Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 77-84 recibido 01-12-2019 revisado 17-01-2020 aceptado 19-01-2020

Reflexive pronouns in Spanish Universal Dependencies

Los pronombres reflexivos en las Dependencias Universales en español

Jasper Degraeuwe, Patrick Goethals
Ghent University (Belgium)

Jasper.Degraeuwe@UGent.be
Patrick.Goethals@UGent.be

Abstract: In this paper, we argue that in current Universal Dependencies treebanks, the annotation of Spanish reflexives is an unsolved problem, which clearly affects the accuracy and consistency of current parsers. We evaluate different proposals for fine-tuning the various categories, and discuss remaining open issues. We believe that the solution for these issues could lie in a multi-layered way of annotating the characteristics, combining annotation of the dependency relation and of the so-called token features, rather than in expanding the number of categories on one layer. We apply this proposal to the v2.5 Spanish UD AnCora treebank and provide a categorized conversion table that can be run with a Python script.
Keywords: reflexive pronouns, *se*, Universal Dependencies, AnCora, Spanish

Resumen: En este trabajo, argumentamos que en los actuales treebanks que aplican el formalismo de las Dependencias Universales, la anotación de los reflexivos españoles es un problema sin resolver, que afecta claramente a la precisión y consistencia de los parsers actuales. Evaluamos diferentes propuestas para afinar las diferentes categorías y discutimos los problemas pendientes. Creemos que la solución para estos problemas se puede encontrar en una anotación en múltiples niveles, combinando la anotación de la relación de dependencia y de las características (*features*) de los tokens, en lugar de ampliar el número de categorías en un solo nivel de anotación. Aplicamos la propuesta a la versión española del treebank UD AnCora (v2.5) y proporcionamos una tabla de conversión categorizada que se puede ejecutar mediante un script Python.
Palabras clave: pronombres reflexivos, *se*, Dependencias Universales, AnCora, español

1 Introduction

In recent years, syntactic parsing, the Natural Language Processing (NLP) technique which assigns a syntactic label to words in a sentence, has been integrated in a wide range of NLP applications. Since these applications do no longer require their users to have an extensive technological expertise, the technique has also become widely accessible to language professionals. In fact, parsers such as spaCy or StandfordNLP can be invoked from simple Python scripts, and generate enriched input for developing intelligent text-based applications. Existing NLP tools are usually trained on reference data (treebanks), which are not only growing in number, but also becoming more and more standardized and comparable within and across languages. The Universal Dependencies (UD) project, launched in 2014, plays a crucial role in this context, as it seeks to develop "cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective" (https://universaldependencies.org/introduction, retrieved 24 January 2020; see also Nivre et al. (2016)). UDv2.5 contains 157 treebanks in 90 languages, including previously built treebanks converted into the UD formalism (e.g. the Spanish AnCora treebank, see Taulé, Martí, and

ISSN 1135-5948. DOI 10.26342/2020-64-9

Recasens (2008) and Martínez Alonso and Zeman (2016)).

However, the UD initiative is also a "constantly improving effort" (Martínez Alonso and Zeman, 2016), meaning that annotation guidelines are constantly being fine-tuned over the successive releases of the treebanks (we will work with the latest 2.5 version). Moreover, several annotation issues, which may be problematic from both a cross-linguistic and an intra-linguistic perspective, remain unsolved. For Spanish, one of these issues is the exact annotation of personal pronouns, more in particular of the potentially reflexive pronouns *me*, *te*, *nos*, *os* and *se* (Marković and Zeman, 2018; Silveira, 2016: 115-144). These are very frequent items in Spanish, with *se* occurring, for example, in more than 20% of the sentences in written genres (percentage obtained from corpus research within the Spanish Corpus Annotation Project (Goethals, 2018)), and in almost 30% of the sentences in the test and training sets of Spanish UD AnCora.

The complexity of this issue is also reflected in the output produced by (the current versions of) publicly available parsers, which is often unreliable, inconsistent and/or very coarse-grained. For example, the indirect object in *se lo dije* ('I said this to him/her') is labeled as a passive marker by StanfordNLP; the inherently reflexive *se acuerdan de ti* ('they remember you') is direct object in spaCy and again a passive marker in StandfordNLP; or, finally, in the reflexive passive *se celebran los cien años del club* ('the 100th anniversary of the club is celebrated') *se* is direct object in both parsers. Incorrect labeling of this kind happens consistently over a wide range of similar constructions with a potentially reflexive pronoun, which implies that it cannot be due to the inherent error rate of the parsers' machine learning algorithm. Rather, faulty annotations in the underlying treebanks are more likely to be at the root of the problem.

Importantly, when trying to solve parser problems, we should not only try to improve annotation consistency and parser accuracy, but also take into account the (cross-)linguistic analyses made in non-computational linguistics (Croft et al., 2017): as parsers become more accessible, more theoretical linguists will use them and evaluate their (linguistic) accuracy and granularity. In this regard, Spanish *se* is a heavily debated subject, with many studies focusing on both the syntactic and the semantic characteristics of the construction (e.g. Mendikoetxea, 1999; Peregrín Otero, 1999; Maldonado, 2008). Although it goes beyond the scope of this paper to discuss all aspects of these analyses, we will briefly come back to this matter in Section 2.3.4.

In the light of the context we have just outlined, with this paper we wish to contribute to a solution for Spanish reflexives by developing an annotation proposal that adheres to the conceptual UD principles (a.o. allowing a satisfactory linguistic analysis, and rapid and consistent human annotation). We will also propose a concrete reannotation of the Spanish UD AnCora treebank, providing exhaustive and categorized conversion tables, and a corresponding Python script to apply these changes to the original treebank files in CoNLL-U format.

Before proceeding to the discussion, it is worth mentioning that we specifically focus on working with UD-based preprocessed corpora in the field of ICALL (Intelligent Computer-Assisted Language Learning), applied to vocabulary learning. Concretely, our purpose is to develop NLP-based corpus query tools that automatically extract authentic usage examples of verbs or nouns, in order to exemplify the constructions in which they are used, and to generate well-targeted vocabulary learning materials.

2 *Reflexives in Spanish Universal Dependencies*

The UD framework provides three key annotation layers by which linguistic constructions can be progressively defined and differentiated: a morphosyntactic Part-of-Speech (POS) tag (limited to a universal set of seventeen tags); a syntactic dependency relation (e.g. subject, direct object, indirect object,…); and a feature set containing additional lexical and grammatical properties (e.g. number or person in the case of pronouns, or tense in the case of verbs).

2.1 Reflexives in current Spanish UD treebanks

The current annotation of reflexives in UD AnCora is as follows:

1. The POS tag is always PRON (which indeed seems the only possible universal tag,

and which we will leave out of the discussion in the remainder of this paper).

2. The feature set includes properties such as “Case” (Acc, Dat), “Person” and “Reflex”, but does not disambiguate “Case”, and only disambiguates “Reflex” in the case of *me, te, nos* and *os*, but not with *se*. As a result, non-coreferential indirect objects such as *el PP no se lo perdona* (‘the PP does not forgive him this’) are still annotated as “Reflex=Yes”. Furthermore, AnCora does not adjust the feature set of the verbal head according to the function of *se* (e.g. by adding the property “Voice=Pass”, see below).

3. Finally, the dependency label is the layer used to actually differentiate between the different uses of reflexive pronouns. Concretely, reflexives can have three values:

- “expl:pass”, used for impersonal constructions such as *en Europa se trabaja mucho* (‘in Europe, people work a lot’) or impersonal passives where there is no subject concordance between the verb and the argument, e.g. *se condena a los culpables* (‘the culprits are convicted’), but not for regular reflexive passives with subject concordance such as *se ve el efecto* (‘the effect is seen’).
- “iobj” (indirect object), used for prototypical coreferential (*Pedro se quita la chaqueta*, ‘Pedro takes his jacket off’) and non-coreferential indirect objects (*no se lo perdono*, ‘I do not forgive him/her this’), but also for some (semi-)lexicalized indirect objects such as *preguntarse si* (‘to ask yourself if’) or *proponerse hacer algo* (‘to intend to do something’).
- “obj” (direct object), used for all cases that are not “expl:pass” or “iobj”, namely regular reflexive passives (see above), prototypical reflexives (*verse a sí mismo*, ‘to see yourself’), and all other (semi-)lexicalized *se* constructions (*materializarse* ‘to become reality’, *morirse* ‘to die’, *moverse* ‘to move yourself’, *...).*

Apart from actual annotation inconsistencies (which are relatively frequent, e.g. ascending up to 30% and 60% of false positives of “expl:pass” and “iobj”, respectively), the main problem with this annotation scheme is its coarse-grained nature. The taxonomy does not allow, for example, distinguishing between passive (*en este volumen se ofrecen textos sobre*, ‘in this volume texts are provided about’) and reflexive uses (*María se ofrece para hacerse cargo del bebé*, ‘María offers herself to take care of the baby’) of the same verb, or between passive (*se incautaron las armas*, ‘the guns were seized’) and inherently reflexive constructions (*la policía se incauta de la armas*, ‘the police seized the guns’). In all these cases, *se* is labeled as “obj”, and the feature sets (both of *se* and of the verbal head) are also equal. For ICALL purposes, this means that the current labels do not enable retrieving targeted examples to illustrate these construction alternations, although they are highly relevant for L2 learners of Spanish.

Interestingly, the multilingual Parallel Universal Dependencies (PUD) treebank for Spanish (created for the 2017 CoNLL shared task and much smaller than AnCora) follows a different strategy: on the one hand, it assigns the same dependency label “compound:prt” to all cases of *se* (which means that all constructions with *se* are conceptualized as a type of multiword expression), but on the other hand, it does introduce a “Voice” feature in the description of the verb and thus manages to distinguish between passives (“Voice=Pass”) and (inherently) reflexives (“Voice=Act”). This solution, however, contrasts with the current UD guidelines for Spanish, which state that “the Voice feature is not used in Spanish because the passive voice is expressed periphrastically” (https://universaldependencies.org/es/index, retrieved 24 January 2020).

2.2 Towards a new annotation of the dependency relations

Recently, Silveira (2016) and Marković and Zeman (2018) formulated several proposals for improving and refining the annotation of reflexives. There seems to be an agreement about the fact that at least the following uses can and should be distinguished:

1. True reflexives, which can be expanded by a focal reflexive *(a/para) sí mismo/a(s)*, or could take other non-coreferential objects (e.g. *le*). In these cases, *se* is assigned the dependency label “obj” (*los participantes tienen que inscribirse*, ‘participants have to

register themselves') or "iobj" (*se reservan el derecho a*, 'they reserve for themselves the right to'), depending on the syntactic function.

2. Passive constructions (e.g. *la noticia no se publicó por razones de seguridad* 'the news was not published for safety reasons' or *se recaudan los ingresos fiscales* 'tax revenues are collected'), where there is verbal concordance with the original object of the corresponding non-reflexive transitive verb, and where a transitive process is evoked in which an (unexpressed and perhaps generic) agent acts upon the object. Here, *se* would be annotated as "expl:pass" (note that this does not cover the same constructions as the current annotation).

3. Impersonal constructions, where *se* is combined with an intransitive verb (*en Europa se trabaja mucho*), or with a transitive verb and a nominal that is explicitly marked as accusative (*se condena a los culpables*). Here, *se* would receive the label "expl:impers" (see also Bouma et al. (2018)).

4. Non-coreferential indirect objects where *se* substitutes *le* when it is combined with accusative *lo/a(s)*, as in *se lo pago* ('I pay it to him/her').

In cases 1, 2 and 3, the reflexive use of the construction activates the same event conceptualization as the non-reflexive counterpart, with *se* occupying one of the "obj" roles, and/or blurring the subject role (in the case of passive and impersonal constructions). However, it is obvious that not all reflexives can be classified into one of these categories (in corpus studies the uncontroversial examples would barely account for 50% of the examples).

Therefore, all proposals also include at least a fifth category, namely inherently reflexive verbs, such as *desmayarse* 'to faint', *parecerse a* 'to resemble' or *negarse a* 'to refuse', which are constructions without a clear transitive counterpart. As stated in the UD guidelines, inherently reflexive verbs "cannot exist without the reflexive clitic, and the clitic cannot be substituted by an irreflexive pronoun or a noun phrase. In many cases, an irreflexive counterpart of the verb actually exists but its meaning is different because it denotes a different action performed by the agent". In these cases, *se* receives the label "expl:pv", meaning that *se* is conceptualized as a lexical morpheme (see also the "compound:prt" label in the Spanish PUD treebank).

Clearly, this set of dependency relations offers a far more subtle way of annotating the reflexive forms. However, there remain several issues, which we will discuss in what follows, and which are mainly related to the annotation of the token features of both the reflexive pronoun and the verbal head.

2.3 What about features?

2.3.1 Voice

First, although the current UD guidelines provide that "Voice=Pass" should not be used for Spanish, we are inclined to follow the PUD practice of adding this property to the feature set of the verbal head in the case of the reflexive passive constructions. It seems counterintuitive to mark the reflexive as "expl:pass", without extending this verbal feature to the head of the reflexive. Moreover, as will become clear from the discussion below, the "Voice=Pass" property also enables us to analyze the "umbrella category" of "expl:pv" in greater detail.

2.3.2 Reflexive / reciprocal

Secondly, the UD guidelines do not make a distinction between reflexive and reciprocal readings. The property "PronType=Rcp" does exist, but it is only applied to cases such as German *einander* 'each other', and as a distinctive feature that contrasts with the broad category of personal pronouns ("PronType=Prs"), to which all reflexives belong by definition. Since the reciprocal use of *se* is only one of its many uses, using "PronType=Rcp" would not be an adequate solution for marking this particular use. However, UD does allow personal pronouns to receive an extra feature called "Reflex", but this takes only one possible value, namely "Yes". We would like to propose that, similarly to the annotation of other features such as "Case", "Reflex" accept two possible values, namely "Reflex" and "Rcp". As a result, it would be possible to distinguish between *es importante quererse (a sí mismo)* 'it is important to love yourself' and *es importante quererse (el uno al otro)* 'it is important to love each other', without jeopardizing the unity of the personal pronoun category. As was the case for "Voice=Pass", the "Reflex=Rcp" property will also prove to be useful for analyzing the "expl:pv" cases.

2.3.3 Comitative case

Thirdly, the feature "Case" for reflexive items could be expanded with "Com" (comitative), which is now exclusively used for describing the pronouns *conmigo/contigo/consigo* ('with me, you, him/herself'). Particularly in the case of the verb *llevar* (*llevarse algo [consigo]*, 'to take something with you'), this seems semantically more appropriate than the "Dat" (dative) value, and it can avoid having to identify two "Dat" arguments in examples such as *el Boca se le llevó un punto al Deportivo* 'Boca took a point from Deportivo with them'.

2.3.4 Features and "expl:pv" constructions

Although the dependency and feature set modifications of sections 2.2-2.3.3 provide a suitable annotation solution for a considerable number of problematic cases, they do not address the annotation of the "expl:pv" category. Clearly, this category covers a wide range of constructions, which, though having a characteristic in common (i.e. that *se* modifies the verbal event structure rather than referring to one of its participants), seem to differ considerably from each other, as is illustrated by the following list:

- *morirse* (adding the nuance of unexpectedness to *morir* 'to die')
- *la gente se manifiesta* ('people are demonstrating')
- *el fenómeno se manifiesta* ('the phenomenon becomes clear')
- *acordarse de algo* ('to remember something')
- *negarse a algo* ('to refuse to do something')
- *se me ocurre que* ('it occurs to me that')
- *ponerse de acuerdo* ('to agree on something')
- *llevarse bien con alguien* ('to get along with someone')

In this regard, it is important to consider a commonly held point of view in Spanish linguistic tradition, namely that reflexives in Spanish activate a so-called "middle voice", in between active and passive voice. One of the most prototypical middle voice contexts are spontaneous processes such as *el problema se manifiesta cada vez más claramente*, which do not carry a truly reflexive (active) meaning, and which exhibit a clear difference with passive constructions, since the agent role has not "faded away" from the profiled event, but is really absent from it. In fact, the middle voice is even considered as the core value of *se*, or, as Maldonado (2008: 155) puts it, "the analysis of the clitic *se* as a reflexive pronoun misrepresents the overall functions that the clitic displays. Instead it is proposed that while having a reduced number of reflexive uses the clitic *se* is a middle voice marker".

One possible solution to capture this middle voice in annotation (a topic which has been left unaddressed in UD guidelines for Spanish) would be to introduce a new dependency relation (e.g. "expl:middle"). However, both Marković and Zeman (2018) and Silveira (2016) take an explicit stance in this matter, pointing out that the distinction between reflexive and passive, on the one hand, and middle voice, on the other, is too subtle and too hard to discern to create a separate "expl:middle" category. Although this may seem a pragmatic rather than a conceptual decision, it should be highlighted that, while in descriptive and theoretical linguistics syntactic categories are often conceptualized as gradual and partially overlapping categories, in the field of NLP tagging and parsing categories are usually of a discrete nature. Therefore, we want to propose an alternative way to handle the diversity of potentially reflexive pronouns (especially of *se*) in this type of construction.

As was already mentioned, the common characteristic of "expl:pv" cases of *se* is that they modify the verbal event rather than referring to one of its participants (or fading away from it, as is the case in "expl:pass"). Starting from the idea that an "expl:pv" modifies an underlying event frame, we believe that an appropriate answer to the problem of accounting for the diversity of the "expl:pv" cases can come from the definition of the features ("Case" and "Reflex" for the reflexive item, and "Voice" for the head). This proposal provides more category distinctions by combining different annotation layers, and not by multiplying the number of tags on one layer.

A good case in point is the verb *manifestar* (Table 1), which has a basic transitive argument structure (*manifestamos nuestros sentimientos*, 'we express our feelings'), and which can be used in a passive frame with an inanimate subject, as in (a) or, more exceptionally, in a true reflexive such as (c). However, there are many examples that would be classified into the category of “expl:pv”, both with inanimate (b) and with animate subjects (d). These two examples are far from being clear-cut passives and reflexives, respectively, and thus would better be labeled as “expl:pv”, but they also clearly differ from each other. Intuitively speaking, the first example seems more passive than the second, and the second more reflexive than the first. We believe that these intuitions can be captured by combining the different annotation layers: (b) and (d) receive the same dependency label “expl:pv”, but their underlying features allow distinguishing between them. Concretely, with transitive verbs “Case” has to be disambiguated between “Acc” and “Dat” (for simplicity we leave “Com” out of the discussion), “Reflex” between “Reflex” and “Rcp”, and “Voice” between “Act” and “Pass”. With *manifestar,* “Case” would be “Acc” in the four cases, since this is the role that the “reflected argument” would play in a non-expletive construction (namely an active transitive construction for (b) or a true reflexive for (d)). “Reflex” would also be “Reflex” in the four cases, but “Voice” would be “Pass” in (b) and “Act” in (d), reflecting the intuition that the core semantic role of the subject is to undergo the process in (b), and to control it in (d).

		Dependency relation	Features reflexive pronoun		Features verbal head
			Case	Reflex	Voice
a	*como se manifestó en el periódico*	expl:pass	Acc	Reflex	Pass
b	*los problemas se manifestaron desde el primer día*	expl:pv			
c	*Dios se manifestó a sí mismo en Cristo*	obj	Acc	Reflex	Act
d	*la gente se manifiesta por tercer día consecutivo; el presidente se manifestó de acuerdo con … (*a sí mismo)*	expl:pv			

Table 1: Feature annotation on passives, reflexives and their corresponding “expl:pv”. Translations: (a) ‘as was said in the newspaper’, (b) ‘the problems became clear from the first day’, (c) ‘God materialized himself in Christ’, (d) ‘people demonstrated for the third consecutive day’; ‘the president said he agreed with the proposal’ (*him/herself)

Crucially, the feature sets link (b) with (a), and (d) with (c), respectively. This means that the expletive reflexive in (b) modifies an inherently passive construction (converting it, prototypically, into a spontaneous process, see the middle voice above), and that in (d), the expletive modifies an inherently reflexive construction, evoking event structures in which it is not relevant to distinguish two separate thematic roles for the reflected argument.

Similarly, the “Case=Acc/Dat” and “Reflex=Reflex/Rcp” properties also enable us to distinguish different underlying structures within the broad category of “expl:pv” examples. As is illustrated in Table 2, the difference between accusative (f) and dative reflexive (h) shows similarities with (e) and (g), respectively, and the difference between accusative reflexive (f) and reciprocal (j) is similar to the difference between (e) and (i).

		Dependency relation	Features reflexive pronoun		Features verbal head
			Case	Reflex	Voice
e	*se ve en el espejo*; *se mete en líos*	obj	Acc	Reflex	Act
f	*se ve amenazado de*; *se mete a hacer algo*	expl: pv			
g	*se quita la ropa*; *se da un baño*	iobj	Dat	Reflex	Act
h	*se da cuenta*	expl:pv			
i	*se saludan*; *se quieren mucho* (*el uno al otro*)	obj	Acc	Rcp	Act
j	*se llevan bien*; *se ponen de acuerdo* (**el uno al otro*)	expl:pv			

Table 2: Feature annotation on accusative/dative reflexives, accusative reflexives/reciprocals and their corresponding “expl:pv”. Translations: (e) ‘he sees himself in the mirror’; ‘he gets himself in trouble’, (f) ‘he is threatened by; ‘he starts doing something’, (g) ‘he takes off his clothes’; ‘he takes a bath’, (h) ‘he realizes something’, (i) ‘they greet each other’; ‘they love each other’, (j) ‘they get along well’; ‘they agree’ (*each other)

2.4 Reannotating Spanish UD AnCora

In Table 3 we present a comprehensive view on the proposed encodings. First, the pronouns were disambiguated according to their general reflexive character, distinguishing between *me veo* (‘I see myself’) and *me ven* (‘they see me’). In the latter group, a distinction is made between “obj” and “iobj” (*me dieron algo*, ‘they gave me something’) at the level of the dependency relation.

Secondly, the reflexive uses were assigned one of the dependency labels “expl:pass”, “obj”, “iobj”, “expl:impers” and “expl:pv”. This means that reflexive and non-reflexive “obj” and “iobj” have the same dependency label but are distinguished by the feature “Reflex”, which is absent in the case of non-reflexives. Reflexive “obj” and “iobj” are further subdivided according to their genuine reflexive versus reciprocal use.

Thirdly, the umbrella category “expl:pv” consists of three subgroups, namely constructions with corresponding transitive verbs, constructions which show an alternation with intransitive verbs, and constructions without corresponding (in)transitive verbs. The first group of “transitivity-based” reflexive constructions is then further subdivided by assigning different combinations of feature sets, as was explained in Section 2.3.4. These feature sets overlap with other “non-expl:pv” constructions, showing their shared characteristics. The proposal also foresees an “expl:pv” category with “Case=Dat”, “Reflex=Rcp” and “Voice=Act”, although this use does not seem to occur in Spanish.

Based on this annotation scheme, we then manually reannotated the AnCora treebank (both the test set and the training set). Table 4 includes a quantitative overview of the original dependency relation labels of all potentially reflexive pronouns (note that “expl:impers” and “expl:pv” do not occur in the original treebank), compared to their new labels after manual reannotation. Apart from the (very numerous) changes in dependency label, it is also worth noting that our reannotation removed the “Reflex” feature from 26 non-coreferential instances of *se*, that adding “Voice=Pass” to the feature set of the verbal head now allows identifying the passive reading of 2715 verb forms, and that, finally, the reciprocal character of 105 pronouns is now reflected in the feature set thanks to the introduction of the “Reflex=Rcp” property.

	Features			
	Pronoun		Verb	
	Case	Reflex	Voice	
Reflexive uses				
expl:pass	Acc	Reflex	Pass	*la noticia se publicó*
obj	Acc	Reflex	Act	*Pedro se ve en el espejo*
	Acc	Rcp	Act	*Pedro y Juan se vieron en la calle*
iobj	Dat	Reflex	Act	*Pedro se quita la ropa*
	Dat	Rcp	Act	*Pedro y Juan se dieron la mano*
expl:impers	-	Reflex	Act	*se trabaja mucho*
expl:pv	***with corresponding non-reflexive transitive verb***			
	Acc	Reflex	Pass	*el fenómeno se manifiesta*
	Acc	Reflex	Act	*la gente se manifiesta*
	Acc	Rcp	Act	*Pedro y Juan se ponen de acuerdo*
	Dat	Reflex	Act	*Pedro se da cuenta*
	Dat	Rcp	Act	*?*
	Com	Reflex	Act	*Pedro se llevó el regalo*
	with corresponding non-reflexive intransitive verb			
	-	Reflex	Act	*Pedro se muere*
	without corresponding non-reflexive verb			
	Acc	Reflex	Act	*Pedro se atreve a …*
Non-reflexive uses				
obj	Acc		-	*me/te/nos/os ven*
iobj	Dat		-	*me/te/nos/os/se lo dijeron*

Table 3: Overview of the annotation scheme for potentially reflexive pronouns in Spanish

reannotated original	expl:impers	expl:pass	expl:pv	iobj	obj	**Total**
expl:pass	285	139	28	1	5	**458**
iobj	1	15	217	142	38	**413**
obj	52	1880	2603	573	618	**5726**
Total	**338**	**2034**	**2848**	**716**	**661**	**6597**

Table 4: Overview of the dependency relation changes in Spanish UD AnCora (test + train)

3 Conclusion

We have argued that in current Spanish Universal Dependencies treebanks, the annotation of reflexives is an unsolved problem. Given the frequency of this construction, occurring for example in more than 20% of the sentences in written texts, this has considerable consequences for parser accuracy and/or granularity. Reflexives, and particularly so-called *se* constructions, have been heavily debated in the tradition of Spanish linguistics. Although it cannot be the aim of morpho-syntactic and dependency parsing to reflect all possible semantic nuances, we have shown that a layered annotation strategy, which combines a relatively limited number of UD dependency relations and feature set properties, can capture both constructional similarities and diversity. We applied this proposal to the v2.5 Spanish UD AnCora treebank and provide categorized conversion tables that can be run as a Python script (see Appendix A and B).

Bibliography

Bouma, G., Hajic, J., Haug, D., Nivre, J., Solberg, P. E., and Øvrelid, L. 2018. Expletives in Universal Dependency Treebanks. In *UDW 2018*, 18-26.

Croft, W., Nordquist, D., Looney, K., and Regan, M. 2017. Linguistic Typology meets Universal Dependencies. In *TLT* 2017: 63-75.

Goethals, P. (2018). Customizing vocabulary learning for advanced learners of Spanish. In T. Read, B. Sedano Cuevas, and S. Montaner-Villalba (Eds.), *Technological innovation for specialized linguistic domains* (pp. 229-240). Berlin: Éditions Universitaires Européennes.

Maldonado, R. 2008. Spanish middle syntax: A usage-based proposal for gramar teaching. In S. De Knop and T. De Rycker (eds.) *Cognitive Approaches to Pedagogical Grammar*, 155-196. Berlin: Mouton De Gruyter.

Marković, S., and Zeman, D. 2018. Reflexives in Universal Dependencies. In TLT 2018.

Martínez Alonso, H. and Zeman D. 2016. Universal Dependencies for the AnCora treebanks. In *Procesamiento de Lenguaje Natural*, 57, 91-98.

Mendikoetxea, A. 1999. Construcciones inacusativas y pasivas. In *Gramática descriptiva de la lengua española*, 2, 1575-1629. Espasa Calpe.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. *LREC* 2016.

Peregrín Otero, C. 1999. Pronombres reflexivos y recíprocos. In *Gramática descriptiva de la lengua española*, 1, 1427-1518. Espasa Calpe.

Silveira, N. 2016. *Designing syntactic representations for NLP: An empirical investigation.* PhD Thesis. Stanford University.

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In LREC 2008.

Appendix A: Conversion table

The conversion table includes all occurrences of *me*, *te*, *nos*, *os* and *se*. Other users can modify or customize the annotation decisions.

Appendix B: Python script

The Python script reads in the original CoNLL-U AnCora files, and applies all the changes to the corresponding dependency relations and feature sets. The appended files are available upon request (by email, to Jasper.Degraeuwe@UGent.be).