Romance Reflexive Constructions Revisited

Verginica Barbu Mititelu[1], **Elena Irimia**[1], **Adriana Pagano**[2], **Ioana Buhnila**[3], **Roxana Ciolăneanu**[4,5]

[1]Romanian Academy Research Institute for Artificial Intelligence, Bucharest, Romania;
[2]Federal University of Minas Gerais, Brazil;
[3]Centre for Data Science in Humanities, Chosun University, South Korea
[4]Centro de Linguística da Universidade de Lisboa, Portugal;
[5]Institute of the Romanian Language, Bucharest, Romania

Abstract

We analyze the current annotation of reflexive constructions, i.e. verbal constructions marked by a reflexive clitic, across five Romance languages (French, Italian, Portuguese, Romanian and Spanish) in several Universal Dependencies treebanks (version 2.17). We discuss the morphologic, syntactic and semantic characteristics of such constructions in each of the languages considered, both from a theoretical perspective and from that of existing annotation. To address inconsistencies in the data and strengthen Universal Dependencies as a scaffold for the automatic conversion of morphosyntactic annotation into semantic representations (Uniform Meaning Representation), we propose a clear distinction between argumental and non-argumental uses of the reflexive clitic, and outline systematic ways to implement this distinction in the annotation guidelines. We also examine how some of the reported inconsistencies can be handled in the treebanks under study and discuss the extent to which these practices can be extended to other treebanks, within the same or across different languages.

Keywords: reflexive constructions, Romance languages, uniform meaning representation, syntactic and semantic arguments

1. Introduction

In this paper we focus on reflexive constructions, by which we refer to a family of verbal constructions marked by reflexive clitics, with forms varying across Romance languages, and with meaning encompassing reflexive, passive, impersonal, middle, and lexically specified uses.

The contributions of this paper are the following: (i) at a theoretical level, we document the types of reflexive constructions that need to be distinguished for annotation in Universal Dependencies (UD) (de Marneffe et al., 2021) treebanks; (ii) we report on the current state of the annotation of these reflexive constructions in UD version 2.17 treebanks for the languages under consideration (see below); (iii) we propose annotation guidelines for such constructions against the presented theoretical aspects; (iv) we consider possible ways of improving the uniformity and consistency of the annotation of such constructions in UD treebanks.

The perspective on reflexive constructions we adopt here is not only morpho-syntactic, but also semantic, in that it envisages automatic conversion of UD trees into their Uniform Meaning Representation (UMR) (Bonn et al., 2023).

UD annotation of reflexive clitics is highly relevant to UMR conversion because UMR relies on syntactic structure to determine semantic roles and argument identity. In UD, reflexive clitics may be annotated either as core arguments (obj, iobj, obl:arg) or as non-argumental markers (expl, expl:pass, expl:impers, expl:pv) (see Section 5). This distinction directly affects whether a reflexive construction is interpreted in UMR as involving role coreference (e.g., the same participant filling two semantic roles) or as a valency-changing or voice-related operation (e.g., passive, impersonal, anticausative). If UD annotation conflates these uses or applies them inconsistently across treebanks, automatic UMR conversion may incorrectly introduce or omit semantic participants. Therefore, consistent and linguistically grounded UD treatment of reflexive clitics is crucial for ensuring accurate semantic representation, especially in multilingual UMR pipelines.

The languages under study here are French, Italian, Portuguese (mainly the Brazilian variant, with occasional references to European[1] one as well), Romanian, and Spanish. They are all Romance

[1]For European Portuguese, CINTIL (Branco et al., 2011) (Mariana Avelãs and others, 2022) is the only available treebank. With approximately 18,000 sentences (around 450,000 tokens), it is an important resource and has been widely used in syntactic and morphosyntactic studies of European Portuguese. It was therefore relevant to assess whether it could be incorporated into our cross-treebank comparison. However, CINTIL could not be productively queried for the purposes of the present study. Although it tokenizes pronominal forms, its annotation scheme does not distinguish clitic from non-clitic pronouns, nor does it encode reflexivity through a dedicated morphological feature such as *Reflex=Yes*. In addition, it does not employ the UD relation *expl* or its subtypes (e.g. *expl:impers, expl:pass, expl:pv*). As a result, reflexive and non-reflexive uses of pronominal forms cannot be reliably separated, and neither reflexive-object queries nor dependency-based filtering of expletive con-

languages, each with its own evolution that has contributed to shaping its current characteristics. Reflexive constructions existed also in Latin, their common ancestor, and have been preserved in the contemporary languages. For each language, we selected one or more UD treebanks (see Section 3) to analyze the current annotation of their reflexive constructions.

In Section 2 we comment on limited body of related work on reflexive constructions in UD treebanks, in Section 3 we briefly present the treebanks we work with and our approach. The characteristics we consider relevant to the analysis of this phenomenon are compared across the five languages in Section 4, which examines the cross-linguistic morphological properties of reflexive clitics, and in Section 5, which addresses their annotation in treebanks at the syntactic and semantic levels. A discussion of the current annotation of reflexives in the UD treebanks considered here can be found in Section 6. We present some considerations relevant to a more consistent annotation of reflexives in Section 7, before concluding the paper.

2. Related Work

Despite the relevance of reflexive constructions for both syntactic analysis and downstream semantic representation, related work on their treatment in Universal Dependencies (UD) treebanks remains limited. Our paper comes as a next step after Marković and Zeman (2018), who showed the unsatisfactory annotation of reflexives in UD treebanks. They examine how reflexive markers are annotated in UD 2.2, with a particular focus on Slavic languages. The paper evaluates whether current UD guidelines are sufficiently clear and applied consistently across languages.

The authors show that reflexives have multiple functions – true reflexive, reciprocal, inherent/pronominal verb marker, passive marker, impersonal marker, and anticausative – and that their annotation in UD is highly inconsistent across treebanks. Problems include uneven use of the feature `Reflex=Yes`, inconsistent distinction between argumental uses (`obj`, `iobj`, `obl`) and non-argumental uses (`expl`, `expl:pass`, `expl:impers`, `expl:pv`), and divergent treatments of inherently reflexive verbs. Some treebanks overgeneralize a single label (e.g., treating all reflexives as `expl` or `compound`), while others fail to distinguish passive and impersonal constructions.

The authors argue that UD should, as an essential first step, distinguish true reflexive arguments from non-argumental reflexive constructions, and they advocate for clearer guidelines, more consistent use of subtypes of `expl`, and harmonization across languages. They conclude that improving reflexive annotation is crucial for cross-linguistic research and for maintaining UD's goal of typological consistency.

Duran et al. (2025) present a corpus-based study of the clitic *se* in Brazilian Portuguese based on 732 occurrences extracted from the UD-annotated Porttinari-base treebank. Their goal is to develop a more consistent and computationally useful annotation scheme for NLP. The analysis shows that traditional grammatical classifications (reflexive, reciprocal, middle, synthetic passive, impersonal, etc.) are difficult to apply consistently in contemporary Brazilian Portuguese, especially the distinction between synthetic passive and impersonal constructions.

To address this, the authors propose a streamlined annotation framework aligned with UD. They argue for simplifying the treatment of *se* in UD by: (i) clearly distinguishing argumental *se* (obj/iobj) from non-argumental uses; (ii) unifying impersonal and synthetic passive constructions under a consistent impersonalization analysis; and (iii) refining the annotation of pronominal and middle constructions. The proposal aims to increase annotation consistency and improve downstream NLP tasks such as information extraction and semantic role labeling.

Both Marković and Zeman (2018) and Duran et al. (2025) address the annotation of reflexive markers in UD, but from complementary perspectives. The former take a cross-linguistic and guideline-oriented approach, showing that reflexive annotation in UD is inconsistent across languages and advocating for clearer distinctions between argumental and non-argumental uses, as well as more systematic use of `expl` subtypes. In contrast, the latter focus specifically on Brazilian Portuguese, demonstrating empirically that traditional fine-grained grammatical classifications are difficult to apply consistently in corpus annotation and proposing a streamlined UD-oriented solution that prioritizes syntactic behavior and annotation reliability. Together, the two works highlight the tension between typological granularity and the need for practical consistency in UD reflexive annotation.

Issues in UD annotation, including those related to reflexive constructions, can be revisited with downstream applications in mind, given recent work highlighting the potential of UD as a scaffold for semantic annotation, particularly in multilingual settings. Gamba et al. (2025) propose a method for bootstrapping partial UMR graphs from UD parses, showing that UD's syntactic structures can be sys-

structions can be formulated in a way comparable to the other treebanks considered here. European Portuguese is therefore included only for contextual and comparative purposes, not as a fully queryable dataset in the empirical analysis.

tematically exploited to derive initial semantic representations that reduce manual annotation effort. Their approach leverages UD relations to construct partial UMR graphs that preserve event structure and core argument configurations, which can then be completed and refined by human annotators. This work demonstrates that, despite UD not being a semantic formalism, its cross-linguistic consistency and wide language coverage make it a practical starting point for scalable semantic annotation.

Building on Gamba et al. (2025), our study similarly exploits UD annotations to guide the identification of reflexive and expletive constructions across languages, while making explicit the limitations imposed by treebank-specific annotation choices. Reflexive constructions such as the one in Figure 1 illustrate how syntactic information encoded in UD can be exploited in the construction of semantic representations such as UMR. In the Portuguese sentence *A menina se olhou no espelho.* ("The girl looked at herself in the mirror."), UD can encode reflexivity through annotating the clitic pronoun (*se*) as an argumental object (obj), with the feature Reflex=Yes, which will be interpreted as coreferential with the subject *A menina*. In UMR, this syntactic configuration naturally maps onto a representation in which a single participant is assigned both actor and undergoer roles in the event, with reflexivity expressed via role identity rather than by introducing an additional argument. This type of mapping exemplifies how UD structures can provide a principled basis for bootstrapping semantic representations, as proposed in recent UD-to-UMR conversion work (Gamba et al., 2025).

3. Methodology

Our methodology involved four main steps: the selection of the languages considered, the selection of the treebanks examined, a cross-linguistic analysis of the relevant properties of these languages, and a cross-treebank analysis of the annotation of reflexive clitics. Within this framework, the identification of reflexive and reciprocal objects was carried out in a way that accommodates cross-linguistic and cross-treebank variation in UD annotation practices.

The selection of the Romance languages considered in this study was guided by the availability of treebanks. For each language selected on this basis, we then identified the following treebanks for analysis:

- French – (i) GSD (Guillaume et al., 2019) (Marie-Catherine de Marneffe and others, 2015), which represents contemporary written French spanning multiple genres (blog, news, reviews, wiki), and contains 16,342 sentences (400,387 tokens); (ii) Sequoia (Candito and Seddah, 2012) (Marie Candito and others, 2017), a treebank of written French composed of 3,099 sentences (70,567 tokens) from various sources: French Europarl, Wikipédia Fr, Newspaper *Est Républicain*, and European Medicines Agency. Both treebanks are regularly updated.
- Italian – (i) VIT (Venice Italian Treebank) (Delmonte et al., 2007; Alfieri and Tamburini, 2016) (Rodolfo Delmonte, 2007), represents contemporary written Italian across multiple genres and contains approximately 10,000 sentences (about 250,000 tokens); and (ii) ISDT (Italian Stanford Dependency Treebank) (Bosco et al., 2012; Sanguinetti and Bosco, 2014) (Bosco et al., 2015), composed primarily of newspaper text, includes approximately 13,000 sentences (about 300,000 tokens). Together, they provide a broad coverage of modern Italian usage.
- Brazilian Portuguese – (i) Porttinari (Duran et al., 2023) (Magali Sanches Duran and others, 2017) contains approximately 5,000 sentences (about 150,000 tokens); and (ii) Petrogold (de Souza and Freitas, 2022) (de Souza and Freitas, 2022) comprises approximately 3,000 sentences (about 80,000 tokens), providing complementary coverage of syntactic phenomena. Both represent contemporary Brazilian Portuguese and include newspaper and academic texts, respectively.
- Romanian – the Romanian Reference Treebank (Barbu Mititelu, 2018) (Verginica Barbu Mititelu and others, 2015) (RRT) was chosen because it represents contemporary written Romanian and includes texts from several genres (literature, medicine, law, news, science, academic writing, Wikipedia, etc.); the corpus contains 9,524 sentences (218,522 tokens).
- Spanish – the AnCora treebank (Taulé et al., 2008) (Héctor Martínez Alonso and Daniel Zeman, 2016), was used, a large manually annotated corpus of contemporary written Spanish covering multiple genres; it contains approximately 17,000 sentences (about 500,000 tokens) and is widely adopted in UD-based research.

The cross-linguistic analysis of the relevant properties of these languages consisted in comparing the main morphological properties of reflexive clitics across the languages considered, in order to establish the dimensions of variation relevant to the study. Syntactic and semantic aspects were then addressed separately through the cross-treebank analysis.

The cross-linguistic analysis of the relevant properties of these languages consisted in comparing

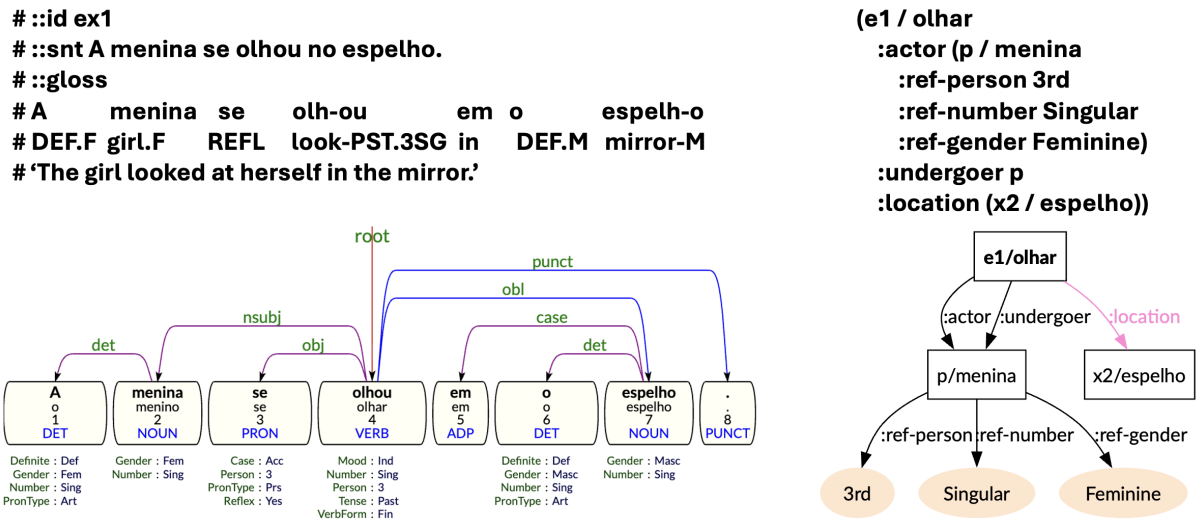

Figure 1: UD and UMR representations of sentence illustrating a reflexive construction with clitic *se* annotated as object in UD and mapped in UMR to a coreferential undergoer participant.

the main morphological properties of reflexive clitics across the languages considered, in order to establish the dimensions of variation relevant to the study. Syntactic and semantic aspects were then addressed separately through the cross-treebank analysis.

To identify reflexive and reciprocal objects, we adopted a strategy-based approach designed to accommodate cross-linguistic and cross-treebank variation in UD annotation practices. On the basis of this approach, we developed a set of language-specific query patterns, presented in Section 5. Direct and indirect objects were identified through agreement in person and number (for the first and second persons) or in number only (for the third person) with the head verb. Since first- and second-person reflexive objects are morphologically indistinguishable from non-reflexive object clitics in Romance languages, reflexivity cannot be identified solely on the basis of clitic form. Instead, it must be detected through form-based queries combined with a person and number agreement heuristic, operationalized by matching the person and number features of the object clitic with those encoded on the governing verb. This is illustrated by the Spanish contrast between *Me pregunté por qué pasó eso* ('1SG.obj ask-PST.1SG why happen-PST.3SG that', 'I asked myself why that happened'), which is retrievable as reflexive because the person and number features of the object clitic (*me*) match those of the governing verb (*pregunté*), and *Te pregunté por qué pasó eso* ('2SG.obj ask-PST.1SG why happen-PST.3SG that', 'I asked you why that happened'), which is excluded because the clitic (*te*) and the verb do not match in person.

By contrast, expletive uses were identified on the basis of dependency relations, following UD guidelines. More specifically, tokens annotated with expletive relations (e.g. `expl`, `expl:pv`, `expl:pass`, `expl:impers`, `expl:poss`) were classified as expletive and excluded from the set of reflexive arguments. This dependency-based filtering was applied across treebanks in order to maintain, as far as possible, a clear distinction between argumental reflexive uses and non-argumental expletive constructions.

In the French and Spanish treebanks considered here, reflexive pronouns are also annotated with the relation `obl:arg`. In these cases, this relation, together with the morphological feature `Reflex=Yes`, proved useful for retrieving argumental reflexive uses.

4. Cross-linguistic morphological properties of reflexive clitics

Reflexive constructions form a heterogeneous class in which the reflexive clitic may function as an argumental reflexive or reciprocal pronoun, a non-argumental marker of voice or impersonality, or a lexically selected element of pronominal verbs. In this section, we comparatively examine the morphological characteristics of this pronoun and their current annotation across the selected UD treebanks.

Table 2, which draws on reference works [2] for the languages under study, shows that in Romance

[2]The following reference works were consulted: Grevisse and Goosse (2011) for French, Prandi and de Santis (2019) for Italian, Godoy and Pinheiro (2023) and Bagno (2024) for Brazilian Portuguese, Raposo et al. (2013) for European Portugese, Pană Dindelegan (2013) for Romanian, and Mendikoetxea (1999) for Spanish.

No.	General patterns	French	Italian	Portuguese	Romanian	Spanish
1	A: upos=PRON deprel=expl|expl:impers|expl:pass|expl:pv|expl:poss head=B; B: upos=VERB|AUX	form=me|m'|m'|te|t'|t'|se|s'|s'| nous|vous|-nous|-vous	form=si	lemma=se	no restrictions	form=se
2	A: upos=PRON deprel=iobj|obj head=B person=B; B: upos=VERB|AUX	form=se|s'|s'	form=si	lemma=se	lemma=sine	form=se, reflex=yes
3	A: upos=PRON deprel=obj|iobj head=B person=B number=B; B: upos=VERB|AUX	form=me|m'|m'|te|t'|t'| nous|vous|-nous|-vous	form=mi|ci|ti|vi	lemma=me|nos|te|vos	lemma=eu|tu	form=me|nos|te|vos
4	A: upos=PRON form=se Reflex=Yes head=B deprel=obl:arg; B: upos=VERB|AUX	yes	no	no	no	yes

Table 1: Strategies used to query the treebanks obtained by combining a general pattern with language specific features. E.g., General pattern 1 + lemma=*se* captures all expletives in Brazilian Portuguese treebanks. Agreement between pronoun and verb is marked through a reference: e.g. person=B means that A has the same person as B.

Characteristics	FR	IT	EuPT	BrPT	RO	SP
clitic is a separate word	yes	yes	yes	yes	yes	yes
clitic can be attached to the verb	yes, in elision with vowel-initial verbs (apostrophe)	yes, with infinitive, gerunds, and affirmative imperatives (direct concatenation)	yes, in enclisis and mesoclisis (hyphen)	yes, in enclisis and mesoclisis (hyphen)	yes, in enclisis (hyphen)	yes, with infinitive, gerunds, and affirmative imperatives (direct concatenation)
proclisis	yes, finite forms; infinitive; gerund; negative imperative	yes, finite forms	yes, finite forms (triggered contexts)	yes, finite forms; infinitive; gerund	yes, finite forms; infinitive	yes, finite forms
enclisis	with imperative	with infinitive, gerund, affirmative imperative	yes	yes	with gerund, affirmative imperative	with affirmative imperatives, gerunds, infinitives
mesoclisis	no	no	future and conditional (finite)	future and conditional (finite)	no	no
distinct clitic form	yes	yes	yes	yes	yes	yes
stressed form	yes	yes	yes	yes	yes	yes
accusative forms	yes	yes	yes	yes	yes	yes
dative forms	yes	yes	yes	yes	yes	yes
acc. different from dat.	no	no	no	no	yes	no
1st pers. identical to pers. pron.	yes	yes	yes	yes	yes	yes
2nd pers. identical to pers. pron.	yes	yes	yes	yes	yes	yes
3rd pers. different from pers. pron.	yes	yes	yes	yes	yes	yes

Note: Stressed forms are subject to language-specific restrictions on co-occurrence with unstressed forms.

Table 2: Morphological characteristics of reflexive constructions in each language of this study

languages the reflexive clitic is a stand alone word[3], though it can also be phonologically dependent on its host and languages use different conventions for this: an apostrophe (as in French), a hyphen (in Romanian and Portuguese) or direct attachment (Italian and Spanish). In UD treebanks it is always tokenized separately, regardless of its written form. It is usually placed before the verbal host, with a few exceptions: affirmative imperatives, gerunds, etc., varying from one language to another. By contrast, in European Portuguese, enclisis (in which the clitic follows the verb and is attached to it by a hyphen) is generally considered the default placement in affirmative main clauses that do not contain proclisis-triggering elements.

Romanian is the only language considered here in which the reflexive pronoun has both stressed and unstressed (also called clitic) forms, with the stressed variants being rarely used.

In all the languages considered here, reflexive clitics occur in both accusative and dative functions, as indicated by their annotation with the *obj* and *iobj* relations. Romanian again stands apart from the other languages in that it has distinct forms for the accusative and the dative.

In all languages under study the reflexive clitic has specific forms only for the third person; its 1st- and 2nd-person forms are homonymous with the corresponding personal pronoun forms.

5. Treebank annotation of reflexive clitics

In most of the treebanks under study, the reflexive marker is tagged as pronoun (PRON) of the personal type (PronType=Prs). Italian uses the feature Clitic=Yes to further distinguish reflexive uses. In Romanian, reflexive clitics are tagged as personal pronouns for persons 1 and 2 (as the reflexive pronoun lacks specific forms for these persons), but they are tagged as reflexive pronouns for the third person, for which there are specific forms. The UD reflexivity feature (`Reflex=Yes`), however, is differently annotated in these languages: in some of them it is marked on all forms that carry the reflexive meaning, while in others (Romanian) it is marked only on the forms that are specialized to express reflexivity, i.e. the 3rd person.

The syntax of reflexive constructions is closely related to their semantics, and this relationship is reflected in the different syntactic relations used in

[3]We use here the notion of *word* in its UD meaning, i.e. a basic syntactic unit, usually corresponding to a whitespace-separated token in text. See https://universaldependencies.org/u/overview/tokenization.html – last accessed 20 Febr. 2026.

	fr_gsd	fr_seq	it_vit	it_isdt	br-pt_pg	br-pt_ptt	ro_rrt	sp_an
obj	y	y	y	y	n	y	y	y
iobj	y	y	y	y	n	y	y	n
expl	n	n	y	y	n	y	y	n
expl :pv	y	y	n	n	y	n	y	y
expl :poss	n	n	n	n	n	n	y	n
expl :impers	n	n	y	y	y	y	y	y
expl :pass	y	y	y	y	y	n	y	y
obl :arg	n	n	n	n	n	n	n	y

Table 3: Dependency relations available to query reflexive constructions in each treebank of this study. y = yes, n = no

their annotation. Table 3[4] presents the current annotation of the reflexive clitic in the treebanks under study[5]. In what follows, we characterize these syntactic relations and relate them to UD annotation practice, with examples from individual Romance languages provided for illustration.

- `obj` – direct object: the reflexive clitic functions as the direct object and is coreferential with the subject of its head verb. The same verb, with the same meaning, can also occur with a non-reflexive object pronoun in other contexts. Compare *Una mattina Pinocchio **si** guarda allo specchio* ('one morning Pinocchio REFL looks.at.the mirror', "One morning Pinocchio looks at himself in the mirror"; reflexive use; *UD_Italian-ISDT*) versus *Una mattina Pinocchio **lo** guarda allo specchio* ('one morning Pinocchio 3SG.M.ACC looks.at.the mirror', "One morning Pinocchio looks at him in the mirror"; non-reflexive use);

- `iobj` – indirect object: the reflexive clitic functions as the indirect object and is coreferential with the subject of its head verb. The same verb, with the same meaning, can also occur with a non-reflexive indirect object pronoun in other contexts. Compare *Isso faz com que as pessoas **se** perguntem o tempo todo o que está acontecendo* ('this make.PRS.3SG with that people REFL ask.SBJV.3PL the time all what be.PRS.3SG happen.GER', 'This makes people ask themselves all the time what is happening'; reflexive use; *UD_Portuguese-Porttinari*) versus *Isso faz com que as pessoas **me** perguntem o tempo todo o que está acontecendo* ('this make.PRS.3SG with that people 1SG.OBJ ask.SBJV.3PL the time all what be.PRS.3SG happen.GER', 'This makes people ask me all the time what is happening'; non-reflexive use);

- `expl` – the reflexive clitic refers to the subject of its head verb when doubling[6] a direct/indirect object, which is also a pronoun (reflexive or personal), but in its stressed form[7]: *se văzu pe el însuși izbind cu o secure exact în moalele capului* ('SE saw on him himself hitting with an ax exactly in soft-the head-of-the' "he saw himself hitting with an ax exactly in the soft spot") (RRT-train-sent#276);

- `expl:pv` – the reflexive clitic is part of an obligatorily pronominal verb and has no argumental value; the verb has no non-reflexive pronominal counterpart: *Pour la majeure partie du pays, il **s'agira d'une éclipse partielle.*** ('For most of the country, it SE will-be-about a partial eclipse.' "For most of the country, it will be a partial eclipse.") (SEQ);

- `expl:poss` – the reflexive clitic refers to the subject while also expressing the possessor of a direct object[8]: *părea să-și fi pierdut capacitatea de a...* ('seemes to-SE have lost capacity of...' "he seemed to have lost his capacity of...") (RRT-dev-sent#10);

- `expl:impers` – the reflexive clitic is non-referential and is not coreferential with any subject, since the verb occurs in an impersonal construction, as in *Como **se** costumava dizer, tudo muda com o tempo* ('as REFL use.to.IPFV.3SG say.INF everything change.PRS.3SG with the time', 'As people used to say, everything changes with time'; impersonal use; *UD_Portuguese-Porttinari*);

[4]The use of dependency relations cannot be assumed to be fully consistent within these treebanks.

[5]It is not relevant what frequency of occurrence each dependency relation has in each treebank; what is relevant is what types of dependency relations are used for annotating the reflexive clitics.

[6]This configuration is specific to Romanian, which morphologically distinguishes between unstressed reflexive clitics and stressed reflexive pronouns. Other Romance languages do not make this distinction in the same way, and therefore do not allow the reflexive clitic to double an independently realized stressed reflexive form.

[7]When this stressed form is not present, the clitic is annotated as `obj` or `iobj`.

[8]This configuration is specific to Romanian. Other Romance languages do not use a reflexive clitic to express possession with abstract direct objects, relying instead on possessive determiners; consequently, the label `expl:poss` is not applicable outside Romanian in our data.

- `expl:pass` – the reflexive clitic marks a passive construction, with the subject interpreted as a patient rather than an agent: *comme cela **se** faisait à l'époque partout* ('as it SE done at the time everywhere' "as it was the custom at the time everywhere") (SEQ-train-sent#frwiki_50.1000_00272);

- `obl:arg` – oblique argument: a core participant selected by the verb that is realized as an oblique rather than as a canonical direct or indirect object; in AnCora, this includes reflexive clitics that are argumental (often with dative features) but are not analyzed as objects. For instance, *Rominger **se** llevó la etapa más bonita de lo que llevamos de carrera* ('Rominger REFL take.PST.3SG the stage most beautiful of what take.PRS.1PL of race', 'Rominger claimed the most beautiful stage of the race so far'; argumental reflexive use; *UD_Spanish-AnCora*). [9]

6. Discussion of Current UD Annotation of Reflexive Constructions

In what follows, we discuss the morphological and syntactic descriptions of reflexive constructions in the treebanks under study, and their interaction with semantic interpretation.

6.1. Morphological Level with Syntactic Enhancement

The morphological feature `Reflex` is typically used for pronouns and determiners that are reflexive, that is, that refer to (or are co-referential with) the subject of the respective clause. This may give rise to misinterpretations in cases where the reflexive clitic is not semantically reflexive.

However, the UD guidelines explicitly state that "the feature `Reflex=Yes` denotes the word type, not its actual function in context (which can be distinguished by dependency relation types)"[10]. This distinction is respected in treebanks that systematically differentiate the various uses of the clitic through dedicated syntactic relations. In the absence of such syntactic differentiation, occurrences marked with `Reflex=Yes` may be incorrectly interpreted as reflexive in meaning.

The reverse case, when a pronoun is used reflexively but the morphologic feature `Reflex` is missing, is also found in the data: e.g., for Romanian, personal pronouns used with a reflexive value (as the reflexive pronoun has forms only for 3rd person, see Section 4) never have this feature: e.g. in the example *Să aluneci și să te lovești în cadă* ('SĂ[11] slip.2nd.Sg and SĂ yourself hurt.2nd.Sg in bathtub' "to slip and hurt yourself in the bathtub")(RRT-test-sent#505), the pronoun *te* lacks the feature `Reflex=Yes`.

It is only through manual investigation that the correct annotation (reflexive or non-reflexive) can be ensured.

6.2. Syntactic Level with Semantic Relevance

As shown in Section 5, the reflexive clitic's syntax can be suggestive of its semantic status as an argument or not:

- argumental structures:
  - when it is annotated as `obj` and `iobj`, the reflexive clitic fills in the syntactic position of direct or indirect object, respectively, which is co-referential with the subject of the same verb;
  - when it is annotated as `obl:arg`, the reflexive clitic also has argumental status, even though it is analyzed as an oblique argument rather than as a canonical object;
  - in reciprocal constructions, the reflexive is suggestive of the fact that each of the entities designated by the subject (in its plural form or lexicalized as an enumeration) occupies both argumental positions (Agent and Patient/Beneficiary) in relation to the verb, in turns, while the other entities occupy the other position: e.g., from the Romanian sentence *Ana și Bob SE bat.* ('Ann and Bob SE fight.' "Ann and Bob fight with each other.") one understands that Ana fights with Bob & Bob fights with Ana, thus Ana is the Agent of the fight when Bob is the Patient, and Ana is the Patient when Bob is the Agent.

    Reciprocal uses of reflexive clitics are not annotated in a distinct way in any of the treebanks considered here. However, given the semantic properties of reciprocal constructions, we suggest that their annotation would benefit from a dedicated syntactic relation that is informative for downstream semantic representation.

[9] Treebanks differ in how they encode non-canonical arguments: UD_Spanish-AnCora may analyze certain clitic arguments (including reflexive/dative *se*) as `obl:arg`, whereas UD_French-Sequoia uses `obl:arg` primarily for prepositional or nominal oblique complements (e.g., selected locatives), treating reflexive clitics separately (e.g., `expl:pv`, `obj`, `iobj`).

[10] https://universaldependencies.org/u/feat/Reflex.html , last accessed 25 January 2026

[11] SĂ is the marker of subjunctive mood in Romanian.

- non-argumental structures:
  - when being an expletive, it is not a semantic argument of the verb. This relation should be used only for those cases when the clitic doubles one argument, i.e. the direct or indirect object; all three words (the subject, the direct/indirect object and the refexive clitic) are coreferential;
  - in inherently reflexive constructions (annotated as `expl:pv` in UD treebanks) doubling by a stressed form is ungrammatical; in such cases we can talk of non-compositionality of the reflexive construction[12];
  - in possessive constructions, it contributes semantic information of the type possessor with respect to a noun which is the direct object of the respective verb; the clitic is coreferential with the subject;
  - in impersonal constructions, it actually blocks the external argument of the verb, i.e. the subject;
  - when having a passive meaning, it signals that the subject is not the Agent of the verb, but rather its Patient.

All these cases show that the appropriate syntactic analysis of the reflexive clitic aligns with important semantic differences across constructions. These distinctions must therefore be adequately encoded in annotation schemes whenever automatic conversion to semantic representation is intended, so as to ensure accurate mapping.

When having a closer look at Table 3, we notice the following:

- Br-Pt-Petrogold does not annotate reflexive constructions with `obj` and `iobj` relations (this is an inconsistency in different treebanks for the same language).
- For French, the impersonal value of the reflexive pronoun has not been annotated in the two treeanks studied, GSD and Sequoia.
- Br-Pt-Porttinari does not use `expl:pass` relation and uses *expl:impers* for all such cases (Duran, 2022) (another inconsistency with respect to the other treebank of Brazilian Portuguese).
- We notice that if `expl` is used, then `exp:pv` is not used and vice versa. Romanian is an exception, as it uses both dependency relations in the syntactic annotation.

7. Towards a Uniform Analysis of Reflexive Constructions

A more uniform analysis of reflexive constructions in UD treebanks is needed both to facilitate cross-linguistic comparison and to support adequate conversion to semantic representation. This requires the following:

- strictly observe the UD guidelines: this means:
  - make adequate use of the morphological feature `Reflex=Yes`,
  - clearly distinguish between argumental and non-argumental positions,
  - go beyond the usage of the universal dependency relation `expl` and use its subtypes: `expl:pv`, `expl:poss`, `expl:impers`, `expl:pass` if the language under description displays these values;
- amend the UD guidelines: in line with what we have presented above, the necessary amendments of the UD guidelines are as follows:
  - recognize the status of all subtypes of the `expl` dependency relation as valid ones, i.e., enumerate them in the official list of dependency relations[13];
  - describe each such subtype of the `expl` relation, by providing documentation for them, that is including their characteristics, as well as examples in several languages displaying them;
  - recognize `expl:rcp` as a new subtype of the `expl` relation, as it is useful in keeping reciprocal values of reflexives apart from the others.

The proposals put forward here are intended to provide guidance for future treebank development and may also be adopted, where appropriate, in the revision of existing treebanks. Some of them have already been implemented in the Romanian treebank. These include:

- Ensuring consistency between the morphologic and syntactic annotation: inconsistencies were automatically identified, such as 1st and 2nd person reflexives that lack the `Reflex=Yes` value. They are personal pronouns

[12] See the PARSEME guidelines in which such cases are annotated as multiword expressions: https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=050_Tests_for_VERBAL_MWEs/040_Inherently_reflexive_verbs__LB_IRV_RB_, last accessed 25 January 2026.

[13] https://universaldependencies.org/u/dep/index.html

that are used as reflexives (see Section 4). Thus, we added this feature automatically to the pronoun annotation in such cases;

- Ensuring the correct semantic interpretation of reflexives by adding one more syntactic relation, which, according to UD principles, is a sub-relation of a universal one: we propose `expl:rcp` as a new relation that can identify the reciprocal meaning of reflexives. The occurrences are relatively easy to spot in the treebanks, when searching for examples with a plural or coordinated subject of the respective verb. Nevertheless, a human-in-the loop approach is mandatory as the same morphological features are not exceptional with pure reflexive values as well; the reciprocity is actually in the semantics of the verb rather than in the morpho-syntactic characteristics of the construction. For Romanian, we have applied a heuristic which builds on the semantic information (in the form of definitions) extracted from an explanatory dictionary[14] to have a list of verbs with reciprocal meanings for which we can rather safely assign the relation `expl:rcp` to the clitic. The same strategy can be applied to the other languages or translation equivalents can be used to find similar examples in other languages, though such a strategy can leave out a lot of examples, as it is highly unlikely that the same reciprocal verbs occur in all or in most treebanks, given that they are not parallel corpora.

Duran et al. (2025) proposed unifying the annotation of synthetic passives (i.e. passives expressed by means of reflexive constructions) and impersonal expletives, on the grounds that the formal indicators traditionally used to distinguish the two – such as number agreement – are becoming increasingly unreliable. This proposal, however, needs to be examined from a cross-linguistic perspective. For instance, the Portuguese impersonal construction *Vende-se casas na ilha* ('Sells-SE houses on island.' "They sell houses on the island.") corresponds in Romanian to a clearly passive sentence: *Se vând case pe această insulă* ('SE sell houses on this island.' "Houses are sold on this island."). In the Romanian example, the passive interpretation is supported by number agreement between the noun *case* and the verb.

Taken together, these observations indicate that the proposal put forward in the Porttinari guidelines holds for Brazilian Portuguese, where the relevant constructions are systematically ambiguous, while its extension to other Romance languages requires careful cross-linguistic assessment. In cases where no syntactic or morphological cues clearly favor one interpretation over the other, adopting a unified analysis – such as `expl:impers` – may nonetheless constitute a practical and consistent solution.

[14] https://dexonline.ro/

8. Conclusions

In this paper we focused on the treatment of reflexive constructions in some UD Romance treebanks, chosen mainly for their representation of contemporary written language. Our observations do not differ from those in Marković and Zeman (2018), which is indicative of the fact that, during the last 8 years, this construction has attracted little interest from the developers and maintainers (at least) of the treebanks under consideration here, with the exception of Duran et al. (2025), as far as we are aware.

As shown above (Section 7), we have taken steps towards ensuring annotation consistency of such constructions. Some of them are easy to implement (i.e., they can be done automatically), others are more difficult (i.e., human supervision and interventions are mandatory), while those concerning treebanks not maintained by us require at least the approval of the actual maintainers, if not also the approval of UD representatives.

9. Limitations

For a generally applicable decision on the annotation of reflexive constructions in UD, a wider perspective is necessary, both in terms of theoretical aspects for all languages (typological studies), and of the practices in more (all) treebanks, not only those of 5 Romance languages.

10. Acknowledgment

This work was supported by a grant of the Ministry of Education and Research, CCCDI – UEFISCDI, project number PN-IV-PCB-RO-MD-2024-0142, within PNCDI IV; and grants of the Brazilian National Council for Scientific and Technological Development (CNPq 406926/2025-5, 404722/2024-5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG), Brazil.

11. Bibliographical References

Linda Alfieri and Fabio Tamburini. 2016. (Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format. In *CEUR Workshop Proceedings*, volume 1749, pages 19–23. Accademia University Press.

Marcos Bagno. 2024. *Gramática pedagógica do português brasileiro*. Parábola.

Verginica Barbu Mititelu. 2018. Modern syntactic analysis of Romanian. In *Clasic şi modern în cercetarea filologică românească actuală*, pages 67–78, Iași, Romania.

Julia Bonn, Andrew Cowell, Jan Hajič, Alexis Palmer, Martha Palmer, James Pustejovsky, Haibo Sun, Zdenka Uresova, Shira Wein, Nianwen Xue, and Jin Zhao. 2023. UMR annotation of multiword expressions. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 99–109, Nancy, France. Association for Computational Linguistics.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2012. Harmonization and merging of two italian dependency treebanks. In *Proceedings of the LREC 2012 Workshop on Language Resource Merging*, pages 23–30. ELRA.

António Branco, João Silva, Francisco Costa, and Sérgio Castro. 2011. Cintil treebank handbook: Design options for the representation of syntactic constituency. Technical Report DI–FCUL–TR–2011–02, Department of Informatics, University of Lisbon, Faculty of Sciences.

Marie Candito and Djamé Seddah. 2012. Le corpus sequoia: annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles*.

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

Elvis de Souza and Cláudia Freitas. 2022. Polishing the gold–how much revision do we need in treebanks? In *Procedings of the Universal Dependencies Brazilian Festival*, pages 1–11.

Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. Vit-venice italian treebank: Syntactic and quantitative features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54. Northern European Association for Language Technol.

Magali Sanches Duran. 2022. Manual de anotação de relações de dependência: versão revisada e estendida.

Magali Sanches Duran, Lucelene Lopes, Maria das Graas Nunes, and Thiago Pardo. 2023. The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, pages 124–133.

Magali Sanches Duran, Adriana Silvina Pagano, and Thiago Alexandre Salgueiro Pardo. 2025. Diving deeper into the waters of "se" as a clitic in brazilian portuguese. *Corpus Linguistics: Studies and Applications*.

Federica Gamba, Alexis Palmer, and Daniel Zeman. 2025. Bootstrapping UMRs from universal dependencies for scalable multilingual annotation. In *Proceedings of the 19th Linguistic Annotation Workshop*, pages 126–136, Online. Association for Computational Linguistics.

Luisa Godoy and Diogo Pinheiro. 2023. A rede gramatical das construções com se no português brasileiro. *Soletras Revista*.

Maurice Grevisse and André Goosse. 2011. *Le bon usage: grammaire française; Grevisse-grammaire langue française*. De Boeck, Duculot.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en universal dependencies [conversion and improvement of universal dependencies french corpora]. *Traitement automatique des langues*, 60(2):71–95.

Sonja Marković and Daniel Zeman. 2018. Reflexives in universal dependencies. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 131–146, Sweden. Linköping University Electronic Press.

Amaya Mendikoetxea. 1999. Construcciones con" se": medias, pasivas e impersonales. In *Gramática descriptiva de la lengua española*, pages 1631–1722. Espasa Calpe España.

Gabriela Pană Dindelegan, editor. 2013. *The Grammar of Romanian*. Oxford University Press.

Michele Prandi and Cristiana de Santis. 2019. *Manuale di linguistica e di grammatica italiana*. UTET.

Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antonia Coelho da Mota, Luisa Segura, and Amalia Mendes. 2013. *Gramática do Português*. Fundação Calouste Gulbenkian.

Manuela Sanguinetti and Cristina Bosco. 2014. Converting the parallel treebank partut in universal stanford dependencies. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa*, pages 316–321. Pisa University Press.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

12. Language Resource References

Bosco, Cristina and others. 2015. *ISDT (Italian Stanford Dependency Treebank)*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-1464.

de Souza, Elvis and Freitas, Cláudia. 2022. *PetroGold*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-4923.

Héctor Martínez Alonso and Daniel Zeman. 2016. *AnCora*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-1699.

Magali Sanches Duran and others. 2017. *Porttinari (PORTuguese Treebank)*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-2184.

Mariana Avelãs and others. 2022. *CINTIL-UDep*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-4923.

Marie Candito and others. 2017. *Sequoia*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-2184.

Marie-Catherine de Marneffe and others. 2015. *UD French GSD*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-1464.

Rodolfo Delmonte, Antonella Bristot, Sara Tonelli. 2007. *VIT (Venice Italian Treebank)*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-2988.

Verginica Barbu Mititelu and others. 2015. *The Romanian Reference Treebank (RoRefTrees or RRT)*. Distributed via Universal Dependencies. PID http://hdl.handle.net/11234/1-1548.