Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 63-72 recibido 31-03-2022 revisado 17-05-2022 aceptado 29-05-2022

Reflexive pronouns in Spanish Universal Dependencies: from annotation to automatic morphosyntactic analysis

Los pronombres reflexivos en las Universal Dependencies en español: desde la anotación hacia el análisis morfosintáctico automático

Jasper Degraeuwe, Patrick Goethals
Ghent University (Belgium)

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

Abstract: In this follow-up article of Degraeuwe and Goethals (2020), we present the annotation scheme used to reannotate the 7298 potentially reflexive pronouns included in the Universal Dependencies Spanish AnCora v2.6 treebank, which resulted in significant modifications for the "Case" feature (100% changed) and dependency relations (87% changed). Next, we evaluate the performance of spaCy v3.2.2 and Stanza v1.3.0 (both trained on AnCora v2.8, and thus based on our reannotations) on the AnCora v2.8 test set, which yielded weighted F1 scores up to 0.88 and 0.98 for the "Case" and "Reflex" features, respectively, and up to 0.71 for the dependency relations. Finally, the error analysis of the spaCy results underlines the (generalisation) potential of the model, but also reveals some of the remaining issues in the automatic morphosyntactic analysis of reflexive pronouns in Spanish, such as determining if expletive relations denote an impersonal, passive or inherently reflexive use.
Keywords: reflexive pronouns, *se*, Universal Dependencies, morphosyntactic tagging and parsing.

Resumen: En este artículo de seguimiento de Degraeuwe y Goethals (2020), presentamos el esquema de anotación utilizado para reanotar los 7298 pronombres potencialmente reflexivos incluidos en el Universal Dependencies Spanish AnCora v2.6 *treebank*, lo cual resultó en un significativo número de modificaciones para la característica (*feature*) de "Case" (el 100% cambiado) y las relaciones de dependencia (el 87% cambiado). A continuación, evaluamos el desempeño de spaCy v3.2.2 y Stanza v1.3.0 (ambos entrenados en AnCora v2.8, y, por tanto, basados en nuestras reanotaciones) en el *set* de prueba de AnCora v2.8, lo cual dio como resultado puntuaciones de F1 ponderado de hasta 0,88 y 0,98 para las características de "Case" y "Reflex", respectivamente, y de hasta 0,71 para las relaciones de dependencia. Por último, el análisis de errores de los resultados de spaCy subraya el potencial (generalizador) del modelo, pero también desvela algunos de los problemas pendientes en el análisis morfosintáctico automático de los pronombres reflexivos en español, como por ejemplo determinar si las relaciones de dependencia expletivas son de carácter impersonal, pasivo o inherentemente reflexivo.
Palabras clave: pronombres reflexivos, *se*, Universal Dependencies, etiquetado y análisis gramatical morfosintáctico.

1 Introduction

As Natural Language Processing (NLP) tools such as spaCy (spacy.io) and Stanza (stanfordnlp.github.io/stanza/) are becoming more and more accessible (also for a non-expert audience, see e.g. Altinok (2021) and Vasiliev (2020)), automatic morphosyntactic analysis has been integrated into a wide range of text-based applications. By means of a simple programming script, for example, raw corpora

ISSN 1135-5948. DOI 10.26342/2022-69-5 © 2022 Sociedad Española para el Procesamiento del Lenguaje Natural

can be transformed into intelligent resources containing morphosyntactic information. These "enriched corpora" can then be used as input for corpus query tools or language learning environments, enabling their users to perform much more fine-grained queries.

To train their (morphosyntactic) taggers and (syntactic) parsers, NLP tools usually make use of treebanks as reference data. One of the most well-known initiatives concerned with the construction of such treebanks is the Universal Dependencies (UD) project, established in 2014. In a nutshell, UD aims at developing "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 22 February 2022; see also Nivre et al. (2016)). Together with the growing number of languages included in the project (v2.9 contains 217 treebanks in 122 languages), the standardised, cross-linguistically consistent approach of UD has led to increasing usage of UD treebanks not only for the development of NLP tools, but also for research purposes (see de Marneffe et al. (2021, p. 304) for an overview).

However, the UD initiative is also a "constantly improving effort" (Martínez Alonso and Zeman, 2016), implying that the annotation guidelines are regularly updated and fine-tuned over the successive releases of the treebanks. Furthermore, within UD there remain several annotation issues, which may be problematic from both a cross-linguistic and an intra-linguistic perspective. In a previous article (Degraeuwe and Goethals, 2020), we addressed one of those pending issues, namely the annotation of the potentially reflexive pronouns *me*, *te*, *nos*, *os* and *se* (see also Marković and Zeman, 2018) in the UD Spanish AnCora treebank (Martínez Alonso and Zeman, 2016; Taulé, Martí and Recasens, 2008). These are very frequent items in Spanish, with *se* alone occurring in almost 30% of the sentences in the AnCora treebank and being ranked eleventh in the list of most common lemmas in CORPES XXI (Real Academia Española de la Lengua, 2022).

The annotation proposal described in Degraeuwe and Goethals (2020) was revised by the UD contributor responsible for the AnCora treebank (see sections 2.1.1 and 2.1.2 for the details of the original and revised annotation schemes), after which all potentially reflexive pronouns in the treebank were reannotated and the resulting 7298 changes were pushed to the UD project (visible from v2.7 onwards). In 2021, both spaCy (with v3.0) and Stanza (with v1.2.0) released thoroughly updated versions of their tools trained on UD v2.7 or higher (thus including the reannotated reflexive pronouns). In this follow-up contribution, we will carry out both a quantitative and qualitative analysis of how well the tools perform on morphosyntactically analysing Spanish reflexive pronouns, and we will identify the key remaining issues (see section 3).

2 Literature overview

2.1 Reflexives in Spanish Universal Dependencies

The UD framework provides three main 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 such as subject or (in)direct object, and a feature set containing additional lexical and grammatical properties (e.g. number or person in the case of pronouns).

2.1.1 Annotation scheme as proposed in Degraeuwe and Goethals (2020)

The original proposal (see Table 1) arose from the notion that, as NLP tools become more accessible, more theoretical linguists will use them and evaluate their linguistic accuracy and granularity. Consequently, we not only focused on improving annotation consistency in order to increase tagger/parser accuracy, but also took into account the (cross-)linguistic analyses made in non-computational linguistics (Croft et al., 2017; Maldonado, 2008; Mendikoetxea, 1999; Peregrín Otero, 1999).

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, the dependency relation is defined as "obj" or "iobj" (*me dieron algo*, 'they gave me something').

Secondly, the reflexive uses were assigned one of the dependency labels "obj", "iobj", "expl:impers", "expl:pass" and "expl:pv". This

means that reflexive and non-reflexive "obj" and "iobj" have the same dependency label but are distinguished by the "Reflex" feature, 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. These feature sets overlap with other "non-expl:pv" constructions, showing their shared characteristics.

With this annotation proposal, many annotation inconsistencies were resolved (e.g. up to 30% and 60% of false positives of "expl:pass" and "iobj", respectively). Furthermore, the proposal also provided a more fine-grained and informative categorisation, as the previous taxonomy (AnCora ≤ v2.6) did not allow distinguishing between, for example, 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 las armas*, 'the police seized the guns'). In all these cases, *se* was labelled as "obj", and no differences were to be found between the feature sets of the *se* instances, nor between the feature sets of their verbal heads.

	Features			
	Pronoun		Verb	
	Case	Reflex	Voice	
Reflexive uses				
expl:pass	Acc	Reflex	Pass	(a) *la noticia se publicó*
obj	Acc	Reflex	Act	(b) *Pedro se ve en el espejo*
	Acc	Rcp	Act	(c) *Pedro y Juan se vieron en la calle*
iobj	Dat	Reflex	Act	(d) *Pedro se quita la ropa*
	Dat	Rcp	Act	(e) *Pedro y Juan se dieron la mano*
expl:impers	-	Reflex	Act	(f) *se trabaja mucho*
expl:pv	***With corresponding non-reflexive transitive verb***			
	Acc	Reflex	Pass	(g) *el fenómeno se manifiesta*
	Acc	Reflex	Act	(h) *la gente se manifiesta*
	Acc	Rcp	Act	(i) *Pedro y Juan se ponen de acuerdo*
	Dat	Reflex	Act	(j) *Pedro se da cuenta de que …*
	Dat	Rcp	Act	[does not occur in Spanish]
	Com	Reflex	Act	(k) *Pedro se llevó el regalo*
	With corresponding non-reflexive intransitive verb			
	-	Reflex	Act	(l) *Pedro se muere*
	Without corresponding non-reflexive verb			
	Acc	Reflex	Act	(m) *Pedro se atreve a …*
Non-reflexive uses				
obj	Acc	-	-	(n) *me/te/nos/os ven*
iobj	Dat	-	-	(o) *me/te/nos/os/se lo dijeron*

Table 1: Overview of the annotation scheme for potentially reflexive pronouns in Spanish as proposed in Degraeuwe and Goethals (2020). Translations: (a) 'the news was published', (b) 'Pedro sees himself in the mirror', (c) 'Pedro and Juan see each other on the street', (d) 'Pedro takes off his clothes', (e) 'Pedro and Juan shake hands', (f) 'a lot of work is being done', (g) 'the phenomenon becomes clear', (h) 'people are demonstrating', (i) 'Pedro and Juan agree', (j) 'Pedro realises that …', (k) 'Pedro took the present with him', (l) 'Pedro dies', (m) 'Pedro dares to …', (n) 'they see me/you/us/you', (o) 'they told it to me/you/us/you/him/her'.

2.1.2 Annotation scheme used for AnCora ≥ v2.7 annotations

As the annotation scheme presented in section 2.1.1 included some drastic modifications, the proposed changes were first revised by the UD contributor responsible for the AnCora treebank before applying any reannotations. Based on this feedback, the reflexive – reciprocal distinction was dropped: since "Reflex" is currently a Boolean feature in all UD treebanks, the addition of a "Rcp" value would lower cross-linguistic consistency. Moreover, the distinction also showed to be a too subtle one to make for machine learning methods (tested with custom models built on spaCy v3.2.2 and Stanza v1.3.0 architectures).

Secondly, changing the "Voice" feature of the verbal head of the potentially reflexive pronoun was also discarded, again to give full priority to cross-linguistic consistency. Although these characteristics do seem to be recognisable for machine learning models (average weighted F1 test set scores of 0.78 for custom model trained with spaCy v3.2.2 architecture and 0.61 with Stanza v1.3.0), the "Voice=Pass" feature was primarily designed for annotating verbal paradigms which distinguish active from passive voice morphologically, which is not the case in Spanish.

Even though the modifications presented above slightly decrease granularity, the annotation proposal (see Table 2) remains very informative (five different dependency labels, accusative/dative/comitative case distinction and reflexive/non-reflexive use distinction). Moreover, the new annotation scheme now adheres very strictly to the UD principles and guidelines.

The reannotation of the potentially reflexive pronouns was implemented from AnCora v2.7 onwards. Since all "Case" values of pronouns were labelled as "{Acc, Dat}" in the v2.6 treebank, all of the 7298 pronouns present in the development, test and training sets received a new "Case" value: 5933 instances were reannotated as "Acc", 893 instances as "Dat" and 472 instances as "NA" (non-applicable, for non-cased instances). The comitative case ("Com") did not occur in the data.

	Features		
	Case	Reflex	
Reflexive uses			
expl:pass	Acc	Yes	(a) *la noticia se publicó*
obj	Acc	Yes	(b) *Pedro se ve en el espejo* (c) *Pedro y Juan se vieron en la calle*
iobj	Dat	Yes	(d) *Pedro se quita la ropa* (e) *Pedro y Juan se dieron la mano*
expl:impers	-	Yes	(f) *se trabaja mucho*
expl:pv	***With corresponding non-reflexive transitive verb***		
	Acc	Yes	(g) *el fenómeno se manifiesta* (h) *la gente se manifiesta* (i) *Pedro y Juan se ponen de acuerdo*
	Dat	Yes	(j) *Pedro se da cuenta de que …*
	Com	Yes	(k) *Pedro se llevó el regalo*
	With corresponding non-reflexive intransitive verb		
	-	Yes	(l) *Pedro se muere*
	Without corresponding non-reflexive verb		
	Acc	Yes	(m) *Pedro se atreve a …*
Non-reflexive uses			
obj	Acc	-	(n) *me/te/nos/os ven*
iobj	Dat	-	(o) *me/te/nos/os/se lo dijeron*

Table 2: Overview of the annotation scheme for potentially reflexive pronouns in Spanish used in AnCora ≥ v2.7.

AnCora ≥ v2.7 / Ancora v2.6	expl:impers	expl:pass	expl:pv	iobj	obj	**Total (Ancora v2.6)**
expl:pass	301	159	35	1	6	**502 (6.9%)**
iobj	1	17	253	152	43	**466 (6.4%)**
obj	54	2052	2927	628	665	**6326 (86.7%)**
other	0	3	0	1	0	**4 (0,1%)**
Total (AnCora ≥ v2.7)	**356 (4.9%)**	**2231 (30.6%)**	**3215 (44.1%)**	**782 (10.7%)**	**714 (9.8%)**	**7298**

Table 3: Overview of the dependency relation changes in Spanish UD AnCora ≥ v2.7 compared to v2.6 (dev + test + train).

Next, the "Reflex" value of 7108 pronouns (97%) remained unaltered: 6483 instances maintained their reflexive annotation ("Yes") and 625 instances their non-reflexive character ("NA"). However, 49 pronouns were changed from reflexive to non-reflexive, while the value of 141 instances was modified the other way around from non-reflexive to reflexive.

Finally, 6322 of the 7298 potentially reflexive pronouns (almost 87%) received a new dependency label. A detailed, quantitative overview of the corresponding changes is presented in Table 3 (note that "expl:impers" and "expl:pv" do not occur in the v2.6 treebank). The statistics show a fundamental shift from "obj" as the predominant label to a more dispersed distribution, with "expl:pv" and "expl:pass" being the most important labels. In other words, the reannotation shows that reflexive pronouns usually express an expletive use, more specifically an inherently reflexive use ("expl:pv") or a passive one which blurs the subject role ("expl:pass").

2.2 Reflexives in machine learning

To our knowledge, to date no studies have been performed which focus on the performance of machine learning models at tagging potentially reflexive pronouns in Spanish based on UD treebank data. On non-UD data, however, some experiments have been carried out. In Aldama García and Barbero Jiménez (2021), for example, a machine learning approach is adopted to predict the dependency label of *se* in a one-per-sentence setup, for which a custom "*se* corpus" was compiled containing 2140 sentences from CORPES XXI (Real Academia Española de la Lengua, 2022). The corpus was annotated according to a four-category annotation scheme, containing the "se-mark" (for cases of valency reduction, such as passive and impersonal constructions), "expl" (for pure pronominal predicates or emphatic contexts), "iobj" (for indirect objects) and "obj" (for direct objects) labels. Next, nine different machine learning classifiers were applied to the test set of the corpus, with pre-trained language models based on a transformers architecture obtaining the best performance (macro F1 score of 0.7).

Results such as these indicate that, to a certain extent, recent machine learning methods are able to successfully distinguish different uses of the potentially reflexive pronoun *se*. To implement them in real-life scenarios, approaches as in Aldama García and Barbero Jiménez (2021), which are based on a language-specific setup and require a self-compiled and annotated set of training and test data, can be integrated as a custom component for that specific language in NLP tools such as spaCy and Stanza. This way, the morphosyntactic information offered by the tool's tagger and parser can be complemented by the output of the task-specific model.

However, the creation of such models is a very time-consuming operation, especially for non-computational linguists. Therefore, in section 3, we will study the potential of the default taggers and parsers included in spaCy and Stanza (which are trained on UD data), and analyse if they would need to be complemented by a task-specific model and where exactly (i.e. for which labels) issues arise.

3 Automatic morphosyntactic analysis of potentially reflexives pronouns

From section 2.1.2 it can be concluded that, in theory, any NLP tool trained on the reannotated treebank as input data should be able to perform a more fine-grained morphosyntactic analysis of potentially reflexive pronouns in Spanish. To evaluate the validity of this claim, we apply the large pretrained Spanish model of spaCy v3.2.2 ("es_core_news_lg") and the default pretrained

		Case			Reflex		Dependency relation				
		Acc	Dat	NA	Yes	NA	expl:impers	expl:pass	expl:pv	iobj	obj
#instances		**452**	**47**	**42**	**504**	**37**	**30**	**171**	**254**	**36**	**50**
spaCy	F1	0.94	0.62	0.57	0.99	0.88	0.51	0.75	0.8	0.6	0.37
	macro avg	0.71			0.93		0.6				
	weighted avg	0.88			0.98		0.71				
Stanza	F1	0.9	0.46	0	0.98	0.78	0.5	0.75	0.76	0.56	0.23
	macro avg	0.45			0.88		0.56				
	weighted avg	0.79			0.97		0.68				

Table 4: Results (macro and weighted F1) of automatic morphosyntactic analysis of potentially reflexive pronouns in the AnCora v2.8 test set using spaCy v3.2.2 and Stanza v1.3.0.

Spanish model of Stanza v1.3.0, which are both trained on the UD Spanish AnCora v2.8 training and development sets, to the corresponding AnCora v2.8 test set. This test data includes 668 potentially reflexive pronouns, of which 127 instances are clitic forms such as *se* in *la gente va a la calle a manifestarse* ('people go to the streets to demonstrate'). As spaCy does not include a multiword tokeniser (which is required to split words with clitics into so-called "subword tokens" and then analyse these separate tokens instead of the entire word form), clitic forms will be excluded from the evaluation in order to obtain comparable results. Table 4 presents a detailed overview of the morphosyntactic analysis, with F1 as the evaluation metric and the number of instances for each label included in the "#instances" row.

Finally, some architectural characteristics of the NLP tools should be highlighted: in the spaCy pipeline, the tagger and parser components listen to the same word embedding component but do not share any information between them, implying that the features and dependency relations are predicted independently of each other. Stanza, however, does take into account information from the tagger when training its dependency parser, which means that the Stanza dependency relation predictions partially depend on the feature predictions.

For the "Case" and "Reflex" features, satisfying results are obtained, especially with spaCy (weighted F1 scores of 0.88 for "Case" and 0.98 for "Reflex"). Stanza, however, does not seem to be able to recognise non-cased uses (see "NA" column), which correspond to reflexive pronouns with "expl:impers" as the dependency label and to "expl:pv" relations with verbs for which a corresponding non-reflexive intransitive counterpart exists (see also Table 2).

As far as the dependency relations are concerned, the automatic morphosyntactic analysis achieves relatively good results as well, with weighted F1 scores of 0.71 for spaCy and 0.68 for Stanza. Compared to the top macro F1 score of 0.7 reached in Aldama García and Barbero Jiménez (2021), both spaCy (0.6) and Stanza (0.56) perform worse, although it should be observed that Aldama García and Barbero Jiménez (2021) distinguish only four instead of five categories and exclusively focus on *se* as potentially reflexive pronoun (and not on *me*, *te*, *nos* and *os*). Next, the low scores for the "expl:impers" and especially "obj" category have to be highlighted. A first possible explanation for this lower performance could be the limited number of training instances in the training and developments sets: 268 for "expl:impers" (5.07%) and 444 for "obj" (8.4%). To gain more in-depth insights into this matter, and into the errors made by NLP tools in general, an additional analysis is performed based on the contingency tables included in Table 5, which zero in on the performance of spaCy, the best-performing tool. The error analysis will also include a qualitative component, with special attention to the generalisation potential of the tool (i.e. if it has learnt to make predictions based on patterns, not just to predict the most frequent label for each word form).

For the "Case" errors, three main findings can be extracted from the results:

1. Predicting the correct case of *me*, *te*, *nos* and *os* when they are used in accusative case is challenging (see (p) and (s) in Table 6 for some examples): together, these pronouns account for 29 of the 452 accusative instances, and 13

Case

correct \ predicted	Acc	Dat	NA	Total
Acc	435	14	3	452
Dat	19	28	0	47
NA	23	1	18	42
Total	477	43	21	541

Reflex

correct \ predicted	Yes	NA	Total
Yes	496	8	504
NA	2	35	37
Total	498	43	541

Dependency relation

correct \ predicted	expl:impers	expl:pass	expl:pv	iobj	obj	other	Total
expl:impers	14	9	5	1	1	0	30
expl:pass	7	138	24	1	0	1	171
expl:pv	4	39	201	7	3	0	254
iobj	0	5	3	25	3	0	36
obj	0	7	16	14	13	0	50
Total	25	198	249	47	21	1	541

Table 5: Results (contingency tables) of automatic morphosyntactic analysis of potentially reflexive pronouns in the AnCora v2.8 test set using spaCy v3.2.2.

of those 29 cases received the wrong "Dat" prediction (which corresponds to 13 of the 17 errors made for the accusative label). Importantly, 12 of those instances had also received a wrong dependency label (namely "iobj" instead of "obj" or "expl:pv").

2. Predicting the correct case of *se* when it is used in dative case also entails challenges (see (q) in Table 6 for an example): *se* accounts for 24 of the 47 dative instances, and 15 of those 24 cases were labelled wrongly as accusative (corresponding to 15 of the 19 errors made for this label).
3. As for the non-cased instances, the errors seem to indicate that the model does not just naïvely link labels to verbal heads, since in sentences with *irse*/*marcharse* ('to leave'), which frequently occur in the training data and under all circumstances receive the "NA" label, 5 incorrect but also 4 correct predictions were to be found. This finding can be considered evidence that the model has developed a kind of generalisation procedure, although it thus results in the introduction of some errors in the case of *irse*/*marcharse*. In this regard, it also appears that the generalisation as such is not entirely successful either, since several of the "Case" errors correspond to reflexive pronoun – verbal head combinations which (almost) do not occur in the training (and development) data, as was the case with *advertir*. For this verb, which only occurs once as a verbal head (i.e. the verbal form on which the potentially reflexive pronoun is syntactically dependent) in the training set, the test sentence (r) (see Table 6) was wrongly predicted as "Acc", meaning that the model was not able to generalise, in this particular case at least, from similar examples with other verbal heads (e.g. *contratar* as in *se contrata a alguien* 'someone was given a contract', which occurs three times in the training data).

Next, the (few) errors made for the "Reflex" feature (see (s) in Table 6) usually correspond to instances of *me*, *te*, *nos* and *os* for which the

model also wrongly predicted both the case (usually dative instead of accusative case) and the dependency relation (usually "iobj" instead of "expl:pv" or "obj"). Especially the wrong "iobj" prediction provides a plausible explanation for the error in the "Reflex" label, as in the training data non-reflexive "iobj" instances are twice as frequent as reflexive "iobj" instances.

Thirdly, the error analysis of the dependency relation predictions led to again three main findings:

1. Errors in one of the "expl" categories almost always correspond to one of the two other "expl" labels (14 of the 16 errors in "expl:impers", 31/33 in "expl:pass" and 43/53 in "expl:pv"). In other words, the model has no problem in identifying expletive uses of potentially reflexive pronouns, but assigning the right expletive subcategory seems to be a less straightforward operation from a machine learning point of view (see sentences (q) and (r) in Table 6).
2. For "iobj", 5 of the 11 errors are "expl:pass" predictions. Looking at the sentences, it appears that the model is not able to predict the "iobj" label for reflexive pronouns which co-occur with an explicit subject and direct object, as in the examples (t) and (u) in Table 6.
3. Predicting the correct dependency label of *me*, *te*, *nos* and *os* when they are used as direct object also poses challenges to spaCy's machine learning model (see (p) and (s) in Table 6): together, these pronouns account for 19 of the 50 "obj" instances, and 17 of those 19 cases received a wrong dependency relation (which corresponds to 17 of the 37 errors made for the "obj" label). Moreover, all of the 14 cases where an "iobj" instance was predicted instead of "obj" were to be found amongst those 17 errors, highlighting the sometimes fuzzy boundary between *me*, *te*, *nos* and *os* acting as direct or indirect object.

Finally, as a general, overarching observation it should be noted that the generalisation potential of the model, which was already briefly addressed in the discussion of the "Case" errors, also comes to the fore with pronoun – verbal head combinations which have multiple possible uses. A good case in point is the combination with the verbal head *tratar*, which occurs 111 times as "expl:impers" and 2 times as "expl:pass" in the training and development data. Despite the imbalanced distribution in the training data, the test sentence containing *los temas se tratarán* ('the topics will be treated') still got labelled correctly as "expl:pass", which hints at the fact that the model has leveraged "knowledge" from other "expl:pass" training examples to arrive at this correct prediction. Still other evidence of

Sentence	Case		Reflex		Dependency relation	
	correct	predicted	correct	predicted	correct	predicted
(p) […] nos trataron muy mal […]	Acc	Dat	NA	NA	obj	iobj
(q) […] las carreteras catalanas se cobraron 16 vidas […]	Dat	Acc	Yes	Yes	expl:pv	expl:pass
(r) Si se hubiera advertido a la gente […]	NA	Acc	Yes	Yes	expl:impers	expl:pass
(s) […] no me engaño a creer en la existencia de […]	Acc	Dat	Yes	NA	obj	iobj
(t) […] Beckenbauer se permite bromear […]	Dat	Acc	Yes	Yes	iobj	expl:pass
(u) […] el affaire Cristo-Rey se tomaba un respiro […]	Dat	Acc	Yes	Yes	iobj	expl:pass

Table 6: Selection of errors in automatic morphosyntactic analysis of potentially reflexive pronouns in the AnCora v2.8 test set using spaCy v3.2.2. Translations: (p) '[…] they treated us really badly […]', (q) '[…] Catalan roads claimed 16 lives […]', (r) 'If people had been warned […]', (s) '[…] I don't delude myself into believing in the existence of […]', (t) '[…] Beckenbauer affords himself to make jokes […]', (u) '[…] the Cristo-Rey affair took a breather […]'.

the generalisation potential can be found in the accuracy rates the model obtains for the 35 pronoun – verbal head combinations which do not occur at all in the training data: 73% for "Case", 97% for "Reflex" and 54% for the dependency relations.

4 Conclusion

In this article, we built upon Degraeuwe and Goethals (2020), in which a proposal was formulated to reannotate the potentially reflexive pronouns (*me*, *te*, *nos*, *os* and *se*) in the Universal Dependencies Spanish AnCora treebank. These items, and in particular *se*, occur very frequently in Spanish, and have also received much attention in non-computational linguistics (Croft et al., 2017; Maldonado, 2008; Mendikoetxea, 1999; Peregrín Otero, 1999). Taking into account that treebanks are used as reference data to train the models offered by state-of-the-art NLP tools such as spaCy and Stanza, we aim to contribute to improving the NLP-driven morphosyntactic analysis of potentially reflexive pronouns, and in doing so, also help creating higher-quality "enriched resources" which can be used as input for, amongst other applications, corpus query tools and language learning environments.

We presented the slightly modified annotation scheme used to reannotate the potentially reflexive pronouns included in the AnCora v2.6 treebank (7298 items in total; changes visible from v2.7 onwards), which resulted in label changes for all "Case" features, 3% of the "Reflex" features and 87% of the dependency relations. The application of spaCy v3.2.2 and Stanza v1.3.0 (both trained on AnCora v2.8, and thus based on our reannotations) to the AnCora v2.8 test set yielded promising results, hinting at the potential of using NLP-driven methods to perform fine-grained morphosyntactic analyses.

Finally, the error analysis on the spaCy results revealed some of the remaining issues in the automatic morphosyntactic analysis of potentially reflexive pronouns in Spanish (e.g. determining the right subcategory of expletive dependency labels), but also underlined the (generalisation) potential of the underlying model.

Although more than satisfactory performance levels were achieved (weighted F1 up to 0.88 for "Case", 0.98 for "Reflex" and 0.71 for dependency relations), there is still room for improvement, especially for the prediction of dependency relations. Therefore, future work could consist in studying if a task-specific model (as in Aldama García and Barbero Jiménez (2021)) can complement the default taggers and parsers of NLP tools in order to push performance. Furthermore, it is worth considering to implement rule-based predictions for a fixed set of verbs which always yield the same labels when functioning as verbal head of a reflexive pronoun (e.g. for *irse* and *marcharse*), and to define rules which determine the feature values for a given dependency relation (e.g. *if* "expl:pv" is predicted as the dependency relation *then* the "Reflex" feature should always be "Yes"). In spaCy, for instance, such specific rules can be easily implemented thanks to the "attribute ruler" component, which manages mappings and exceptions at token level.

Acknowledgements

This research has been carried out as part of a PhD fellowship on the IVESS project (file number 11D3921N), funded by the Research Foundation – Flanders (FWO).

References

Altinok, D. 2021. *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*. Packt.

Croft, W., D. Nordquist, K. Looney, and M. Regan. 2017. Linguistic Typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63-75, Indiana University (Bloomington, Indiana).

de Marneffe, M-C., C.D. Manning, J. Nivre and D. Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2): 255-308.

Degraeuwe, J., and P. Goethals. 2020. Reflexive pronouns in Spanish Universal Dependencies. *Procesamiento del Lenguaje Natural*, 64: 77-84.

Maldonado, R. 2008. Spanish middle syntax: A usage-based proposal for grammar teaching. In S. De Knop and T. De Rycker (eds.) *Cognitive Approaches to Pedagogical*

Grammar, 155-196. Mouton De Gruyter, Berlin.

Marković, S. and D. Zeman. 2018. Reflexives in Universal Dependencies. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 131-146, Oslo University (Oslo).

Martínez Alonso, H. and D. Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del 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. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659-1666, European Language Resources Association (Portorož).

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

Real Academia Española de la Lengua. 2022. Banco de datos (CORPES XXI) [online]. Corpus del Español del Siglo XXI (CORPES) https://www.rae.es/recursos/banco-de-datos/corpes-xxi. Accessed date: 22/02/2022

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (Marrakech).

Vasiliev, Y. 2020. *Natural Language Processing with Python and spaCy: A Practical Introduction.* No Starch Press.