. Upon full inspection of the flagged sentence, we identified and corrected the erroneous annotation found by the heuristic. In addition, while searching for systematic errors, the expert may encounter isolated cases of errors that do not follow any formal pattern and thus can be identified only by a thorough manual inspection. These are cases of random errors, also corrected by the experts when found. Note that the manual revision process targeted erroneous annotations at all annotation levels. Namely, the Part-Of-Speech (POS) of the dependent d of a dependency relation, its syntactic head h, the type of dependency connecting d to h, and the lemma and morphological features of d and h.

After completing the manual revision process, a validation step was taken in phase 3 to assess the overall quality of the ParlaMint-It corpus. For this purpose, a random sample of sentences amounting to 20% of the total number of sentences was selected for manual validation. Note that we excluded from this subset sentences that had already undergone manual validation. This allowed us to focus exclusively on sentences that were presumed to be error-free based on previous stages of the process. The objective of the evaluation was to determine the effectiveness of our revision methodology in identifying sentences in ParlaMint-It that were most likely to contain annotation errors, which would make the revision process faster and more efficient than revising the entire corpus. By validating this subset of sentences, we ensure that the final version of ParlaMint-It is a high-quality resource for researchers and developers. In fact, while it may not reach the level of a '24-karat gold' resource, which would require full manual revision, it can be regarded as an '18-karat gold' resource.

Furthermore, to ensure the compliance of the corpus with the formal requirements of the UD annotation schema, we relied on the official UD validator.[9] ParlaMint-It is currently accessible in the official UD repository.[10] Note that the IDs of sentences in UD ParlaMint-It align with those of the original ParlaMint data. The IDs of sentences subject to manual revision are reported in Appendix A.

[9] UD validation rules are described at https://universaldependencies.org/validation-rules.html.

[10] https://universaldependencies.org/treebanks/it_parlamint

3.1 LISCA algorithm

LISCA (Dell'Orletta et al., 2013a) is an unsupervised algorithm aimed at assigning a *plausibility* score to each dependency relation in a Target Corpus based on the statistics acquired from a Reference Corpus. As described in Alzetta et al. (2020), the algorithm operates in two steps:

1. collection of statistics about a set of linguistically motivated features extracted from an annotated corpus obtained through automatic dependency parsing to build a statistical model of the language;
2. assignment of a plausibility score to each dependency link in a target corpus on the basis of the statistical model built in the previous step. Note that a dependency relation is defined here as a triple d(ependent), h(ead), and t(ype) of dependency linking d to h.

Note that in the present work, the Target corpus corresponds to the automatically parsed ParlaMint-It corpus; the Reference corpus is a portion of the Italian Wikipedia (around 40 million tokens) meant to be representative of the ordinary Italian language.

The statistics collected by the LISCA model concern a wide set of linguistically motivated features with respect to local and global properties of the dependency tree. Hence, the plausibility score assigned to each dependency relation is sensitive to changes in the context of occurrence of each specific relation: it reflects the degree of similarity of the linguistic context in which a given dependency relation occurs in the reference and target corpora. On the other hand, the score is based on the assumption that more frequently occurring syntactic structures are more likely to be correct than less frequent ones. From this, it follows that higher LISCA scores are assigned to dependency relations associated with linguistic contexts more frequently occurring in the reference corpus; conversely, lower scores identify relation instances occurring in less typical contexts, which possibly correspond to annotation errors.[11] Based on these operational principles, the difference in textual genres between the Target and Reference corpora enables the attribution of lower plausibility scores to linguistic constructions specific to parliamentary speeches (representing the Target corpus in this study). This is because such constructions are likely less prevalent in both the Wikipedia pages (the Reference corpus) and the training data of the Stanza parsing model.

As a last step, all dependency relations of ParlaMint-It are ordered by decreasing LISCA score, thus obtaining a ranking of relations that reflects the gradient of the annotation plausibility, from higher to lower. Note that, for the specific concerns of this study, we used LISCA in its de-lexicalized version: this allows us to abstract away from variations resulting from lexical effects.

3.2 Filtering heuristics

The goal of filtering heuristics is to detect anomalies in the annotation that occur systematically and that need to be manually inspected and, if needed, corrected. As

[11] Refer to Alzetta et al. (2020) for a detailed description of the formula used to compute the LISCA score.

Table 2 Distribution (in percentages) of erroneous dependency relations matched by each group of filtering heuristics and corrected in the revision process over ParlaMint-It and the 2015 and 2020 sub-sections

Group	Heuristics	ParlaMint-It	2015	2020
General purpose	non-projective	36.09%	37.28%	34.12%
	nominal mod	22.62%	27.46%	14.69%
	aux-verb	0.72%	0.87%	0.47%
	adj-mod	0.18%	0.29%	0.00%
	Group total	59.61%	65.90%	49.29%
Corpus specific	honorifics	10.95%	9.25%	13.74%
	nsubj-obj	4.85%	4.91%	4.74%
	art-pron	2.69%	1.45%	4.74%
	che-homograph	1.97%	1.45%	2.84%
	Group total	20.47%	17.05%	26.07%
Guidelines based	conjuncts	16.88%	13.58%	22.27%
	punct	1.80%	2.02%	1.42%
	iobj	1.26%	1.45%	0.95%
	Group total	19.93%	17.05%	24.64%

mentioned, the heuristics are operationalised through Python scripts designed to match specific sub-trees formalized using the values of the fields in the CoNLL-U file, i.e. Stanza parser output. These heuristics are manually defined on the basis of recurring patterns of erroneous annotation emerging from the inspection of the lower part of the ranking of dependencies ordered by decreasing LISCA score, roughly corresponding to one third of the complete ranking. This ranking segment contains dependency relations that exhibit the lowest LISCA scores in the entire corpus.

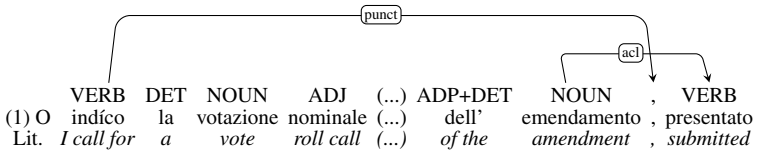
The identified cases of systematic errors encountered in ParlaMint-It can be categorized into three main classes based on the typology of linguistic phenomena that were targeted:

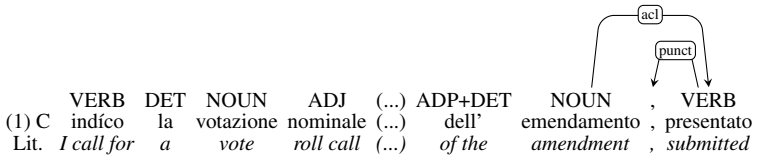
- erroneous annotation of morpho-syntactic areas which usually pose difficulties for the automatic annotation of Italian sentences. We referred to these heuristics as *General purpose* and are partially derived from the original study conducted by Alzetta et al. (2017);
- erroneous annotation of phenomena specific to the language variety used in the parliamentary debates, which we referred to as *Corpus specific* heuristics;
- annotation resulting erroneous according to the criteria set by the UD guidelines. These heuristics can be viewed as a linguistically-motivated expansion of the formal requirements that are verified by the UD-validator. We referred to this class of heuristics as *Guidelines based*.

Each of these three classes includes a set of finer-grained heuristics that we describe in what follows. Note that, in the examples, the original wrong sentence will be marked with the progressive number of the example between parenthesis and using the letter *O* to refer to the Original sentence annotation, while the letter *C* will indicate the Corrected version.

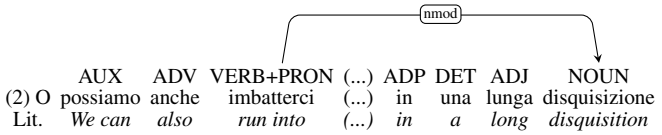
General Purpose Heuristics

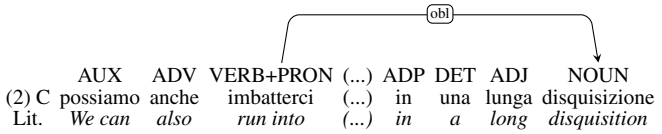
– **Non-projective dependencies** (*non-projective* in Table 2). This pattern refers to cases where the projectivity of one or more dependency relation(s) is violated since there is no path from the head to every word that lies between the head and the dependent, such as in example (1) where the comma was erroneously headed by *indíco* 'I call for' thus creating a non-projective relation `punct` that crosses the `acl` relation, while, on the contrary, according to the UD guidelines a punctuation mark preceding or following a dependent unit (represented here by *presentato* 'submitted') is attached to the unit, as reported in the correct version:



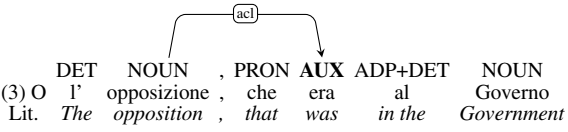


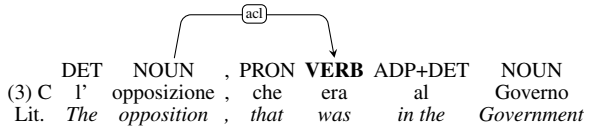
– **Nominal modifiers** (*nominal mod*). It refers to cases where an oblique modifier `obl` was erroneously annotated as a nominal modifier `nmod`, such as in example (2) where the noun *disquisizione* 'disquisition' was incorrectly linked by the dependency relation `nmod` (allowed only when the head is a noun or noun phrase) rather than `obl`:



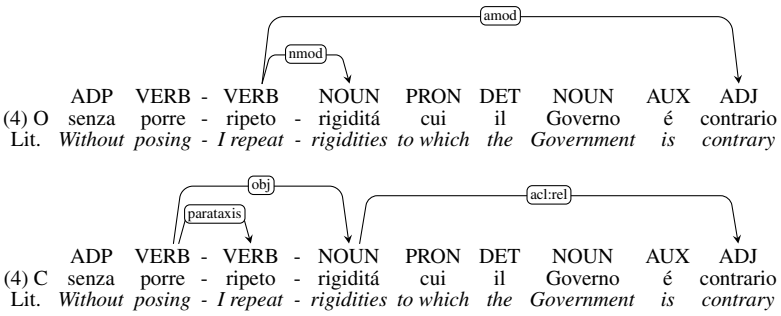


– **Auxiliary verbs** (*aux-verb*). It refers to the erroneous morpho-syntactic categorization of auxiliary verbs (e.g. *essere* 'to be', *avere* 'to have') mislabelled as auxiliaries (`AUX`) rather than main verbs (`VERB`), such as in (3) where *era* 'was' erroneously labeled as `AUX` rather than `VERB` since it is used in a presentational construction where it has a locative meaning:



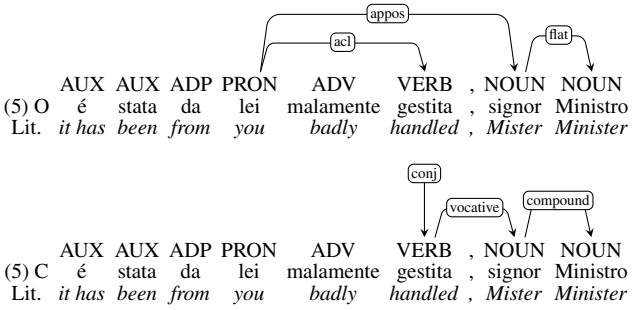


– **Adjectival modifiers** (*adj-mod*). It refers to cases where adjectives serving as clausal modifiers (`acl`) were erroneously annotated as adjectival modifiers (`amod`), as in (4) where the adjective *contrario* 'opposed' part of the lexical predicate *é contrario* 'is contrary' was erroneously headed by the verb *ripeto* 'repeat', and linked by the relation `amod`, rather than by the noun *rigiditá* 'rigidities' thus functioning as the head of the relative clause that modifies the noun, labeled as `acl:rel`:



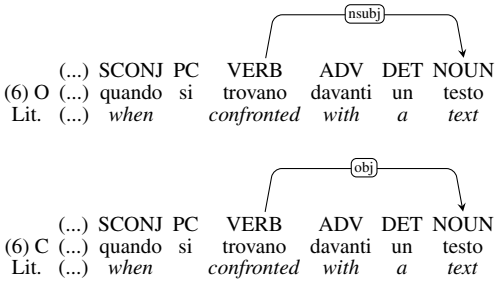
Corpus specific Heuristics

– **Vocatives** (*honorifics*). It refers to cases where honorific titles addressed in the speech (e.g. 'Mister President') were erroneously treated as appositions (`appos`) of the addressee's name but function clearly as vocatives (`vocative`), possibly due to a wrong head assignment, as in (5) where *signor* 'Mister' was erroneously annotated as apposition headed by *lei* 'you' rather than vocative linked to its host sentence:



Such use of vocative is quite typical of parliamentary speeches since it serves to attract the attention of specific parliament members or chairpersons.

– **Nominal subjects and objects** (*nsubj-obj*). It refers to cases where either nominal subjects (`nsubj`) or objects (`obj`) were erroneously annotated. To exemplify, consider (6), where *testo* 'text' functions as a direct object (`obj`) rather than a nominal subject (`nsubj`) headed by the verb *trovano* 'confronted':



It is widely acknowledged that the disambiguation between subjects and objects is generally critical for parser (McDonald et al., 2007). This results particularly challenging in ParlaMint's transcriptions where sentences are characterized by subordinate clauses embedded in complex syntactic structures[12].

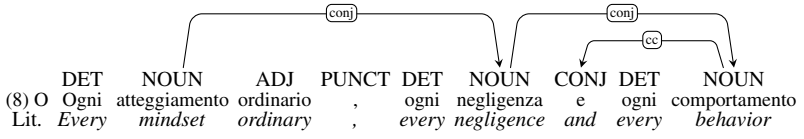
– **Homographs** (*art-pron* and *che-homograph*). It refers to cases where the particle *che* 'that', homograph of both the relative pronoun and the subordinating conjunction, or forms of determinative articles that are homographs of clitic pronouns (or vice-versa), have been erroneously lemmatized, such as in (7) where the clitic pronominal form *lo* 'this', homograph of the masculine singular determinative article used before a consonant, has been erroneously lemmatized as *il* which is the base form of the masculine determinative article 'the' instead as *lo*, as according to the UD guidelines the lemma of a clitic pronoun is the repeated form:

	CCONJ	PRON	VERB	SCONJ
	e	**il**	dire	perchè
(7) O	e	lo	diciamo	perchè
Lit.	*and*	*this*	*we say*	*since*

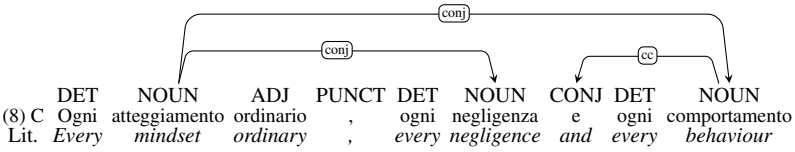
	CCONJ	PRON	VERB	SCONJ
	e	**lo**	dire	perchè
(7) C	e	lo	diciamo	perchè
Lit.	*and*	*this*	*we say*	*since*

Guidelines based Heuristics

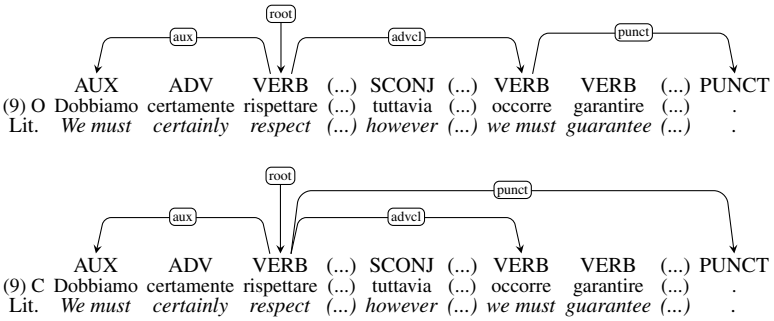
– **Conjuncts** (*conjuncts*). It refers to cases where a sequence of coordinating elements has not been correctly recognized. Consider (8), where the noun *comportamento* 'behavior' has been erroneously headed by the second conjunct *negligenza* 'negligence' rather than by the first one, i.e. *atteggiamento* 'mindset', in accordance to the UD guidelines:



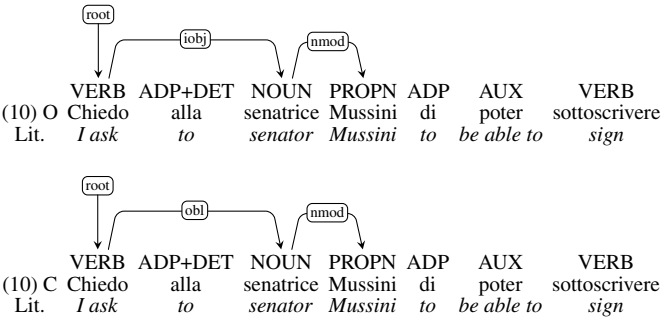
[12] The example was extracted by the following sentence: *Dobbiamo stare molto attenti, perchè ogni giudice ed ogni avvocato, quando si trovano davanti un testo, per ogni bambino sono poi tenuti a valutare e a soppesare tutto quello che é contenuto nel testo.* ('We must be very careful, because every judge and every lawyer, when confronted with a text, is then required to evaluate and weigh everything in the text.')



– **Punctuation** (*punct*). It refers to cases where the last punctuation mark has not been headed by the `ROOT` thus violating the UD guidelines, such as in (9) where the full stop of the sentence was headed by *occorre* 'we must' rather than by *rispettare* 'respect' which is the `ROOT` of the sentence:



– **Indirect objects** (*iobj*). It refers to cases where nominals that serve as oblique arguments or adjuncts were erroneously labeled as `iobj`, which should only be used (following the UD guidelines) when the function is carried out by a clitic pronoun. In (10), the syntactic role of the indirect object `iobj` holding between the noun *senatrice* 'senator' and the head verb *chiedo* 'ask' was erroneously identified instead of the correct oblique nominal relation `obl` that here functionally corresponds to an adverbial attachment (introduced by an articulated preposition) to a verb:



4 Building the 18K treebank

In this section, we will present the '18-karat gold' ParlaMint-It treebank, discussing, in particular, the errors encountered in the treebank during the revision process. Before delving into these details, let us first present the results of the evaluation process,

which allows us to assess the effectiveness of the revision methodology employed. As outlined in Sect. 3, our revision primarily targeted sentences with a higher likelihood of containing systematic errors, as well as random inconsistencies. Consequently, our evaluation aimed to determine whether the remaining sentences (i.e., those not directly targeted by the filtering heuristics) were correctly annotated or, at most, contained only a few random errors. To achieve this goal, we randomly selected and manually validated 20% of the sentences of ParlaMint-It, which amounted to 75 sentences. These sentences were then subjected to fully manual validation. Note that we made sure to exclude from this subset any sentences that had already undergone the revision process.

Such an evaluation revealed that 68% of the sentences were correctly annotated, demonstrating that the revision methodology we adopted is both efficient and reliable. In fact, while not all sentences underwent manual revision, the correction process targeted the portions of the corpus most likely to contain annotation errors, resulting in a high-quality resource, which justifies deeming it as '18-karat gold' treebank. Analyzing these results in more detail, we notice not only that a few sentences (24) required a manual correction, but also that their errors were mostly random and sparse. In fact, only 45 tokens were corrected, accounting for 5.7% of the tokens in the modified sentences. On average, there were 2.33 corrections per sentence, which is quite low considering the average sentence length in this subset is approximately 33 tokens.

Overall, the evaluation confirmed the quality of the ParlaMint-It resource and the effectiveness of the revision process. The semi-automatic revision process successfully removed recurrent errors in the annotations, which arguably affect the most the quality of the annotation, leaving behind only instances of low-impact sparse errors. Note that the statistics reported in the next sections refer to the final release of ParlaMint-It, thus they comprise also the errors corrected during the evaluation process.

4.1 Revision statistics

Overall, we corrected 339 sentences, which correspond to 48.36% of ParlaMint-It sentences (see Table 1). More specifically, we modified 1,378 tokens, corresponding approximately to 10% of the total amount of tokens in the set of modified sentences. For each token, the annotation was corrected at one or more annotation levels depending on the type of error encountered, as we will show in Sect. 4.3. Overall, we observe that the distribution of random and systematic errors is quite evenly distributed: 41.87% of the revised tokens reported systematic errors and 58.13% random ones. The 2015 sub-section is the one showing the higher percentage of errors: approximately 63% of the erroneously annotated tokens were found in this portion.

The distribution of systematic and random errors in the set of corrected tokens indicates that the correction methodology, although based on heuristics that target recurring errors, is effective for identifying both types of incorrect annotations. This effectiveness stems from the methodology's ability to isolate sentences that are most likely to report annotation errors. To illustrate, we note that each modified sentence in ParlaMint-It received an average of 5.12 corrections, where a correction is defined as any manual change made to any level of a token annotation. Consistent with the

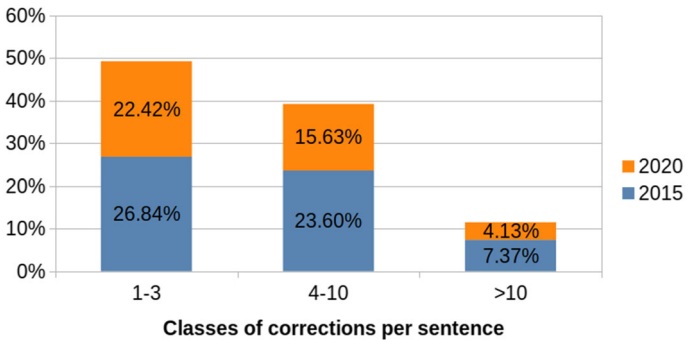


Fig. 2 Sentence distribution with respect to the number of annotation corrections

observations above, the 2015 subsection shows a higher incidence of corrections per sentence (5.54) than the 2020 subsection (4.54). The difference in the distribution of corrections between the two subsections is particularly evident when examining Fig. 2, which displays the distribution of modified sentences of ParlaMint-It with respect to the number of corrections they received. As the figure illustrates, around 50% of the modified sentences received more than four corrections, with those requiring a higher number of corrections (> 10) mainly located in the 2015 portion. Conversely, sentences that required only a few corrections are more evenly distributed between the two subsections, mainly corresponding to cases where the error concerned a local phenomenon identified using the error heuristics, as discussed below. Before moving to the in-depth discussion about the typologies of systematic errors encountered during the revision process, we provide insights into the various types of random errors that were corrected during the revision.

Random errors mainly concern head attachments (around 80%), mostly involving cases of `punct` (61.67%) and `nmod` (10.31%). Within this subset of head changes, approximately 20% of corrections involve simultaneous modifications to both the syntactic head and the associated dependency relation type. Random errors involving modifications to the dependency type, regardless of changes to other annotation levels, constitute about 28% of the overall corrections. Around half of these cases cover changes from nominal modifiers (`nmod`) to oblique (`obl`) and vice-versa, or clausal complements (i.e., `xcomp` to `ccomp`).

4.2 Typologies of systematic errors

Table 2 reports the distribution of the different typologies of systematic errors corrected in the entire corpus and the two sub-sections.

As it can be noted, errors falling into the general purpose group are the most frequently occurring in ParlaMint-It, both in the full corpus and sub-corpora. Among them, the heuristic capturing cases of *non-projective* dependency relations was the most effective for pointing to annotation errors. While non-projectivity may not be considered a proper error in the UD schema, especially in languages with free word order where relation heads can be discontinuous (Kuhlmann & Nivre, 2006), it can

hinder the performance of parsing algorithms and lead to incorrect parses (Jurafsky & Martin, 2023). Therefore, it is recommended to minimize the use of non-projective dependencies in UD.

The annotation of oblique nominals (captured by the *nominal mod* heuristic) represents another frequent source of error in ParlaMint-It annotation. In this case, non-core arguments of verbs were erroneously annotated as `nmod`, a type of relation reserved to nominal dependents of nouns or noun phrases. The relatively recent revision to the guidelines concerning the annotation of these relations caused the inconsistent use of `obl` and `nmod` in the Italian treebanks, which thus lead to inconsistent parses, as already observed in Alzetta et al. (2017). Notably, the somewhat minimal effectiveness of certain heuristics, such as *aux-verb* and *adj-amod*, may be attributed to revisions made to the Italian UD treebanks used for building the Stanza model. These revisions were guided by the insights from the work of Alzetta et al. (2017), where these patterns demonstrated considerable effectiveness.[13] The revisions made on the Italian UD Treebank appeared to have improved the homogeneity and internal coherence of annotations for these specific constructions, which is reflected in the output of the parser.

The heuristics in the other groups, namely corpus-specific and guidelines-based, exhibit a similar distribution, yielding a comparable number of errors. Table 2 shows, in particular, the effectiveness of the *honorifics* and *conjunct* heuristics for the correction of vocatives and coordinating constructions respectively, especially in the 2020 portion. Such a difference in the distribution of these error types in the two sub-corpora may be due to the different topics accounted for during the parliamentary sessions covered by the two portions. As introduced in Sect. 2, 2015 reports the transcripts of a parliamentary debate about the appropriateness of the adoption and foster care institutions. Such an in-depth discussion on legal amendments is characterized by a lower emotional involvement as all speeches were full of legal jargon and specific references to laws. In contrast, the 2020 sub-corpus pertains to discussions surrounding COVID-19 and jail uprisings, which are more urgent and rhetorical in tone. Although the structure of the discussions is more linear and therefore possibly easier to analyze for a parser, as indicated by the lower amount of errors pinpointed by most of the other heuristics, the 2020 section shows a more spontaneous and speech-like language, featuring rhetorical constructions and orality-related traits.

As such, each of the two sub-corpora presented its own peculiarities and, consequently, different types of most frequently occurring errors.

4.3 Annotation element involved in the revisions

We now move to assess the extent to which the annotation revision process impacted each annotation level. This evaluation is based on an analysis of both systematic and random errors across the whole ParlaMint-It corpus given their comparable distribution in the two sub-corpora.

[13] In Alzetta et al. (2017), the *aux-verb* heuristic accounted for 13.32% of systematic errors, while errors captured by the *adj-mod* heuristic covered 12.52% of systematic errors.

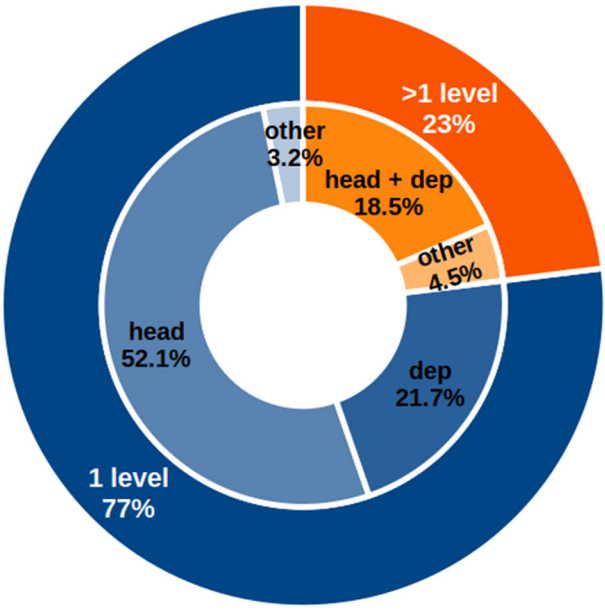


Fig. 3 Errors distribution by annotation levels in the ParlaMint-It corpus: the external circle shows how many annotation levels are involved in the corrections; the inner circle indicates the levels involved (*'head'*: identifier of the head of the token; *'dep'*: type of Universal dependency relation to the head)

As depicted in Fig. 3, corrections concerned a single level of the annotation in 77% of cases.[14] Notably, we observe that the syntactic annotation of ParlaMint-It was commonly subject to correction, particularly regarding the head assignment of dependency relations, which cover 72% of cases of errors found in ParlaMint-It. These corrections include cases of single-level corrections (marked as *'head'* in Fig. 3) as well as corrections involving also dependency type assignment (*'head+dep'*) or some other annotation level (*'other'* in the *'>1 level'* portion).

Upon closer examination of these results, we observe that the head assignment errors were primarily related to the annotation of punctuation markers (43.29%).

The *punct* and, most importantly, the *non-projective* heuristics played a key role in identifying those errors, respectively belonging to the guidelines-based and general purpose groups. Indeed, the annotation of punctuation markers is prone to generate non-projective relations (see guidelines for `punct`[15]) and is fairly inconsistent both within and across treebanks possibly due to ambiguous instructions and guideline revisions across versions, concerning in particular coordinated structures. This may lead to inconsistent parses that we corrected in order to improve the coherence of punctuation annotation in ParlaMint-It, especially for parenthetical phrases.

When we move to the inspection of the types of dependency relation we corrected, we can observe that the nominal modifiers (`nmod`) is the most modified typology when the revision process concerned both only the type of dependency (*'dep'* in Fig. 3), and

[14] The distribution of corrections for dependence type is reported in Appendix B.

[15] https://universaldependencies.org/u/dep/punct

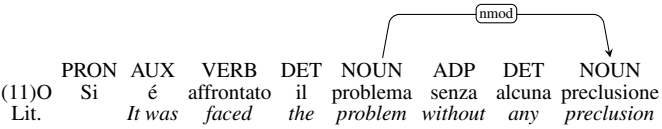
the syntactic head and the type (*'head+dep'*). By inspecting the corresponding correct relations, we noticed that the `nmod` instances were mostly changed into oblique arguments (`obl`) and mainly emerged from the *nominal modifier* heuristic. As mentioned in Sect. 4.2, the `obl` relation was introduced in UD 2.0 to annotate oblique arguments of verbs, previously marked using the `nmod` relation, which now identifies nominal dependents of nouns. Possibly due to such a revision of the universal annotation guidelines, we found many errors in the use of these two relations, which, in most cases, turn out to be inverted.

A less common, but still relevant, typology of labeling errors involves the annotation of `vocative` dependencies that were detected using the *vocatives* heuristic. This could be due to the fact that this type of syntactic relation is rarely used in Italian UD treebanks (less than 1% of instances in ISDT) thus representing a peculiarity of the ParlaMint-It textual genre. In fact, vocatives are typical of speech-like parliamentary debates as they are used to address the other person and attract his/her attention during the speech. The majority of mislabeled dependencies of this type occur in the 2020 section of ParlaMint-It in line with the greater rhetorical emphasis and spontaneity of the speeches held during the COVID-19 pandemic period. As a result, a related type of erroneously annotated dependency concerns the honorific titles of the people being addressed. They are annotated with a different dependency relation depending on their internal composition, e.g. 'Mister President' (`compound`), 'Senator Falanga' (`nmod`) and 'Falanga, the Senator' (`appos`), with titles followed by a proper noun composition often annotated as `appos` instead of `nmod`.

Lastly, we observe inconsistent annotation of conjunction groups, which were in some cases not recognised as such. The *'conjunct'* heuristic allowed identifying those cases and restoring their correct annotation.

4.4 Head-dependent variations after revision

Considering that the corrections involving head assignment, either alone or in combination with other annotation levels, are the most frequent in ParlaMint-It, we investigated further the nature of those changes. To this aim, we focused on relation lengths (namely, the linear distance in tokens between a head and its dependent) and how they vary after correction. Specifically, dependency length variation is computed as the difference between the length of the erroneous relation and its length after correction. To clarify, consider the example below. In the automatic annotation, *preclusione* 'preclusion' is linked to the head *problema* 'problem' by a 3-token relation, while when we manually modified the sentence and we moved the head to *affrontato* 'faced' the relation became 5-token long. Thus, the dependency length variation between them is +2 tokens.



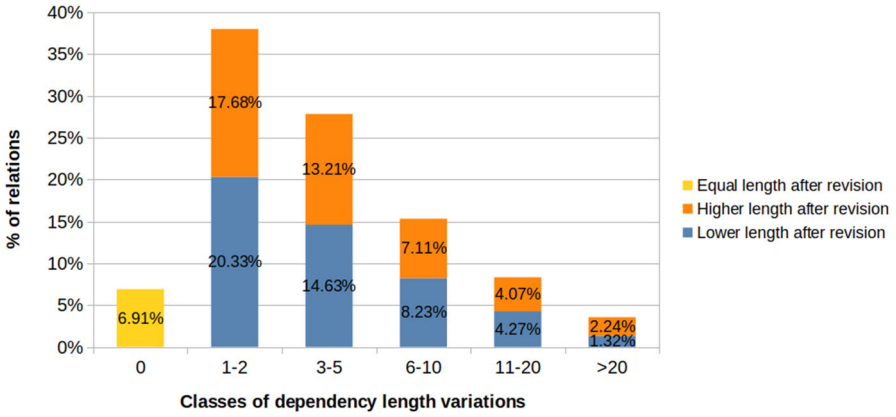


Fig. 4 Relations distribution with respect to their dependency length variation after revision

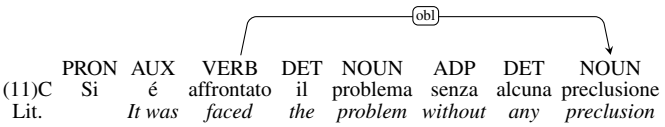
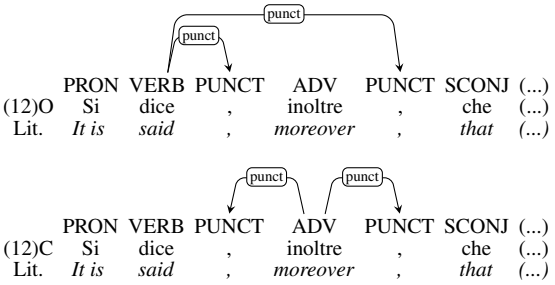


Figure 4 depicts the distribution of relations with respect to the variation of their dependency length after the head attachment revision. The majority of corrections resulted in minimal variations, with 65.85% of cases showing a variation from 1 to 5 tokens maximum. This often occurred when the parser erroneously assigned a token's head to a word near the correct head in the sentence's linear order, as in example (11) above. Interestingly, when the difference in length between the original and revised versions was $\leq$10 tokens, corrections tended to decrease the distance between the head and its dependent (43.19%) rather than increase it (38%). Notably, this pattern reversed when the dependency length variation was $\geq$20, i.e. instances where the two elements were distant in the sentence, posing a challenge for the parser to determine the correct dependency. Figure 3 also presents a 6.91% of relations that maintained an unchanged length. These are cases where the number of tokens between the erroneous and corrected head-assignment pairs remained the same, resulting in a dependency length variation of 0. To exemplify, consider example (12): the length of the `punct` relation connecting the first comma to the head *dice* 'said', remains unaltered after the revision process. Conversely, the other `punct` relation became shorter after revision, passing from 3-token long to 1-token long, showing a 2-tokens variation.

We conducted a thorough analysis to examine possible variations in the head/dependent order of modified dependencies, regardless of the variation in dependency length. The results revealed that in most cases (57%) the word order was not modified by the manual revision process. This further seems to confirm that the majority of manual revisions did not modify the structure of sentences. Yet, punctuation, the main relation type with head/dependent order changes, is the most altered dependency type during revisions. As proof, consider again example (12), where one of the two links had its direction reversed.



5 Discussion and conclusions

In this paper, we have presented the results of the building process of a new resource for the Italian language, ParlaMint-It, morpho-syntactically annotated in accordance with the Universal Dependencies (UD) formalism. One of the main peculiarities of the resource concerns the language variety it covers which is quite underrepresented in the current UD repository, i.e. parliamentary speeches, featuring characteristics specific both to the government and legal textual genres.

The semi-automatic revision methodology we adopted yielded an '18-karat gold' treebank of 20,460 tokens partially manually corrected for systematic and random annotation errors at different levels of linguistic description. The analysis of the set of recurrent patterns of errors showed that the two most common typologies correspond to cases of i) non-projective dependency relations yielding uncorrected annotations and ii) oblique verbal arguments erroneously annotated as nominal modifiers. Quite interestingly, they are followed by systematic errors targeting characteristics specific to the parliamentary speeches, i.e. the use of honorific titles functioning as vocatives and of coordinating constructs occurring within sequences either of normative references or of elements listed by the speakers to give a sense of emphasis to their speeches. In addition, our analysis revealed that most of the corrections made were local and involved only one level of annotation. In fact, 52% of cases concerned only the head assignment of dependency relations, which primarily involved punctuation markers, and consisted of minimal variations of the head-dependent distance measured in terms of dependency length. The random errors encountered during revision mostly concerned head attachments, however, these errors appeared in sparse and diverse contexts that we could not systematize using heuristics.

These outcomes seem to confirm the high level of annotation accuracy currently achieved by the automatic linguistic annotation pipelines and seem to suggest a primary focus on the revision of local aspects of sentence structure rather than on global ones. This, in turn, minimizes an extensive manual revision that is only necessary for correctly representing textual genre peculiarities or for dealing with phenomena (e.g. non-projective relations and punctuation markers) whose annotation is possibly not homogenous in the gold treebanks and that can result in additional errors.

In our opinion, the presented initiative can have multiple future directions. One obvious direction would involve expanding the semi-automatic revision process that we introduced to the whole Italian section of the ParlaMint corpus to integrate the full

section of the ParlaMint corpus into UD. Starting from the correction of sentences instantiating the constructions that correspond to the filtering heuristics introduced in this study, we might additionally enlarge the heuristics thus finding new patterns of errors. This will result in an '18-karat gold' quite large resource featuring domain-specific texts. Additionally, a similar semi-automatic revision process could be applied to the other corpora collected during the European ParlaMint project, thus creating new gold multilingual treebanks of parliamentary speeches. Multilingual corpora sharing the same formalism of linguistic annotation might represent the starting point of a large-scale cross-linguistic initiatives ranging from Machine Translation applications to studies based on the principles of Computational Stylometry. In fact, ParlaMint treebanks can serve to identify patterns in the content and style of parliamentary speeches, such as the length of the speech, distribution of lexicon, use of subordinating constructions, marked order of specific dependency relations. This would provide insights into characteristics unique to an individual speaker or group of speakers sharing the same metadata (e.g. political party or parliamentary group, the government in charge).

Furthermore, ParlaMint-It sentences could prove valuable when used as signals for selecting new training data for zero-shot scenarios, as highlighted by Müller-Eberstein et al. (2021a). This is specifically because the genre information is reported at the sentence level rather than the treebank level, which represents a main limit to current approaches to new target data selection. Quite interestingly, the hybrid nature of parliamentary speeches, being between two genres, represents a further challenge in such a direction.

Appendix A: Identifiers of revised sentences

The following are the identifiers of sentences in UD ParlaMint-It manually corrected using the semi-automatic error revision methodology outlined in the paper.

2015 Section Add `ParlaMint-IT_2015-03-11-LEG17-Sed-407_` in front for proper referencing													
4	45	65	80	98	121	147	185	210	227	244	266	295	331
15	47	66	81	99	122	148	186	211	228	245	267	296	339
21	49	67	83	101	128	153	190	212	230	247	268	297	340
22	50	68	85	102	129	154	192	213	232	248	275	299	341
25	51	69	87	104	130	157	193	214	233	249	276	302	342
26	55	70	88	105	131	158	198	215	234	250	278	305	343
28	57	71	89	106	134	159	201	216	235	251	280	308	344
31	58	72	90	107	136	161	202	217	236	253	286	312	346
35	59	73	91	109	137	168	203	218	237	254	287	314	350
36	60	75	92	110	138	169	204	221	238	255	288	319	352
38	61	76	93	111	139	170	205	223	240	258	290	321	353
39	62	77	94	113	141	175	207	224	241	261	291	326	355
40	63	78	95	114	143	183	208	225	242	262	293	327	358
41	64	79	96	115	145	184	209	226	243	265	294	330	359

2020 Section Add `ParlaMint-IT_2020-03-11-LEG18-Sed-200_` in front for proper referencing												
6	34	53	74	104	126	147	199	229	244	264	285	311
10	35	54	77	106	127	148	201	230	245	268	286	314
12	42	55	85	107	128	151	207	231	246	272	290	320
13	43	56	86	108	129	152	210	232	248	273	292	323
15	44	57	88	110	133	169	214	233	249	275	293	324
21	45	59	90	111	135	175	216	234	251	276	294	325
23	48	66	95	115	140	176	218	237	254	277	295	326
24	49	68	96	118	141	181	221	238	258	279	299	328
26	50	70	97	119	142	184	222	239	259	281	300	331
30	51	72	99	121	145	189	225	241	260	283	306	333
33	52	73	101	124	146	197	227	242	261	284	310	339

Appendix B: Corrections distribution per dependency type

Table 3 The table illustrates the distribution of manual corrections per dependency type in UD ParlaMint-it

Dependency type	Level modified during revision				Total count of links modified	% over links for dep. type in corpus
	Head	Dep. Type	Head+ dep. type	Other		
acl	29	5	18	2	54	8.94%
advcl	29	17	7	3	56	15.91%
advmod	31	2	3	6	42	4.21%
amod	4	3	8	6	21	2.35%
appos		24	7		31	48.44%
aux	5		1		6	0.84%
case	17	2	6	8	33	1.22%
cc	24		1	2	27	4.58%
ccomp		7	13		20	10.64%
compound		2	3	1	6	7.14%
conj	39	6	11	4	60	8.49%
cop	4	1	9		14	5.38%
csubj		2	2		4	8.33%
det	11	2	1	6	20	0.62%
discourse		2			2	33.33%
expl	1	2			3	1.14%
fixed			10		10	10.64%
flat	2	10		1	13	20.97%
iobj		8		9	17	28.33%
mark	6	4	3	2	15	2.08%
nmod	43	107	60	9	219	14.78%
nsubj	9	21	15	12	57	6.21%

Table 3 continued

Dependency type	Level modified during revision				Total count of links modified	% over links for dep. type in corpus
	Head	Dep. Type	Head+ dep. type	Other		
nummod	1	1	2		4	2.15%
obj	5	29	12	14	60	6.67%
obl	14	12	28	7	61	5.3%
parataxis	3	1	10		14	46.67%
punct	425	1		4	430	19.65%
reparandum				1	1	100%
root			8	8	16	2.28%
vocative	3		9	12	24	31.17%
xcomp		32	5		37	17.45%

It outlines the number of links that were changed during the revision for each dependency type, specifying the annotation level corrected. To enhance comparability, the last column presents the percentage of modified links relative to the total number of links for that type in the automatically parsed corpus, i.e., before the revision

Acknowledgements We thank Professor Maria Simi for her valuable contribution to the annotation revision of the ParlaMint-It UD resource.

Funding Open access funding provided by ILC - PISA within the CRUI-CARE Agreement. This work was supported by the Dipartimento di Informatica at University of Pisa – Borsa di studio e approfondimento (Fondo "Google Gift") assigned to Marta Sartor.

Declarations

Conflict of interest The authors declare that they do not have any conflict of interest.

Ethical approval Our work has limited ethical implications since we mainly introduced a novel treebank enriched with morpho-syntactic annotations compliant with the Universal Dependencies standard. The ParlaMint treebank from which ParlaMint-It originates was used in compliance with the Terms of Use and the resources and materials produced during this study will be distributed in compliance with the license agreement of the UD project.

References

Agnoloni, T., Bartolini, R., & Frontini, F., et al. (2022). Making Italian parliamentary records machine-actionable: The construction of the parlamint-it corpus. In *Proceedings of the workshop ParlaCLARIN*

III within the 13th language resources and evaluation conference. European Language Resources Association, Marseille, France, pp. 117–124.
Agrawal, B., Agarwal, R., Husain, S., et al. (2013). *An automatic approach to treebank error detection using a dependency parser* (pp. 294–303). Springer.
Alzetta, C., Dell'Orletta, F., & Montemagni, S., et al. (2017). Dangerous relations in dependency treebanks. In *Proceedings of the 16th international workshop on treebanks and linguistic theories* (pp 201–210).
Alzetta, C., Dell'Orletta, F., Montemagni, S., et al. (2020). Linguistically-driven selection of difficult-to-parse dependency structures. *IJCoL Italian Journal of Computational Linguistics, 6*(6–2), 37–60.
Ambati, B. R., Agarwal, R., & Gupta, M. et al. (2011). Error detection for treebank validation. In *Proceedings of 9th international workshop on Asian Language Resources (ALR)*.
Arnardóttir Þ, Hafsteinsson, H., & Sigurðsson, E. F., et al. (2020). A universal dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the fourth workshop on universal dependencies* (UDW 2020). Association for Computational Linguistics, Barcelona, Spain (Online) (pp. 16–25).
Bosco, C., Montemagni, S., & Simi, M. (2013). Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*. Association for Computational Linguistics, Sofia, Bulgaria (pp. 61–69).
Boyd, A., Dickinson, M., & Meurers, W. D. (2008). On detecting errors in dependency treebanks. *Research on Language & Computation, 6*(2), 113–137.
Croft, W. B., Nordquist, D., & Looney, K., et al. (2017). Linguistic typology meets universal dependencies. In *International workshop on treebanks and linguistic theories*
Dell'Orletta, F., Venturi, G., & Montemagni, S. (2013). Linguistically-driven selection of correct arcs for dependency parsing. *Computaciòn y Sistemas, 2*, 125–136.
Dell'Orletta, F., Venturi, G., & Montemagni, S. (2013). Linguistically-driven selection of correct arcs for dependency parsing. *Computación y Sistemas, 17*(2), 125–136.
Dickinson, M., & Meurers, W. D. (2003). Detecting inconsistencies in treebank. In *Proceedings of the second workshop on treebanks and linguistic theories* (TLT 2003).
Dickinson, M., & Meurers, W. D. (2005). Detecting errors in discontinuous structural annotation. In *Proceedings of the 43rd annual meeting of the ACL* (pp. 322–329).
Erjavec, T., & Pančur, A. (2019). Parla-CLARIN TEI guidelines for corpora of parliamentary proceedings. https://doi.org/10.5281/zenodo.3446164
Erjavec, T., Ogrodniczuk, M., Osenova, P., et al. (2022). The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0
Fišer, D., Eskevich, M., de Jong, F. (eds.) (2020). *Proceedings of the second ParlaCLARIN Workshop*, European Language Resources Association (ELRA), Marseille, France. https://www.aclweb.org/anthology/2020.parlaclarin-1.0
Fišer, D., Eskevich, M., & Lenardič, J. et al. (eds.). (2022). *Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation conference*. European Language Resources Association, Marseille, France. https://aclanthology.org/2022.parlaclariniii-1
Fišer, D., Eskevich, M., de Jong, F. (eds.). (2018). *Proceedings of LREC 2018 workshop ParlaCLARIN Creating and using parliamentary corpora*, European Language Resources Association (ELRA), Paris, France. http://lrec-conf.org/workshops/lrec2018/W2/pdf/book_of_proceedings.pdf
Fort, K., Nazarenko, A., & Rosset, S. (2012). Modeling the complexity of manual annotation tasks: A grid of analysis. *Proceedings of COLING, 2012*, 895–910.
Hladká, B., Hajic, J., Hana, J., et al. (2008). The czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics, 89*, 41.
Ilie, C. (2015). Parliamentary discourse. *The International Encyclopedia of language and social interaction* (pp. 1–15).
Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing (3rd edn)*. Prentice-Hall.
Kondratyuk, D., & Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China (pp. 2779–2795). https://doi.org/10.18653/v1/D19-1279, https://aclanthology.org/D19-1279
Kríž, V., Hladká, B., & Urešová, Z. (2016) .Czech legal text treebank 1.0. In *Proceedings of the tenth international conference on language resources and evaluation* (LREC'16) (pp. 2387–2392).

Kuhlmann, M., & Nivre, J. (2006). Mildly non-projective dependency structures. In *Proceedings of the COLING/ACL 2006 main conference poster sessions* (pp. 507–514).

Lynn, T., & Foster, J. (2016). Universal dependencies for Irish. In *Proceedings of the second Celtic language technology workshop*.

de Marneffe, M., Grioni, M., & Kanerva, J., et al. (2017). Assessing the annotation consistency of the universal dependencies corpora. In *Proceedings of the 4th international conference on dependency linguistics* (Depling 2007), Pisa, Italy (pp. 108–115).

de Marneffe, M. C., Manning, C. D., Nivre, J., et al. (2021). Universal dependencies. *Computational Linguistics, 47*(2), 308. https://doi.org/10.1162/coli_a_00402

McDonald, R., & Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning* (EMNLP-CoNLL). Association for Computational Linguistics, Prague, Czech Republic (pp. 122–131). https://www.aclweb.org/anthology/D07-1013

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2021a). Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (pp. 4786–4802)https://doi.org/10.18653/v1/2021.emnlp-main.393, https://aclanthology.org/2021.emnlp-main.393

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2021b). How universal is genre in universal dependencies? In *Proceedings of the 20th international workshop on treebanks and linguistic theories* (TLT, SyntaxFest 2021). Association for Computational Linguistics, Sofia, Bulgaria (pp. 69–85). https://aclanthology.org/2021.tlt-1.7

Nencioni, G. (1976). Parlato-parlato, parlato-scritto, parlato-recitato. Strumenti critici (29).

Nivre, J., de Marneffe, M. C., Ginter, F., et al. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the twelfth language resources and evaluation conference*. European Language Resources Association, Marseille, France, pp. 4034–4043. https://aclanthology.org/2020.lrec-1.497

Pyysalo, S., Kanerva, J., & Missilä, A., et al. (2015). Universal Dependencies for Finnish. In *Proceedings of NoDaLiDa* 2015. NEALT, pp 163–172, https://aclweb.org/anthology/W/W15/W15-1821.pdf

Qi, P., & Zhang, Y., Zhang, Y., et al. (2020). 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*. Association for Computational Linguistics, Online, pp. 101–108, https://doi.org/10.18653/v1/2020.acl-demos.14, https://aclanthology.org/2020.acl-demos.14

Sanguinetti, M., & Bosco, C. (2015). Parttut: The Turin University parallel treebank. In *Italian natural language processing within the PARLI project*

Volokh, A, & Neumann, G. (2011). Automatic detection and correction of errors in dependency treebanks. In *Proceedings of ACL-HLT* 2011.