Cómo investigar en ASL mediante la combinación de métodos de corpus y experimentales: Una introducción para investigadores y profesores de lenguas

How to do research in SLA by combining corpus and experimental methods: an introduction for language researchers and teachers

Resumen

El objetivo de este capítulo es introducir al lector en el proceso de investigación y en el del manejo de un software que le permita analizar corpus de datos y diseñar un experimento de seguimiento. Nos centraremos en la adquisición de pronombres del español como segunda lengua (L2).

Presentaremos una panorámica de los métodos de investigación en la adquisición de segundas lenguas (ASL), centrándonos en los métodos de corpus y experimentales. Tras describir el fenómeno lingüístico con el que trabajamos (pronombres), presentamos conceptos básicos en investigación (preguntas de investigación, hipótesis, variables, constantes, métodos de investigación, etc.) y posteriormente mostramos cómo se anota un corpus y se hace estadística con el software UAM Corpus Tool. Finalmente, partiendo de los resultados del corpus, mostramos como se diseña un experimento de seguimiento para confirmar/refutar los hallazgos del corpus. Defendemos que mediante la triangulación de datos de corpus y experimentales, podemos obtener una visión más completa de la lengua del aprendiz, lo cual en última instancia puede proporcionar una base más sólida para la enseñanza de lenguas.

Abstract

The aim of this chapter is to introduce the reader into the research process and to show them the use of a software to analyse corpus data and to design a follow-up experiment. To do so, we take a particular linguistic phenomenon as a case in point: the acquisition of pronouns in Spanish as a second language (L2).

We present a brief overview of research methods in Second Language Acquisition (SLA) research, with a focus on corpus and experimental methods. After describing the linguistic phenomenon we are to be working with (pronouns), we discuss basic concepts in research (research questions, hypotheses, variables, constants, research methods, etc.). We finally take a hands-on approach to practising how to annotate a corpus and how to do statistics in the UAM Corpus Tool software. Finally, based on the corpus findings, we show how to design an experiment to confirm/refute the corpus findings. We argue that it is by triangulating corpus and experimental methods that we can get a fuller understanding of L2 learners' language, which can ultimately provide a solid base for language teaching.

Palabras clave: Adquisición de segundas lenguas; Corpus de aprendices; Experimentos
Keywords: Second language acquisition; Learner corpora; Experiments

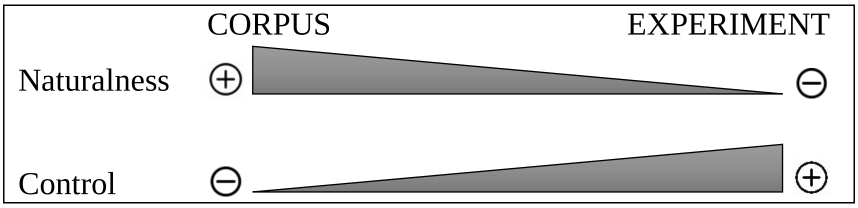
I Brief overview of research methods in SLA: experiments and corpora

Second Language Acquisition (SLA) research aims at describing (and also explaining) how learners acquire their second language (L2). We will use the term L2 to refer to non-native language acquisition independently of the context of learning (naturalistic or instructed). SLA researchers as well as language teachers and instructors should have noticed that L2 learners often produce linguistic items and structures that may be either (i) a direct consequence of the influence from their first language (L1), or (ii) a reflection of what they have learnt from being exposed to the written/oral input of the target language (L2) they are learning, or even (iii) the product of other universal or cognitive mechanisms that arise as a by-product of language development and that cannot be accounted for by transfer or input alone. The language produced by learners is often referred to as *interlanguage* (Selinker, 1972, 1991), which has been the focus of SLA research for near half a century now (Hawkins, 2001, 2018; Liceras, 1991; Liceras et al., 2008; Slabakova, 2016; White, 1989, 2003).

Different research methods can be used to systematically investigate interlanguages and different approaches to SLA have favoured one method over another. For example, in the generative-linguistics tradition, experiments have been typically favoured. In descriptive and functional approaches, corpora have been widely used. Such dichotomy is more apparent than real, since, for example, many generative and formal approaches to SLA have started to use corpus methods (Lozano, 2020b; Myles, 2005, 2007, 2015; Rankin, 2015).

There is no ideal research method to investigate interlanguage. Your method will depend on several factors: whether you are testing a particular hypothesis; whether you want to describe the data; your research questions and hypotheses; the participants; the degree of control (vs. naturalness) you want over your linguistic data; etc. Figure 1 represents the two research methods we will explore in this chapter (corpus and experiments) along two graded dimensions: the degree of naturalness of your linguistic items (i.e., how natural the language is) vs. the degree of control (i.e., how much you want to control for your variables and for your linguistic items). There is always a trade-off between naturalness and control. Corpus data typically sample ecologically valid language (i.e., language that is highly contextualised and represents natural production) though there is often little or no control over the linguistic structures/items the informants will produce. By contrast, in an experiment, the experimenter has a high degree of control over the stimuli (i.e., the linguistic structures/items) as well as other SLA-relevant variables (e.g., proficiency level of the learners, their language dominance, their working memory index, and any other variables that presumably may affect the results). Corpus data measure production, whereas the range of data measured by experiments can be wider: comprehension, online processing and controlled production. There are additional differences between corpora and experiments, e.g., the form-message dichotomy since when the focus is on the form (experiments), learners are more likely to resort to their explicit knowledge than when the focus is on the message (corpus). Nevertheless, these two methods are ultimately complementary. In some cases, a well-designed corpus can contain SLA-relevant variables, so the degree of control can increase when the researcher selects only those texts that meet certain criteria (Lozano, 2020a, 2020b), as we will see with the *Corpus Escrito del Español L2* (CEDEL2 corpus) below.

Figure 1: Research methods in SLA: contrasting corpus and experiments



In this chapter, we argue for triangulation, i.e., the combination of research methods, since many L2 acquisition phenomena are better understood if we triangulate corpus and experimental methods in a systematic way (Gilquin & Gries, 2009; Mendikoetxea & Lozano, 2018).

II A case in point: pronouns and discourse in SLA

In null-subject languages like Spanish (and also Italian, Greek, Arabic, etc.), null pronominal subjects (marked by Ø) are syntactically allowed to alternate with overt pronouns in subject position, as in (1). Full Noun Phrases (NPs) as well as proper nouns

(Ns) can also appear in such syntactic position. By contrast, non null-subject languages like English (and also German, French, etc.) do not allow null pronominals in subject position, (2), except in very limited contexts, as will be discussed later.

(1) $\left\{\begin{array}{c}\text{Sofía Vergara}\\ \text{Ella}\\ \text{Ø}\end{array}\right\}$ nació en Colombia.

(2) $\left\{\begin{array}{c}\text{Sofía Vergara}\\ \text{She}\\ \text{*Ø}\end{array}\right\}$ was born in Colombia.

An interesting question is how and when learners of L2 Spanish acquire not only the fact that null pronominal subjects are syntactically allowed in Spanish but also the discourse-pragmatic conditions that regulate them. In particular, the overt/null alternation in Spanish is constrained mainly (but not exclusively) by two major discursive factors having to do with information status: topic continuity vs. shift. First, *topic continuity* (also known as *topic maintenance*) is marked via a null pronoun, which is pragmatically more felicitous than an overt subject pronoun (*ella* 'she'), which would be pragmatically redundant (marked by the '*#*' symbol). This is so when there is only one antecedent as in sentence (3), where the null pronoun refers back to its antecedent *Sofía Vergara* in subject position in the preceding clause (as the correference $_{i}$ index indicates). This also holds when there are two antecedents, as in sentence (4). Second, in *topic shift* scenarios as in (4), an overt pronoun (*ella* 'she') is pragmatically required to change the topic since a null pronoun would be pragmatically infelicitous as it would lead to ambiguity, i.e., it could potentially refer to the subject antecedent *Antonio Banderas* or the non-subject antecedent *Sofía Vergara*. In short, null pronouns typically mark topic continuity whereas overt pronouns normally mark topic shift.

(3) *Sofía Vergara*$_{\text{i}}$ nació en Colombia. $\left\{\begin{array}{c}\textbf{\#Ella}_{\text{i}}\\ \text{Ø}_{\text{i}}\end{array}\right\}$ llegó a Estados Unidos para trabajar en el mundo de la televisión.

'*Sofía Vergara* was born in Colombia. (**She**) arrived in the USA to work in the world of TV.'

(4) *Antonio Banderas*$_{\text{i}}$ vivía con *Sofía Vergara*$_{\text{j}}$. $\left\{\begin{array}{c}\textbf{Ella}_{\text{j}}\\ \text{Ø}_{\mathbf{i/\#j}}\end{array}\right\}$ Llegó a Estados Unidos para trabajar en el mundo de la televisión.

‘*Antonio Banderas* used to live with *Sofía Vergara*. (**She**) arrived in the USA to work in the world of TV.’

In those cases where the two antecedents share the same gender (5), an overt pronoun *ella* would be ambiguous as it could refer to both *Elsa Pataki* (in subject position) or *Sofía Vergara* (in non-subject position). In this cases, Spanish prefers proper names or noun phrases for disambiguation purposes (Lozano, 2009b, 2016).

(5) *Elsa Pataki*$_{i}$ vivía con *Sofía Vergara*$_{j}$. $\left\{\begin{array}{c}\textbf{Elsa}_{\mathbf{i}}\\ \textbf{Sofía}_{\mathbf{j}}\\ \textbf{Ella}_{\mathbf{i/j}}\\ \text{Ø}_{\mathbf{i/\#j}}\end{array}\right\}$ Llegó a Estados Unidos para trabajar en el mundo de la televisión.

Early experimental evidence showed that L1 English-L2 Spanish learners acquire from early stages of development the fact that Spanish syntactically licenses null pronominal subjects (Liceras, 1989), but later experimental work showed that learners have problems acquiring the information-status conditions that constrain the overt/null alternation in discourse: they tolerate overt pronouns in topic continuity but are more aware of the infelicity of null pronouns in topic shift because a null pronoun would cause ambiguity (Lozano, 2018; Pérez-Leroux & Glass, 1999). This has been also reported for other L2 Romance languages such as Italian (Sorace, 2016 and references therein). More recent corpus work has shown that learners are indeed redundant by producing not only overt pronouns but also NPs when they are not pragmatically required, and that the overt/null pronoun alternation in discourse is constrained by additional factors not reported in the experimental literature, like the number of antecedents, the syntactic environment, etc. (Blackwell & Quesada, 2012; Lozano, 2009b, 2016; Martín-Villena & Lozano, 2020).

III How to work with corpora in SLA and SLT: A hands-on approach

In this section I will discuss how to conduct corpus-based and experimental research in SLA at small scale. I will depart from a typical classroom observation, turn this observation into a research question and a hypothesis, and finally I will use two research methods to confirm (or reject) the hypothesis: (i) corpus methods (we will show how to annotate a small sample from the CEDEL2 corpus with the help of a software, UAM Corpus Tool, which will lead to how tag the corpus and perform statistics on those tags);

(ii) experimental methods (we will use the insights from the corpus to design a basic experiment).

III.1 Introduction: working with pronouns and discourse in L2 Spanish; formulating research questions; setting up hypotheses

Those who have some experience in teaching a language like Spanish (or Italian, Greek, Arabic, etc.) where null and overt pronominal subjects are grammatically allowed, as in (1), (3) and (4), might have noticed that learners often produce an overt pronoun instead of a null pronoun in cases where they are not required (6a). They also produce full NPs or proper Ns in these scenarios, e.g., (6b). In other words, learners tend to be redundant or overexplicit.

(6) a. *Jennifer Lopez*$_i$ es cantante y actriz famosos. **Ella**$_i$ también va por el nombre de JLO. (*CEDEL2 corpus, beginner: EN_WR_15_30_1_2_TT*)

‘*Jennifer Lopez* is a famous singer and actress. **She** is also known as JLO.’

b. *Denzel Washington*$_i$ es muy guapo. **Senor Washington**$_i$ te gusta pelicula. (*CEDEL2 corpus, beginner: EN_WR_9_26_0_2_NLP*)

‘*Denzel Washington* is very handsome. **Mr Washington** likes film.’

Such classroom observation can be stated as in (7). Departing from such observation, we can formulate a research question, (8), which will be the driving force of our predictions below.

(7) **Classroom observation**: L2 learners of Spanish often overuse overt personal pronouns when they are not required, as in the examples in (6) above.

(8) **Research question**: Do L2 Spanish learners redundantly use pronominal subjects?

Typically, in research we make predictions about the differences between two groups (e.g., learners vs. natives) or between two linguistic elements (e.g., null vs. overt pronouns). A prediction is formulated as an experimental hypothesis (known as the alternative hypothesis, H_1). H_1 is set up to reject the null hypothesis (H_0) that there will be no difference between the two (or more) elements being contrasted. H_1 typically states the direction of the prediction of the elements the researcher is investigating (known as the *independent variables*, IVs) and how these IVs will have an impact on what we are measuring (the *dependent variable*, DV, also known as the response variable). It is important to include these basic ingredients (IVs, DV) when formulating a hypothesis. We will return to IVs and DVs below when discussing experimental design.

(9) Hypothesis formulation:

H_1: $IV_1 \neq IV_2$ (in relation to DV)

H_0: $IV_1 = IV_2$ (in relation to DV)

If we turn our attention now to the phenomenon under study, we can make the corresponding prediction in (10), where our IV is *group* (which contains two levels: *L2 learners vs. Spanish natives*) and our DV is the rate of overt pronominal subjects, which is what we are measuring. The way we measure it will depend on our research method, as we will see in detail below.

(10) H_1 (abbreviated format): L2 learners > Spanish natives (overt pronominal subjects)

H_1 (long format): L2 Spanish learners will be more redundant than Spanish natives by producing overt pronouns in subject position when maintaining the topic.

In the subsequent sections, H_1 will be the driving force that will shape our corpus research and our experimental research demonstrations.

III.2 Analysing the CEDEL2 corpus with UAM Corpus Tool

In this subsection, we will introduce the corpus we will use (CEDEL2) and then present the software we will use to analyse it (UAM Corpus Tool). We will explain, step by step, how to tag (i.e., annotate) a small sample of CEDEL2 so that researchers can (i) implement the observations and research questions in UAM Corpus Tool and (ii) confirm (or reject) the predictions we have formulated in the preceding section.

III.2.1 Learner corpora and the CEDEL2 corpus

L2 learners' interlanguage has been extensively used with the help of corpus data over the past decades (Callies & Paquot, 2015; Díaz-Negrillo & Thompson, 2013; Granger et al., 2015). Particularly fruitful has been the approach known as Contrastive Interlanguage Analysis (CIA) (Granger, 2015), whereby researchers typically contrast learners' varieties of interlanguage in several ways, e.g., across different proficiency levels (beginners vs. intermediate vs. advanced levels), against the native norm (e.g., advanced level vs. natives), etc. The use of L2 corpora in SLA research has been extensive in the case of L2 English corpora (Granger et al., 2015). In the case of L2 Spanish corpora, research is still in its infancy, though it is a burgeoning field (Alonso-Ramos, 2016; Mas Álvarez & Gil Martínez, 2018; Mendikoetxea, 2014; Sánchez Rufat, 2017).

CEDEL2 (Lozano, in prep., 2009a; Lozano & Mendikoetxea, 2013) is a multi-L1 corpus of L2 Spanish learners. It is freely available online for browsing and downloading (http://cedel2.learnercorpora.com). It is an ongoing project and in its current version (version 2.0 beta, February 2020) it contains data from over 4,000 learners from a variety of L1s (English, German, Dutch, Portuguese, Italian, French, Greek, Japanese, Chinese, Arabic and Russian), as well as a series of native subcorpora (Spanish, English, Japanese, Greek, Portuguese and Arabic) that serve as 'control' or normative corpora. CEDEL2 samples a variety of learners in terms of proficiency level, learning environment and other SLA-relevant variables, so it has been argued to be an SLA-motivated, well-designed corpus for SLA research (Gilquin, 2015; Sánchez Rufat, 2017). In the next subsections, we will contrast a sample of L1 English-L2 Spanish learners (beginner level) with Spanish natives from the CEDEL2 corpus with the help of a tagging software: UAM Corpus Tool.

III.2.2 UAM Corpus Tool: Defining a tagset and tagging

UAM Corpus Tool (http://www.corpustool.com) (O'Donnell, 2009) is a free software for the manual (and also automatic) annotation of text corpora. The user first defines a hierarchy of tags (i.e., a tagset) and then manually assigns tags to elements in each text. We can annotate multiple texts at multiple levels, which can range in size, i.e., we can tag a letter, a morpheme, a word, a phrase, a clause, a sentence, a paragraph, or even an entire text. We can then compare groups of texts (e.g., learners vs. natives) based on specific tags (e.g., learners vs. natives' use of null pronouns; or learners vs. natives' use of null pronouns in coordination vs. subordination). In this respect, UAM Corpus Tool very versatile and allows the researcher to explore multiple combinations.

First, we will download UAM Corpus Tool and work with the current latest version (version 3.3.) for either Windows or Mac. After installing the software, we will follow the steps in (11) and create a project with a view to learning how to analyse our sample texts from CEDEL2. The reader is referred to additional resources to learn how to use UAM Corpus Tool: cf. Lozano (2020a) and several video tutorials in YouTube.

(11) **Starting a new project in UAM Corpus Tool:** Click on *Start New Project* > Click on *Next* > Type in the name of the project, e.g., "Pronominal subjects project" > Select the folder where the project should be saved > Click on *Finalise*.

Next, let's add texts to our project. For this exercise, we will download four texts from L1 English-L2 Spanish learners from CEDEL2 (see the list of texts in Table 1 in Appendix 1) and two Spanish native texts (see the texts in Table 2 in Appendix 2). Follow the steps in (12). Needless to say, six texts are too few to do a solid analysis, but they will be enough to practically illustrate how to tag and analyse the corpus. The texts are based on the topic *Persona famosa* 'Famous person', in which the participants have to write about a famous person, which creates a lot of opportunities for topic continuity.

(12) **Downloading corpus files:** Go to the CEDEL2 webpage (http://cedel2.learnercorpora.com) and search for the following files:

L1 English-L2 Spanish subcorpus: EN_WR_9_26_0_2_NLP; EN_WR_12_18_2_2_MLS; EN_WR_13_38_1_2_JAR; EN_WR_15_30_1_2_TT.

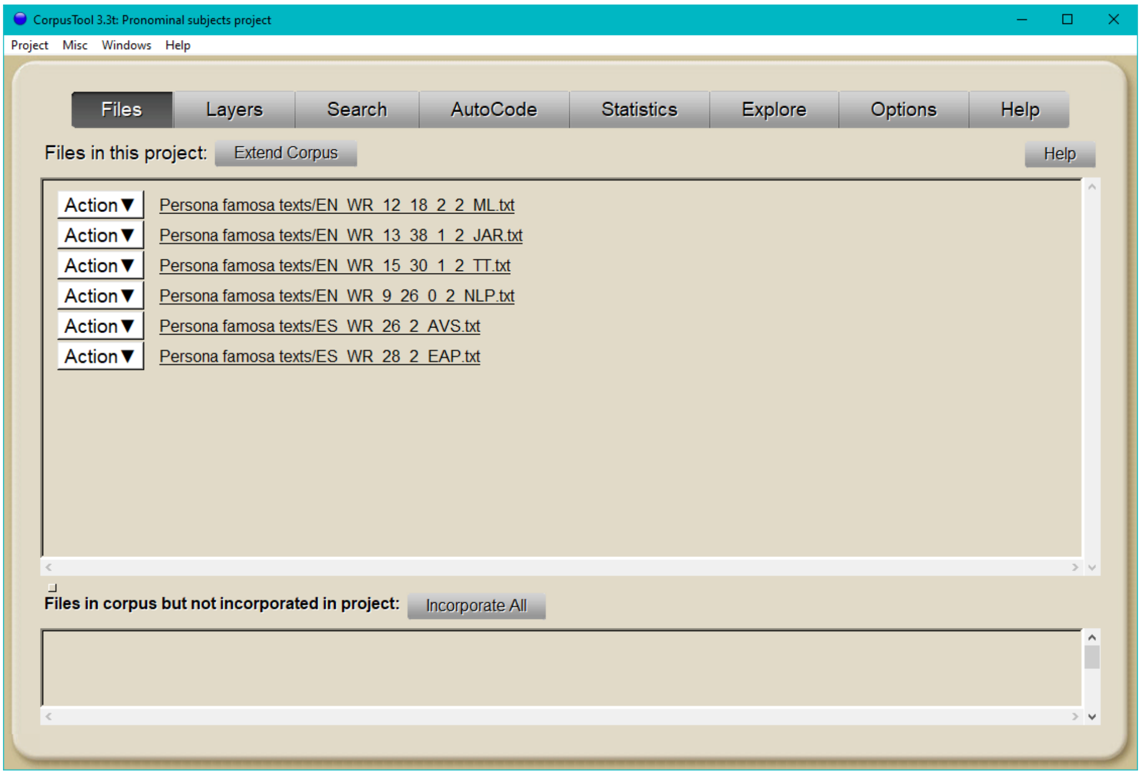
Spanish native subcorpus: ES_WR_28_2_EAP; ES_WR_26_2_AVS.

NOTE FROM CRISTÓBAL TO THE EDITORS: IF POSSIBLE, I'D RATHER ARANZADI UPLOADED THESE FILES AS SUPPLEMENTARY MATERIAL TO THE CHAPTER ONTO THEIR WEBPAGE. IT IS FASTER FOR THE READER TO DOWNLOAD THE SIX FILES IN ONE GO RATHER THAN BROWSING THE CORPUS CEDEL2 AND FINDING THEM ONE BY ONE.

Once the six texts have been downloaded, we need to go back to the main interface in UAM Corpus Tool and add those text files to our project, as instructed in (13). You should end up with the UAM Corpus Tool interface looking something like Figure 2.

(13) **Add corpus files to a project in UAM Corpus Tool**: Click on the button *Extend Corpus* > In the *Corpus Location* window, select tehe option *I want to add a folder of text files* > Click on the three-dotted button > Select the folder where you have downloaded your texts > Click on *Finalize* > Finally, you need to incorporate all the files onto the final project by clicking on *Incorporate All* (click on *OK* should you get messages about text encoding).

Figure 2: Main interface in UAM Corpus Tool



Persona famosa texts/EN_WR_12_18_2_2_ML.txt
Persona famosa texts/EN_WR_13_38_1_2_JAR.txt
Persona famosa texts/EN_WR_15_30_1_2_TT.txt
Persona famosa texts/EN_WR_9_26_0_2_NLP.txt
Persona famosa texts/ES_WR_26_2_AVS.txt
Persona famosa texts/ES_WR_28_2_EAP.txt

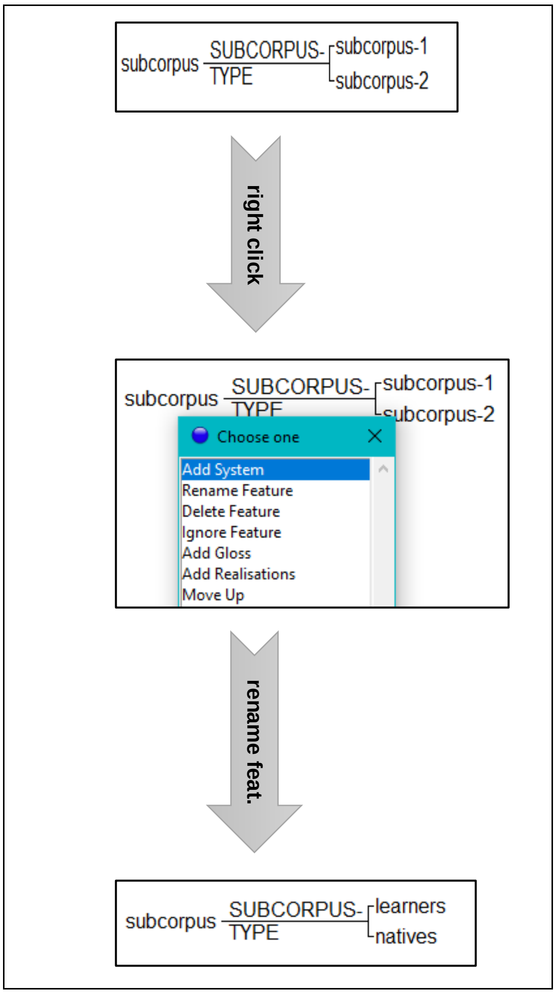
Once our files are incorporated in the project, we need to define two layers of annotation: one layer to annotate each text so as to categorise it into learners/natives, and another layer to annotate every subject representing topic continuity.

The first layer will be used to classify our texts into learners (the four files starting with EN, representing English, the L1 of our learners) vs natives (the two files starting with ES, representing *Español* ‘Spanish’, the L1 of our natives).

(14) **Create a layer (annotation of whole documents):** Button *Layers* > *Add Layers* > *Start* > Name the layer as “subcorpus” > *Manual Annotation* > *Design Your Own* > *Whole Document* so as to annotate each text as belonging to the learners or natives subcorpus > *Create Layer*.

We need to define now the tags for the *subcorpus* layer. When clicking on *Edit scheme*, the software will generate two tags by default, *subcorpus-1* and *subcorpus-2*, as in Figure 3. We need to rename each of them by right-clicking on each tag at a time, then selecting *Rename Feature* from the pop-up window and assigning the names *learners* and *natives* respectively.

Figure 3: The *subcorpus* tagset: Renaming process



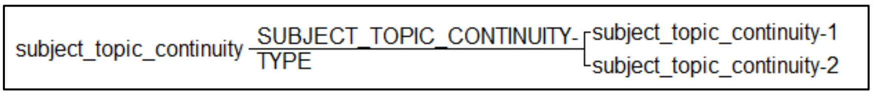
In our second layer we will define a tagset that will contain the tags to annotate segments containing topic continuity, as in (15).

(15) **Create a layer (annotation of segments):** Button *Layers* > *Add Layers* > *Start* > Name layer as "subject_topic_continuity" > *Manual Annotation* > *Design Your Own* > *Segments within a Document* (we will be annotating segments this time) > In the options for special segment, click on either *No* (default option) or on *Error* (if you want to specify the correct version of the error produced by the learner, just for informative purposes) > In the automatic segment options, click on *No* > *Create Layer*.

We need to edit the scheme now. By default, UAM Corpus Tool generates a scheme with a root (*subject_topic_continuity*) and two branches (*subject_topic_continuity-1* and *subject_topic_continuity-2*), as shown in Step 1 in Figure 4. The square bracket option represents an OR logical operator, i.e., we can choose either *subject_topic_continuity-1* or *subject_topic_continuity-2*. This is visually shown as a square bracket bifurcation in Step 1. What we need at this stage is to create a scheme with two systems or branches: we need to tag for the *syntax* of the subject (either null_pronoun or overt_pronoun or NP) and also for its *pragmatics* (felicitous or infelicitous), as shown in Steps 2 and 3. So, the root will contain an AND logical operator, visually shown as a curly bracket, which indicates that we will first assign a tag to the branch *syntax* and then another tag to the branch *pragmatics*. For example, the overt pronoun *ella* 'she' in sentence (6) above would be assigned the terminal tags *overt_pronoun, infelicitous,* and *overt-when-null,* as explained after Step 4 in Figure 4.

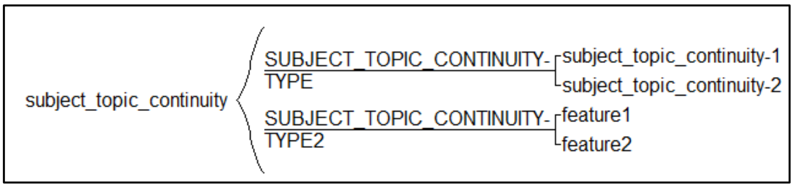
Figure 4: The *subject_topic_continuity* tagset: Creation and renaming process

Step 1: Default tagset:


subject_topic_continuity
SUBJECT_TOPIC_CONTINUITY-TYPE
subject_topic_continuity-1
subject_topic_continuity-2

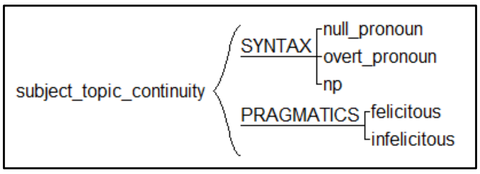
Right click on the root *subject_topic_continuity* so as to add an AND operator (visually shown as a curly bracket). Select "Add system" from the pop-up window and you should end up with this:

Step 2: Tagset with two systems:


subject_topic_continuity
SUBJECT_TOPIC_CONTINUITY-TYPE
subject_topic_continuity-1
subject_topic_continuity-2
SUBJECT_TOPIC_CONTINUITY-TYPE2
feature1
feature2

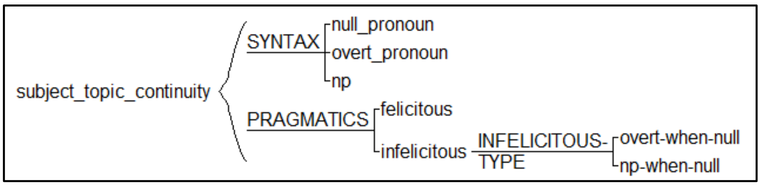
Right click on each branch and on each terminal tag to rename them as shown in Step 3. Additionally, right click on the newly renamed branch *SYNTAX* to add an additional feature (NP). You should end up with something like this:

Step 3: Edited tagset with two systems:



Finally, if we want to specify the type of pragmatic infelicity, right click on the *infelicitous* tag and select "Add a system" to create an 'OR' branch: *overt-when-null* or *NP-when-null*. We can specify now whether an overt pronoun has been used instead of a null pronominal subject (e.g., *Sofía Vergara$_i$ nació en Colombia* y #***ella****$_i$ se fue a EEUU en 2000*) or an NP instead of a null pronominal subject (*Sofía Vergara$_i$ nació en Colombia* y #***Sofía****$_i$ se fue a EEUU en 2000*). You should end up with something like this:

Step 4: Final tagset with two systems and one subsystem:



Once we have created the two tagsets in Figure 3 and Figure 4, we need start tagging, as instructed in (16). For additional information on the tagging of segments for the layer *subject_topic_continuity*, the reader is referred to the appendices at the end of this chapter, where the referential expressions in topic-continuity position have been highlighted. In those cases where we need to tag a null pronominal subject, we will simply tag the verb for convenience, given that null/elliptical elements are not visible in the original text.

(16) **Tagging**: Make sure you are on the project main page (click on the button *Files*) so as to visualize the list of files, each with its layer to its left, as shown in Figure 5 > Click on the name of the layer to be tagged (*subcorpus*|*subject_topic_continuity*), which will be shown in dark orange (this indicates that the file is ready to be tagged; when the file has been fully tagged, the layer colour will change to light orange).

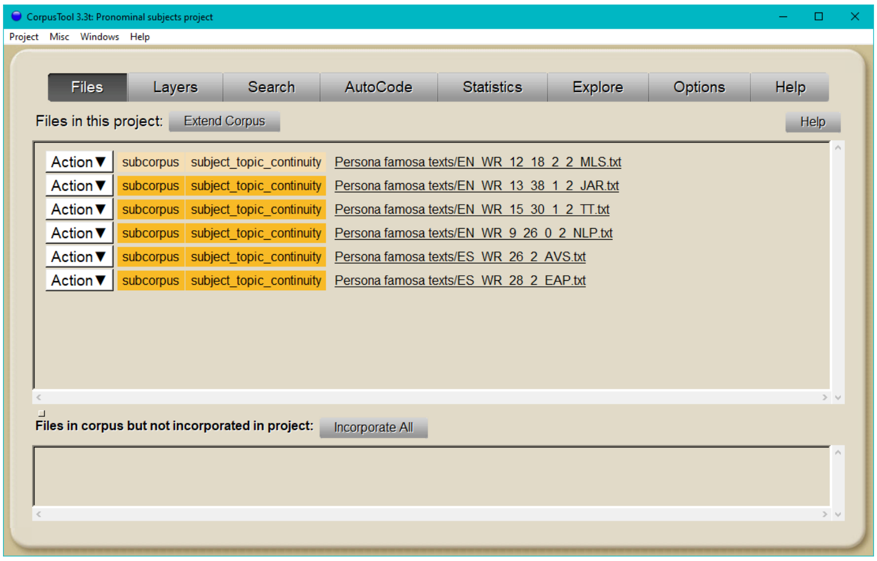
Tagging a whole text (*subcorpus* layer) > Click on the *subcorpus* layer button > Double click on the corresponding tag (*learners*|*natives*) to assign a tag to the text > Close the window > Save changes.

Tagging a segment (*subject_topic_continuity* layer): Click on layer button > Click on the beginning of the segment to be tagged and drag the mouse while still pressing the mouse button until the end of the segment > Release the mouse button > The selected segment will be marked with green underlining > Assign the corresponding tags relating to the branches *Syntax* and *Pragmatics* by double clicking on each tag (see the result in Figure 6) > Close the window > Save changes.

Automatically tagging repeated segments: In those cases where the segments to be tagged are the same, by default UAM Corpus Tool will automatically assign the same tags when you highlight a repeated segment. For example, every time we highlight *El* 'He' in example (25) in Appendix 1, it will be assigned the same tags: overt_pronoun, infelicitous, overt-when-null. If you want to disable this option, click on the button *Options* > *Coding Options* > Tick as 'False' the option *Automatically code repeated segments*.

Untagging a segment: If you assign an incorrect tag to a segment by accident, simply double click on the assigned tag in the *Selected* window (where you can see the already assigned tags) and the tag will be unassigned.

Figure 5: UAM Corpus Tool main project window with layers and files ready to be tagged

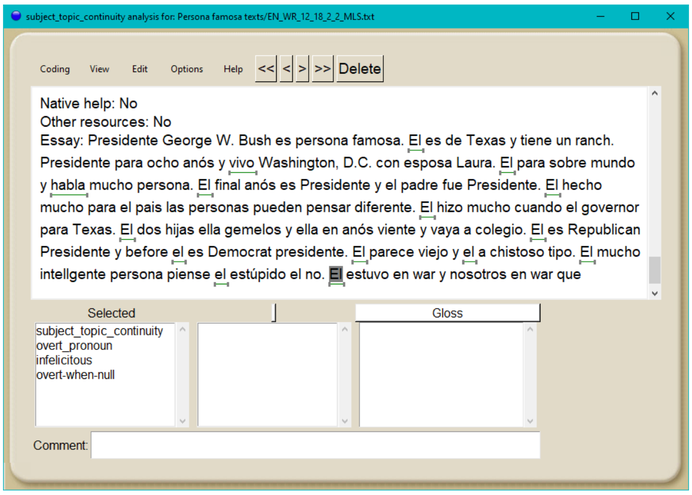

CorpusTool 3.3t: Pronominal subjects project
Project Misc Windows Help
Files Layers Search AutoCode Statistics Explore Options Help
Extend Corpus
Help
Incorporate All

Files in this project:

Action▼	subcorpus	subject_topic_continuity	Persona famosa texts/EN_WR_12_18_2_2_MLS.txt
Action▼	subcorpus	subject_topic_continuity	Persona famosa texts/EN_WR_13_38_1_2_JAR.txt
Action▼	subcorpus	subject_topic_continuity	Persona famosa texts/EN_WR_15_30_1_2_TT.txt
Action▼	subcorpus	subject_topic_continuity	Persona famosa texts/EN_WR_9_26_0_2_NLP.txt
Action▼	subcorpus	subject_topic_continuity	Persona famosa texts/ES_WR_26_2_AVS.txt
Action▼	subcorpus	subject_topic_continuity	Persona famosa texts/ES_WR_28_2_EAP.txt

Files in corpus but not incorporated in project:

Figure 6: Example of the tagging process


subject_topic_continuity analysis for: Persona famosa texts/EN_WR_12_18_2_2_MLS.txt
Coding View Edit Options Help
<< < > >> Delete
Selected
Gloss
Comment:

Native help: No
Other resources: No
Essay: Presidente George W. Bush es persona famosa. El es de Texas y tiene un ranch. Presidente para ocho anós y vivo Washington, D.C. con esposa Laura. El para sobre mundo y habla mucho persona. El final anós es Presidente y el padre fue Presidente. El hecho mucho para el pais las personas pueden pensar diferente. El hizo mucho cuando el governor para Texas. El dos hijas ella gemelos y ella en anós viente y vaya a colegio. El es Republican Presidente y before el es Democrat presidente. El parece viejo y el a chistoso tipo. El mucho intellgente persona piense el estúpido el no. El estuvo en war y nosotros en war que

subject_topic_continuity
overt_pronoun
infelicitous
overt-when-null

Once we have tagged all the texts as belonging to either the *learners* or *natives* subcorpus and also tagged all the segments within each text (see the Appendices for a solution to this), we can start doing statistics in UAM Corpus Tool.

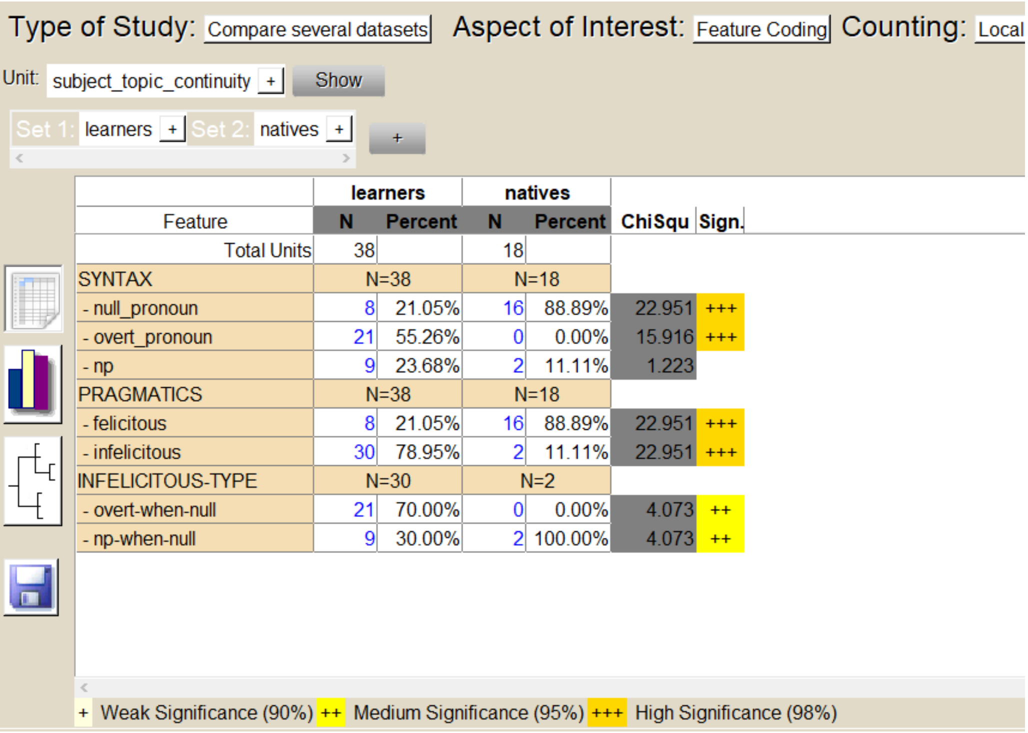
III.2.3 UAM Corpus Tool: Doing statistics on the tags

Let's go back to our research question in (8) and hypothesis H_1 in (10). In order to (dis)confirm our hypothesis, we need to compare two subcorpora (learners vs. natives) and determine whether learners are more redundant (i.e., whether they produce more overt pronominal subjects) than Spanish natives. We also need to check whether the difference is statistically significant, i.e., whether the difference is strong enough from a mathematical point of view to reject the null hypothesis H_0 and, therefore, accept H_1. The concept of statistical significance is beyond the scope of this chapter, so the reader is referred to readings on statistics in SLA to better understand the concept of statistical significance: introductory readings (Mackey & Gass, 2016 Chapter 10) and detailed textbooks (Larson-Hall, 2016; Lowie & Seton, 2013).

(17) **Statistics (comparing two sets):** Button *Statistics* > In *Type of Study* select *Compare several datasets* to compare two groups or two subcorpora (learners vs. natives); in *Aspect of Interest* select *Feature Coding* to perform statistics on the tags > in *Counting* select *Local* (see the next paragraph for an explanation of *Local* vs. *Global* counting) > in *Unit* select the unit to analyse, in our case: the root of the tagset, *subject_topic_continuity*, although we could have chosen any other branch or tag; in *Set 1* choose the first subcorpus to be compared ('learners') and in *Set 2* choose the second subcorpus ('natives') > Click on *Show* to see the statistics (Figure 7).

Local vs. Global counting. The difference between local vs. global counting is simple: Local counting represents the count out of the total count in a given sub-branch or subsystem, whereas in global counting, for every branch or system, we will get the counting out of the total number of counts in the root. To illustrate, in Figure 7, for the natives we have a total of 18 cases for the root *subject-topic-continuity*. The count for the infelicitous tag *np-when-null* is 2. The local count is 2 out of 2 infelicitous cases (100%), but the global counting would be 2 out of 18 root cases (11.11%). While the raw frequency is always the same (i.e., 2 counts in this case), the percentage will vary depending on whether we do a local or global counting.

Figure 7: UAM Corpus Tool statistics: comparing two sets (learners vs. natives)



Type of Study: Compare several datasets Aspect of Interest: Feature Coding Counting: Local

Unit: subject_topic_continuity

Set 1: learners Set 2: natives

	learners		natives			
Feature	N	Percent	N	Percent	ChiSqu	Sign.
Total Units	38		18			
SYNTAX	N=38		N=18			
- null_pronoun	8	21.05%	16	88.89%	22.951	+++
- overt_pronoun	21	55.26%	0	0.00%	15.916	+++
- np	9	23.68%	2	11.11%	1.223	
PRAGMATICS	N=38		N=18			
- felicitous	8	21.05%	16	88.89%	22.951	+++
- infelicitous	30	78.95%	2	11.11%	22.951	+++
INFELICITOUS-TYPE	N=30		N=2			
- overt-when-null	21	70.00%	0	0.00%	4.073	++
- np-when-null	9	30.00%	2	100.00%	4.073	++

+ Weak Significance (90%) ++ Medium Significance (95%) +++ High Significance (98%)

As we can see in Figure 7, there is a very strong statistical difference (marked by three plus symbols) between natives and learners in their use of null pronouns: learners 21.05% vs. natives 88.89%. This difference is statistically significant (p<0.02), which means that the probability (p) value of the results having been due to chance is 2% or lower, as indicated by the three plus symbols in the figure. Therefore we can reject H_0 with a 98% or higher degree of confidence and therefore accept our H_1 with a 2% of lower degree of error; cf. also the Chi Square result (also represented as χ^2), which would need to be reported in your study. This finding entails that we can be highly confident that learners' production of null pronominal subjects to mark topic-continuity (21.05%) is significantly lower than natives' (88.89%). As we can see, learners' production is redundant since they mark topic continuity mostly by producing overt pronouns (55.26%), whereas natives never do so (0%), at least in our corpus sample.

Additionally, we can see that natives are significantly more felicitous (88.89%) than our beginner learners (21.05%), as expected. Most cases of such infelicity correspond to the overuse of overt pronouns instead of null pronouns (70% out of all infelicitous cases) and a few cases of NP instead of null (30%), as expected.

Importantly, we are dealing with very low frequencies per cell in this exercise. The χ^2 test assumes a minimum of ≥5 observations per cell. For example, in Figure 7 the frequency of natives' production of NPs is 2, which is less than 5. In this case, the reader is advised to use Fisher's Exact Test (cf. the online statistical calculator Graphpad: www.graphpad.com).

III.2.4 Beyond statistics in UAM Corpus Tool: Reformulating hypotheses and back

An interesting question from our statistical results is why our beginners occasionally produce a pragmatically felicitous null pronoun (21.05%) to mark topic continuity. It may be the case that they are starting to pick up from the input the fact that Spanish allows a null pronominal subject in Topic Continuity scenarios. Or it may be the case that they are transferring null pronominal subjects from their L1 English, since it is well known that null pronominal subjects are allowed in coordination in native English in general (Oh, 2006 for an overview). However, coordination is a necessary but not a sufficient condition since there must be a topic continuity in such coordination and the antecedent of the null subject must be in subject (preverbal) position, and both English natives and English learners are sensitive to this (Quesada & Lozano, accepted). To illustrate, let's consider (18): in sentence *a* a null pronominal subject is not allowed in sentence-initial position, in *b* it is allowed in topic-continuity coordinate clause and in *c* it is not allowed in a subordinate clause. By contrast, given that Spanish is a null-subject language, null pronouns are allowed in all syntactic scenarios, as the equivalent sentences in (19) show.

(18) a. *Sofía Vergara*$_i$ was born in Colombia. ***Ø**$_i$ came to the USA to work in TV serials.

b. *Sofía Vergara*$_i$ was born in Colombia and **Ø**$_i$ came to the USA to work in TV serials..

c. *Sofía Vergara*$_i$ was 30 when ***Ø**$_i$ came to the USA to work in in TV serials.

(19) a. *Sofía Vergara*$_i$ nació en Colombia. **Ø**$_i$ vino a los EEUU para trabajar en series de TV.

b. *Sofía Vergara*$_i$ nació en Colombia y **Ø**$_i$ vino a los EEUU para trabajar en series de TV.

c. *Sofía Vergara*$_i$ tenía 30 cuando **Ø**$_i$ vino a los EEUU para trabajar en series de TV.

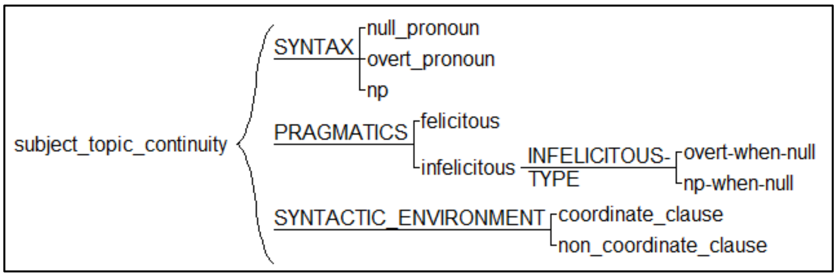
Based on these syntactic and pragmatic observations, we can reformulate our initial hypothesis in (10) above as two hypotheses in (20). Incidentally, note that H_2 technically corresponds to a null hypothesis and H_3 to an alternative hypothesis, but we keep the numbering H_2 and H_3 for illustrative purposes.

(20) H_2 (no transfer hypothesis): IF learners produce Ø in all syntactic scenarios, THEN they are not transferring Ø from English → coordination ≈ non-coordination (Ø)

H_3 (transfer hypothesis): IF learners produce Ø only in coordinate scenarios, THEN they are transferring Ø from English → coordination > non-coordination (Ø)

We need to do additional tagging: for every subject we have tagged, we need to specify now its syntactic environment: coordinate clause vs. non-coordinate clause. We need to add a new system or branch to our final tagset (cf. step 4 in Figure 4), namely, *syntactic environment: coordinate_clause|non_coordinate_clause*. Follow the instructions in Step 1-Step 2 (cf. Figure 4 above) to define such a new system. The final tagset should look like Figure 8. We need to include our new tag to every text. Follow the instructions for tagging a segment in (16) above.

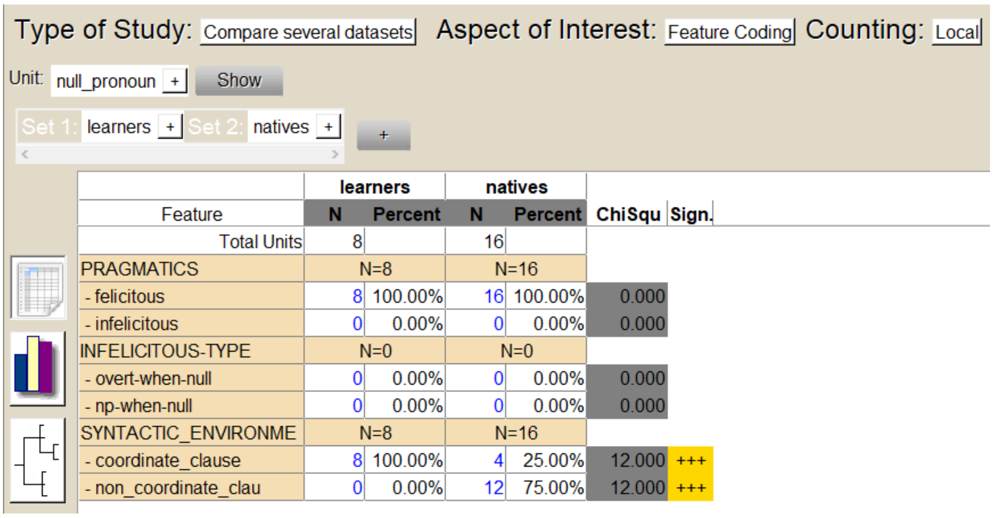
Figure 8: Final tagset with three systems or branches



We can now proceed to confirming H_2 or H_3. We ask for comparative statistics again, as we did in (17). Importantly, our unit of analysis now is not the entire root (*subject_topic_continuity*) but the tag *null_pronoun* since we are interested now in whether null pronominal subjects are produced by learners in all syntactic scenarios (coordination and non coordination), as predicted by H_2 (no transfer hypothesis), or rather in coordination only, as predicted by H_3 (transfer hypothesis). The results output (Figure 9) shows that learner produce null pronouns only in coordination (100%), whereas

Spanish natives produce them across the board (coordination: 25%; non coordination: 75%). This confirms H_3 and rejects H_2. The sample for this exercise is small but results are suggestive: Beginner's production of null pronouns in Spanish is not necessarily a consequence of acquisition, as originally thought in the early experimental literature, but it rather seems to be caused by transfer of null pronouns in coordinated topic-continuity structures from their L1 English (cf. Martín-Villena & Lozano (2020) for further details on coordination and topic continuity in L1 English-L2 Spanish).

Figure 9: A statistical comparison taking into account *syntactic_environment*



Type of Study: Compare several datasets Aspect of Interest: Feature Coding Counting: Local

Unit: null_pronoun

Set 1: learners Set 2: natives

	learners		natives			
Feature	N	Percent	N	Percent	ChiSqu	Sign.
Total Units	8		16			
PRAGMATICS	N=8		N=16			
- felicitous	8	100.00%	16	100.00%	0.000	
- infelicitous	0	0.00%	0	0.00%	0.000	
INFELICITOUS-TYPE	N=0		N=0			
- overt-when-null	0	0.00%	0	0.00%	0.000	
- np-when-null	0	0.00%	0	0.00%	0.000	
SYNTACTIC_ENVIRONME	N=8		N=16			
- coordinate_clause	8	100.00%	4	25.00%	12.000	+++
- non_coordinate_clau	0	0.00%	12	75.00%	12.000	+++

III.3 Designing an offline experiment

The empirical insights we have gained from analysing corpus data can be summarised as follows. At beginning level, L1 English-L2 Spanish learners (i) redundantly produce overt pronominal subjects to mark topic continuity; (ii) correctly produce null pronominal subjects to mark topic continuity only in coordination, though this appears to be a result of L1 transfer and not L2 acquisition proper. We can summarise these corpus findings as research questions (RQs) in (21), which we can test now via an experiment. Our experimental RQs are still based on our earlier hypotheses, H_1 in (10) and H_3 in (20), but rather than making predictions about production (corpus data), we are making predictions now about acceptability data (experimental data). In this way, we will triangulate corpus and experimental data to (dis)confirm our hypotheses. If confirmed, the combination of two different research methods and data types will provide a more solid basis to our L2 acquisition predictions.

(21) RQs: In an experimental setting, will learners accept redundant pronouns to mark topic continuity? Will they accept null pronouns in coordination more than in subordination?

Given that in an experiment we have a high degree of control over the linguistic factors that we want to test, we can manipulate them. In our case, these factors (also known as *independent variables*, IVs) are the type of referential expression (RE), which has three levels (Ø|overt|N) and the type of syntactic scenario, with two levels (coordination|non-coordination), as shown in (22).

(22) IV$_1$: Type of RE (Ø|overt|N)

IV$_2$: Type of syntactic scenario (coordination|non-coordination)

The IVs will determine our *experimental design.* Given the number of IVs and their levels above, we have a 3 x 2 experimental design, as illustrated in Figure 10, which yields 6 experimental *conditions* (i.e., sentence types) that need to be tested.

Figure 10: A 3x2 experimental design with 6 conditions

		IV$_1$: **RE form**		
		Ø	**overt**	**N**
IV$_2$: **Syn. scenario**	**Coord.**	1	2	3
	Non coord.	4	5	6

Another IV we are manipulating is nativeness (learners vs. natives), as shown in (23). This variable is known as a *between-group* IV since a given participant belongs to only one level (either you are a learner or you are a native, but you cannot be in both groups). By contrast, the linguistic variables above (IV$_1$, IV$_2$) are *within-group* IVs since every participant is measured for every level of the IV or, in other words, everybody is measured on the 6 conditions.

(23) IV$_3$: Nativeness (learners|natives)

As researchers, we will feel naturally tempted to introduce into our experimental design too many IVs. For example, we may wish to include (i) between-group IVs like proficiency level {beginner|intermediate|advanced}, the learners' L1 {English|Italian}, etc; and (ii) within-group IVs like the information status of the RE {topic continuity|topic shift}, the number of antecedents in the preceding discourse {1|2}, etc. Imagine the complex experimental design we would end up having. Sometimes, it is not viable to include too many IVs simultaneously in the same experiment.

An important concept in experimental design are *constants*. By definition, unlike IVs, constants do not vary. In our case we have several constants: the number of antecedents (there is only 1 antecedent); the information status of the RE (the subject always marks topic continuity); the person of the subject (third person singular); the proficiency level of the learners (beginners). Basically, we want the results of our experiment to be due to the manipulation of the IVs and nothing else, hence the need to keep everything as constant as possible except for the IVs.

We should avoid *confounding variables* or *unwanted factors*. We always need to saturate all the cells in an NxN table. In our case, we need to test every cell in the 3x2 table in Figure 10. Imagine we are interested only in IV_1 (RE form). It would make no sense to test the three RE forms in coordination and just overt pronouns in subordination. When we obtain the results, we will be unsure of whether they are due to the form itself or they are partly explained by a confounding variable (overt pronouns in subordination). In order to eliminate that confounding variable, it is advisable to either exclude it altogether or to include it and turn it into IV_2 by ensuring it is represented at the three levels of the IV_1 RE form, as the table shows.

The next step is to construct *linguistic stimuli.* Once we have a clear idea about our experimental design (IVs, constants), we will construct stimuli for each condition. A participant will have to see an equal amount of sentences in each condition. Our 6 conditions are illustrated in (24), which are called an *item.* In general, depending on the experimental method used, we will need every participant to see 5 to 10 stimuli per condition in order to have enough statistical power in the results. That means creating from 5 to 10 items in *offline* experiments like the ones shown here, i.e., experiments where participants read sentences and then rate them, but in *online* experiments (i.e., psycholinguistic experiments where we measure participants' reaction time in real time as the sentence unfolds), the number will increase. The reader is referred to additional readings for offline experimental methods (Cowart, 1997; Ionin, 2012; Phakiti, 2014;

Roever & Phakiti, 2017; Rose et al., 2019; Sorace, 1996; Spinner & Gass, 2019) and online experimental methods (Keating & Jegerski, 2015; Marinis, 2010; Roberts, 2012).

Stimulus design is an art in itself. We need to take into account different issues, particularly when investigating non-native speakers. For example, we need to ensure that the vocabulary used will be understood by the learners since we want the results to be due to our IVs and not to vocabulary-related issues. Another stimulus design issue is *counterbalancing*. In our case, given that we are using 3rd person singular personal pronouns, we could use a masculine pronoun (*él* 'he') in 50% of our stimuli and a feminine pronoun (*ella* 'she') in the remaining 50%. Counterbalancing can be applied in many ways, e.g., by using 50% of the sentences in the present tense and 50% in the past tense, etc.

(24) *Condition:* *Stimulus:*

1. Antonio estudió en Madrid y trabajó en Barcelona.
2. Antonio estudió en Madrid y él trabajó en Barcelona.
3. Antonio estudió en Madrid y Antonio trabajó en Barcelona.
4. Antonio dice que trabajó en Barcelona.
5. Antonio dice que él trabajó en Barcelona.
6. Antonio dice que Antonio trabajó en Barcelona.

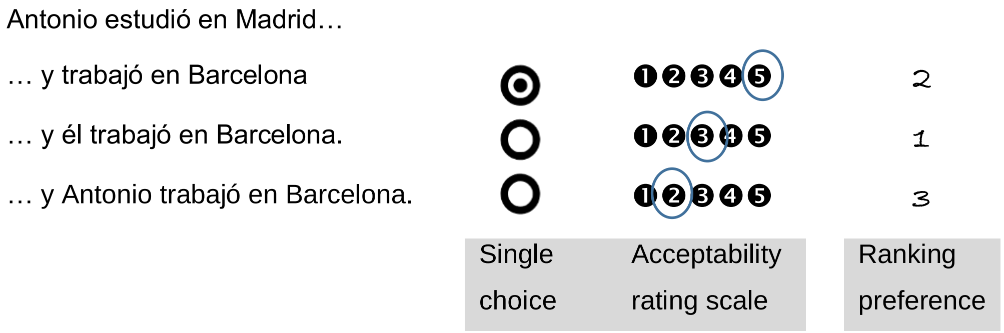
Finally, we need to think about the actual *experimental method* we want to implement. In Figure 11 we can see three typical experimental methods:

i. Multiple-choice answer: In our case, given the preceding sentence (*Antonio estudió en Madrid…*), which serves as context, the participant is presented with three target sentences to choose from. Obviously, we will lose information in case that the participant completely tolerates one option but partially tolerates another option. This can be compensated with (ii).

ii. Acceptability rating scale: We can use a Likert scale to measure the participant's acceptability about each target sentence by using either a positive scale ranging from 1…5 or a positive-negative scale ranging from -2 … +2, or any other ranges. In this case, we can get a more fine-grained measure of the participants' acceptability, as it may be the case that two target sentences are equally acceptable/unacceptable.

iii. A ranking preference test, whereby the participants see the three target sentences and have to rate each of them in order of preference.

Figure 11: Three typical experimental methods


Antonio estudió en Madrid…
… y trabajó en Barcelona
… y él trabajó en Barcelona.
… y Antonio trabajó en Barcelona.
1 2 3 4 5
2
1
3
Single choice
Acceptability rating scale
Ranking preference

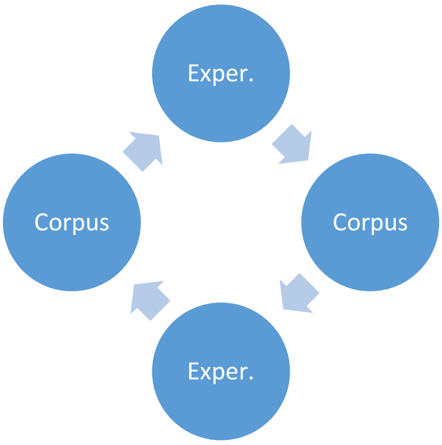
The reader is referred to books on SLA research methods for more details on different types of experiments (Blom & Unsworth, 2010; Ionin, 2012; Mackey & Gass, 2012; Rose et al., 2019; Spinner & Gass, 2019). These experiments can be ultimately implemented in different ways (pen-and-paper forms, platforms for data collection via online forms like Google Forms or LimeSurvey, etc.).

Finally, we also need to think about the *dependent variable* (also known as the response variable), which is what we are measuring. In our case, based on the RQs and Hypotheses, we want to measure the participants' behaviour on the IVs. The precise method we will use to measure the DV will determine the nature of the data, which in turn will determine the statistical method needed to analyse the data: frequency data with a single-choice test, ordinal numeric data as in the ranking-preference test, or scale data as in the acceptability test. The reader is referred to books on statistical analysis in SLA (Larson-Hall, 2016; Lowie & Seton, 2013; Phakiti, 2014; Roever & Phakiti, 2017). Given space limitations in this chapter, the reader is now invited to implement a basic experiment to (dis)confirm the corpus data from the CEDEL2 corpus.

III.4 Closing the circle: Triangulation in a cyclic fashion

Triangulation in science is not new and it has been used in SLA in relation to the combined used of corpus and experimental data (Gilquin & Gries, 2009; Mendikoetxea & Lozano, 2018). However, the combination of different research methods to investigate the same phenomenon should not be envisaged as a static process (i.e., we use corpus and experimental data simultaneously to investigate a phenomenon), but as something dynamic and cyclic whereby we may depart from, e.g., corpus evidence, which can give us insights to be tested in an experiment, which in turn may give us ideas to be searched in the corpus, which may provide new evidence that can be tested in an experiment, and so on. This is known as triangulation in a cyclic fashion (Mendikoetxea & Lozano, 2018), as illustrated in Figure 12. Doing this type of cyclic research can provide solid and detailed insights into the nature of L2 learners' interlanguage.

Figure 12: The research cycle (based on Mendikoetxea & Lozano, 2018)



IV Conclusion: empirical evidence, acquisition and teaching

We have departed from a classic classroom observation about personal pronouns and turned it into a research question and its corresponding hypothesis. Triangulation (i.e., the systematic combination of corpus and experimental methods) has provided insights into the knowledge of L2 Spanish learners' use of personal pronouns. The scientific knowledge attained up to here is useful for SLA researchers, who can understand how learners' competence (i.e., their knowledge) of the L2 is shaped and acquired. But L2 learners acquire linguistic knowledge mostly when they are cognitively and developmentally ready for it, as widely attested in the scientific literature – cf. the Teachability Hypothesis (Pienemann, 2005), which states that learners should not be taught what they are not developmentally ready to acquire. The recommendation for language teachers is to teach what is teachable (or, in other words, to teach what is acquirable). In our case, the recommendation for teachers is to implement teaching materials that, based on the preliminary results we have seen in this chapter, focus on how overt and null pronominal subjects are constrained by topic continuity vs. shift and by other factors (syntactic scenario). The bottom line is that teaching recommendations should rest on solid corpus and experimental evidence, bearing also in mind the well-known fact that learners are often capable of acquiring subtle and complex properties from the input that have never been taught.

V Appendices

V.1 Appendix 1 (CEDEL2 learner sample)

These are the L1 English-L2 Spanish learners' texts that have been tagged in section III.2. The name of the file is indicated between parentheses. Table 1 below summarises the tagging.

(25) *Presidente George W. Bush*$_i$ es persona famosa. **El**$_i$ es de Texas y **tiene**$_i$ un ranch. [Fue] Presidente para ocho anós y **vivo**$_i$ Washington, D.C. con esposa Laura. **El**$_i$ para sobre mundo y **habla**$_i$ mucho persona. El final anós es Presidente y el padre fue Presidente. **El**$_i$ [ha] hecho mucho para el pais las personas pueden pensar diferente. **El**$_i$ hizo mucho cuando el governor para Texas. **El**$_i$ [tiene] dos hijas ella gemelos y ella en anós viente y vaya a colegio. **El**$_i$ es Republican Presidente y before **el**$_i$ es Democrat presidente. **El**$_i$ parece viejo y **el**$_i$ [es] a chistoso tipo. **El**$_i$ mucho intellgente persona piense **el**$_i$ [es] estúpido **el**$_i$ no [lo es]. **El**$_i$ estuvo en war y nosotros en war que […] (*CEDEL2 corpus, beginner: EN_WR_12_18_2_2_MLS*)

(26) *Steve Irwin*$_i$ es una persona famosa de Australia. **El**$_i$ [es] hombre muy activo y **trabaja**$_i$ con grande y pequeño animales salvajes y el enviromente. **El**$_i$ es blanco. **El**$_i$ es muy guapo y blonde. **Sénior Irwin**$_i$ morir 4 de Septiembre. (*CEDEL2 corpus, beginner: EN_WR_13_38_1_2_JAR*)

(27) *Jennifer Lopez*$_i$ es cantante y actriz famosos. **Ella**$_i$ también va por el nombre de JLO. Sus padres son de Puerto Rico pero **ella**$_i$ nació en Nueva York. **Jeniffer**$_i$ fue a las escuelas católicas y **ruega**$_i$ regulary. **Ella**$_i$ comenzó su carrera mientras que un bailarín para "en color vivo" y **ella**$_i$ también bailó para Janet Jackson. **Ella**$_i$ ha aparecido en muchas películas y **ha**$_i$ registrado cuatro álbumes. **Jennifer**$_i$ tiene su propia línea de la ropa. **Jennifer**$_i$ es la primera actriz y cantante para tener una película y un álbum en el número uno de la misma semana. […] (*CEDEL2 corpus, beginner: EN_WR_15_30_1_2_TT*)

(28) *Denzel Washington*$_i$ es muy guapo. **Senor Washington**$_i$ te gusta pelicula. **Senor Washington**$_i$ esta alto y mucho **tiene**$_i$ deniro. **Senor Washington**$_i$ vive en Los Angeles California con el novia en mucho encasa. **Denzel Washington**$_i$ encanta football y basebol. **Washington**$_i$ Tiene tres ninos. el nino$_j$ camine Morehouse Universidad y **juege**$_j$ football. [...] (*CEDEL2 corpus, beginner: EN_WR_9_26_0_2_NLP*)

Table 1: Summary of the tagging of the learner corpus sample.

Text (L1 English-L2 Spanish)	**Null pronoun**	**Overt pronoun**	**N or NP**
EN_WR_12_18_2_2_MLS	3	13	0
EN_WR_13_38_1_2_JAR	1	3	1
EN_WR_15_30_1_2_TT	2	5	3
EN_WR_9_26_0_2_NLP	2	0	5
TOTAL count **TOTAL % (raw frequency)**	**8** **21.05% (8/38)**	**21** **55.26% (21/38)**	**9** **23.68% (9/38)**

V.2 Appendix 2 (CEDEL2 native sample)

These are the CEDEL2 native texts that have been used in the tagging in section III.2. Table 2 summarises the tagging.

(29) Como muchos saben, *Rafa Nadal*$_i$ es uno de los mejores tenistas del momento y quizás en un futuro, de la historia. ¶

Rafael Nadal$_i$ es un tenista de unos 24 años nacido en Manacor, un pueblo de las Islas Baleares. Ya **creció**$_i$ en un ambiente deportivo ya que su tío, "Nadal", fue un famoso jugador del futbol club Barcelona. **Rafa Nadal**$_i$ comenzó su carrera deportiva desde muy pequeño jugando al tenis y al fútbol. **Creció**$_i$ en un ambiente muy familiar: siempre **ha estado**$_i$ muy arropado por su familia y su entrenador es su tío. ¶

Desde pequeño **comenzó**$_i$ a despuntar en el tenis ganando incluso a contrincantes de categorías superiores a la suya aunque cuando **se dio**$_i$ a conocer para el gran público español y por extensión, extranjero, fue cuando **alcanzó**$_i$ por primera vez la final de Roland Garros. **Se convirtió**$_i$ en el jugador más joven en lograrlo y **consiguió**$_i$ volver a "renganchar" al público español al tenis, un deporte en el que siempre hemos tenidos grandes jugadores y ganadores pero del que nos habíamos despegado desde la retirada de Arantxa Sánchez-Vicario. […] (*CEDEL2 corpus, native: ES_WR_26_2_AVS*)

(30) Me gustaría hablar de una persona que me hace mucho reír cuando la veo en televisión: *Sofía Vergara*$_i$. **Nació**$_i$ en Colombia y **llegó**$_i$ a Estados Unidos para trabajar en el mundo de la televisión hace más de veinte años. No obstante, la serie que la catapultó a la fama fue Modern Family, en la que **interpreta**$_i$ a un personaje que se caracteriza mucho con su forma de ser, ya que **es**$_i$ una mujer colombiana, muy llamativa, con un fuerte acento colombiano al hablar en inglés y con un carácter fuerte. **Sofía Vergara**$_i$ tiene un hijo de unos veinte años y **se casó**$_i$ hace un año con otro famoso actor estadounidense. **Es**$_i$ actualmente una de las actrices mejor pagadas de Hollywood y constantemente se habla de ella en revistas, programas de televisión y redes sociales. Sin lugar a dudas su atractivo físico le ha

> ayudado a ganarse un hueco en televisión, pero sin su talento para el mundo de la televisión no **habría llegado**$_i$ a alcanzar la posición privilegiada de la que **goza**$_i$ en la actualidad. (*CEDEL2 corpus. native: ES_WR_28_2_EAP*)

Table 2: Summary of the tagging of the native corpus sample.

Text (Spanish natives)	**Null pronoun**	**Overt pronoun**	**N or NP**
ES_WR_26_2_AVS	8	0	1
ES_WR_28_2_EAP	8	0	1
TOTAL count	**16**	**0**	**2**
TOTAL % (raw frequency)	**88.89% (16/18)**	**0% (0/18)**	**11.11% (3/18)**

VI References

Alonso-Ramos, M. (Ed.). (2016). *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. John Benjamins. https://doi.org/10.1075/scl.78

Blackwell, S. E., & Quesada, M. L. (2012). Third-person subjects in native speakers' and L2 learners' narratives: Testing (and revising) the Givenness Hierarchy for Spanish. In K. L. Geeslin & M. Díaz-Campos (Eds.), *Selected Proceedings of the 14th Hispanic Linguistics Symposium* (pp. 142–164). Cascadilla Press. http://www.lingref.com/cpp/hls/14/paper2662.pdf

Blom, E., & Unsworth, S. (Eds.). (2010). *Experimental Methods in Language Acquisition*. John Benjamins.

Callies, M., & Paquot, M. (2015). Learner Corpus Research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research*, *1*(1), 1–6.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage.

Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards the future. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data.* (pp. 9–29). John Benjamins.

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 9–34). Cambridge University Press.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1–26.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Hawkins, R. (2001). *Second Language Syntax: A Generative Introduction*. Blackwell.

Hawkins, R. (2018). *How Second Languages are Learned: An Introduction*. Cambridge University Press. https://doi.org/10.1017/9781108565875

Ionin, T. (2012). Formal Theory-based Metodologies. In A. Mackey & S. M. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 30–52). Wiley-Blackwell.

Keating, G. D., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition, 37*, 1–32.

Larson-Hall, J. (2016). *A Guide to Doing Statistics in Second Language Research Using SPSS and R* (2nd ed.). Routledge. https://www.routledge.com/A-Guide-to-Doing-

Statistics-in-Second-Language-Research-Using-SPSS-and/Larson-Hall/p/book/9781315775661

Liceras, J. M. (1989). On some properties of the “pro-drop” parameter: Looking for missing subjects in non-native Spanish. In S. M. Gass & J. Schachter (Eds.), *Linguistic Perspectives on Second Language Acquisition* (pp. 109–133). Cambridge University Press.

Liceras, J. M. (1991). *La adquisición de las lenguas extranjeras: Hacia un modelo de análisis de la interlengua*. Visor.

Liceras, J. M., Zobl, H., & Goodluck, H. (Eds.). (2008). *The Role of Formal Features in Second Language Acquisition*. Lawrence Erlbaum Associates.

Lowie, W., & Seton, B. (2013). *Essential Statistics for Applied Linguistics*. Palgrave Macmillan. http://www.palgrave.com%2Fpage%2Fdetail%2Fessential-statistics-for-applied-linguistics-wander-lowie%2F%3FK%3D9780230304819

Lozano, C. (in prep.). CEDEL2: The design and creation of an online corpus for L2 Spanish acquisition research. *About to Be Submitted*.

Lozano, C. (2009a). CEDEL2: Corpus Escrito del Español como L2. In C. M. Bretones & et al (Eds.), *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente* (pp. 197–212). Universidad de Almería.

Lozano, C. (2009b). Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus. In Y.-I. Leung, N. Snape, & M. Sharwood-Smith (Eds.), *Representational Deficits in Second Language Acquisition* (pp. 127–166). John Benjamins. https://doi.org/10.1075/lald.47.09loz

Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: Advanced English learners of Spanish in the CEDEL2 corpus. In M.

Alonso Ramos (Ed.), *Spanish Learner Corpus Research: Current Trends and Future Perspectives* (pp. 235–265). John Benjamins. https://doi.org/10.1075/scl.78.09loz

Lozano, C. (2018). The development of anaphora resolution at the syntax-discourse interface: Pronominal subjects in Greek learners of Spanish. *Journal of Psycholinguistic Research*, *47*(2), 411–430. https://doi.org/10.1007/s10936-017-9541-8

Lozano, C. (2020a). Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2. In M. Cruz Piñol (Ed.), *E-Research y español LE/L2: Investigar en la era de las tecnologías*. Routledge.

Lozano, C. (2020b). Generative approaches. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora*. Routledge.

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. (pp. 65–100). John Benjamins. https://doi.org/10.1075/scl.59.06loz

Mackey, A., & Gass, S. M. (Eds.). (2012). *Research Methods in Second Language Acquisition: A Practical Guide*. Wiley-Blackwell.

Mackey, A., & Gass, S. M. (2016). *Second Language Research: Methodology and Design* (2nd ed.). Routledge.

Marinis, T. (2010). Using on-line processing methods in language acquisition research. In E. Bloom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition* (pp. 139–162). John Benjamins.

Martín-Villena, F., & Lozano, C. (2020). Anaphora resolution in topic continuity: Evidence from L1 English–L2 Spanish data in the CEDEL2 corpus. In J. Ryan & P. Crosthwaite (Eds.), *Referring in a Second Language: Studies on Reference to Person in a Multilingual World* (pp. 119–141). Routledge. http://doi.org/10.4324/9780429263972-7

Mas Álvarez, I., & Gil Martínez, A. (2018). Los corpus de aprendices: Un terreno en expansión para la enseñanza del español. In M. Ellison, M. Pazos Anido, P. Nicolás Martínez, & S. Valente Rodrigues (Eds.), *As línguas estrangeiras no ensino superior: Propostas didácticas e casos em estudo* (pp. 35–55). FLUP. https://sigarra.up.pt/flup/pt/pub_geral.pub_view?pi_pub_base_id=240721

Mendikoetxea, A. (2014). Corpus-based research in second language Spanish. In K. L. Geeslin (Ed.), *The Handbook of Spanish Second Language Acquisition* (pp. 11–29). Wiley-Blackwell.

Mendikoetxea, A., & Lozano, C. (2018). From corpora to experiments: Methodological triangulation in the study of word order at the interfaces in adult late bilinguals (L2 learners). *Journal of Psycholinguistic Research*, *47*(4), 871–898. https://doi.org/10.1007/s10936-018-9560-0

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, *21*(4), 373–391.

Myles, F. (2007). Using electronic corpora in SLA research. In D. Ayoun (Ed.), *Handbook of French Applied Linguistics* (pp. 377–400). John Benjamins.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 309–332). Cambridge University Press.

O'Donnell, M. (2009). The UAM CorpusTool: Software for corpus annotation and exploration. In C. M. Bretones & et al (Eds.), *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente* (pp. 1433–1447). Universidad de Almería.

Oh, S.-Y. (2006). English zero anaphora as an interactional resource II: *Discourse Studies*, *8*(6), 817–846. https://doi.org/10.1177/1461445606067332

Pérez-Leroux, A. T., & Glass, W. R. (1999). Null anaphora in Spanish second language acquisition: Probabilistic versus generative approaches. *Second Language Research*, *15*(2), 220–249.

Phakiti, A. (2014). *Experimental Research Methods in Language Learning*. Bloomsbury. https://www.bloomsbury.com/us/experimental-research-methods-in-language-learning-9781441122407/

Pienemann, M. (Ed.). (2005). *Cross-linguistic Aspects of Processability Theory*. John Benjamins.

Quesada, T., & Lozano, C. (accepted). Which factors determine the choice of referential expressions in L2 English discourse? A multifactorial study from the COREFL corpus. *Studies in Second Language Acquisition*.

Rankin, T. (2015). Learner corpora and grammar. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 231–254). Cambridge University Press.

Roberts, L. (2012). Psycholinguistic techniques and resources in second language acquisition research. *Second Language Research*, *28*(1), 113–127. https://doi.org/10.1177/0267658311418416

Roever, C., & Phakiti, A. (2017). *Quantitative Methods for Second Language Research: A Problem-Solving Approach*. Routledge. https://www.routledge.com/Quantitative-Methods-for-Second-Language-Research-A-Problem-Solving-Approach/Roever-Phakiti/p/book/9780415814027

Rose, H., McKinley, J., & Briggs Baffoe-Djan, J. (2019). *Data Collection Research Methods in Applied Linguistics*. Blomsbury. https://www.bloomsbury.com/uk/data-collection-research-methods-in-applied-linguistics-9781350025851/

Sánchez Rufat, A. (2017). Análisis contrastivo de interlengua y corpus de aprendientes: Precisiones metodológicas. *Pragmalingüística*, *23*, 191–210. https://doi.org/10.25267/Pragmalinguistica.2017.i25

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *X*(3), 209–231.

Selinker, L. (1991). *Rediscovering Interlanguage*. Addison Wesley Publishing Company.

Slabakova, R. (2016). *Second Language Acquisition*. Oxford University Press. https://global.oup.com/ukhe/product/second-language-acquisition-9780199687275?cc=us&lang=en&

Sorace, A. (1996). The use of acceptability judgments in second language acquisition research. In W. C. Ritchtie & T. K. Bhatia (Eds.), *Handbook of Second Language Acquisition*. Academic Press.

Sorace, A. (2016). Referring expressions and executive functions in bilingualism. *Linguistic Approaches to Bilingualism*, *6*(5), 669–684. https://doi.org/10.1075/lab.15055.sor

Spinner, P., & Gass, S. M. (2019). *Using Judgments in Second Language Acquisition Research*. Routledge. https://www.crcpress.com/Using-Judgments-in-Second-Language-Acquisition-Research/Spinner-Gass/p/book/9781138207035

White, L. (1989). *Universal Grammar and Second Language Acquisition*. John Benjamins.

White, L. (2003). *Second Language Acquisition and Universal Grammar*. Cambridge University Press.