Leveraging UD data for Downstream Semantic Tasks: Reflexive Clitics Across Languages

Adriana Pagano[1], **Verginica Barbu Mititelu**[2], **Daniel Zeman**[3], **Federica Gamba**[3], **Patricia Chiril**[4], **Diego Alves**[5], **Cristina Bosco**[6], **Irina Lobzhanidze**[7], **Petya Osenova**[8], **Kaja Dobrovoljc**[9], **Daša Farkaš**[10], **Elena Irimia**[2], **Ioana-Madalina Silai**[12]
[1]Federal University of Minas Gerais, Brazil; [2]RACAI, Bucharest, Romania;
[3]Charles University, Czech Republic; [4]Télécom Paris, Institut Polytechnique de Paris, France;
[5]Saarland University, Germany; [6]University of Turin, Italy; [7]Ilia State University, Georgia;
[8]Sofia University, IICT-BAS, Bulgaria; [9]University of Ljubljana, Slovenia; [10]University of Zagreb, Croatia;
[11]Institute of Mathematics and Computer Science, University of Latvia; [12]Université Paris Nanterre, France

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

We focus on reflexive constructions, i.e verbal constructions marked by a reflexive clitic, with meaning encompassing reflexive, passive, impersonal, middle, and lexically specified uses. The corpora we work with for their analysis is represented by several Universal Dependencies (UD) (de Marneffe et al., 2021) treebanks for several languages: Belarusian, Bulgarian, Croatian, Czech, French, Georgian, Italian, Polish, Brazilian Portuguese, Romanian, Russian, Old Church Slavonic, Slovak, Slovene, Upper Sorbian, Spanish, Ukrainian. Here are two examples of such constructions in Spanish:

(1) *La paciente se quejaba de todo.*

la	paciente	se	quejaba	de todo
DEF.F.SG	patient	REFL	complain.PST.IPFV.3SG	of all

'The patient complained about everything.'

(2) *La paciente se culpaba por todo.*

la	paciente	se	culpaba	por todo
DEF.F.SG	patient	REFL	blame.PST.IPFV.3SG	for all

'The patient blamed herself for everything.'

The interest in these constructions is justified, on the one hand, by the desirability to ensure consistency within and across UD treebanks, and on the other, by the interest in automatic conversion of UD treebanks into their Uniform Meaning Representation (UMR) (Bonn et al., 2023), with benefits for comparative linguistics and the PARSEME guidelines[1] as well, as a particular reflexive clitic construction, namely Inherently Reflexive Verbs, is relevant to MWE annotation. In this context, recent work on a method for bootstrapping (partial) UMRs from UD trees (Gamba et al., 2025) provides a practical framework for such conversion. The approach involves iterating over all nodes in each UD tree and processing them sequentially, allowing for a systematic mapping from syntactic structures to their corresponding semantic representations.

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 are annotated either as core arguments or as non-argumental markers (see Section 3.2). This distinction directly affects how a reflexive construction is interpreted in UMR: as involving coreference between semantic roles, with the same participant filling more than one role; as part of the lexical meaning of a pronominal verb; or as a valency-changing or voice-related operation, such as passive, impersonal, or anticausative marking. The contrast between the first two uses is illustrated by the UMR graphs for examples (1) and (2) below.

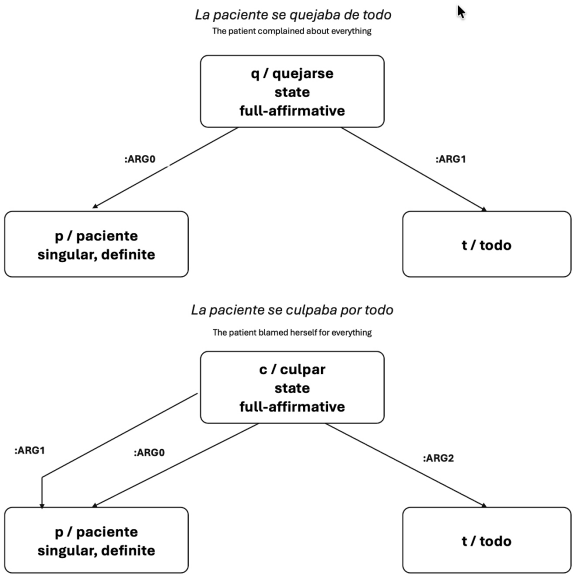

Figure 1: Two different uses of Spanish reflexive se in UMR

[1]https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

In (1) the clitic is part of the lexical predicate "quejarse" (to complain), not an independent argument, whereas in (2) the same participant ("paciente") fills both ARG0 and ARG1, the blamer and blamed roles, in the predicate "culpar" (to blame).

If UD annotation conflates these uses or applies them inconsistently across treebanks, automatic UMR conversion may incorrectly introduce or omit semantic participants. Therefore, consistent (Marković and Zeman, 2018) and linguistically grounded (Duran et al., 2025) UD treatment of reflexive clitics is crucial for ensuring accurate semantic representation, especially in multilingual UMR pipelines .

Our aim is to determine how syntactic annotation can support the identification of semantically relevant contrasts, especially the distinction between argumental and non-argumental reflexive uses, and how they may be leveraged in converting syntactic annotation into semantic representation in UMR. Thus, a broader goal is identifying the kinds of information in treebanks that are most useful for supporting semantically informed annotation and for formulating recommendations for prospective treebank annotators and researchers interested in syntax-semantics interfaces.

2 Methodology

We have taken the following steps so far: (i) offer a uniform description of the reflexive system across the languages considered on two linguistic levels: morphologic and syntactic; (ii) document UD treebank practices for annotation of reflexive constructions. We present some insights from this below.

2.1 Morphology

The reflexive clitic is a stand alone word, though it can also be phonologically dependent on its host. In UD treebanks it is tokenized separately, regardless of its written form. Clitic placement varies across languages and follows language-specific patterns. 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, for emphasis. In most of the languages considered here, reflexive clitics occur mostly in accusative and dative cases (though in some languages other cases are also possible). The reflexive clitic has specific forms only for the third person; its 1st- and 2nd-person forms, when existing, are homonymous with the corresponding personal pronoun forms.

2.2 Syntax

The clitic may not have a fixed position with respect to the verb, though in some languages it may be fixed to second (also called Wackernagel) position. It sometimes climbs over a modal(-like) verb, it may co-occur with some non-finite forms, but not with others. The semantic value of the clitic is sometimes distinguished only by means of the presence or absence of the person and number agreement between the subject and the verb (e.g., for impersonal or passive readings).

3 UD annotation

3.1 Morphological level

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. However, this is not used systematically, especially for forms that are not originally reflexive but acquire such value contextually.

3.2 Syntactic level

The dependency relations assigned to reflexive pronouns are argumental (`obj`, `iobj`, `obl:arg`) or non-argumental (mostly `expl`, but subtypes of this relation are also used: `expl:pv` for inherently reflexives, `expl:pass` for passive values of the clitic, `expl:impers` for impersonal meanings of the clitic, `expl:poss` for a possesive value). Following UD guidelines, the reciprocal value is not annotated by a distinct relation in any treebank.

4 Discussion

Treebank inspection shows that current UD annotation captures some semantically relevant distinctions more clearly than others. Canonical argumental reflexives are often retrievable when reflexive markers are morphologically identifiable and consistently annotated as `obj` or `iobj`. In such cases, the syntactic annotation provides useful evidence for semantic interpretation, including for UMR, since the reflexive marker may correspond to a participant role co-referential with the subject. However, non-argumental reflexives display greater variation in annotation and are therefore

more difficult to retrieve systematically and map consistently onto semantic representations.

A clear distinction of reflexive clitic annotation for argumental and non-argumental roles is therefore desirable. The above Spanish examples in Spanish illustrate this.

To distinguish these two uses of the reflexive clitic *se*, it is desirable for the clitic to be annotated as `expl:pv` in (1) and as `obj` in (2) above. This contrast captures the difference between a non-argumental reflexive marker associated with a pronominal verb and an argumental reflexive marker functioning as an object co-referential with the subject. The distinction is directly relevant for UMR: in the latter case, the reflexive marker corresponds to a semantic participant within the event structure, whereas in the former it is better treated as part of the lexical or constructional profile of the predicate, without necessarily introducing a separate semantic argument.

At a broader level, the cross-linguistic comparison shows variation between object-like and expletive-like annotations. Similar surface patterns are currently annotated as `obj` in some contexts and as `expl` or `expl:pv` in others, either across treebanks or within the same treebank. This affects not only retrieval, but also comparability across languages and the interpretation of argument structure. From a UMR perspective, this variation has direct implications for automatic conversion from CoNLL-U representations, This distinction matters because annotation decisions in treebanks may determine whether a reflexive construction is interpreted as part of the lexical meaning of the verb, as in quejarse 'to complain', or as involving coreference between semantic roles, as in culparse 'to blame oneself'. This is illustrated by the UMR graphs of examples (1) and (2) below.

5 Conclusions

Reflexive clitics provide a particularly revealing test case for the relation between syntactic annotation and semantic analysis, especially in view of semantic frameworks such as UMR. Current treebanks already encode some distinctions that are highly relevant for semantic interpretation, especially in the domain of argumental reflexivity, but they do so unevenly. As a result, treebank annotation can support a more semantically oriented analysis and contribute to UMR-oriented interpretation only when the relevant distinctions are explicitly and consistently represented.

Acknowledgements

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

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.

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

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 (LAW-XIX-2025)*, pages 126–136, Vienna, Austria. Association for Computational Linguistics.

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.