Miriam Bouzouita, Johnatan E. Bonilla & Rosa Lilia Segundo Díaz

Gaming for Dialects: Creating an Annotated and Parsed Corpus of European Spanish Dialects through GWAPs

1 Introduction

Despite the recent increase in interest in dialectal grammars by both dialectologists (e.g., Fernández-Ordóñez & Pato 2020) and generative linguists (e.g., Castillo *et al.* 2020), it is well-known that, for Spanish, diatopic microvariation in morpho-syntax has been much less explored in comparison to lexical and phonetic matters. Though there exist sociolinguistic corpora that focus on spoken Spanish, such as those that are part of the *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA, "Project for the Sociolinguistic Study of Spanish from Spain and America"; Moreno Fernández 2005) which encompasses more than 40 research groups gathering oral data from various cities in the Hispanic-speaking world, or the *Corpus Oral de Lenguaje Adolescente* (COLA, "Oral Corpus of Adolescent Language"; Jørgensen *et al.* 2002–2017; Jørgensen & Eguía Padilla 2015) which contains samples of youth speech from Madrid, Buenos Aires, Santiago de Chile and Managua, few of these corpora are morpho-syntactically annotated and/or parsed. There are some notable exceptions though, such as the Spanish part of the *Integrated Reference Corpora for Spoken Romance Languages* (C-ORAL-ROM; Moreno-Sandoval *et al.* 2005; Moreno-Sandoval & Guirao 2006), the *Corpus Oral de Español como Lengua Extranjera* (CORELE, "Oral Corpus of Spanish as a Foreign Language"; Campillos Llanos 2016), and the *Corpus del Habla de Baja California* (CHBC, "Corpus of Baja California Speech"; Rico-Sulayes *et al.* 2017).

However, none of these corpora focus on the European Spanish diatopic varieties. In recent years, there have also been some initiatives that center on spontaneous web speech, such as the *Latin American Spanish Discussion Forum Treebank* (LAS-DisFo; Taulé *et al.* 2015), which is not an open access tool and is lodged behind a pay wall. Although this morpho-syntactically annotated corpus also contains non-standard fragments and thus shares the problem with oral corpora that Natural Language Processing (NLP) tools trained on standard written texts perform badly when applied to this type of data (see also section 5.2.3), it should not be forgotten that LAS-DisFo's focus is on written spontaneous language, which has its own idiosyncrasies, not necessarily shared with its oral counterpart (e.g., typos, emoticons, see Taulé *et al.* 2015).

https://doi.org/10.1515/9783110781465-005

In sum, up until now there are no morpho-syntactically annotated and parsed corpora available for spoken Spanish dialects. In what follows, we will present an interdisciplinary crowd-sourced project that aims to fill this gap in order to stimulate and enhance more fine-grained dialectal morpho-syntactic research.[1]

More specifically, the general aim of this project consists in creating a morpho-syntactically annotated and parsed corpus of the European Spanish dialects, the so-called *Corpus Oral y Sonoro del Español Rural – Anotado y Parseado* (COSER-AP, "Annotated and Parsed Audible Corpus of Spoken Rural Spanish") and later renamed COSER-UD (COSER-Universal Dependencies, Bonilla *et al.* 2022, 2023). As the name of this new annotated and parsed corpus indicates, its basis is the *Corpus Oral y Sonoro del Español Rural* (COSER, "Audible Corpus of Spoken Rural Spanish", Fernández-Ordóñez 2005-present), currently the largest collection of spoken Spanish data. Recently, similar initiatives to create annotated and parsed corpora for spoken dialects have been undertaken for other languages too, such as the PARLARS corpus for Valencian Catalan (Esplà-Gomis & Sentí in prep.; Montserrat & Segura 2020) and the *Gesproken Corpus van de zuidelijk-Nederlandse Dialecten* for Southern Dutch (GCND, "Spoken Corpus of Southern Dutch Dialects"; Farasyn *et al.* 2022; Breitbarth *et al.* 2020; Ghyselen *et al.* 2020), among others.

As regards the objectives of this paper, we will present the project design and the challenges that have been encountered in the first two project phases (see section 3) while constructing the COSER-UD through the interdisciplinary approach that was used, in which the fields of Dialectology, NLP and Human-Computer Interaction (HCI) intertwine. As will become clear, this project also employs a citizen science methodology given that linguistic confirmations and corrections are obtained from the general public through various online Games With A Purpose (GWAPs), collectively referred to as *Juegos del español* (Bouzouita et al. 2022).

As concerns the structure of this paper, first, we will introduce the COSER corpus on which this interdisciplinary crowd-sourced project is based (section 2). Subsequently, we will present the different project stages and detail the various tasks involved in each stage (section 3). In section 4, we briefly introduce the various GWAPs of *Juegos del español* that have been created for the confirmation and correction of the automatically generated morpho-syntactic tags, while section 5 deals with various tasks of Phase I and II carried out by the Linguistics team, such as the pre-processing of the transcriptions of the COSER corpus (section 5.1), morpho-syntactic annotation of the COSER corpus (section 5.2), which provides

1 This international research project, titled "A (Respeaking and) Collaborative Game-Based Approach to Building a Parsed Corpus of European Spanish Dialects" (I000418N; PI: M. Bouzouita), has been financed by the Flemish Research Fund (*Fonds voor wetenschappelijk onderzoek*, FWO; 2018–2023).

details on the framework used (section 5.2.1), the creation of the COSER-PoS (COSER-Parts-of-Speech; section 5.2.2), on the results of a study evaluating the tagging accuracy of one of the automatic taggers that have been tested (section 5.2.3) and of the human annotators (section 5.2.4), and on the post-tagging knowledge transfer of the data obtained by the players of *Juegos del español* (section 5.2.5). In the final part, we will draw some conclusions (section 6).

2 The COSER Corpus: Contents and Transcription Protocol

As mentioned before, the COSER corpus (Fernández-Ordóñez 2005-present) is the beating heart of the COSER-UD. In view of this, this section will provide more details on this corpus, its contents, and goals. The primary objective of the COSER corpus is to document diatopic variation, especially morpho-syntactic one, in rural areas of Spain. This database of spoken Spanish is constructed using transcriptions of semi-directed sociolinguistic interviews with elderly men and women living in rural parts of Spain, who have little to no formal education and who enjoyed limited mobility throughout their lives. In other words, the COSER informants coincide largely with those used in traditional dialectology, i.e., the so-called NORM (Non-mobile Older Rural Men, Chambers & Trudgill 1998), though women's speech has also been included in the COSER corpus. As of December 2022, 2.961 informants have been interviewed, 1.415 men and 1.546 women to be precise, for 1.415 locations in 55 provinces and islands. As the COSER focuses on the speech of elderly, the average age of its informants is quite high, 74.2 years to be exact: 75 years for the men and 73.6 for the women. In total, 1.772 interviews have been conducted, for which 1.910 hours of spoken Spanish have recorded, and of which 218 interviews have been transcribed, corresponding to 295 hours and 48 minutes. In other words, at the moment, a mere 12.3% of the total interviews has been transcribed (or 15.4% of the total recorded hours), though we are currently exploring ways to improve the transcription rate significantly using newly developed automatic tools that use large-scale weak supervision, such as Whisper (Radford *et al.* 2023).

As regards the COSER's transcriptions, although that the conventions have changed a few times due to newly acquired insights gained during the corpus construction process, currently, the transcription protocol mixes orthographic and non-standard considerations. More precisely, they try to reflect the pronunciation of mostly phonological (and not phonetic) phenomena that can be found in spoken rural Spanish, while also adopting commonly used spelling rules. This decision distinguishes COSER from other spoken Spanish corpora, such as PRESEEA

and COLA, that have adopted orthographic transcription protocols (see Ghyselen *et al*. 2020 for the advantages and disadvantages of the various types of transcription protocols). The two main phonological changes that have been included in the COSER transcriptions are the omission and addition of phonological segments (de Benito Moreno *et al*. 2016: 79). Let us consider the transcription in example (1), in which the speech of the informant (I1) contains both types of phonological changes (unlike the one of the interviewer, E1):

(1) I1: *me pagaban el rebaje y ganaba ochocientas cincuenta pesetas* ***to*** *los meses*.
E1: *¿El rebaje qué es?*
I1: *La* ***comía****. Lo que cuesta la* ***comía*** *en el cuartel.*
[…] *y yo, ¿qué gastaba? Si diba al bare o al aquello que salía con las chicas, le compraba una peseta* ***e*** *golosinas o tal, que una peseta daban... buah. Y, o echaba un café en el bar o cosas de esas.* ***Demás****, no tenía que gastar. ¿****Pa*** *qué?* ***Comía*** *tenía, pues... Y, ya digo, traje el doble dinero del que llevé.*

I1: 'They paid me the rebate and I earned eight hundred and fifty pesetas every month.'
E1: 'What is the rebate?'
I1: 'The food. What the food costs in the barracks.'
'[…] and I, what did I spend? If I went to the bar or to there where I went out with girls, I would buy her a peseta of candy or something, and they would give a peseta... buah. And, or I would have a coffee at the bar or things like that. Besides, I didn't have to spend. For what? Food I had, well... And, I'm telling you, I brought twice as much money back as I took.'
(COSER-5506-01: M, 88 years, El Remo, Los Llanos de Aridane, in La Palma)

(2) I1: "*Coño, [NP], ¿qué* ***t'ha*** *pasao?*". *Digo: "Pues, tonto este, que* ***s'ha*** *dejao el jefe la escopeta* […]
'"Damn, [Name], what happened to you?" I say: "Well, this fool, the boss left the shotgun […]"'
(COSER-0222-01: M, 82 years, Povedilla, in Albacete)

As can be seen, the COSER transcription presented in example (1) respects, on the one hand, (i) the suppression of phonological segments, represented in bold, as in *to* for *todos* 'all', *comía* for *comida* 'food', *e* for *de* 'of', *demás* for *además* 'besides', *pa qué* for *para qué* 'what for', and, on the other, (ii) the increase in phonological segments in the underlined items, as in *diba* for *iba* 'I went' and *bare* for *bar* 'bar', the latter of which is a typical case of paragogic *-e*, commonly found among elderly speakers in the Canary Islands (see Castillo Lluch *et al*. 2022 for more de-

tails). Similarly, non-standard stress changes and the concatenation of sounds are represented in the COSER transcriptions, as for instance in the pronunciation of *pajáro* [pa'xaro] instead of the standard *pájaro* ['paxaro] 'bird', or *t'ha* and *s'ha* for *te ha* and *se ha* respectively, as illustrated in example (2) (de Benito Moreno *et al.* 2016: 79, Bonilla *et al.* 2022).

Features that are typical of spontaneous conversations are also represented in the COSER transcriptions, such as overlaps, interruptions, and self-corrections. Concretely, overlaps are inserted within the transcription of the speech of the first speaker at the point where the overlapping fragment starts (though no indication is given on where it ends), and are marked with square brackets followed by HS, which stands for *Habla simultánea* 'simultaneous speech', and a specification of who the second interlocutor is, as in [HS:E1] (see the transcription section on the COSER website for more information on other types of overlaps and their transcription). Interruptions are signalled by a hyphen (-), while self-corrections by a vertical bar (|) that indicates that the interruption is followed by a sequence that does not repeat the interrupted fragment but instead is self-corrected (de Benito Moreno *et al.* 2016: 80).

Although these transcription decisions were taken with the aim to represent the original pronunciation as closely as possible, they create additional challenges for the construction of a morpho-syntactically annotated and parsed corpus, such as the tokenization and lemmatization process. To ease this burden, the COSER also adapted a special transcription rule, the so-called disambiguation convention, whereby phonological reductions are restored to eliminate ambiguous interpretations of the item in question. To exemplify, as in spontaneous speech the deletion of the final /-r/ of the infinitive of verbs of the 1st conjugation class (i.e., verbs ending in *-ar*) can give rise to a form that can also be interpreted as the feminine past particle, the speech is disambiguated between brackets using the equals sign (=), as in *cant*(*á=ar*) 'to sing' or *cant*(*á=ada*) 'sung' (de Benito Moreno *et al.* 2016: 80–81; Fernández-Ordóñez & Pato 2020: 76). As the COSER is still under development, the transcription team continues to transcribe the recorded audio and video material in addition to work on the disambiguations.

3 The Creation of COSER-UD: Project Phases and Tasks

Now that the base corpus and its transcription protocol have been introduced, we will briefly discuss the different phases of the current research infrastructure project. The goal of this project consists in creating a morpho-syntactically anno-

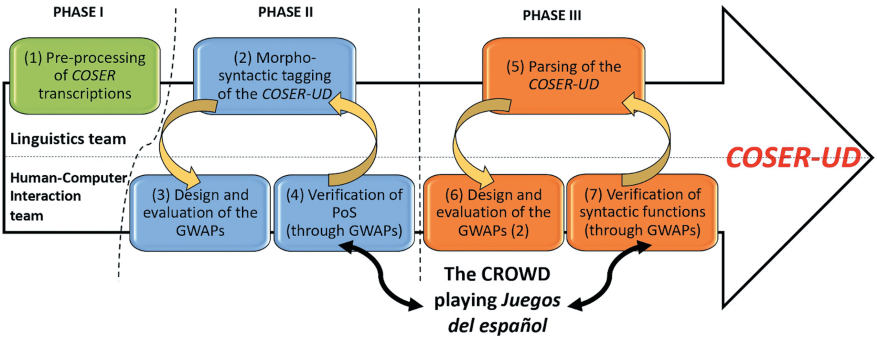

Figure 1: The Creation Process of the COSER-UD.

tated and parsed version of the COSER corpus, originally called COSER-AP (Bonilla *et al.* 2022) and later renamed COSER-UD, due to the Universal Dependencies (UD; Nivre *et al.* 2016, 2020) framework employed for the current research (see section 5.2.1). The resulting COSER-UD, which is presented in a treebank format following the UD guidelines (Bonilla 2022), is the first treebank for oral Spanish.

As illustrated in Figure 1 (adapted from Bonilla *et al.* 2022: 81), we can identify three different project phases, represented by the different colors in the workflow: namely, (i) the COSER pre-processing phase (Phase I), shown in the green rectangle; (ii) the morpho-syntactic annotation stage (Phase II) in light blue; and (iii) the parsing phase (Phase III) in orange. Each task within the different phases is described in the corresponding rectangle, with a numerical sequence indicating the ordering of the tasks. To create the COSER-UD treebank, two specialist teams have been working in parallel, as can be seen by the different levels in the workflow. More specifically, the upper level details the tasks of the Linguistics team, which is responsible for the NLP and dialectology matters of the project, that comprise (1) the pre-processing of the COSER transcriptions, (2) the manual and automatic morpho-syntactic annotation, and (5) the parsing of the corpus, needed to build as an output the COSER-UD treebank (Bonilla 2024a, submitted). Morpho-syntactic annotation or PoS tagging is the procedure whereby a word or a token is assigned a label, which either indicates its grammatical category (e.g., noun, adjective) or its status as a punctuation mark, symbol, or incomplete word. A reference corpus, the so-called Gold Standard (GS) dataset, also termed the COSER-PoS (Bonilla 2024a, 2024b) is created in Phase II. Parsing refers to the process whereby the syntactic function (e.g., subject, direct object, etc.) of a word is identified.

This project also includes an HCI team, given that a collaborative game-based approach has been adopted, whereby the crowd, i.e., (non-expert) members of the public, helps to review the automatic morpho-syntactic annotation and pars-

ing (Bonilla *et al*. 2022; see also Segundo Díaz *et al*. 2023a, 2023b, 2024). As can be seen, the HCI team focuses on the design and evaluation of the various sets of GWAPs (tasks (3) and (6)), that have been specifically developed to verify the automatically annotated PoS and the syntactic functions (tasks (4) and (7)).

These various GWAPs have been collectively referred to as *Juegos del español* (Bouzouita et al. 2022). The collaborative aspect of the project is represented in Figure 1 by the two bidirectional black arrows that point to these tasks and the crowd that plays one of the GWAPs included in *Juegos del español*. In other words, the public confirms or corrects the grammatical categories or functions that have been assigned automatically to the words while playing the GWAPs of *Juegos del español*. Importantly, both teams need to collaborate closely and exchange different types of results, as illustrated by the yellow arrows between the two project teams. For example, the results of the morpho-syntactic annotation carried out by taggers and the manual expert validation of the tags are incorporated into the design of the first series of GWAPs. In turn, the verifications (i.e., both corrections and confirmations) by the players of this automatic annotation can, in theory, serve to retrain the language model to improve the accuracy of the automatic PoS tagging. Likewise, the results produced by the parsers can form the basis for the second series of GWAPs, whose goal is to confirm and correct the syntactic functions that have been generated automatically by the parsers. These crowd-sourced verifications can then, in turn, enhance the accuracy of automatic parsing.

It is important to keep in mind that the workflow of the creation process of the COSER-UD presented in Figure 1 is an abstract representation of the various phases and tasks involved in this project and that various of these tasks are complex ones and thus comprise sub-tasks, this is the case for the tasks of both teams. The pre-processing phase of the COSER, for instance, includes the selection of transcriptions based on geographical distribution criteria, the ridding of these transcriptions of marks typical of the COSER transcription protocol, as well as the sentence extraction from the selected texts (see section 5.1).

Similarly, the morpho-syntactic annotation of the COSER-UD, i.e., task (2) (see also section 5.2), can be subdivided into tasks (2.1) the automatic pre-annotation by the best performing tagger, whose pipeline includes the sentence segmentation, tokenization, lemmatization, PoS tagging, and parsing of the data, (2.2) the manual and semi-automatic tag verification by members of the Linguistics team for the creation of the reference model or GS corpus, COSER-PoS, (2.3) the evaluation of the automatic tagging using the developed reference model, and (2.4), the fine-tuning of the language model of the best performing tagger, which, in theory, can be done on the basis of the confirmative and corrective feedback by the members of the general public once obtained and processed.

Likewise, task (3), which concerns the design and evaluation of the GWAPs of *Juegos del español*, carried out by the HCI team, contains various sub-tasks, such as (3.1) the conceptual design of low-fidelity prototypes of the GWAPs (version 1.0), (3.2) the development of high-fidelity prototypes of the GWAPs, (3.3) the evaluation of GWAPs in terms of Player Enjoyment (PE) and the Game Design Elements (GDEs) integrated in the games to study their influence on the PE, (3.4) the improvement of the GWAPs (version 2.0) based on the results of the previous evaluation, (3.5) the implementation of a mechanism to assess the inter-annotator agreement, which automatically accepts confirmations and corrections of tags provided by the players, and (3.6) the implementation of the crowdsourcing environment, in which the GWAPs are launched to the crowd (for more details, see Segundo Díaz 2024: chapters 5–8 and 10).

Finally, it should be pointed out that although Figure 1 represents the various project stages as conceived originally in the project proposal, not all phases and tasks have been concluded during the funded period (until April 2023). This is due to several reasons, such as lesser obtained funding, which resulted in hiring fewer project members, as well as delays due to the physical and mental consequences of the pandemic outbreak on the project members. Although the Linguistics team was able to evaluate the PoS-taggers and fine-tune the language model for the PoS tagging (Bonilla 2024a: chapter 5, 2024b), this was not done using data resulting from the verified tokens provided by the crowd and the HCI team, but based on expert manual validation. Similarly, as regards the parsing phase, though task (5) has very recently been completed by the Linguistics team (see Bonilla 2024a: chapters 7–9, submitted), no new GWAPs have been designed nor evaluated for the verification of the syntactic functions (tasks (6) and (7)) by the HCI team.

In view of this, not all tasks of the workflow will be discussed. In this contribution we will focus mainly on the tasks carried out by the Linguistics team, such as the pre-processing of the COSER transcriptions (task (1); section 5.1) and the morpho-syntactic annotation tasks (tasks (2) and (5); section 5.2). Though the various created game concepts will be introduced in section 4, we refer to Segundo Díaz *et al.* (2022, 2023a, 2023b, 2024) for more details on the results obtained by the HCI team. For more details on the various GWAPs, see Segundo Díaz *et al.* (2023a, 2024).

4 HCI Team's Tasks: Design and Evaluation of the GWAPs for PoS Verification

While the Linguistics team worked on the pre-processing of the COSER transcriptions (task (1), see Figure 1), the HCI team started designing and evaluating the three GWAPs of *Juegos del español* through which the crowd could help confirm and correct the automatically generated PoS tags (task (3) in Figure 1). As explained in section 3, the design and evaluation of the GWAPs involves various stages, going from creating low-fidelity prototypes of the GWAPs, then high-fidelity ones to the creation of the GWAPs version 2.0 (Segundo Díaz 2024: chapters 5–8). To build engaging games, the HCI team carried out several studies on the GDEs and examined their correlation to PE (see also Segundo Díaz *et al.* 2022). Later, the HCI team also researched the correlations between GDEs, PE and the Personality Traits of the players (PT; for more details, see Segundo Díaz *et al.* 2023b, 2024, Segundo Díaz 2024: chapter 7). Once the Linguistics team completed the automatic tagging process in Phase II (see sections 5.1 to 5.2.2), the relevant data was passed to the HCI team to integrate into the three GWAPs of *Juegos del español*. The HCI team then moved to the fine-tuning, testing, and eventually launching of the GWAPs to the crowd (for further details on the various iterations, see Segundo Díaz 2024: chapters 5–8). In what follows, we will briefly outline the various games included in *Juegos del español*.

Figure 2: The GWAPs in *Juegos del español*: *Agentes*, *Tesoros* and *Anotatlón.*

Three different game concepts have been designed to contrastively examine the effect of various GDEs on PE. As shown in Figure 2 on the left, *Agentes* centers its narrative around the topic of secret agents. This GWAP is a clicker game in which various PoS tags are presented around a sentence with a highlighted word (in this case the determiner *la*) and in which players need to confirm or correct the PoS tag by either clicking on the appropriate one or dragging the word to it.

The second game is called *Tesoros*, in which players need to gather coins and win treasure chests by building a path for an avatar named Gummy, who needs to walk or jump along the constructed path (Figure 2 on the upper right). The path is built every time the player identifies the PoS of a highlighted word.

The third GWAP is a racing game called *Anotatlón*, in which players drive a car to avoid obstacles and reach the finish line. At the finish line, the player must select the appropriate PoS tag for a highlighted word, as illustrated in Figure 2 on the second line.

The three games of *Juegos del español* contain two distinct session types, to wit a training and a playing mode. In the training mode, the highlighted word and its corresponding PoS tag are displayed in the same color to help players become acquainted with the various tags, as shown for *Agentes* and *Tesoros* in Figure 2. Some adjustments to the PoS tags have been introduced though. To illustrate, the *ADP* tag is used in the UD framework for labelling both prepositions and postpositions (see Table 1 in section 5.2.1). Nonetheless, as Spanish does not have the latter, it was decided to present players with a preposition PoS tag (which internally corresponds to the *ADP* tag), given that this is also the denomination used in the Spanish-speaking educational systems.

A mechanism to disambiguate certain tags was also implemented, for example, when a token with a *SCONJ* tag appears in the game, a *CCONJ* and *PRON* tag are also be added (for more details, see Segundo Díaz 2024: chapter 8). Note that the training mode also offers a definition and examples of each PoS to assist players in recognizing the words and their respective PoS tags more effectively (see the examples on the first line of Figure 2).

Besides familiarizing players with the PoS tags and the game mechanics of each GWAP, the training session also serves to establish the confidence score of each player. This score is subsequently used in assessing the inter-annotator agreement, which is needed for automatically accepting confirmations and corrections of the PoS tagging and the extrapolation of the verified data in the morpho-syntactic annotation phase (sections 5.2.3 and 5.2.5).

Once the training session has been completed and the player passes to the playing mode, color cues and PoS definitions are no longer provided, thus challenging players to identify the grammatical categories without help, as shown by the screenshot of *Anotatlón* on the second line in Figure 2, in which no color cue with corre-

sponding definition appears. Finally, note that the games show sentences from the geographic variety that the players have selected when they register to play (Segundo Díaz 2024: chapter 8). This feature was introduced to investigate whether the geographic origin of the players influences the annotation quality, as demonstrated in Bonilla *et al.* (2023).

5 Linguistics Team's Tasks

5.1 Phase I: Pre-Processing of the COSER Transcriptions

As concerns the pre-processing of the COSER transcriptions, firstly, they have been classified into different regional zones, based on the administrative-political division of Spain into autonomous communities or, in some cases, on the grouping of some of these autonomous communities, as has been done for the Principality of Asturias and the Community of Castile and León due to their shared linguistic heritage (cf. Menéndez Pidal 1906; Tuten *et al.* 2016; for recent work on varieties from this region, see d'Andrés Díaz *et al.* 2017). This regional classification is of utmost importance when aiming to construct a geographically balanced reference model, the COSER-PoS, and ultimately the treebank, COSER-UD, that is representative for the European Spanish rural varieties. Secondly, per region between 500 to 600 conversational turns have been randomly extracted, their transcriptions have been altered to remove XML tags and features that are typical of the COSER transcription protocol, such as, for instance, the abbreviations and punctuation marks to indicate overlapping speech, interruptions and self-corrections (see section 2). Concatenations, represented by the apostrophe, have been detached to ensure the success of the subsequent tokenization of each lexical item, a task belonging to the morpho-syntactic annotation phase (see section 5.2). To illustrate, *s'h*a and *t'ha* in example (2) in section 2 have been divided each into three parts, to wit *s* and *t* respectively, the apostrophe, followed by *ha* (Bonilla *et al.* 2022: 79; for further details, see Bonilla 2024a: chapter 5, 2024b).

5.2 Phase II: Morpho-Syntactic Annotation

Now that the pre-processing of the COSER interviews, the first phase of the project, has been introduced, we will discuss the morpho-syntactic annotation process. Before discussing the details of the PoS tagging, it should be pointed out that this project is not the first to morphologically annotate the COSER corpus. Indeed, as

described in de Benito Moreno *et al.* (2016: 81–82), FreeLing (Carreras *et al.* 2004), an open-source tool, has been used for this in an earlier project. In total, around 180h of transcribed material have been lemmatized and annotated using this tool. Unfortunately, no information exists on the accuracy of the tagging of the COSER corpus with FreeLing, which prevents us from comparing results.

5.2.1 The UD Project

In order to select the most accurate PoS tagger, the morpho-syntactic annotation phase of the project started with a study evaluating the tagging accuracy of three different state-of-the-art open-source taggers, which are based on neural network architectures: to wit, spaCy (Honnibal *et al.* 2020), Stanza NLP (Qi *et al.* 2020), and UDPipe (Straka *et al.* 2016). These taggers have been trained on UD treebanks (Nivre *et al.* 2020), created by the open UD community, which aims to create a cross-linguistically consistent treebank annotation system (see the UD website: https://universaldependencies.org/, 15-10-2023).

This immensely successful project currently comprises around 200 treebanks in over 100 languages. Various levels of representation exist within the morpho-syntactic UD annotation scheme: apart from a lemma representing the base form of the word, tokens can get assigned a coarse-grained PoS tag that indicates the word's grammatical category (as in Table 1) and (ii) a more fine-grained label, which describes lexical and grammatical properties associated with the form in question (Nivre *et al.* 2020: 4035; de Marneffe *et al.* 2021).[2] To exemplify, the PoS tag assigned to the Spanish word *mujer* 'woman' is *NOUN*, whereas the fine-grained features (FEATS) common noun, gender, and number specify a more specific subcategory of the PoS *NOUN* and that the word in question is feminine and singular (for a more elaborate example in treebank format, see Segundo Díaz *et al.* 2023a: 139; Bonilla 2024a chapter 5).

Due to the UD project's general aim to achieve cross-linguistic consistency, the annotation system proposed by the UD project is a universal one. As such, there is a list of universal PoS tags which includes 17 different grammatical categories question (Nivre *et al.* 2020: 4036; de Marneffe *et al.* 2021: 261). Note that, although languages are not required to use all categories, this list cannot be ex-

2 The UD project follows traditional grammar by considering words as the primary units, which are interconnected by dependency relations (de Marneffe *et al.* 2021: 259). Observe, however, that this morpho-syntactic notion does not always coincide with the phonological nor orthographic one, as is the case for clitics, which cannot appear without a phonological host, such as those in *t'ha* and *s'ha* in example (2) in section 2.

tended with language-specific PoS tags (cf. section 4 for a discussion on *ADP*). Indeed, as can be seen in Table 1, the morpho-syntactic annotation of Spanish data requires 16 tags out of the full set of 17.[3]

Table 1: Set of Spanish PoS Tags.

PoS Tag	Grammatical Category
ADJ	Adjective
ADP	Adposition
ADV	Adverb
AUX	Auxiliary
CCONJ	Coordinate conjunction
DET	Determinant
INTJ	Interjection
NOUN	Noun
NUM	Number
PRON	Pronoun
PROPN	Proper noun
PUNCT	Punctuation sign
SCONJ	Subordinate conjunction
SYM	Symbol
VERB	Verb
X	Other

While most of the PoS tags in Table 1 are self-explanatory, some are less so. For instance, *X* is used for incomplete words, as well as unanalyzed lexical items from other languages, which, unsurprisingly, can be found more frequently in the bilingual regions of Spain (e.g., Basque Country, Galicia, Catalonia, Balearic Islands, etc.). The PoS tag *AUX* is used in UD more widely than is usual (among linguists and the general public), given that this tag not only encompasses auxiliary verbs, such as *haber* 'to have' in perfect tenses and *ser* and *estar* 'to be' in passive and progressive constructions, but also modal verbs, such as *poder* 'to can' and *deber* 'to must', *soler* 'to tend to', as well as *ser* and *estar* 'to be' that are used as copulas (see also Bonilla *et al.* 2022: 89–90; Bonilla 2024a: chapter 5). All instances of *ser* (and *soler*) are always classified as auxiliary verbs, irrespective of

3 The COSER-UD only uses 16 of these tags, unlike the UD Spanish AnCora treebank, which also uses the particle tag (*PART*), as for instance for tagging *no* 'no(t)' in *no obstante* 'notwithstanding'. In COSER-UD, *no* is classified as an *ADV* when modifying a verb and as an *INTJ* when appearing alone. *No obstante* is in COSER-UD classified as a Multi-Word Expression (MWE), which is tagged as a *CCONJ* (https://universaldependencies.org/treebanks/es_ancora/es_ancora-pos-PART.html, 15-03-2024; Bonilla 2022).

their actual status as a verb that supports another one, as for instance in *soy profesora* 'I am a teacher'. As de Marneffe *et al.* (2021: 273) indicate, the distinction between copula and auxiliary verbs is restored at the parsing level, given that auxiliary verbs are treated as dependents of the main verb through the *aux* relation, whereas copula as dependents of a non-verbal predicate, such as an adjective or a noun, through the *cop* relation. The variants of existential *haber*, in contrast to the use of *haber* in perfect tenses, are classified as *VERB*.

5.2.2 Creation of the COSER-PoS: Automatic Tagging, Data Review and Creation of a Reference Model

As concerns Spanish treebanks, the most widely used one is AnCora_Es (ANnotated CORporA; Taulé *et al.* 2008), which has been developed using Spanish newspapers articles (from the Spanish EFE news agency and the newspaper *El Periódico*) and material from the *Léxico Informatizado del Español* corpus (LexEsp, "Computerized Lexicon of Spanish", Sebastián Gallés *et al.* 2000).

The LexEsp corpus is a balanced corpus of 6 million words which includes various literary genres, news articles, scientific texts, etc., all written in European Spanish. For the coarse-grained tagging, an accuracy rate above 98% has been reported for the taggers trained with the AnCora_Es treebank (e.g., https://spacy.io/models/es and https://stanfordnlp.github.io/stanza/performance.html). Nevertheless, their performance with data from other types of language varieties, such as spoken speech, has not been evaluated yet. It is expected though that the accuracy rate with spoken speech will be lower, considering that these taggers have been trained using written language, predominantly from the journalistic domain.

One of the challenges is indeed that there is insufficient labelled and publicly accessible data from other domains. More representative language models are thus needed. In view of this, we aim in this project to fill this gap by evaluating the taggers through the construction of a reference corpus, COSER-PoS, which is a 200.000-word sub-corpus of European rural spoken Spanish, based on a geographically balanced sample from the COSER corpus (Bonilla 2022, 2024a: chapter 5, 2024b). The creation of this reference corpus, the GS, will allow us to measure the accuracy of the current taggers, trained on written varieties close(r) to the standard variety, when dealing with spoken data, as well as to calculate the players' confidence and resulting inter-annotator agreement scores in the training mode of the GWAPs (see section 4).

The evaluation of these taggers can in turn help determine which features are at the basis for the flaws of the current models when tagging non-standard oral speech. For the creation of the COSER-PoS, three taggers based on neural network

architectures and trained with AnCora_Es (Taulé *et al.* 2008), have been selected: namely, spaCy (Honnibal *et al.* 2020), Stanza NLP (Qi *et al.* 2020), and UDPipe (Straka *et al.* 2016). A geographically balanced sample of around 200.000 tokens has been selected for the construction of the COSER-PoS (see section 5.1), given that the (transcribed part of the) COSER corpus contains more than 4 million words, and that the manual revision of the morpho-syntactic annotation would be too labor-intensive. To ensure that this reference model can be reused and replicated, the annotation criteria and labels from the UD project have been followed (section 5.2.1). This reference corpus has then been used for the accuracy evaluation tasks of the tagging and for the subsequent retraining of the taggers (see section 5.2.3).

Once the geographically balanced sample has been selected, adequately delimited and the changes in the transcriptions discussed in section 5.1 have been implemented (task (1) in Figure 1), various automatic procedures have been carried out using the spaCy NLP library, to wit, sentence segmentation, tokenization, lemmatization, and morpho-syntactic tagging, as will be discussed now.

As regards sentence segmentation, during this process the conversational turns extracted during the pre-processing of the sample (see 5.1) have been separated into sentences, which have then been given a unique identifier, composed of the first four letters of the name of the region and an incremental integer (e.g., extr=Extremadura). The use of this type of identifier makes geographical classification (and thus searching) of the data possible. Note that, even though a similar number of conversational turns has been included per region, the total number of sentences can still vary due to differences in the length of the conversational turns (see Bonilla *et al.* 2022: 85 for specific details per region; Bonilla 2024a: chapter 5, table 5.2).

Once the different sentences have been separated, the words (or tokens) contained in these sentences have been extracted through tokenization. Subsequently, all tokens have been lemmatized, a procedure whereby the inflectional complexity of words is reduced to a common base form, i.e., the lemma, and then morpho-syntactically tagged. The results of all this have been adapted to the CoNLL-X format (Buchholz & Marsi 2006; Computational Natural Language Learning), which is an UD adaptation in which CoNLL-U format properties are assigned to a document, its sentences, and tokens, using the spaCy_conll library (version 3.0, Vanroy 2021).

The spaCy library offers three Spanish models that have been created using convolutional neural networks and which vary in size (small *sm*, medium *md*, and large *lg*) and one model that uses Transformer (*trf*) architecture, in this case the Spanish version of BERT (Cañete *et al.* 2020). According to data published on spaCy's website, the models *sm*, *md*, and *lg* (version 3.0) have an PoS tagging accuracy of 0.98, while *trf* achieves 0.99. All these models have used the AnCora_Es treebank for their training. In this project, we used the large model (es_core_-

news_lg) given that the *trf* model had not been released yet when the Linguistics team started the morpho-syntactic tagging task.

After having transformed the geographically balanced COSER sample into the CONLL-U format, the lemmas, coarse-grained PoS tags, and more fine-grained FEATS have been validated both manually and semi-automatically (using regular expressions). This corrected dataset then served to establish the COSER-PoS, a reference corpus or GS, which is freely available for consultation on GitHub (Bonilla 2022).[4]

5.2.3 Automatic Taggers' Accuracy Evaluation

Once the data has been reviewed and the COSER-PoS has been created as a GS, the accuracy evaluation of the various taggers took place. For this sub-task, each of the sentences has been tokenized and tagged using the various versions of the spaCy tagger (*sm*, *md*, *lg*, and *trf*), Stanza NLP, and UDPipe, which all use neural network architecture and have been trained with the AnCora_Es corpus. The results of these processes have then been verified with those of the corrected reference corpus to assess the accuracy rate of each tagger. For the accuracy calculations, the *scikit-learn* library has been used to evaluate the models in terms of precision, recall, F1-score, and accuracy (Pedregosa *et al*. 2011). These reference statistics are important to determine whether the domain adaptation of these taggers, trained on written standard language, to oral-dialectal data, as found in the COSER corpus, improves the tagging accuracy. The comparison of the different accuracy rates of the various taggers reveals that the differences are not significant. Nonetheless, the spaCy's *trf* model outperforms the others with an accuracy rate of 0.927, while the other models' rates range between 0.90 (UDPipe) to 0.920 (Stanza NLP), with the *sm*, *md* and *lg* models of spaCy occupying an intermediate position with an 0.913 accuracy rate (Bonilla *et al*. 2022).

Interestingly, minor differences in the tagger performance can be observed depending on the geographic origin of the text that is tagged, as observed by Bonilla *et al*. (2022: 87). For instance, UDPipe obtained the least accuracy rate for Andalusian and Murcian Spanish (0.89) and the best for Balearic Spanish (0.92). Though the spaCy *trf* model achieved consistently higher results than the other models, for some varieties, such as Castilian, Basque Country and Aragonese

4 Note that the size of the COSER-PoS has evolved over time due to various data cleaning phases (cf. Bonilla *et al*. 2022: 13.402 sentences, 204.899 tokens vs. Bonilla 2024a: chapter 5, 13.219 sentences, 196.372 tokens).

Spanish, Stanza NLP performed equally well. Indeed, the difference between these two models is never greater than 1% for the various Spanish regions, whereas the maximum difference between UDPipe and spaCy *trf* can reach 3%, as is the case for Andalusian Spanish. Given the small differences between the geographical areas, it is not possible to draw conclusions on which dialectal zones are closer or more removed from the journalistic language on which the various models were trained.

As the spaCy *trf* model consistently achieved the highest accuracy rates, the next part of the analysis only used this Transformer model, whereby the performance of this tagger is reviewed for each PoS (see also Table 2 in section 5.2.4). Summarizing the most important observations made by Bonilla *et al*. (2022: 88–89), the lowest F1-scores are found in the tagging of incomplete words (*X*: 0) and interjections (*INTJ*: 0.53). These findings are in line with those of Moreno-Sandoval and Guirao (2006: 201–206) for the C-ORAL-ROM corpus. In this project, however, there is no need to increase the size of the lexicon nor to implement a grammar to disambiguate certain linguistic aspects, as machine learning techniques were used for the training of the tagger, whereby it learns from the data input without having to rely on predefined lexicon or rules.

The highest F1-scores, in contrast, are obtained for numbers (*NUM*: 0.96), punctuation signs (*PUNCT*), coordinate conjunctions (*CCONJ*) and prepositions (*ADP*), the latter three of which received an F1-score of 0.99. These results are expected given that, on the one hand, incomplete words and interjections are typical of colloquial speech and, as such, are not found in written journalistic genres, and thus unknown features for the tagger, and that, on the other, the categories with the highest F1-scores are all invariable ones, and thus do not present a challenge for the tagger. Note further that the low F1-score for the interjections can be in part attributed to the UD guidelines (Bonilla 2024a: 76, 2024b), given that words used in exclamations obtain the original PoS. Consequently, *Dios* 'God' in *Dios mío* 'my God' will be classified as *NOUN*, while the whole construct is also tagged as a Multi-Word Expression (MWE).

Conversely, grammatical categories that exhibit morphological variation associated with oral and/or dialectal idiosyncrasies, such as adjectives (*ADJ*: 0.84), verbs (*VERB*: 0.91), and auxiliaries (*AUX*: 0.75), present the tagger with more difficulties and thus obtain lower F1-scores than the invariable ones. For instance, adjectives with diminutives, deverbal adjectives that coincide with past particles (e.g., *los calderos todos ahumaos* 'lit. the cauldrons all smoked'), and verbs that reflect dialectal pronunciation and are transcribed differently from the standard form (e.g., *trabajábanos* = *trabajábamos* 'we worked') are real hurdles for the tagger as they present unknown morphological characteristics.

5.2.4 Human Annotators' Accuracy Evaluation

Apart from verifying the accuracy of the automatic tagging process, the accuracy rate of the human annotators, i.e., the players of the GWAPs included in *Juegos del español*, was also examined. As already mentioned in section 4, the players' performance in the training mode is used to assign a confidence score for each participant, whereby the accuracy of PoS tagging is compared with the GS. This score serves to calculate the inter-annotator agreement to automatically accept PoS verifications. For a specific token to be assigned an inter-annotator agreement score, at least three players need to have verified the token in question and the score for the tag should have a coefficient of at least 0.75 (Bonilla *et al.* 2023). Once a token receives an inter-annotation agreement score, annotation stops. In what follows, we will report on a study that examines the accuracy of human annotators.

The overall human accuracy rate is 0.80 in the study reported on in Bonilla *et al.* (2023), in which 121 participants took part, who managed to verify 5.976 tokens. This is considerably lower than the accuracy rates of the automatic taggers (spaCy's *trf* model: 0.927; Stanza NLP: 0.920; spaCy *sm*, *md* and *lg* models: 0.913; UDPipe: 0.90; see section 5.2.3; Bonilla *et al.* 2022). However, this difference in general accuracy rate should not lead to an abandonment of the citizen science approach implemented in this research infrastructure project. On the contrary, the comparative analysis of the F1-scores per tagged PoS, shown in Table 2,[5] in which the bold items indicate the highest F1-score per PoS, reveals that for certain grammatical categories the human tagging accuracy score is higher than for the automatic tagger spaCy's *trf*, thus indisputably demonstrating that human input is not obsolete and, more generally, that citizen science projects can positively contribute to advancing knowledge and technological progress.[6]

Indeed, humans outperform the Transformer tagger for the interjections (*INTJ*: 0.86 vs 0.53) and proper nouns (*PROPN*: 0.93 vs 0.69). These results are expected given that interjections tend to be absent in certain genres, such as journalist texts, on which the automatic taggers were trained (see section 5.2.2). Similarly, humans tend to be better at inferencing that a given token in a certain linguistic context is a proper noun, such as a name or place name, without necessity of previously having

5 For additional details, such as the precision and recall scores for the spaCy *trf* model and for the human annotators, we refer the interested reader to Bonilla *et al.* (2022) and Bonilla *et al.* (2023), respectively. For a study on the variables that significantly influence the human annotators' tagging accuracy, such as educational level, the field of study, and geographic upbringing, see Bonilla *et al.* (2023).

6 This said, we do not claim that taggers cannot be trained to achieve as high accuracy rates as human annotators or even higher. Though, human input will be needed for this task too.

Table 2: Comparison of Human and Automatic Tagging Accuracy (F1-scores per PoS).

PoS	F1-scores	
	Human annotators	**SpaCy *trf* model**
ADJ	0.76	**0.84**
ADP	0.83	**0.99**
ADV	0.86	**0.91**
AUX	0.61	**0.75**
CCONJ	0.83	**0.99**
DET	0.84	**0.94**
INTJ	**0.86**	0.53
NOUN	**0.94**	**0.94**
NUM	0.89	**0.96**
PRON	0.69	**0.91**
PROPN	**0.93**	0.69
PUNCT	/	0.99
SCONJ	0.46	**0.87**
SYM	/	0.91
VERB	0.69	**0.91**
X	**0.79**	0

heard or being familiar with the proper noun in question. The F1-score of the automatic tagger for incomplete words (*X*) is 0 given that this PoS is absent in written language models, whereas human annotators are completely used to it due to its high frequency in oral speech. As such, humans performed significantly better (0.79).

As can be seen, the differences between the human and automatic tagging accuracy scores are for some PoS very great. For the PoS *NOUN*, in contrast, the F1-scores for the tagging accuracy are the same for the *Juegos del español* participants and the spaCy *trf* model: to wit, 0.94, which is the highest F1-score for human tagging when comparing all the PoS scores. The lowest human F1-score is obtained by the *SCONJ* (0.46), which was in almost half of the cases confused with its coordinate counterpart, *CCONJ*, probably due insufficient technical linguistic knowledge by the participants (Bonilla *et al.* 2023). Equally, the *AUX* category received a low human F1-score (0.61), which is quite likely due to the classification used in the UD guidelines (see section 5.2.1), which diverges from the one taught in schools and appears to lead to confusion among the players. Concretely, *ser* is always regarded as an *AUX* (e.g., in its copular and passive function) regardless of its status as a supporting verb to another one, whereas *estar* 'to be' receives the *AUX* tag when being a copula (e.g., *Lili está enferma* 'Lili is ill') and part of the progressive construction (e.g., *Johnatan está estudiando* 'Johnatan is studying'), but can also function as a

VERB, as in *Miriam está en el despacho* 'Miriam is in the office' (Bonilla 2022). However, the same cannot be said verbal periphrases, whereby *voy* in *voy a cantar* 'I am going to sing' [*ir* 'to go'+ preposition *a* + infinitive] and in *voy cantando* 'I'm going (while) singing' [*ir* 'to go' + gerund], or *ando* in *ando cantando* 'I walk around singing' [*andar* 'to walk' + gerund] are regarded as a *VERB*, despite the clear parallelism between these cases with the *estar*-constructions with a gerund.

5.2.5 Post-Tagging Knowledge Transfer

As we will see in this section, it is not necessary to verify the PoS tag for each token in the corpus given that the knowledge gained from the players of *Juegos del español*, who confirmed or corrected PoS tags, can be extrapolated to unverified PoS tags on condition that certain requirements are met. To exemplify, the preposition *de* 'of/from' can only ever be a preposition (*ADP* in UD). In view of this, the human annotators' confirmation of the accuracy of this PoS tag can be extrapolated to all 4.699 cases in COSER-UD, thus tremendously upscaling the crowd-sourced input. In contrast, the PoS tag verifications of *bueno*, which can function either as an *ADJ* 'good' or as an *INTJ* 'well', but which both have the same lemma, namely *bueno*, cannot be extrapolated because of its lemma's inherent polyfunctionality. Notwithstanding this, homographs that have been assigned different lemmas can be extrapolated, as is the case for instance for *la* which can be the feminine article 'the' (*DET*) but also a feminine object pronoun 'her' (*PRON*), the latter of which has been attributed the lemma *él* 'he', while the former *el* 'the'.

The knowledge transfer of verified PoS tags has been implemented in a semi-automated manner. Initially, all text underwent conversion to lowercase to ensure uniformity across the dataset, crucial for consistent text analysis, while eliminating duplicate entries. Next, PoS corrections suggested by human annotators have been reviewed by experts, who adjusted the PoS tags when needed. For instance, all cases of *ser* 'to be' and *ir* 'to go' are considered *AUX* and *VERB* respectively in UD (see also section 5.2.4). Additionally, erroneous corrections, such *umbilical* 'umbilical' or *última* 'last', which were tagged as *NOUN* instead of *ADJ*, have been rectified. Furthermore, the lemmas of the tokens have been automatically extracted and then manually confirmed. This step is needed as the original database included only IDs, which complicates the large-scale knowledge transfer of verified tags. The final stage of the post-tagging knowledge transfer consists in an extrapolation process of the verified PoS tags. First, the set [token + lemma + PoS tag] of the verified data is matched with their counterparts in the automatically tagged data of the COSER-UD treebank. As they coincide, the latter cases can thus be regarded as veri-

fied by extension. Table 3 provides the number of matches of this extrapolation operation per PoS tag, counting MWEs as single entities.

Table 3: Post-Tagging Knowledge Transfer per PoS tag.

PoS tag	Extrapolated cases
ADJ	180 (0.25%)
ADP	15.499 (21.44%)
ADV	8.812 (12.19%)
AUX	1.465 (2.03%)
CCONJ	10.329 (14.29%)
DET	14.413 (19.93%)
INTJ	1.052 (1.45%)
NOUN	4.657 (6.44%)
NUM	633 (0.88%)
PRON	8.806 (12.18%)
PROPN	178 (0.25%)
SCONJ	4.120 (5.7%)
SYM	0
VERB	2.139 (2.96%)
X	21 (0.03%)
Total	**72.304**

Leaving punctuation signs aside, the knowledge transfer of the 467 verified tokens by players of *Juegos del español* yielded the confirmation of a total of 72.304 tokens. In other words, while human annotators confirmed or corrected a mere 0.29% of COSER-UD's tokens, the knowledge transfer upscaled these results to 45.1% (72.304/160.321) of the treebank, which is not an unsignificant feat.[7] However, lemma-wise the results of this extrapolation process are a lot less impressive as it only affects 3.34% (189/5.662) of all lemmas.

As can be seen in Table 3, the extrapolation of prepositions (*ADP*: 21.44%) is responsible for more than a fifth of all cases, closely followed by the determiners (*DET*: 19.93%), and then the coordinate conjunctions (*CCONJ*: 14.29%), adverbs (*ADV*: 12.19%) and pronouns (*PRON*: 12.18%). These results indicate that, though maintaining players' interest in GWAPs is not an easy task (Chamberlain *et al.*

7 In summary, 147 players confirmed and/or corrected 8.215 PoS annotations, which resulted in the full verification of 467 tokens (using the inter-annotator agreement score; see section 5.2.4), while 3.428 tokens have been verified partially (up until the 23rd of February 2024). It should not be forgotten that these PoS annotations only include those obtained in the playing mode (see section 4). In the training session, players provided 10.576 annotations.

2013; Poesio *et al*. 2017), upscaling verification results by knowledge transfer can advance the PoS verification process considerably, leaving more time for the human annotators to handle the polyfunctional, more difficult cases.

6 Conclusions

In summary, this contribution outlines an interdisciplinary research infrastructure project integrating the fields of Dialectology, NLP, and HCI aimed at developing a morpho-syntactically annotated and parsed corpus of the diatopic varieties of European Spanish, known as COSER-UD. Employing a citizen science approach, this project engages the public through various GWAPs, collectively termed as *Juegos del español*, to verify the automatic PoS tags. In addition to delineating the three games comprising *Juegos del español*, this discussion detailed the tasks undertaken by the Linguistics team during the initial two phases of the project, which encompass the transcription pre-processing and morpho-syntactic annotation of the COSER. Regarding the latter, we elaborated on the creation of COSER-PoS as a reference model and the automated tagging process. Subsequently, we provided a comparative analysis of tagging accuracy rates between the spaCy Transformer model and human annotators. While human annotators generally exhibit lower tagging accuracy scores, they notably surpass the Transformer model for specific PoS categories, such as interjections (*INTJ*), proper names (*PROPN*), and incomplete words (*X*). This underscores the effectiveness of the collaborative approach employed in this corpus creation endeavor, wherein human expertise and NLP algorithms complement each other. The final phase of morpho-syntactic annotation involves the post-tagging transfer of verified data, resulting in a significant expansion of the verification rate (from 0.29% to 45.1% of the treebank). Prepositions (*ADP*) and determiners (*DET*) emerge as the primary PoS categories driving this extrapolative procedure, collectively constituting more than 40% of the instances therein.

Regarding future tasks pertaining to the morpho-syntactic annotation of the COSER-UD, approximately half of the PoS tags of the treebank are yet to be verified. Given the difficulty of sustaining player engagement in gaming, alternative data verification methods are currently under exploration, as well as variations on the post-tagging knowledge transfer.

Bibliography

Bonilla, Johnatan E. (2022): COSER-UD <https://github.com/johnatanebonilla/UD_Spanish-COSER> (21-10-2023).

Bonilla, Johnatan E. (2024a): *Universal Dependencies for Spoken Spanish*. Doctoral Dissertation. Ghent University/Humboldt University of Berlin.

Bonilla, Johnatan E. (2024b): "Spoken Spanish PoS Tagging: Gold Standard Dataset", in Language Resources and Evaluation. <https://doi.org/10.1007/s10579-024-09751-x>.

Bonilla, Johnatan E. (submitted): "Development of the First Spoken Spanish Treebank within the Universal Dependencies Framework".

Bonilla, Johnatan E., Miriam Bouzouita and Rosa Lilia Segundo Díaz (2022): "La construcción del *Corpus Oral y Sonoro del Español Rural – Anotado y Parseado* (COSER-AP): avances en el etiquetado de partes del discurso", in *Revista Internacional de Lingüística Iberoamericana*, 22(40), pp. 77–96.

Bonilla, Johnatan E., Rosa Lilia Segundo Díaz and Miriam Bouzouita (2023): "Using GWAPs for Verifying PoS Tagging of Spoken Dialectal Spanish", in *10th International Conference on Behavioural and Social Computing (BESC)*. Institute of Electrical and Electronics Engineers, pp. 1–7. <https://doi.org/10.1109/BESC59560.2023.10386542>.

Breitbarth, Anne, Melissa Farasyn, Anne-Sophie Ghysele and Jacques Van Keymeulen (2020): "Het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten (GCND)", in *Handelingen* (KZM), LXXII, pp. 23–38. <https://doi.org/10.21825/kzm.v72i0.17914>.

Buchholz, Sabine and Erwin Marsi (2006): "CoNLL-X Shared Task on Multilingual Dependency Parsing", in Lluís Màrquez and Dan Klein (eds). *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York City: Association for Computational Linguistics, pp. 149–164. <https://aclanthology.org/W06-2920.pdf> (21-10-2023) .

Campillos Llanos, Leonardo (2016): "PoS-Tagging a Spanish Oral Learner Corpus", in Margarita Alonso-Ramos (ed.). *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins, pp. 78–89.

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang and Jorge Pérez (2020): "Spanish Pre-trained BERT Model and Evaluation Data", in *Practical Machine Learning for Developing Countries at the International Conference on Learning Representations 2020*, pp. 1–10.

Carreras, Xavier, Isaac Chao, Lluís Padró and Muntsa Padró (2004): "FreeLing: An Open-Source Suite of Language Analyzers", in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.). *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon: European Language Resources Association (ELRA), pp. 239–242.

Castillo, Lorena, M.ª Pilar Colomina and Irene Fernández (2020): "El *Atlas Sintáctico del Español* (ASinEs): una herramienta para codificar la variación", in Ángel J. Gallego and Francesc Roca Urgell (eds.). *Dialectología digital del español*. Santiago de Compostela: Verba: Anuario Galego de Filoloxía, Anexo 80, pp. 47–69. <https://dx.doi.org/10.15304/9788418445316>.

Castillo Lluch, Mónica, Cristina Peña Ruedo and Michiel de Vaan (2022): "¿'Pronunciar' o 'pronunciare'? Esa es la cuestione", in Ana Estrada, Beatriz Martín and Carlota de Benito (eds.). *Como dicen en mi pueblo: el habla de los pueblos españoles*. Madrid: Pie de Página, pp. 63–75.

Chamberlain, Jon, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade and Massimo Poesio (2013): "Using Games to Create Language Resources: Successes and Limitations of the Approach", in Iryna

Gurevych and Jungi Kim (eds.). *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Berlin, Heidelberg: Springer, pp. 3–44. <https://doi.org/10.1007/978-3-642-35085-6_1>.

Chambers, Jack and Perter Trudgill (1998): *Dialectology*. Cambridge: Cambridge University Press.

COSER = Fernández-Ordóñez, Inés (dir.) (2005-present): "Corpus Oral y Sonoro del Español Rural". <www.corpusrural.es> (11-08-2024).

d'Andrés Díaz, Ramón, Fernando Álvarez-Balbuena García, Xulio Miguel Suárez Fernández, and Miguel Rodríguez Monteavaro (2017): *Estudiu de la transición llingüística na zona Eo-Navia, Asturies (ETLEN). Atles llingüísticu dialectográficu - horiométricu - dialectométricu*. Trabe: Universidá d'Uviéu.

de Benito Moreno, Carlota, Javier Pueyo and Inés Fernández-Ordóñez (2016): "Creating and Designing a Corpus of Rural Spanish", in Stefanie Dipper, Friedrich Neubarth and Heike Zinsmeister (eds.). *Proceedings of the 13th Conference on Natural Language Processing KONVENS 2016*. Bochum: Bochumer Linguistische Arbeitsberichte, pp. 78–83.

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

Esplà-Gomis, Miquel and Andreu Sentí (in preparation): "An Annotated Informal and Spoken Corpus for Dialectal Conversations: The Parlars Corpus for Valencian Catalan".

Farasyn, Melissa, Anne-Sophie Ghyselen, Jacques Van Keymeulen and Anne Breitbarth (2022): "Challenges in Tagging and Parsing Spoken Dialects of Dutch", in *Journal of Historical Syntax*, 6, pp. 4–11. <https://doi.org/10.18148/hs/2022.v6i4-11.92>.

Fernández-Ordóñez, Inés and Enrique Pato (2020): "El *Corpus Oral y Sonoro del Español Rural* (COSER) y su contribución al estudio de la variación gramatical del español", in Ángel J. Gallego and Francesc Roca Urgell (eds.). *Dialectología digital del español*. Santiago de Compostela: Verba: Anuario Galego de Filoloxía, Anexo 80, pp. 71–100. <https://dx.doi.org/10.15304/9788418445316>.

Gelbukh, Alexander, Sulema Torres and Hiram Calvo (2005): "Transforming a Constituency Treebank into a Dependency Treebank", in *Procesamiento del Lenguaje Natural*, 35, pp. 145–152.

Ghyselen, Anne-Sophie, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen and Arian van Hessen (2020): "Clearing the Transcription Hurdle in Dialect Corpus Building: The Corpus of Southern Dutch Dialects as Case Study", in *Frontiers in Artificial Intelligence*, 3. <https://doi.org/10.3389/frai.2020.00010>.

Honnibal, Matthew, Ines Montani, Sofie van Landeghem and Adriane Boyd (2020): "spaCy: Industrial-strength Natural Language Processing in Python". <https://github.com/explosion/spaCy> (11-08-2024).

Juegos del español = Bouzouita, Miriam, Johnatan E. Bonilla, Rosa Lilia Segundo Díaz, Véronique Hoste, Karin Coninx and Gustavo Rovelo Ruiz (2022): "Project *Juegos del español*". <www.juegosdelespanol.com> (11-08-2024).

Jørgensen, Annette Myre and Esperanza Eguía Padilla (2015): "Presentación de COLA, un corpus oral de lenguaje adolescente en línea", in Sigrún A. Eriksdottir (ed.). *Actes du XIXème Congrès des romanistes scandinaves*. Reykjavik: Institute of Foreign Languages. <https://conference.hi.is/rom14/rom-lectures/> (11-08-2024).

Jørgensen, Annette Myre, Esperanza Eguía Padilla, Anna-Brita Stenstrom, Juan Antonio Martínez López, Eli Marie Drange Danbolt, Mariano Reyes Tejedor, Anna Acevedo, Giovanna Angela Mura, Stine Huseby, Lise Holmvik, Solfrid Hernes, Evert Jakobsen, Kristine Eide and Marie Espeland (2004–2017): "Proyecto COLA. Corpus Oral de Lenguaje Adolescente". <https://blogg.hiof.no/colam-esp/> (21-09-2021).

Martínez Alonso, Héctor and Daniel Zeman (2016): “Universal Dependencies for the AnCora Treebanks”, in *Procesamiento del Lenguaje Natural*, 57, pp. 91–98.

Menéndez Pidal, Ramón (1906): “El dialecto asturleonés”, in *Revista de Archivos, Bibliotecas y Museos*, 2–3, pp. 128–172 and pp. 294–311.

Montserrat, Sandra and Carles Segura (2020): “Un corpus col·loquial i dialectal del valencià: PARLARS”, in *Zeitschrift für Katalanistik*, 33, pp. 9–44. <https://doi.org/10.46586/ZfK.2020.9-44>.

Moreno Fernández, Francisco (2005): “Corpus para el estudio del español en su variación geográfica y social: el corpus PRESEEA”, in *Oralia: Análisis del discurso oral*, 8, pp. 123–140.

Moreno-Sandoval, Antonio, Guillermo de la Madrid, Manuel Alcántara, Ana González, José M. Guirao and Raúl de la Torre (2005): “The Spanish Corpus”, in Emanuela Cresti and Massimo Moneglia (eds.). *C-ORAL-ROM: Integrated Reference Corpus for Spoken Romance Languages*. Amsterdam: John Benjamins, pp. 135–161.

Moreno-Sandoval, Antonio and José M. Guirao (2006): “Morphosyntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation”, in Yuji Kawaguchi, Susumu Zaima and Toshihiro Takagaki (eds.). *Spoken Language Corpus and Linguistic Informatics*. Amsterdam: John Benjamins, pp. 199–218.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty and Daniel Zeman (2016): “Universal Dependencies v1: A Multilingual Treebank Collection”, in Nicoletta Calzolari *et al.* (eds.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož: European Language Resources Association (ELRA), pp. 1659–1666.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers and Daniel Zeman (2020): “Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection”, in Nicoletta Calzolari *et al.* (eds.). *Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association (ELRA)*. Marseille: European Language Resources Association, pp. 4034–4043.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard Duchesnay (2011): “Scikit-learn: Machine Learning in Python”, in *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Poesio, Massimo, Jon Chamberlain and Udo Kruschwitz (2017): “Crowdsourcing”, in Nancy Ide and James Pustejovsky (eds.). *Handbook of Linguistic Annotation*. Dordrecht: Springer. <https://doi.org/10.1007/978-94-024-0881-2_10>.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher Manning (2020): “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages”. <https://arxiv.org/abs/2003.07082> (11-08-2024).

Radford, Alec, Jing Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey and Ilya Sutskever (2023): “Robust Speech Recognition via Large-Scale Weak Supervision”, in *Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 202, pp. 28492–28518. <https://proceedings.mlr.press/v202/radford23a/radford23a.pdf> (11-08-2024).

Rico-Sulayes, Antonio, Rafael Saldívar-Arreola and Álvaro Rábago-Tánori (2017): “Part-of-Speech Tagging with Maximum Entropy and Distributional Similarity Features in a Subregional Corpus of Spanish”, in *Ingeniería y Competitividad*, 19(2), pp. 55–67.

Sebastián Gallés, Núria, M. Antònia Martí Antonín, Manuel Francisco Carreiras Valiña and Fernando Cuetos Vega (2000): *LEXESP: Léxico informatizado del español*. Barcelona: Edicions Universitat de Barcelona.

Segundo Díaz, Rosa Lilia (2024): *Juegos del español – Iterative Design, Evaluation and Implementation of Games with a Purpose to Enhance Parts-of-Speech Tagging in a Corpus of European Spanish Dialects*. Doctoral Dissertation. Hasselt University/Ghent University.

Segundo Díaz, Rosa Lilia, Gustavo Rovelo, Miriam Bouzouita and Karin Coninx (2022): "Building Blocks for Creating Enjoyable Games – A Systematic Literature Review", in *International Journal of Human – Computer Studies*, 159. <https://doi.org/10.1016/j.ijhcs.2021.102758>.

Segundo Díaz, Rosa Lilia, Johnatan E. Bonilla, Miriam Bouzouita and Gustavo Rovelo Ruiz (2023a): "Juegos con propósito para la anotación del *Corpus Oral Sonoro del Español Rural*", in *Dialectologia et Geolinguistica*, 31, pp. 135–164. <https://doi.org/10.1515/dialect-2023-0007>.

Segundo Díaz, Rosa Lilia, Gustavo Rovelo, Miriam Bouzouita, Véronique Hoste and Karin Coninx (2023b): "The Influence of Personality Traits and Game Design Elements on Player Enjoyment: A Demo on GWAPs for Part-of-Speech Tagging", in Mads Haahr, Alberto Rojas-Salazar and Stefan Göbel (eds.). *Serious Games, 9th Joint International Conference, JCSG 2023, Lecture Notes in Computer Science (LNCS, volume 14309)*. Cham: Springer, pp. 353–361. <https://doi.org/10.1007/978-3-031-44751-8_28>.

Segundo Díaz, Rosa Lilia, Gustavo Rovelo, Miriam Bouzouita, Véroniqe Hoste and Karin Coninx (2024): "The Influence of Personality Traits and Game Design Elements on Player Enjoyment: An Empirical Study in GWAP for Linguistics", in Pierpaolo Dondio *et al.* (eds.). *Games and Learning Alliance, 12th International Conference, GALA 2023, November 29-December 1, 2023, Lecture Notes in Computer Science (LNCS, volume 14475)*. Cham: Springer, pp. 1–10. <https://doi.org/10.1007/978-3-031-49065-1_20>.

Straka, Milan, Jan Hajič and Jana Straková (2016): "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing", in Nicoletta Calzolari *et al.* (eds.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association (ELRA), pp. 4290–4297.

Taulé, Mariona, M. Antònia Martí and Marta Recasens (2008): "AnCora: Multilevel Annotated Corpora for Catalan and Spanish", in: Nicoletta Calzolari *et al.* (eds.). *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech: European Language Resources Association (ELRA), pp. 96–101.

Taulé, Mariona, M Antònia Martí, Ann Bies, Montserrat Nofre, Aina Garí, Zhiyi Song, Stephani Strassel and Joe Ellis (2015): "Spanish Treebank Annotation of Informal Non-standard Web Text", in Florian Daniel and Oscar Diaz (eds.). *Current Trends in Web Engineering, ICWE 2015, Lecture Notes in Computer Science*. Cham: Springer, pp. 15–27. <https://doi.org/10.1007/978-3-319-24800-4_2>.

Tuten, Donald N., Enrique Pato and Ora R. Schwarzwald (2016): "Spanish, Astur-Leonese, Navarro-Aragonese, Judaeo-Spanish", in Adam Ledgeway and Martin Maiden (eds.). *The Oxford Guide to the Romance Languages*. Oxford: Oxford University Press, pp. 382–410.

Vanroy, Bram (2021): spaCy_connll 3.0. <https://github.com/BramVanroy/spaCy_conll> (11-08-2024).